

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**

Membres du jury :

**… …**

**… …**

**… …**

Année académique :

**2022/2023**

Présenté par :

**Omar Abd Al Wahab DIASSE**

Encadreurs :

**Pr Jean Marie DEMBELE**

**Dr Alioune Badara Mbengue**

Présentée le : ..**/02/2025**

# 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](#_Toc188130188)

[Sommaire II](#_Toc188130189)

[Liste des figures III](#_Toc188130190)

[Liste des tableaux III](#_Toc188130191)

[Liste des équations III](#_Toc188130192)

[Liste des sigles et des acronymes IV](#_Toc188130193)

[Introduction de générale 1](#_Toc188130194)

[Partie 1 : Fondements théoriques de l’intelligence artificielle appliquée à la finance 3](#_Toc188130195)

[Introduction de partie 3](#_Toc188130196)

[Chapitre 1 : Généralités et théories de l’intelligence artificielle 4](#_Toc188130197)

[Chapitre 2 : Revue des travaux de recherche de l’intelligence artificielle appliquée à la finance 29](#_Toc188130198)

[Conclusion de partie 35](#_Toc188130199)

[Partie 2 : Conception et développement des outils d’IA appliquée à l’analyse des états financiers 36](#_Toc188130200)

[Introduction de partie 36](#_Toc188130201)

[Chapitre 1 : Analyse de développement de modèles prédictifs 37](#_Toc188130202)

[Chapitre 2 : Conception du Chatbot pour l’interrogation des états financiers 56](#_Toc188130203)

[Conclusion de partie 70](#_Toc188130204)

[Conclusion générale et perspectives 71](#_Toc188130205)

[Bibliographies Webographies A](#_Toc188130206)

[Annexes A](#_Toc188130207)

[Table des matières B](#_Toc188130208)

# Liste des figures

# Liste des tableaux

# Liste des formules

# 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 | Excèdent Brute d’Exploitation |
| F CFA | Franc Communauté Financière Africain |
| 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 de 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 trouver 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 ce mémoire. C’est ainsi que nous avons choisi comme sujet de mémoire « Développement d’un modèle de machine pour faire une analyse financière (historique et prédictive) et le développement 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 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. 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 et de l’IA s’est renforcée puisque nous avons eu la chance de démystifier ce domaine complexe et très intéressant. 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. 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, déjà en 1998 le professeur Seydina Moussa Ndiaye nous parler de Reinforcement Learning dans l’un de ses articles les plus cités. 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 : dans quelle mesure l’intelligence artificielle va pouvoir aider les entreprises cotées à la Bourse régionale des Valeurs Mobilières (BRVM) à faciliter leur analyse financière et tout ce qui tourne au tour de leurs états financiers ? Afin de mener cette mission à bien, nous adopterons une démarche bien spécifique, tout d’abord nous nous attèlerons à trouver des données avec lesquelles nous allons travailler. 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. Deux familles de modèles vont être utilisées, il y a les modèles de prédiction et aussi les modèles 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, communiquer avec un Chatbot 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 chapitres. Nous allons d’abord voir toute la littérature sur l’intelligence artificielle et la finance, les définitions, l’histoire, les recherches scientifiques (Chapitre I). En plus de cela nous réserverons une partie spéciale pour découvrir l’analyse financière, ses méthodes, son fonctionnement, ses intérêts (Chapitre II). L’une des parties les plus intéressantes de ce document, sera la partie des théories derrière l’intelligence artificielle (Chapitre III). Nous allons terminer avec la présentation des modèles ou nous allons voir de manière pratique comment crée-t-on une intelligence artificielle mais aussi la présentation d’une interface graphique conviviale pour accueillir les clients (Chapitre IV).

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

## Introduction de partie

Apres notre introduction générale, il d’abord être fait la partie une de notre document qui va une aventure vers le soubassement de l’IA. Il serait bien entendu impératif de connaitre quelques notre liées à l’IA afin de mieux aborder la suite les 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 chapitre, d’abord les généralités et théories de l’IA ou nous allons évoquer la question de la définition de l’IA. Puis la revue de la littérature scientifiques de l’IA appliquée à la finance, dans nous aborderons les applications de l’IA de sur la finance dans un premier ensuite les travaux de l’IA sur l’analyse des états financiers.

Une fois tout ceci fait de la manière la plus adéquate possible, vous allez la suite va être plus agréable à suivre. Puisque la deuxième partie sera dédiée au développement de modèles, mais avant cela, faisons d’abord connaissance avec l’IA.

## Chapitre 1 : Généralités et théories de l’intelligence artificielle

En premier lieu, nous allons aborder quelles 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. La haute importe que relève ce chapitre est de lever l’ambigüité sur notre science car certaines personnes usent de ses termes à tort et à travers. Nous allons sans plus tarder commencer l’explication liée à tous cela.

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

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

#### 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 c’est quoi une intelligence artificielle. Cependant, il va être difficile de définir une IA sans savoir c’est quoi l’intelligence, sans savoir c’est quoi artificiel. Une fois s’être bien familiarisé avec ces deux termes, cela va s’avérer bien plus simple de parler de l’IA. Fort heureusement, beaucoup de recherches, études ont été faites par les scientifiques académiciens sur ses termes que nous nous donnons la tache 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, ou est son siège dans le cerveau et on ne peut pas vraiment voir de différence notable entre le cerveau d’une personne intelligente et celle 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 définition 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 n’est pas seulement les psychologues qui ont tentés de définir l’intelligence, les chercheurs en IA aussi, c’est le cas de James S. Albus qui le 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 réagi en conséquence (Outputs).

* **C’est quoi artificielle ?**

La plupart des gens penseraient que l’artificielle est un terme sans ambigüité et donc facile à définir, et à juste titre d’ailleurs. Mais nous savons qu’il y a des subtilités qui se cache derrière ce terme surtout dans le domaine de l’IA.

Des définitions simples comme « l’artificielle est tout ce qui est créé par l’homme » ou « c’est l’œuvre de l’Homme s’inspirant de la nature » ne sont plus d’actualité.

C’est pourquoi nous allons donner la définition de Francesco Bianchini dans son article « A New Definition of “Artificial” for Two Artificial Sciences » : “The artificial is what is humanly constructed, often in a natural model, also through the manipulation of natural systems and processes, and maintains existing and acting/operating/behaving in an open-ended context or environment without human control, regardless the substance or materials of its constituent parts”.

Cette définition prend l’ensemble des aspects relatifs à l’IA mais aussi quelque peu à la biologie artificielle. Elle met aussi en avant non seulement la construction de l’Homme mais aussi sa non-intervention dans les opérations dites artificielles.

Mais un problème se pose, selon Francesco Bianchini lui-même avec cette définition peut-on considérer une IA créée par une IA toujours artificielle ?

* **Proposition de définitions de l’intelligence artificielle**

Nous y voilà, les termes que nous avons décidé 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 donnait Haugeland en 1885 : “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”. Dans la suite du document, nous allons parler du teste de Turing.

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 & Erik Schrijvers ont copublié un article « Artificial Intelligence : Definition and Background » ou 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 recherches. Mais à la fin ils se rejoignent tous.

Au vu de tout cela, nous pouvons conclure que l’AI a pour objectif d’imiter l’intelligence humaine en faisant des taches qui auraient pu être jugées impossible à faire pour les machines. Cette dernière n’est pas une proposition de définition.

#### Historique de l’intelligence artificielle

L’IA est une science intrigante voire mystérieuse qui pour certaine personne c’est une boite noire, et bien des choses se sont passer pour qu’elle devient l’une science les populaire et qui fait peur, disons-le, à beaucoup de professionnels. Son histoire est riche et rocambolesque, nous pourrions facilement en faire un film. Pour ce qui suit il va être fait un bref historique des évènements les frappant de l’IA de 1943 (naissance présumée de l’IA) à nos jours.

##### Genèse de l’IA : le premier neurone artificiel

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

Durant cette année, une étude menée par Warren S. McCulloch & 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 ses deux scientifiques était, pour chaque étape de la transmission d’information d’un neurone a un autre, de trouver une fonction mathématique qui pourra le répliquée. A 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 au 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 neurone, après Frank Rosenblatt va créer le perceptron en s’appuyant sur les travaux de ses prédécesseurs et ainsi de suite.

