



UFR de Sciences Appliquées et de Technologie (SAT)

UFR de Sciences Economiques et de Gestion (SEG)

Département : Informatique

Département : Gestion

Master II

Filière : Méthodes Informatiques Appliquées à la Gestion (MIAGE)

----- **Mémoire de recherche de master II** -----

Sujet : Développement d'un modèle de Machine Learning pour faire une analyse financière (historique et prédictive) et le développement d'un Chatbot pour interroger les états financiers

Présenté par :

Omar Abd Al Wahab DIASSE

Membres du jury :

... ...

Encadreurs :

Pr Jean Marie DEMBELE

... ...

Dr Alioune Badara MBENGUE

Année académique :

Présentée le : ../**02/2025**

2022/2023

Remerciements

Allah

Pr Dembélé et Dr Mbengue

Tous les établissements et professeur

Parent

Frères et sœur

Camarades de classe

Camarade de chambre

Moi même

Sommaire

Remerciements	I
Sommaire	II
Liste des figures	III
Liste des tableaux.....	IV
Liste des formules	V
Liste des sigles et des acronymes.....	VI
Introduction générale	1
Partie 1 : Fondements théoriques de l'intelligence artificielle appliquée à la finance.....	4
Introduction de partie	4
Chapitre 1 : Généralités et théories de l'intelligence artificielle.....	5
Chapitre 2 : Revue des travaux de recherche de l'IA appliquée à la finance	30
Conclusion de partie	37
Partie 2 : Conception et développement des outils d'IA appliquée à l'analyse financières	38
Introduction de partie	38
Chapitre 3 : Analyse et développement de modèles prédictifs	39
Chapitre 4 : Conception du Chatbot pour l'interrogation des états financiers.....	60
Conclusion de partie	75
Conclusion générale et perspectives	76
Bibliographies	A
Webographies	D
Annexes.....	E
Table des matières.....	CC

Liste des figures

Figure 1: Le neurone biologique (Source : Reche, 2019)	9
Figure 2 : Zoom sur la dérivée d'une fonction (Source : mathisfun)	13
Figure 3 : Les variations de Learning Rate (Source : Cayla, 2021).....	17
Figure 4 : Simple reseau de neurone (Source : Kumar, 2020).....	18
Figure 5 : La fonction sigmoid (Source : Saleem, 2023).....	19
Figure 6 : Différents degrés de la régression polynomiale (Source : Graph of polynomial functions, 2020)	20
Figure 7 : Structure d'un arbre de decision (Source : Zhao, 2021)	23
Figure 8 : Une foret aleatoir (Source : Montalvo, 2023)	25
Figure 9 : Representation de la fonction XOR (Source : Oman, 2016).....	27
Figure 10 : Structure de réseau de neurones (Source : Ahmad, 2020)	27
Figure 11 : Reseau d'un CNN (Source : Shahriar, 2023).....	28
Figure 12 : Réseau d'un RNN (Source : Poudel, 2023)	29
Figure 13 : Analyse verticale du compte de résultat.....	43
Figure 14 : Analyse du comportement des flux de trésorerie	44
Figure 15 : Représentation graphique des ratios de profitabilité	45
Figure 16 : Page d'accueil de l'application	49
Figure 17 : Analyse des ratios de profitabilité	50
Figure 18 : Analyse prédictive des équilibres financiers	50
Figure 19 : La prédition du compte de résultat sur 5 ans	51
Figure 20 : Le dataset.....	53
Figure 21 : Exemple conversation 1	69
Figure 22 : Exemple conversation 2	69
Figure 23 : Exemple conversation 3	70
Figure 24 : Exemple conversation 4	70
Figure 25 : Architecture globale du travail	71

Liste des tableaux

Tableau 1 : Bilan	40
Tableau 2 : Compte de résultat	41
Tableau 3 : Tableau des flux de trésorerie	42
Tableau 4 : Les familles de ratios	45
Tableau 5 : Résultat chiffre d'affaires	57
Tableau 6 : Résultat excès brut d'exploitation	57
Tableau 7 : Résultat du résultat net.....	57
Tableau 8 : Résultat total actif	58
Tableau 9 : Résultat capitaux propres.....	58
Tableau 10 : Résultat SVM.....	72
Tableau 11 : Résultat random forest	72
Tableau 12 : Résultat arbre de décision	73
Tableau 13 ; Résultat naive bayes.....	73
Tableau 14 : Résultat gradient boost.....	73
Tableau 15 : Résultat KNN.....	74
Tableau 16 Ratios de probabilité de défauts	O
Tableau 17 Zones d'Altman	P
Tableau 18 Resultat XOR	S
Tableau 19 Exemple données pour arbre de décision.....	T
Tableau 20 Exemple données pour K-means.....	V

Liste des formules

Équation 1 : Mean Square Error	16
Équation 2 : Mise à jour des poids	17
Équation 3 : Le chainage des dérives partielles	18
Équation 4 : Calcul de la sortie observee.....	19
Équation 5 : Équation 5 : La fonction sigmoid.....	19
Équation 6 : Calcul de probabilité	22
Équation 7 : Probabilité conditionnelle.....	22
Équation 8 : Normalisation de probabilité	23
Équation 9 : Entropie	24
Équation 10 : Gain d'information.....	24
Équation 11: Calcul de la croissance	44
Équation 12 : Calcul TF-IDF	64
Équation 13 Zscore d'Altman	P

Liste des sigles et des acronymes

AF	Analyse financière
ANN	Artificial Neural Network
API	Application Programming Interface
ANSD	Agence Nationale des Statistiques et de la Démographie
BAC	Baccalauréat
BFG	Besoin de Financement Global
BFR	Besoin de Fond de Roulement
BRVM	Bourse Régionale des Valeurs Mobilières
CA	Chiffres d'affaires
CNN	Convolutional Neural Network
CSV	Comma Separated Values
CTIC	Incubateur des Startups
DL	Deep Learning
EBE	Excès Brute d'Exploitation
F CFA	Franc Communauté Financière Africaine
FTAF	Flux de Trésorerie des Activités de Financement
FTAI	Flux de Trésorerie des Activités d'Investissement
FTAO	Flux de Trésorerie des Activités Opérationnelles
GAFAM	Google, Amazon, Facebook, Apple, Microsoft
GI	Gain d'Information
GUI	Graphical User Interface
HAO	Hors Activité Ordinaire
IA	Intelligence Artificielle
IEEE	Institute of Electrical and Electronics Engineers
JSON	JavaScript Object Notation
LLM	Large language model
LR	Learning Rate
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MIAGE	Méthodes Information Appliquées à la Gestion
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MSE	Mean Squared Error
MVC	Model View Controller
MySQL	My Structure Query Language
NB	Naïve Bayes
NLP	Natural Language Processing
PDF	Portable Document Format
REP	Résultat Exceptionnel
REST	REpresentational State Transfer
REX	Résultat d'exploitation
RF	Résultat Financier

RN	Résultat net
RNN	Recurrent Neural Network
SGBD	Système de Gestion de Base de Données
SIG	Solde Intermédiaire de Gestion
SOAP	Simple Object Access Protocol
SML	Supervised Machine Learning
SQL	Structure Query Language
SVM	Support Vector Machine
TFT	Tableau des Flux de Trésorerie
UEMOA	Union Economique et Monétaire Ouest-Africaine
UML	Unsupervised Machine Learning
VA	Valeur Ajoutée
XAMPP	Cross-Platform, Apache, MySQL, PHP, and Perl
XML	eXtensible Markup Language
XOR	eXclusive OR

Introduction générale

Dans les années 50, s'est tenue une conférence qui avait rassemblé plusieurs chercheurs de l'époque sur un domaine dont eux-mêmes n'avaient pas conscience de comment cela allait révolutionner le monde. Parmi ce florilège de scientifiques se trouvait un mathématicien du nom de John McCarthy, organisateur par ailleurs de cette conférence, qui s'est illustré d'une manière simple : il a tout simplement proposé le terme « *Artificial Intelligence* » pour décrire cette nouvelle science qui était en train d'émerger. C'était la conférence de Dartmouth dans l'Etat du New Hampshire aux Etats-Unis en 1956. Toutes les personnes qui étaient présentes dans cette conférence venaient d'assister non pas à la naissance de l'intelligence artificielle mais au baptême de cette dernière.

Ce domaine qui est l'intelligence artificielle s'applique aujourd'hui dans plusieurs secteurs de nos vies notamment la finance qui va nous intéresser pour notre mémoire. C'est ainsi que nous avons choisi comme sujet de mémoire « **Développement d'un modèle de Machine Learning pour faire une analyse financière (historique et prédictive) et le développement d'un Chatbot pour interroger les états financiers** ». Nous serons amenées à utiliser des termes techniques comme Machine Learning qui est l'apprentissage des machines, Deep Learning qui représente quant à lui l'apprentissage profond des machines, NLP qui regroupe les méthodes mathématico-informatiques pour faire du calcul avec du texte et aussi l'analyse financière qui est un sous-domaine de la finance d'entreprise nous permettant de consulter la santé financière d'une entreprise. Tous ces concepts sont plus qu'important pour mener à bien notre mission dans ce travail de rédaction.

Ayant toujours eus une affection particulière pour l'informatique, nous nous sommes naturellement orientés vers ce domaine. Après les premiers cours d'intelligence artificielle, l'affection de l'informatique s'est renforcée puisque nous avons eu la chance de démystifier ce domaine complexe et très intéressant. Cette amour de l'informatique a engendré une passion de l'IA et ses technologies et nous avons naturellement suivi le chemin qui nous mène dans ce sens. Etant un étudiant de la MIAGE (**Méthodes Informatiques Appliquées à la Gestion**), il s'est avéré être pertinent de faire appliquer l'IA au domaine de la finance, c'est ainsi que nous avons choisi avec l'aide de nos professeurs encadreurs la finance d'entreprise comme domaine d'application de l'IA.

L'intelligence artificielle n'est pas une science nouvelle comme nous l'avons déjà vue même si cette dernière gagne beaucoup de popularité ces derniers temps. Et notre pays le Sénégal n'est pas en reste par rapport à tout cela, plusieurs scientifiques s'illustrent dans ce domaine depuis fort longtemps. Mais plus proche de nous, beaucoup d'initiatives ont été prise dans le sens de l'IA, il y a l'Agence Nationale des Statistiques et de la Démographie (ANSD) qui a ouvert un bureau d'IA pour la prédiction démographique. Cela ne s'arrête pas là, même le gouvernement du Sénégal a lancé un programme appelé « La stratégie IA » à travers le ministère de la communication des télécommunications et du numériques. En plus de tout cela vient s'ajouter un bon nombre de chercheurs et de jeunes passionnés qui essayent tant bien que mal de faire bénéficier ses technologies intelligentes à la population sénégalaise.

La question principale que ce travail de mémoire aura pour but de répondre est : **Quels modèles de Machine Learning pour une analyse et une interrogation des états financiers des sociétés cotées à la bourse régionale des valeurs mobilières (BRVM)** ? A cette question, va venir s'imbriquer d'autres questions subsidiaires qui vont rendre plus fluide le travail qui va être abattu. A partir de cette question, nous allons façonnner tout le travail de tel sorte que chaque partie du travail réponde à une question subsidiaire. De ce fait, la réponse à notre question de recherche apparaitra en entier une fois tous les blocs mis ensemble. De là, nous allons être en mesure de savoir si nos entreprise ouest-africaines cotées à la BRVM pourront réellement bénéficier des avantages qu'apporte l'IA. Si nous parvenons à répondre par l'affirmative à cette question centrale de recherche, il sera convenu de manière presque unanime de l'importance que cela représente pour les entreprises africaines, mais aussi l'entreprenariat en général.

L'intérêt du travail sera d'utiliser les modèles intelligents afin qu'ils aident à faciliter le travail des financiers en général. L'analyse financière se fait généralement par des logiciels généralistes qui font un peu de tout, mais dans notre cas, il suffira d'entrer les états financiers et de constater les résultats. En plus de la possibilité d'une analyse avec les valeurs historiques, il doit aussi être possible d'effectuer une analyse prédictive. Ajouté à tout cela, le fait que, par le biais d'une question, de recevoir des éléments des états financiers à travers un Chatbot.

Afin de mener cette mission à bien, nous adopterons une démarche bien spécifique, tout d'abord nous nous attelerons à trouver des données avec lesquelles nous allons travailler. La donnée est là, mais est-elle disponible, pour ce qui est des données financières, nous savons déjà où aller, mais

de type texte, il y a du travail à faire. Une fois les données collectées et traitées, nous passerons par la suite à ce qu'on appelle la recherche en grille qui est une méthode utilisée en Machine Learning pour déterminer le meilleur modèle, celui qui sera le plus adapté à nos données. Il y a plusieurs modèles de ML et notre travail consistera à trouver le plus adapté à nos données en association avec les hyperparamètres qui conviennent le mieux. Deux familles de modèles vont être utilisées, il y a les modèles de prédiction et de classification mais aussi les techniques de NLP (faire comprendre le texte à un ordinateur).

C'est ainsi à la fin de ce travail deux applications vont être produites sous forme de logiciel que les entreprises pourront utiliser pour faire leur analyse financière, prédire leurs états financiers, interroger un Chatbot sur leurs états financiers etc. Et tout cela dans un environnement cousu à la taille de leur finance. Ces deux applications vont être déployées dans le réseau local de l'entreprise pour l'intégrité des données.

Dans le but de réaliser ce mémoire de manière efficace, notre document va être divisé en partie, il va y avoir deux parties : fondements théoriques de l'intelligence artificielle appliquée à la finance (partie 1) et Conception et développement des outils d'IA appliquée à l'analyse financières (partie 2). Ensuite chaque partie se verra divisée en chapitre, ce qui va nous faire quatre chapitres. Nous allons d'abord voir les généralités et fondement théoriques de l'IA (Chapitre I). En plus de cela nous réservons une partie spéciale pour toute la littérature sur l'intelligence artificielle appliquée à la finance, les découvertes, les tendances, les perspectives et bien d'autres (Chapitre II). Pour commencer le développement des modèles, il va être consacrer un chapitre sur l'analyse prédictive des états financiers (Chapitre III). Nous allons terminer avec de développement du Chatbot et toutes les technologies à découvrir à travers son implémentation (Chapitre IV).

Partie 1 : Fondements théoriques de l'intelligence artificielle appliquée à la finance

Introduction de partie

Après notre introduction générale, il devra d'abord être fait la partie 1 de notre document qui se veut être une aventure vers le soubassement de l'IA. Il serait bien entendu impératif de connaître quelques notions liées à l'IA afin de mieux aborder la suite des aspects pratiques. De ce fait, cette partie aura pour objectif de faire connaissance avec l'IA dans toutes les dimensions qu'elle regorge.

Dans le but d'atteindre cet objectif, nous allons diviser cette partie en chapitres, d'abord les généralités et théories de l'IA où nous allons évoquer la question de la définition de l'IA. Puis la revue de la littérature scientifique de l'IA appliquée à la finance, dans laquelle nous aborderons les applications de l'IA sur la finance dans un premier temps, puis ensuite les travaux de l'IA sur l'analyse des états financiers.

Chapitre 1 : Généralités et théories de l'intelligence artificielle

En premier lieu, nous allons aborder quelques notions théoriques de l'IA, il serait bien évidemment intéressant et important de comprendre la signification de l'intelligence artificielle, son fonctionnement, ses origines et bien d'autres avant d'aborder les aspects pratiques. L'importance que relève ce chapitre est de lever l'ambiguité sur notre science, car certaines personnes usent de ces termes à tort et à travers. Nous allons sans plus tarder commencer l'explication liée à tout cela.

Section 1 : L'intelligence artificielle : Définitions, origines et évolutions

Dans cette section, il sera fait un inventaire des définitions proposées par des scientifiques de toutes origines et toutes époques. Nous allons voir qu'il y a beaucoup de subtilités liées à certains termes que nous entendons tous les jours. Une fois ce travail fait, nous allons prendre la machine à voyager dans le temps afin d'établir une timeline des grands faits qui ont marqué l'évolution de l'IA.

1. *Définitions de l'intelligence artificielle*

Avant d'entrer dans les détails, dans les aspects techniques et scientifiques ou dans l'implémentation d'une IA, il serait bien de donner une vue globale de ce qu'est une intelligence artificielle. Fort heureusement, beaucoup de recherches et d'études ont été faites par les scientifiques académiciens sur ces termes que nous nous donnons la tâche de définir.

- **C'est quoi l'intelligence ?**

L'intelligence humaine est un concept qui est difficile à définir car on ne sait pas comment il fonctionne, où est son siège dans le cerveau, et on ne peut pas vraiment voir de différence notable entre le cerveau d'une personne intelligente et celui d'une personne qui l'est moins.

Néanmoins, cette difficulté n'a pas empêché les chercheurs en psychologie d'essayer de définir la chose selon leur entendement et la manière qu'a l'Homme d'interagir avec son environnement. Les avis des scientifiques sont divers et variés. Dans son article de 1993 publié au British Journal of Psychology « On What Intelligence Is », Robert W. Howard nous fait un inventaire de définitions proposées par des psychologues avant lui.

“The word ‘intelligence’ labels three different major concepts: g, the sum of an individual’s knowledge and skills, and the specific mental abilities important in a given culture” (Jensen, 1987).

“Intelligence is not an entity within the organism but a quality of behavior” (Anastasi, 1986).

Mais ce ne sont pas seulement les psychologues qui ont tenté de définir l'intelligence, les chercheurs en IA aussi, c'est le cas de James S. Albus qui la définit comme : “. . . the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success, and success is the achievement of behavioral subgoals that support the system's ultimate goal”.

“Intelligence constitutes the state of equilibrium towards which tend all the successive adaptations of a sensori-motor and cognitive nature, as well as all assimilatory and accommodatory interactions between the organism and the environment” (Piaget, 2005).

Cette définition nous renvoie à l'individu et son environnement, cet individu prend les données de l'environnement (Inputs) et réagit en conséquence (Outputs).

- **Proposition de définitions de l'intelligence artificielle**

Nous y voilà, les termes que nous avons décidés de donner des définitions vont nous permettre de définir ce sur quoi porte notre sujet de mémoire. Dans cette partie, nous allons essayer de répondre à la question : c'est quoi une IA ? Nous allons voir que plusieurs scientifiques ont donné des définitions, mais à la fin, c'est plus ou moins les mêmes.

Déjà en 1988, Asa SIMMONS et Steven CHAPPEL avaient publié un article dans le IEEE Journal of Oceanic Engineering sur lequel ils nous rappelaient la définition qu'avait donnée Haugeland en 1985 : “The fundamental goal of this research IS not merely to mimic intelligence or produce some clever fate. Not at all. AI wants only the genuine article: machines with minds, in the full and literal sense”.

Vingt-quatre (24) ans plus tard, en 2012, le mathématicien et docteur en IA américain Matt L. Ginsberg donnait, dans son ouvrage « Universal intelligence : A definition of machine intelligence », la définition suivante : “Artificial Intelligence is the enterprise of constructing an artefact that can reliably pass the Turing test”.

Maintenant plus proche de nous, au moment où ce mémoire est en train d'être écrit, de nouvelles définitions émergent. En janvier 2023, Haroon Sheikh, Corien Prins et Erik Schrijvers ont copublié un article « Artificial Intelligence : Definition and Background » où ils ont défini la chose comme

suit : “Systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.”

Nous voyons que plusieurs auteurs à travers le temps ont donné leurs définitions de l'IA selon leur entendement de la chose et leurs domaines de recherche. Au vu de tout cela, nous pouvons conclure que l'IA a pour objectif d'imiter l'intelligence humaine en faisant des tâches qui auraient pu être jugées impossibles à faire pour les machines.

2. *Historique de l'intelligence artificielle*

L'IA est une science intrigante, voire mystérieuse, qui, pour certaines personnes, est une boîte noire. Et bien des choses se sont passées pour qu'elle devienne l'une des sciences les plus populaires et qui fait peur, disons-le, a beaucoup de professionnels. Son histoire est riche et rocambolesque ; nous pourrions facilement en faire un film. Pour ce qui suit, il sera fait un bref historique des événements marquants de l'IA, de 1943 (naissance présumée de l'IA) à nos jours.

2.1. Genèse de l'IA : le premier neurone artificiel

Pour beaucoup de chercheurs dans le domaine de l'IA, la naissance de notre science pourrait remonter en 1943 avec le premier article publié dans ce domaine.

Durant cette année, une étude menée par Warren S. McCulloch et Walter Pitts pour une expérimentation mathématique du neurone biologique a vu le jour. L'objectif de cette recherche était de mettre en évidence la simulation du fonctionnement du neurone biologique avec de l'analyse mathématique (*Calculus*). La méthodologie suivie par ces deux scientifiques était, pour chaque étape de la transmission d'information d'un neurone à un autre, de trouver une fonction mathématique qui pourrait le répliquer. À la fin, ils ont pu trouver les calculs nécessaires pour reproduire un tant soit peu le fonctionnement du neurone biologique. Les implications de cette étude sont énormes, car elles sont à la base de tous les réseaux de neurones que nous utilisons aujourd'hui, du perceptron aux ANN (*Artificial Neural Network*).

Cette étude, l'une des toutes premières dans notre domaine, va avoir un impact considérable sur les futurs réseaux de neurones. Cependant, elle n'est pas exhaustive ; plus tard, Donald O. Hebb crée l'apprentissage pour les réseaux de neurones, et après, Frank Rosenblatt va créer le perceptron en s'appuyant sur les travaux de ses prédécesseurs, et ainsi de suite.

2.2. Evolution

Depuis la création du premier neurone qui constitue le premier article scientifique s'inscrivant dans l'intelligence artificielle, bien des choses se sont passées pour aboutir au statu quo. Voici une timeline d'événements marquants qui marque l'évolution de notre science.

1949 : Donald Hebb développe le premier algorithme d'apprentissage dans les réseaux de neurones.

1950 : Alan Turing publie un article « Computing Machinery and Intelligence » où il sort le Turing test, qui se veut être une mesure d'intelligence d'une machine et une réponse à la question « Can machines think ? ».

1965 : L'un des premiers chatbots (ELIZA), capable de s'exprimer, a été créé par Joseph Weizenbaum (MIT) et est l'une des premières machines à avoir quelque peu réussi le test de Turing.

1980 : Lisp Machine développe et commercialise le premier système expert.

La rétropropagation commence à être largement utilisée dans les réseaux de neurones.

1993 : Rodney Brooks et ses collaborateurs développent le premier robot humanoïde.

2005 : Honda développe ASIMO : un robot humanoïde et artificiellement intelligent, capable de faire des tâches propres à l'homme.

2016 : AlphaGo devient la première société à avoir réussi à créer une IA impossible à battre dans le jeu du Go.

2022 : OpenAI lance ChatGPT et Google lance Google Bard (actuellement Gemini).

2.3. Les sciences qui ont impulsé sa dynamique

• **Les mathématiques**

“These considerations show that there is a tremendous need for mathematics in the area of artificial intelligence. And, in fact, one can currently witness that numerous mathematicians move to this field, bringing in their own expertise” (Kutyniok, 2022).

Les mathématiques constituent le soubassement de l'IA, comme vient de le rappeler Gitta Kutyniok, c'est-à-dire que tous les algorithmes d'IA reposent sur des théories mathématiques. D'ailleurs, nous allons voir cela en détails dans la suite du document, mais pour le moment, donnons quelques exemples : les dérivées, l'algèbre linéaire, les probabilités et statistiques.

- **La biologie**

Pour parler de l'impact de la biologie dans l'intelligence artificielle, il nous faut forcément parler du neurone biologique. Le neurone est une cellule spécialisée dans le traitement et la transmission de l'information, ce qui est reflété par sa morphologie très particulière. Il se compose d'un corps cellulaire, le soma, et d'expansions : les dendrites d'une part, et l'axone d'autre part. Les dendrites forment des ramifications qui rentrent en contact avec d'autres neurones, typiquement de l'ordre de 10000, et dont le rôle est de recevoir des informations, électriques ou chimiques. L'axone est un prolongement de la cellule, typiquement long de quelques millimètres, qui conduit un signal électrique jusqu'à son arborisation terminale, où il peut alors entrer en contact avec les dendrites d'autres neurones. La jonction axone-dendrite est appelée synapse (Brette, 2003). C'est cette structure du neurone biologique que les réseaux de neurones artificiels vont essayer de répliquer à travers des calculs mathématiques.

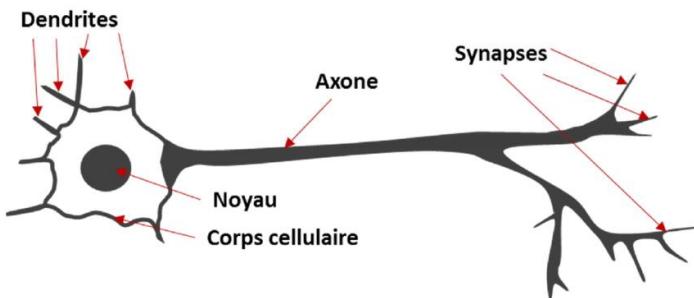


Figure 1: Le neurone biologique (Source : Reche, 2019)

- **L'informatique**

Là, un non-initié pourrait dire que l'IA est un sous-domaine de l'informatique, que nenni ! Nous allons voir qu'il y a des subtilités ; on peut dissocier ces termes, même s'ils sont bien liés et vont le rester dans le temps.

C'est ainsi que nous allons donner un petit aperçu de comment marche un algorithme d'IA pour comprendre cela. Un réseau de neurones prend des données en entrée, les normalise, puis il va les

passer à la prochaine couche à travers la multiplication avec les poids. Cette opération va se répéter autant de fois que nécessaire, et enfin, on va avoir une sortie.

Dans ce processus, il n'y a que des calculs mathématiques, ce qui veut dire qu'on peut faire cela sur une feuille (chose qui va certainement prendre beaucoup, beaucoup de temps), ou le faire aussi avec une calculette, ou tout support nous permettant de faire des calculs.

Tout cela pour dire que l'ordinateur nous sert tout simplement à faciliter les calculs et à afficher les résultats dans une interface graphique, du fait qu'il est la calculatrice la plus puissante. Mais encore, nous pouvons noter que Warren S. McCulloch et Walter Pitts n'avaient pas utilisé d'ordinateur pour réaliser le premier réseau de neurones.

- **La cybernétique**

La cybernétique est une science qui dérive de plusieurs autres sciences ; c'est le mathématicien américain Norbert Wiener à qui l'on attribue la création de cette science. La cybernétique essaie de répondre à la question de savoir comment les systèmes peuvent se contrôler eux-mêmes. L'exemple qui est très souvent donné est celui du thermostat, qui récupère les informations de son environnement, c'est-à-dire la température ambiante, et se régule lui-même en s'ajustant à la température désirée.

Cette science a beaucoup apporté à l'IA, car le fonctionnement de plusieurs modèles intelligents s'inspire des théories de la cybernétique, comme le fait d'utiliser sa propre erreur pour se rectifier soi-même dans un réseau de neurones ou encore le fait de donner un bonus ou malus à un agent dans un environnement s'il fait le bon ou le mauvais choix dans un *Reinforcement Learning*.

Section 2 : Fondements théoriques des algorithmes d'intelligence artificielle

À ce moment du travail, nous avons une certaine compréhension de certains aspects de l'IA. C'est déjà bien ; cependant, nous pouvons aller plus loin en ouvrant la boîte noire que constitue cette science. Dans cette section, il s'agira de la prise de conscience des fondamentaux mathématiques et informatiques de l'IA ; ensuite, nous allons voir le fonctionnement de certains algorithmes des modèles intelligents qui façonnent nos vies.

1. Les prérequis de l'intelligence artificielle

Bien évidemment, il y a quelques prérequis, deux pour être précis : les mathématiques et l'algorithmique. À part les mathématiques et l'informatique, il y a quelques autres prérequis qui ne sont pas nécessaires mais peuvent aider dans notre objectif, et tous ces autres prérequis vont être classés dans le domaine de l'intelligence sociale.

1.1. Les mathématiques

Quand on parle de mathématiques, la plupart des gens vont prendre peur, abandonner, voire même fuir. Mais ici, nous allons voir les concepts mathématiques qui nous seront utiles pour l'IA, mais de manière simple et concise.

Les mathématiques peuvent être compliquées, mais quand on leur trouve une application, c'est là que ça devient intéressant, et l'une des plus belles applications des mathématiques, c'est l'IA. Nous allons vous montrer à quel point il est fascinant de résoudre des problèmes mathématiques pour créer des modèles intelligents. Les mathématiques sont plus que nécessaires pour l'IA, elles sont vitales. D'ailleurs, mon professeur d'intelligence artificielle nous disait à la fin d'un cours : « L'intelligence artificielle n'est ni plus ni moins que des calculs mathématiques. »

En dépit du fait qu'il y a plusieurs domaines mathématiques qui nous seront utiles dans l'IA, pour ce travail de mémoire, nous allons nous concentrer sur seulement trois (3) domaines des mathématiques : les statistiques et probabilités, l'algèbre linéaire et l'analyse.

• Les statistiques et probabilités

Is everything on this planet determined by randomness? This question is open to philosophy debate. What is certain is that every day thousands and thousands of engineers, scientists, business persons, manufacturers, and others are using tools from probability and statistics. (Dekking, Frederik Michel, 2005).

Cette citation de Michel nous renvoie à l'importance de ces domaines dans nos vies de tous les jours, et l'IA ne fait pas exception. La statistique est un domaine des mathématiques qui travaille sur des données en les faisant parler, ce qui nous permet de mieux comprendre les valeurs d'une base de données. C'est ce qu'on appelle les statistiques descriptives. Il y a aussi les statistiques inférentielles qui, comme son nom l'indique, vont nous permettre de faire des inférences, c'est-à-dire faire des estimations. Et c'est là que réside le lien entre les probabilités et les statistiques, car

les statistiques inférentielles vont avoir besoin des probabilités. La probabilité est l'étude de la chance qu'un évènement se produise, pour faire simple.

Ceci étant dit, comment ces deux sont utiles en Machine Learning et Deep Learning ? Ils interviennent tous les deux avant et après le développement du modèle d'IA.

