Université Hassan II de Casablanca
Faculté des sciences Ben M'sik
Département de Mathématiques et Informatique
Filière SMI

THÈME

Machine-Learning pour la prédiction de l'espérance de vie postopératoire des patients atteints d'un cancer des poumons

Présenté par :

- Mlle Nouhaila MOUSSAMMI
- Mlle Chaimaa M'SALA
- Mlle Amina MONTASSIR

Soutenu Le 10-06-2019 devant les membres de jury :

Mr Abdelwahed NAMIR	: Encadrant
Mme Fatima TAIF	: CO- Encadrante
Mr Saïd NOUH	: Examinateur
Mr Mohammed AIT DAOUD	: Examinateur

Remerciements

Nous tenons à remercier toutes les personnes qui ont contribué au succès de notre projet fin d'études et qui nous ont aidées lors de la rédaction de ce mémoire.

Nous voudrions dans un premier temps remercier, notre directeur de mémoire Mr NAMIR, notre cher encadrant et professeur à l'université Hassan II Casablanca faculté des sciences Ben M'SIK, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter notre réflexion.

Nous remercions également notre chère Co-encadrante Taif Fatima pour son aide à la réalisation de ce travail ou elle a partagé ses connaissances et ses expériences dans ce milieu, tout en nous accordant sa confiance et une large indépendance dans l'exécution de missions valorisantes.

Nos parents, pour leurs soutiens constants et leurs encouragements.

Table des matières

Liste des figures

Liste des tableaux

Résumé

L’application de le machine Learning à la médecine offre une perspective essentielle à l’essor de ces nouvelles technologies, qu’il s’agisse de renforcer le lien entre patients et médecins, de poser des diagnostics plus rapides et plus précis, ou encore d’optimiser la création de nouveaux traitements. L’innovation a pour objectif de combattre la mort et la maladie. Quoi de plus noble ? L'ML permet aux médecins de gagner du temps en laissant la machine analyser elle-même les données et fournir des estimations. Le but à plus long terme : réussir à prédire de nombreuses maladies, plus précisément au suivi des cancers, afin que les médecins puissent intervenir le plus tôt possible.

Ce document décrit les travaux réalisés dans le cadre du projet PFE. L’objet de ce projet est de fournir des outils pour prédire l’espérance de vie postopératoire des patients atteints de cancer du poumon à l’aide des méthodes de ML En utilisant la régression logistique.

Mot clés :

Prédiction, cancer, machine Learning, régression logistique.

Introduction

Le cancer est une maladie très répandue (Chez les hommes, le cancer des poumons est en tête de liste avec 22% des cas diagnostiqués) et critique (c'est la première cause de décès avec 149 000 décès) mais qui donne lieu à des innovations considérables.

Le retrait de la tumeur via la chirurgie ne présente un bénéfice que pour les patients dont le cancer du poumon est à un stade précoce. Parfois, il demeure possible d'effectuer une intervention chirurgicale si le cancer ne s'est propagé qu'aux ganglions lymphatiques avoisinants. Certains patients peuvent recevoir une chimiothérapie après une intervention chirurgicalev (chimiothérapie adjuvante) pour réduire le risque de récidive du cancer.

La chirurgie n'est pas utile si le cancer s'est propagé à l'autre poumon ou au-delà du thorax, dans une autre partie du corps (métastasé), car une intervention, sur un site, n'empêchera pas le cancer de se développer sur les autres sites du corps où il s'est déjà propagé. Retirer une partie du cancer n'est pas utile, car la tumeur restante continue de se développer.

Par ailleurs, il est difficile pour des chirurgiens de prédiction la durée de vie d'un patient si ce dernier est atteint d'un cancer du poumon après une intervention chirurgicale.

L'utilisation du machine Learning ML dans le domaine médical a pour ambition de lever cette difficulté.

Le Machine Learning est une branche de l'intelligence artificielle. Ce domaine de l'informatique vise à développer des algorithmes permettant de modéliser les données. Les statistiques classiques sont quant à elles un domaine de l'analyse en mathématiques. Elles permettent de rejeter une hypothèse, au profit d'une hypothèse alternative. Appliqué au domaine de la santé, et plus précisément au suivi des cancers, le Machine Learning peut offrir de nouvelles possibilités aux chirurgiens pour le suivi de l'évolution de la maladie.

Le but de notre travail était de concevoir un système d'aide aux chirurgiens. Pour ce faire, un outil a été développé pour prédire l'espérance de vie postopératoire des patients atteints de cancer du poumon à l'aide des méthodes de ML. La méthode appliquée dans ce projet est : la régression logistique. Cette méthode a été utilisée spécifiquement pour prédire si un patient atteint du cancer du poumon survivra un an après une chirurgie thoracique. Les résultats de

cette technique ont ensuite été mesurés et comparés en fonction de leur précision et de leurs performances.

Le manuscrit est organisé en trois principaux chapitres en plus de l'introduction générale.

Le 1er chapitre présente un aperçu général sur le cancer en citant les causes, le diagnostic, les traitements et la prévention.

Le 2ème chapitre résume l'essentiel du Machine Learning dans les différents domaines toutes en donnant certains exemples de son utilisation soit pour la prédiction ou pour la réalisation d'autres taches ou il faut absolument passer par plusieurs étapes en commençant par la précision des données.

Le 3ème chapitre cite le résultat graphiquement de l'ensemble de personnes vivants et morts après un an de chirurgie concernant le cancer de poumons en utilisant Machine Learning précisément la technique de la régression logistique.

Chapitre 1 : Généralités sur le cancer

Résumé

Dans ce chapitre nous verrons le concept de cancer et ses types en se concentrant sur ce des poumons. Le cancer du poumon prend naissance dans les cellules du poumon. La tumeur cancéreuse (maligne) est un groupe de cellules cancéreuses qui peuvent envahir et détruire le tissu voisin. Elle peut aussi se propager (métastases) à d'autres parties du corps. Quand le cancer débute dans les cellules du poumon, il est appelé cancer primitif du poumon. Les cancers du poumon sont divisés en cancer du poumon non à petites cellules et en cancer du poumon à petites cellules selon le type de cellule à partir duquel ils se développent. Les facteurs de risque du cancer du poumon sont classés en deux types : facteurs de risque connus (tabagisme, radon, amiante, fumée secondaire et pollution de l'air extérieur..) et facteurs de risque possibles (exposition professionnelle à certaines substances chimiques, mutation génétiques...)

Introduction :

Un homme sur cinq et une femme sur six dans le monde développeront un cancer au cours de leur vie, et un homme sur huit et une femme sur 11 meurent de cette maladie. C'est ce qu'affirme le Centre international de recherche sur le cancer (CIRC) – un organisme relevant de l'Organisation mondiale de la santé (OMS) – dans un communiqué de presse rendu public le 12 septembre. Une déclaration réalisée à l'occasion de la publication des dernières données mondiales sur le cancer fournissant des estimations de l'incidence et de la mortalité dans 185 pays et pour 36 types de cancers ainsi que pour tous les sites de cancers combinés **[1].**

Au Maroc, le taux d'incidence du cancer est de 139,6 cas par 100.000, et le taux de mortalité est estimé à 86,9 cas par 100.000. Les Marocains ont, selon le CIRC, 14,67% de probabilité de contracter un cancer avant l'âge de 75 ans, et 9,28% de risque d'en mourir avant le même âge. En 2018, 52.783 nouveaux cas de cancer ont été recensés dans le Royaume. Les cancers du sein (10.136, 20,73%), du poumon (6.488, 13,27%) et de la prostate (3.990, 8,16%) étant les plus répandus. En revanche, 32.962 de décès ont été comptabilisés. Les cancers du poumon (6.937), du sein (3.518) et du col de l'utérus **(2.465) se révèlent être les plus mortels.**

En tout et pour tout, le CIRC estime le fardeau mondial du cancer à 18,1 millions de nouveaux cas et 9,6 millions de décès en 2018. Cette augmentation est, selon la même source, due à plusieurs facteurs, notamment à « la croissance démographique et le vieillissement, ainsi qu'à l'évolution de la prévalence de certaines causes de cancer associées au développement social et économique ». Ainsi, le CIRC observe dans les économies à croissance rapide « une évolution des cancers liés à la pauvreté et aux infections vers des cancers associés aux modes de vie plus typiques des pays industrialisés » **[2].**

Les cancers du poumon, du sein et du côlon-rectum restent les trois types de cancer touchant le plus de personnes dans le monde, et figurent parmi les cinq les plus mortels – respectivement premier, cinquième et deuxième -. « Pris ensemble, ces trois types de cancer sont responsables d'un tiers de l'incidence du cancer et de la mortalité dans le monde » , note le communiqué.

I. Définition du Cancer :

Maladie provoquée par la transformation de cellules qui deviennent anormales et prolifèrent de façon excessive. Ces cellules déréglées finissent par former une masse qu'on appelle tumeur maligne. Les cellules cancéreuses ont tendance à envahir les tissus voisins et à se détacher de la tumeur. Elles migrent alors par les vaisseaux sanguins et les vaisseaux lymphatiques pour aller former une autre tumeur (métastase).

1. Les types du cancer

Il y a Plus de 20 types de cancers recensés :

- Nasopharynx ou cavum
- oesophage
- estomac
- colon
- rectum
- foie
- vésicule et voies biliaires
- Pancréas
- larynx
- **Poumon**
- sein
- col utérin
- corps
- utérin

- ovaires
- prostate
- vessie
- rein
- system nerveux central
- thyroïde
- hémopathies malignes (lymphome hodgkinien, lymphome non hodgkinien, leucémie…)
- peau, cerveau, primitif…

2. Statistiques

D'après ces statistiques, on voit bien qu'au Maroc les hommes entre 50 et 69 ans sont les plus touchés ainsi que les femmes entre 40 et 59 ans, on peut dire alors que ça a une relation directe avec l'âge (La vieillesse) , D'où la possibilités de l'élimination de nombres des personnes qui souffrent de cette maladie est possible , tout d'abord en évitant les causes possibles (pendant sa jeunesse) ou bien de choisir le bon traitement, le choix du bon traitement dépend bien des cas et de plusieurs conditions où ce n'est pas facile de tomber sur le traitement convenable directement et par hasard , pour cela il faut absolument étudier les données et passer par la prédiction, on va clairement parler dans ce rapport des différentes méthodes et techniques de prédiction **[3].**

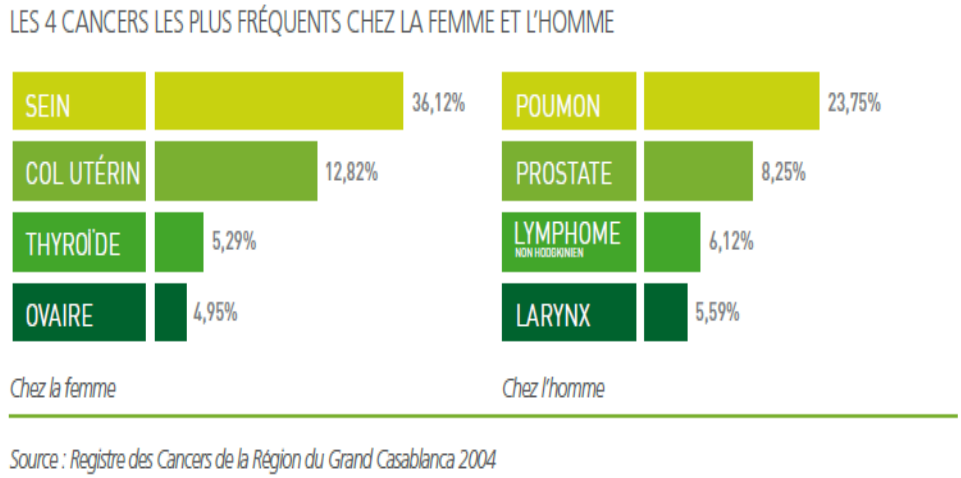


Figure 1: Les cancers les plus fréquents chez les hommes et les femmes

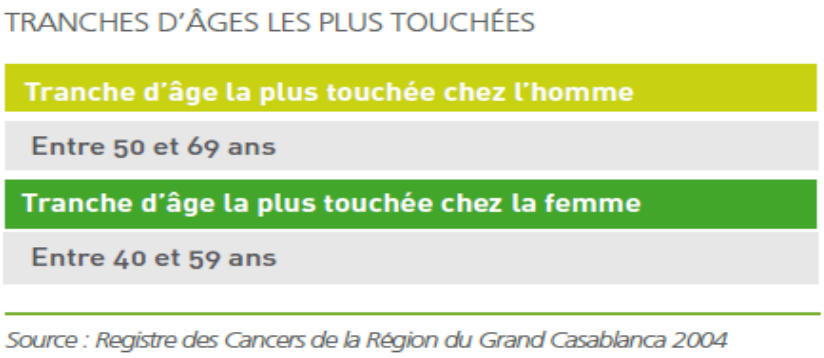


Figure 2 : la tronche d'âge les plus touchées [4]

//Ces statistiques montrent bien que le cancer des seins est le plus fréquenté chez les femmes d'un pourcentage de 36,12%, cependant, on trouve que chez les hommes le cancer des poumons prend le premier classement d'un pourcentage de 23,75% , ça nous encourage de faire plus de recherches qui permettent l'élimination de cette terrible maladie et de trouver plutôt le traitement efficace surtout de ces deux genres de cancer car ce sont les plus répandus.

- **Dans ce rapport, on va s'intéresser par l'étude de celui des poumons.**

II. Définition du cancer du poumon :

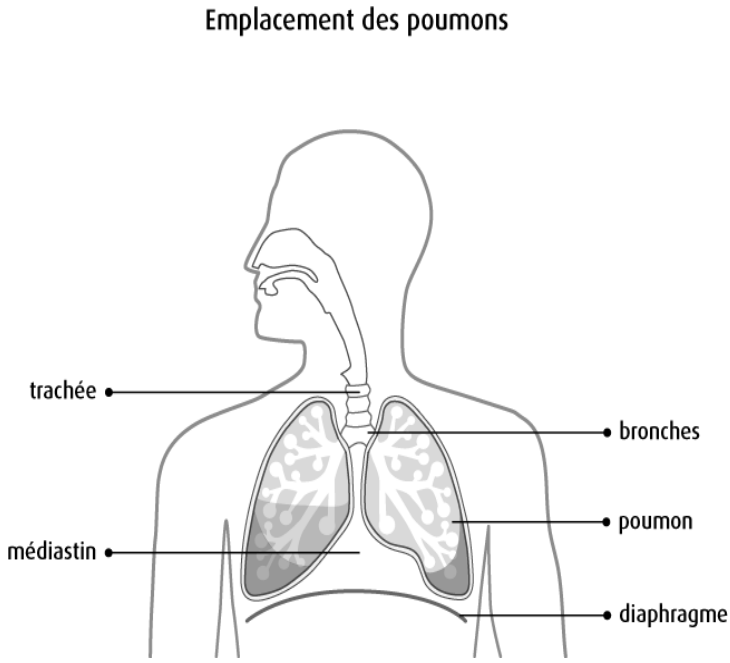
Le cancer du poumon prend naissance dans les cellules du poumon. La tumeur cancéreuse (maligne) est un groupe de cellules cancéreuses qui peuvent envahir et détruire le tissu voisin. Elle peut aussi se propager (métastases) à d'autres parties du corps. Quand le cancer débute dans les cellules du poumon, il est appelé cancer primitif du poumon **[5].**

Le poumon fait partie de l'appareil respiratoire. Vous utilisez vos poumons quand vous respirez. Les poumons sont situés dans le thorax, de chaque côté du cœur. Le poumon droit est constitué de 3 sections principales appelées lobes. Le poumon gauche, un peu plus petit, comporte 2 lobes. Les poumons sont enveloppés d'une fine membrane protectrice appelée plèvre, qui constitue un genre de coussin pour ces organes.

Les cellules du poumon subissent parfois des changements qui rendent leur mode de croissance ou leur comportement anormal. Ces changements peuvent engendrer la formation de tumeurs non cancéreuses (bénignes), comme l'hamartome et le papillome. Mais dans certains cas, les changements qui se produisent dans les cellules pulmonaires peuvent causer le cancer.

Les cancers du poumon sont divisés en cancer du poumon non à petites cellules et en cancer du poumon à petites cellules selon le type de cellule à partir duquel ils se développent.