##### Evolution

Depuis la création du premier neurone qui constitue de premier article scientifique qui s’inscrit dans le l’intelligence artificielle, bien des choses se sont passées pour aboutir au status quo. Voici un timeline d’évènement frappant 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 machine think ? ».

**1965** : L’un des premier Chatbot (ELIZA) capable de s’exprimer a été créé par Joseph Weizenbaum (MIT) et est l’une des premières machines a 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 taches propres à l’homme.

**2016** : Alpha Go devient la première société à avoir réussi de créer une AI impossible à battre dans le jeu du Go.

**2022** : Open AI lance Chat GPT et Google lance Google Bard (actuellement Gemini)

##### Les sciences qui ont impulsé sa dynamique

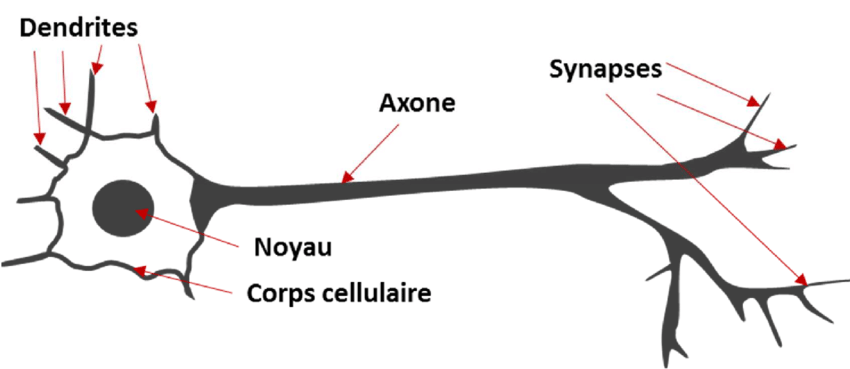
* **Les mathématiques**

“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 le vient de rappeler Gitta Kutyniok, c’est-à-dire que tous les algorithmes d’IA reposent derrière des théories mathématiques. D’ailleurs nous allons voir cela dans la suite du document en détails mais pour le moment donnons des 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.



* **L’informatique**

Là, un non-initié peut 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 faire donner un petit aperçu de comment marche un algorithme d’IA pour comprendre cela. Un réseau de neurones prend des données en entré, les normaliser, 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), on peut 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 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 est dérivée de plusieurs autres sciences, c’est le mathématicien américain Norbert Wiener que l’on attribue la création de cette science. La cybernétique essaie de répondre à la question de savoir comment les systèmes peuvent se contrôler elles-mêmes. L’exemple qui est très souvent donné c’est le 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 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

A 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 boite noire que constitue cette science. Dans cette section il sera la prise de conscience des fondamentaux mathématiques et informatiques de l’IA, ensuite nous allons voir le fonctionnement de certains algorithmes des modelés intelligents qui façonnent nos vies.

#### Les prérequis de l’intelligence artificielle

Avant tout, il va falloir préciser un certain nombre de choses. L’IA n’est pas facile, ce n’est pas un domaine auquel n’importe qui sans le maximum de volonté peut y entrer. Bien évidemment il y a quelques prérequis, deux pour être précis que sont les mathématiques et l’algorithme. Et pour ces deux domaines il va falloir être excellent dans un et avoir un bon niveau dans l’autre ou maitriser les deux. A 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 ses autres prérequis vont être classés dans le domaine de l’intelligence sociale.

##### Les mathématiques

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

Les mathématiques peuvent être compliquées, mais quand on lui 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 comment c’est fascinant de résoudre des problèmes mathématiques pour créer des modèles intelligents. Les mathématiques sont plus que nécessaire pour l’IA, elles sont vitales. D’ailleurs mon professeur d’intelligence artificielle nous disait à la fin d’un cours : « l’intelligence artificielle ce n’est ni plus ni moins que des calculs mathématiques ».

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

* **Les statistiques et probabilités**

Is everything in 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, manufactures, and others are using tools from probability and statistics. (Dekking, Frederik Michel, 2005).

Cette citation de Michel nous renvoie à comment sont important ses domaines dans nos vies de tous les jours et l’IA ne fait pas exception. La statistique est un domaine des mathématiques qui travaillent 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 statistique car les statistiques inférentielles vont avoir besoin des probabilités. La probabilité est l’étude de la chance pour qu’un évènement se produise pour faire simple.

Ceci étant dit, comment ses deux sont utiles en Machine Learning et Deep Learning ? Ils interviennent tous les deux avants et après le développement de 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.
* **Apres 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. Nous pouvons aussi faire des graphs comme la matrice de confusion par exemple.
* **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 travail a priori. Pour le modèle il y a deux parties le Feed-forward et le Back-propagation et L’algèbre linéaire vont intervenir dans ces deux parties.

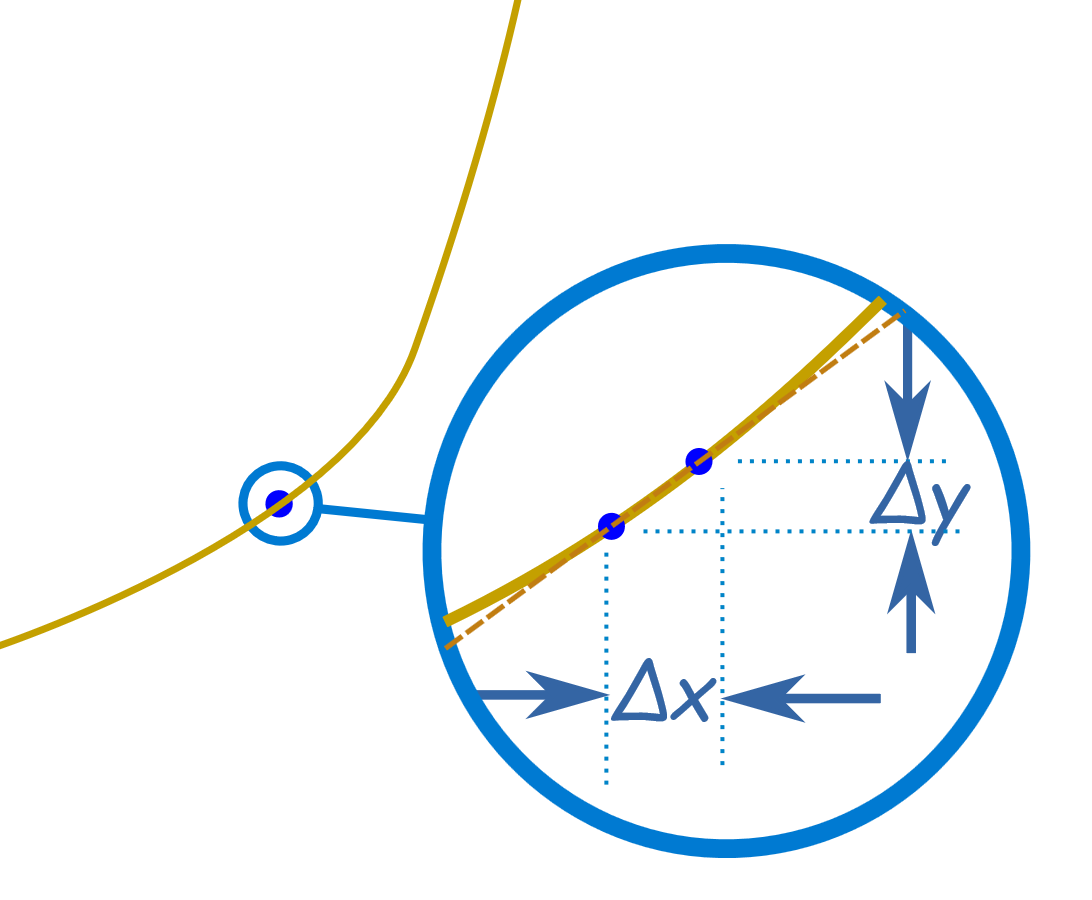
L’algèbre linéaire est la branche des mathématiques qui s'intéresse à l'étude des espaces vectoriels (ou espaces linéaires), de leurs éléments les vecteurs, des transformations linéaires et des systèmes d'équations linéaires (théorie des matrices). (Algèbre linéaire - Définition, 2024)

Ainsi, la plus grande utilité de l’algèbre linéaire est le calcul de poids, elle va nous permettre d’automatiser les calculs lourds et couteux, de ce fait nous permettant 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’inputs, plusieurs couches cachées qui peuvent avoir elle-même des milliers de neurones. C’est juste impossible de calculer tout ceci d’une manière séquentielle.