- **Avant le développement du modèle** : les statistiques nous aident à comprendre les données, car très souvent les données brutes ne sont pas exploitables. Ici, nous vérifions le maximum des valeurs, le minimum, la moyenne, les outliers et l'une partie des plus importantes du « Feature Engineering » la mise à l'échelle etc.
- **Après le développement du modèle** : il va bien falloir calculer la fiabilité du modèle, ce qu'on appelle « accuracy », il faut calculer aussi, la précision, le *f1-score*, le *recall*... Ces derniers nous permettent d'apprécier la robustesse du modèle une fois déployer.

- **L'algèbre linéaire**

Au fait, il y a trois grandes parties dans le développement d'un réseau de neurones et, à titre illustratif, nous pouvons dire qu'il y a le travail a posteriori, le développement du modèle et le travail a priori. Pour le modèle, il y a deux parties : le Feed-forward et le Back-propagation, et l'algèbre linéaire va intervenir dans ces deux parties.

L'algèbre linéaire est la branche des mathématiques qui s'intéresse à l'étude des espaces vectoriels (ou espaces linéaires), de leurs éléments, les vecteurs, des transformations linéaires et des systèmes d'équations linéaires (théorie des matrices). (Algèbre linéaire - Définition, 2024)

Ainsi, la plus grande utilité de l'algèbre linéaire est le calcul des poids ; elle va nous permettre d'automatiser les calculs lourds et coûteux, nous permettant ainsi de gagner du temps. Sans elle, nous aurions passé beaucoup de temps sur ces calculs. Je rappelle qu'un réseau de neurones a des milliers de neurones d'input, plusieurs couches cachées qui peuvent elles-mêmes avoir des milliers de neurones. C'est juste impossible de calculer tout cela de manière séquentielle.

- **L'analyse**

Quand on parle d'apprentissage en IA, ou du terme très populaire en anglais « Learning », c'est grâce au domaine des mathématiques, l'analyse, et au calcul des dérivées. Si l'IA est devenue ce

qu'elle est aujourd'hui avec les performances qu'on lui connaît, c'est en très grande partie grâce aux calculs d'analyse mathématique.

La dérivée d'une fonction nous informe sur la variation de la fonction en un point donné. Pour être plus clair, elle nous permet de calculer la pente en n'importe quel point de la fonction. Le calcul de la dérivée est très important dans de nombreux domaines, notamment dans le Deep Learning.

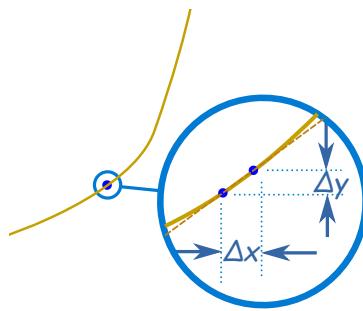


Figure 2 : Zoom sur la dérivée d'une fonction (Source : [mathisfun](#))

Comment se passe l'apprentissage dans un réseau de neurones ? Nous allons rectifier les erreurs commises par l'IA durant son entraînement à l'aide de la dérivée. Si nous répétons cela autant de fois que nécessaire, l'erreur d'assomption sera réduite au minimum et la « accuracy » sera maximisée. Ce qu'il faut comprendre ici, c'est que le calcul de la dérivée de la fonction d'erreur nous permet de rectifier cette erreur.

De manière pratique, on calcule d'abord l'erreur, puis on calcule la dérivée de la fonction d'erreur. La manière dont la rectification se fait est que chaque poids reçoit une valeur correspondant à sa responsabilité dans l'erreur. C'est cela le Back-propagation, ou rétropropagation en français.

1.2. L'informatique

L'informatique, c'est la science de l'automatisation de l'information, d'ailleurs son nom vient de là : une contraction entre 'information' et 'automatique'. Chez les anglo-saxons, on parle plutôt de 'Computer Science', qui se traduit littéralement par 'science de l'ordinateur'. Plus haut, nous avons mentionné que l'IA est une science purement mathématique, avec uniquement des calculs que l'on pourrait même faire sur une feuille. Dès lors, que représente l'informatique pour l'IA ? Elle joue le rôle d'une calculatrice géante, capable d'effectuer des super calculs en un temps record, tout en servant également d'interface graphique

- **L'algorithme**

Bien évidemment, la première des choses que nous allons aborder est l'algorithme. Nous pouvons affirmer, sans prendre trop de risques, que l'algorithme est l'informatique, et que l'informatique est l'algorithme. Un algorithme est un ensemble d'étapes à suivre pour résoudre un problème informatique. L'analogie fréquemment utilisée est celle de la recette de cuisine, et c'est tout à fait pertinent.

La raison pour laquelle il est crucial de maîtriser l'algorithme est la suivante : pour implémenter un problème mathématique dans un ordinateur, il est essentiel de savoir comment s'y prendre et quelles étapes suivre. Sans cela, beaucoup de frustration nous attend.

Exemple : écrivons un algorithme qui résout un polynôme du second degré.

- Afficher : Donner les valeurs a, b et c.
- Stocker a, b et c dans des variables.
- Calculer delta ($\delta = b^2 - 4 * a * c$)
- Si delta positif alors $x_1 = (-b - \sqrt{\delta}) / 2 * a$ et $x_2 = (-b + \sqrt{\delta}) / 2 * a$
- Si delta nul alors $x = \sqrt{\delta} / 2 * a$
- Si delta négatif alors il n'y a pas de solution dans \mathbb{R} .

Voici ci-dessus un algorithme qui marche pour un polynôme du second degré et cette même manière de réflexion peut nous permettre d'implémenter n'importe quel problème déjà résolu en mathématique en algorithme informatique.

• **Les structures de données**

D'abord, les structures de données désignent les différentes façons de modéliser les données avec lesquelles nous travaillons en informatique. Très souvent, pour ne pas dire tout le temps, nous n'avons pas directement la méthode optimale de gestion des données.

Ces structures peuvent aller d'un simple tableau dans un langage de programmation jusqu'aux graphes (une structure de données complexe et très puissante). Comme nous l'avons souligné à plusieurs reprises, l'IA travaille sur des données. Citons quelques exemples de structures de données : les listes chaînées, les tables de hachage, les arbres, les piles et files, les graphes, etc.

- **Les langages de programmation**

Les langages de programmation, aussi appelés langages informatiques, sont des syntaxes qui traduisent les algorithmes d'une manière compréhensible par l'ordinateur. Il est important de préciser que l'ordinateur ne comprend pas le texte ; il comprend seulement les chiffres, c'est-à-dire le langage binaire. Ce que le langage de programmation fait, c'est de convertir sa syntaxe en langage binaire compréhensible par l'ordinateur, et chaque langage a sa propre syntaxe.

L'importance des langages de programmation est évidente pour tout le monde. Parmi les langages les plus utilisés en IA, nous pouvons citer : C/C++ (important pour l'IA), Python (très utilisé en IA), Java, PHP, JavaScript, etc.

2. Les algorithmes d'intelligence artificielle

Nous avons parlé de l'IA dans ce document, mais cette fois, nous allons voir comment elle fonctionne en parcourant différents des plus importants algorithmes d'IA, ceux qui sont vraiment utilisés par les grandes entreprises. Donc, pour cette partie, je vais vous demander une attention particulière, car ce sera très intéressant. Alerte âme sensible !!! Il y aura beaucoup de calculs mathématiques dans cette partie.

2.1. Machine Learning

Littéralement, Machine Learning veut dire apprentissage des machines. Comme nous, êtres humains, nous naissions sans connaissance dans notre tête, mais en regardant notre environnement et en imitant nos parents, nous apprenons. Ce processus peut être répliqué sur un ordinateur, c'est le Machine Learning, il y en a deux : Supervised Machine Learning (SML) et Unsupervised Machine Learning (UML).

- **Supervised learning**

Si nous reprenons l'analogie de l'enfant, dans sa phase d'apprentissage, ses parents vont être derrière lui et le guider. Si l'enfant commet des erreurs, ses parents vont le rectifier ; s'il fait une bonne chose, ses parents vont le récompenser ou l'encenser.

Dans le domaine des ordinateurs, pour faire en sorte qu'une machine apprenne, on aura besoin de données, beaucoup de données. Et chaque ligne de données va être étiquetée, on parle d'input et

d'output. Maintenant, le modèle va essayer de s'adapter à tous les inputs et leurs outputs. Nous allons voir dans la suite les différents types d'apprentissage supervisé et leurs algorithmes.

- **La régression**

La régression est une méthode statistique qui nous permet d'approximer la valeur d'une variable à partir des valeurs déjà présentes et connues. Elle va se faire en traçant une courbe qui représente le mieux la relation des points dans un repère orthonormal. Il y a plusieurs types de régression, mais nous allons en voir trois (3).

- **La régression linéaire**

La régression linéaire nous permet de tracer une droite qui va au mieux s'adapter aux données d'une courbe. Maintenant, si nous voulons tracer une droite qui va au mieux représenter l'évolution de ces points, qu'allons-nous faire ? Il y a la méthode des moindres carrés, élaborée par le légendaire Carl Friedrich Gauss, qui est une méthode purement statistique, mais nous allons utiliser une méthode d'IA avec la descente des gradients. Cette dernière méthode peut être divisée en trois parties :

- ❖ Forward propagation (essaie au hasard)

D'abord, la courbe que l'on veut tracer va être de la forme $y = ax + b$, mais dans le jargon, on va parler de w_0 et w_1 qui vont représenter les poids respectifs. L'équation devient $y = w_0.x + w_1$. Le but du jeu est de trouver les w_0 et w_1 qui vont au mieux correspondre à nos points. Dans un premier temps, on va leur donner des valeurs aléatoires, d'où l'essai au hasard.

- ❖ Calculer l'erreur

Très souvent, pour ne pas dire jamais, le premier essaie va être une erreur, de ce fait, il faut calculer l'erreur, nous allons utiliser la fonction suivante :

$$MSE = \frac{1}{n} \sum_{i=1}^n (y - \hat{y})^2$$

Équation 1 : Mean Square Error

MSE : Mean Square Error (la moyenne des erreurs au carré)

Y : la sortie attendue

\hat{Y} : la sortie observée

N : le nombre d'élément dans le tableau

Cependant, les plus curieux vont se demander pourquoi éléver l'erreur au carré. C'est une bonne question. La raison est simple : une erreur de -1 est égale à une erreur qui vaut 1. Et le fait de l'élèver au carré va nous aider dans la mise à jour des poids, où nous allons utiliser l'algorithme de la descente des gradients.

❖ Back-propagation (rétropropagation qui met à jour les poids)

Maintenant que nous avons l'erreur, nous pouvons enfin mettre à jour nos poids w_0 et w_1 . Cela veut dire que chacun va prendre une part de l'erreur qui est égale à sa responsabilité dans cette dernière et se rectifier lui-même. Pour ce faire, nous allons calculer la dérivée de toutes les fonctions qui nous ont menés à cette erreur de manière suivante :

$$w_0 = w_0 - lr * \frac{\partial MSE}{\partial w_0}$$

Équation 2 : Mise à jour des poids

$$w_1 = w_1 - lr * \frac{\partial MSE}{\partial w_1}$$

Dans la descente des gradients, il y a ce qu'on appelle le pas. Il va déterminer à quelle vitesse la descente va se faire. Si le pas est trop petit, l'apprentissage va être lent, et si le pas est trop grand, nous allons dépasser le point qui minimise l'erreur. Ce pas, c'est le « LR » dans les deux fonctions, cela signifie « Learning Rate ».

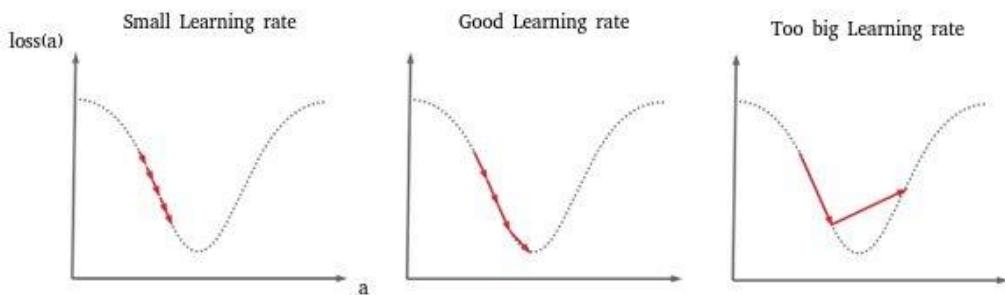


Figure 3 : Les variations de Learning Rate (Source : Cayla, 2021)

$\frac{\partial MSE}{\partial w}$ Représente quant à elle, la dérivée de la fonction MSE par rapport au poids concerné, c'est ce qu'on appelle une dérivée partielle. Exemple :

$$f(x, y) = x \cdot y$$

$$\frac{\partial f}{\partial x} = y, \quad \frac{\partial f}{\partial y} = x$$

Donc

$$\frac{\partial MSE}{\partial w_0} = \frac{\partial MSE}{\partial Y} * \frac{\partial y}{\partial w_0}$$

Équation 3 : Le chainage des dérives partielles

Voir annexes 2 pour un exemple de calcul de régression linéaire.

▪ La régression logistique

La régression logistique, contrairement à celle dite linéaire, n'a pas pour vocation de prédire une valeur future. Sa prédiction est du type binaire : oui ou non, bon ou mauvais, 0 ou 1, etc. Ceci va s'avérer être très important dans beaucoup de domaines. Nous l'utilisons dans nos vies de tous les jours sans nous en rendre compte. Par exemple, détecter si un email est un spam ou non, si une information est un fake news ou non, si un investissement va être rentable ou pas...

❖ Forward propagation

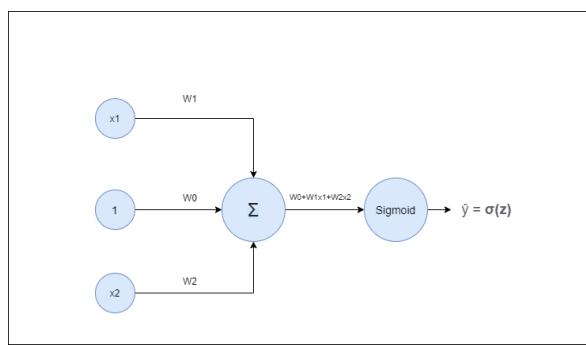


Figure 4 : Simple réseau de neurone (Source : Kumar, 2020)

Voici à quoi va ressembler notre réseau de neurones, on va ajouter un autre input en plus x_1 et x_2 , c'est le biais qui va toujours être égale à 1, son utilité est d'éviter que certains neurones ne meurent durant l'entraînement si $x_1 = 0$ et $x_2 = 0$.

$$y = \sum_{i=1}^n w_i \cdot x_i$$

Équation 4 : Calcul de la sortie observée

Puisqu'on dit que les valeurs de sortie doivent être 0 ou 1, nous devons trouver un moyen de toujours mettre à l'échelle la sortie observée, c'est là qu'intervient la fonction d'activation. Pour les problèmes de régression logistique il y en a deux très populaires : la fonction à seuil et sigmoid.

Fonction	Formule	Sortie possible
Seuil(x)	1 si $x > 1, 0$ sinon	0, 1
Sigmoid $\sigma(x)$	$\frac{1}{1 + e^{-x}}$ <i>Équation 5 : Équation 5 : La fonction sigmoid</i>	Tout réel compris en 0 et 1

Nous allons continuer avec la fonction sigmoid :

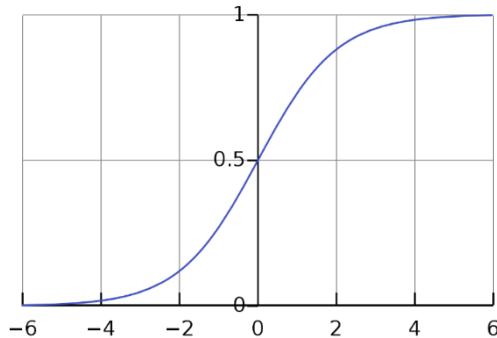


Figure 5 : La fonction sigmoid (Source : Saleem, 2023)

❖ Calculer l'erreur

Pour l'erreur rien ne va changer nous allons utiliser la Mean Square Error :

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

❖ Backpropagation

Nous voici près pour la rétropropagation, seulement ici nous allons mettre à jour trois poids à savoir w_0, w_1, w_2 .

$$w_0 = w_0 - lr * \frac{\partial MSE}{\partial w_0}, \quad w_1 = w_1 - lr * \frac{\partial MSE}{\partial w_1}$$

$$w_2 = w_2 - lr * \frac{\partial MSE}{\partial w_2}$$

Même si les formules restent les mêmes, ne prenons encore rien pour acquis. Ici, la valeur de la dérivée partielle va changer étant donné qu'on a introduit une nouvelle fonction, celle d'activation. Nous allons de facto nous retrouver avec trois membres dans le calcul de la dérivée partielle.

$$\frac{\partial MSE}{\partial w_1} = \frac{\partial MSE}{\partial \sigma} * \frac{\partial \sigma}{\partial Y} * \frac{\partial Y}{\partial w_1}$$

Voir annexe 2 pour un exemple de calcul de régression logistique.

▪ La régression polynomiale

La régression polynomiale nous permet de représenter une courbe de données qui adapte une forme exponentielle. Les étapes de régression polynomiale restent les mêmes que pour les autres algorithmes, mais ses calculs vont changer.

❖ Forward-propagation

Pour le Forward-pass de la régression polynomiale, nous allons utiliser une fonction quadratique, c'est-à-dire qui admet une puissance dans la variable. On va parler du degré de la fonction. Plus le degré est élevé, plus la fonction pourra être en mesure d'aller chercher des variations.

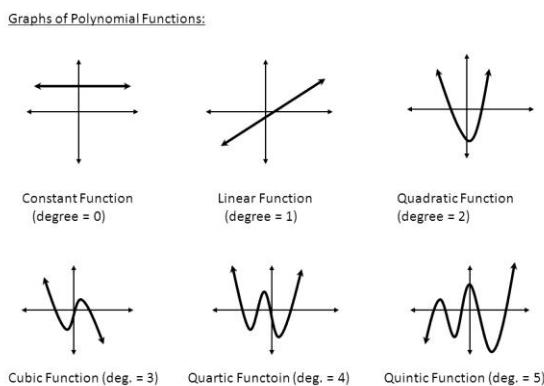


Figure 6 : Différents degrés de la régression polynomiale (Source : Graph of polynomial functions, 2020)

Du fait que nous n'avons pas beaucoup de variation dans le tableau, nous allons utiliser le deuxième degré, ainsi notre formule se présente comme suit :

$$y = w_0 + w_1 * x + w_2 * x^2$$

❖ Calculer l'erreur

La fonction d'erreur ne change toujours pas, c'est le MSE.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y - \hat{y})^2$$

❖ Backpropagation

$$w_0 = w_0 - lr * \frac{\partial MSE}{\partial w_0}, \quad w_1 = w_1 - lr * \frac{\partial MSE}{\partial w_1}, \quad w_2 = w_2 - lr * \frac{\partial MSE}{\partial w_2}$$

La valeur de la dérivée partielle pour w_1 se présente comme suit :

$$\frac{\partial MSE}{\partial w_2} = \frac{\partial MSE}{\partial Y} * \frac{\partial Y}{\partial w_2}$$

Nous avons décidé de prendre w_2 car il a la dérivée partielle la plus compliquée à calculer. Avec ce calcul établi, nous pouvons passer à l'étape des mises à jour des poids. N'oubliez pas de prendre un Learning Rate.

Voir annexe 2 pour un exemple de calcul de régression polynomiale.

○ **La classification**

La classification est un problème qui existe depuis longtemps dans le domaine de l'intelligence artificielle. Les académiciens ont fait beaucoup de recherches sur le sujet et nous ont proposé un certain nombre de méthodes.

La classification a pour objectif de déterminer les éléments qui différencient les données dans une base de données, ainsi que de ranger chacune dans sa classe de prédilection, et aussi, mais surtout, de prédire les classes pour des données non encore observées. Différents algorithmes sont aujourd'hui disponibles pour nous permettre de résoudre les problèmes de classification, mais nous allons en voir trois (3).

▪ Naïve Bayes

Le modèle de Naïve Bayes (NB) est un algorithme de ML qui nous vient des statistiques et des probabilités. Selon les cas, il peut être très puissant avec un mécanisme simple de calcul de probabilité. Il fonctionne en calculant les probabilités de toutes les valeurs d'attributs avec la variable cible.

❖ Probabilité des variables cibles

Tout d'abord il faut calculer la probabilité de toutes les variables cibles afin de savoir nos chances de tomber sur l'un ou l'autre (il est possible d'utiliser le NB dans une multi-classe classification aussi).

$$P(C_i) = \frac{\text{nombre } C_i}{N}$$

Équation 6 : Calcul de probabilité

❖ La probabilité conditionnelle des valeurs d'attribut

Pour chaque valeur d'attribut, il nous faut calculer sa probabilité conditionnelle par rapport aux valeurs cibles.

$$P(C_k|x) = \frac{P(C_k) * P(x|C_k)}{P(x)}$$

Équation 7 : Probabilité conditionnelle

Cela semble peu, mais on a presque tout le travail qui est fait. En pratique, il y aura beaucoup de calculs à faire. Maintenant, nous pouvons classer un nouvel individu en calculant sa probabilité de se trouver dans une classe ou une autre. Ensuite, nous allons normaliser les probabilités et classer dans celle qui a la plus grande valeur.

$$P(C_k|N) = P(C_k) * \sum_{i=1}^n P(\text{val attri} = N \text{ attrib}|Ck)$$

Pour normaliser les probabilités :

$$Pn(C_k) = \frac{P(C_k)}{\sum_{i=1}^n P_{ci}}$$

Équation 8 : Normalisation de probabilité

▪ L'arbre de décision

L'arbre de décision, ou "*decision tree*" en anglais, est aussi une méthode de classification avec un concept qui lui est bien particulier. Comme son nom l'indique, il prend des décisions en se basant sur les attributs des données. D'abord, l'arbre vérifie l'attribut le plus indicatif et prend la direction d'une de ses valeurs, puis le deuxième attribut le plus significatif et prend la direction d'une de ses valeurs, ainsi de suite, jusqu'à classer un nouvel enregistrement.

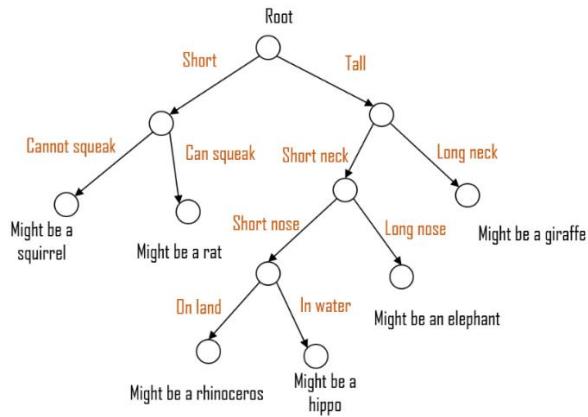


Figure 7 : Structure d'un arbre de décision (Source : Zhao, 2021)

En revanche, le fait de distinguer un attribut significatif ne se fait pas arbitrairement, sinon ce ne serait pas de l'intelligence artificielle. Il y a un certain nombre de calculs (oui, encore des maths) à faire pour trouver la bonne structure de l'arbre et nous allons les voir tout de suite.

❖ Entropie

L'entropie nous renseigne sur la pureté d'un attribut. Si deux classes sont équitablement représentées dans un attribut, on dit que le nœud est impur. En conséquence, l'entropie est maximale (égale ou proche de 1). Si une seule classe est représentée, le nœud est pur et l'entropie est minimale (égale ou proche de 0).

$$Entropy(S) = - \sum_{i=1}^n p_i * \log_2 p_i$$

Équation 9 : Entropie

- ❖ Gain d'information (GI)

La première des choses à faire c'est de calculer le gain d'information c'est-à-dire de tous les attributs, quel est celui qui nous renseigne le plus si l'individu est dans une classe ou l'autre.

$$GI(S, A) = Entropy(S) - \sum_{i=1}^n p(S_A) * Entropy(S_A)$$

Équation 10 : Gain d'information

Voir annexe 3 pour un exemple de calcul d'un arbre de décision.

- **Random forest**

Nous venons juste de parler du *decision tree*, donc nous avons déjà les bases pour comprendre ce qu'est le *random forest*, sans les calculs. Ce terme peut être traduit en français par forêt aléatoire. Pour mieux comprendre, considérons un *dataset*. Pour entraîner le modèle, nous allons utiliser plusieurs modèles de *decision tree*, disons cinq (5) modèles. À partir de là, nous allons diviser notre *dataset* en cinq (5) parties différentes. À ce moment, nous avons 5 datasets et 5 modèles de DT. Ce que nous allons faire, c'est entraîner les 5 datasets avec les 5 modèles de DT.

Nous venons juste de créer de manière simple un *random forest*, "random" (aléatoire) car le *dataset* initial se voit divisé de manière aléatoire. Cependant, ce n'est pas fini, nous avons un modèle certes, mais vous l'aurez remarqué, ce dernier va avoir 5 sorties. Néanmoins, ceci ne constitue pas un souci. Il va être procédé à un vote, et la majorité va l'emporter. Supposons que trois votent pour une classe et deux pour l'autre. Dans ce cas, nous considérons la classe sortie par les trois : la majorité l'emporte. Le *random forest* va pallier aux deux grands problèmes de l'arbre de décision, comment ? En interrogeant plusieurs d'entre eux. Ces deux problèmes sont le biais et *l'overfitting*.

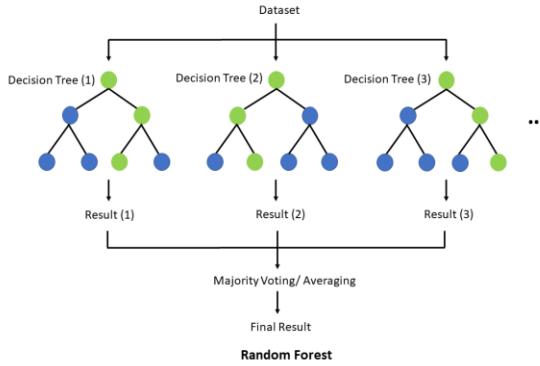


Figure 8 : Une foret aleatoire (Source : Montalvo, 2023)

- **Unsupervised learning**

Pour ce qui est de l'apprentissage non supervisé, ici, nous n'aurons pas d'output pour les inputs. Dans ce cas de figure, nous aurons seulement des données d'entrée, mais on ne sait pas comment réagir en conséquence. C'est le modèle qui va, à lui seul, trouver une représentation générale qui correspond le plus aux données qui lui sont présentées.

Pour ce faire, il y a ce qu'on appelle le clustering : c'est un modèle dans lequel nous allons essayer de regrouper en clusters les individus qui se ressemblent le plus, en utilisant plusieurs variables qui décrivent les données.

- **Le clustering**

Le clustering est une méthode d'apprentissage non supervisé dans laquelle le but est de rassembler les individus qui se ressemblent le plus. Le principe est simple : nous avons des données, mais qui ne sont pas étiquetées, donc c'est au modèle de trouver la représentation la plus fidèle des données. Il existe plusieurs algorithmes de clustering, mais nous allons voir le fameux k-means (le k de k-means représente le nombre de classes ou clusters).

- ❖ Définir le nombre de cluster

En premier lieu, il faut définir le nombre de clusters. Ce choix peut relever du libre arbitre de l'ingénieur ou peut être défini en fonction de méthodes spécifiques.

- ❖ Le centre de gravité

Pour chaque cluster, il faut calculer son centre de gravité et affecter chaque point de la base de données à la classe la plus proche. De là, tous les individus appartiennent à une classe et c'est là que le travail commence.

- ❖ Calcul de distance

Maintenant, nous allons calculer toutes les distances de tous les individus par rapport à tous les centres de gravité de chaque cluster. Nous allons nous apercevoir que certains individus sont mal classés, car ils sont plus proches d'un autre cluster que celui auquel ils appartiennent. Il suffit de les mettre à jour. Cette étape va être répétée autant de fois que nécessaire pour avoir des clusters les plus représentatifs des données possibles.

Voir annexe 4 pour un exemple de calcul de clustering.

- Les règles d'associations

Les règles d'association, ou en anglais "*association rules mining*", sont des méthodes non supervisées qui nous permettent de trouver la corrélation entre une donnée et les autres. Ces règles permettent de répondre à des questions comme : dans quelle mesure B et C vont apparaître sachant que A est apparu ? Ces calculs vont se faire avec un ensemble de sous-ensembles. Nous allons parler ici de itemset pour désigner les sous-ensembles. Les règles d'association sont très fréquentes dans les marchés et supermarchés pour déceler les produits qui sont souvent achetés en même temps par les clients. Une fois que nous avons des règles intéressantes, les dirigeants peuvent prendre de bonnes décisions.

2.2. Deep Learning

Frank Rosenblatt a créé le perceptron qui nous a permis de résoudre des problèmes, notamment le OU et le ET logiques, mais quand ils l'ont essayé pour le XOR, ils se sont rendus compte que le perceptron ne convergeait pas. Le problème était simple : un perceptron traçait des séparateurs linéaires, or ce n'était pas possible pour le problème du XOR.

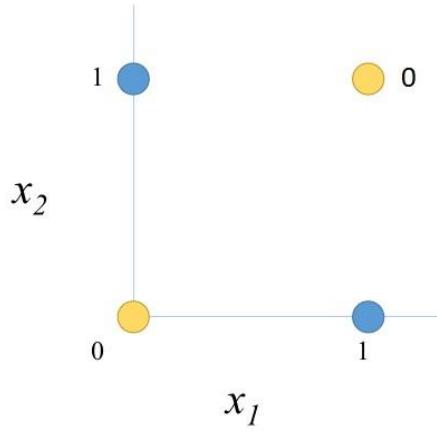


Figure 9 : Representation de la fonction XOR (Source : Oman, 2016)

Allez-y ! Essayez de tracer une seule droite qui soit capable de séparer les 0 et les 1, une droite, ce n'est pas possible. Bienvenue dans le monde du non linéaire, un monde qui fut un casse-tête pour les chercheurs pendant longtemps, jusqu'à ce qu'ils découvrent les solutions qui vont être présentées ici.