Le **cancer du poumon non à petites cellules** prend habituellement naissance dans les cellules glandulaires situées dans la partie externe du poumon. Ce type de cancer porte le nom d'adénocarcinome. Le cancer du poumon non à petites cellules peut aussi prendre naissance dans les cellules minces et plates appelées cellules squameuses. Celles-ci tapissent les bronches qui sont les grosses voies respiratoires se ramifiant de la trachée jusqu'aux poumons. On parle alors d'un carcinome épidermoïde du poumon. Le carcinome à grandes cellules est un autre type de cancer du poumon non à petites cellules, mais il est moins fréquent. Il existe également plusieurs types rares de cancer du poumon non à petites cellules, dont le sarcome et le carcinome sarcomatoïde.



Le **cancer du poumon à petites cellules** prend habituellement naissance dans les cellules qui tapissent les bronches situées au centre des poumons. Les principaux types de cancer du poumon à petites cellules sont le carcinome à petites cellules et le carcinome mixte à petites cellules (tumeur mixte formée entre autres de cellules squameuses ou glandulaires) **[6].**

D'autres types de cancer peuvent se propager au poumon, mais il s'agit alors d'une maladie différente du cancer primitif du poumon. Le cancer qui prend naissance dans une autre partie du corps et qui se propage au poumon est appelé métastase pulmonaire ; celle-ci est traitée

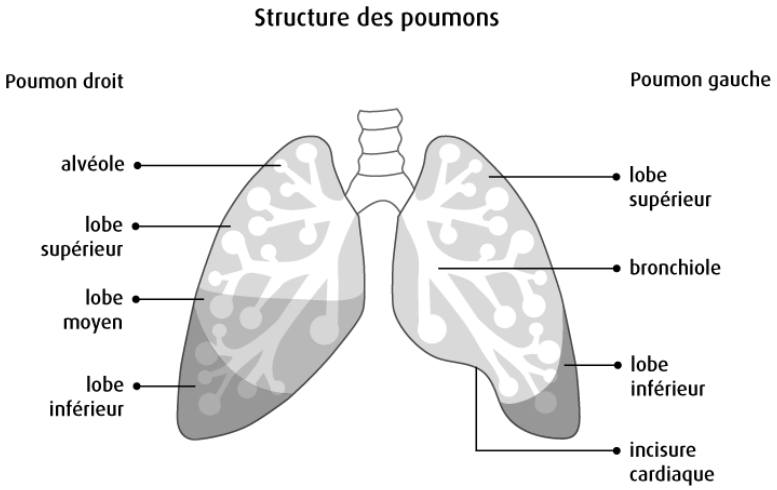
différemment du cancer primitif du poumon. Apprenez-en davantage sur les métastases pulmonaires.

Un type rare de cancer, le mésothéliome pleural, est souvent appelé à tort cancer du poumon. Bien que le mésothéliome se forme dans la plèvre qui recouvre le poumon, il est très différent d'un cancer qui prend naissance dans le poumon.

1) Les poumons :

Les poumons sont situés dans le thorax et font partie de l'appareil respiratoire.

Les poumons prennent presque tout l'espace à l'intérieur du thorax. Ils sont entourés de la paroi thoracique, qui est composée des côtes et des muscles entre les côtes. Les poumons sont séparés par le médiastin, qui contient le cœur et d'autres organes. Sous les poumons se trouve le diaphragme, un muscle mince qui sépare la cavité thoracique de l'abdomen.



Chaque poumon est divisé en lobes (sections).

- Le poumon gauche a 2 lobes. Le cœur s'appuie sur un creux (incisure cardiaque) dans le lobe inférieur.
- Le poumon droit a 3 lobes et est légèrement plus large que le poumon gauche.

La trachée est la voie respiratoire en forme de tube qui se trouve dans le cou et le thorax. Elle se divise en 2 tubes, ou branches, appelés bronches souches. Chacune entre dans un poumon. La région où chaque bronche entre dans le poumon est appelée hile.

La plèvre est une membrane fine qui recouvre les poumons et tapisse la paroi thoracique. Elle protège les poumons et forme tout autour un coussin. La plèvre produit un liquide qui agit comme un lubrifiant afin que les poumons puissent bouger sans problème dans la cavité thoracique. La plèvre est constituée de 2 couches:

- couche interne (plèvre viscérale) – couche entourant le poumon
- couche externe (plèvre pariétale) – couche qui tapisse la paroi thoracique

La région qui se trouve entre ces 2 couches est appelée espace pleural.

Chaque bronche souche se divise en bronches plus petites, dont les parois contiennent des glandes menues et du cartilage. Ces bronches plus petites se divisent en tubes encore plus petits appelés bronchioles, qui n'ont ni glandes ni cartilage. À l'extrémité des bronchioles se trouvent des millions de sacs minuscules appelés alvéoles. Autour des alvéoles se trouvent de très minuscules vaisseaux sanguins (capillaires).

Les bronches sont tapissées de cellules présentant des prolongements semblables à des cheveux très fins appelés cils.

Les poumons produisent un mélange de matières grasses et de protéines appelé surfactant pulmonaire. Le surfactant recouvre la surface des alvéoles, ce qui rend leur expansion et leur contraction plus facile à chaque inspiration/expiration.

Différents groupes de ganglions lymphatiques, qui font partie du système lymphatique, évacuent le liquide normalement produit dans les poumons :

- ganglions bronchiques – autour des bronches souches
- ganglions hilaires – dans la région où la trachée se divise en bronches souches
- ganglions médiastinaux – le long de la trachée, entre les 2 poumons
- ganglions médiastinaux sous-carénaires – juste sous la trachée, là où elle se divise en bronches souches

2. Facteurs de risque du cancer du poumon

Un facteur de risque est quelque chose, comme un comportement, une substance ou un état, qui accroît le risque d'apparition d'un cancer. La plupart des cancers sont attribuables à de nombreux facteurs de risque. Fumer du tabac est le plus important facteur de risque du cancer du poumon.

Le risque d'être atteint d'un cancer du poumon augmente avec l'âge. Plus de la moitié de tous les nouveaux cas de cancer du poumon sont diagnostiqués chez des personnes âgées de 60 ans ou plus. Les hommes sont atteints de ce cancer légèrement plus souvent que les femmes.

Les facteurs de risque sont habituellement classés du plus important au moins important. Mais dans la plupart des cas, il est impossible de les classer avec une certitude absolue.

Facteurs de risque connus	Facteurs de risque possibles
Tabagisme Fumée secondaire Radon Amiante Exposition professionnelle à certaines substances chimiques Pollution de l'air extérieur Antécédents personnels ou familiaux de cancer du poumon Antécédents personnels d'affection pulmonaire Exposition à la radiation Arsenic dans l'eau potable Polluants issus de la cuisson et du chauffage Système immunitaire affaibli Lupus Prise de suppléments de bêta-carotène chez les fumeurs	Exposition professionnelle à certaines substances chimiques Mutations génétiques Fumer du cannabis Usage de la marijuana Inactivité physique Alimentation faible en fruits et légumes

Tableau 1 : les facteurs de risque

2.1. Facteurs de risque connus

Des preuves convaincantes permettent d'affirmer que les facteurs suivants font augmenter votre risque de cancer du poumon.

- ✓ ***Tabagisme***

Fumer du tabac, en particulier sous forme de cigarette, est la principale cause du cancer du poumon. La fumée de tabac contient de nombreuses substances chimiques. Certaines de ces substances sont carcinogènes, ce qui signifie qu'elles causent des changements génétiques dans les cellules pulmonaires qui mènent à l'apparition d'un cancer du poumon.

Plus de 85 % des cas de cancer du poumon au Canada sont liés au tabagisme. Le risque d'être atteint d'un cancer du poumon augmente en fonction de la durée du tabagisme, de l'âge auquel vous avez commencé à fumer et du nombre de cigarettes fumées chaque jour. Le risque est aussi plus élevé si vous fumez et présentez d'autres facteurs de risque.

La pipe, le cigare, les cigarettes aux herbes, le houka, le tabac à priser ainsi que les cigarettes à faible teneur en goudron et les cigarettes à faible teneur en nicotine engendrent aussi le cancer et ne sont pas considérés comme inoffensifs.

- ✓ ***Fumée secondaire***

La fumée secondaire est celle qui est expirée par les fumeurs et celle qui se répand dans l'air depuis l'extrémité d'une cigarette, d'une pipe ou d'un cigare allumé. La fumée secondaire est aussi appelée fumée de tabac ambiante (FTA). L'inhalation de fumée secondaire est appelée tabagisme passif ou inhalation involontaire de fumée.

Aucune exposition à la fumée secondaire ne peut être considérée comme sécuritaire. La fumée secondaire contient les mêmes substances chimiques que la fumée directement inhalée. Les personnes qui sont exposées à la fumée secondaire risquent davantage d'être atteintes d'un cancer du poumon. La fumée secondaire est un important facteur de risque du cancer du poumon chez les non-fumeurs.

- ✓ ***Radon***

Le radon est un gaz incolore, inodore et sans goût qui provient de la désintégration naturelle de l'uranium présent dans le sol et la pierre. Il n'y a habituellement pas lieu de s'inquiéter du radon quand on est à l'extérieur, car il est dilué dans l'atmosphère. Cependant, le radon peut s'infiltrer dans un bâtiment par des planchers en terre battue ou des fissures dans les fondations. Il peut ainsi atteindre des niveaux dangereux dans des espaces fermés ou peu aérés. L'inhalation du radon risque d'endommager les cellules qui tapissent les poumons.

L'exposition au radon fait augmenter le risque de cancer du poumon. Le radon est la principale cause de cancer du poumon chez les non-fumeurs, et la deuxième plus importante cause de ce cancer chez les fumeurs.

Le risque d'apparition du cancer du poumon dépend de la quantité de radon à laquelle vous êtes exposé et de la durée de l'exposition. Le risque de cancer du poumon engendré par le radon est beaucoup plus élevé chez les fumeurs que chez les non-fumeurs.

- ✓ ***Amiante***

L'amiante est le nom donné à un groupe de minéraux naturels. Ces minéraux peuvent être séparés en longues fibres minces qui sont très fines. Lorsqu'une personne inhale ces fibres, celles-ci peuvent être emprisonnées dans les poumons.

L'amiante a été très utilisé dans les matériaux de construction et de nombreuses industries. Les personnes qui sont le plus à risque d'être exposées à l'amiante sont entre autres les suivantes :

- Travailleurs de mines d'amiante
- Travailleurs de l'industrie de l'automobile, dont les réparateurs de freins et d'embrayages
- Travailleurs de chantiers navals
- Travailleurs de cimenteries
- Travailleurs dans le domaine de la plomberie et du chauffage
- Travailleurs de la construction, peintres, charpentiers et électriciens

Des études montrent que les fumeurs qui sont exposés à l'amiante risquent encore plus d'être un jour atteints d'un cancer du poumon.

- ✓ ***Exposition professionnelle à certaines substances chimiques***

Certaines substances chimiques sont carcinogènes, ce qui signifie qu'elles causent le cancer. Ces substances chimiques peuvent causer le cancer du poumon chez les personnes qui sont exposées à ces substances en milieu de travail. En général, le risque de cancer du poumon est encore plus élevé chez les fumeurs qui sont exposés à ces substances chimiques.

Une exposition aux substances chimiques suivantes en milieu de travail engendre une augmentation du risque de cancer du poumon :

- Arsenic et composés inorganiques d'arsenic
- Béryllium et composés de béryllium
- Cadmium et composés de cadmium
- Substances chimiques utilisées dans la fabrication du caoutchouc, dans les fonderies de fer et d'acier et dans la peinture
- Éther chlorométhylique et éther bis(chlorométhylique)
- Composés de chrome (VI)
- Gaz d'échappement de moteur diésel

- Gaz moutarde
- Minerais radioactifs comme l'uranium et le plutonium
- Poudre de silice et silice cristalline
- Certains composés de nickel
- Certains types d'hydrocarbures aromatiques polycycliques (HAP)
- Bitume utilisé dans les toitures
- Cobalt-carbure de tungstène
- Fumées de soudage

Certaines industries utilisent de nombreuses substances chimiques, c'est pourquoi il est difficile pour les chercheurs de savoir lesquelles parmi ces substances font augmenter le risque de cancer du poumon. Les personnes qui travaillent dans les industries suivantes présentent le plus grand risque d'être atteintes d'un cancer du poumon :

- Fabrication du caoutchouc
- Fonderies de fer et d'acier
- Gazéification du charbon
- Production de coke
- Ramonage de cheminée
- Peinture commerciale
- Couverture de toit et asphaltage
- Production de carbure de silicium à partir du procédé Acheson

✓ ***Pollution de l'air extérieur***

La pollution de l'air, ce sont les substances chimiques, les particules et autres éléments présents dans l'air en quantités susceptibles de causer du tort à l'environnement ou de nuire à la santé ou au confort des humains, des animaux et des plantes. Les types de polluants présents dans l'air varient d'un endroit à l'autre selon les sources d'émissions locales. Les émissions peuvent aussi provenir d'autres régions.

On a des preuves solides pour dire que l'exposition à la pollution de l'air extérieur cause le cancer du poumon. Plus votre exposition à la pollution atmosphérique est grande, plus votre risque est grand d'être atteint d'un cancer du poumon.

La recherche démontre que les différents composants de la pollution de l'air extérieur causent le cancer. Ces composants sont entre autres les gaz d'échappement des moteurs au diésel, le benzène, la matière particulaire et certains hydrocarbures aromatiques polycycliques (HAP).

- ✓ ***Antécédents personnels ou familiaux de cancer du poumon***

Les personnes qui ont déjà été atteintes d'un cancer du poumon risquent davantage de développer un autre cancer du poumon. Vous pourriez aussi présenter un risque légèrement plus élevé de cancer du poumon si vous avez un parent au premier degré (frère, sœur, enfant, mère ou père) qui a déjà été atteint d'un cancer du poumon. Cette hausse du risque pourrait être attribuable à un certain nombre de facteurs, dont des habitudes de vie communes (comme le tabagisme) ou le fait de vivre dans un même endroit où il y a des carcinogènes (comme le radon).

- ✓ ***Antécédents personnels d'affection pulmonaire***

Certaines affections ou maladies pulmonaires peuvent laisser des cicatrices aux poumons et faire augmenter le risque d'apparition d'un cancer du poumon. Ce sont entre autres celles qui suivant:

- Maladie pulmonaire obstructive chronique (MPOC), qui se caractérise par des dommages à long terme aux poumons et qui est souvent causée par le tabagisme
- Tuberculose, qui est une infection pulmonaire causée par le bacille de la tuberculose
- Infection pulmonaire causée par *Chlamydophila pneumoniae*

- ✓ ***Exposition à la radiation***

Les personnes ayant reçu une radiothérapie au thorax pour certains cancers, comme un lymphome hodgkinien ou un cancer du sein, risquent davantage d'être atteintes d'un cancer du poumon. Ces personnes présentent un risque encore plus élevé si elles fument.

Les personnes qui ont été exposées aux rayonnements ionisants lors d'explosions de bombes atomiques ou d'accidents nucléaires risquent davantage d'être atteintes d'un cancer du poumon.

- ✓ ***Arsenic dans l'eau potable***

L'arsenic peut s'infiltrer dans l'eau potable à partir de sources naturelles dans le sol ou de certains types d'industries, comme l'exploitation minière. Bien que les experts ne comprennent pas totalement comment l'arsenic cause des changements aux cellules, les études de nombreuses parties du monde montrent que des taux élevés d'arsenic dans l'eau potable font augmenter le risque de cancer du poumon. Le risque est encore plus élevé chez les personnes qui fument.

- ✓ ***Polluants issus de la cuisson et du chauffage***

Certains types d'appareils de cuisson et de chauffage peuvent libérer des polluants qui font augmenter le risque de cancer du poumon. Les taux de ces polluants peuvent être très élevés dans des espaces où il y a une mauvaise circulation d'air.

Faire brûler du charbon à l'intérieur pour cuisiner ou chauffer a surtout été lié au cancer du poumon. La combustion de bois et d'autres combustibles, comme le fumier ou l'herbe, et la friture d'aliments dans l'huile à température élevée peuvent aussi accroître le risque de cancer du poumon.

- ✓ ***Système immunitaire affaibli***

L'infection au VIH et le sida peuvent affaiblir le système immunitaire. Les personnes atteintes du VIH ou du sida risquent davantage de développer plusieurs types de cancer, dont le cancer du poumon.

Les personnes qui ont eu une greffe d'organe prennent des médicaments pour freiner leur système immunitaire afin que le corps ne rejette pas l'organe greffé. Le fait d'avoir un système immunitaire affaibli fait augmenter le risque de cancer du poumon.