* **L’analyse**

Quand on parle d’apprentissage en IA ou le terme très populaire anglais « Learning », c’est à cause du domaine des mathématiques l’analyse avec le calcul des dérivés. Si l’IA est devenue ce qu’elle est devenue aujourd’hui avec les performances qu’on la connait, c’est en très grande partie à cause des calculs d’analyse mathématique.

La dérivée d’une fonction nous informe sur la variation de la fonction en un point. Pour être plus claire, elle nous permet de calculer la pente sur n’importe quel point de la fonction. Le calcule de dérivée est très important dans beaucoup de domaines notamment dans de Deep Learning.



Maintenant comment se passe l’apprentissage dans un réseau de neurones ? Nous allons rectifier les erreurs commises par l’IA dans son entrainement à travers la dérivée. Si nous répétons cela autant de fois que nécessaire, l’erreur d’assomption va se réduire au minimum et l’ « accuracy » va se maximiser. Ce qu’il faut comprendre par-là, c’est que le calcul de la dérivée de la fonction d’erreur va nous permettre de rectifier cette erreur.

De manière pratique, on calcule l’erreur en premier lieu, puis on calcule la dérivée de la fonction d’erreur. La manière dont la rectification va se faire, c’est qu’on va donner à chacun des poids une valeur correspondante à sa responsabilité dans l’erreur et c’est ça le Back-propagation ou la rétropropagation en français.

##### 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, ils parlent plutôt de de Computer Science qui se traduit littéralement par science de l’ordinateur. Plus haut, nous avions attesté que l’IA est une science purement mathématique avec que des calculs que l’on pourrait même faire sur feuille. Dès lors, que représente l’informatique pour l’IA : il a le rôle d’une calculatrice géante capable de faire des super-calculs en un temps record et aussi faire office d’interface graphique.

* **L’algorithme**

Bien évidemment, la première des choses que nous allons voir c’est l’algorithme. Nous pouvons attester sans prendre beaucoup de risques que l’algorithme est l’informatique, et que l’informatique est l’algorithme. L’algorithme est l’ensemble des étapes auxquelles il va falloir passer pour résoudre un problème informatique. Il est souvent fait l’analogie de la recette de cuisine pour illustrer l’algorithme et à juste titre.

La raison pour laquelle il faut maitriser l’algorithme c’est que : pour implémenter un problème mathématique dans un ordinateur, il faut savoir comment s’y pendre et savoir quelles étapes à suivre, sinon 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 = b2 – 4 \* a \* c)
* Si delta positif alors x1 = (-b – racine(delta) / 2 \* a) et x2 = (-b + racine(delta) / 2 \* a)
* Si delta nul alors x = racine(delta) / 2 \* a
* Si delta négatif alors il n’y a pas de solution dans R.

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

* **Les structure de données**

D’abord, les structures de données sont le terme utilisé pour représenter toutes les différentes façons en informatique pour modéliser les données avec lesquelles nous travaillons. Très souvent, pour ne pas dire tout le temps, nous n’avions pas directement la façon optimale de gestion de données.

Ces structures peuvent partir d’un simple tableau dans un langage de programmation jusqu’à atteindre les graphs (structure de données complexe et très puissante). Comme nous l’avons dit et redit l’IA travaille sur des données. Citons quelques exemples de structures de données : les listes chainées, les tables de hachages, les arbres, les piles et files, les graphs etc.

* **Les langages de programmation**

Les langages de programmation aussi appelés langages informatiques sont les syntaxes qui traduisent les algorithmes d’une manière compréhensible à l’ordinateur. Il faut préciser que l’ordinateur ne comprend pas le texte, il comprend seulement les chiffres (nombre binaire en l’occurrence). 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 va être évidente pour tout le monde, nous pouvons citer : C/C++ (important pour l’IA), Python (important pour l’IA), Java, PHP, JavaScript …

#### Les algorithmes d’intelligence artificielle

Nous y voilà, l’une des parties les plus importantes de ce travail de mémoire. Nous avons parlé de l’IA dans ce document mais cette fois nous allons voir comment elle fonctionne en parcourant différant 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.

##### Machine Learning

Littéralement, Machine Learning veut dire apprentissage des machines. Comme nous êtres humains, nous naissons 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), Unsupervised Machine Learning (UML).

* **Supervised learning**

Si nous reprenons l’analogie de l’enfant, dans sa phase d’apprentissage ses parents vont être derrière et le guider. Si l’enfant commet des erreurs ses parents vont de 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 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 repaire 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, élaboré 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 propogation (essaie au hasard)

D’abord la courbe que l’on veut tracer va être de la forme , mais dans le jargon on va parler w0 et w1 qui vont représenter les poids respectifs, l’équation devient . Le but du jeu est de trouver les w0 et w1 qui vont au mieux correspondre à nos points. Dans un premier temps on va les 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 : Mean Square Error (la moyenne des erreurs au carré)

Y : la sortie attendue

Ŷ : la sortie observée

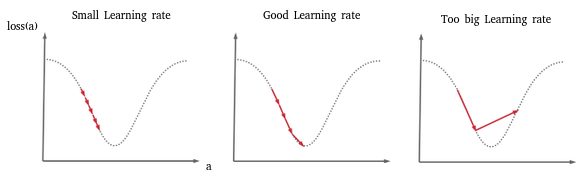
N : le nombre d’élément dans le tableau

Cependant, les plus curieux vont se demander pourquoi élever l’erreur au carré. C’est une bonne question. La raison est simple car une erreur de -1 est égale à une erreur qui vaut 1. Et le fait de l’élever au carrée 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 w0 et w1. Cela veut dire que chacun va prendre une part de l’erreur qui est égale à sa responsabilité de cette dernière et se rectifier lui-même. Pour se faire nous allons calculer la dérivée de toutes les fonctions qui nous ont mené à cette erreur de manière suivante :

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* ».



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 :

Donc

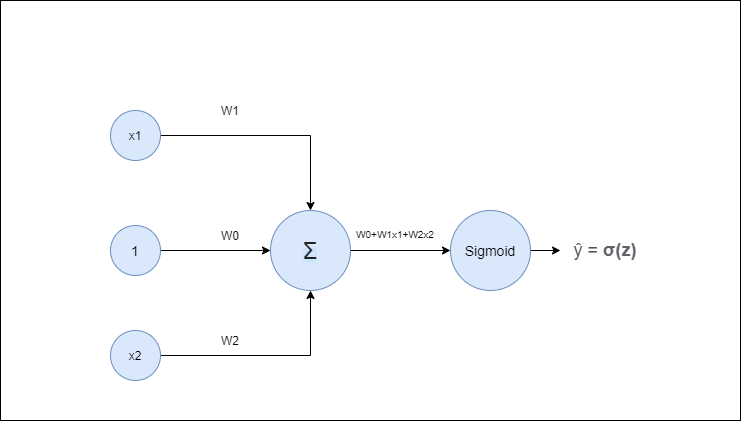
Équation 1 Le chainage des dérives partielles

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 s’en rendre compte. Par exemple, détecter si un email est un spam ou non, si une information est un fake new ou non, si un investissement va être rentable ou pas…

* Forward propogation

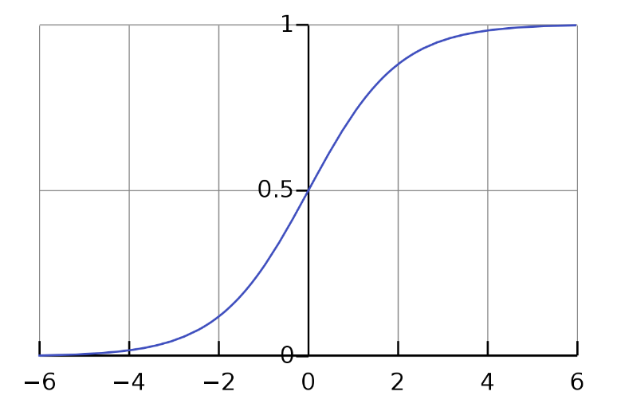


Voici à quoi va ressembler notre réseau de neurones, on va ajouter un autre input en plus x1 et x2, c’est le biais qui va toujours être égale à 1, son utilité est d’éviter que certains neurones ne meurent durant l’entrainement si x1 = 0 et x2 = 0.