- **Artificial neuron network (ANN)**

Si vous vous rappelez la partie portant sur la régression logistique, vous avez déjà quelques notions sur les ANN. Là-bas, nous faisions un apprentissage avec une couche d'entrée et la couche de sortie, mais ici, il sera question d'une couche d'entrée, une ou plusieurs couches cachées et la sortie. Plus il y a de couches cachées, plus c'est profond : apprentissage profond (Deep Learning).

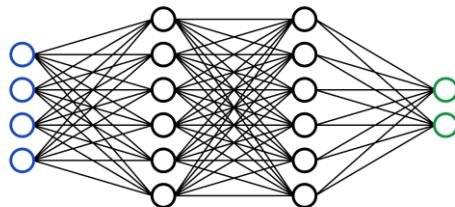


Figure 10 : Structure de réseau de neurones (Source : Ahmad, 2020)

- **Convolutional neuron network (CNN)**

S'il y a un domaine où l'humain a toujours dépassé la machine, c'est la vision : reconnaître des choses, des éléments de la nature et les classer. Mais depuis quelque temps, les scientifiques ont réalisé d'énormes avancées dans le domaine appelé Computer Vision (ou vision par ordinateur).

Et l'un des premiers algorithmes qui a permis de réaliser cela reste le CNN que l'on va voir tout de suite.

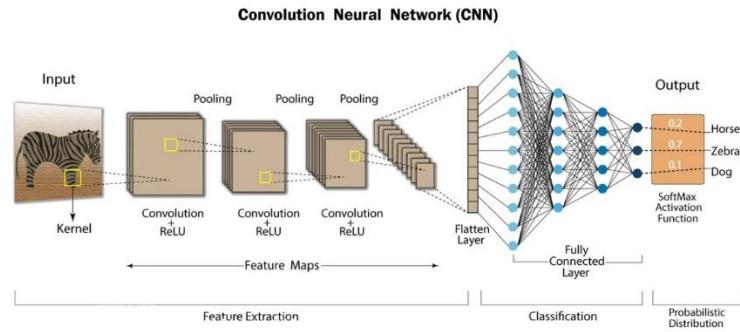


Figure 11 : Reseau d'un CNN (Source : Shahriar, 2023)

- **Recurrent neuron network (RNN)**

Nous venons juste de parler des ANN et de leurs utilités, mais dans tout domaine, il y a toujours des limites. Le principal reproche que l'on peut faire aux ANN, c'est qu'ils n'ont pas de mémoire. Imaginons un jeu de données avec 60 000 inputs. De la première ligne du premier epoch jusqu'à la dernière ligne du dernier epoch, le modèle va oublier tout ce qui s'est passé et se concentrer seulement sur les caractéristiques principales. Mais très souvent, il est nécessaire de savoir ce qui s'est passé pour décider de ce que l'on va prédire.

Exemple : Le Sénégal est un pays qui se trouve en Afrique et dont l'ethnie principale est composée de ...

Nous voulons prédire ce qui va arriver et nous avons trois propositions : ashantis, masaïs, wolofs. Et bien évidemment, c'est les wolofs. Le mot qui nous a permis de décider, c'est "Sénégal", bien sûr. Or, ce mot se trouve au début de la phrase et donc, ce modèle doit avoir une certaine mémoire pour bien prédire.

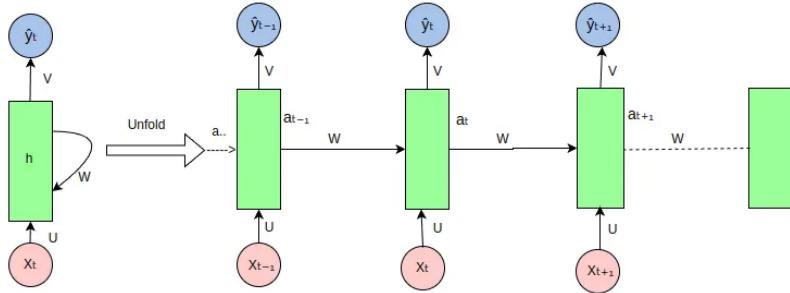


Figure 12 : Réseau d'un RNN (Source : Poudel, 2023)

Voici la structure générale d'un RNN. Il y a, en fait, une seule couche et il représente l'évolution dans le temps. Pour ce qui est de l'erreur et du backpropagation, ce sera la même chose que pour les ANN que nous avons déjà vues.

Chapitre 2 : Revue des travaux de recherche de l'IA appliquée à la finance

Ce chapitre va constituer la revue de la littérature scientifique des travaux de l'IA appliquée à la finance. La finance fait partie des sciences qui font le plus appel à l'application de l'IA, car la finance travaille sur des chiffres et l'IA apprécie les chiffres. Beaucoup d'entreprises de finance utilisent les technologies intelligentes pour faciliter leur travail. En outre, les scientifiques s'illustrent aussi à travers leurs recherches et nous allons ici en voir certains.

Section 1 : Application générale de l'intelligence artificielle sur la finance

Pour ce qui est de cette section, la chose qui sera faite est une présentation des domaines financiers et de gestion que l'IA est en train de révolutionner. Comme déjà dit, la finance fait partie des plus belles applications de l'IA et cette dernière peut aider à parfaire la finance à bien des égards. De l'analyse prédictive à l'analyse des tendances boursières, nous allons voir plusieurs domaines d'application de l'IA en finance.

1. Analyse prédictive

L'analyse prédictive regroupe l'ensemble des méthodes que les financiers peuvent utiliser pour avoir un aperçu de leurs données futures en corrélation avec leurs données historiques grâce à l'aide de l'IA. Ces analyses peuvent prendre plusieurs formes et s'appliquent dans plusieurs domaines de la finance.

Dans ce passage, nous allons avoir une vue d'ensemble des analyses prédictives à base d'IA afin de traiter chaque cas de manière individuelle. C'est dans cette optique que le scientifique Daniel Broby, en 2022, a publié un article où il fait l'inventaire des algorithmes pour les différents domaines de la finance. L'objet de cet article est de présenter une revue des méthodes basées sur la littérature scientifique en se focalisant sur les domaines d'application de l'analyse prédictive. Pour y parvenir, il a fait une étude comparative de différents modèles tels que la classification, la régression, le clustering, les règles d'association, et les modèles de time series. Les résultats étaient les suivants : pour les prédictions sur l'économie, les time series sont les plus performants ; pour les prédictions sur des gains potentiels, plusieurs modèles offrent de bonnes performances, notamment le Naive Bayes, les ANN, les modèles non linéaires et enfin, pour l'optimisation de portfolio, ce sont les modèles de ML. Cependant, son travail ne s'arrête pas là, il a utilisé d'autres algorithmes sur d'autres domaines.

Nous avons constaté que dans cet article, il y a plusieurs modèles qui peuvent être intéressants pour une analyse prédictive. Néanmoins, le plus important est d'être à l'écoute des données dont nous disposons et de choisir le modèle le plus adapté.

2. Gestion des risques

De manière simple, le risque peut être considéré comme une probabilité de perte. Dans le domaine financier, il est d'autant plus pertinent de le considérer et de réagir en conséquence, c'est ce qu'on appelle la gestion des risques. En finance, nous avons plusieurs types de risques : le risque de défaut, qui se produit lorsqu'un emprunteur ne respecte pas ses obligations de paiement. Il existe aussi des risques liés au marché, dans le cas d'achat d'actions, d'obligations ou d'autres produits financiers, où il y a toujours une part d'incertitude dans ce genre de transactions. Il y a également le risque inhérent, qui n'est pas spécifique mais lié à l'exécution de l'activité en tant que telle : il y a un risque dès qu'il y a une activité.

De là, les modèles de ML, de DL, de NLP, voire même de BI, vont nous permettre soit de pallier, contrecarrer ou même prévenir ces risques. Les modèles de ML peuvent nous aider à trouver les relations qui existent entre les données des clients et la potentialité de comportements licites ou douteux à travers les calculs mathématiques. Le NLP peut nous permettre de comprendre et de traiter les données de type financier, comme les rapports d'activité, les contrats financiers, etc. Ce domaine nous intéressera particulièrement ici, car nous allons développer un modèle NLP dans ce document.

Nous pouvons poursuivre en donnant des exemples d'application de l'IA dans la finance : ALLADIN (BlackRock), Bloomberg AIM (Bloomberg), Marcus AI (Goldman Sachs), COiN (J.P. Morgan).

3. Services clients

Est-ce que les IA vont remplacer les humains dans les services des entreprises ? Voilà l'une des questions qui se posent le plus dans le domaine. Les services clients ou services après-vente sont très coûteux pour les organisations et demandent des ressources humaines et financières. Or, l'IA peut aider à améliorer cela d'une manière plus qu'efficace, et nous allons voir comment.

Dans un article de 2022, Mengmeng Song, Xinyu Xing, Yucong Duan, Jason Cohen et Jian Mou ont essayé d'apporter des réponses à ces questionnements. Ils nous parlent dans leur article des

impacts que l'IA pourrait avoir dans les services clients. Ils ont fait différentes expérimentations, des tests, des comparaisons et des hypothèses pour aboutir aux résultats. Et ces résultats étaient que les technologies d'IA allaient sans doute prendre une grande place dans les services clients, mais il ne faut pas complètement supprimer l'interaction humaine.

Dans le même cadre, d'autres études ont été faites sur le domaine, comme celle de 2022 de Dimitrios Buhalis et Iuliia Moldavskaya dans le *Journal of Hospitality and Tourism Technology*. Cet article se donnait pour objectif d'investiguer les interactions entre les hôtels et les hôtes dans le contexte de « l'hospitalité ». La méthodologie que cette étude a suivie a consisté à faire des entretiens avec le personnel des hôtels sur l'utilisation des assistantes vocales dans ces structures. Ce qu'ils ont trouvé, c'est que l'utilisation de ces technologies est d'un grand apport autant pour le personnel que pour les clients. Selon les auteurs, cela peut leur permettre d'explorer les domaines du "smart hospitality" et de l'écosystème du tourisme avec l'IA.

4. Détection de fraudes

La détection de fraude est un procédé qui nous permet d'identifier des activités frauduleuses exécutées ou tentées au sein d'une organisation. Les fraudes, si elles ne sont pas identifiées et réglées, peuvent être un cauchemar pour les organisations. C'est là que l'IA peut intervenir avec ses algorithmes.

Pour mettre en évidence l'aide non négligeable que l'IA peut apporter dans la détection de fraude, Muhammad Farman et Muzamil Abbas ont très récemment publié un article sur le sujet intitulé "*Artificial Intelligence for Fraud Detection and Prevention*". L'objet de cette étude était de révéler l'apport de l'IA sur la détection de fraude, mais aussi sur la prévention. Pour ce faire, ils ont d'abord présenté les algorithmes d'IA et leur fonctionnement spécifique dans la détection de fraude, puis ils ont abordé l'IA dans la finance. Les résultats de cette étude ont été conséquents, car ils ont pu prouver l'apport considérable de l'IA dans ce domaine sans rien implémenter.

Un peu plus tôt, en 2018, deux chercheurs (Dahee Choi, Kyungho Lee) avaient fait une étude similaire mais plus pratique sur la détection de fraude. Ici, l'étude se voulait non seulement théorique mais aussi pratique en implantant des modèles qui allaient fonctionner pour la détection. Pour ce qui est des modèles, ils ont utilisé dans un premier temps des algorithmes de Machine Learning, puis de Deep Learning. Les observations faites à la suite des expérimentations

sont claires : les modèles de ML ont été plus performants que les modèles de DL, comme quoi parfois le simple est plus efficace.

5. La bourse et les marchés financiers

La bourse et les marchés financiers n'échappent pas à l'IA, c'est l'une des applications les plus challengeantes de l'IA dans la finance, du fait de la volatilité des marchés financiers et de l'incertitude qui y règne. Mais une telle chose n'a pas empêché les scientifiques de s'y aventurer.

C'est en l'occurrence le cas de Mahinda Mailagaha Kumbure, Christoph Lohrmann, Pasi Luukka et Jari Porras qui ont collaboré à la rédaction d'un article qui se voulait être une revue des techniques de ML dans la prédiction du cours boursier. Cet article avait pour objectif principal de faire une investigation sur la littérature, mais aussi d'avoir connaissance des types de variables utilisées pour cette tâche. Une fois ceci fait, il a été procédé à l'application de ces variables sur le ML afin de prédire les cours boursiers des marchés financiers. Les données qu'ils ont exploitées proviennent d'articles de presse de 2000 à 2019. Ils ont adapté la méthodologie suivante : d'abord la collecte de données boursières sur les marchés financiers du monde entier avec les *time series*, ensuite le traitement, enfin l'étude comparative des modèles. Le résultat le plus frappant était que les ANN ont été très performants par rapport aux autres, surtout sur les marchés comme le S&P 500 (86,81 %) et le DJIA (88,98 %).

Il y a beaucoup d'autres recherches dans le domaine, elles sont toutes aussi intéressantes les unes que les autres. Cependant, faisons un tour chez les professionnels, ils s'illustrent mieux car disposant de plus de fonds, mais aussi il y a une meilleure praticabilité de leur produit. Parlant des produits provenant des professionnels, nous pouvons notamment parler de *EquBot* qui est un Chatbot qui prodigue des informations boursières sur les marchés financiers, il y a aussi *Trade Ideas* qui, comme son nom l'indique, est utilisé pour des conseils en investissement sur les marchés financiers, *TrendSpider* quant à lui est utilisé pour faire la prédiction des prix de titres financiers et bien évidemment il y a d'autres applications. Néanmoins, toutes ces applications ne sont pas des produits miracles, elles viennent avec leur lot de qualités et de défauts.

Section 2 : L'intelligence artificielle dans l'analyse des états financiers

Ici, nous allons entrer dans le vif du sujet. Il y a un certain nombre de travaux scientifiques sur l'analyse des états financiers. Il existe trois principaux états financiers que nous allons voir dans

la partie 2 (bilan, compte de résultat et tableau des flux de trésorerie). La principale difficulté sera la modélisation des états financiers en un format compréhensible par les modèles que nous avons déjà vus. Ensuite, va s'opérer la magie de l'IA et nous verrons comment les scientifiques procèdent.

1. Les travaux de l'intelligence artificielle sur l'analyse financière

Dans notre travail, nous nous sommes donnés pour objectif d'appliquer l'IA dans l'analyse financière des entreprises ouest-africaines. Nous allons voir que d'autres travaux ont été réalisés ailleurs et peuvent ressembler à ce que nous allons faire ici, nous allons les examiner.

Ce sujet est tellement d'actualité que des articles très intéressants sont en train d'être rédigés et publiés au moment où j'écris ces lignes. C'est le cas de cet article de Ewerton Alex Avelar et Ricardo Vinícius Dias Jordão, publié en juillet 2024. Cet article a pour objectif d'analyser la performance de différents algorithmes d'IA dans la prédiction des mouvements des plus grands marchés financiers du monde. L'approche qu'ils ont adoptée est de tester différents algorithmes sur des données empiriques, et pour cela, ils vont utiliser neuf (9) indicateurs. Le résultat qu'ils ont trouvé est que les modèles d'IA sont plus performants que les techniques utilisées par les analystes, et que parmi tous les algorithmes, le Random Forest est le plus efficace. Les enseignements que l'on peut tirer de cette étude, selon les auteurs, sont que l'IA joue un rôle réel dans l'analyse financière et la prise de décision, et que les managers doivent prendre cela en considération.

D'autres recherches ont aussi été menées sur le domaine, comme cette étude de 2000 de Ning Yang sur l'IA et le Data Mining dans le Financial Big Data. L'objet de cette recherche était de développer des modèles basés sur le Data Mining en entrant dans les détails du fonctionnement des algorithmes sur des données financières de très grande taille. Il a suivi la démarche suivante : d'abord expliquer le fonctionnement des modèles, puis les appliquer sur des données empiriques. Il a réussi à obtenir des résultats satisfaisants, puisque l'application parvenait à répondre à des tâches importantes dans la finance. Avec le développement rapide des données financières, l'application de ces modèles va devenir plus que nécessaire.

Comme pour les autres domaines, les chercheurs font des avancées, mais les ingénieurs et les professionnels aussi. De ce fait, il y a beaucoup de modèles sur l'analyse financière qui ont déjà

été développés et mis en ligne pour une utilisation générale du grand public. Il peut être donné l'exemple de [*Vena Insights*](#) qui s'illustre dans la budgétisation, la prédition et le planning, etc. Il y a aussi [*Domo*](#), qui est une application à base d'IA pour intégrer des données en temps réel et gérer l'ensemble. Nous pouvons aussi parler de [*Datarails FP&A Genius*](#), qui est un Chatbot pour la gestion financière en intégrant la possibilité de se connecter à une source de donnée pour recevoir des informations en temps réel.

2. *Limites des travaux actuels*

Dans tous les travaux de recherche, il y a toujours des limites, des défauts ou des points d'amélioration. C'est d'ailleurs pourquoi les chercheurs continuent toujours de faire des recherches. Dans le domaine de l'IA appliquée à la finance, il y a un certain nombre d'endroits où il y a lieu d'amélioration pour parfaire la science, et nous allons les examiner.

- **Qualités des données financières**

Quand on parle de qualité des données, on peut faire référence à plusieurs aspects des données financières. Pour ces types de données, nous faisons souvent face à des bruits et à des valeurs incomplètes. La conséquence de cela, c'est qu'il faut faire un grand travail de *preprocessing*. En outre, les données historiques peuvent être biaisées : la finance change, les méthodes aussi, et les politiques financières évoluent également. Cela peut constituer un problème lors de l'entraînement des modèles.

Pour illustrer cela, en 2018, dans la zone UEMOA, une nouvelle réglementation a vu le jour, imposant à toutes les entreprises de la BRVM de présenter leurs états financiers selon les normes IFRS (*International Financial Reporting Standards*). Mais cela concerne surtout les grandes entreprises. Pour les petites entreprises, la situation est encore plus difficile, car il est très compliqué d'accéder à leurs données. C'est l'une des raisons pour lesquelles les petites et moyennes entreprises ne bénéficient pas encore de ces technologies.

Enfin, nous pouvons parler du contexte : les données peuvent changer en fonction du contexte géopolitique, économique, social, etc., ce qui peut biaiser les modèles en prenant en compte des périodes avec des mesures exceptionnelles (comme la crise des subprimes en 2008 ou la pandémie de COVID-19 en 2020, par exemple).

- **L'éthique**

Pour ce qui est de l'éthique des IA dans le contexte financier, elle peut être abordée sous plusieurs angles. Nous allons parler des manières de développer les modèles. Le comportement des modèles sera toujours basé sur les données d'entraînement, ce qui veut dire que les développeurs peuvent intentionnellement programmer l'IA pour qu'elle se comporte d'une manière ou d'une autre. De plus, l'IA ne fera que ce pour quoi elle a été développée ; à moins d'une mise à jour, elle ne pourra pas contextualiser les situations pour prendre la meilleure décision.

Il y a aussi un autre problème qui subsiste : une IA qui a été programmée d'une certaine manière se comportera et fonctionnera de cette même manière. Pour être plus clair, il y a un risque que l'IA donne toujours le même conseil. Si elle le fait et que tout le monde (ou la majorité des gens) suit ses conseils, cela pourrait mener à ce que les gens prennent les mêmes décisions, ce qui n'est pas souhaitable. Imaginons juste un marché avec seulement des vendeurs mais pas d'acheteurs.

Nous pouvons aussi nous questionner sur la confidentialité et la sécurité des données, ainsi que sur les délits d'initié. L'IA pourra-t-elle faire preuve de discernement et savoir quelle information donner à qui ?

On peut répondre à toutes ces interrogations en disant que l'IA est un programme, et qu'il est possible de programmer ces aspects dans son développement (même si cela ne sera pas facile). Par exemple, certains LLM ne répondent pas à certains sujets sensibles comme la religion, la race, les ethnies, etc.

- **L'insuffisance des études dans le contexte africain**

Personnellement, la limite qui nous interpelle le plus est l'absence de recherches qui se focalisent sur l'Afrique. Et même pour être plus précis, dans la zone UEMOA, il n'y a pas assez d'études approfondies sur le domaine de la finance des entreprises ouest-africaines. Quand nous savons comment certaines entreprises africaines fonctionnent avec toutes leurs difficultés, l'IA leur sera d'une très grande utilité.

C'est d'ailleurs l'une des grandes raisons pour lesquelles nous nous sommes lancés dans ce domaine, en plus de notre grand intérêt pour l'IA. Il n'est pas seulement intéressant de constater

les problèmes et les manquements, il faut aussi proposer des solutions avec les moyens dont nous disposons.

Conclusion de partie

En guise de conclusion pour cette partie, nous avons vu deux chapitres qui nous révèlent des éléments importants de l'intelligence artificielle : les généralités et théories, mais aussi les travaux de recherche scientifiques. Nous pouvons tous convenir que, sans ces connaissances acquises durant cette partie, il serait difficile de s'aventurer dans le développement pratique des modèles, que ce soit pour la prédition ou le Chatbot. Comme constaté dans cette partie, beaucoup de choses ont déjà été faites et nous n'allons pas réinventer la roue, mais nous allons la façonne à notre guise pour qu'elle soit en mesure de rouler de la manière qui nous convient.

En ce qui concerne cette suite, nous allons, comme les anglophones diraient, "*get our hands dirty with some code*", ce qui signifie se salir les mains avec du code. La partie suivante sera plus technique, car nous passerons à la pratique après toute cette théorie. Et comme nous l'avons dit ici, nous allons entrer dans les détails pour que la première étape ne soit pas vaine.

Partie 2 : Conception et développement des outils d'IA appliquée à l'analyse financières

Introduction de partie

Dans la partie précédente, nous avons fait connaissance avec l'IA, nous avons vu les généralités, les théories et les travaux scientifiques. Après cela, tous les outils nécessaires sont dans notre trousse afin d'implémenter des modèles pratiques d'IA. L'objectif de cette partie est simple : utiliser toutes les connaissances acquises durant la partie précédente pour développer les modèles qui permettront d'augmenter l'efficacité des analystes dans leur travail.

Cette partie sera également subdivisée en plusieurs chapitres : le premier sera consacré au développement des modèles prédictifs. Ici, nous ferons d'abord une présentation de la démarche d'analyse financière, puis le développement des modèles prédictifs. Après cela, nous aborderons le chapitre sur le développement du Chatbot, où la collecte de données et le développement en tant que tel, seront les deux sections qui nous intéresseront.

Chapitre 3 : Analyse et développement de modèles prédictifs

Dans ce chapitre, nous allons commencer le développement pratique de l'interface graphique et des modèles de prédiction. Néanmoins, il est évident qu'il faut être en mesure de faire une analyse financière avant de développer une application ou des modèles. Nous allons aussi présenter les outils utilisés pour le développement de l'interface et expliquer comment cette dernière sera réalisée.

Section 1 : Mise en œuvre d'une application d'analyse financière

Nous voilà dans la section du document où nous allons parler du développement des modèles. Elle se divise en deux blocs : l'analyse financière traditionnelle (où nous allons voir comment les analystes financiers font habituellement leur travail) et l'amélioration grâce à l'intégration des outils intelligents (une interface graphique avec la possibilité de faire des prédictions).

1. Démarche d'une analyse financière

D'abord, nous allons procéder à une présentation des outils que l'analyse financiers va utiliser pour faire convenablement son travail, il s'agit du bilan, du compte de résultat et du tableau des flux de trésorerie avant de passer à l'analyse proprement dite.

• Le bilan

Il est souvent entendu que le bilan est la photographie d'une entreprise à un instant T. Ce document comptable révèle les actifs et les passifs d'une entreprise.

BILAN		
	N	N-1
ACTIFS		
<i>Charges immobilisées</i>	-	-
<i>Immobilisations incorporelles</i>	-	-
<i>Immobilisations corporelles brutes</i>	-	-
<i>Immobilisations corporelles brutes</i>	-	-
<i>Immobilisations corporelles brutes</i>	-	-
<i>Immobilisations corporelles brutes</i>	-	-
<i>Amortissements et provisions</i>	-	-
TOTAL ACTIF IMMOBILISE	-	-
<i>Stock</i>	-	-
<i>Fournisseurs, avances versées</i>	-	-
<i>Clients</i>	-	-

	<i>Autres créances</i>	-	-
	TOTAL ACTIF CIRCULANT	-	-
	TOTAL TRESORERIE ACTIF	-	-
	TOTAL ACTIF	-	-
PASSIF			
	<i>Capital</i>	-	-
	<i>Primes et réserves</i>	-	-
	<i>Report à nouveau</i>	-	-
	<i>Résultat net</i>	-	-
	TOTAL CAPITAUX PROPRES	-	-
	<i>Emprunts et dettes financières</i>	-	-
	<i>Provisions financières</i>	-	-
	TOTAL DETTES FINANCIERES	-	-
	<i>Dettes circulants</i>	-	-
	<i>Clients, avances reçues</i>	-	-
	<i>Fournisseurs d'exploitation</i>	-	-
	<i>Dettes fiscales</i>	-	-
	<i>Dettes sociales</i>	-	-
	<i>Autres dettes</i>	-	-
	TOTAL PASSIF CIRCULANT	-	-
	TOTAL TRESORERIE PASSIF	-	-
	TOTAL PASSIF	-	-

Tableau 1 : Bilan

- **Le compte de résultat**

Le compte de résultat est un document, comme le bilan, qui permet de juger de la santé des finances d'une entreprise. Cet état financier retrace les charges d'une entreprise et les éléments qui les composent, les produits et les éléments qui les composent, et éventuellement les soldes intermédiaires de gestion.

COMPTE DE RESULTAT

		N	N-1
	<i>Ventes de marchandises</i>	-	-
	<i>Production vendue</i>	-	-
	<i>Travaux, services vendus</i>	-	-
	<i>Produits accessoires</i>	-	-
	CHIFFRE D'AFFAIRES	-	-
	<i>Production stockée</i>	-	-
	<i>Autres produits</i>	-	-

TOTAL PRODUITS	-	-
<i>Achats et frais sur achats</i>	-	-
<i>Variation de stock</i>	-	-
<i>Transports</i>	-	-
<i>Services extérieurs</i>	-	-
<i>Impôts et taxes</i>	-	-
<i>Charges et pertes diverses</i>	-	-
TOTAL CHARGES	-	-
VALEUR AJOUTEE	-	-
<i>Frais de personnel</i>	-	-
EXCEDENT BRUTE D'EXPLOITATION	-	-
<i>Transfert de charges</i>	-	-
<i>Dotations aux amortissements</i>	-	-
<i>Dotations aux provisions</i>	-	-
RESULTAT D'EXPLOITATION	-	-
<i>Produits financiers</i>	-	-
<i>Charges financières</i>	-	-
RESULTAT FINANCIER	-	-
<i>Produits HAO</i>	-	-
<i>Charges HAO</i>	-	-
RESULTAT EXCEPTIONNEL	-	-
<i>Impôt sur la société</i>	-	-
RESULTAT NET	-	-

Tableau 2 : Compte de résultat

- **Le tableau des flux de trésorerie**

Selon le site [L'expert-comptable](#), le tableau de flux de trésorerie est un outil financier qui permet de déterminer la rentabilité d'un projet, d'évaluer le besoin en fonds de roulement et d'anticiper ses besoins en fonds propres. Il indique les entrées et sorties d'argent de l'entreprise sur une période.

TABLEAU DES FLUX DE TRESORERIE

	N	N-1
TRESORERIE INITIALE	-	-
<i>Capacité d'autofinancement</i>	-	-
<i>Variation actif circulant HAO</i>	-	-
<i>Variation des stocks</i>	-	-
<i>Variation des créances</i>	-	-
<i>Variation du passif circulant</i>	-	-
<i>Variation du BFR</i>	-	-

FLUX DE TRESORERIE DES ACTIVITES OPERATIONNELLES	-	-
Décaissements liés aux acquisitions d'immobilisations incorporelles	-	-
Décaissements liés aux acquisitions d'immobilisations corporelles	-	-
Décaissements liés aux acquisitions d'immobilisations financières	-	-
Encaissements liés aux acquisitions d'immobilisations incorporelles et corporelles	-	-
Encaissements liés aux acquisitions d'immobilisations financières	-	-
FLUX DE TRESORERIE DES ACTIVITES D'INVESTISSEMENT	-	-
Augmentation du capital par apports nouveaux	-	-
Subvention d'exploitation	-	-
Prélèvement sur le capital	-	-
Dividendes versés	-	-
Flux de trésorerie provenant des capitaux propres	-	-
Emprunts	-	-
Autres dettes financières	-	-
Remboursement des emprunts et autres dettes financiers	-	-
Flux de trésorerie provenant des capitaux étrangers	-	-
FLUX DE TRESORERIE DES ACTIVITES DE FINANCEMENT	-	-
Variation de trésorerie nette	-	-
TRESORERIE FINALE	-	-

Tableau 3 : Tableau des flux de trésorerie

- **Les étapes d'une analyse financière**

Une fois en possession de ses trois états financiers, nous sommes fin prêt pour faire une analyse financière, il faut prendre conscience qu'il y a plusieurs moyens d'en faire une. Pour ce document et pour l'application que nous allons développer nous allons suivre une procédure en quatre (4) étapes.

- **Vérification des états financiers**

Avant de pouvoir commencer l'analyse financière, il faut d'abord effectuer certaines vérifications pour éviter que l'analyse soit biaisée par certaines valeurs.

L'actif et le passif : il faut vérifier que le total des actifs est égal au total des passifs.

Le résultat net : il faut vérifier que le résultat du compte de résultat est égal au résultat du bilan.

La trésorerie nette : il faut également vérifier que la différence entre la trésorerie-actif et la trésorerie-passif est bien égale à la trésorerie nette au 31 décembre de l'année N.

Une fois ces vérifications faites et que tout est OK, on peut procéder à l'analyse financière proprement dite. Il faut ajouter qu'il existe d'autres types de vérifications que l'on peut effectuer, mais celles-ci sont les plus importantes.