- ✓ ***Lupus***

Le lupus érythémateux disséminé (LED), ou simplement lupus, est une maladie auto-immune. Il peut affecter diverses parties du corps et causer l'inflammation de la peau, des articulations, des vaisseaux sanguins, du système nerveux et des organes internes comme le cœur, les poumons et les reins. Les personnes atteintes de lupus risquent davantage de développer un cancer du poumon.

- ✓ ***Prise de suppléments de bêta-carotène chez les fumeurs***

Le bêta-carotène est un type d'antioxydant. Certains essais cliniques ont révélé que les personnes qui ont fumé plus d'un paquet de cigarettes par jour et qui prenaient des

suppléments à haute dose de bêta-carotène présentent un risque plus élevé de cancer du poumon.

2.2. Facteurs de risque possibles

On a établi un lien entre les facteurs qui suivent et le cancer du poumon, mais on ne possède pas suffisamment de preuves pour dire qu'ils sont des facteurs de risque connus. On doit faire plus de recherches pour clarifier le rôle de ces facteurs dans le développement du cancer du poumon.

- ✓ ***Exposition professionnelle à certaines substances chimiques***

Les chercheurs tentent de savoir si les produits chimiques suivants font augmenter le risque de cancer du poumon :

- Bitume utilisé pour l'asphaltage
- Dioxine utilisée dans les pesticides
- Brumes d'acide chimique fort
- Carbure de silicium fibreux

- ✓ ***Mutations génétiques***

La recherche démontre que certaines familles ont des antécédents importants de cancer du poumon, ce qui peut signifier qu'elles présentent une mutation dans un certain gène pouvant causer le cancer du poumon. Les chercheurs tentent de savoir si un gène ou des gènes en particulier pourraient faire augmenter le risque de cancer du poumon. Certains chercheurs tentent aussi de trouver de très petits changements dans des gènes (polymorphisme génétique), qui pourraient faire augmenter le risque de cancer du poumon ou rendre les personnes, tout particulièrement les non-fumeurs, plus sensibles aux risques connus de cancer du poumon.

- ✓ ***Fumer du cannabis***

Les éléments de preuve qui laissent croire à la présence d'un lien entre l'usage du cannabis (marijuana) à long terme et le cancer du poumon ne sont pas aussi solides ou nombreux que ceux qui associent le tabagisme au cancer. Lors de certaines études, on a constaté qu'un usage récréatif prolongé du cannabis peut accroître le risque de cancer du poumon.

- ✓ ***Inactivité physique***

La recherche laisse entendre que les personnes qui ne pratiquent pas d'activités physiques peuvent présenter un risque accru de cancer du poumon, que ces personnes fument ou non.

- ✓ ***Alimentation faible en fruits et légumes***

Certaines études indiquent que les personnes dont l'alimentation est riche en fruits et légumes présentent un risque moins élevé de cancer du poumon.

2.3.Facteurs de risque inconnus

On ne sait pas s'il y a un lien entre les facteurs qui suivent et le cancer du poumon. C'est peut-être parce que les chercheurs ne parviennent pas à établir définitivement ce lien ou que les études ont engendré différents résultats. On doit faire plus de recherches afin de savoir si les éléments suivants sont des facteurs de risque du cancer du poumon :

- ✓ Exposition professionnelle à des fibres synthétiques (comme la laine de verre)
- ✓ Exposition professionnelle au chlorure de vinyle
- ✓ Arthrite rhumatoïde

Les chercheurs se penchent sur le rôle que pourrait jouer, chez les femmes, l'œstrogène dans l'apparition du cancer du poumon. Des études démontrent que des facteurs reproducteurs, comme le nombre d'enfants auquel une femme a donné naissance, l'âge auquel une femme a eu sa ménopause ou une ablation des ovaires, peuvent accroître le risque d'une femme d'être atteinte du cancer du poumon.

Conclusion

Nous avons vu dans ce chapitre les types de cancer en se concentrant sur ce des poumons, les différents traitements de ce dernier 'cancer des poumons' et tests ainsi que les complications dues à cette maladie. Même s'il existe des méthodes de prévention qui permettent de réduire le risque d'avoir ce genre de cancer, parfois il est impossible de l'éviter ou de trouver la bonne solution de traitement direct. Dans ces cas-là, la seule solution est de passer par une opération le plus tôt possible et faire tout son possible pour combattre les complications.

Chapitre 2 : Machine Learning (ML)

Résumé

Nous allons développer dans ce chapitre les principaux axes de connaissance ci-dessous sur lesquels on a travaillé pour structurer nos recherches et nécessaires pour mieux comprendre le projet et ses enjeux techniques :

- Présentation et domaines d'application du machine learning
- les differents types d'apprentissage
- la préparation des données
- les modèles générés par les algorithmes d'apprentissage et leur évaluation

Introduction

Le **Machine Learning** est un ensemble de techniques utilisées par les Data Scientists, qui a grandement fait parler de lui ces dernières années. Car ses applications sont variées et très prometteuses !

Une fois que le Data Scientist a effectué son travail de collecte, de nettoyage et d'exploration des données, il peut passer à la partie "**modélisation"**. C'est ce processus que nous allons explorer ensemble dans ce cours d'initiation au Machine Learning.

Vous allez découvrir un ensemble de techniques puissantes permettant de créer des modèles **prédictifs** à partir de données, **et qui apprennent par eux-mêmes !**

I. Présentation et domaines d'application

Le machine learning (ML), ou apprentissage automatisé, est un champ de l'intelligence artificielle. Il peut être défini comme couvrant l'étude, la conception et le développement d'algorithmes donnant la possibilité à des machines d'apprendre sans avoir été explicitement programmées (définition d'Arthur Samuel en 1959). Au lieu d'écrire un programme à la main, un algorithme de ML va analyser de nombreux cas d'exemple pour produire un programme ou modèle pour effectuer la tâche illustrée par les exemples. Un tel algorithme peut combiner un très grand nombre de données et règles différentes, et être mis à jour en lui fournissant de nouveaux exemples. Si les exemples fournis sont représentatifs du processus à modéliser,

l'algorithme fonctionnera aussi bien sur de nouveaux cas d'exemple. La forte augmentation des capacités de calcul et de données disponibles rendent le ML préférable au développement manuel de tel traitement et permet la création nouveaux programmes auparavant irréalisables sachant la complexité et le temps de développement qu'il aurait fallu.

Une tâche de machine learning peut ainsi apprendre à : classifier, prédire, recommander, optimiser, détecter des motifs ou anomalies, filtrer… Les tâches de machine learning peuvent évidemment être combinées avec d'autres algorithmes ML ou non pour produire des algorithmes plus avancés.

Les applications sont nombreuses:

- ✓ prédire la météo, le prix d'une action, une prédisposition génétique, la page la plus pertinente pour une question (moteurs de recherche), la probabilité de survenue d'un virus, d'un incident actuel ou future.
- ✓ prédire la performance : d'une pub ou campagne marketing (ex. Adwords, Criteo), d'une page web, d'un canal de distribution, des implantations de magasins, d'un produit pour un profile de client.
- ✓ recommander par la prédiction : des produits les plus complémentaires ou qui seront les plus appréciés par un client (Amazon, Netflix), la stratégie la plus adéquate (gestion de portfolio en trading), l'action de maintenance la plus adaptée (ex. gestion automatisée des datacenters comme pour Office 365).
- ✓ faire de la reconnaissance : vocale, faciale, de mouvement (ex. Kinect de Microsoft), d'écriture (ex. les codes postaux sont identifiés dans les centres de tri depuis très longtemps par des réseaux de neurones).
- ✓ l'exploration de données ou data mining.
- ✓ classer ou identifier :
    - des éléments dans une image ou vidéo pour la classer, la commenter, l'indexer, réagir (ex. système de sécurité embarqués dans les voitures).
    - les sujets (ex. Google News) et les sentiments dans des textes (ex. analyses Twitter en temps réel pour trading et marketing) ou même le niveau de langue.
    - les prospects susceptibles de devenir clients (meilleures opportunités) ou de redevenir client (winback).
    - les clients susceptibles de partir (churn prediction) ou d'être intéressés par une offre.
    - les réclamations clients pour un processus adapté (ex. réponse automatique ou manuelle).

- ✓ découvrir et identifier les meilleures stratégies : de jeu (ex. Google DeepMind pour les jeux Atari), de stabilisation d'un robot ou d'un aéronef.
- ✓ détecter : des fraudes potentielles, des anomalies réseaux, une anomalie de capteurs (ex. alertes préventives centrales nucléaires), des objets ou comportements suspects (ex. aéroports), des erreurs d'orthographe ou de codage.
- ✓ optimiser l'exploitation par la prédiction de l'utilisation des ressources (ex. réseau de vélos et voitures, réseau électrique…).
- ✓ La traduction automatique.

II. Fonctionnement de l'apprentissage automatisé

1. Objectif et caractéristiques du problème d'apprentissage

La définition de l'objectif de l'apprentissage automatisé par Tom M. Mitchell, directeur du département de Machine Learning à la Carnegie Mellon University, clarifie le problème à maximiser **[7]** :

- Étant donné : l'expérience E, une classe de tâches T, une mesure de performance P.
- On dit qu'une machine apprend si : sa performance sur une tâche de T mesurée par P augmente avec l'expérience E.

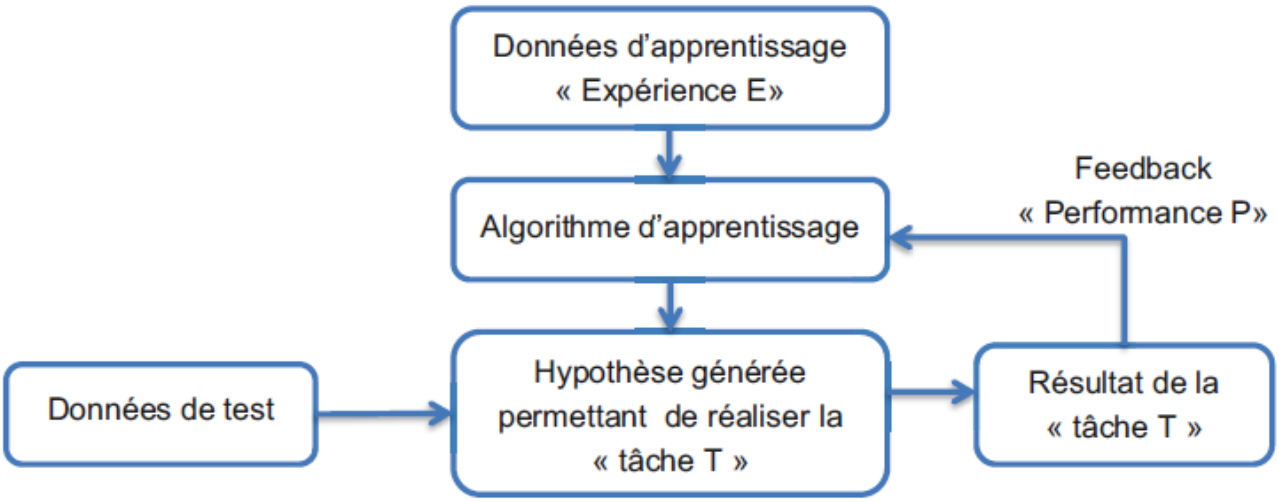


Figure 3 : principe du machine learning

Dans notre cas :

- L'expérience sera nos données d'apprentissage souvent appelées exemples ;
- L'hypothèse générée par l'algorithme d'apprentissage sera appelée « Modèle » ;
- Les mesures de performances sont couramment : la précision, le taux d'erreur, la variance entre le résultat donnée par l'algorithme et le résultat attendu.
- Un problème d'apprentissage peut ainsi se caractériser par :
- Un type d'apprentissage définissant la façon d'interagir avec l'environnement.
- Une sortie dont on va mesurer l'erreur généralement sous forme d'une « fonction de

cout » à minimiser.

- Un modèle et ses paramètres.
- Un algorithme pour créer et adapter le modèle en utilisant les exemples d'apprentissage issus de l'environnement, de façon à optimiser la fonction coût.

2. Les principaux types d'apprentissage

Pour s'orienté et strucutrer la résolution du problème, il est essentiel d'identifier le type d'apprentissage. Les types classiques d'apprentissage sont les suivants :

2.1. L'apprentissage supervisé

En analysant une base d'exemples contenant chacun des données d'entrée « avec » un résultat cible en sortie, une fonction de prédiction sera produite généralisant la règle d'association entre les données d'entrée et de sortie.Ainsi la fonction générée aura pour paramètre des données d'entrée similaires aux exemples d'apprentissage ou nouvelles et retournera le résultat iféré sur la base de la règle d'association produite **[8].**

2.2. L'apprentissage non-supervisé (ex. clustering)

En analysant une base d'exemples contenant des données d'entrée « sans » un résultat cible en sortie, une fonction sera produite classant en groupes homogènes ces données selon une régle d'association généralisant en classement. Ainsi la fonction générée aura pour paramètre des données d'entrée similaires aux exemples d'apprentissage ou nouvelle et retournera le résultat inféré sur la base de la règle d'association produite. La différence majeure avec l'apprentissage supervisé est qu'il n'y a ici pas de résultat à priori, le but est de découvrir le meilleur classement des données fournies et des structures invisibles.

2.3. L'apprentissage semi-supervisé

L'apprentissage semi-supervisé est une technique d'apprentissage supervisé exploitant l'apprentissage non-supervisé pour analyser les données d'exemple qui n'ont pas de données de sortie. En apprentissage supervisé, avoir des données d'exemple avec la sortie souhaitée peut être difficile ou onéreux. Ainsi, il est courant d'avoir qu'une fraction des données disponibles avec la valeur de sortie et la majorité des données sans **[9].**

Un exemple serait une banque de données de millions de photos que nous voudrions classer : nous pouvons classer 5% des photos manuellement et analyser automatiquement le reste des photos pour découvrir les traits majeurs distinctifs de ces photos.

Le processus d'apprentissage non-supervisé, en ajoutant aux données d'apprentissage des informations de classement et de structure découvertes sur un volume de données bien plus important, aidera l'apprentissage supervisé à produire une règle d'association qui se généralisera beaucoup mieux sur l'ensemble des données.

2.4. L'apprentissage par renforcement (ex. q-learning)

L'apprentissage par renforcement apprend un comportement décisionnel optimal à partir de récompenses ponctuelles. L'algorithme optimisera le comportement pour obtenir la récompense quantitative maximale au cours du temps. Il est courant de l'illustrer par un agent autonome plongé dans un environnement qui doit apprendre de ses décisions permanentes en ayant ponctuellement des récompenses positives ou négatives, il doit alors comprendre et apprendre de son expérience passée pour optimiser sa stratégie.

3. Autres caractéristiques du problème d'apprentissage

En complément de celles définies par le type d'apprentissage, d'autres caractéristiques majeures facilitent la compréhension et la formalisation le problème :

3.1. Sortie : classification ou régression (ou autre) ?

Le résultat à produire en sortie de l'algorithme est classiquement une classification ou une régression. La classification a pour but de catégoriser les données d'entrée fournies dans des classes simples (ex. vrai / faux) ou multiples (ex. les éléments d'une scène), la régression doit prédire des valeurs numériques de sortie (ex. probabilité, score, valeur). Il est aussi possible de produire une structure, une séquence, des règles d'association.

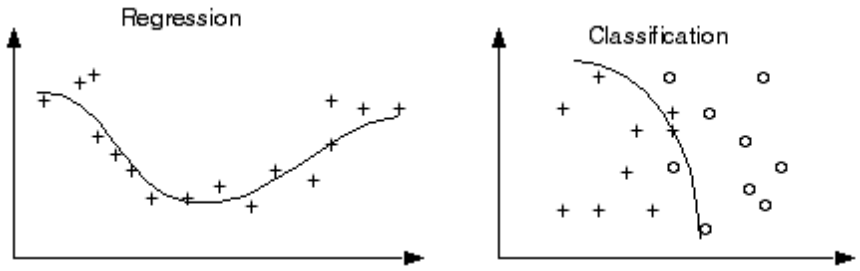


Figure 4: classification vs régression

3.2. Apprentissage “hors-ligne” vs “en-ligne”

L’apprentissage hors-ligne est la méthode la plus classique où l’on apprend dans un premier temps avec les exemples disponibles pour créer un modèle statique puis on utilise ce modèle pour prédire. L’apprentissage en-ligne met à jour le modèle au fur et à mesure que de nouvelles données d’apprentissage arrivent. L’apprentissage hors-ligne atteint en général de meilleures performances pour un même jeu d’exemples donné **[10].** L’apprentissage en-ligne est utilisé quand il y a un grand nombre de données d’apprentissage, car elles peuvent être traitées au fil de l’eau sans nécessité de stockage, et qu’avoir des données récentes dans le modèle peut significativement améliorer le modèle.