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) |  | 0, 1 |
| Sigmoid (x) |  | Tout réel compris en 0 et 1 |

Nous allons continuer avec la fonction sigmoid :



* Calculer l’erreur

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

* Backpropagation

Nous voici près pour la rétropropagation, seulement ici nous allons mettre à jour trois poids à savoir w0, w1, w2.

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 dérivée partielle.

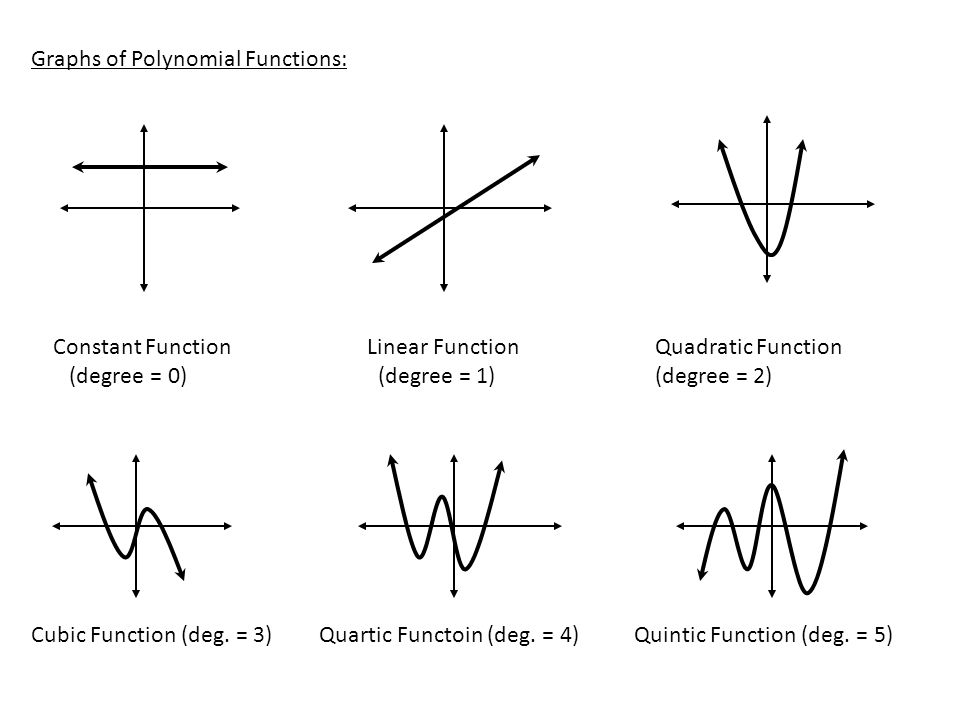
Voir annexe 2 pour un exemple de calcul de régression logistique.

* + - **La régression polynomiale**

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

* Forward-propogation

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 de degré de la fonction. Plus le degré est élevé plus la fonction pourra être en mesure d’aller chercher des variations.



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 :

* Calculer l’erreur

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

* Backpropagation

La valeur de la dérivée partielle pour w1 se présente comme suit :

Nous avons décidé de prendre w2 car il a la dérivée partielle la plus compliqué à calculer, avec ce calcul établi, nous pouvons passer à l’étape des mises à jour des poids, ne pas oublier de prendre un Learning Rate.

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

* + **La classification**

La classification est un problème qui est là 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 ranger chacune dans sa classe de prédilection et aussi mais surtout prédire les classes pour des données non encore observées. Différents algorithmes sont aujourd’hui là pour nous permettre de régler les problèmes de classification mais nous allons en voir trois (3).

* + - **Naive Bayes**

Le modèle de Naïve Bayes (NB) est un algorithme de ML qui nous vient des statistiques et 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).

* La probabilité conditionnelle des valeurs d’attribut

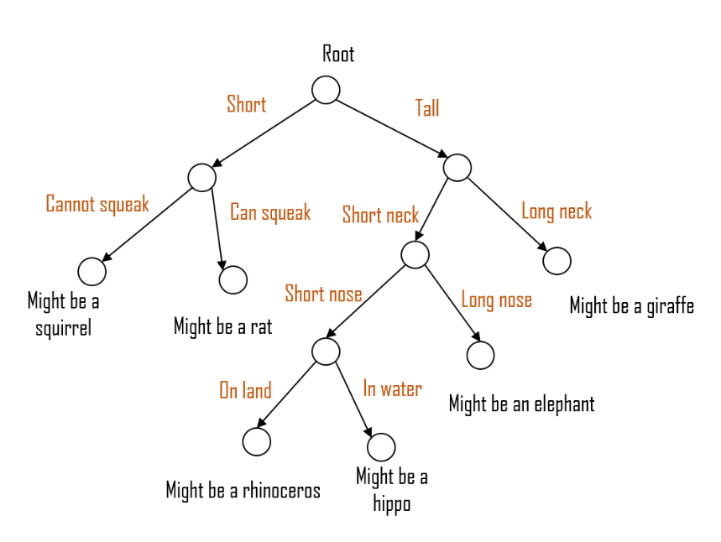
Pour chaque valeur d’attribut, il nous faut calculer sa probabilité conditionnelle par rapport aux valeurs cibles.

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.

Pour normaliser les probabilités :

* + - **L’arbre de decision**

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 elle prend des décisions en se basant sur l’attribut des données. D’abord l’arbre vérifie l’attribut le plus indicatif et prends la direction d’une de ses valeurs, puis le deuxième attribut le plus significatif et prends la direction d’une de ses valeurs, ainsi de suite jusqu’à classer un nouvel enregistrement.



En revanche, le fait de distinguer un attribut significatif ne se fait pas arbitrairement, sinon ce ne serait pas une 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, conséquence l’entropie est maximale (égale ou proche de 1), si une seule classe est représentée le nœud est pure et l’entropie est minimale (égale ou porche de 0).

* Gain d’information (GI)

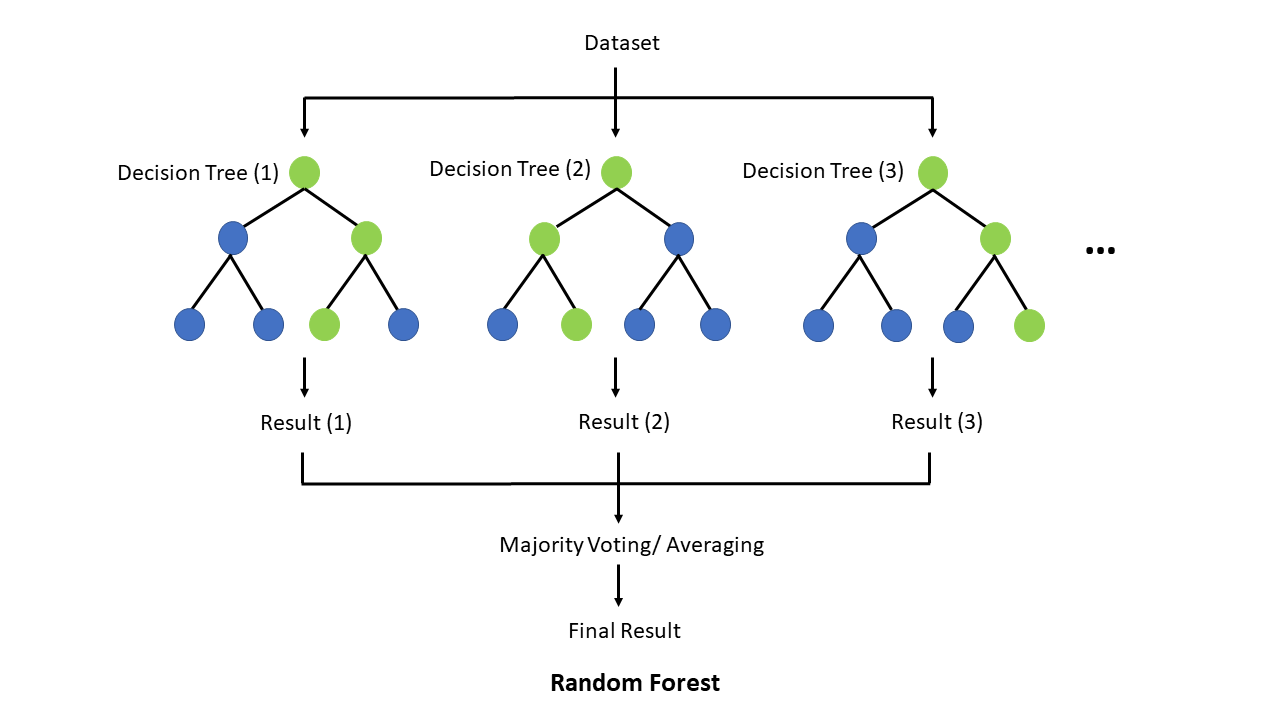
La première des choses à faire c’est de calculer le gain d’information c’est-à-dire de tous les attributs, quel est celui qui nous renseigne le plus si l’individu est sénégalais ou pas.