- **Analyse des états financiers**

Dans cette partie, nous allons faire l'analyse financière des états financiers, à savoir le compte de résultat, le bilan et le tableau des flux de trésorerie. Pour chaque état financier, il y aura deux types d'analyse : une analyse verticale et une analyse horizontale.

Analyse verticale

L'analyse verticale consiste à rapporter tous les éléments d'un état financier à une valeur pivot. C'est-à-dire la part de chaque élément d'un état financier par rapport à la valeur choisie. Qu'allons-nous faire exactement ? Nous allons tout simplement diviser chaque rubrique de l'état financier en question par la valeur constante et nous obtiendrons ainsi notre tableau pour l'analyse verticale.

Analyse du compte de résultat						
Analyse verticale						
Ventes de marchandises	0.03%	0.04%	0.04%	0.05%	0.06%	0.09%
Achats de marchandises	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Variation de stocks	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Marge brute sur marchandises	0.03%	0.04%	0.04%	0.05%	0.06%	0.09%
Marge brute sur matières	82.13%	82.49%	81.06%	0.00%	0.00%	0.00%
Marge commerciale	0.03%	0.04%	0.04%	0.05%	0.06%	0.09%
Ventes de produits fabriqués	54.71%	56.31%	61.39%	60.26%	58.00%	64.50%
Travaux, services vendus	43.44%	41.73%	36.31%	37.55%	39.84%	32.70%
Produits accessoires	1.82%	1.92%	2.26%	2.15%	2.10%	2.71%
Chiffre d'affaires	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Production stockée (ou déstockage)	-0.11%	0.06%	0.15%	1.10%	0.79%	-0.88%
Production immobilisée	3.03%	2.39%	1.95%	2.17%	3.88%	5.57%
Subventions d'exploitation	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Autres produits	3.10%	3.03%	1.93%	0.46%	0.58%	0.50%
Transferts de charges d'exploitation	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Achats des matières premières et fournitures	15.57%	20.69%	18.75%	20.26%	23.79%	22.95%

Figure 13 : Analyse verticale du compte de résultat

Analyse horizontale

Pour ce qui est de l'analyse horizontale du compte de résultat, du bilan et du tableau des flux de trésorerie, nous allons tout simplement calculer la croissance de chaque élément de l'état financier par rapport à l'année précédente.

$$Croissance(X) = \frac{X_n - X_{n-1}}{X_{n-1}}$$

Équation 11: Calcul de la croissance

○ Analyse de l'activité et des relations de trésorerie

Ce point de l'analyse financière est la partie qui va nous permettre d'apprécier l'évolution des chiffres clés d'une entreprise. Ce sont toutes les valeurs que l'entreprise doit maximiser pour rester en bonne santé et durer dans le temps. Il y aura trois (3) analyses qui seront faites dans cette partie :

Analyse du cycle de vie de l'activité

Analyse du comportement des flux de trésorerie

Analyse des équilibres financiers et de la relation de trésorerie

Flux de trésorerie provenant des activités opérationnelles	3,207,708,967	(1,933,016,664)	(14,699,842,866)	(4,064,191,206)	10,480,831,071	6,291,120,26
Flux de trésorerie provenant des activités d'investissement	6,321,327,641	9,870,347,776	9,692,145,003	7,945,553,506	9,772,611,323	8,630,238,76
Flux de trésorerie provenant des activités de financement	(1,114,129,941)	9,851,203,738	(3,354,482,017)	(3,550,829,906)	(67,659,787)	(2,102,156,78)
Variation de la trésorerie nette de l'exercice	(4,203,923,485)	(1,433,184,396)	(27,739,146,063)	(16,194,039,240)	640,559,961	(4,441,275,24)
Trésorerie nette au 31 décembre	2,587,596,703	1,154,412,307	(26,584,733,756)	(42,778,772,996)	(42,138,213,035)	(46,579,488,24)

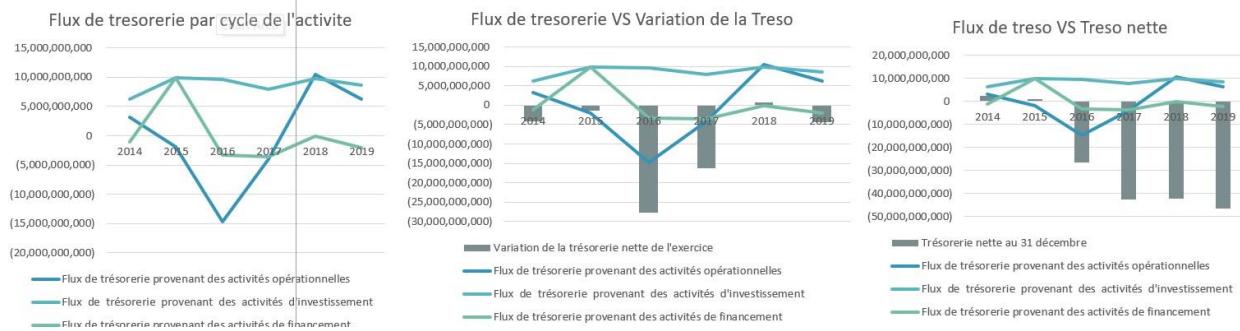


Figure 14 : Analyse du comportement des flux de trésorerie

○ Analyse tendancielle par la méthode des ratios

Tout d'abord, l'analyse financière avec la méthode des ratios nous permet de poser un diagnostic sur l'entreprise à travers un certain nombre de ratios. Ces ratios sont de tout type et font intervenir les trois (3) états financiers. Ici, nous sommes à un point très important de l'analyse financière,

c'est pourquoi cette partie va particulièrement intéresser les investisseurs, en plus de l'entreprise elle-même.

Ces calculs de ratios vont nous permettre de noter l'entreprise à la fin de l'analyse, afin de déterminer si elle court un risque de défaut ou de faillite. Nous allons utiliser un système de notation développé par Altman en 2005. Finalement, l'évaluation de l'entreprise viendra clore cette analyse par les ratios. Ci-après un tableau récapitulatif des familles et de certains de leurs ratios.

Famille	Ratios
Profitabilité	Taux de marge nette, taux de valeur ajoutée, taux bruts d'exploitation ...
Rentabilité	Rentabilité économique et rentabilité financière.
Politique comptable	Ratio de vétuste, taux de provision de stocks, taux de provision de créances.
Liquidité	Liquidité (générale, réduite et immédiate)
Gestion de la dette	Levier, taux d'endettement, maturité de l'endettement
Flux de trésorerie	Capacité à investir, taux d'investissement net, taux de réinvestissement ...
Efficacité des actifs du BFG	Variation CA, variation VA, taux de valeur ajoutée
Valorisation	Capitalisation boursière, valorisation boursière, capitalisation du CA ...
Probabilité de défaut	Zscore d'Altman, score de la banque de France.

Tableau 4 : Les familles de ratios

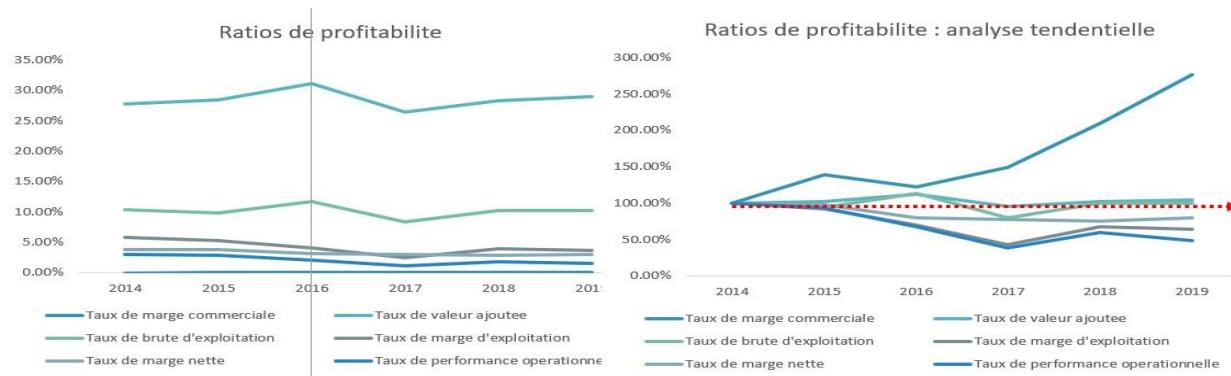


Figure 15 : Représentation graphique des ratios de profitabilité

2. Le frontend

Pour ce qui est du développement de l'interface graphique, nous allons utiliser Qt. Ce dernier nous offre un certain nombre de fonctionnalités pour réaliser des logiciels robustes et très avancés, mais

nous allons en sélectionner quelques-unes qui nous ont été utiles dans notre travail, afin de les expliquer et de montrer comment nous les avons utilisées dans nos modèles.

- Widget

Les widgets font partie des éléments de base de Qt. En effet, tout est un widget, et il existe des widgets pour tout faire. Pour mieux comprendre ce concept, un widget est un élément visible avec lequel l'utilisateur peut interagir. Nous pouvons donner l'exemple d'un bouton, d'une zone de texte, d'un label, voire même d'une page entière.

L'un des widgets les plus importants est ce que Qt appelle *QMainWindow*, qui est le widget principal sur lequel tous les autres vont s'accorder. Il faut préciser que chaque widget a un parent, sauf le widget principal, ce qui donne une structure sous forme d'arbre.

Dans notre logiciel, nous allons utiliser toutes sortes de widgets, en commençant bien évidemment par le *QMainWindow*, puis en passant par les autres qui permettent de présenter les résultats de nos modèles de manière simple.

- Layout

Le deuxième élément sur lequel nous allons un peu nous attarder ce sont les layouts. Ils représentent toutes les manières de disposer nos widgets. Par défaut, un widget va empiler tous ses enfants à son point (0, 0) en haut à gauche dans une fenêtre d'application.

Avec l'aide des layouts, nous pouvons disposer les widgets de différentes manières, ce qui permet d'obtenir un logiciel plus soigné. Il existe trois types de layouts sur Qt : QVBoxLayout (disposition verticale), QHBoxLayout (disposition horizontale) et QGridLayout (disposition en grille). Ces trois héritent tous de la classe mère QLayout.

Bien que nous ayons principalement utilisé les dispositions verticale et horizontale dans notre logiciel, la disposition en grille peut être très pertinente dans certains cas. Il est impressionnant de constater que tous les interfaces utilisateurs que nous voyons dans les logiciels sont construites à partir de ces trois layouts.

- Signal and Slot

Les *signals* et slots sont les deux concepts clés de la programmation événementielle que Qt met à notre disposition. Un signal est un déclencheur, cela peut être un clic, un focus, un survol ou tout changement quelconque. Il existe des *signals* prédéfinis, comme ceux que nous venons de citer, mais nous avons également la possibilité d'en créer de nouveaux. Cependant, dans la plupart des cas, les *signals* prédéfinis suffisent.

Les *slots* sont les fonctions que nous utilisons pour réagir aux *signals*. Mais pourquoi les *slots* ne sont pas des méthodes comme les autres ? Parce que les *signals* n'acceptent que les slots comme fonctions ; une méthode qui ne porte pas la mention slot ne sera pas acceptée dans la programmation événementielle de Qt.

Nous allons illustrer cela avec l'exemple de l'utilisateur qui veut ouvrir les paramètres. S'il clique sur le bouton des paramètres (*signal*), l'application va ouvrir la boîte de dialogue des paramètres (*slot*). Ce qui est intéressant avec Qt, c'est qu'il nous permet de faire cela en une seule ligne de code grâce à la fonction *connect* de la classe *QObject*.

3. Le Backend

Le backend d'une application c'est tout ce que l'utilisateur ne voit pas, et qui fait toute la puissance d'une application. Tout ce qui est base de données, classes, appelle de fonctions sont classer dans le backend. Pour ce qui s'agit de nos applications nous aurons une base de données MySQL, des classes qui vont interagir avec la base de données.

- La base de données

Nous allons mettre en place une base de données locale pour stocker les valeurs des états financiers. Cette base contiendra trois tables : Bilan, Compte de résultat, Tableau des flux de trésorerie.

Chaque enregistrement dans ces tables représentera une année financière. Il s'agit d'une base de données relativement simple, où nous n'effectuerons pas fréquemment des insertions ou des modifications. La plupart des opérations consisteront en des requêtes SELECT pour récupérer les données.

- Le contrôleur

Bien que nous ne suivions pas strictement le modèle MVC (Modèle, Vue, Contrôleur), nous n'allons pas permettre un accès direct aux requêtes de la base de données. Cette approche permet une meilleure gestion des données et plus de flexibilité dans l'architecture. Nous utiliserons des classes pour gérer la logique métier des états financiers.

Il y aura donc trois classes, une pour chaque état financier. Chaque classe comprendra :

Des méthodes pour récupérer les états financiers.

Des méthodes pour effectuer les analyses verticales.

Des *getters* et *setters* pour chaque élément des états financiers.

- Serveur

Pour interagir avec la base de données, nous avons besoin d'un **SGBD** (Système de Gestion de Base de Données). Nous avons choisi **MySQL** comme SGBD, et pour faciliter la gestion de la base de données localement, nous utiliserons **XAMPP**. XAMPP est un serveur qui permet de configurer facilement une base de données **MySQL** ainsi qu'Apache et d'autres services nécessaires pour le développement local.

4. Le Web server

Un web server est un programme informatique qui nous permet une encapsulation de données avec les protocoles http ou https. Avec un web server, on peut déployer une fonction qui peut être invoquée depuis plusieurs *endpoints*.

De là, on peut voir comment cela peut nous être utile, les modèles ont été développés en langage Python et l'interface graphique avec C++, il nous faut les lier. C'est là qu'intervient la puissance des web servers. Il est possible d'écrire une *Application Programming Interface* (API) dans un langage A et de l'invoquer dans un langage B avec l'aide d'un navigateur.

Il y a deux types de web servers : étendu avec SOAP et REST, nous allons utiliser REST avec la bibliothèque de Python *FastAPI*.

Nous pouvons donner l'exemple d'un cas d'utilisateur du Chatbot. L'utilisateur pose sa question depuis l'application Qt, une requête REST va être invoquée avec la question comme paramètre, le web server va la prendre, la traiter et renvoyer la réponse sous format JSON ou XML.

Cette méthode est aussi appelée le développement multi-tiers. Il y a deux programmes différents (deux tiers), mais ils peuvent communiquer tout en étant indépendants l'un de l'autre.

5. Présentation de l'application

Toute l'explication a été faite, donc ici nous allons juste montrer à quoi ressemble l'application, l'interface graphique de l'utilisateur. Toutes les pages ne seront pas montrées ici, mais seulement les plus pertinentes, celles que l'utilisateur va ouvrir très souvent.

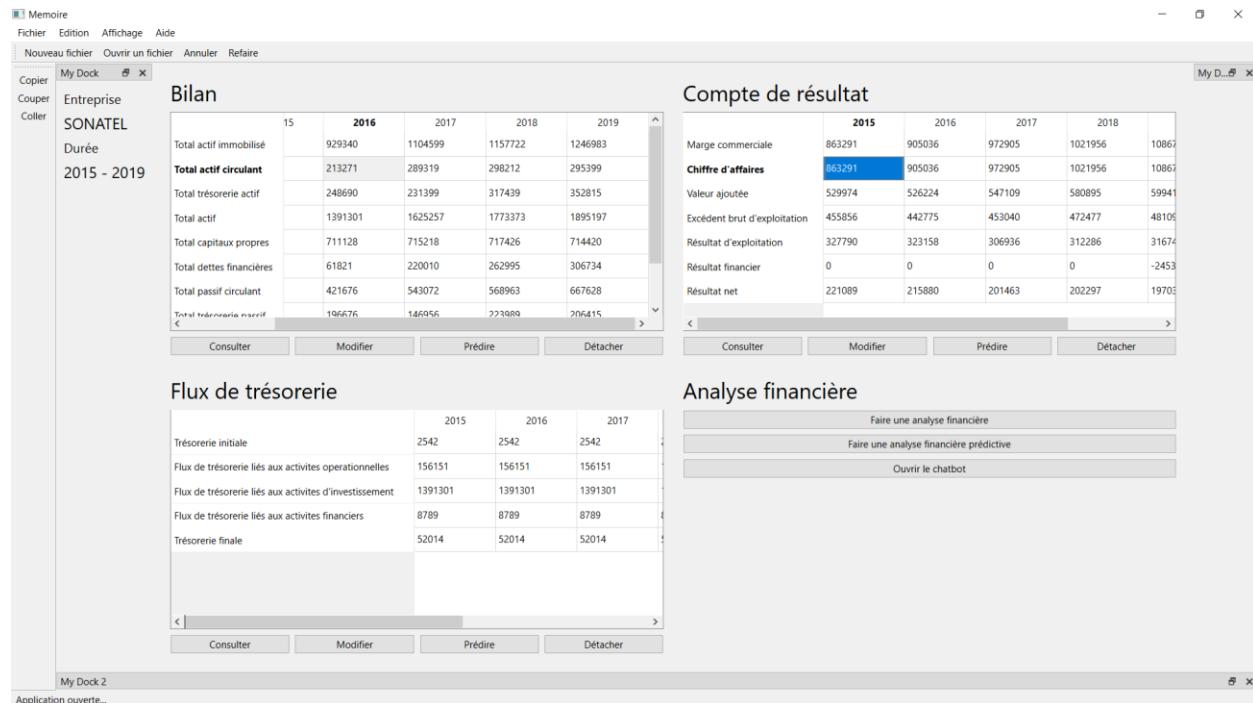


Figure 16 : Page d'accueil de l'application

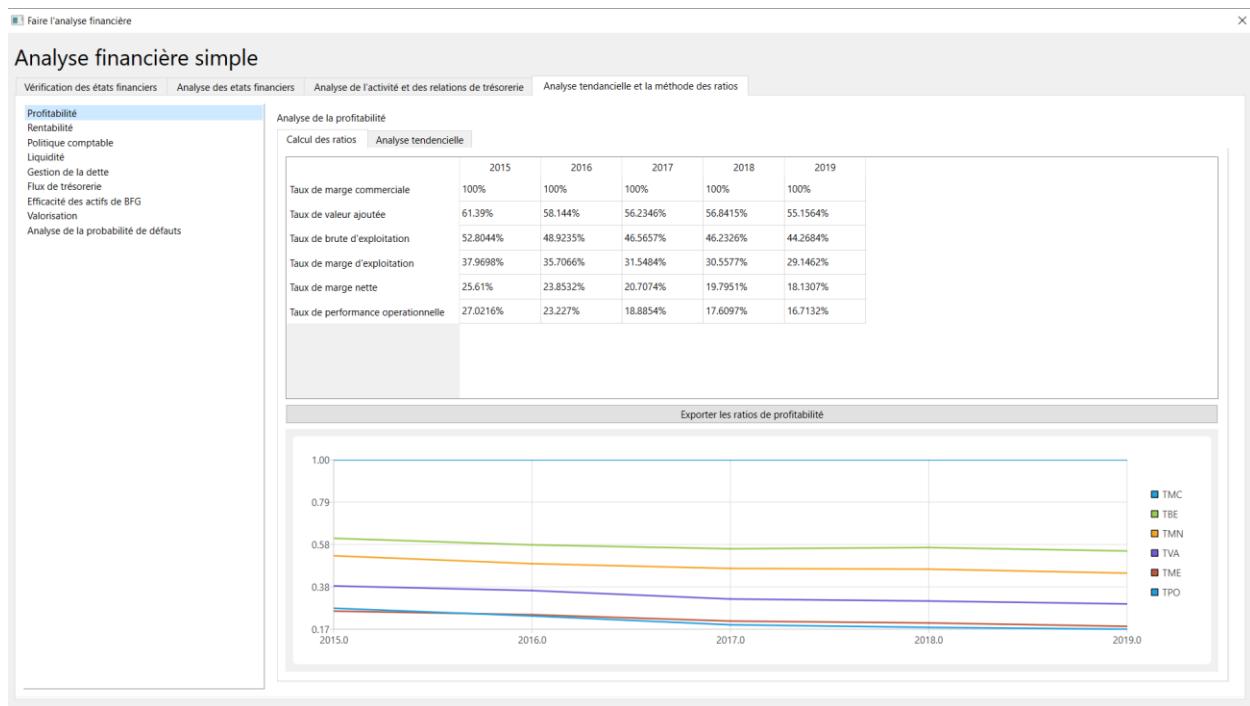


Figure 17 : Analyse des ratios de profitabilité

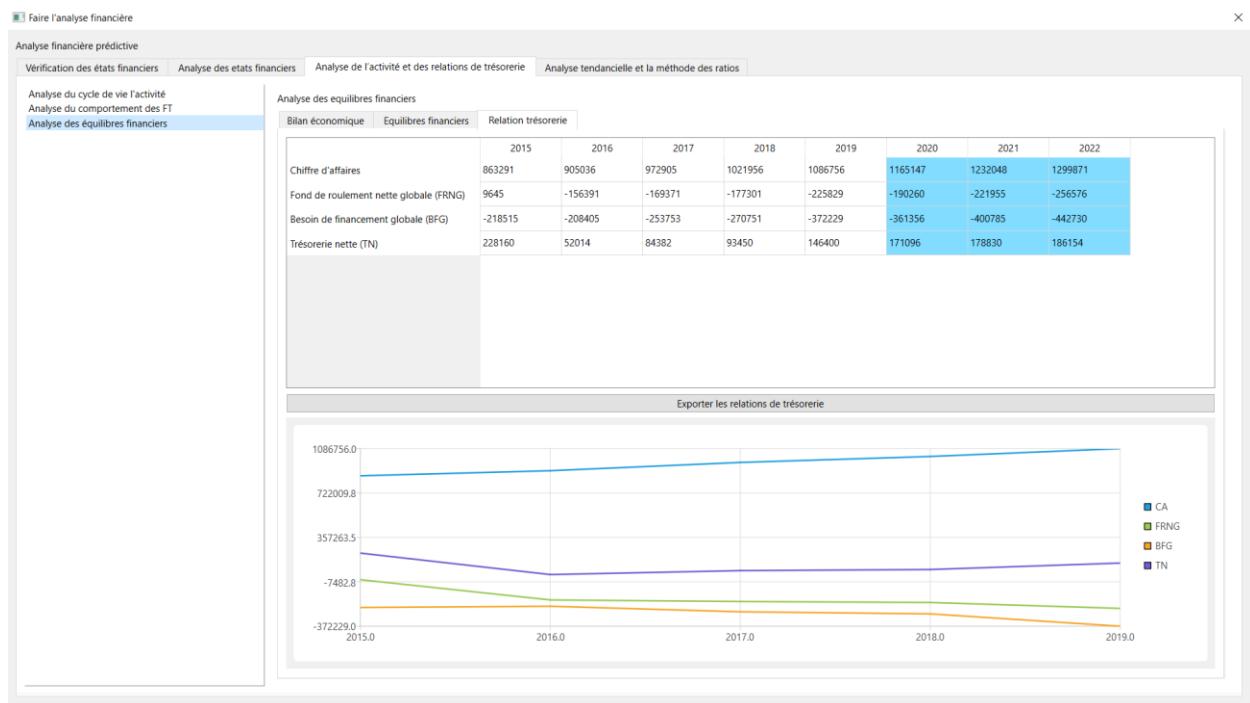


Figure 18 : Analyse prédictive des équilibres financiers

Prendre le compte de résultat

Prédition du compte de résultat										
	2015	2016	2017	2018	2019	2020	2021	2022	2020	2021
Marge commerciale	863291	905036	972905	1021956	1086756	1093448	1142024	1190608	1239184	1287760
Chiffre d'affaires	863291	905036	972905	1021956	1086756	1165147	1232048	1299871	1369129	1439822
Production stockée	0	0	0	0	0	0	0	0	0	0
Production immobilisée	3426	3877	4687	2040	0	4349	4614	4880	5145	5410
Autres produits	23926	28582	21602	37175	31045	32688	34041	35394	36747	38100
Production de l'exercice	890642	937495	999194	1061170	1117801	1129248	1179336	1229432	1279528	1329624
Achats consommés	52334	53420	61746	63338	63335	68417	71457	74497	77537	80577
Services extérieurs	308334	357851	390339	416938	354688	483667	520259	558166	597388	637927
Consommation de l'exercice	360668	411271	452085	480276	0	558812	599501	642116	685635	730569
Valeur ajoutée	529974	526224	547109	580895	599415	641576	668436	695296	722156	749016
Charges personnels	74119	83449	94069	108417	118325	117609	125068	132756	140672	148816
Excédent brut d'exploitation	455856	442775	453040	472477	481090	535388	557644	579904	602160	624420
Dotations aux amortissements	138736	131426	161192	168955	182825	176422	184059	191697	199334	206972
Reprises de provisions	10671	11809	15088	8764	8827	13548	14580	15660	16790	17969
Résultat d'exploitation	327790	323158	306936	312286	316748	371654	386788	401924	417058	432194
Produits financiers	11522	9879	10312	11236	8827	11742	12144	12546	12949	13351
Charges financières	8510	20305	18962	24753	33366	27273	29933	32731	35668	38744
Résultat financier	0	0	0	0	-24539	-15953	-18328	-20915	-23585	-26403
Résultat des activités ordinaires	330801	312732	298285	298769	292209	363136	377752	392368	406984	421600
Produits HOA	0	0	0	0	10618	0	0	0	0	0
Charges HAO	0	0	0	0	8128	0	0	0	0	0
Résultat HOA	-1811	3648	-1662	-3380	2490	-3179	-3269	-3360	-3450	-3540
Résultat avant impôts	328911	316379	296624	295390	294699	362098	376680	391262	405844	420426
Impôts sur le résultat	110105	101655	93502	95567	100133	123201	131236	139520	148053	156836
Impôts différés	2203	1156	-1659	2474	2469	2173	2268	2363	2458	2553

Figure 19 : La prédition du compte de résultat sur 5 ans

Section 2 : Développement de modèles prédictifs et l'analyse des données

Maintenant que nous savons faire une analyse financière, il est maintenant possible d'en faire une analyse financière prédictive. Mais avant toute tentative de développement de modèles, nous aurons besoin de données. Donc, nous allons voir dans un premier temps comment va se faire la collecte, l'analyse et le traitement des données, ensuite nous procéderons au développement des modèles, notamment de régression.

1. La collecte des données

Le travail que nous faisons va porter sur des entreprises cotées à la Bourse Régionale des Valeurs Mobilières (BRVM). La BRVM est le marché financier de l'Union Économique et Monétaire Ouest-Africaine (UEMOA). C'est ici que l'on peut échanger des actions et des obligations pour le compte des entreprises et des États de la zone UEMOA. La BRVM regroupe les 46 entreprises et banques les plus performantes de la zone.

Eu égard à cela, toute entreprise cotée à la BRVM a le devoir, chaque année, de publier ses états financiers pour le compte des actionnaires, des États, des investisseurs, de potentiels investisseurs ou de n'importe quelles personnes physiques ou morales. Leurs états financiers sont aussi publiés

dans le [site de la BRVM](#) ainsi, il est possible de les télécharger et de faire notre travail, on choisit une entreprise et c'est bon.

Une fois téléchargés, les états financiers d'une seule année se présentent comme suit :

- Le rapport d'activité
- Le résultat financier
- La synthèse des rapports de gestion

Mais ces données ne sont pas pour le moment exploitables, il va falloir faire un certain nombre de transformations, c'est-à-dire extraire les données qui nous intéressent, les mettre sous format CSV, JSON ou autre avant de pouvoir passer au Feature Engineering. Avec des états financiers, on peut faire une analyse financière, vérifier la rentabilité de l'entreprise, prédire des valeurs, etc.

2. La prédiction des valeurs

La prédiction est un vaste domaine qui ne relève pas seulement de l'IA. En effet, beaucoup de domaines scientifiques essaient de faire des prédictions sur les valeurs avec lesquelles ils travaillent, c'est notamment le cas de l'économie, la météo, la bourse, etc.

Les intérêts de faire des prédictions dépendent du domaine où elles sont faites. Pour ce qui nous concerne, qui est le domaine de la finance d'entreprise, cela nous permet d'approximer la future santé financière de l'entreprise en se basant sur les valeurs actuelles et passées.

Dans le chapitre passé, nous avons montré les techniques intelligentes qui permettent de prédire des valeurs. Ici, nous allons voir comment cela fonctionne en pratique.

D'abord, il nous faut des données. Elles sont collectées, et nous avons vu comment dans la précédente partie. Nous allons faire la prédiction de chaque élément de chaque état financier. Ce qui nous fait une centaine de prédictions à faire.