3.3. Et les systèmes de recommandation ?

Les systèmes de recommandation visent à prévoir le score ou la préférence d’un utilisateur pour un produit ou plus généralement le score évaluant la qualité d’association entre des entités. Ils apprennent depuis des scores existants, explicites (ex. note d’un utilisateur) ou implicites (ex. temps sur une page, click...), pour prédire les autres scores entre les entités afin de recommander les meilleures associations. C’est donc en général un problème d’apprentissage supervisé avec des techniques propres aux systèmes de recommandation (ex. filtrage collaboratif “item-item” ou “user-user”).

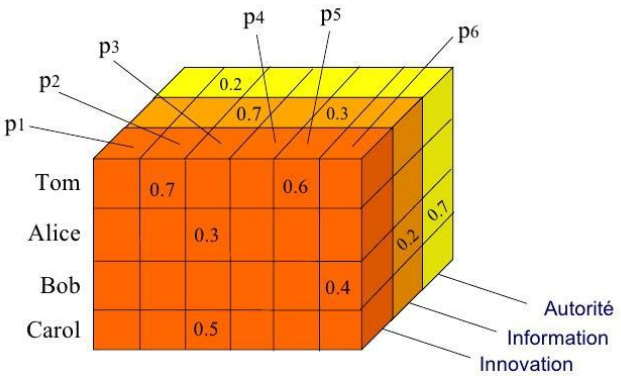


Figure 5 : matrice de recommandation multiple

3.4. Optimisation combinatoire : parcours de graphe, heuristique

Les problèmes de machine Learning sont souvent couplés à des problèmes d'optimisation combinatoire comme la recherche du chemin le plus court, la tournée optimale du voyageur de commerce.

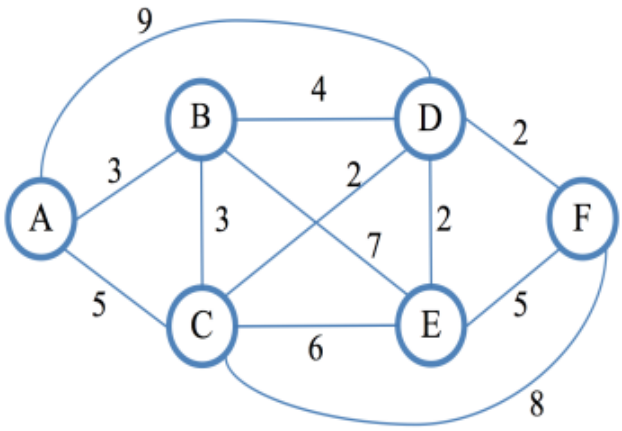


Figure 6: exemple de formalisation de graphe pour recherche du chemin le plus court

On peut résoudre ces problèmes avec les algorithmes classiques liés aux parcours de graphe suivant le problème : recherche de chemin le plus court (Dijkstra, A*, DFS, BFS), recherche des flots maximum (Ford-Fulkerson, BFS), calcul d'un arbre recouvrant (Prim), trouver la séquence la plus probable (Viterbi).

Pour résoudre des problèmes où les méthodes classiques ne fonctionnent pas ou sont trop longues à converger, de nombreuses heuristiques, recherchent la solution par différentes règles de recherche itératives :

- Les algorithmes génétiques,
- Les algorithmes utilisant l'intelligence globale et l'auto-organisation : « colonies de fourmis », cartes de Kohonen (non supervisé), Optimisation par Essaim Particulaire (OEP)
- Le renforcement apprenant de manière différé sur l'expérience (Q-learning, Sarsa)
- Les algorithmes de parcours de l'espace de recherche et spécialisation (ex. convergence par tâtonnement) : "recuit simulé", "recherche tabou".

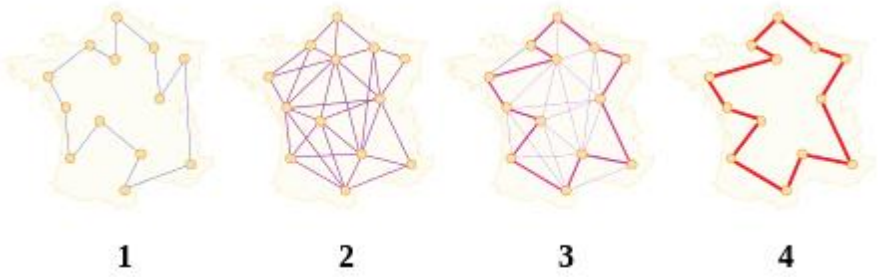


Figure 7: Illustration de l'algorithme d'optimisation « colonies de fourmis » qui explore dans demultiples zones locales et converge vers une solution optimale par intelligence collective

4. Choix et préparation des données d'apprentissage (attributs ou variables explicatives)

Les données d'entrée des exemples d'apprentissage (hors valeur de sortie) ou de prédiction sont individuellement appelés attributs. La sélection et l'amélioration des attributs interviennent continuellement dans le processus de création et d'amélioration d'un traitement de machine Learning. Fournir des données pertinentes est évidemment primordial mais leur qualité l'est tout autant **[11]** :

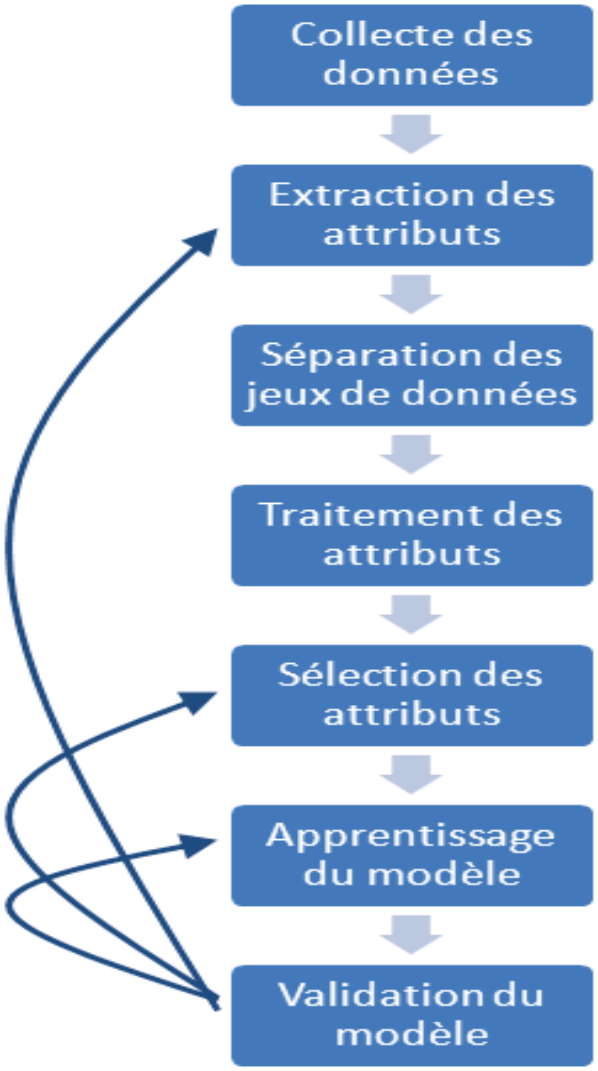


Figure 8: sélection des attributs à différentes phases

4.1. Malédiction de la dimension : réduction et sélection

Le nombre d'attributs (valeurs) fournis à l'algorithme d'apprentissage a grandement influer sur la complexité, le temps de calcul, notre propre travail de préparation des données, il faut ainsi essayer de réduire le nombre d'attributs pour pouvoir se concentrer sur les intéressants. Heureusement il existe différents algorithmes aidant à la sélection des attributs les plus pertinents comme l'analyse en composante principale (PCA) **[12]** et d'autre.

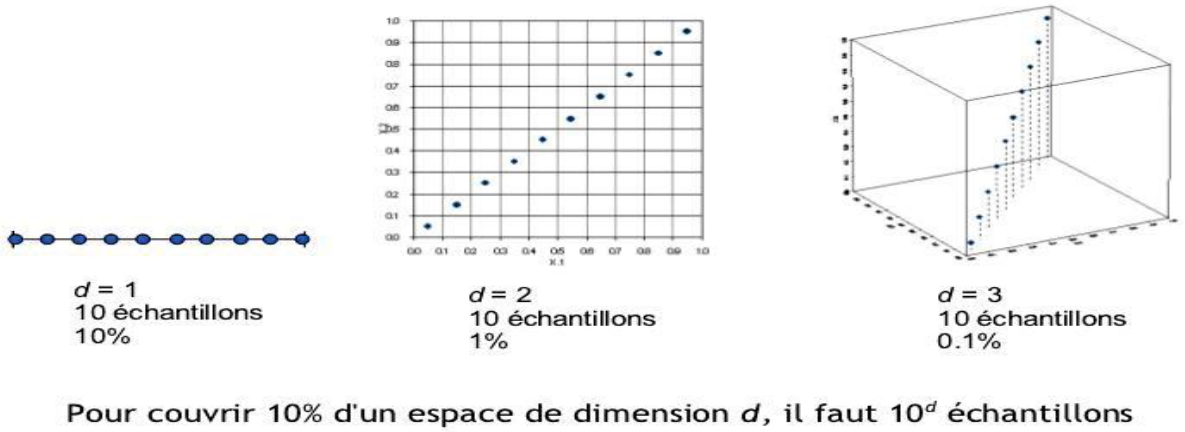


Figure 9: Illustration de la malédiction de la dimension montrant l'accroissement rapide de la complexité par l'ajout de dimensions (ajout d'attributs dans les exemples d'apprentissage)

4.2. Attributs manquants ou "sparse data", erronés, bruités, redondants ou dépendants

Il est rare d'avoir des données parfaites. Pour chaque cas des solutions existent :

- attributs manquants ou clairsemés (sparse) : remplacer par 0, la moyenne, valeur probable par analyse statistique des autres exemples. Il existe de nombreuses techniques d'imputation et si de nombreuses données sont manquantes par nature (sparse data), il faudra opter pour un algorithme d'apprentissage qui accepte ce type de données.
- attributs erronés ou bruités : l'analyse de la distribution des données est une bonne méthode pour découvrir les valeurs erronées ou bruitées à supprimer ou corriger. Si des données sont bruitées et qu'elles ne peuvent pas être supprimée ou corrigée, il est important de choisir un algorithme d'apprentissage qui soit robuste face au bruit.
- attributs redondants et dépendants : il est important d'éviter d'avoir des attributs qui soient redondants ou dépendants notamment pour les algorithmes qui supposent que les données sont statistiquement indépendantes. Si besoin, l'analyse de corrélation des attributs rend possible la découverte automatique d'attributs trop fortement corrélés.

4.3. Normalisation

De nombreux algorithmes sont plus performants suivant la plage des données, il est ainsi assez courant d'avoir à les ré échantillonner sur [0, 1], [-1, 1], ou autre plage et de les centrer sur 0.

4.4. Discrétisation

Des données numériques réelles peuvent être transformées en valeur discrètes selon différents stratégies (ex. plage de distribution) pour être gérés comme des attributs catégoriques, cela peut conduire à une importante amélioration de la performance.

4.5. Amélioration ou création de données calculées

La majorité des algorithmes d'apprentissage ne peuvent comprendre qu'une date correspond à un jour de la semaine, qu'utiliser le log d'une valeur améliorera sa capacité à faire des liens de distribution, qu'il est préférable d'utiliser le produit de 2 valeurs que les 2 valeurs indépendantes (ex. surface = données 1 x données 2). Créer de ces données calculées sur des données pourtant existantes peut largement améliorer la performance de prédiction en aidant l'algorithme à faire de nouveaux liens pertinents entre les données d'entrée et la sortie.

4.6. Données temporelles et d'état

La plus part des algorithmes d'apprentissage supposent les exemples comme indépendants des uns des autres et ne peuvent prendre en compte un quelconque lien. Si des exemples ont un lien temporel ou un autre lien de séquence, il peut être crucial de créer des nouveaux attributs caractérisant l'évolution temporelle ou séquentielle : moyenne pondérée, delta, variable à état sur différents échelles de temps…

5. Les familles d'algorithmes

Il n'existe pas encore d'algorithme magique répondant à tous les besoins et de manière optimale. Il faut ainsi avoir une bonne connaissance des différentes familles d'algorithme pour s'orienter dans les algorithmes à tester, les paramétrer correctement, les combiner (voir section « super modèles »), transformer les données si besoin et mieux comprendre les résultats pour gagner du temps.

5.1. Résolution ou approximation de systèmes linéaires simple ou multiple

De nombreuses méthodes mathématiques classiques rendent possible la résolution ou l'approximation de systèmes linéaires et non-linéaires, ou classification des données. Les algorithmes correspondant ont ainsi été largement implémentés et optimisés pour différents cas

d'utilisation. Les fonctions de régressions sont très utilisées car fournissent rapidement un résultat approximatif avec une implémentation simple mais pour des problèmes complexes ont tendance à produire des modèles qui se généralisent mal (sur-apprentissage).

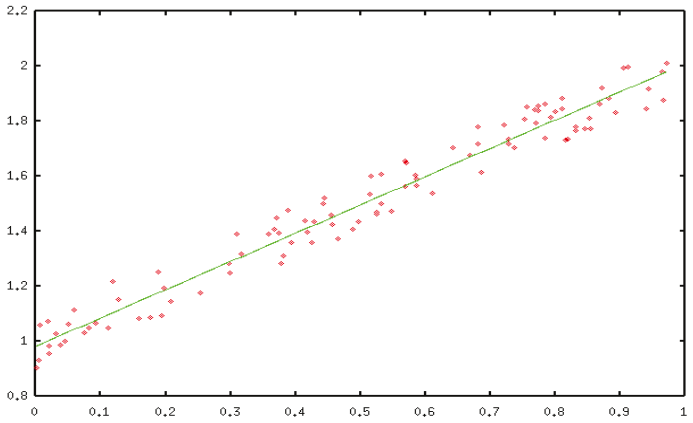


Figure 10: simple droite de régression linéaire

5.2. Arbres de décision

Ces algorithmes construisent des arbres de décision liant les données d'entrée jusqu'à une sortie plus au moins optimale. Ils sont majoritairement dédiés aux problèmes de classification mais peuvent être adaptés pour de la régression.

Un objectif majeur des algorithmes ML étant leur capacité de généralisation, le choix se fera les critères de segmentation utilisés, les méthodes d'élagage implémentées pour éviter le surapprentissage, et la manière de gérer les données manquantes dans le parcours de l'arbre **[13].**

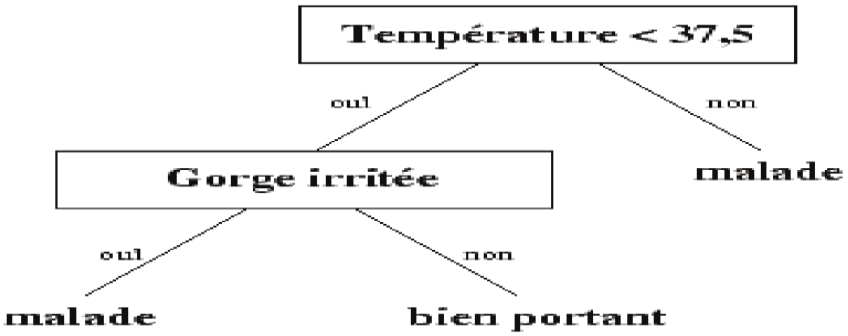


Figure 11: arbre de décision

5.3. Réseaux de neurones artificiels et “Deep learning”

Inspiré du fonctionnement des neurones biologiques, les réseaux de neurones artificiels offrent une très grande flexibilité mais nécessitent beaucoup de calcule et de données d’exemple pour l’apprentissage ainsi qu’un paramétrage parfois complexe. Ils bénéficient d’un important regain d’intérêt depuis quelques années, à travers le « Deep Learning », grâce à l’augmentation des capacités de calcul, les nombreuses données disponibles et la découverte de méthodes d’apprentissage raccourcissant le temps de convergence et des structures complexes très performantes

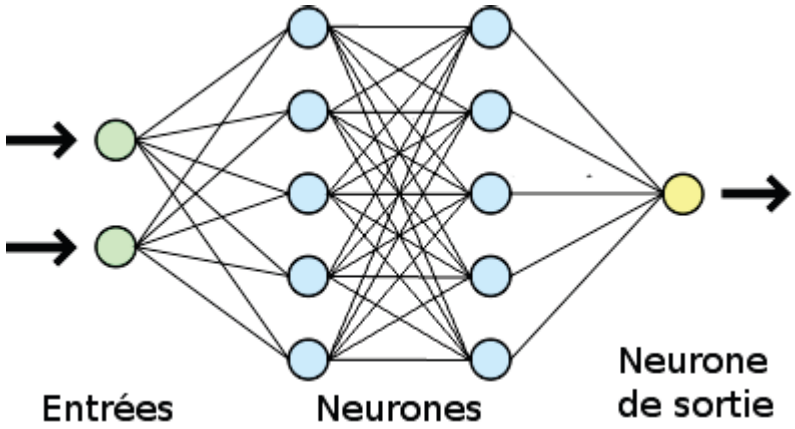


Figure 12: réseau de neurones artificiels

Le « Deep Learning » utilise des réseaux de neurones plus profonds (plusieurs étages de réseaux de neurones). Empile et connecte différents réseaux neurones spécialisés (ex. segmentation d’une image et découverte de caractéristiques) et améliorer les prédictions par la découverte de concepts intermédiaires pour l’apprentissage final.