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

* + - **Random forest**

Nous venons juste de parler du desicion tree, donc nous avons déjà les base pour comprendre ce que c’est le random forest sans les calculs. Ce terme peut être traduit en français comme foret aléatoire. Pour mieux comprendre, considérons un dataset, pour entrainer le modèle, nous allons utiliser plusieurs modèles de decison tree, disons cinq (5) modèles. A partie de la, nous allons diviser notre dataset en cinq (5) parties différentes. A ce moment nous avons 5 dataset et 5 modèles de DT, ce que nous allons faire c’est d’entrainer les 5 dataset avec les 5 modeles de DT.

Nous venons juste de créer de manière simple un random forest, random (aleatoire) car le dataset initiale se voit être diviser 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 par un souci. Il va être procéder à un vote et la majorité va l’emporter. Supposons que les trois votent pour une classe et les deux pour l’autre, dans ce cas nous considère la classe sortie par le 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.*



* **Unsupervised learning**

Pour ce qui est de l’apprentissage non supervisé, c’est qu’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 cluster les individus qui se ressemble le plus en utilisant plusieurs variables qui décrivent les données.

* + **Le clustering**

Le clustering est une méthode d’apprentissage non supervisé dans lequel 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 y a plusieurs algorithmes de clustering mais nous allons voir le fameux k-means (le k de k-means représente le nombre de classe ou cluster).

* Définir le nombre de cluster

En premier lieu, il faut définir le nombre de cluster, ce choix peut relever du libre arbitre de l’ingénieur ou peut-être défini en fonction de méthodes.

* Le centre de gravité

Pour chaque cluster il faut calculer son centre de gravité et on affecte 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 où ils sont, il suffit de les mettre à jour. Cette étape va être répéter autant de fois que nécessaire pour avoir des clusters les plus représentatifs des données que possible.

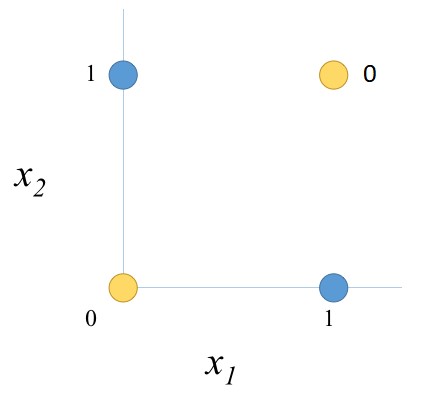
Voir annexe 4 pour un exemple de calcul de clustering.

* + **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. Ses règles permettent de répondre à des questions comme : dans quelle mesure B et C vont apparaitre sachant que A est apparu ? Ces calculs vont se faire avec un ensemble de sous-ensemble, nous allons parler ici de itemset pour désigner les sous ensemble. Les règles d’associations sont très fréquentes dans les marchés et supermarché pour desceller les produits qui sont souvent acheter en un temps par les clients. Une fois que nous avons des règles intéressants les dirigeants peuvent prendre de bonne décision.

##### Deep Learning

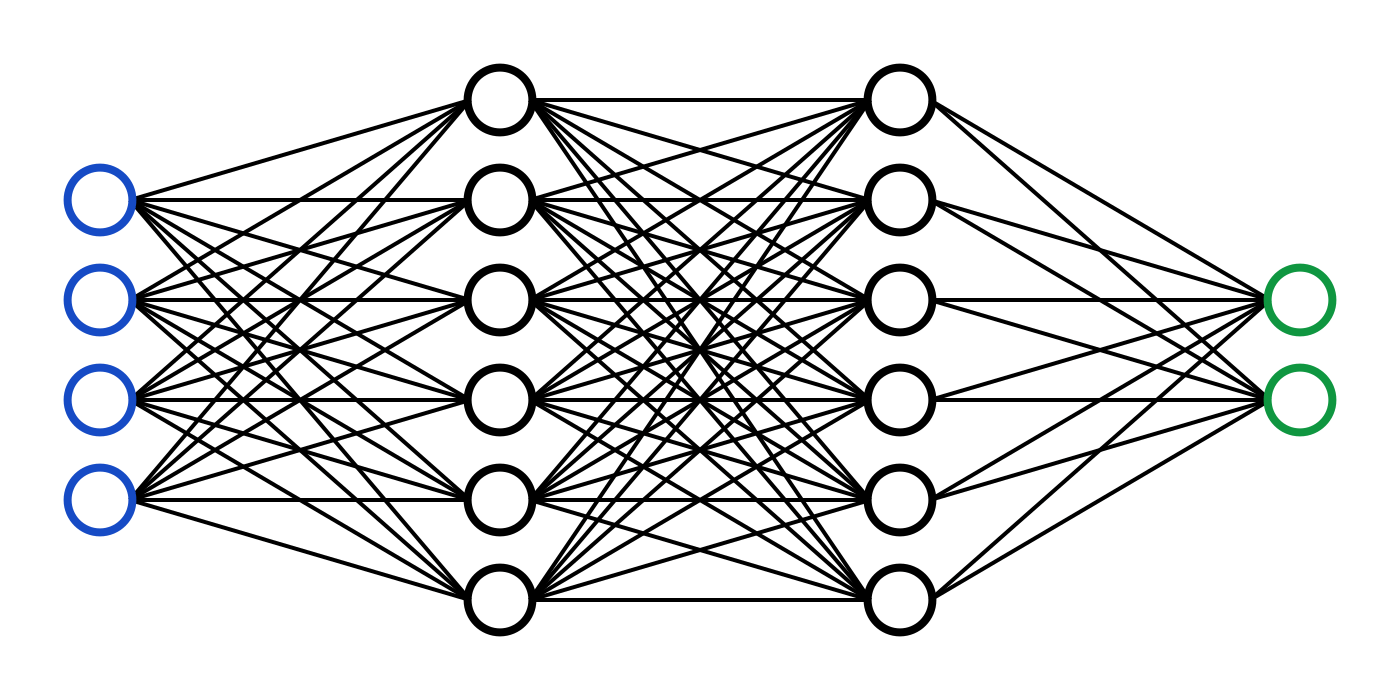
Frank Rosenblatt a créé le perceptron qui nous a permis de résoudre des problèmes, notamment le OU et le ET logique, mais quand ils l’ont essaiés pour le XOR, ils se sont rendu compte que le perceptron ne convergeait pas. Le problème était simple, un perceptron traçait des séparateurs linéaires, or ce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

1. 

Allez-y ! essayer de tracer une seule droite qui est capable de séparer les 0 et le 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 le longtemps, jusqu’à qu’ils découvrent les solutions qui vont être présentées ici.

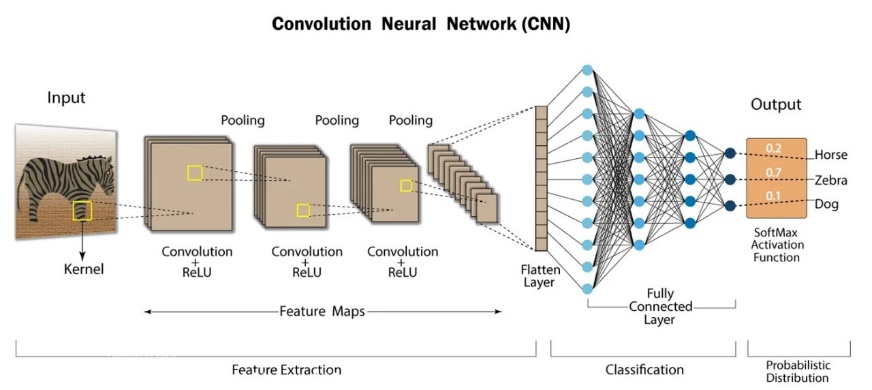
* **Artificial neuron network (ANN)**

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, c’est 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 couche cachée, plus c’est profond : apprentissage profond : Deep Learning. D’ailleurs nous avons très probablement tous déjà vu la représentation d’un réseau de neurones profonds.