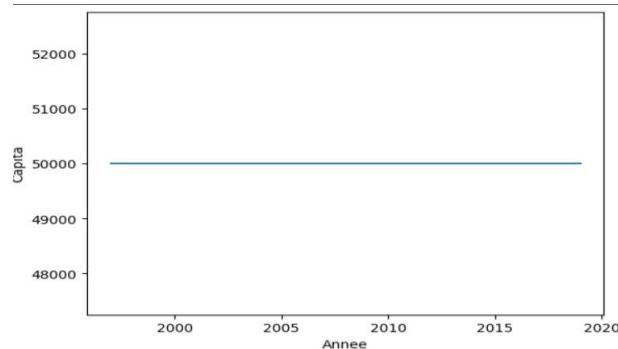
Ce que nous allons faire par la suite, c'est mettre toutes les valeurs dans un fichier CSV. Il faut rappeler que les documents téléchargés dans le site de la BRVM sont sous format PDF et donc non exploitables. Une fois sous le format CSV, nous aurons trois (3) fichiers, à savoir les bilans, les comptes de résultats et les tableaux des flux de trésorerie.

elements	annee 1	annee 2	annee 3	annee 4	annee 5	annee 6	annee 7	annee 8	annee 9	annee 10	annee 11
annee	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
charges immobilisees	2428	2474	2482	1877	1287	1015	102	73	84	287	1199
immobilisations incorporelles	488	628	2569	2940	8028	39169	40218	36369	42342	42633	60285
immobilisations corporelles	88698	121013	159338	189503	204902	215730	224943	246391	278246	333088	399672
immobilisations financieres	13661	13997	13680	14319	14516	12524	13147	14386	13001	13951	17901
amortissements et provisions											
total actif immobilises	105275	138065	178071	208639	228733	268438	278410	297220	333674	389960	479059
stock	2114	2119	3189	3670	6177	4445	3649	7600	5698	7264	8313
fournisseurs avances versees	2531	2896									
creances et emplois assimiles											
clients	12699	23401	29647	35347	40902	44797	37205	40202	89137	92928	83203
autres creances	37255	30080	27917	45499	46844	48647	59369	74976	38414	51425	57791
total actif circulant	52483	56378	60753	84516	93923	97889	100223	122599	133250	151617	149308
total tresorerie actif	48171	32115	26451	15756	30925	29673	43325	72576	98924	113732	114807
ecart de conversion actif	0	1059									
total actif	205929	227617	265274	308911	353581	396000	421958	492394	565849	655308	743175
capital	50000	50000	50000	50000	50000	50000	50000	50000	50000	50000	50000
primes et reserves	61619	42000	82277	93945	106190	126177	143464	154827	170239	206086	247969
report a nouveau	0	0	0	0							
ecart de conversion											
resultat net	29981	36246	40783	42512	47451	46682	55504	76312	105548	130628	140967
autres capitaux propres	38839	33126	0	0							

Figure 20 : Le dataset

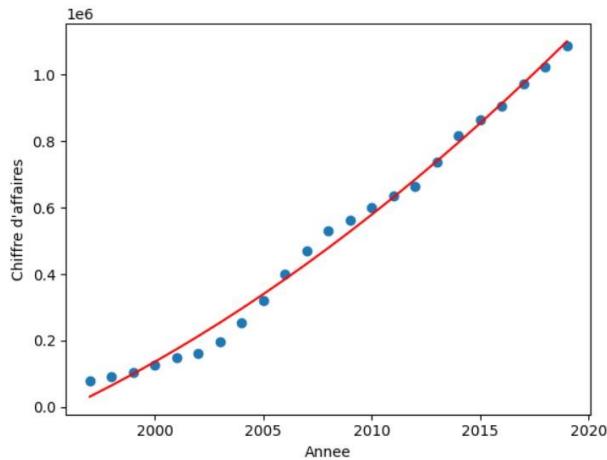
Du fait que nous avons plusieurs prédictions à faire, nous allons seulement en présenter quatre (4) du plus simple au plus intéressant.

- Le capital



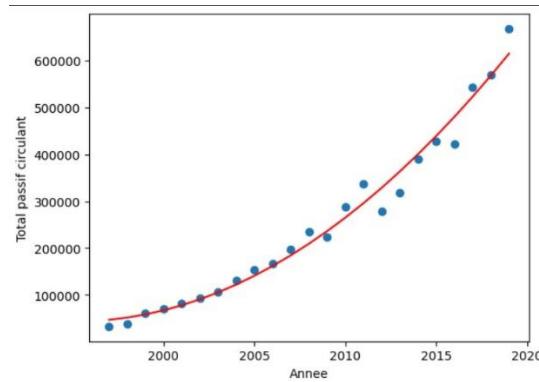
Pour ce qui est du capital de cette société, nous voyons qu'il n'a pas évolué de 1997 à 2019, donc ici il n'y a pas de prédiction à faire, puisque nous savons que le capital de cette société ne varie pas. Sur toute cette période, il reste à 50 000 000 000 de F CFA. Pour les valeurs futures et pour les calculs futurs qui vont faire intervenir le capital, nous allons choisir cette même valeur.

- Le chiffre d'affaires



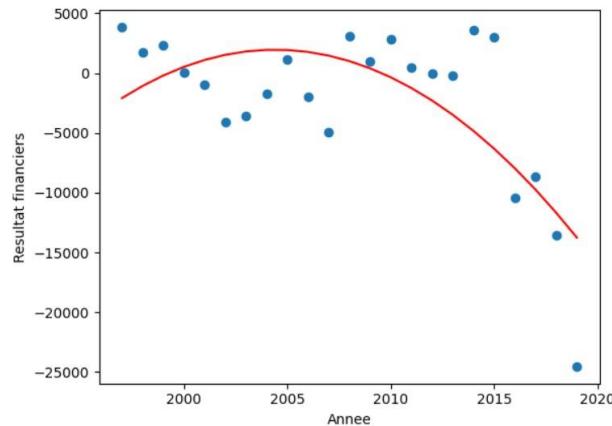
Il y a deux choses à voir ici : les points en bleu qui représentent les valeurs réelles du chiffre d'affaires en fonction des années et la droite en rouge qui est la droite de régression. La régression polynomiale a bien fonctionné ici puisqu'elle épouse à la presque perfection les données. Cette droite va renvoyer les coefficients et la constante qui représentent nos w_0 et w_1 que nous avons déjà expliqués pour faire des prédictions pour les années à venir.

- Le passif circulant



Ce cas est un tantinet plus intéressant que le précédent : les données ne suivent pas une forme linéaire, alors il nous faut autre chose. Cette autre chose, c'est évidemment la régression polynomiale que nous avons aussi déjà vue. Pour ce cas de figure, nous n'aurons pas seulement w_0 et w_1 , mais aussi w_2 et w_3 (le biais), puisque le degré de notre polynôme est égal à 2.

- Le résultat financier



Le modèle des résultats financiers est l'un des plus intéressants pour trois (3) raisons : d'abord, les valeurs sont très dispersées par rapport aux autres, ensuite les valeurs décroissent de manière exponentielle et enfin, il y a des valeurs négatives. C'est normal qu'il y ait des valeurs négatives pour le résultat financier dans la mesure où les charges financières sont supérieures aux produits financiers. Nous pourrions être tentés de penser que ce modèle va être plus compliqué que les autres, mais il n'en est rien. Nous allons simplement le faire passer dans une régression polynomiale et la magie va opérer.

Il faut noter que pour tous les modèles, nous allons suivre exactement les mêmes procédures que pour ces quatre. Tous les autres vont se classer dans l'un de ces cas de figure.

Maintenant que nous avons prédit nos modèles, il est possible de les déployer pour les utiliser dans une interface graphique. Pour ce modèle, le déploiement peut être simple du fait qu'il y a seulement des nombres. On peut les stocker dans un fichier et l'appeler depuis le frontend.

3. Validation des modèles

La validation intervient dans plusieurs étapes du processus de développement, son but est de permettre de choisir les meilleurs modèles avec les meilleurs paramètres. Dépendamment des modèles, il y aura différentes couches de validation. Il en va de soi que les modèles de régression n'auront pas autant de niveaux de validation que les ANN, RNN ou le Reinforcement Learning par exemple. Pour nos modèles, nous allons voir comment pour certains la régression linéaire est plus adaptée et pour d'autres la régression polynomiale l'est.

- Séparation en données d'entraînement et de test

Généralement, la première des choses que l'on va faire, c'est de diviser le dataset en données d'entraînement et données tests. Ceci aura pour but de s'assurer que le modèle que l'on a créé soit adaptable à de nouvelles données non connues par notre modèle. Par contre, la séparation des données pose un dilemme auquel tous les *data scientists* ont plusieurs fois été confrontés, c'est : il me faut assez de données pour faire l'entraînement, donc le maximum des données doit se retrouver à l'entraînement, mais il me faut aussi assez de données pour faire le test. Dans ce cas, il faut des arbitrages et aussi tester plusieurs divisions. En règle générale, on prend 80 % des données pour l'entraînement et 20 % pour le test. Il y a aussi un autre type de division : 60 % pour l'entraînement, 20 % pour le test et 20 % pour la validation. Cette dernière technique est utilisée pour les données de très grande taille.

Pour ce qui est de notre travail, nous avons adapté la division 80 %, 20 % des données car nous n'avions que des données de 1997 à 2019. Dans certains cas, il est fait ce qu'on appelle la cross-validation, où le dataset va être divisé en folds, disons 5 sous-ensembles, et chaque sous-ensemble va être entraîné séparément. Cela reprend un peu l'idée du random forest. Cependant, nous n'avons pas assez de données pour tester cette technique.

- Métrique de la validation

À ce moment, nous allons présenter les métriques utilisées pour la comparaison des différents modèles. Ces métriques peuvent être considérées comme des fonctions d'erreur qui nous servent à calculer la distance entre les valeurs prédites et la réalité des valeurs, et tout ceci s'applique bien évidemment aux données tests. Pour ce qui est de la régression, il y a trois métriques assez populaires que sont :

- *Root Mean Squared Error* (RMSE) : l'écart type entre les valeurs prédites et les valeurs réelles
- *Mean Absolute Error* (MAE) : l'erreur absolue entre les valeurs prédites et les valeurs réelles
- *R-squared* (R^2) : la variance entre les valeurs prédites et les valeurs réelles

Puisque nous avons plusieurs modèles, prenons l'exemple du chiffre d'affaires et calculons ses différentes métriques pour les modèles utilisés et le meilleur des modèles va être évident.

	<i>RMSE</i>	<i>MAE</i>	<i>R</i> ²
<i>Régression linéaire</i>	60299.543411	44766.908198	0.945807
<i>Régression polynomiale</i>	37793.955849	30096.976848	0.978711
<i>Random forest</i>	39616.84609	27628.576000	0.970016

Tableau 5 : Résultat chiffre d'affaires

Ce tableau ci-dessus résume bien les résultats dont nous disposons et nous permet de savoir clairement que la régression polynomiale est le modèle le plus adapté à nos chiffres d'affaires. Cette dernière présente un RMSE et un MAE plus faibles et un R² qui se rapproche le plus de 1. Donc, c'est la raison pour laquelle c'est la régression polynomiale qui a été choisie pour la prédiction du chiffre d'affaires. Il sera abordé les résultats des autres modèles dans ce qui suit.

Résultats et discussion

- **Résultats**

Pour ce qui est de ce travail, nous avons étudié plusieurs modèles de régression et quelques modèles de classification. Nous avons commencé par la collecte des données, puis avons passé au *Feature Engineering*, c'est-à-dire (l'analyse et le traitement des données). Il faut rappeler que pour les éléments du compte de résultat et du bilan, nous avons fait ce travail et les résultats ont été satisfaisants pour certains, mais laissent à désirer pour d'autres. Quoi qu'il en soit, voici les résultats de certains des modèles que nous avons développés.

L'excèdent brute d'exploitation

	<i>RMSE</i>	<i>MAE</i>	<i>R</i> ²
<i>Régression linéaire</i>	20590.648902	17408.292125	0.975251
<i>Régression polynomiale</i>	26508.796376	22186.520542	0.958980
<i>Random forest</i>	22382.350571	17977.178000	0.970757

Tableau 6 : Résultat excès brut d'exploitation

Le résultat net

	<i>RMSE</i>	<i>MAE</i>	<i>R</i> ²
<i>Régression linéaire</i>	21191.289213	16698.198544	0.886268
<i>Régression polynomiale</i>	22892.405651	18745.964893	0.867276
<i>Random forest</i>	19647.888037	16322.178000	0.902232

Tableau 7 : Résultat du résultat net

Total actifs

<i>Régression linéaire</i>	151944.961321	137918.854288	0.771658
<i>Régression polynomiale</i>	94550.704900	64463.893510	0.911582
<i>Random forest</i>	93707.317637	96959.800000	0.901471

Tableau 8 : Résultat total actif

<i>Total des capitaux propres</i>		
<i>Régression linéaire</i>	19838.776782	19273.393123
<i>Régression polynomiale</i>	46235.199281	30806.219956
<i>Random forest</i>	25752.790772	20621.528000

Tableau 9 : Résultat capitaux propres

À chaque élément du compte de résultat comme du bilan, un certain nombre de modèles a été testé et nous pouvons maintenant apprécier la pertinence de certains modèles par rapport à d'autres. Il est temps de passer à la discussion.

- **Discussion**

Maintenant que nous avons développé nos modèles et obtenu les résultats, il est enfin possible de faire une analyse de nos résultats, un peu comme les analyses de matchs après un tournoi de basket.

Concernant les résultats, nous avons constaté qu'il y a principalement trois modèles qui ont été testés : la régression linéaire, la régression polynomiale et le random forest. Et comme il est de coutume dans l'IA, ce sont les données qui définissent les modèles. Il n'existe pas encore d'outil permettant de trouver le modèle parfait du premier coup, donc il faut en tester plusieurs. Le modèle de régression linéaire, bien que simple, nous a permis de représenter plusieurs données, comme le total des capitaux propres, le report à nouveau, les impôts différés, mais bien d'autres éléments, avec bien évidemment les métriques pour le confirmer. Le random forest, bien qu'étant un modèle de classification, s'en est très bien sorti, avec, par endroit, des métriques qui avaient plus de valeur que celles des modèles de régression. C'est notamment le cas pour les primes et réserves.

De là, puisque nous avons pris l'analogie du basketball, le MVP désigné n'est autre que la régression polynomiale, qui a été la plus performante pour le maximum de données et des plus importantes, à savoir le CA, la VA, l'EBE, et j'en passe.

Néanmoins, tout ne s'est pas déroulé comme prévu, nos modèles comportent bien entendu des limites et des points d'amélioration. Le grand défi est d'avoir plus de données, car les données dont nous disposons commencent en 1997 et se terminent en 2019, avec beaucoup de données manquantes. Mais nous ne pouvons pas en vouloir à la BRVM, qui a commencé ses activités seulement en 1998. La qualité des données en finance est un défi majeur de l'IA appliquée à la finance, comme mentionné dans la section des limites des travaux de recherche. L'autre difficulté

personnelle était le nombre de modèles que nous devions créer. Pour tous les éléments du compte de résultat et tous les bilans, nous arrivons facilement à 100 modèles.

Toutes ces difficultés en valaient la peine, car nous avons attaqué un gros morceau qu'est la finance à travers l'analyse financière. En fin de compte, nous avons réussi, autant que faire se peut, à développer des modèles de prédiction acceptables. Pour toujours rester sur l'analogie du basketball, notre prochain adversaire, celui pour la finale, est le Chatbot : un adversaire coriace.

Chapitre 4 : Conception du Chatbot pour l'interrogation des états financiers

Nous voilà dans le chapitre le plus intéressant du mémoire en termes de développement de modèles. Beaucoup de choses sur le NLP vont être apprises dans ce chapitre. Nous allons suivre le *pipeline* qui nous mènera de la collecte des données jusqu'au Chatbot que nous allons utiliser pour interroger les états financiers.

Section 1 : La collecte des données

De tous les types de données que l'on va utiliser pour des modèles intelligents, bien qu'étant les plus difficiles à exploiter, les données textuelles sont les plus faciles à collecter. Nous aurons besoin de données textuelles pour le développement du Chatbot. Notre principal objectif ici est de comprendre les questions que l'utilisateur du Chatbot peut poser. Bien sûr, le Chatbot que nous allons développer sera spécialisé dans nos états financiers, donc nous aurons besoin de textes qui traitent de la finance des entreprises. Il y a plusieurs manières de collecter ce genre de texte :

- Les sites de finance

Il y a un certain nombre de sites web traitant de la finance en générale, et représentent une aubaine pour la collecte de données. Car là-bas, on peut trouver des définitions, des formules et calculs, des textes relatifs à la finance, chose qui va nous permettre d'avoir un contexte c'est-à-dire de savoir dans quelle mesure un mot va apparaître sachant qu'un autre est apparu.

- Les sites d'informations

À côté des textes relatifs à la finance, il va nous falloir évidemment du texte généraliste qui traite de tout et de rien. Les sites d'informations représentent une bonne ressource pour trouver du texte de qualité. Cela sera important pour comprendre les objets d'une question que l'utilisateur peut éventuellement poser au Chatbot.

- Les bases de données

Il existe beaucoup de datasets disponibles sur le web, mais aussi dans les bases de données. Ce sont d'autres développeurs qui créent ces datasets et qui les stockent dans des bases de données. Il y a des sites web comme [kaggle](#) que l'on peut prendre comme exemple. Dans certaines

organisations comme les GAFAM (Google, Amazon, Facebook, Apple, Microsoft), elles collectent des données et payent des personnes pour faire l'étiquetage.

- Les sites de génération de textes

Les sites de génération de textes vont nous être d'une grande utilité pour notre travail, car ils vont nous permettre d'avoir des données spécifiques à notre projet. Dans le cas d'un Chatbot spécialisé, il est très difficile de trouver des données qui nous conviennent pour l'*Intent classification* pour des raisons évidentes. Donc, l'un des meilleurs moyens d'avoir des données spécifiques à notre travail, c'est de générer, voire même de créer nos propres données.

- Le Web Scraping

Le *web scraping* est une technique en informatique où l'on développe des programmes qui vont recueillir des informations sur le web sans l'intervention directe d'une personne. Ces programmes fonctionnent de manière similaire aux robots des moteurs de recherche, bien qu'ils n'aient rien à voir avec eux, mais l'analogie est pertinente. Il est nécessaire de préciser que certaines grandes organisations comme Google n'aiment vraiment pas le *web scraping*, mais ce dernier n'est pas illégal.

Nous avons utilisé un condensé de tout cela pour obtenir les données textuelles dont nous avons besoin. Il y aura d'abord les textes qui vont nous permettre de faire l'*Intent classification*, ensuite nous utiliserons des techniques pour reconnaître les entités d'une question posée, afin de pouvoir répondre avec la plus grande précision.

Section 2 : Le développement du Chatbot

L'année 1964 à 1967 fut le temps nécessaire pour développer le premier Chatbot nommé ELIZA par un scientifique du MIT. Ce Chatbot était capable de tenir une conversation en utilisant un algorithme de reconnaissance de mots, ce qui implique qu'il ne comprenait pas vraiment le texte, mais recherchait des mots clés qu'il utilisait pour répondre.

Mais c'est quoi un Chatbot ? C'est un programme informatique comme les autres, mais qui a des aspects bien particuliers. Il sait mener une conversation qui se rapproche de celle des humains. Au début, les Chatbots étaient créés pour reproduire le comportement humain chez les machines, mais aussi pour faire avancer le NLP. Aujourd'hui, les Chatbots sont là pour nous aider dans nos vies.

Comment fonctionnent les Chatbots ? Le fonctionnement des Chatbots est simple : c'est comme une conversation entre deux personnes, l'une parle, l'autre écoute puis répond. Cependant, dans le cas des Chatbots, très souvent c'est : je pose une question et le Chatbot répond, mais il est possible d'avoir des banalités avec les Chatbots. Tout dépend de la manière dont les scientifiques ont décidé de développer le Chatbot. Il faut préciser que, même si le fonctionnement est simple, le développement ne l'est pas et nous allons voir pourquoi.

Mettons-nous à la place d'un Chatbot qui reçoit une question : comment ferions-nous pour répondre ? Oui, c'est évident pour nous, humains, mais pour un programme informatique, c'est très compliqué. C'est là que nous allons introduire les deux (2) concepts de base de tous les Chatbots : les *intents* et la détection d'entités (*entity detection*).

1. *La modélisation des textes*

Maintenant, il y a une phase qui précède l'*Intent Classification* et la détection d'entités (*Entity Detection*), c'est la représentation du texte. L'ordinateur ne comprend pas le texte et ne peut pas faire de calculs sur du texte. Or, nos données sont de type texte et nous devons les comprendre, comment faire ? Il existe un certain nombre de méthodes qui nous permettent de modéliser le texte en nombres, c'est-à-dire sous un format que l'ordinateur va comprendre.

- Label encoding

Le *label encoding* est une technique de représentation de texte où l'on associe chaque mot du *dataset* à un nombre bien particulier. Cela implique que nous aurions autant de labels que de mots.

Si nous prenons la phrase « L'argent fait le bonheur », nous aurions : L' => 1, argent => 2, fait => 3, le => 4, bonheur => 5. Par conséquent, notre phrase devient « 1 2 3 4 5 » et nous pouvons faire des calculs sur ce résultat.

- One hot encoding

En ce qui concerne le *one-hot encoding*, il est plus sophistiqué que le *label encoding*, mais un peu plus compliqué. Ici, nous allons créer un tableau à deux dimensions, avec autant de lignes que de mots. Quant aux colonnes, elles correspondront au nombre de mots dans le texte. Pour chaque mot, il sera noté avec un 1 à la place qui lui est attribuée dans la colonne, et toutes les autres valeurs seront à 0. Prenons encore la phrase « L'argent fait le bonheur » comme exemple.

	<i>L'</i>	<i>Argent</i>	<i>Fait</i>	<i>Le</i>	<i>Bonheur</i>
<i>L'</i>	1	0	0	0	0
<i>Argent</i>	0	1	0	0	0
<i>Fait</i>	0	0	1	0	0
<i>Le</i>	0	0	0	1	0
<i>Bonheur</i>	0	0	0	0	1

Le problème qui se pose ici, c'est ce qu'on appelle le *curse of dimensionality* ou la malédiction de la dimension, ce qui signifie que les dimensions vont facilement augmenter à mesure que le texte s'allonge. Imaginons un texte de 5000 mots, nous aurions 5000 éléments dans la colonne.

- Bag of Word

Le bag of words (BoW) s'inscrit dans le même cadre que le OHE, mais apporte des améliorations à ce dernier. Ici, la taille des colonnes ne varie pas, ce sont les lignes qui vont changer. La différence avec le OHE, c'est qu'ici on ne modélise pas un mot, mais une phrase. Le fonctionnement est le suivant : après avoir mis tous les mots en colonne, on compte le nombre de fois que chaque mot apparaît dans une phrase.

Exemple : « L'argent fait le bonheur », « Le bonheur fait le bonheur », « L'argent est le pouvoir »

	<i>L'</i>	<i>Argent</i>	<i>Fait</i>	<i>Le</i>	<i>Bonheur</i>	<i>Est</i>	<i>Pouvoir</i>
<i>L'argent fait le bonheur</i>	1	1	1	1	1	0	0
<i>Le bonheur fait le bonheur</i>	0	0	1	1	2	0	0
<i>L'argent est le pouvoir</i>	1	0	1	0	0	1	1

Il y a moins d'entrées et plus de signification. Le grand avantage ici, c'est qu'on modélise par phrase et non par mot. Le BoW connaît également le *curse of dimensionality*, la taille de la colonne étant proportionnelle aux données. En plus de cela, il y a aussi le problème de l'OOV (*Out Of Vocabulary*) pour les mots qui peuvent apparaître dans les données de test et qui étaient absents dans les données d'entraînement.

- Bag of n-grams

Le bag of n-grams est une variation, voire une amélioration du BoW. Dans ce dernier, on comptait mot par mot, mais avec le bag of n-grams, on compte deux mots par deux mots, trois mots par trois mots, ou enfin n mots par n mots : on parle de n-grams. Cela nous permet de donner du sens au texte : « je suis » est plus compréhensible que « je » et « suis ».

	<i>L'argent</i>	<i>Argent</i>	<i>Fait</i>	<i>Le</i>	<i>Argent</i>	<i>Est le</i>	<i>Le</i>
	<i>fait</i>	<i>le</i>	<i>bonheur</i>	<i>est</i>			<i>pouvoir</i>
<i>L'argent fait le bonheur</i>	1	1	1	1	0	0	0
<i>L'argent est le pouvoir</i>	1	0	0	0	1	1	1

Nous voyons ici, on compte par deux mots, cela fait plus de sens que de compter mot par mot. Mais il n'est pas interdit de mettre plusieurs grams. Si le gram est égal à 1 donc on revient au BOW.

- TF-IDF

Le TF-IDF, qui signifie *Term Frequency – Inverse Document Frequency*, est une méthode qui calcule la fréquence d'un mot dans l'ensemble du document. Pour ce faire, nous allons effectuer deux calculs : le TF et le IDF, puis nous allons prendre le produit des deux.

$$TF = \frac{\text{occurrence mot}}{\text{nombre total mot}}$$

$$IDF = \frac{\text{nombre total document}}{\text{nombre document qui contient le mot}}$$

$$TF - IDF = TF * IDF$$

Équation 12 : Calcul TF-IDF

A la suite de cela nous obtenons une seule valeur pour chaque mot, ce qui est plus facile à travailler avec. Le TF-IDF quant à lui souffre seulement du OOV.

- Word Embedding

Le *word embedding* est un ensemble de méthodes qui propose des moyens de transformer le texte en vecteurs. Le plus populaire d'entre eux est le *word2vec*, qui, à partir d'un mot, lui attribue un vecteur associé. Le bénéfice du *word2vec* est que des mots qui se ressemblent vont avoir des vecteurs similaires, et il est possible de trouver la ressemblance entre deux mots en calculant la similarité cosinus.

Exemple :

Père : [1, 1, 0, 0]

Mère : [0.9, 1, 0.01, 0]

Maison : [0.2, 0.54, 2, -0.5]

Bâtiment : [0.21, 0.5, 1.9, -0.49]

Nous voyons ici que les vecteurs de "père" et "mère" se ressemblent, tout comme ceux de "maison" et "bâtiment".

Maintenant, les valeurs de ces vecteurs ne tombent pas du ciel : elles sont obtenues avec le Machine Learning en utilisant ce qu'on appelle le *Self-Supervised Learning*.

- Fast text

Dans le bag of n-grams, on comptait l'occurrence des mots dans une phrase, ce qui peut bien fonctionner, mais souffre du problème de l'OOV. Le *FastText* est venu pour pallier à cela. Ce qu'il fait, c'est diviser le mot en n-grams et faire la représentation de chaque n-gram.

Si nous prenons le mot "bonheur" avec des n-grams de taille 3 : {bo, bon, onh, nhe, heu, eur, ur} et le mot "bonheur" lui-même, cela aura pour effet que, si un mot ressemblant à "bonheur" et qui n'était pas présent dans le dataset apparaît, nous pourrons tout de même travailler avec.

Il peut arriver que l'on ait besoin de faire l'inverse, c'est-à-dire qu'après avoir fait la représentation du texte et nos calculs, on obtient des résultats sous forme de nombres. Dans ce cas, il faut effectuer l'opération inverse : transformer les nombres en texte.

2. *Les Intents*

Ils représentent l'objet de la question ou la phrase ou tout ce que l'utilisateur donne comme input au Chatbot : c'est la phase de compréhension de la question. Ci-après quelques phrases et leurs objets :

Comment vas-tu ? **Salutation**

Quel temps va-t-il faire demain ? **Prédiction**

Pourquoi le gouvernement ne peut-il pas imprimer plus de monnaies ? **Explication**

Nous avons trois questions et trois objets différents, c'est-à-dire que chaque question s'attend à une réponse différente. Encore une fois, pour nous humains, cela paraît évident, mais ce n'est pas

le cas pour les machines. Alors, comment les machines font-elles pour comprendre l'objet d'une question ? Il faut faire appel à quelqu'un que nous connaissons déjà : le Machine Learning. Tout à l'heure, nous avons parlé de la collecte de données texte, maintenant nous allons les utiliser en réalisant ce que l'on appelle un *Intent Classification*.

Nous allons procéder à un apprentissage supervisé, avec les données texte qui seront étiquetées par leur objet. Les trois phrases avec leurs objets que nous avons mentionnées représentent également des exemples de ce à quoi va ressembler le dataset. Ce qui suit, c'est la représentation du texte : transformer le texte en un format compréhensible par la machine. Après avoir effectué cette transformation, nous le passons dans un algorithme de Machine Learning. Si nous obtenons une bonne accuracy, cela signifie que le modèle est bon et que notre programme est désormais capable de comprendre le texte. L'*Intent Classification* est un moyen très puissant qui nous permet de comprendre l'objet d'une question posée par un utilisateur.

3. L'*entity detection*

Maintenant que l'on comprend le sens des questions, il faut bien répondre, mais pour répondre, il nous faut des arguments. Qu'y a-t-il dans la question qui peut nous permettre de répondre ? Deux cas de figure se présentent. Il y a des questions auxquelles nous pouvons répondre directement, sans recherche d'arguments, comme lorsque l'utilisateur salue ou remercie le Chatbot. Par exemple, s'il dit « Salut » ou « Merci beaucoup », si le Chatbot parvient à bien détecter que l'un a pour objet une salutation et l'autre un remerciement, alors nous pouvons renvoyer la réponse associée.

Cependant, si l'utilisateur pose la question suivante : « Quel sera notre chiffre d'affaires de 2026 ? », et que le Chatbot a bien compris que l'objet de cette question est une prédiction, c'est là qu'intervient l'*entity detection*.

« Quel sera notre chiffre d'affaires de 2026 ? » Quelles sont les informations de cette question que nous allons utiliser pour y répondre ? D'abord, il y a le chiffre d'affaires, c'est l'élément que l'on veut prédire, et aussi l'année 2026, c'est la période pour laquelle on veut faire une prédiction. C'est cela, l'*entity detection*. Une fois que nous avons extrait « chiffre d'affaires » et « 2026 » de la question, nous appelons la fonction qui permet de faire les prédictions en lui passant ces arguments, et ensuite, nous envoyons la réponse à l'utilisateur.

Il peut arriver que l'utilisateur pose une question où il n'y a pas tous les arguments, mais le modèle parvient quand même à comprendre l'objet, exemple : « quelle est la prédition de l'actif total », c'est une prédition mais il n'y a pas de période. Dans ce cas, le Chatbot va lui envoyer un message pour préciser la période ou l'élément sur lequel va être faite la prédition si c'est ce dernier qui manque.

4. *La gestion des réponses*

Une fois toutes ces parties gérées, il faut bien répondre à l'utilisateur, c'est le sens même d'un Chatbot. En ce qui concerne notre Chatbot, nous l'avons évoqué tout à l'heure, il s'agira de réponses fixes et aléatoires. Pour chaque *intent*, nous aurons un certain nombre de réponses associées, afin de rendre l'expérience de l'utilisateur plus agréable. Comment fonctionnent les réponses fixes mais aléatoires ? Dans le cas d'une salutation, par exemple, nous allons avoir 3 réponses comme suit : « Bonjour, comment allez-vous ? », « Bien le bonjour, mon très cher », « Je vous salue ». À partir de là, une fois une salutation détectée, nous lui renvoyons l'une de ces trois réponses au hasard.