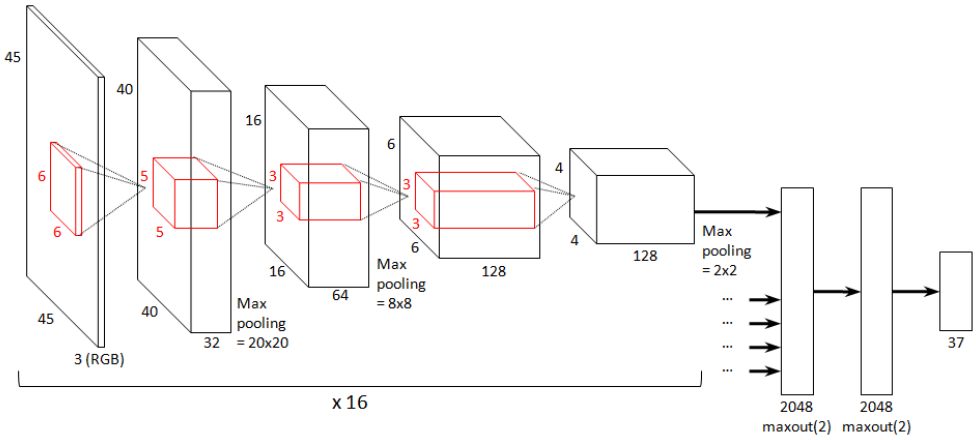


***Figure 13*: exemple de réseau "deep learning**

Il automatise ainsi une partie d'une des tâches les plus complexes dans un projet de machine learning : la sélection et l'enrichissement des données d'entrée permettant à l'algorithme d'extraire différentes informations et concepts clés pour améliorer la performance d'apprentissage. Le temps d'apprentissage et de convergence pour cette famille d'algorithme reste très élevé en comparaison des autres familles d'algorithmes.

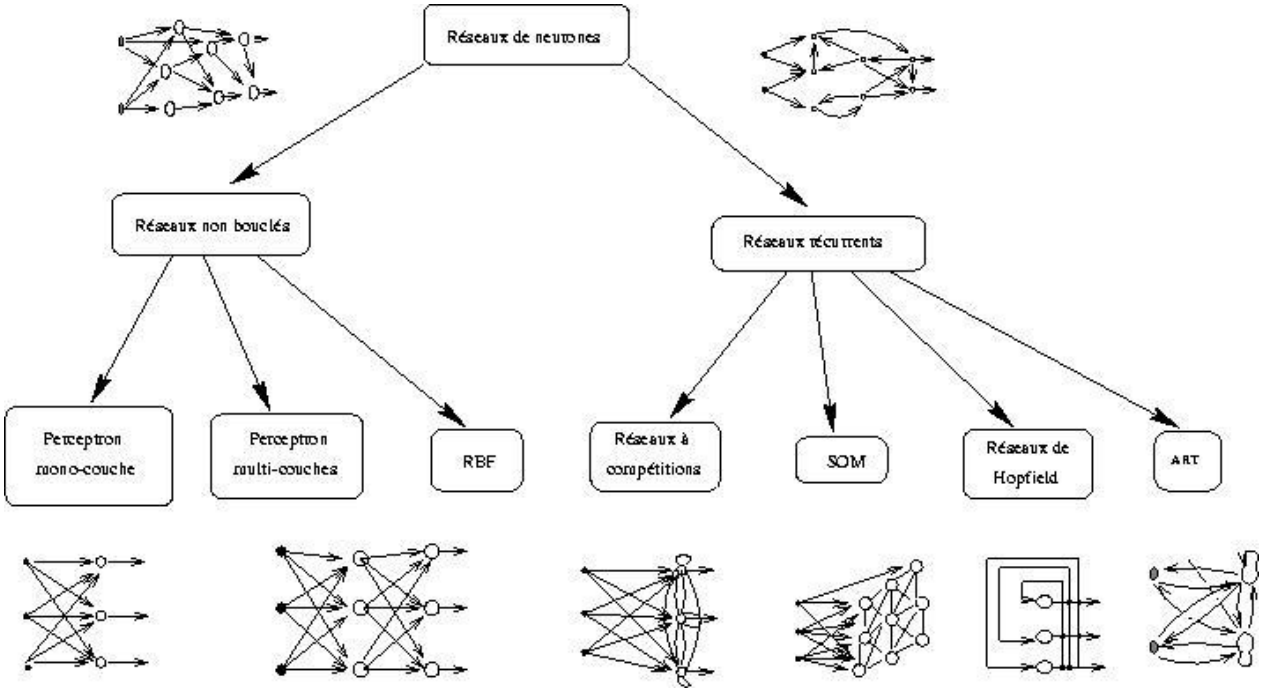


Figure 14: les principaux types de réseaux de neurone

5.4. Machine à vecteurs de support (SVM) : noyaux et marges maximales

Les machines à vecteurs de support sont une généralisation des classifieurs linéaires. Elles rendent notamment possible la classification de données non linéairement séparables en utilisant un “noyau” (fonction de transformation) et de classifier sur N dimensions (les hyperplans). Elles maximisent la marge, ou distance moyenne, entre les données et les frontières de séparation de telle sorte qu’elles trouvent les hyperplans optimaux séparant ces données pour une bonne généralisation. Les SVM supportent des données de grande dimension en entrée (nombreuses variables), nécessitent peu de paramètres, leurs garanties théoriques et leurs résultats pratiques sont bons.

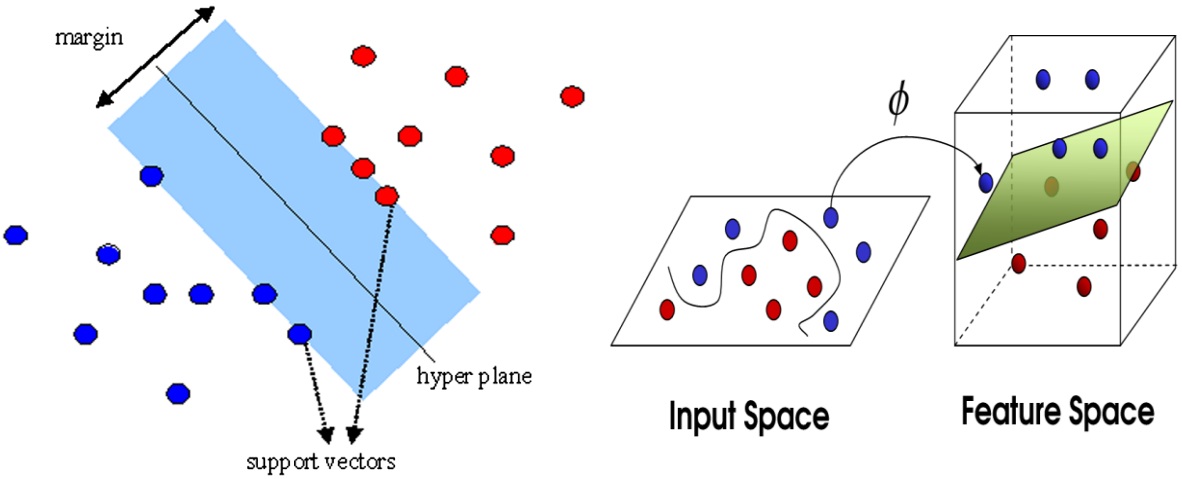


Figure 15: illustration des concepts des algorithmes SVM

C’est donc une famille d’algorithme performante et polyvalente dont la vitesse d’apprentissage est le seul bémol mais qui reste plus rapide que les réseaux de neurone.

5.5. Méthodes probabilistes et graphes

Basées sur les probabilités et hypothèses de distribution des données d’entrée et de sortie, pouvant aussi intégrer les liens entre données, la prédiction par l’utilisation de modèles probabilistes exploitent majoritairement les règles de Bayes et les différentes analyses de distribution. C’est un pilier du machine learning. Cela va du modèle très simple “bayésien naïf” sur les probabilité issues des statistiques des valeurs d’entrée par rapport à la donnée de sortie jusqu’aux différentes modèles dérivés des réseaux bayésiens et des propriétés de Markov et de

Gauss intégrant les graphes de probabilités, les notions de séquences et de temporalité. C'est une famille trop vaste pour donner un jugement globale : à part les algorithmes "bayésiens naïfs" qui sont simples à mettre en oeuvre et efficaces s'il y a des dépendances de distribution, beaucoup nécessitent une bonne compréhension pour les exploiter efficacement mais sont la base des systèmes les plus aboutis en intelligence artificielle et en recherche d'information (ex. algorithme PageRank de Google).

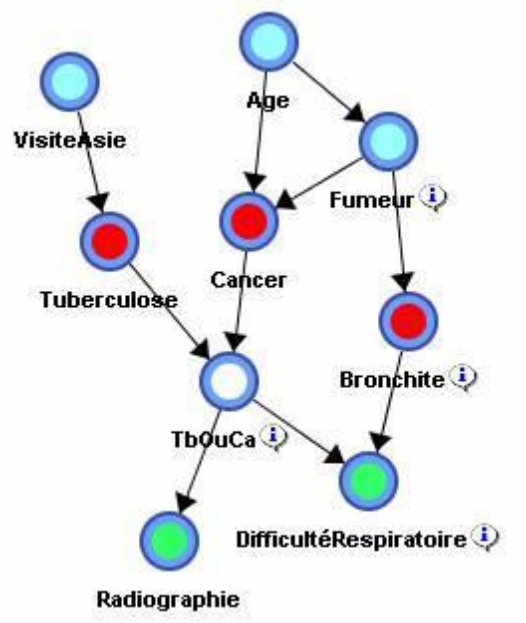


Figure 16: exemple de réseau Bayésien

5.6. Par analogie (ex. kNN)

Ces algorithmes utilisent un modèle de prédiction par analogies sur les exemples d'apprentissage. Le plus connu de cette famille est le k plus proches voisins qui stock l'ensemble des exemples d'apprentissage et au moment d'une prédiction recherche les k exemples les plus similaires pour décider par moyenne ou vote de la prédiction optimale.

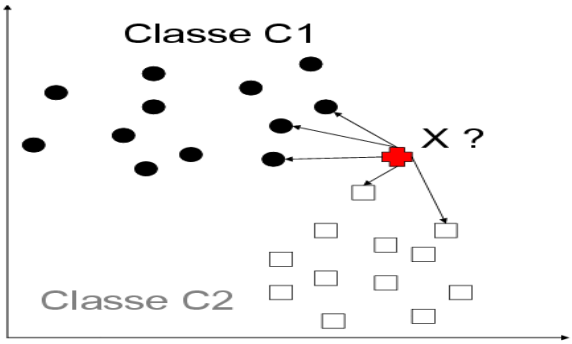


Figure 17: k plus proche voisin

5.7. Générateur de règles

Des algorithmes de génération de règles (DTNB, M5, Jrip, DecisionTable, NNGE, PART, OneR...) génèrent des règles optimales pour la classification ou la régression des données. Beaucoup s'appuient sur un algorithme d'arbre de décision ou naïve bayésien pour en déduire les règles. L'intérêt est de produire un modèle facilement compréhensible, portable. A noter, que OneR qui génère une seule règle basique et ZeroR qui ne retourne que la valeur la plus fréquente constatée pendant l'apprentissage sont très utiles comme référence de comparaison dans un processus de sélection de modèle : un modèle devrait avoir au minimum une performance supérieure à ZeroR et OneR.

- SI p(gène A) >= 0.32 ALORS tissu = bronches
- OU SI p(gène B) >= 0.9 AND p(gène A) < 0.1 ALORS tissu = cerveau
- OU SI p(gène C) >= 0.15 ALORS tissue = rein

Figure 18: exemples de règles simples générées par JRip

5.8. Agrégation de modèles ou méthodes d'ensemble - "super modèles"

Combiner des modèles parfois individuellement moyennement performants autorise la construction de "super modèles" plus performants et plus stables. On peut voir cela comme la création d'un comité d'experts avec différentes stratégies pour la prise de décision collégiale. Les méthodes d'ensemble les plus connues sont :

Bagging (ou agrégation boostrap) : consiste à ré-échantillonner au hasard avec doublons les exemples d'apprentissage et faire générer à l'algorithme voulu un modèle pour chaque sous-échantillon. On obtient ainsi un ensemble de modèles dont les différentes prédictions doivent être moyennées si c'est une régression ou choisies par vote si c'est une classification.

● Avantages : performant, rapide et facilement parallélisable.

● Inconvénients : perte de compréhension du modèle créé (commun à la majorité des méthodes ensemblistes)

Boosting : plusieurs modèles apprennent en parallèle et en séquence pour chaque exemple. L'apprentissage d'un modèle est boosté, en appliquant un poids plus fort sur l'exemple à apprendre, à chaque fois que le précédent modèle a prédit avec erreur après apprentissage. Cela créé ainsi un ensemble de modèle chacun spécialisés sur certaines erreurs.

- Avantages : garantie théorique d'amélioration, généralement plus performant que bagging si pas de bruit. Une version améliorée de l'algorithme appelée « Gradient Boosting Machine » (GBM ou GBRT) est un des algorithmes les plus performants et polyvalents actuellement.
- Inconvénients : sensible au bruit, algorithme séquentiel, possible sur-apprentissage (non prouvé).

Stacking : consiste à appliquer un algorithme de Machine Learning à des modèles générés par un autre algorithme de machine Learning. On va donc prédire quels seront les meilleurs modèles à utiliser suivant les données en entrée.

- Avantage : très performant (fameux vainqueur du Netflix prize) et polyvalent par la possibilité de combiner des algorithmes très différents.
- Inconvénients : temps d'apprentissage, temps de prédiction et taille du modèle final.

Forêts aléatoires (Random Forest) : ce modèle très performant nécessitant la création d'un nombre important de modèles à combiner (ex. 500) ajoute à la technique de bagging le fait d'échantillonner au hasard aussi les variables d'entrée.

- Avantages : parmi les algorithmes les plus performants et polyvalents, forte élimination du biais et de la variance.
- Inconvénients : temps d'apprentissage et taille du modèle final.

6. Choix d'algorithme

Chaque famille d'algorithme a des avantages et inconvénients liés au type de problème, a sa complexité, sa famille, au temps et ressources nécessaires. De même chaque algorithme a souvent différents paramètres appelés « hyper-paramètres » qui ont une grande influence sue les performances de l'algorithme pour les données fournies. Il faut aussi pouvoir définir de bons outils de comparaison, comprendre les résultats pour corriger les données d'entrée ou l'algorithme, concevoir le modèle ou combinaison des modèles finaux.