* **Convolutional neuron network (CNN)**

S’il y a un domaine où l’humain a toujours dépassé la machine, c’est la vision reconnaitre des choses, des éléments de la nature et de les classer. Mais depuis quelque temps les scientifiques ont réalisé d’énormes avancés sur 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.

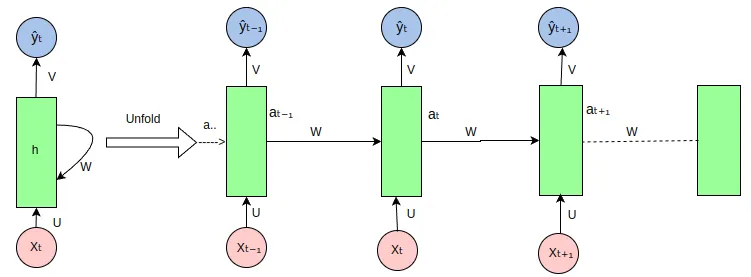


* **Recurrent neuron network (RNN)**

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 au ANN, c’est ils n’ont pas de mémoire. Imaginons un jeu de données avec 60 000 inputs, de la première ligne du premier epochs jusqu’à la dernière ligne du dernière epochs, le modèle va oublier tout ce qui s’est passé et se concentre seulement sur les caractéristiques principales. Mais très souvent, il est nécessaire de savoir ce qui s’était 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.



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

## Chapitre 2 : Revue des travaux de recherche de l’intelligence artificielle appliquée à la finance

Ce chapitre va constituer la revue de la littérature scientifique de travaux de l’IA appliquée à la finance. La finance fait des sciences qui font le plus l’application de l’IA car la finance travail sur des chiffres et l’IA aime les chiffres, beaucoup d’entreprise 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 y voir certains.

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

Pour ce qu’il s’agit ce cette section, la chose qui sera faite c’est une présentation des domaines financiers et de gestion que l’IA est train de révolutionner. Comme déjà dit la financer fait partie des plus belles applications de l’AI et cette dernière peut aider à parfaire la finance a bien des égards. De l’analyse prédictive a l’analyse des tendance boursières, nous allons voir plusieurs endroits d’application de la finance par l’IA.

#### Analyse prédictive

L’analyse prédictive regroupe l’ensemble des méthodes que les financiers peuvent utiliser pour avoir 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’applique dans plusieurs domaines de la finance.

Dans ce passage nous allons avoir une vue d’ensemble des analyses prédictive à base d’IA pour ensemble traiter cas par cas. C’est optique que les scientifiques 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 présenter une revue des méthodes basées sur la littérature scientifique en focalisant sur les domaines d’application de l’analyse prédictive. Pour y parvenir il a fait une étude comparative de diffèrent modelés tels que la classification, la régression, le clustering, les règles d’associations, et les modèles de *times series*. Les résultats étaient les suivants : pour les prédictions sur l’économie les *tims series* sont les plus performants, pour les prédictions sur des gains potentiels il y a plusieurs qui offrent une bonne personne performent notamment le *naives bayes*, les ANN, les modèles non linéaires et enfin pour l’optimisation de portfolio ce sont les modèles de ML. Cependant sont travail de 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 intéressant pout une analyse prédictive. Néanmoins le plus important d’être a l’écoute des données dont nous disposons et de choisir le modèle le plus adapté.

#### 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).

#### Services clients

Est-ce que les IA vont remplacer les humains dans les services des entreprises ? Voilà l’un des questions qui se posent le plus dans le domaine. Les services clients ou services après-vente sont très couteux pour les organisations et demande 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 d 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 ses 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 humaines.

Dans le même cadre d’autres études ont été faite sur le domaine comme celle de 2022 de Dimitrios Buhalis et Iuliia Moldavska dans le Journal of Hospitality and Tourism Technology. Cet article se donner comme objectif d’investiguer dans les interactions entre les hôtels et ôtes dans le contexte de « l’hospitalité ». La méthodologie que cette étude a suivie de faire des entretiens avec le personnel des hôtels sur l’utilisation des assistantes vocales dans ses structure. Ce qu’ils ont trouvé c’est l’utilisation de ses technologies sont d’un grand apport autant pour eux personnel que pour les clients. Selon les auteurs cette peut les permettre d’explorer domaines du "smart hospitality" et de l’écosystème du tourisme avec l’IA.

#### 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 fraude, si elles ne sont pas identifiées et régler 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, Muzamil Abbas ont très récemment publié un article sur le sujet "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 ensuite ils ont parlé de l’IA dans la finance. Les résultats de cette étude ont été conséquents car ils ont pu prouver l’apport considérable de l’IA sur ce domaine sans rien implémenter.

Un peu plut 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 implémentant des modèles qui vont fonctionner pour la détection. Pour ce qui est des modèles ils ont utilisé dans un premier temps des algorithmes de Machine Learning, ensuite de Deep Learning. Les observations faites à la suite des expérimentations sont claires : les modèles de ML ont été plus performent que les modelé DL, comme quoi des fois le simple est plus efficace.

#### La bourse et les marchés financiers

La bourse et les marchés financiers non plus n’est échappent pas à l’IA, c’est l’une des applications les plus challengeant de l’IA sur la finance du fait de la volatilité des marchés financiers et de l’incertitude qui y règne. Mais tel chose n’a 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 été en collaboration dans rédaction d’un article qui se voulait être une revue des techniques de ML dans prédiction du cours boursier. Cet article avait pour but principale était de faire une investigation sur la littérature mais aussi les avoir connaissance des types de variable utilisées pour cette tâche. Une fois ceci fait, il va être fait l’application de ses variables sur le ML afin de prédire les cours boursiers des marchés financiers. Les données qu’ils ont exploitées proviennent d’article de presse de 2000 à 2019. Ils ont adaptée la méthodologie suivante : d’abord la collecte de données boursiers sur plus marchés financiers du monde entier avec les *times series*, ensuite le traitement, enfin l’étude comparative de modèles. Le résultat le plus frappant était les ANN ont été très performent par rapport aux autres surtout les marchés comme S&P 500 (86.81%) et DJIA (88.98%).

Il y a encore beaucoup d’études qui se font sur la bourse car les investisseurs en demandent de plus en plus donc entendons nous a des modèles de plus en plus performants dans l’avenir sur les marchés boursiers.

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

Ici nous allons entrer dans le vif de notre sujet, il y a un certain nombre de travaux des scientifiques sur l’analyse des états financiers. Il y a trois principaux états financiers que nous allons voir la partie 2 (bilan, compte de résultat et le tableau des flux de trésorerie). La principale difficulté sera la modélisation des états financier en un format compréhensible par les modelés que nous avons déjà vu. Ensuite, va s’opérer la magie de l’IA et allons comment les scientifiques procèdent.

#### Les travaux de l’intelligence artificielle sur l’analyse financière

Dans notre travail, nous nous sommes donnés comme objectifs de faire appliquer l’IA dans l’analyse financière des entreprises ouest-africaines. Nous allons voir que d’autres choses ont été faites ailleurs qui peuvent ressembler à ce que nous allons faire ici, nous allons les voir.

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

D’autres recherches ont aussi été faites sur le domaine comme cette étude de 2000 de Ning Yang sur l’IA et de 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 grandes tailles. Il a suivi la démarche suivante : d’abord expliquer le fonctionnement des modèles puis de les appliquer sur des données empiriques. Il a réussi à obtenir des résultats satisfaisants puisque l’application parvenait à répondre dans des taches importantes dans la finance. Avec le développement rapide des données financières, l’application de ces modèles vont être plus que nécessaire.