Cependant, ce subterfuge ne suffit pas vraiment à convaincre l'utilisateur de l'"humanité" du chatbot. En effet, à un moment donné, l'utilisateur se rendra compte que ce dernier ne renvoie que les mêmes réponses, bien que cela ne constitue pas un problème pour ce genre de Chatbot. C'est ici qu'intervient l'un des éléments les plus importants dans le développement de Chatbots actuellement, à savoir les *Large Language Models* (LLM). Dans ce document, nous avons vu comment réaliser une classification des intentions, qui est un moyen de faire passer du texte dans un modèle de ML, ce que l'on appelle également dans le jargon un *Language Model* (LM), c'est-à-dire de la classification de texte.

Maintenant, et c'est là que cela devient intéressant, notre *dataset* se constituait d'environ 1000 lignes étiquetées, d'accord. Imaginons simplement que nous avons un *dataset* qui se compose de l'ensemble des informations disponibles sur Internet (ce qui est juste énorme) ou d'un *dataset* qui se rapproche d'une telle abondance de données ; là, on parle de LLM. Il existe aujourd'hui de nombreux LLM, dont GPT, Gemini, Llama, etc.

Il était pertinent de faire ce détour, car nous allons utiliser un LLM, en l'occurrence *Llama* (LLM gratuit de Meta), qui va nous aider dans la gestion des réponses. Comment ? *Llama* va nous

permettre tout simplement de formuler les réponses de manière à ce qu'elles soient moins redondantes. Ce LLM pourrait aussi nous permettre d'optimiser l'*intent classification*.

5. Test de fonctionnement du Chatbot

Maintenant donnons l'exemple de fonctionnement du Chatbot. Ce que nous allons faire, c'est entrer dans la "tête" du Chatbot qui reçoit une question, la traite et renvoie la réponse.

Utilisateur : quelle est la valeur présente du total actif ?

- La première chose à faire, c'est la représentation de texte. Utilisons pour cette phrase le BOW et nous aurons un vecteur qui compte les mots qui se trouvent dans cette phrase.
- Après avoir obtenu le vecteur, nous allons le faire passer dans un algorithme de Machine Learning pour obtenir l'intent, le Naive Bayes en l'occurrence, et la sortie sera un « *get* », c'est-à-dire recevoir une valeur dans la base de données.
- Après l'intent, on passe à l'entity detection. Ici, nous avons comme entités l'actif total et « présente » qui signifie l'année actuelle. Nous aurons (actif total, 2024), ce seront nos paramètres pour la fonction de « *get* ».
- Ensuite, on appelle la fonction avec les paramètres qui nous renvoient la valeur dont on a besoin, disons 5000.
- Nous avons tout ce qu'il nous faut pour répondre. Parmi toutes les réponses de « *get* » fixes, on en choisit une au hasard, que l'on concatène avec la valeur, et la réponse est prête.

Chatbot : la valeur actuelle du total actif est de 5000.

Ci-après des exemple chat de notre Chatbot avec un utilisateur.

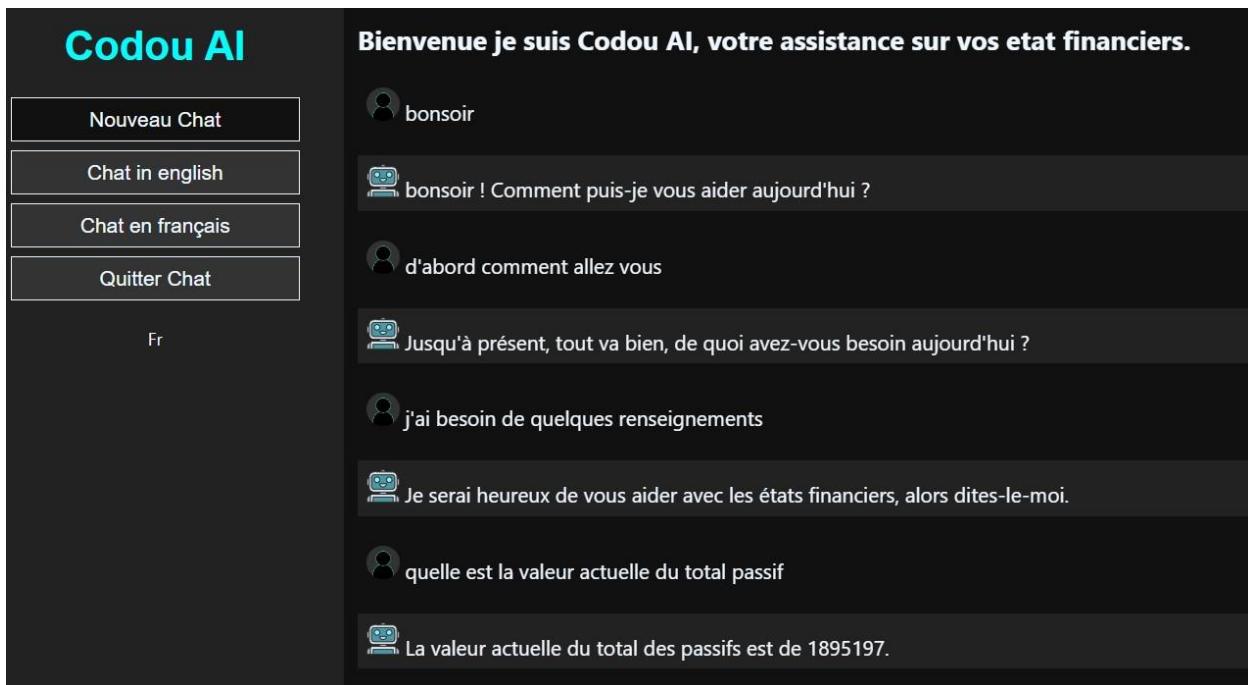


Figure 21 : Exemple conversation 1

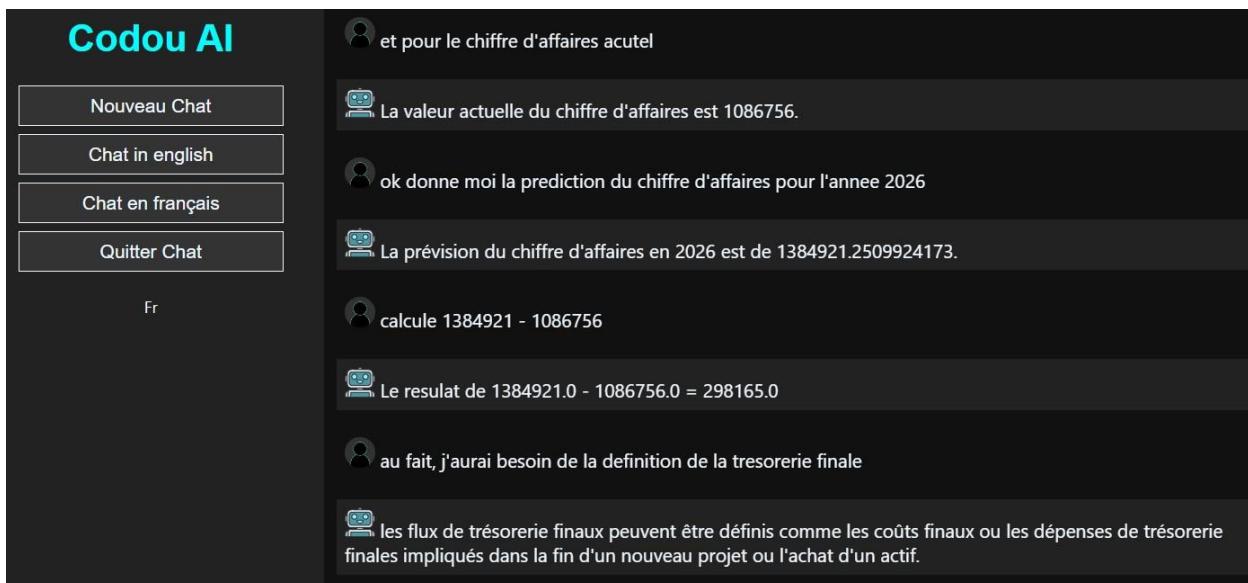


Figure 22 : Exemple conversation 2

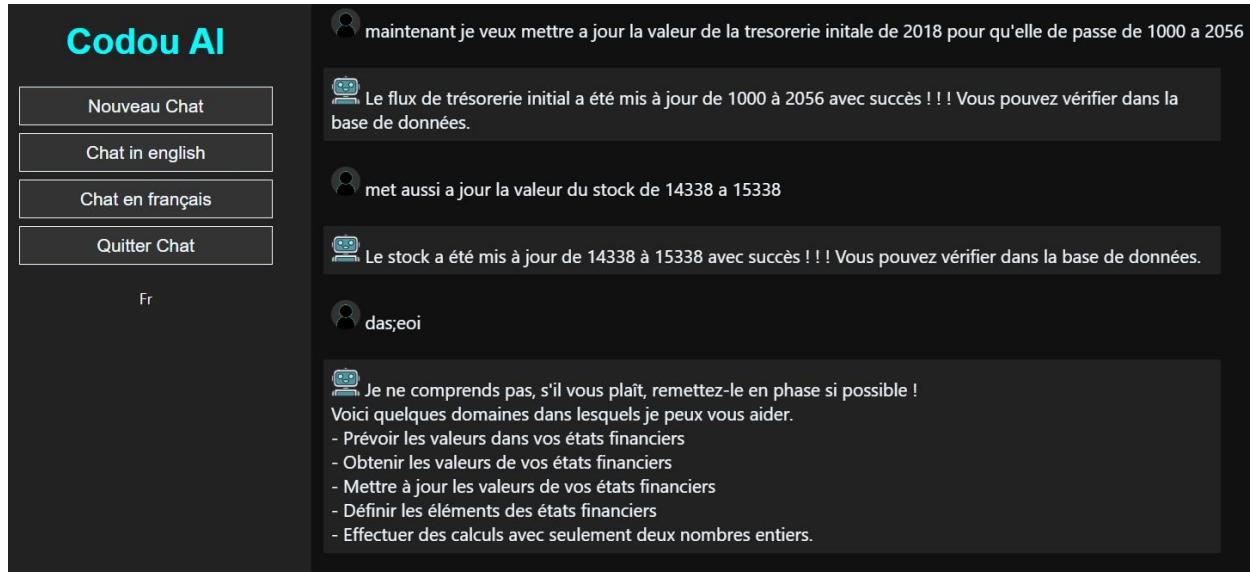


Figure 23 : Exemple conversation 3

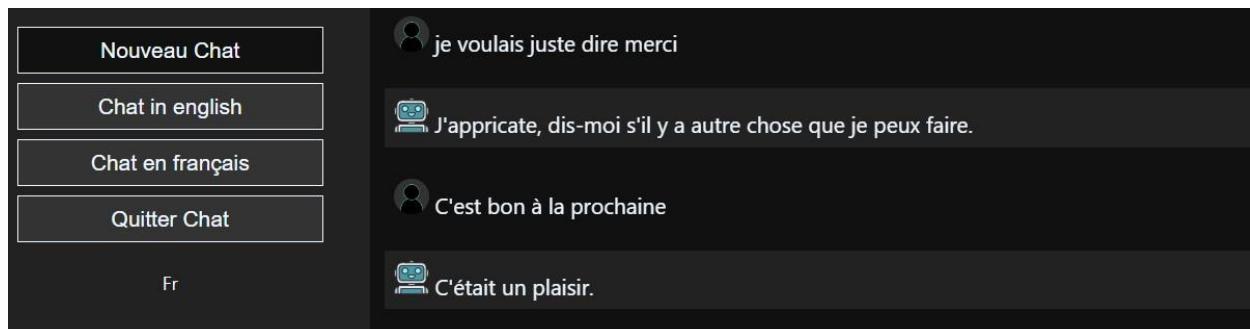


Figure 24 : Exemple conversation 4

Nous pouvons conclure cette partie sur le Chatbot en disant que pour son développement, la question est plus importante que la réponse. Si nous parvenons à comprendre parfaitement la question, nous parviendrons à donner la bonne réponse. Ci-après, une image qui illustre l'architecture du processus global.

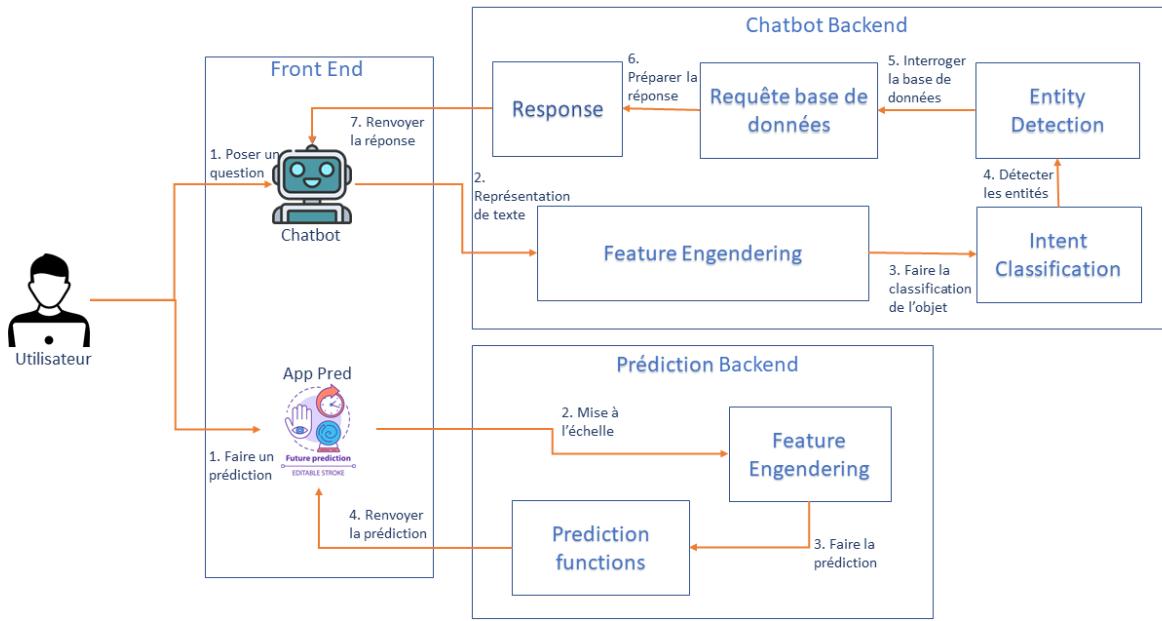


Figure 25 : Architecture globale du travail

Résultats et discussion

• Résultats

Pour dire vrai, développer ce Chatbot n'est pas une mince affaire, cela nous demandait beaucoup de compétences en IA, ce qui n'a pas toujours été facile, mais la qualité qu'il nous fallait le plus et dont nous nous sommes armés, c'était la patience. C'est ce que nous allons présenter : les résultats que nous avons obtenus durant le développement de ce Chatbot. Ce qu'il faut savoir quand on manipule du texte, c'est qu'il y a deux niveaux de difficulté. D'abord, il y a la représentation de texte : ce qui veut dire transformer le texte en nombres dans le but qu'il puisse être passé dans un modèle. Le deuxième niveau de difficulté, c'est le développement en tant que tel. La manière dont nous avons procédé est la suivante : pour chaque méthode de représentation, nous l'avons fait passer sur tous les modèles choisis. Il y aura trois méthodes de représentation, à savoir le bag of words, le bag of n-grams et le TF-IDF, ajoutées à cela six modèles que nous avons choisis.

Les métriques que nous avons utilisées sont la précision, le recall, et le f1-score, que nous avons déjà abordés dans la partie des prérequis mathématiques. Nous sommes fin prêt pour présenter les résultats obtenus.

- *SVM*

Intents	Précision			Recall			F1-score			Support
Banalité	Bow	Bon	Tf-	Bow	Bon	Tf-	Bow	Bon	Tf-	200
Remercier	0.97	0.99	0.98	0.99	0.99	0.99	0.98	0.99	0.99	200
Aide	1.00	1.00	1.00	0.98	0.98	0.99	0.99	0.99	0.99	200
Prédiction	0.99	1.00	0.99	0.97	1.00	0.98	0.98	1.00	0.99	200
Valeur	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Mis à jour	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	200
Calculer	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Définition	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Quitter	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Accuracy							0.99	1.00	1.00	1800

Tableau 10 : Résultat SVM

- *Random forest*

Intents	Précision			Recall			F1-score			Support
Banalité	Bow	Bon	Tf-	Bow	Bon	Tf-	Bow	Bon	Tf-	200
Remercier	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	200
Aide	0.98	0.97	0.99	1.00	1.00	1.00	0.99	0.99	1.00	200
Prédiction	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	200
Valeur	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	200
Mis à jour	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Calculer	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Définition	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Quitter	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Accuracy							1.00	0.99	1.00	1800

Tableau 11 : Résultat random forest

- *Decision tree*

Intents	Précision			Recall			F1-score			Support
Banalité	Bow	Bon	Tf-	Bow	Bon	Tf-	Bow	Bon	Tf-	200
Remercier	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	200
Aide	0.98	0.97	0.99	1.00	1.00	1.00	0.99	0.99	1.00	200
Prédiction	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	200
Valeur	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	200
Mis à jour	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Calculer	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Définition	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Quitter	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Accuracy							1.00	0.99	1.00	1800

Tableau 12 : Résultat arbre de décision

- *Naive baiye*

Intents	Précision			Recall			F1-score			Support
Banalité	Bow	Bon	Tf-	Bow	Bon	Tf-	Bow	Bon	Tf-	200
Remercier	0.79	0.86	0.83	0.94	0.98	0.93	0.86	0.91	0.88	200
Aide	0.97	0.99	0.97	0.96	0.98	0.97	0.97	0.98	0.97	200
Prédiction	0.99	0.99	0.98	0.94	0.98	0.95	0.96	0.99	0.96	200
Valeur	0.96	1.00	0.99	0.98	1.00	0.99	0.97	1.00	0.99	200
Mis à jour	0.96	0.99	0.97	0.98	0.99	1.00	0.97	0.99	0.99	200
Calculer	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Définition	0.99	1.00	0.98	0.84	0.86	0.86	0.91	0.93	0.92	200
Quitter	0.97	0.99	0.99	1.00	1.00	1.00	0.99	1.00	0.99	200
Accuracy							0.95	0.97	0.96	1800

Tableau 13 ; Résultat naive bayes

- *Gradient boost*

Intents	Précision			Recall			F1-score			Support
Banalité	Bow	Bon	Tf-	Bow	Bon	Tf-	Bow	Bon	Tf-	200
Remercier	0.97	0.98	0.99	0.98	0.98	0.99	0.97	0.98	0.99	200
Aide	0.98	0.97	1.00	0.98	0.99	0.99	0.98	0.98	0.99	200
Prédiction	0.99	1.00	0.99	0.97	1.00	1.00	0.98	1.00	1.00	200
Valeur	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	200
Mis à jour	1.00	0.99	1.00	0.99	0.99	1.00	1.00	0.99	1.00	200
Calculer	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Définition	1.00	0.99	1.00	0.84	1.00	0.86	0.91	1.00	0.93	200
Quitter	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Accuracy							0.97	0.99	0.98	1800

Tableau 14 : Résultat gradient boost

- *KNN*

Intents	Précision			Recall			F1-score			Support
Banalité	Bow	Bon	Tf-	Bow	Bon	Tf-	Bow	Bon	Tf-	200
Remercier	1.00	0.95	0.97	0.99	0.99	0.99	1.00	0.97	0.98	200
Aide	1.00	0.97	0.99	0.98	1.00	0.96	0.99	0.99	0.97	200
Prédiction	0.99	0.99	1.00	1.00	0.97	0.96	0.99	0.98	0.98	200
Valeur	0.99	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	200
Mis à jour	1.00	1.00	1.00	0.93	0.89	0.94	0.96	0.94	0.97	200
Calculer	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	200
Définition	0.98	0.95	0.91	1.00	1.00	1.00	0.99	0.97	0.95	200
Quitter	0.98	0.98	0.99	1.00	1.00	1.00	0.99	0.99	1.00	200

Accuracy		0.99	0.98	0.98	1800
-----------------	--	------	------	------	------

Tableau 15 : Résultat KNN

Nous voyons bien comment différentes représentations de texte peuvent influencer complètement les résultats des modèles. C'est toute l'importance de choisir une technique de modélisation appropriée. Enfin, il y a beaucoup à dire avec tous ces chiffres, mais nous allons les analyser et les interpréter, tout en définissant les limites de nos modèles dans la partie qui suit.

- **Discussion**

Puisque nous avons tous les résultats des modèles, il est maintenant temps de se poser, prendre du recul et les interpréter de manière autant objective que possible. Mais avant cela, il faut expliquer le tableau des résultats. Il y a des métriques dans la première ligne et les *intents* dans la première colonne. Ainsi, pour chaque métrique, il y a les trois méthodes de représentation. Le support représente le nombre d'individus allouant un *intent* en particulier. La première des choses à observer, c'est que pour la représentation, le tf-idf est la meilleure, car les modèles présentent des métriques plus performantes avec lui. Cela se comprend car c'est une version améliorée du BOW et du bag of n-gram. C'est pourquoi nous allons continuer cette discussion avec le résultat de cette méthode-là. S'agissant des modèles, il y en a trois qui se démarquent clairement des autres : c'est le SVM, le *random forest* et le *decision tree*.

À ce moment, il fallait choisir et cela n'allait pas se faire par hasard, donc il nous fallait un moyen de trancher. C'est ainsi qu'est venue l'idée de commencer à tester le Chatbot, en quoi faisant : se mettre à la place de l'utilisateur et lui donner des textes que les trois modèles allaient ensuite classer dans les différents *intents*. C'est un niveau supérieur de données de test. Après avoir donné plusieurs phrases, il était devenu clair que c'est le *random forest* qui se démarquait le plus. Donc, l'idée selon laquelle, quand il s'agit de texte, on commence toujours par le *naive bayes* et ensuite on teste les autres, s'est avérée être vraie ici. Donc, le modèle choisi est le *random forest* et la méthode de représentation est le tf-idf.

Les difficultés rencontrées pour ce travail restent toujours les données. Autant pour la prédiction des modèles, les données étaient disponibles sur le site de la BRVM, mais ici il fallait faire la « création » du *dataset* soi-même. Ce qui, en soi, n'est pas du tout une tâche facile. Après, nous avons aussi rencontré beaucoup de limites liées aux hardwares. Même si les données ne sont pas

gigantesques, l’entraînement pose beaucoup de problèmes, surtout dans le cas de recherche en grille.

Est-il possible d’améliorer ce travail ? La réponse est absolument oui. Dans chaque segment de développement, il y a lieu d’amélioration. Nous pouvions utiliser des techniques plus sophistiquées pour la *feature engineering* de texte, utiliser des LLM pour faire la reconnaissance des *intents* ou encore tester des modèles de LSTM ou GRU. Tout ceci serait intéressant, mais l’objectif était de faire ce travail *from scratch* et de comprendre ce qu’il se passe derrière ces technologies que nous commençons à utiliser tous les jours.

Conclusion de partie

Dans cette partie qui constitue la partie pratique de notre travail, il a été fait la présentation financière de l’analyse. Ensuite, il a été expliqué la prédiction des modèles avec les algorithmes utilisés, notamment la régression, avant de passer enfin au développement du Chatbot. Pour chacune de ces étapes, nous avons essayé tant bien que mal d’expliquer la démarche utilisée au maximum afin d’éviter de tomber dans le simplisme.

Nous arrivons à un moment charnière de notre travail. Nous avons fait un tour de l’univers de l’IA et nous allons dire, sans prendre beaucoup de risques, qu’il n’a pas été exploré 1 % du monde de l’IA, mais nous avons fait notre possible. Maintenant, il est temps de conclure ce travail fastidieux et chronophage. (**Voir la première section des annexes pour un guide d’utilisation complet des deux modèles**).

Conclusion générale et perspectives

Pour synthétiser ce travail dans lequel nous nous étions lancés, il faut rappeler la question que ce mémoire a pour but de répondre à savoir **quels modèles de Machine Learning pour une analyse et une interrogation des états financiers des sociétés cotées à la bourse régionale des valeurs mobilières (BRVM)**. Cette question centrale de recherche nous a amené à explorer les dimensions de l'IA qui pourrons nous permettre d'améliorer la performance des entreprises de l'UEMOA en leur prédisant leur santé future.

Par conséquent, répondre à cette question va, à notre sens, être évident car l'IA peut aider la finance a bien des égards. C'est ainsi que nous avons choisi la prédiction de valeur qui est un domaine de prédilection de l'IA et de l'appliquer à la finance, mais surtout le NLP pour le développement de Chatbot. Il y a bien d'autres domaines d'application de l'IA sur la finance mais pour un début ses deux peuvent s'avérer être suffisants.

Parlons des résultats, au début nous avions une centaine de prédiction à faire à savoir tous les éléments du bilan, du compte de résultat et du tableau des flux de trésorerie. Et chaque élément nécessitait une attention particulière, la raison est que les données ne se ressemblent jamais. Nous avions utilisé la régression linéaire, mais aussi la régression polynomiale par moments pour faire les prédictions. Comme nous l'avons vu certains éléments ont été faciles de travailler avec, mais d'autres plus compliquées. Il faut aussi préciser que les données avec lesquelles nous avons travaillé sur la prédiction des modèles nous viennent du site de la BRVM.

Pour ce qui s'agit du Chatbot, il y avait trois (3) niveau de difficulté d'abord il fallait trouver des données avec lesquelles il faut travailler, puis faire ce qu'on appelle un *Intent Classification*, enfin finir de faire le *Entity Detection*. La collecte de données de type texte ne fut pas un challenge de taille puisque le texte est disponible en quantité et en qualité, là où cela devient intéressant c'est quand il faut transformer le texte en un format compréhensible par l'ordinateur et par ailleurs un modèle d'IA. Nous avons vu qu'il y avait plusieurs moyens de faire cette représentation de texte, et une fois ce travail fait on peut le faire passer au *Intent Classification*. Là aussi plusieurs modèles de Machine Learning s'offrent à nous, nous avons fait la recherche en grille et choisi celui qui donne le plus grand *accuracy*. L'*Entity Detection* n'est pas encore une fois bien compliqué, nous

avons une liste de tous les éléments des états financiers, nous faisons juste une recherche de ses éléments et aussi de la période si nécessaire.

Et bien évidemment, après tout ce travail, il faut donner une interface de communication aux utilisateurs. C'est ce que nous avons fait pour terminer le travail pratique de développement, l'interface a été fait sous forme de logiciel qui regroupe les deux applications, c'est-à-dire celle des prédictions et le chatbot.

En guise de perspectives, nous avons descellé plusieurs mais ici, va être listé les plus importants. D'abord parlons de la gestion des fichiers, si cela venait à être implémentée, il serait possible d'avoir des fichiers pour plusieurs années. Il serait aussi intéressant d'avoir la possibilité de faire une rédaction complète et rigoureuse de rapport d'analyse. Pour que le chatbot soit moins robotique, augmenter les *Intents* va s'avérer être une bonne idée, ce serait vraiment intéressant si le chatbot pouvait nous répondre sur différent domaine de la finance. Il serait aussi intéressant de ne plus avoir des réponses fixes comme l'état actuel des choses, mais générer des réponses, c'est le travail de la *Generative AI*, il y a les LSTM (Long Short Term Memory) qui peuvent nous aider par rapport à cela. Actuellement, pour entrer ou modifier des valeurs il faut le faire manuellement, il est possible d'automatiser cela en implantant des CNN qui vont aller directement récupérer les informations sur des fichiers PDF voir même des images. Le travail que nous avons fait ici s'applique à l'analyse financière, nous pouvons aussi intégrer d'autres domaines de la finance pour avoir un produit encore plus puissant, d'autres domaines comme la finance de marché, la finance d'entreprise, l'analyse du risque etc. Voici de manière résumée les perspectives de l'application.

- La gestion des fichiers
- Rédaction complète et rigoureuse d'un rapport d'analyse (RNN)
- Utiliser les CNN pour extraire les informations directement sur les fichiers PDF
- Pousser le travail avec du Deep Learning sur toute l'étendue de la finance d'entreprise
- Augmenter les *Intents* pour avoir un Chatbot plus performant
- Générateur de texte pour les réponses (LSTM)

Ce qui met fin à ce travail passionnant de mémoire, nous avons appris tellement de choses en rapport avec l'IA et la finance. Ceci nous motive à poursuivre nos études de recherche pour pouvoir creuser encore plus profond dans le domaine de l'IA.

Bibliographies

Howard, W. R. (1993). On What Intelligence Is. *British Journal of Psychology*, 84(1), 27-37.

Piaget, J. (2005). *The psychology of intelligence*. Routledge.

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, 157, 17.f

Simmons, A. B., & Chappell, S. G. (1988). Artificial intelligence-definition and practice. *IEEE journal of oceanic engineering*, 13(2), 14-42.

Ginsberg, M. (2012). *Essentials of artificial intelligence*. Newnes.

Sheikh, H., Prins, C., Schrijvers, E. (2023). Artificial Intelligence: Definition and Background. In: Mission AI. Research for Policy. Springer, Cham. https://doi.org/10.1007/978-3-031-21448-6_2

Kutyniok, G. (2022). The mathematics of artificial intelligence. arXiv preprint arXiv:2203.08890.

Brette, R. (2003). *Modèles impulsionnels de réseaux de neurones biologiques* (Doctoral dissertation, Université Pierre et Marie Curie-Paris VI).

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5, 115-133.

Collins, G. S., & Moons, K. G. M. (2019). Reporting of artificial intelligence prediction models. *The Lancet*, 393(10181), 1577–1579. doi:10.1016/s0140-6736(19)30037-6

Miah, M. B. A., Hossain, M. Z., Hossain, M. A., & Islam, M. M. (2015). Price prediction of stock market using hybrid model of artificial intelligence. *International Journal of Computer Applications*, 111(3).

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, 47, 1-6.

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. In IFIP international conference on artificial intelligence applications and innovations (pp. 373-383). Springer, Cham.

Okuda, T., & Shoda, S. (2018). AI-based chatbot service for financial industry. Fujitsu Scientific and Technical Journal, 54(2), 4-8.