6.1. Biais, Variance, Généralisation

Un enjeu majeur d'un bon processus d'apprentissage est la capacité à bien se généraliser et non à produire des règles qui se seraient valables que pour les exemples d'apprentissage donnés. Pour cela, il faut trouver un bon compromis entre biais et variance :

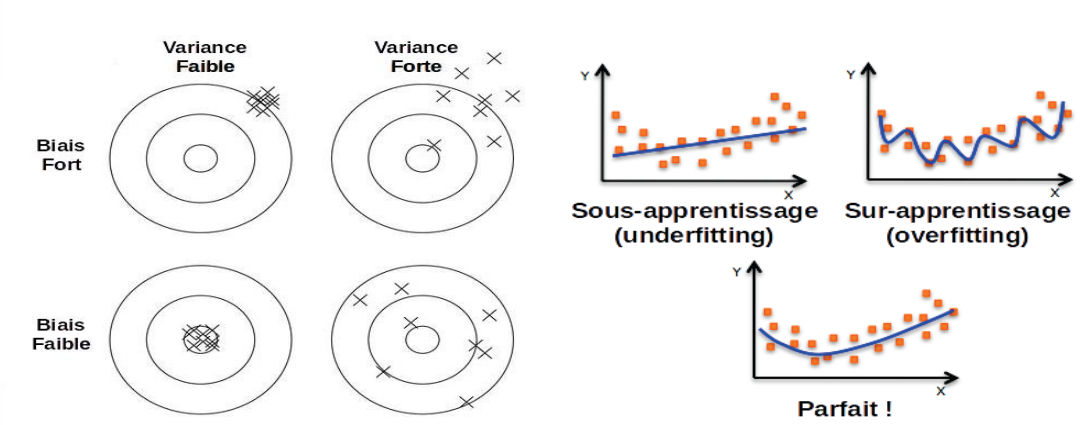


*Figure 19: **Biais, Variance***

- ***Biais :*** le biais est la différence entre l'espérance des données réelles et celle qui ont été prédites. Si il est de 0, le modèle est sans biais. Les moyennes correspondent entre prédiction et réalité mais les données ont peut-être systématiquement un écart important et se retrouvent finalement que sur la moyenne, c'est typique du sous-apprentissage (underfitting). Il faut donc aussi mesurer la variance.
- ***Variance :*** la variance mesure la dispersion entre les données réelles et celle qui ont été prédites, c'est la moyenne de l'écart au carré. A vouloir trop réduire la variance, on peut arriver au surapprentissage.

- ***Généralisation vs surapprentissage :*** l'apprentissage sur des exemples a pour but de pouvoir prédire sur d'autres données, c'est la généralisation. Un modèle qui a trop appris sur les données d'exemple sans dégager les tendances générales créé un modèle proche des données d'exemple mais sans capacité de généralisation.

6.2. Fonction de coût, descente de gradient et régularisation

Les algorithmes d'apprentissage optimisent leur modèle en minimisant l'erreur qui est mesurée classiquement par une fonction de coût et un algorithme de minimisation comme la descente de gradient.

Dans ce processus d'optimisation automatique, l'algorithme peut tendre vers un modèle non généralisable dû à un surapprentissage (overfitting).

Pour éviter cela, on utilisera une méthode de régularisation pénalisant la fonction de coût pour éviter les valeurs extrêmes afin de favoriser une fonction plus générale.

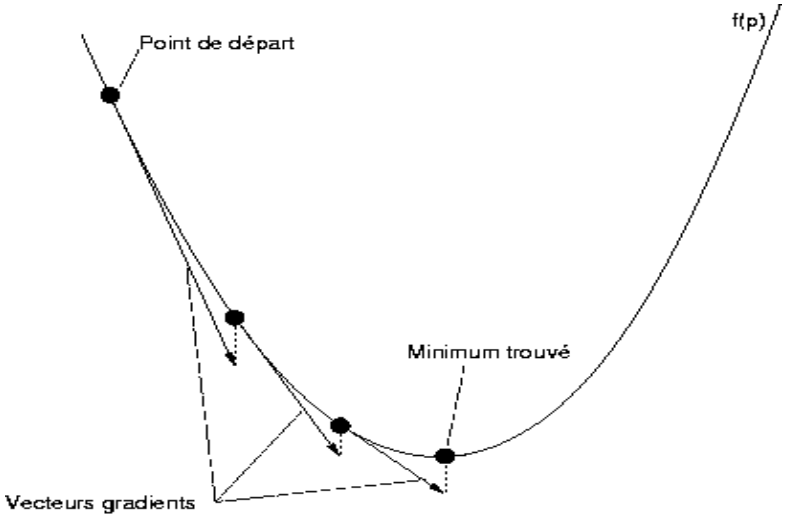


Figure 20: illustration du processus de descente de gradient pour la recherche du minimum local d'une fonction

6.3. Critères de comparaison

Pour comparer la performance des algorithmes il faut choisir un critère correspondant à l'objectif souhaité.

Imaginons que nous souhaitions évaluer la performance d'une prédiction d'un évènement qui à 2% de chance de se produire. En prédisant systématiquement "faux" l'évènement n'arrivera pas, on devrait tendre vers un beau taux de succès de 98% mais quelle utilité ? Dans ce cas il serait plus intéressant d'avoir un autre critère d'évaluation comme : "le rappel de la classe faux'", le coefficient de confiance Kappa, l'aire AUC en-dessous de la courbe ROC.

a-Classification

Pour les classements binaires de type vrai / faux on différencie les :

- VP (TP) : vrai positif (true positive), c'est vrai on ne s'est pas trompé.
- FP : faux positif (false positive), c'est une fausse alarme !
- VN (TN) : vrai négatif (true negative), c'est faux on ne s'est pas trompé.
- FN : faux négatif (false négative), ce n'est pas faux !

	Malade	**Non Malade**
Test +	Vrai Positif (VP/TP)	Faux Positif (FP)
Test -	Faux Négatif (FN)	Vrai Négatif (VN/TN)

Tableau 2: Classification

Une classe est une des possibilités de classification pour une prédiction. Pour une classification binaire, les classes sont "vrai" et "faux", mais cela pourrait être un classement multiple comme : "nul", "moyen", "bien", "super".

Les mesures élémentaires les plus utilisées sont la précision, la sensibilité et la spécificité que j'ai illustrées ci-dessous dans le cas d'un test de maladie :

Nom de la Mesure	**Formule**	**Non Malade**
Précision (Precision)	TP/(TP+FP)	Probabilité qu'une personne soit malade si le test est positif. Un test peut tricher et obtenir 100% en retournant juste une fois positive en choisissant la prédiction la plus facile.
Rappel (Recall) / Sensibilité/ Taux de vrais positifs	TP/(TP+FN)	Probabilité que le test soit positif chez les malades. Un test peut tricher et obtenir 100% en retournant toujours « positif ».
Spécificité /Taux de vrais négatives	TN/(FP+TN)	Probabilité que le test soit négatif chez les non malades. Un test peut tricher et obtenir 100% en retournant toujours « négatif ».

Tableau 3: Les mesures élémentaires les plus utilisées dans le cas d'un test de maladie

On remarque que l'on doit faire une combinaison de critères d'évaluation pour qu'une évaluation soit représentative de la performance de prédiction sur les différents cas possibles. Il existe donc des mesures plus équilibrées sur les différents cas possibles comme :

● ***La courbe ROC (Receiver Operating Characteristic) :*** est une courbe de mesure de la performance d'un classifieur binaire facilitant la visualisation de son spectre d'action en affichant les taux de vrais positifs en fonction du taux de faux positifs.

AUC : c'est l'aire totale en-dessous de la courbe ROC, plus elle est élevée, plus le classifieur est globalement performant.

Le F1-score : score global de performance d'un classifieur intégrant la précision et le rappel.

Le coefficient de Kappa : score d'accord réciproque de classement entre différents observateurs, ce score est un bon indicateur de la qualité globale sur les différentes possibilités de classement.

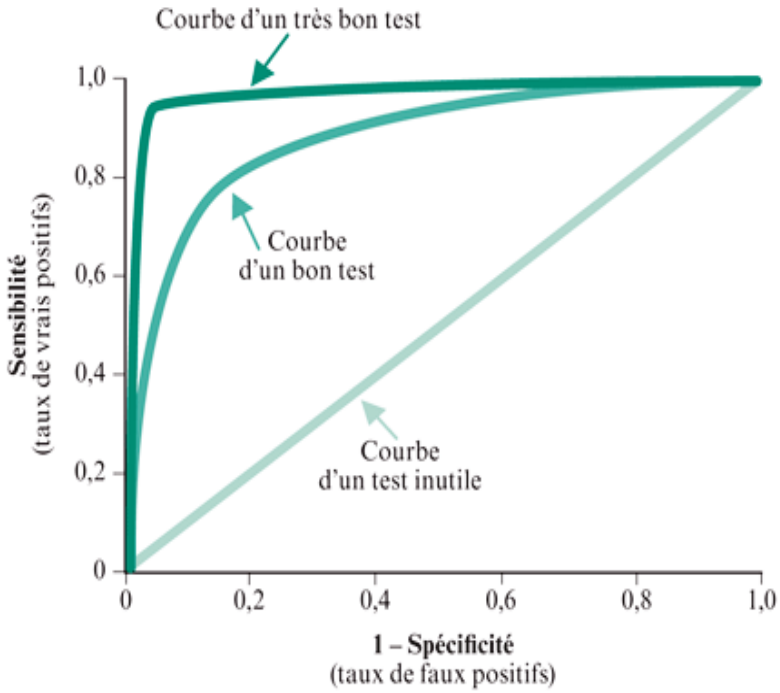


Figure 21: exemple de courbe ROC

b-Régression

Les mesures les plus utilisées en régression sont :

- ✓ ***RMSE (root mean squared error) :*** l'erreur quadratique moyenne est la mesure la plus utilisés pour les problèmes de régression.
- ✓ ***MAE (mean absolite error) :*** la moyenne de l'erreur.

6.4. Méthodes de test : jeux de données, validation croisée, backstesting

a-Jeux de données

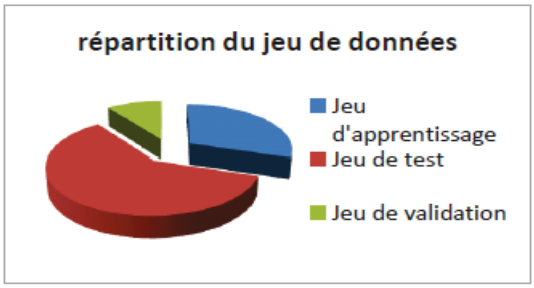


Figure 22: Répartition du jeu de données

b-Validation croisée (Cross-Validation)

Une technique, appelée validation croisée (CV, cross-validation), découpe automatiquement et test successivement différents jeux d'apprentissage et jeux de test. C'est un bon moyen pour valider la capacité de généralisation d'un algorithme d'apprentissage. On divise le jeu de données en k échantillons, un des échantillons est utilisé pour l'apprentissage, le reste pour le test. On répète l'opération avec un autre échantillon jusqu'à avoir testé l'apprentissage avec chacun des échantillons. Les scores finaux sont les moyennes des scores de chacun des k tests (voir illustration ci-dessous).

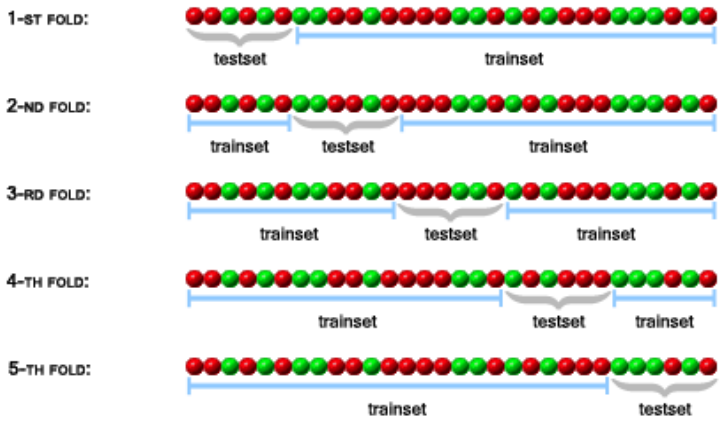


Figure 23: exemple de k-cross fold validation (validation croisée) avec k=5. Le jeu de données est séparé également 5 fois différemment, sans réutiliser les mêmes données d'apprentissage, et testé à chaque fois. Le résultat moyen des 5 tests est le résultat final

Pour des données temporelles, pour ne pas biaiser l'évaluation, il est important de faire attention à ce que le jeu de test soit composé uniquement d'exemples postérieurs aux exemples du jeu d'apprentissage. La validation croisée n'est donc pas possible dans ce cas.
Une technique très utilisée en trading algorithmique est de simuler le processus dans l'ordre historique pour voir l'évolution d'un algorithme ou du système complet, c'est le backtesting.

6.5. Recherche des hyper-paramètres optimaux : grid, random, bayesian optimization.

Chaque algorithme à des paramètres appelés hyper-paramètres qu'il faut adapter à chaque nouveau jeu de données. Avec un peu d'expérience, on peut s'orienter plus rapidement dans la définition de bons hyper-paramètres initiaux mais il y a souvent de nombreuses combinaisons à tester. Heureusement, il existe des algorithmes pouvant tester automatiquement ces combinaisons.

- Gridsearch testera itérativement, avec des intervalles définis, toutes les combinaisons des différents hyper-paramètres.
- Randomsearch effectue ces mêmes tests mais de manière aléatoire, cette solution est généralement plus rapide pour trouver des combinaisons performantes sans pour autant être optimale.
- L'optimisation bayésienne commence par une approche aléatoire puis s'oriente vers les combinaisons dont les probabilités sont les plus élevées pour converger vers la solution optimale.

Conclusion

La mise en oeuvre de solutions de machine learning est grandement facilitée par les implémentations open source des différents algorithmes. Il n'est alors plus nécessaire de connaître le fonctionnement exact de ces derniers ainsi que les mathématiques poussées dont ils proviennent.

Néanmoins, une bonne connaissance des données sur lesquelles on travaille est nécessaire pour concevoir des modèles efficaces.

Il est facilement oublié que le machine learning est l'étape finale du processus d'analyse des données.

Avant de considérer faire des prédictions, il faut déjà savoir collecter les données, les pré-traiter (car les sources sont bien souvent inconstantes), les stocker, les visualiser pour mieux les connaître et enfin seulement il est intéressant d'envisager le machine learning.

Dans le chapitre suivant, nous présentons les méthodes du machine learning afin de prédire l'espérance de vie après la chirurgie de cancer du poumon.

Chapitre 3 : Implémentation

Résumé

Dans ce chapitre, on a utilisé une des méthodes les plus précises de ML : La régression logistique pour prédire le nombre de vivants et de morts après un an de chirurgie qui concerne le cancer des poumons. Python, le logiciel utilisé après l'import de certaines bibliothèques et outils qui permettent de bien analyser, travailler, faire des calculs mathématiques et donner des résultats graphiques facilement d'un ensemble de données (DATASET).
Le travail passe par deux étapes, la première consiste à tester (essais) et la deuxième à apprendre, la régression logistique donne un résultat final binaire CAD soit un 0 pour les vivants soit un 1 pour les morts représenté graphiquement (en barres) sous forme de nombres et de statistiques (pourcentage).

Introduction

L'intégration du machine Learning dans le domaine médical a eu une incidence directe sur la productivité et la précision des médecins.
ML peut soutenir à la fois la création et l'utilisation de connaissances médicales. ML est généralement destiné à soutenir les hospitaliers, personnels de santé dans l'exercice normal de leurs fonctions, dans les tâches qui dépendent de la manipulation des données et des connaissances. Un ML peut être exécuté dans un système de dossier médical électronique, par exemple, et alerter un clinicien lorsqu'il détecte une contre-indication à un traitement planifié. Il pourrait également alerter le clinicien lorsqu'il a détecté des tendances dans les données cliniques suggérant des changements significatifs dans l'état du patient. Le cancer est l'une des principales causes de décès dans la plupart des pays. Actuellement, le cancer du poumon est le cancer le plus fréquent en chirurgie thoracique. Dans ce chapitre nous avons utilisé la méthode ML : la régression logistique pour prédire si un patient atteint du cancer du poumon survivra un an après

une chirurgie thoracique. Le résultat de cette technique a ensuite été mesuré et comparé en fonction de la précision et de la performance.

I. Etat de l'art

Plusieurs travaux ont été réalisés dans la littérature pour le diagnostic médical en utilisant les méthodes connexionnistes, et évolutionnaires. En ce qui concerne :

1. L'approche connexionniste

1/Cheng et al. [46] ont utilisé deux modèles des réseaux de neurones artificiels.

Le premier est un réseau à rétro propagation d'erreur avec une topologie à trois couches.

Le second est un réseau utilisant des fonctions à base radiale.

Le taux de reconnaissance obtenu avec le premier est de 72%, et avec le second est de 65%, bien que la démarche est intéressante, le temps d'exécution est long entrainant un sur apprentissage et ceci a donné des résultats faibles.

2/ Guo et Nandi [54], (2006), ont proposé un perceptron multicouche (PMC) en tant que classifieur avec l'algorithme de retro propagation d'erreur pour le diagnostic du cancer du sein avec la base de données WDBC (Wisconsin Diagnosis Breast Cancer) ils ont obtenu un taux de classification de 96.21 %.