#### Limites des travaux actuels

Dans tous les travaux de recherche, il y a toujours des limites, des défauts ou des points d’améliorations. 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’endroit où il y a lieu d’amélioration pour parfaire la science et nous allons les voir.

* **Qualités des données financières**

Quand on parle de qualité des données, on peut faire référence sur plusieurs aspects des données financières. Pour les données financières nous faisons souvent face à des bruits et aussi des valeurs incomplètes. La conséquence de cette chose c’est qu’il falloir faire un grand travail de preprocessing. En outre, les données historiques peuvent être biaisées, la finance change et les méthodes aussi changent, les politiques de finance aussi. Ce qui peut constituer un problème dans l’entrainement. Pour illustrer cela, en 2018 dans la zone UEMOA une nouvelle réglementation a vu le jour, en imposant à toutes les entreprises de la BRVM de présenter leurs états financiers avec les normes IFRS (International Financial Reporting Standards). Mais cela c’est pour les grandes entreprises, pour les petites entreprises c’est encore pire puisque qu’il est très difficile d’accéder à ses données, c’est l’une des raisons pour lesquelles les petites et moyennes entreprises ne bénéficient pas de ses technologies pour le moment. En fin nous pouvons parler du contexte, des données peuvent changer selon contexte géopolitique, économique, sociale etc. ce qui peut biaiser les modèles en prenant en compte des périodes avec des mesures exceptionnelles (la crise des subprimes en 2008 ou le 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 sur plusieurs angles. Nous allons parler des manières de développement des modèles, le comportement de modèles vont être toujours basé sur les données d’entrainement. Ce qui veut dire que les développeurs peuvent intentionnellement programmés l’IA à se comporter d’une manière ou d’une autre. Mais aussi l’IA ne va faire seulement la chose pour laquelle elle a été développée, à moins de mise à jour, elle ne pourra pas contextualiser pour prendre la meilleure décision. De plus, il y a un autre problème qui subsiste, une IA qui a été programmée d’une manière, donc elle va se comporter et fonctionner de cette même manière. Pour être plus claire, il y a un risque que l’IA donne le même conseil, si elle le fait et que tout le monde (ou la majorité des gens) suit ses conseils, les gens peuvent prendre les mêmes décisions, ce qui n’est pas bien. 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, les délits d’initié, IA pourra-t-elle faire preuve de discernement et savoir quelle information donner à qui.

On peut répondre à tous ces questionnements en disant que l’IA est un programme, il est possible de programmer cela dans son développement (chose qui ne va pas être facile). Par exemple il y a des LLM qui ne vont pas répondre sus certains sujets qui fâchent 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-africains. Quand nous savons comment certaines entreprises africaines fonctionnent avec toutes leurs difficultés, l’IA leurs sera d’une très grande utilité.

C’est d’ailleurs l’une des grandes raisons pour lesquelles nous nous sommes lancés dans ce domaines en plus de notre grand intérêt pour l’IA. Il n’est pas seulement intéressant de constater les dégâts 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 révèle des parties importantes de l’intelligence artificielle que sont les généralités et théories mais aussi les travaux de recherche scientifiques. Nous avons tous convenir que sans ses connaissances durant cette partie il serait difficile de s’aventurer dans le développement pratiques des modèles que ce soit pour la prédiction ou le Chatbot.

Comme constate dans cette partie beaucoup ont déjà été faites et nous allons réinventer la roue mais nous allons la façonner à notre guise pour qu’elle soir en mesure de rouler de la manière qui nous convient. Tous les choses vus dans ce chapitre nous ont permis d’avoir le courage d’attaquer la suite.

Pour ce qu’il s’agit cette suite, nous allons comme les anglophones diraient *"get our hand dirty wih some code"* ce qui veut se salir les mains avec du code. La partie suivante va être plus attrayant et intéressant car nous passer aux pratiques après toute cette théorie. Et comme nous l’avons ici nous allons entrer dans les moindres détails pour la premier ne soit vaine.

# Partie 2 : Conception et développement des outils d’IA appliquée à l’analyse des états financiers

## Introduction de partie

Dans la partie précédant, nous avons faire connaissance avec l’IA, nous avons vu les généralités, les théories, les travaux scientifiques. Apres cela, tous les outils nécessaires sont dans nos trousses afin d’implémenter des modèles pratiques d’IA. L’objectif de cette partie est simple utiliser toutes les connaissances engendrer durant la partie précédente pour développer les modèles qui permettre d’augmenter l’efficacité des analystes dans leur travail.

Cette partie aussi va se voir être subdiviser en chapitre : le premier qui sera consacrer au développement des modes prédictifs, ici il sera fait d’abord fait une présentation de la démarque d’une analyse financière, ensuite le développement des modèles prédictifs. Il en va de soit qu’il faut des domaines d’applications avant de s’aventurer aux implémentations d’algorithmes. Apres cela, nous allons attaquer le chapitre sur le développement du Chatbot où la collecte de données et le développement en tant que tels vont être les deux sections qui vont nous intéresser.

A la fin de cette partie nous aurons deux modèles fonctionnels que les utilisateurs vont pouvoir s’en servir pour faciliter leur travail. Et bien entendu il y aura aussi une interface graphique de communication qui permettra à n’importer interagir avec les modèles.

## Chapitre 1 : 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 d’avant d’en faire une application ou des modèles. Nous allons aussi présenter les outils utiliser pour le développement de l’interface et comment cette dernière va être faires.

### Section 1 : Mise en œuvre d’une application d’analyse financière

Nous voilà dans la section du document ou allons parler des développements des modèles. Elle se diviser en deux bloque : l’analyse financière traditionnelle (ou nous allons voir les analyses financières font d’habitude leur travail) et l’amélioration grâce à l’intégration des outils intelligence (une interface graphique avec la possibilité de faire des prédiction).

#### Démarche d’une analyse financière

D’abord, nous allons précéder 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** | - | - |
| Charges immobilisées | - | - |
| Immobilisations incorporelles | - | - |
| Immobilisations corporelles brutes | - | - |
| Immobilisations corporelles brutes Immobilisations financières | - | - |
| Amortissements et provisions | - | - |
| **TOTAL ACTIF IMMOBILISE** | - | - |
| Stock | - | - |
| Fournisseurs, avances versées | - | - |
| Clients | - | - |
| Autres créances | - | - |
| **TOTAL ACTIF CIRCULANT** | - | - |
| **TOTAL TRESORERIE ACTIF** | - | - |
| **TOTAL ACTIF** | - | - |
| **PASSIF** |  |  |
| Capital | - | - |
| Primes et réserves | - | - |
| Report à nouveau | - | - |
| Résultat net | - | - |
| **TOTAL CAPITAUX PROPRES** | - | - |
| Emprunts et dettes financières | - | - |
| Provisions financières | - | - |
| **TOTAL DETTES FINANCIERES** | - | - |
| Dettes circulants | - | - |
| Clients, avances reçues | - | - |
| Fournisseurs d’exploitation | - | - |
| Dettes fiscales | - | - |
| Dettes sociales | - | - |
| Autres dettes | - | - |
| **TOTAL PASSIF CIRCULANT** | - | - |
| **TOTAL TREORERIE PASSIF** | - | - |
| **TOTAL PASSIF** | - | - |

* **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 le composent, les produits et les éléments qui le composent et éventuellement les soldes intermédiaires de gestion.

|  |  |  |
| --- | --- | --- |
| COMPTE DE RESULTAT | | |
|  | N | N-1 |
| Ventes de marchandises | - | - |
| Production vendue | - | - |
| Travaux, services vendus | - | - |
| Produits accessoires | - | - |
| **CHIFFRE D'AFFAIRES** | - | - |
| Production stockée | - | - |
| Autres produits | - | - |
| **TOTAL PRODUITS** | - | - |
| Achats et frais sur achats | - | - |
| Variation de stock | - | - |
| Transports | - | - |
| Services extérieurs | - | - |
| Impôts et taxes | - | - |
| Charges et pertes diverses | - | - |
| **TOTAL CHARGES** | - | - |
| **VALEUR AJOUTEE** | - | - |
| Frais de personnel | - | - |
| **EXCEDENT BRUTE D’EXPLOITATION** | - | - |
| Transfert de charges | - | - |
| Dotations aux amortissements | - | - |
| Dotations aux provisions | - | - |
| **RESULTAT D'EXPLOITATION** | - | - |
| Produits financiers | - | - |
| Charges financières | - | - |
| **RESULTAT FINANCIER** | - | - |
| Produits HAO | - | - |
| Charges HAO | - | - |
| **RESULTAT EXCEPTIONNEL** | - | - |
| Impôt sur la société | - | - |
| **RESULTAT NET** | - | - |