Muhammad Farman, M. A. (2023, Novembre). Artificial Intelligence for fraud detection and prevention. RearchGate.

Broby, D. (2022). The use of predictive analytics in finance. The Journal of Finance and Data Science, 8, 145-161.

Choi, D., & Lee, K. (2018). An artificial intelligence approach to financial fraud detection under IoT environment: A survey and implementation. Security and Communication Networks, 2018(1), 5483472.

Song, M., Xing, X., Duan, Y., Cohen, J., & Mou, J. (2022). Will artificial intelligence replace human customer service? The impact of communication quality and privacy risks on adoption intention. Journal of Retailing and Consumer Services, 66, 102900.

Buhalis, D. and Moldavská, I. (2022), "Voice assistants in hospitality: using artificial intelligence for customer service", Journal of Hospitality and Tourism Technology, Vol. 13 No. 3, pp. 386-403. <https://doi.org/10.1108/JHTT-03-2021-0104>

Alex Avelar, E., & Jordão, R. V. D. (2024). The role of artificial intelligence in the decision-making process: a study on the financial analysis and movement forecasting of the world's largest stock exchanges. Management Decision.

Yang, N. (2022). Financial big data management and control and artificial intelligence analysis method based on data mining technology. Wireless Communications and Mobile Computing, 2022(1), 7596094.

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. Expert Systems with Applications, 197, 116659.

Mbili Landu, A. (2021). Chapitre 3. La trésorerie actif. Dans : , A. Mbili Landu, *Le mémo d'un comptable: Approche par le SYSCOHADA revisé* (pp. 301-324). Paris: L'Harmattan.

- Altman, E. I. (2005). An emerging market credit scoring system for corporate bonds. *Emerging Markets Review*, 6(4), 311–323. doi:10.1016/j.ememar.2005.09.007
- Reche, J. (2019, Octobre). 7 : Schématisation d'un neurone biologique. Paris.
- Ahmad, B. (2020, Juillet). Basic design of a neural network.
- Cayla, B. (2021, Mars 7). The Stochastic Gradient Descent (SGD) & Learning Rate.
- Friedman, D. (2020). Relu Activation.
- Graph of polynomial functions. (2020, Juin 20).
- Zhao, O. (2021, Mars 24). What is a Decision Tree.
- Montalvo, N. (2023, November 9). When representation learning meets Random Forest: Deep Neural Decision Forest.
- Kumar, V. (2020, Mai 4). Neural Network.
- mathisfun. (n.d.). Introduction to Derivatives.
- Oman, S. (2016, Janvier 11). Xor representation.
- Poudel, S. (2023, Aout 28). RNN Unfolded.
- Saleem, M. (2023, Juillet). The standard logistic function.
- Shahriar, N. (2023, Fevrier 1). Convolutional Neural Network — CNN architecture.
- Kanade, V. (2022, Setembre 29). A Pictorial Representation of the Reinforcement Learning Model.

Webographies

<https://www.lafinancepourtous.com/decryptages/entreprise/gestion-et-comptabilite/comptes-de-l-entreprise/comprendre-le-bilan-le-compte-de-resultat-et-l-annexe/le-bilan/>

Tout savoir sur la trésorerie passive de l'entreprise. (2024, 5 2). Récupéré sur AGICAP:
<https://agicap.com/fr/article/tresorerie-passive-definition-utilisation/>

Marge commerciale : définition simple, exemple, calcul. (2024, 2 6). Retrieved from JDN:
<https://www.journaldunet.fr/business/dictionnaire-comptable-et-fiscal/1198457-marge-commerciale-definition-exemple-formule/>

Marge commerciale : définition simple, exemple, calcul. (2024, 2 6). Récupéré sur JDN:
<https://www.journaldunet.fr/business/dictionnaire-comptable-et-fiscal/1198457-marge-commerciale-definition-exemple-formule/>

MARCHAL, J. (2024, 2 8). *Comment élaborer un tableau de flux de trésorerie ? Intérêts et analyse.* Retrieved from L'expert comptable:

Annexes

Le guide d'utilisation de l'application :

- Le Chatbot

Pour le Chatbot, l'utilisation est simple, c'est comme tous les autres Chatbot : l'utilisateur pose des questions et le Chatbot répond dans la mesure du possible. De ce fait si vous avez déjà utiliser un Chatbot de n'importe quelle sorte, celui-ci ne sera pas différent.

Donc la nouveauté ici, c'est que le Chatbot est spécialisée dans vos états financiers, donc il peut répondre qu'aux questions relative à votre finance. Voici les domaines auxquels il peut vous aider.

- La prédiction

Vous pouvez utiliser le Chatbot pour prédire les valeurs des éléments qui se trouve dans vos états financiers.

Exemple de questions :

Quelle est la prédiction du chiffre d'affaires pour l'année 2026 ?

Quelle sera la valeur des actifs total dans 5 ans ?

Fais-moi la prédiction du résultat d'exploitation en 2028.

Donne-moi la valeur prédictive du résultat net pour 2025 ?

D'ici 2027, que pouvons-nous espérer pour la valeur du flux de trésorerie finale.

- Récupérer une valeur

Vous pouvez aussi récupérer une valeur de vos états financiers dans la période sur lequel vous êtes en train de travailler.

Exemple de questions :

Quelle est la valeur actuelle de la valeur ajoutée ?

Donne-moi la valeur des capitaux propres total pour l'année 2019.

Affiche la valeur du total passif pour l'année 2018.

Quelle était la valeur de l'excédent brute d'exploitation en 2017 ?

Récupère le résultat financier de l'année actuelle.

- Mettre à jour une valeur

Vous pouvez en outre changer une valeur dans la base de données directement avec le Chatbot.

Attention : il faut être bien sûr de soi avant d'utiliser cette fonctionnalité.

Exemple de questions :

Je veux mettre à jour la valeur du résultat exceptionnel pour qu'il passe de 5000 à 6000.

Change la valeur de la marge commerciale de 5000 à 6000.

Peux-tu ajuster la valeur du flux de trésorerie initiale de 2000 à 3000.

Pour ce qui est du stock, change la valeur de 6000 au lieu de 7000.

Modifie la valeur du fournisseur d'exploitation de 8000 à 9000.

- Définition

Vous pouvez obtenir des définitions des éléments des états financiers et les états financiers eux-mêmes.

Exemple de questions :

Quelle est la définition du bilan ?

Qu'est-ce qu'un compte de résultat ?

C'est quoi la signification du tableau des flux de trésorerie ?

Que veut dire l'actif circulant ?

Explique-moi c'est quoi la variation de stock ?

- Faire des calculs

Oui, il est possible de faire des calculs simples directement dans le Chatbot. Il est seulement possible des opérations avec deux membres.

Exemple de questions :

Calcule 45 * 98

Le résultat de 12 + 96

Fait le calcul de 45 divisé par 89

Additionne 45 et 95

898 multiplier par 9

- Les banalités

En plus de tout cela, il est possible de le saluer, le remercier, lui demander de l'aide ou même de dire au revoir.

Exemple de formules :

Bonjour, comment vas-tu ?

Bonsoir.

Je vous remercie.

Merci beaucoup.

J'ai besoin d'aide.

Peux-tu apporter une assistance.

Au revoir.

A la prochaine.

Voilà, nous avons fait le tour du fonctionnement du Chatbot, Nous espérons que c'est clair et que le Chatbot va vous être utile.

- L'application d'analyse et de prédition

The screenshot shows the 'My Dock' application interface with four windows docked:

- Bilan**: Displays a summary financial statement for SODECI for the period 2014 - 2019. It includes columns for 2014, 2015, 2016, 2017, and 2018. The total assets section shows:

	2014	2015	2016	2017	2018
Total actif immobilisées	22997090380	27031096255	30731112626	32618511653	35729306
- Compte de résultat**: Shows the profit and loss statement for the same period. Key figures include:

	2014	2015	2016	2017	2018
Marge commerciale	81229774842	87928603572	87982585098	91269534900	9843
- Flux de trésorerie**: Displays cash flow statements. For example, for 2014:

	2014	2015	2016
Tresorerie initiale	6791520188	2587596703	1154412307
- Analyse financière**: Provides analytical tools. Buttons include:
 - Faire une analyse financière
 - Faire une analyse financière prédictive
 - Ouvrir le chatbot

Quant à l'application, on observe dans la partie principale une division en quatre parties. Il y a une partie sur le bilan, le compte de résultat, le tableau des flux de trésorerie et les actions. Les trois premières parties font les mêmes choses pour les différents états financiers.

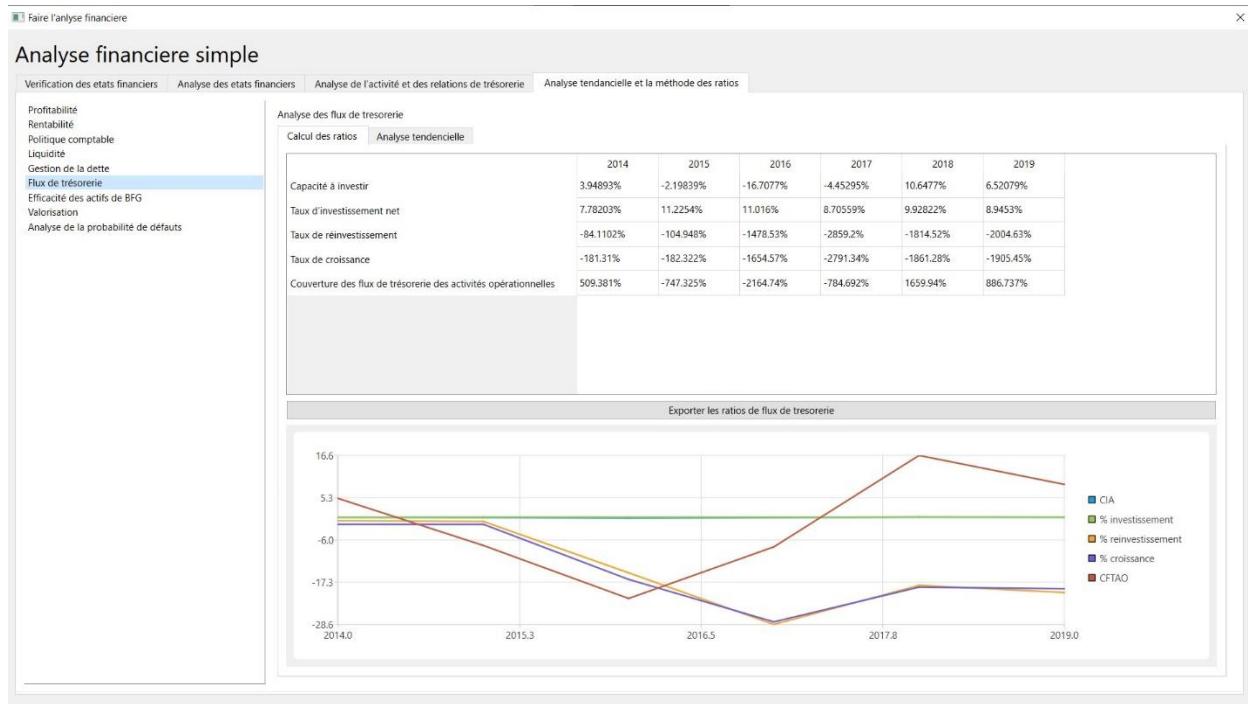
- L'affichage

Prenons le bilan comme exemple : il y a d'abord un tableau et quatre boutons en bas. Le tableau affiche un bilan résumé, les quatre boutons ont les fonctions suivantes :

- Consulter le bilan en entier
- Modifier le bilan
- Prédire le bilan : quand vous appuyait sur ce bouton un popup va apparaître pour vous demander le nombre d'année sur lequel vous voulez faire la prédition.
- Détacher : pour ouvrir une fenêtre seulement sur le bilan avec ses fonctionnalités.
- Les actions

Ensuite pour ce qu'il s'agit de la partie des actions, il y en a trois que sont : faire une analyse simple, faire une analyse prédictive, ouvrir le Chatbot (dont nous avons déjà parler plus haut).

Une fois cliquée sur le bouton faire une analyse financière, une fenêtre va s'ouvrir, c'est la fenêtre de l'analyse financière qui ne sera en rien différent de celle de l'analyse financière prédictive, seulement les années de prédictions.



Cette fenêtre a quatre onglets pour les différentes parties de l'analyse financières comme vous analyste financier en avez l'habitude.

La chose qui est intéressante ici, c'est que pour chaque analyse on aura la possibilité de l'exporter dans un fichier CSV (que l'on peut ouvrir avec Excel) avec le bouton associé.

Comment faire l'analyse d'une nouvelle entreprise ?

- D'abord il faut appuyer sur le bouton « Nouveau fichier » qui se trouve dans le menu ou en haut à gauche de l'application.
- Ensuite il faut donner le nom de l'entreprise
- Puis donnez la durée de l'analyse
- Après l'année de début d'analyse
- Ce qui suit c'est les fichiers CSV des données des états financiers de la manière suivantes (bilan, compte de résultat, tableau des flux de trésorerie)

- C'est bon, vous êtes prêt à faire une nouvelle analyse pour une nouvelle entreprise, ici un popup va vous indiquer de ne pas le fermer, il ne faut pas le fermer. La raison est que l'analyse de l'entreprise première n'est pas fermée mais cachée. Une fois terminée avec la nouvelle vous pouvez fermer le popup et retourner à l'entreprise première.

Comment créer les fichier CSV pour les données.

- Il faut créer trois fichiers avec l'extension CSV suivantes (bilan, compte de résultat, tableau des flux de trésorerie). Exemple bilan.csv, compte de resultat.csv, flux de tresorerie.csv.
- Ouvrir ce fichier avec Excel.
- Les données doivent suivre ce format (des erreurs peuvent se produire dans le cas inverse) : les éléments en colonne et les valeurs de chaque année en dessous.

Elément 1	Elément 2	Elément 3	Elément 4	Elément N
Valeur année 1				
Valeur année 2				
Valeur année 3				
Valeur année N				

Exemple du bilan

Charges immobilisées	Immobilisations incorporelles	Immobilisations corporelles	Immobilisations financières	Amortissements et provisions	Total actifs
13	31137	565288	141977	0	73841
1200	77558	595046	225046	0	92940
0	296606	625255	155950	0	110441
0	287573	706798	163352	0	115725
0	323538	759475	163970	0	124693

Annexe 1 : Exemple d'analyse financière sur Excel

Analyse horizontale du compte de résultat

Analyse du compte de résultat						
Analyse verticale						
Analyse horizontale						
Ventes de marchandises	50.29%	-12.19%	27.17%	50.99%	29.46%	
Achats de marchandises	8.72%	-1.67%	-100.00%			
Variation des stocks						
Marge brute sur marchandises	50.29%	-12.19%	27.17%	50.99%	29.46%	
Marge brute sur matières	8.72%	-1.67%	-100.00%			
Marge commerciale	50.29%	-12.19%	27.17%	50.99%	29.46%	
Ventes de produits fabriqués	11.42%	9.08%	1.83%	3.80%	9.01%	
Travaux, services vendus	3.99%	-12.93%	7.26%	14.43%	-19.56%	
Produits accessoires	13.98%	18.33%	-1.51%	5.44%	26.55%	
Chiffre d'affaires	8.25%	0.06%	3.74%	7.85%	-1.99%	
Production stockée (ou déstockage)	-161.43%	135.33%	674.47%	-22.78%	-209.02%	
Production immobilisée	-14.84%	-18.17%	15.49%	92.56%	40.67%	
Subventions d'exploitation						
Autres produits						
Transferts de charges d'exploitation	5.84%	-36.23%	-75.35%	35.31%	-15.41%	
Achats des matières premières et fournitures liées	43.78%	-9.33%	12.14%	26.63%	-5.44%	

Analyse verticale du bilan

Stocks et encours	6.15%	7.51%	7.52%	7.47%	7.59%	6.69%
Fournisseurs, avances versées	4.29%	4.73%	4.52%	2.66%	2.94%	3.91%
Clients	67.38%	61.28%	63.63%	58.96%	59.76%	53.52%
Autres créances	2.46%	2.40%	2.59%	9.42%	9.55%	17.27%
Créances et emplois assimilés	74.13%	68.41%	70.75%	71.04%	72.24%	74.70%
Total actif circulant	80.28%	75.92%	78.27%	78.51%	79.83%	81.39%
Titres de placement	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Valeurs à encaisser	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Banques, chèques postaux, caisses et ass	4.70%	7.59%	3.78%	4.75%	3.26%	2.71%
Total trésorerie actif	4.70%	7.59%	3.78%	4.75%	3.26%	2.71%
Ecart de conversion actif	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Total Actif	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Dettes fiscales et sociales	0.00%	0.00%	0.00%	0.00%	8.44%	7.19%
Dettes fiscales	30.42%	22.06%	12.15%	5.80%	0.00%	0.00%
Dettes sociales	2.26%	2.27%	1.78%	1.98%	0.00%	0.00%
Autres dettes	3.09%	2.91%	3.15%	2.80%	2.60%	2.97%
Provision pour risques à court terme (risqu	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Total Passif circulant	70.68%	61.52%	50.29%	47.28%	51.33%	54.73%
Banques, crédits d'escompte	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Banques, établissements financiers et cré	3.01%	6.88%	19.31%	26.70%	23.20%	22.24%
Total Trésorerie passif	3.01%	6.88%	19.31%	26.70%	23.20%	22.24%
Ecart de conversion - Passif	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Total passif	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Analyse horizontale du bilan

Dettes fiscales et sociales	0.00%	0.00%	0.00%	0.00%	8.44%	7.19%
Dettes fiscales	30.42%	22.06%	12.15%	5.80%	0.00%	0.00%
Dettes sociales	2.26%	2.27%	1.78%	1.98%	0.00%	0.00%
Autres dettes	3.09%	2.91%	3.15%	2.80%	2.60%	2.97%
Provision pour risques à court terme (risqu	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Total Passif circulant	70.68%	61.52%	50.29%	47.28%	51.33%	54.73%
Banques, crédits d'escompte	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Banques, établissements financiers et cré	3.01%	6.88%	19.31%	26.70%	23.20%	22.24%
Total Trésorerie passif	3.01%	6.88%	19.31%	26.70%	23.20%	22.24%
Ecart de conversion - Passif	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Total passif	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Total Passif circulant		-6.84%	-14.60%	7.03%	17.75%	20.33%
Banques, crédits d'escompte						
Banques, établissements financiers et crédits de trésorerie		144.87%	193.19%	57.46%	-5.78%	8.18%
Total Trésorerie passif		144.87%	193.19%	57.46%	-5.78%	8.18%
Ecart de conversion - Passif		7.03%	4.47%	13.86%	8.44%	12.85%
Total passif						

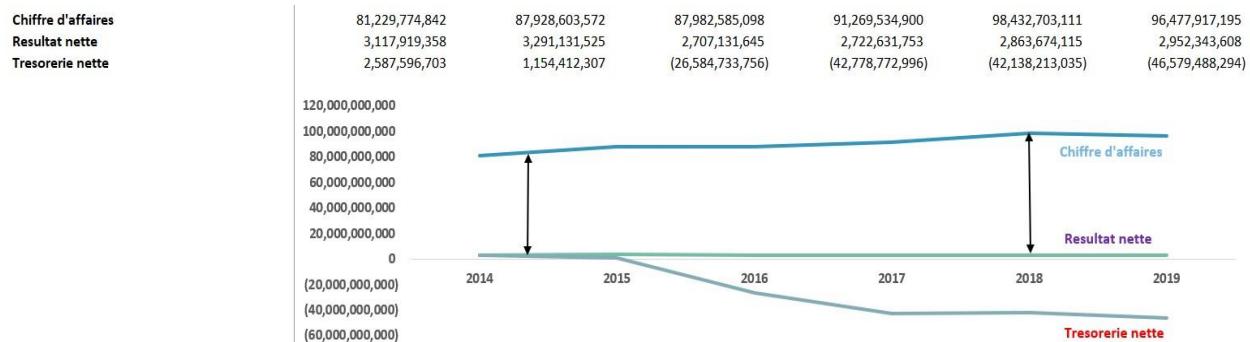
Analyse verticale du tableau des flux de trésorerie

Flux de trésorerie provenant des capitaux	-93.91%	-265.07%	12.19%	6.31%	6.41%	5.80%
(+) Emprunts	22.15%	958.02%	-6.18%	0.00%	0.00%	0.00%
(+) Autres dettes financières	57.03%	236.46%	-9.79%	-9.85%	-9.04%	-4.95%
(-) Remboursement des emprunts et autres dettes financières	28.33%	76.06%	-16.40%	-11.84%	-2.79%	-3.67%
Flux de trésorerie provenant des emprunts	50.85%	1118.42%	0.43%	1.99%	-6.25%	-1.28%
Flux de trésorerie provenant des activités de financement	-43.06%	853.35%	12.62%	8.30%	0.16%	4.51%
Variation de la trésorerie nette de l'exercice	-162.46%	-124.15%	104.34%	37.86%	-1.52%	9.53%
Trésorerie nette au 31 décembre	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Analyse horizontale du tableau des flux de trésorerie

Flux de trésorerie provenant des capitaux propres	25.93%	5.88%	-16.67%	0.00%	0.00%
(+) Emprunts	1829.42%	-85.15%	-100.00%		
(+) Autres dettes financières	84.98%	-4.61%	61.84%	-9.63%	-39.42%
(-) Remboursement des emprunts et autres dettes financières	19.78%	396.61%	16.16%	-76.78%	45.32%
Flux de trésorerie provenant des emprunts	881.19%	-100.89%	643.20%	-409.39%	-77.29%
Flux de trésorerie provenant des activités de financement	-984.21%	-134.05%	5.85%	-98.09%	3006.95%
Variation de la trésorerie nette de l'exercice	-65.91%	1835.49%	-41.62%	-103.96%	-793.34%
Trésorerie nette au 31 décembre	-55.39%	-2402.88%	60.91%	-1.50%	10.54%

Analyse du cycle de vie de l'activité



Analyse des équilibres financiers et la relation de trésorerie

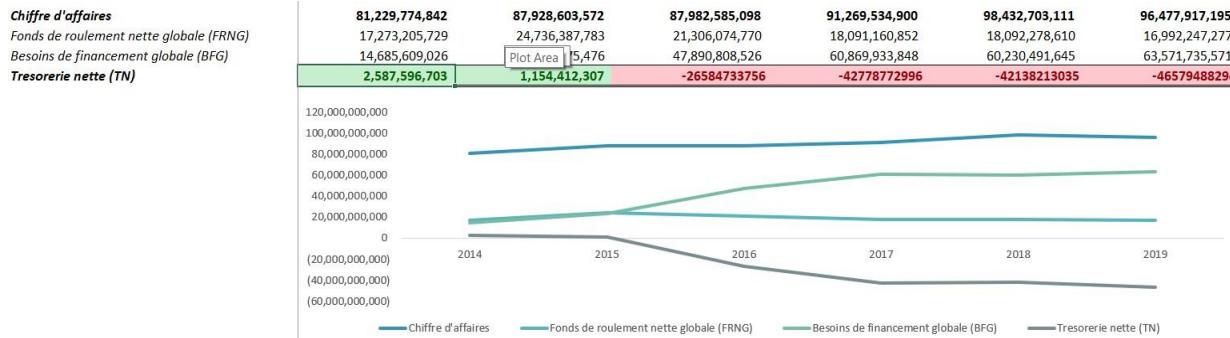
Bilan économique

Actif immobilisé (valeur nette)	22,997,090,380	27,031,096,255	30,731,112,626	32,618,511,653	35,729,308,904	37,927,999,062
Besoins de financement global (BFG)	14,685,609,026	23,581,975,476	47,890,808,526	60,869,933,848	60,230,491,645	63,571,735,571
Actif économique	37,682,699,406	50,613,071,731	78,621,921,152	93,488,445,501	95,959,800,549	101,499,734,633
Capitaux propres et ressources assimilées	13,387,703,261	13,618,834,786	13,085,966,431	13,169,476,097	13,727,894,920	14,461,067,669
Endettement nette	24,294,996,145	36,994,236,945	65,535,954,721	80,318,969,404	82,231,905,629	87,038,666,964
Capitaux investis	37,682,699,406	50,613,071,731	78,621,921,152	93,488,445,501	95,959,800,549	101,499,734,633
Verification	OK	OK	OK	OK	OK	OK

Les équilibres financiers

Fonds de roulement nette globale (FRNG)	17,273,205,729	24,736,387,783	21,306,074,770	18,091,160,852	18,092,278,610	16,992,247,277
Besoins de financement global (BFG)	14,685,609,026	23,581,975,476	47,890,808,526	60,869,933,848	60,230,491,645	63,571,735,571
Tresorerie nette (TN)	2,587,596,703	1,154,412,307	-26584733756	-42778772996	-42138213035	-46579488294

La relation de trésorerie



Analyse tendancielle et la méthode des ratios

Profitabilité

Calcul des ratios

Taux de marge commerciale	0.03%	0.04%	0.04%	0.05%	0.06%	0.09%
Taux de valeur ajoutée	27.74%	28.42%	31.05%	26.41%	28.27%	28.99%
Taux de brute d'exploitation	10.35%	9.79%	11.67%	8.30%	10.28%	10.29%
Taux de marge d'exploitation	5.75%	5.31%	4.02%	2.48%	3.89%	3.67%
Taux de marge nette	3.84%	3.74%	3.08%	2.98%	2.91%	3.06%
Taux de performance opérationnelle	3.05%	2.85%	2.07%	1.16%	1.81%	1.48%

Analyse tendancielle

Taux de marge commerciale	100.00%	138.84%	121.83%	149.36%	209.11%	276.20%
Taux de valeur ajoutée	100.00%	102.46%	111.93%	95.21%	101.92%	104.51%
Taux de brute d'exploitation	100.00%	94.62%	112.84%	80.19%	99.35%	99.43%
Taux de marge d'exploitation	100.00%	92.46%	69.91%	43.13%	67.60%	63.76%
Taux de marge nette	100.00%	97.51%	80.16%	77.72%	75.79%	79.72%
Taux de performance opérationnelle	100.00%	93.50%	67.72%	38.06%	59.33%	48.60%

Rentabilité

Calcul des ratios

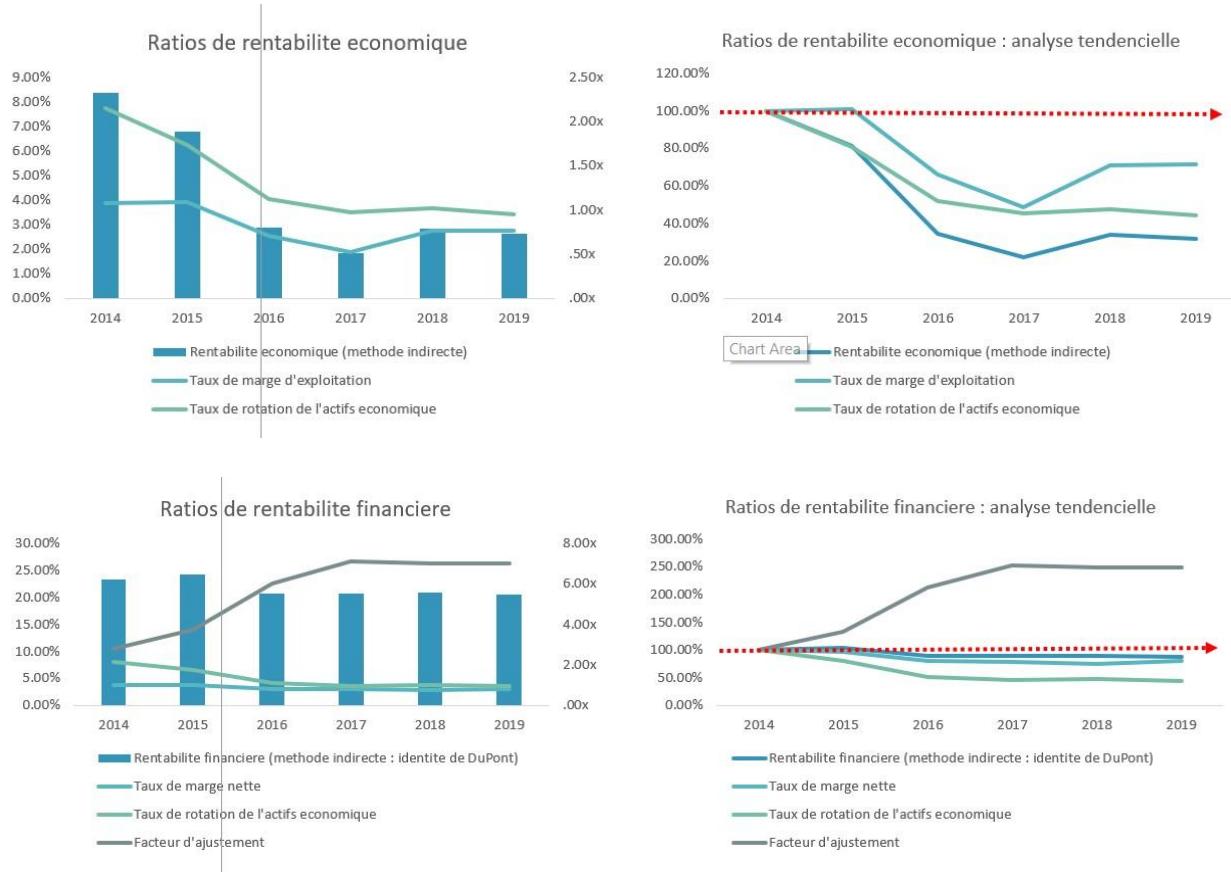
Rentabilité économique (methode directe)	8.36%	6.80%	2.87%	1.85%	2.83%	2.64%
Rentabilité économique (methode indirecte)	8.36%	6.80%	2.87%	1.85%	2.83%	2.64%
Taux de marge d'exploitation	3.88%	3.91%	2.56%	1.89%	2.76%	2.77%
Taux de rotation de l'actifs économique	2.16x	1.74x	1.12x	.98x	1.03x	.95x
Rentabilité financière (methode directe)	23.29%	24.17%	20.69%	20.67%	20.86%	20.42%
Rentabilité financière (methode indirecte : ident)	23.29%	24.17%	20.69%	20.67%	20.86%	20.42%
Taux de marge nette	3.84%	3.74%	3.08%	2.98%	2.91%	3.06%
Taux de rotation de l'actifs économique	2.16x	1.74x	1.12x	.98x	1.03x	.95x
Facteur d'ajustement	2.81x	3.72x	6.01x	7.10x	6.99x	7.02x