3/ Verma [45] a proposé un nouvel algorithme d'apprentissage pour les réseaux de neurones appelé SCNN(Soft cluster neural network) basé sur la technique de soft clustering pour l'optimisation des poids des réseaux de neurones en utilisant la base de données de mammographie DDSM (Digital Database of Screening Mammography). Ce système a donné un taux de classification 94%.

2. L'approche évolutionnaire.

1/ Zhang et al. ont employé un algorithme génétique avec réseau de neurones pour la classification du cancer du sein avec un taux de reconnaissance de 90.5%

2/ Sekkal et al. ont appliqué les algorithmes génétiques combinés avec les réseaux de neurones pour l'amélioration de l'architecture sur la base de données d'arythmies cardiaques de la base MIT BIH, ils ont obtenu un taux de classification de 98.86%, une sensibilité de 99.09% et une spécificité de 98.66%.

Ils ont utilisé les algorithmes génétiques fondés sur un classifieur neuronal pour les poids de connexions sur la même base de données .Cette approche a donné de très bons résultats avec un taux de classification correcte de 98,72% et une sensibilité de 97,33% par rapport à classificateur classique qui a un taux de classification de 95,71% et 87,98% de sensibilité.

II– Matériels et Méthodes

1. Le modèle de régression logistique

a) Présentation

Le modèle de régression permet d'exprimer sous forme de probabilité la relation entre une variable dichotomique, dite variable dépendante ou expliquée, et des variables explicatives, quantitatives ou qualitatives. Dans notre cas, on cherchera à exprimer la probabilité que l'individu rachète son contrat au cours de l'année en connaissant les caractéristiques de cet individu.

L'utilisation de ce modèle est largement répandue en médecine par exemple, pour isoler les facteurs qui séparent les individus sains des individus malades.

Un tel modèle permet d'analyser plus précisément l'impact de chaque variable explicative retenue, et de pouvoir quantifier cette relation.

b) Formalisation mathématique

La régression logistique nous permet donc d'exprimer la probabilité que l'évènement se réalise en fonction des variables explicatives, à l'aide de la fonction Logit **[14]**. Cette fonction est la suivante :

$$\boldsymbol{logit(p) = \ln(\frac{p}{1-p})}$$

Cette fonction est la fonction qui est à la base de la régression logistique. Lorsque p varie dans]0;1[, la fonction Logit prend ses valeurs dans l'intervalle]-∞;+∞[tout entier.

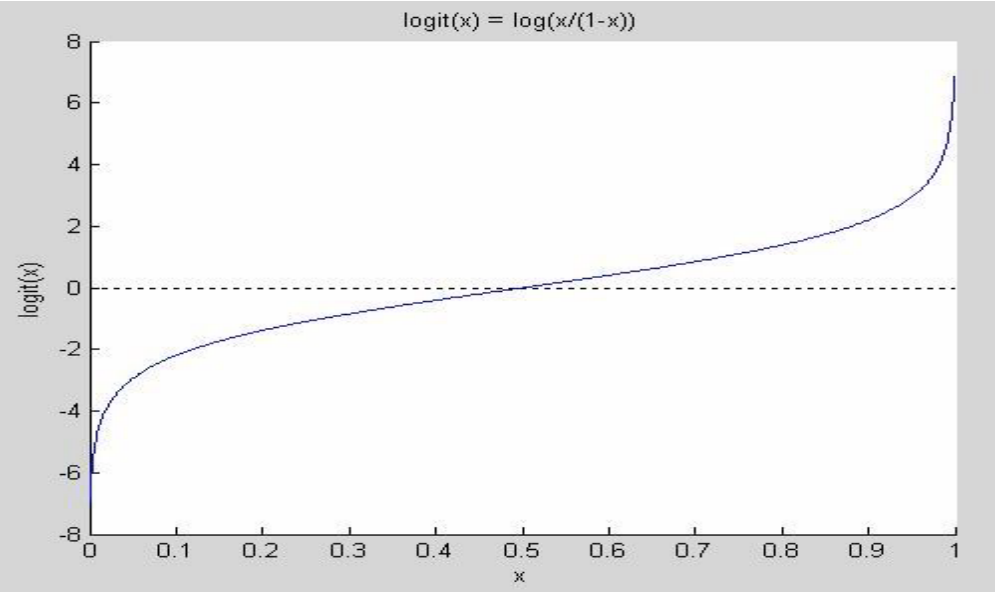


Figure 24:Représentation de la fonction Logit

La formulation mathématique de la régression logistique est la suivante :

$$\boldsymbol{logit\big(p(Y=1/X)\big) = \ln\left(\frac{p(Y=1/X)}{1-p(Y=1/X)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j}$$

$\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_j})$ Représente les variables explicatives, qui peuvent être qualitatives ou quantitatives, et qui doivent permettre de caractériser le phénomène étudié.

Y est la variable à expliquer qualitative valant **1** si un rachat est observé sur le contrat au cours de l'année et **0** sinon. On note alors **p(1|X)** la distribution conditionnelle de **X** sachant la valeur prise par **Y**, ce qui nous donne la relation suivante :

$$p(1/X) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j}}$$

Avec X représentant les variables explicatives et les étant les coefficients de la régression à estimer. Les coefficients permettent donc de mesurer l'influence de chaque variable et ainsi de déterminer les plus discriminantes.

2. Outils utilisés

Python : Est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la programmation orientée objet. Parce que sa syntaxe est élégante, que son typage est dynamique et qu'il est facile à interpréter, Python est un langage idéal pour l'écriture de scripts et le développement rapide d'applications dans de nombreux domaines et sur la plupart des plateformes.

L'interpréteur Python et sa vaste bibliothèque standard sont disponibles librement, sous forme de sources ou de binaires, pour toutes les plateformes majeures depuis le site Internet https://www.python.org/ et peuvent être librement redistribués. Ce même site distribue et pointe vers des modules, des programmes et des outils tiers. Enfin, il constitue une source de documentation.

L'interpréteur Python peut être facilement étendu par de nouvelles fonctions et types de données implémentés en C ou C++ (ou tout autre langage appelable depuis le C).

Python est également adapté comme langage d'extension pour personnaliser des applications.

Scikit-Learn : Est une bibliothèque qui fournit une gamme d'algorithmes d'apprentissage supervisés et non supervisés via une interface cohérente en Python. La vision de la bibliothèque est un niveau de robustesse et de support requis pour une utilisation dans les systèmes de production. Cela signifie qu'il faut se concentrer sur des préoccupations telles que la simplicité d'utilisation, la qualité du code, la collaboration, la documentation et les performances.

Pandas : Est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques. Pandas est un logiciel libre sous licence. Les principales structures de données sont :

Les séries (pour stocker des données selon une dimension - grandeur en fonction d'un index).

Les Dataframes : pour manipuler des données aisément et efficacement avec des index pouvant être des chaines de caractères (stocker des données selon 2 dimensions - lignes et colonnes).

Les Panels pour représenter des données selon 3 dimensions.

Format de lecture et écriture des données structurées en mémoire depuis et vers différents formats : fichiers CSV, fichiers textuels, fichier du tableur Microsoft Excel, base de données SQL [W4].

Spyder : Est un environnement de développement pour Python, libre et multiplateforme (Windows, Mac OS, GNU/Linux), il intègre de nombreuses bibliothèques d'usage scientifique. Spyder a un ensemble unique de fonctionnalités - multiplateforme, open-source, écrit en Python et disponible sous une licence non-copyleft. Spyder est extensible avec des plugins, comprend le support d'outils interactifs pour l'inspection des données et incorpore des instruments d'assurance de la qualité et d'introspection spécifiques au code Python.

Il offre une combinaison unique de fonctionnalités avancées d'édition, d'analyse, de débogage et de profilage d'un outil de développement complet avec l'exploration de données, l'exécution interactive, l'inspection approfondie et les capacités de visualisation d'un package scientifique.

Éditeur : Travaillez efficacement dans un éditeur multilingue avec un navigateur de fonctions / classes, des outils d'analyse de code, l'achèvement automatique du code, la division horizontale / verticale et la définition de la définition.

Console IPython : Exploitez la puissance d'autant de consoles IPython que vous le souhaitez dans la flexibilité d'une interface graphique complète ; exécutez votre code par ligne, cellule ou fichier ; et restitue les graphiques directement en ligne.

Explorateur de variables : Interagir avec et modifier les variables à la volée ; tracer un histogramme ou une série chronologique, éditer une image de données par rapport à un tableau Numpy, trier une collection, creuser dans des objets imbriqués, et plus encore !

Débogueur : Tracez chaque étape de l'exécution de votre code de manière interactive.

Python-Seaborn (Visualisation de données statistiques pour python) : Seaborn est une bibliothèque pour créer des graphiques statistiques attrayants et informatifs en Python.

Certaines des fonctionnalités offertes par Seaborn sont :

Des outils pour choisir des palettes de couleurs pour faire de belles parcelles qui révèlent des motifs dans vos données.

Fonctions pour visualiser des distributions in variées et bi variées ou pour comparer entre des sous-ensembles de données.

Outils qui adaptent et visualisent des modèles de régression linéaire pour différents types de variables indépendantes et dépendantes.

Une fonction pour tracer des données statistiques représentation de l'incertitude autour de l'estimation.

Des abstractions de haut niveau pour structurer des grilles de tracés qui permettent de construire facilement des visualisations complexes.

II. Modélisation

1. Etude de la démarche data science utilisée

Comme tout projet de développement informatique, avoir une méthodologie rigoureuse de travail permet d'avancer rapidement et d'augmenter les chances d'atteindre le résultat souhaité. Les problèmes de machine Learning ne dérogent pas à la règle et découpent la problématique en étapes successives en facilitera la résolution.

Un projet ML se décompose en plusieurs phases :

En effet, avant de se focaliser sur les outils, il faut procéder aux points suivants :

1. Nettoyez les données brutes à l'aide de diverses techniques de nettoyage, telles que l'imputation, la normalisation et la transformation, puis effectuez une analyse exploratoire des données.

2. Divisez les données en un ensemble d'apprentissage et un ensemble d'essais et effectuez la sélection des variables sur les données d'apprentissage, si nécessaire.

3. Développez le modèle en utilisant les données d'apprentissage.

4. Appliquez le modèle aux données de test en tant qu'outil de prédiction.

5. Enfin, mesurez l'exactitude, la précision, ainsi que les métriques de rappel et autres pour déterminer les performances du modèle.

2. Ensemble de données

2.1. Description de la population

La première étape de l'étude consiste à récupérer des données cohérentes afin de former la base de données sur laquelle le modèle de prédiction sera appris. La plus grande partie des données a été acquise via d'une base de données du Centre de chirurgie thoracique de Wroclaw pour des patients consécutifs âgés de 21 à 87 ans ayant subi une résection pulmonaire majeure pour un cancer primitif du poumon entre 2007 et 2011

Les problèmes d'apprentissage sont énoncés sous forme de données. Ces séries caractérisant une série d'instances du phénomène à apprendre, que l'on nomme patient **[15]**.

Chaque patient Pa est constitué d'une description D et d'une sortie S

Pa = (D, S)

Où : D ∈ X = {DGN, PRE4, PRE5, PRE6, PRE7, PRE7, PRE9, PRE10, PRE11, PRE14, PRE17, PRE19, PRE25, PRE30, PRE32, AGE}

S∈Y = {Risk1} = (0,1)

Avant de commencer le développement du modèle, toutes les variables numériques ont été normalisées et toutes les variables catégorielles et factorielles ont été binarisées. La normalisation convertissait essentiellement toutes les variables en nombres décimaux compris entre 0 et 1, 0 étant le minimum de toute valeur et 1, le maximum pouvant être pris

Attribut	Description
DGN	Diagnosis - specific combination of ICD-10 codes for primary and secondary as well multiple tumors if any
PRE4	Forced vital capacity
PRE5	Volume that has been exhaled at the end of the first second of forced expiration
PRE6	Performance status -Zebra scale
PRE7	Pain before surgery
PRE9	Dyspnoea before surgery
PRE10	Cough before surgery
PRE11	Weakness before surgery
PRE14	T in clinical TNM - size of the original tumor
PRE17	Type 2 DM - diabetes mellitus
PRE19	MI up to 6 months
PRE25	PAD - peripheral arterial diseases
PRE30	Smoking
PRE32	Asthma
AGE	Age at surgery

Risk1 Y	1 year survival period - True value if died, False if alive

Tableau 4: le résumé la description de notre ensemble de données :

Selon la documentation des données, quelques-unes des variables mesurées pourraient être ignorées. Le DGN était uniquement un numéro d'identification de diagnostic, il n'a donc joué aucun rôle dans le résultat de la classification. Par conséquent, il a été exclu des données.

L'ensemble de données contient un total de 470 patients, avec 17 valeurs idéalement enregistrées pour tous les patients. Les données comprenaient environ 7 990 points de données, ce qui était une bonne quantité pour développer un modèle.

Cette base de données contient des patients de tout genre et âge confondus. Les valeurs sont de différents types.

3. Choix des algorithmes utilisés dans ce projet

La technique de modélisation détaillée utilisée dans ce projet était notamment : la régression logistique.

La régression logistique est une technique statistique couramment utilisée dans l'analyse de jeux de données.

La régression logistique est principalement utilisée pour expliquer la relation entre une variable binaire dépendante et d'autres variables indépendantes, basée sur une courbe logistique. Depuis la chirurgie thoracique.

Les données avaient une variable de réponse binaire (1 pour les morts, 0 pour les vivants), la régression logistique était considérée comme un outil pertinent pour l'analyse prédictive **[16].**

4. Apprentissage

Le nettoyage des données était une partie importante de l'analyse de l'ensemble de données. Premièrement, comme mentionné précédemment, la variable DGN a été supprimée des données conformément à la documentation de l'UCI.

Un facteur surprenant de cet ensemble de données était l'absence de valeurs manquantes. Après avoir parcouru l'ensemble du jeu de données, toutes les valeurs ont été prises en compte, ce qui a été très bénéfique pour la mise au point d'un prédicteur précis **[17].**

4.1.Visualisation des packages

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
Import seaborn as sns
```

4.2.Chargement de la base

```
df = pd.read_csv('C:/Users/Fatima/Desktop/csv_result-ThoracicSurgery (1).csv')
```

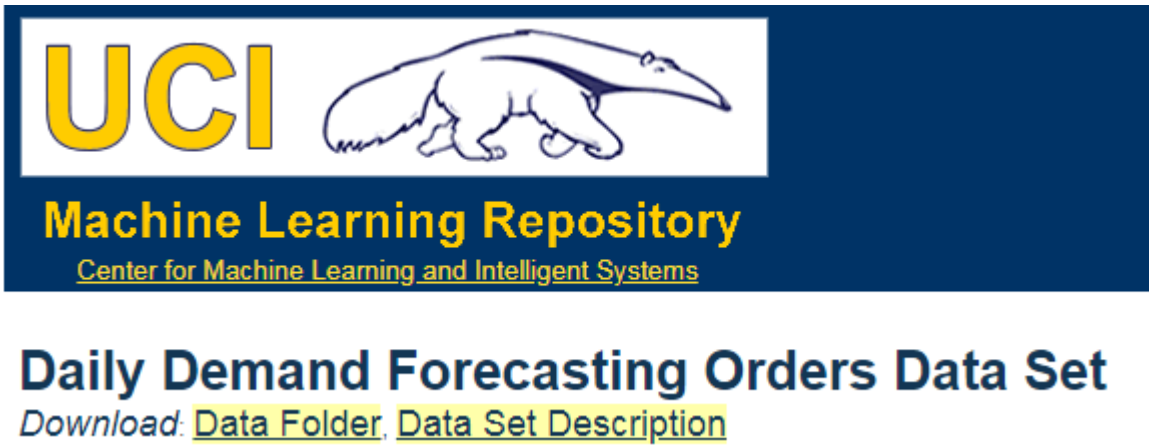


Figure 25: la source de base de données utilisée dans ce projet

4.3.Découverte du contenu de la base

df.head ()

Runfile ('C:/Users/Fatima/.spyder-py3/machine.py', wdir='C:/Users/Fatima/.spyder-py3')

```
Entrée [78]: df.head()
```

Out[78]:

	id	DGN	PRE4	PRE5	PRE6	PRE7	PRE8	PRE9	PRE10	PRE11	PRE14	PRE17	PRE19	PRE25	Smoking	PRE32	AGE	Risk1Yr
0	1	DGN2	2.88	2.16	PRZ1	F	F	F	T	T	OC14	F	F	F	T	F	60	F
1	2	DGN3	3.40	1.88	PRZ0	F	F	F	F	F	OC12	F	F	F	T	F	51	F
2	3	DGN3	2.76	2.08	PRZ1	F	F	F	T	F	OC11	F	F	F	T	F	59	F
3	4	DGN3	3.68	3.04	PRZ0	F	F	F	F	F	OC11	F	F	F	F	F	54	F
4	5	DGN3	2.44	0.96	PRZ2	F	T	F	T	T	OC11	F	F	F	T	F	73	T

Tableau 5: une description de l'entête données obtenu :

Entrée [88]: `pd.get_dummies(df1)`

Out[88]:

	id	PRE4	PRE5	AGE	DGN_DGN1	DGN_DGN2	DGN_DGN3	DGN_DGN4	DGN_DGN5	DGN_DGN6	...	PRE19_F	PRE19_T	PRE25_F	PRE25_T	Smoki
0	1	2.88	2.16	60	0	1	0	0	0	0	...	1	0	1	0	
1	2	3.40	1.88	51	0	0	1	0	0	0	...	1	0	1	0	
2	3	2.76	2.08	59	0	0	1	0	0	0	...	1	0	1	0	
3	4	3.68	3.04	54	0	0	1	0	0	0	...	1	0	1	0	
4	5	2.44	0.96	73	0	0	1	0	0	0	...	1	0	1	0	
5	6	2.48	1.88	51	0	0	1	0	0	0	...	1	0	1	0	
6	7	4.36	3.28	59	0	0	1	0	0	0	...	1	0	1	0	
7	8	3.19	2.50	66	0	1	0	0	0	0	...	1	0	0	1	
8	9	3.16	2.64	68	0	0	1	0	0	0	...	1	0	1	0	
9	10	2.32	2.16	54	0	0	1	0	0	0	...	1	0	1	0	
10	11	2.56	2.32	60	0	0	1	0	0	0	...	1	0	1	0	
11	12	4.28	4.44	58	0	0	1	0	0	0	...	1	0	1	0	
12	13	3.00	2.36	68	0	0	1	0	0	0	...	1	0	1	0	
13	14	3.98	3.06	80	0	1	0	0	0	0	...	1	0	1	0	
14	15	1.96	1.40	77	0	0	1	0	0	0	...	1	0	1	0	
15	16	4.68	4.16	62	0	0	1	0	0	0	...	1	0	1	0	
16	17	2.21	1.88	56	0	1	0	0	0	0	...	1	0	1	0	
17	18	2.96	1.67	61	0	1	0	0	0	0	...	1	0	1	0	

Tableau 6: Convertir une variable catégorique en variable factice / indicatrice

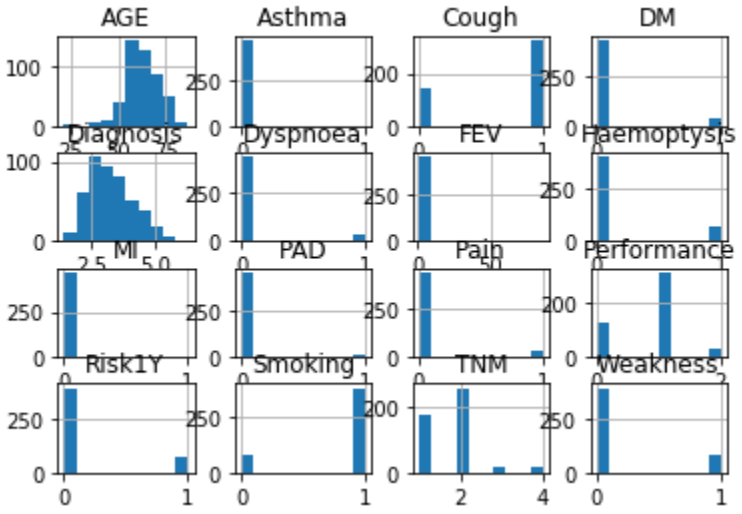


Figure 26: représente les données

Après avoir représenté et analysé les données récoltées, il nous est apparu qu'une corrélation était nécessaire afin de répertorier les différentes relations entre ces données.

	Diagnosis	FEV	Performance	Pain	Haemoptysis	Dyspnoea	Cough	Weakness	TNM	DM	MI	PAD	Smoking	Asthma	AGE	Risk1Y

	Diagnosis	FEV	Performance	Pain	Haemoptysis	Dyspnoea	Cough	Weakness	TNM	DM	MI	PAD	Smoking	Asthma	AGE	Risk1Y
Diagnosis	1.000	0.033	-0.091	0.020	-0.096	0.056	-0.053	-0.100	0.034	-0.115	-0.009	-0.036	-0.012	-0.061	-0.290	-0.046
FEV	0.033	1.000	-0.143	0.162	0.103	0.260	-0.100	-0.086	0.016	-0.022	-0.014	-0.025	-0.101	-0.017	-0.116	-0.043
Performance	-0.091	-0.143	1.000	0.093	0.123	0.093	0.685	0.418	0.090	0.025	0.027	0.023	0.172	-0.034	0.215	0.093
Pain	0.020	0.162	0.093	1.000	0.256	0.068	-0.024	-0.072	0.100	0.023	-0.017	-0.035	-0.077	-0.017	0.045	0.057
Haemoptysis	-0.096	0.103	0.123	0.256	1.000	0.134	0.082	0.060	0.060	-0.001	-0.027	0.086	-0.045	-0.027	0.087	0.066
Dyspnoea	0.056	0.260	0.093	0.068	0.134	1.000	0.050	-0.072	0.076	-0.043	-0.017	0.098	-0.077	-0.017	-0.015	0.106
Cough	-0.053	-0.100	0.685	-0.024	0.082	0.050	1.000	0.202	0.145	0.017	0.044	0.018	0.200	-0.026	0.150	0.089
Weakness	-0.100	-0.086	0.418	-0.072	0.060	-0.072	0.202	1.000	-0.036	0.070	0.059	0.030	0.119	-0.029	0.208	0.086
TNM	0.034	0.016	0.090	0.100	0.060	0.076	0.145	-0.036	1.000	0.037	-0.022	-0.021	0.038	-0.022	0.016	0.174
DM	-0.115	-0.022	0.025	0.023	-0.001	-0.043	0.017	0.070	0.037	1.000	-0.019	0.025	-0.037	-0.019	0.085	0.109
MI	-0.009	-0.014	0.027	-0.017	-0.027	-0.017	0.044	0.059	-0.022	-0.019	1.000	-0.009	0.030	-0.004	-0.030	-0.027
PAD	-0.036	-0.025	0.023	-0.035	0.086	0.098	0.018	0.030	-0.021	0.025	-0.009	1.000	0.061	-0.009	0.058	0.037
Smoking	-0.012	-0.101	0.172	-0.077	-0.045	-0.077	0.200	0.119	0.038	-0.037	0.030	0.061	1.000	-0.055	0.069	0.086
Asthma	-0.061	-0.017	-0.034	-0.017	-0.027	-0.017	-0.026	-0.029	-0.022	-0.019	-0.004	-0.009	-0.055	1.000	-0.019	-0.027
AGE	-0.290	-0.116	0.215	0.045	0.087	-0.015	0.150	0.208	0.016	0.085	-0.030	0.058	0.069	-0.019	1.000	0.039
Risk1Y	-0.046	-0.043	0.093	0.057	0.066	0.106	0.089	0.086	0.174	0.109	-0.027	0.037	0.086	-0.027	0.039	1.000

Tableau 7: Représente la corrélation entre les données

Lorsqu'on a essayé de prédire la durée de vie, nous avons travaillé avec un ensemble d'entrainement et un ensemble de test. Nous avons construit un modèle en utilisant l'ensemble de données d'entrainement et nous avons évaluez sa performance en utilisant l'ensemble de test. La plupart des fonctions de scikit-learn sont faites pour fonctionner dans ce cadre.

```
from sklearn.linear_model import LogisticRegression
# Don't be confuse about the name it is a classification algorithm not a regression
ModelLogit = LogisticRegression ()
Skb = SelectKBest (score_func=chi2, k=4)
Skbfit = skb.fit (Xfeatures, Ylabel)
```

4.4.Vérification de resultat de prédiction des résultats cibles

```
From sklearn.feature_selection import RFE
Rfe = RFE (modelLogit,3)
rfe_fit = rfe.fit (Xfeatures, Ylabel)

Print ("Number of Features", rfe_fit.n_features_)
Print ("Feature Names", list (df1.columns) [0:15])
Print ("Selected Features", rfe_fit.support_)
Print ("Features Ranking", rfe_fit.ranking_)
```

Number of Features 3

Feature Names ['Diagnosis', 'FEV', 'Performance', 'Pain', 'Haemoptysis', 'Dyspnoea', 'Cough',

'Weakness', 'TNM', 'DM', 'MI', 'PAD', 'Smoking', 'Asthma', 'AGE']

Selected Features [False False False False False True False True False True False False

False False False]

Features Ranking [5 13 11 3 10 1 9 1 2 1 6 8 4 7 12]

4.5. La prediction

```
From sklearn.linear_model import LinearRegression
From sklearn.datasets import make_regression
# generate regression dataset
X, y = make_regression (n_samples=100, n_features=2, noise=0.1)
# fit final model
Model = LinearRegression ()
model.fit(X, y)
# define one new data instance
Xnew = [[2.88,2.66,1,1,1,0,1,1,1,4,0,0,1,0,0,50]]
# make a prediction
Ynew = model.predict (Xnew)
Pinrt (new predict)
```

4.6. la precision de l'algorithme

```
# Using our LogisticRegression
num_folds = 10
Seed = 7
Kfold = KFold (n_splits=num_folds, random_state=seed)
Results = cross_val_score (modelLogit, Xfeatures, Ylabel, cv=kfold)
Print (results)
Print ("Mean Accuracy:", results.mean()*100.0)
Print ("STD Accuracy:", results.std()*100.0)
```

```
[0.78723404  0.85106383  0.85106383  0.85106383  0.80851064  0.82978723
  0.85106383  0.82978723  0.87234043  0.93617021]
Mean Accuracy: 84.6808510638
STD Accuracy: 3.78221039035
```

Une fois les données nettoyées, le nombre final de patients décédés par rapport aux patients, nous avons déterminé à l'aide d'un graphique en barres le risque de décès après 1 an de chirurgie thoracique.

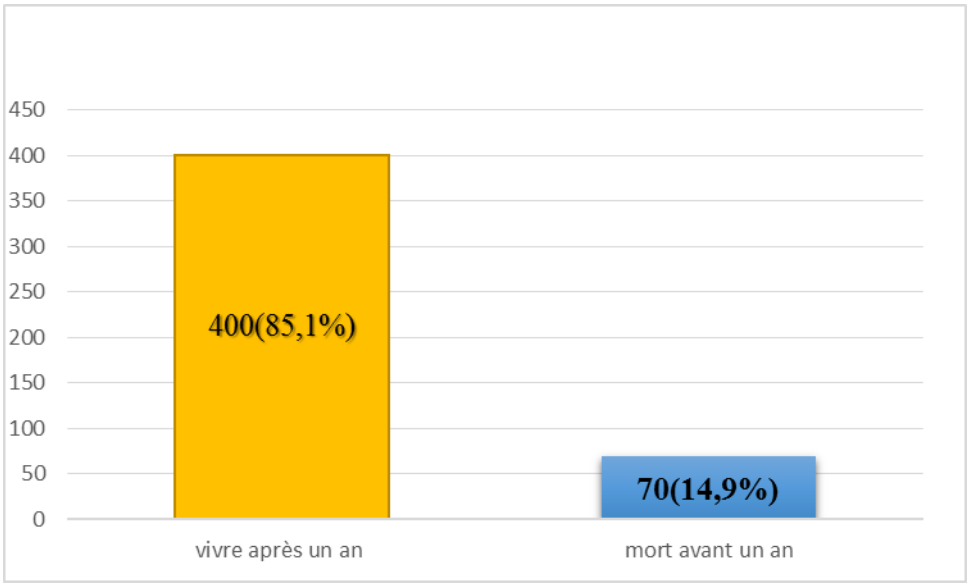


Figure 27 : Représentation graphique de risque de décès après 1 an de chirurgie thoracique

Conclusion

Pour résumer le processus, les données ont d'abord été soigneusement nettoyées afin d'éliminer tous les problèmes au sein du jeu de données. Des modifications ont été apportées aux données conformément à la documentation, notamment en supprimant les colonnes inutiles. De plus,

toutes les variables ont été normalisées ou binarisées en fonction de leur type, pour avoir une valeur comprise entre 0 et 1, 0 étant le minimum absolu et 1 le maximum absolu.

Conclusion générale et perspectives

Le cancer des poumons est considéré actuellement comme la maladie du siècle. Au Maroc, le cancer des poumons arrive en troisième position avec 6.488 cas suivi de celui du colon (4.118) et de la prostate (3.990).

Ces cancers ne représentent pas les cas les plus fréquents, mais plutôt ceux qui causent un nombre important de décès. Notamment à cause du retard dans leur diagnostic.

Dans ce travail, nous avons mené une étude critique des travaux proposés sur la prédiction du cancer des poumons. Pour ce faire, nous avons commencé par la définition des critères d'évaluation des différentes solutions existantes. Ensuite, nous avons fait une comparaison des travaux passés en revue, dans laquelle nous avons repris l'essentiel des avantages et inconvénients des travaux proposés. Pour finir par proposer notre méthode qui est une amélioration des travaux étudiés. Nous avons présenté une méthode pour la prédiction du durée de vie des patient atteint au cancer des poumons après une opération chirurgicales cet études est basée sur l'application des algorithmes d'apprentissage automatique supervisé.

Nous avons testé les algorithmes d'apprentissage supervisé sur une base de données, à savoir celle du CHU et du cabinet privé de Dr Djamel MEHIDI. Par la suite, nous avons amélioré le meilleur algorithme d'apprentissage à savoir si les personnes (souffrant d'un cancer de poumons) qui ont passé par une opération vont survivre ou non après un an où le résultat final a été représenté graphiquement (graphe de bâtonnets) en utilisant le ML.

En termes de perspectives, la prédiction du cancer à l'aide des méthodes d'apprentissage peut être élargie en utilisant les méthodes de base de connaissance pour augmenter l'interopérabilité du diagnostic.

Bibliographie

[1]. Moro-Sibilot, D., Tumeurs du poumon, primitives et secondaires. 2010.

[2]. Inserm, I., Dynamique d'évolution des taux de mortalité dans les principaux cancers en France. Plan Cancer, 2010.

[3]. INVS, F.H.c.d.L.I.I., Projection de l'incidence et de la mortalité par cancer en France en 2010. 2011.

[4]. Brambilla, E., [Responses to targeted therapies: lung cancer]. Ann Pathol, 2009. 29 Spec No 1: p. S77-80.

[5]. Flieder, D., Common Non-Small-Cell Carcinomas and Their Variants, in Pulmonary Pathology, M. Springer, Editor. 2008.

[6]. INSERM. http://www.inserm.fr/thematiques/cancer/dossiers/cancer-du-poumon.

[7]. http://eric.univ lyon2.fr/~ricco/cours/cours_regression_logistique.html

[8]. PennState « STAT 504 Analysis of discrete data »
https://onlinecourses.science.psu.edu/stat504/

[9]. Répertoire exhaustif des librairies, logiciels, framework du machine learning :
https://github.com/josephmisiti/awesome-machine-learning

[10]. Bibliothèques d'algorithmes de Machine learning en flux avec gestion du concept drift (évolution des modèles et mutation des données) - Université de Waikato :
http://moa.cms.waikato.ac.nz/details/

[11]. Distribuée pour Machine learning en flux - Yahoo et université de Waikato : http://samoa-project.net/

[12]. Bibliothèques et framework pour systèmes de recommandation - Université du Minnesota: http://lenskit.org

[13]. Bibliothèque C++ de machine learning online out-of-core très rapide:
https://github.com/JohnLangford/vowpal_wabbit

[14]. Lung Cancer Statistics. Centers for Disease Control and Prevention.
https://www.cdc.gov/cancer/lung/statistics/. Published 2016. Accessed November 24, 2016.

[15]. Thoracic Surgery Data Set. UCI Machine Learning Repository: Data Set.

https://archive.ics.uci.edu/ml/datasets/Thoracic Surgery Data. Accessed November 24, 2016.
[16]. Jahnke L, Asher A, Kermis SDC. The Problem of Data. Council Library on Information Resources; 2012. https://www.clir.org/pubs/reports/pub154/pub154.pdf
[17]. Machine Learning: What it is and why it matters. What it is and why it matters | SAS. http://www.sas.com/en_id/insights/analytics/machine-learning.html. Accessed November 24, 2016.