Analyse tendancielle

Rentabilité économique (methode directe)	100.00%	81.25%	34.26%	22.12%	33.80%	31.53%
Rentabilité économique (methode indirecte)	100.00%	81.25%	34.26%	22.12%	33.80%	31.53%
Taux de marge d'exploitation	100.00%	100.82%	65.99%	48.84%	71.03%	71.51%
Taux de rotation de l'actifs économique	100.00%	80.59%	51.91%	45.29%	47.59%	44.10%

Rentabilité financière (méthode directe)	100.00%	103.76%	88.83%	88.77%	89.57%	87.66%
Rentabilité financière (méthode indirecte : identité de DuPont)	100.00%	103.76%	88.83%	88.77%	89.57%	87.66%
Taux de marge nette	100.00%	97.51%	80.16%	77.72%	75.79%	79.72%
Taux de rotation de l'actifs économique	100.00%	80.59%	51.91%	45.29%	47.59%	44.10%
Facteur d'ajustement	100.00%	132.03%	213.45%	252.20%	248.34%	249.36%

Représentation graphique



Politique comptable

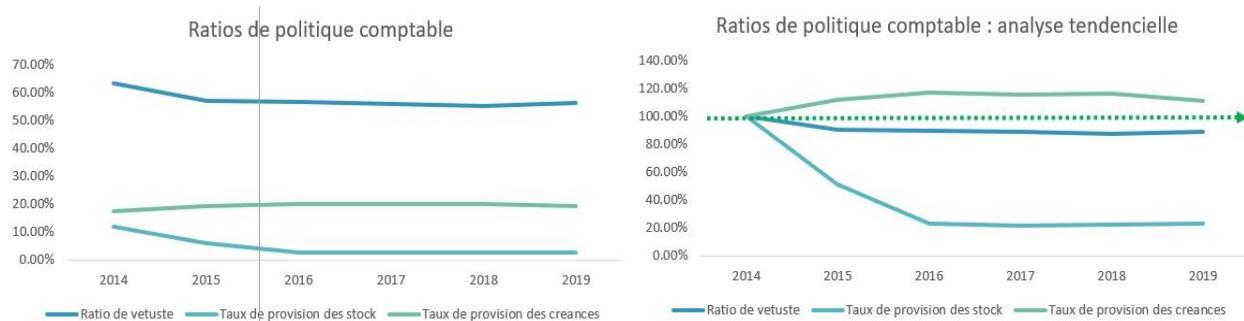
Calcul des ratios

Ratio de vétusté	63.11%	57.07%	56.57%	55.93%	55.17%	56.09%
Taux de provision des stock	11.94%	6.11%	2.77%	2.58%	2.71%	2.76%
Taux de provision des créances	17.29%	19.26%	20.16%	19.92%	20.14%	19.17%

Analyse tendancielle

Ratio de vétusté	100.00%	90.43%	89.63%	88.62%	87.42%	88.89%
Taux de provision des stock	100.00%	51.13%	23.20%	21.58%	22.65%	23.12%
Taux de provision des créances	100.00%	111.43%	116.63%	115.22%	116.53%	110.88%

Représentation graphique



Liquidité

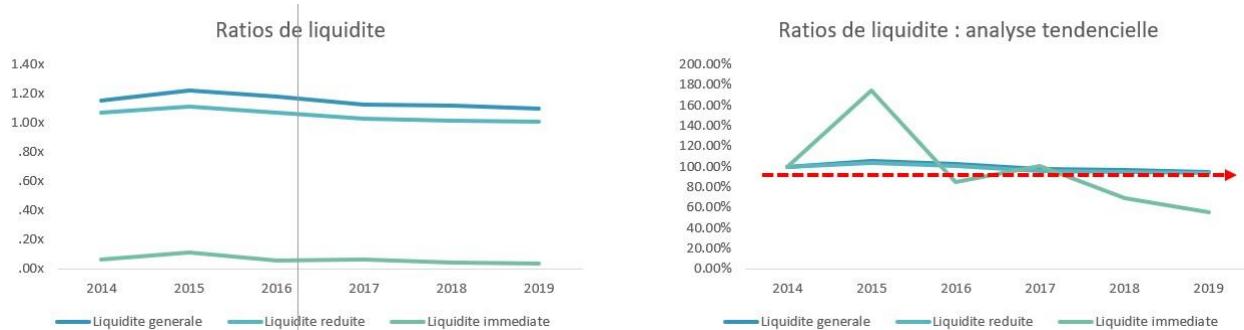
Calcul des ratios

Liquidité générale	1.15x	1.22x	1.18x	1.13x	1.11x	1.09x
Liquidité réduite	1.07x	1.11x	1.07x	1.02x	1.01x	1.01x
Liquidité immédiate	.06x	.11x	.05x	.06x	.04x	.04x

Analyse tendancielle

Liquidité générale	100.00%	105.86%	102.23%	97.60%	96.68%	94.75%
Liquidité réduite	100.00%	103.86%	100.10%	95.78%	94.70%	94.01%
Liquidité immédiate	100.00%	173.93%	85.11%	100.75%	68.62%	55.20%

Représentation graphique



Analyse de la probabilité de défauts

Tableau 16 Ratios de probabilité de défauts

Constante	Poids
1	

<i>besoin de financement globale</i>	6,56
$\frac{\text{total actifs}}{\text{report à nouveau}}$	3,26
$\frac{\text{total actifs}}{\text{resultat d'exploitation}}$	6,72
$\frac{\text{total actifs}}{\text{capitalisation boursiere}}$	1.05
$\frac{\text{total actifs}}{\text{zscore}} = \sum_{i=1}^n X_i W_i$	

Équation 13 Zscore d'Altman

Tableau 17 Zones d'Altman

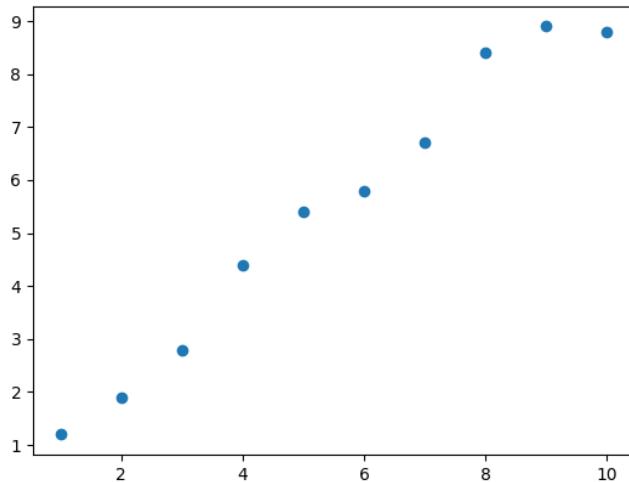
Zscore	Rating	Zone
> 8,15	AAA	
7,60 - 8,15	AA+	
7,30 - 7,60	AA	
7,00 - 7,30	AA-	
6,85 - 7,00	A+	Zone de sécurité
6,65 - 6,85	A	
6,40 - 6,65	A-	
6,25 - 6,40	BBB+	
5,85 - 6,25	BBB	
5,65 - 5,85	BBB-	
5,25 - 5,65	BB+	
4,95 - 5,25	BB	
4,75 - 4,95	BB-	Zone d'incertitude
4,50 - 4,75	B+	
4,15 - 4,50	B	
3,75 - 4,15	B-	
3,20 - 3,75	CCC+	
2,50 - 3,20	CCC	Zone de détresse
1,75 - 2,50	CCC-	
< 1,75	D	

Paramètre	Coef						
Constante	3.25	1	1	1	1	1	1
X1	6.56	0.096	0.144	0.280	0.312	0.285	0.267
X2	3.26	0.004	0.004	0.004	0.004	0.004	0.004
X3	6.72	0.031	0.029	0.021	0.012	0.018	0.015
X4	1.05	0.537	0.571	0.444	0.273	0.165	0.106
Zscore		4.662	5.000	5.705	5.676	5.427	5.222
Implication	Zone d'incertitude						
Rating	B+	BB	BBB-	BBB-	BB+	BB	

Annexe 2 : Régression exemple

Linéaire

X	1	2	3	4	5	6	7	8	9	10
Y	1.2	1.9	2.8	4.4	5.4	5.8	6.7	8.4	8.9	8.8



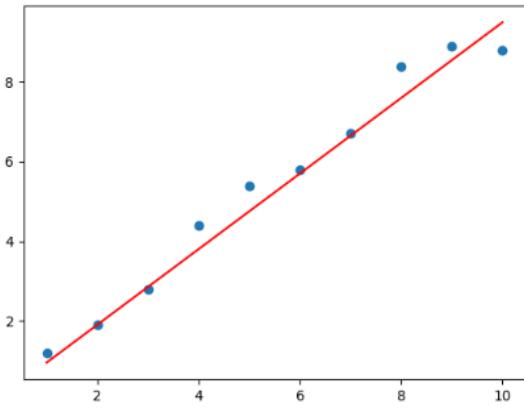
$$w_0 = 1 \text{ et } w_1 = 1$$

$$\text{Pour } x = 1, \quad y = 1 * 1 + 1, \quad y = 3$$

$$y = 3, \quad \hat{y} = 1.2, \quad MSE = (1.2 - 3)^2, \quad MSE = 3.24$$

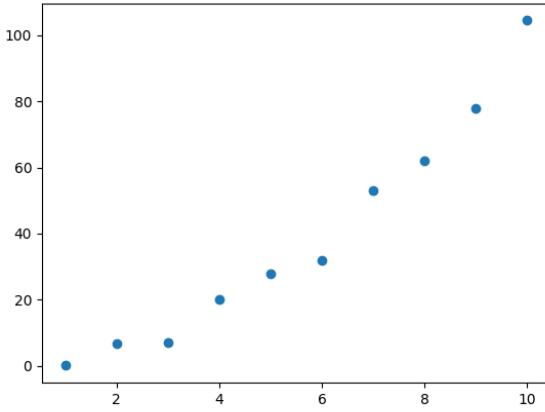
$$\frac{\partial MSE}{\partial Y} = -2 * 1(1.2 - 3), \quad \frac{\partial y}{\partial w_0} = 1$$

$$\frac{\partial MSE}{\partial w_0} = 3.6$$



Polynomiale

X	1	2	3	4	5	6	7	8	9	10
Y	0.1	6.9	7.2	20	28	32	53	62	78	104.5



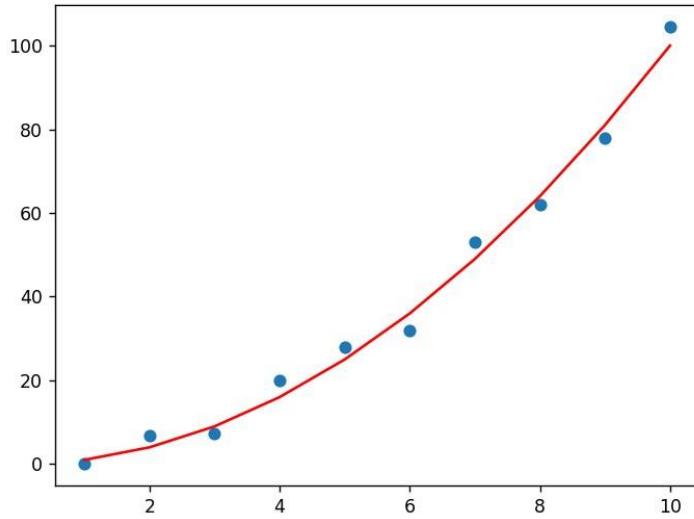
$$y = w_0 + w_1 * x + w_2 * x^2$$

$$w_0 = 1, \quad w_1 = , \quad w_2 = 1$$

$$\text{pour } x = 3, \quad y = 1 + 1 * 3 + 1 * 3^2, \quad y = 13$$

$$MSE = (7.2 - 3)^2, MSE = 17.64$$

$$\frac{\partial MSE}{\partial w_2} = -2 * (x^2) * (y - (w_0 + w_1 * x + w_2 * x^2))$$



Logistique

X1	X2	OU
0	0	0
0	1	1
1	0	1
1	1	1

$$Pour \ x1 = 1, \quad x2 = 0, \quad y = 1 * 1 + 1 * 1 + 1 * 0 = 2$$

$$\sigma(2) = 0.12$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y - \hat{y})^2$$

$$MSE = (1 - 0.12)^2$$

$$MSE = 0.77$$

Tableau 18 Resultat XOR

X0	X1	X2	W0	W1	W2	Y	$\sigma(Y)$	OU
1	0	0	-2.2121	5.41528	5.41528	-2.2121	0.099	0
1	0	1				3.20318	0.961	1
1	1	0				3.20318	0.961	1
1	1	1				8.61846	0.999	1

Annexe 3 : Classification exemple

Arbre de décision

Tableau 19 Exemple données pour arbre de décision

Numéro	Plat	Teint	Taille	Si sénégalais
1	Riz	Sombre	Grande	Oui
2	Attiéké	Claire	Petite	Non
3	Mafé	Sombre	Grande	Non
4	Riz	Sombre	Grande	Oui
5	Attiéké	Sombre	Petite	Non
6	Mafé	Claire	Grande	Oui
7	Riz	Sombre	Grande	Oui

Calculons l'entropie générale

$$Entropy(S) = -\frac{4}{7} * log_2(\frac{4}{7}) - \frac{3}{7} * log_2(\frac{3}{7})$$

$$Entropy(S) = 0.985$$

Gain d'information de l'attribut plat

$$Entropy(S_{Riz}) = -\frac{3}{3} * log_2\left(\frac{3}{3}\right) - \frac{0}{3} * log_2\left(\frac{0}{3}\right) = 0$$

$$Entropy(S_{Attieke}) = -\frac{2}{2} * log_2\left(\frac{2}{2}\right) - \frac{0}{2} * log_2\left(\frac{0}{2}\right) = 0$$

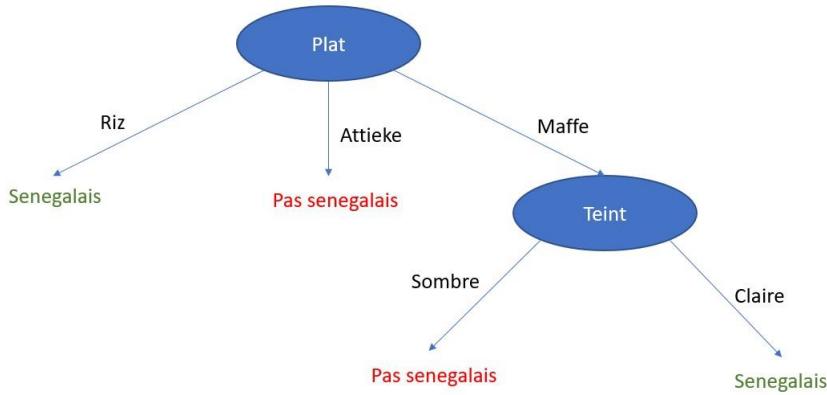
$$Entropy(S_{Maffe}) = -\frac{1}{2} * log_2\left(\frac{1}{2}\right) - \frac{1}{2} * log_2\left(\frac{1}{2}\right) = 1$$

$$GI(S, Plat) = 0.985 - \frac{3}{7} * 0 - \frac{2}{7} * 0 - \frac{2}{7} * 1 = 0.7$$

Si nous répétons les calculs avec les attributs nous allons trouver que

$$GI(S, Teint) = 0.006$$

$$GI(S, Taille) = 0.249$$



Naive Bayes

Numéro	Plat	Teint	Taille	Si sénégalais
1	Riz	Sombre	Grande	Oui
2	Attiéché	Claire	Petite	Non
3	Mafé	Sombre	Grande	Non
4	Riz	Sombre	Grande	Oui
5	Attiéché	Sombre	Petite	Non
6	Mafé	Claire	Grande	Oui
7	Riz	Sombre	Grande	Oui

Probabilité des valeurs cibles

$$P(\text{oui}) = \frac{4}{7} = 0.57, \quad P(\text{non}) = \frac{3}{7} = 0.43$$

Les probabilités des valeurs d'attributs

Plat	Oui		Non		Teint	Oui		Non	
	Riz	$\frac{3}{4}$	0	$\frac{1}{3}$		$\frac{3}{4}$	$\frac{2}{3}$	Claire	$\frac{1}{4}$
Attiéché	$\frac{0}{4}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$					
Maffé	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$					
Taille	Oui	Non							
Grande	$\frac{4}{4}$	$\frac{1}{3}$							
Petite	$\frac{0}{4}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$					

Avec ce tableau nous avons tout ce qu'il nous faut pour classer un nouvel individu. D'ailleurs c'est

ce que nous allons faire, classons I1 (Plat = riz, Teint = sombre, Taille = Grande) et I2 (Plat = Attiéché, Teint = claire, Taille = Petite).

I1 :

$$P(\text{oui}|I1) = P(\text{oui}) * P(\text{Plat} = \text{riz}|\text{oui}) * P(\text{Teint} = \text{sombre}|\text{oui}) * P(\text{Taille} = \text{petite}|\text{oui})$$

$$P(\text{oui}|I1) = \frac{4}{7} * \frac{3}{4} * \frac{3}{4} * \frac{4}{4} = 0.32$$

$$P(\text{non}|I1) = P(\text{non}) * P(\text{Plat} = \text{riz}|\text{non}) * P(\text{Teint} = \text{sombre}|\text{non}) * P(\text{Taille} = \text{petite}|\text{non})$$

$$P(\text{non}|I1) = \frac{3}{7} * \frac{0}{4} * \frac{2}{3} * \frac{2}{3} = 0$$

Annexe 4 : Unsupervised Learning exemple

Clustering (k-means)

Tableau 20 Exemple données pour K-means

	P1	P2	P3	P4	P5	P6	P7	P8	P9
X	1	1	2	5	5	6	1	1	2
Y	1	2	1	5	6	5	9	10	9
C1 = {P1},									

C2 = {P2},

C3 = {P3, P4, P5, P6, P7, P8, P9}.

$$\text{Cg}(C1) = (1, 1), \text{Cg}(C2) = (1, 2), \text{Cg}(C3) = (3.42, 6.42)$$

	P1	P2	P3	P4	P5	P6	P7	P8	P9
C1	0.0	1.0	1.0	5.66	6.40	6.40	8.0	9.0	8.06
C2	1.0	0.0	1.41	5.0	5.66	5.83	7.0	8.0	7.07
C3	5.94	5.04	5.60	2.12	1.63	2.94	3.54	4.21	2.94

C1 = {P1, P2},

C2 = {P3},

C3 = {P4, P5, P6, P7, P8, P9}.

Règles d'association

T1 = {A, B, C},

T2 = {E, F},

T3 = {A, C, E},

T4 = {A, E,}

T5 = {B}.

$$\begin{aligned}\text{sup}(A) &= \frac{3}{5} & \text{sup}(B) &= \frac{2}{5} & \text{sup}(C) &= \frac{2}{5} & \text{sup}(E) &= \frac{2}{5} & \text{sup}(F) &= \frac{1}{5} & \text{minsup} &= \frac{2}{5} \\ \text{sup}(AB) &= \frac{1}{5} & \text{sup}(AC) &= \frac{2}{5} & \text{sup}(AE) &= \frac{2}{5} & \text{sup}(BC) &= \frac{1}{5} & \text{sup}(BE) & \\ &= \frac{0}{5} & \text{sup}(CE) &= \frac{1}{5}\end{aligned}$$

$A \rightarrow C$, conf = 2/3, sup = 2/5 : cette règle n'est pas intéressante car son support > minsup et sa confiance < minconf.

$C \rightarrow A$, conf = 1, sup = 2/5 : cette règle est intéressante car son support > minsup et sa confiance > minconf.

Annexe 5 : Les outils utilisés

C++

Le C++ est un langage de programmation créé en 1985 par l'informaticien danois Bjarne Stroustrup pour pallier aux manquements du langage C qui n'est pas orienté objet. Le C++ est un langage de programmation très utilisé par les développeurs, notamment en ce qui concerne les applications. Il permet d'aborder le développement sous plusieurs paradigmes : programmation générique, procédurale et orientée objet. C'est un langage compilé, ce qui signifie que le code source est traduit en code objet ou binaire pour que la machine puisse l'exécuter. (C++ : présentation du langage de programmation, 2024)

Ce langage de programmation est un langage orienté objet ce qui veut dire il permet de créer des classes. Il est si populaire, on peut donner l'exemple de Google qui l'utilise pour son moteur de recherche, Microsoft qui l'utilise pour Word, Excel ou PowerPoint et aussi Autodesk qui l'utilise pour Maya. Pour ce qui est de l'IA, avant l'avènement de Python, les ingénieurs l'utilisaient pour écrire les codes mais son impact est toujours présent. Car derrière presque tous les Framework de Python, qui nous aident dans l'IA, il y a le C++ ou C, le cas de Numpy, Pandas ou Matplotlib.

Les avantages de C++ :

- La performance, rapidité
- La popularité
- La portabilité dans les OS
- L'abondance de bibliothèques
- La programmation orienté objet

Les inconvénients :

- Syntaxe compliquée
- Langage pas du tout pour les débutants

Python

Python est un langage de programmation créé par Guido Van Rossum. La première version publique du langage est sortie en 1991. Son nom provient de la troupe de comiques anglais les Monty Python.

Python est un langage de programmation dit de “très haut niveau”. Cela signifie qu'il possède un haut niveau d'abstraction par rapport au langage machine. Pour le dire très simplement : plus un langage de programmation est de “haut niveau”, plus sa syntaxe se rapproche de notre langage (l'anglais) plutôt que du langage machine. Un langage de haut niveau est donc plus facile à comprendre et à utiliser qu'un langage de plus bas niveau.

Certains langages (comme Python) utilisent un interpréteur comme traducteur tandis que d'autres utilisent un compilateur.

Un interpréteur se distingue d'un compilateur par le fait que, pour exécuter un programme, les opérations d'analyse et de traductions sont réalisées à chaque exécution du programme (par un interprète) plutôt qu'une fois pour toutes (par un compilateur). (Introduction à Python, 2024)

Les avantages de Python :

- Facile à utiliser
- Sécuriser
- Très populaire
- Compatibilité avec d'autres langages
- Possède beaucoup de bibliothèque pour le Machine Learning

Les limites de Python

- Temps d'interprétation très lent
- Mauvaise présentation des erreurs

SQL

Structured Query Language (SQL) est un langage de gestion de données sous forme de base de données. Il est utilisé pour gérer des bases de données relationnelles avec ces quatre (4) actions principales que sont le CRUD (CREATE, RETREIVE, UPDATE ET DELETE).

De manière simple SQL va nous permettre de créer des bases de données en utilisant un système de gestion de base de données comme PostgreSQL, Oracle, Maria DB mais nous allons utiliser MySQL. Il va être créer une base de données locale pour stocker les états financiers avec lesquels nous allons travailler.

Les bibliothèques et Framework

Un Framework est en ensemble de fonction prédéfinie dans un langage de programmation nous permettant de faire une action bien précise. Pour une tache bien définie, si nous avons un Framework, il n'est pas nécessaire de commencer de zéro puisque certaines fonctionnalités sont déjà implémentées. Les Framework ont été créer pour les tâches complexes qui nécessite beaucoup de compétence, ainsi même les développeurs de niveau moyen peuvent créer des programmes avancés, ce qui va servir à la productivité.

Il y a différents Framework pour différents domaines informatiques (développement web, mobile, logiciel...), mais nous allons présenter les Framework qui vont nous aider dans le développement de modèle intelligent et le développement d'interface graphique, ils sont tous liés soit à Python ou à C++.

Scikit-learn

Scikit-learn est une bibliothèque de Python qui a commencé en 2007 avec le Google Summer of Code Project par David Carpanneau.

Ce Framework s'est spécialisé dans le Machine Learning (supervisé et non supervisé) et nous donne des fonctions pratiques pour le développement de modèles. Scikit-learn supporte parfaitement des domaines comme la classification, la régression, le clustering ...

Pour ce qui est de nos modèles, nous allons l'utiliser pour faire la prédiction des éléments des états financiers, la représentation de texte et bien d'autres.

Tensorflow

Si Scikit-Learn est une bibliothèque de Machine Learning, Tensorflow en est une spécialisée sur le Deep Learning. Cette bibliothèque a été développée par Google dans le but de permettre aux experts mais aussi de débutant d'avoir un environnement pour travailler dans le Machine Learning en général.

De tous les Framework que nous avons présentée, Tensorflow est très probablement le plus puissant car nous permettant de faire ce que tous les autres font. C'est un outil tout en un avec ses forces et ses faiblesses.

En plus de tout cela, il nous permet de faire du Computer Vision avec les CNN, du NLP avancée avec la RNN, les LSTM, de créer des API pour le déploiement et bien d'autres.

Pandas

Pandas est un Framework Python très pratique dans le développement de modèle. Il est utilisé dans le travail a priori, le Feature Engineering. Avec pandas, nous pouvons importer des fichiers CSV, vérifier les données manquantes, les outliers...

Avant chaque développement de modèle, Pandas va certainement intervenir, ce Framework supporte les statistiques qui pourront nous permettre de mettre les données dans un format acceptable par l'ordinateur.

Numpy

Nativement, les structures de type tableau n'existe pas en Python, il y a des listes en Python pour le remplacer. La différence entre ces deux c'est que les tableaux acceptent un seul type de donnée et sa taille ne varie pas, or, les listes acceptent plusieurs types et sa taille peut varier.

C'est là qu'intervient Numpy pour permettre d'utiliser des tableaux des Python, qui sont bien plus rapide à exécuter. En plus de cela, Numpy a un excellent support de l'algèbre linéaire, les matrices, les vecteurs et autres domaines mathématiques.

Il est nécessaire d'ajouter que Numpy, bien qu'utilisé en Python est écrit en langage C qui est plus puissant et plus rapide que le Python.

Matplotlib

Matplotlib est un Framework de visualisation avec Python, il sert à tracer des courbes en utilisant Numpy ou Pandas. La visualisation peut intervenir avant et après le modèle, soit pour les comprendre les données brutes, soit pour vérifier les résultats.

La visualisation est en train de devenir une science à part entière, donc Matplotlib est utilisé dans des domaines autres que le Machine Learning, notamment dans le développement d'interface graphique que nous allons voir.

Qt

Qt est une bibliothèque de C++ cross plateforme lancée en 1995, et qui est complètement gratuit. Qt nous permet de créer des interfaces très avancées et dans un IDE et simple à comprendre et à utiliser. Avec cette bibliothèque, il n'est pas nécessaire de savoir coder pour créer des interfaces graphiques car il y a la possibilité de créer des widgets avec du glisser-déposer.

Pour ce qui nous concerne nous allons bien évidemment l'utiliser pour l'interface qui va accueillir les clients. Mais un logiciel mais pas du web.

Puisque cette application n'a pas pour vocation d'être déployé dans le cloud, le web n'est pas nécessaire. De plus, développer l'interface graphique de cette manière nous donne une certaine sécurité car il n'y aura pas de brèche que des personnes extérieures à l'organisation peuvent utiliser pour accéder aux données sensibles.

Table des matières

Remerciements	I
Sommaire	II
Liste des figures	III
Liste des tableaux.....	IV
Liste des formules	V
Liste des sigles et des acronymes.....	VI
Introduction générale	1
Partie 1 : Fondements théoriques de l'intelligence artificielle appliquée à la finance.....	4
Introduction de partie	4
Chapitre 1 : Généralités et théories de l'intelligence artificielle.....	5
Section 1 : L'intelligence artificielle : Définitions, origines et évolutions.....	5
1. Définitions de l'intelligence artificielle	5
2. Historique de l'intelligence artificielle	7
2.1. Genèse de l'IA : le premier neurone artificiel	7
2.2. Evolution	8
2.3. Les sciences qui ont impulsé sa dynamique	8
Section 2 : Fondements théoriques des algorithmes d'intelligence artificielle.....	10
1. Les prérequis de l'intelligence artificielle.....	11
1.1. Les mathématiques	11
1.2. L'informatique.....	13
2. Les algorithmes d'intelligence artificielle	15
2.1. Machine Learning.....	15
2.2. Deep Learning	26
Chapitre 2 : Revue des travaux de recherche de l'IA appliquée à la finance	30

Section 1 : Application générale de l'intelligence artificielle sur la finance	30
1. Analyse prédictive	30
2. Gestion des risques	31
3. Services clients.....	31
4. Détection de fraudes	32
5. La bourse et les marchés financiers	33
Section 2 : L'intelligence artificielle dans l'analyse des états financiers	33
1. Les travaux de l'intelligence artificielle sur l'analyse financière	34
2. Limites des travaux actuels	35
Conclusion de partie	37
Partie 2 : Conception et développement des outils d'IA appliquée à l'analyse financières	38
Introduction de partie	38
Chapitre 3 : Analyse et développement de modèles prédictifs	39
Section 1 : Mise en œuvre d'une application d'analyse financière	39
1. Démarche d'une analyse financière	39
2. Le frontend.....	45
3. Le Backend	47
4. Le Web server	48
5. Présentation de l'application.....	49
Section 2 : Développement de modèles prédictifs et l'analyse des données	51
1. La collecte des données	51
2. La prédiction des valeurs	52
3. Validation des modèles	55
Résultats et discussion	57
Chapitre 4 : Conception du Chatbot pour l'interrogation des états financiers.....	60

Section 1 : La collecte des données	60
Section 2 : Le développement du Chatbot	61
1. La modélisation des textes	62
2. Les Intents	65
3. L'entity detection	66
4. La gestion des réponses	67
5. Test de fonctionnement du Chatbot	68
Résultats et discussion	71
Conclusion de partie	75
Conclusion générale et perspectives	76
Bibliographies	A
Webographies	D
Annexes.....	E
Table des matières.....	CC