* **Le tableau des flux de trésorerie**

Selon le site [L’expert-comptable](https://www.l-expert-comptable.com/a/6312-comment-elaborer-un-tableau-de-flux-de-tresorerie-interets-et-analyse.html#:~:text=flux%20de%20tr%C3%A9sorerie%20%3F-,D%C3%A9finition,ann%C3%A9e%20etc...).), 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** | - | - |
| Capacite d’autofinancement | - | - |
| Variation actif circulant HAO | - | - |
| Variation des stocks | - | - |
| Variation des créances | - | - |
| Variation du passif circulant | - | - |
| Variation du BFR | - | - |
| **FLUX DE TRESORERIE DES ACTIVITES OPERATIONNELLES** | - | - |
| Décaissements lies aux acquisitions d’immobilisations incorporelles | - | - |
| Décaissements lies aux acquisitions d’immobilisations corporelles | - | - |
| Décaissements lies aux acquisitions d’immobilisations financières | - | - |
| Encaissements lies aux acquisitions d’immobilisations incorporelles et corporelles | - | - |
| Encaissements lies 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 verses | - | - |
| **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** | - | - |

* **Les étapes d’une analyse financière**

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

* + **Vérification des états financiers**

Avant de pouvoir commencer l’analyse financière en question, il va falloir faire des vérifications pour ne pas avoir une analyse biaiser 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 au compte de résultat est égal au résultat du bilan

La trésorerie nette : il va falloir aussi vérifier la différence entre la trésorerie actif et la trésorerie passif soit bien égale à la trésorerie nette au 31 décembre de l’année N.

Une fois ses vérifications faites et que tout est OK, on peut faire l’analyse financière proprement dite. Il faut rajouter qu’il y a d’autres types de vérification que l’on peut faire mais ses dernières 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 a une valeur pivote. C’est-à-dire la part de chaque élément d’un état financier par rapport a 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 notre tableau pour l’analyse verticale.

Analyse horizontale

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

* + **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. C’est toutes les valeurs que l’entreprise doit maximiser pour rester en bonne santé et durée dans le temps. Il y aura trois (3) d’analyse qui va être faite dans cette partie.

Analyse du cycle de vie de l'activité

Analyse du comportement des flux de trésorerie

Analyse des équilibres financiers et la relation de trésorerie

* + **Analyse tendancielle et la méthode des ratios**

Tout d’abord, l’analyse financière avec la méthode des ratios nous permet de faire le diagnostic de l’entreprise à travers un certain nombre de ratios. Ces ratios sont de tout type et vont faire intervenir les trois (3) états financiers. Ici, nous sommes à un point très important de l’analyse financière c’est pour cela que cette partie va particulièrement intéressée les investisseurs en plus de l’entreprise elle-même.

Ces calculs de ratios vont nous permettre de noter à la fin l’entreprise de voir si elle en risque de défaut ou risque de faillite. Nous allons utiliser un système de notation développé par Altman en 2005. Finalement, l’évaluation de l’entreprise va mettre fin à cette analyse par les ratios. Ci-après un tableau récapitulatif des familles et 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’entêtement, maturité de l’endettement |
| Flux de trésorerie | Capacite 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. |

#### Le frontend

Qt nous donne un certain nombre de fonctionnalités pour réaliser des logiciels robustes et très avancés mais nous allons en prendre quelques-uns, qui nous ont été utiles dans notre travail, pour les expliquer et faire comprendre comment nous avons utilisé dans nos modèles.

* Widget

Les Widgets font partie des éléments de base sur Qt, au fait tout est un widget et il y a des Widgets pour tout faire. Pour mieux comprendre ce concept, un Widget est un élément visible que l’utilisateur peut interagir avec. Nous pouvons donner l’exemple d’un bouton, une zone de texte, un label ou voir même une page entière.

L’un des Widget les plus importante c’est ce que Qt appelle QMainWidget qui est le Widget principale sur lequel tous des autres vont s’accoler. Il faut préciser chaque Widget a un parent sauf le Widget principale. Ce qui nous donne une structure sous forme d’arbre.

Dans notre logiciel nous allons utiliser toute sorte de Widget commençant bien évidemment par le QMainWidget en passant par les autres qui permettent de présenter le résultat de nos modèles d’une manière simple.

* Layout

Le deuxième élément sur lequel nous allons un peu nous attarder ce sont les Layout. Ils représentent tous les manières de disposer nos Widgets. Par défaut, un Widget va empiler tous ses enfants sur son point (0, 0) en haut en gauche dans un ordinateur.

Avec l’aide des Layouts, nous pouvons avoir plusieurs manières de présenter les Widgets, ceci aura pour effet d’avoir un logiciel plus sérieux. Il y a trois sortes de Layout sur Qt : QVBoxLayout (disposition verticale), QHBoxLayout (disposition horizontale) et QGridLayout (disposition en grille). Ses trois héritent tous de la classe mère QLayout.

Bien que nous ayons plus utilisé les dispositions verticales et horizontales dans notre logiciel, la disposition en grille peut être très pertinente par moments. C’est impressionnant de se dire quand même que tous les frontends que nous voyons dans les logiciels ne sortent pas de ces trois Layout.

* Signal and Slot

Les Signal and Slot sont les deux concepts de la programmation événementielle que Qt met à notre disposition. Un Signal c’est un signal c’est-à-dire un déclencheur, cela peut être une clique, un focus, un survol un changement quelconque. Il y a des Signal prédéfinis comme ceux que nous venons de citer, mais nous avons la possibilité d’en créer des nouveaux. Néanmoins, les prédéfinis vont nous suffire dans la plupart des cas.

Les Slot ce sont les fonctions que nous allons utiliser pour réagir par rapport au Signal. Mais pourquoi ce ne sont pas des méthodes comme les autres ? C’est que les Signal n’acceptent que les Slot comme fonction, une méthode qui ne portent pas la mention Slot ne va pas être acceptée dans la programmation évènementielle de Qt.

Nous allons illustrer cela en donnant l’exemple de l’utilisateur qui veut ouvrir les paramètres. S’il clique sur le bouton paramètre (Signal), l’application va ouvrir la boite de dialogue paramètre (Slot). Ce qui intéressant avec Qt, c’est qu’il nous permet de faire cela en une seule ligne de code avec la fonction « connect » de la classe QObject.

#### 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 qu’il 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 avoir une base de données locale qui va nous servir à stocker les valeurs des états financiers. Il y aura trois tables dans la base de données à savoir le bilan, le compte de résultat et aussi le tableau des flux de trésorerie et chaque enregistrement de chaque table va représenter une année. C’est une base de données simple, nous allons rarement y insérer des données ou même modifier. Dans la plupart des cas il sera seulement fait des requêtes SELECT pour obtenir des données.

* Le contrôleur

Même si nous ne sommes pas vraiment dans un modèle MVC (Model, View, Controller), car nous ne pouvons-nous permettre de laisser les requêtes directement accéder à notre base de données, cela nous permet aussi une bonne gestion des données. Ce modèle va se présenter sous forme de classe, ici aussi nous aurons trois (3) classes pour nos trois états financiers. Pour chaque classe, nous aurons les méthodes qui renvoient les états financiers, les analyses verticales, les getters et les setters pour chaque élément des états financiers.

* Serveur

Avant de pouvoir utiliser une base de données, il nous faut un SGBD (système de gestion de base de données) qui va nous permettre d’interagir avec la base, nous l’avons déjà dit c’est MySQL que nous allons utiliser avec XAMPP, qui est un serveur qui nous permet la connexion avec une base de données MySQL, Apache etc.

#### 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 Endpoint.

De là, on peut voir comment cela peut nous être utile, nous avons déjà développé nos modèles intelligents et ce qui suit c’est de les lier avec l’interface graphique. 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 server. Il est possible d’écrire un Application Programmation Interface (API) dans un langage A et de l’invoque dans un langage B avec l’aide d’un navigateur.

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

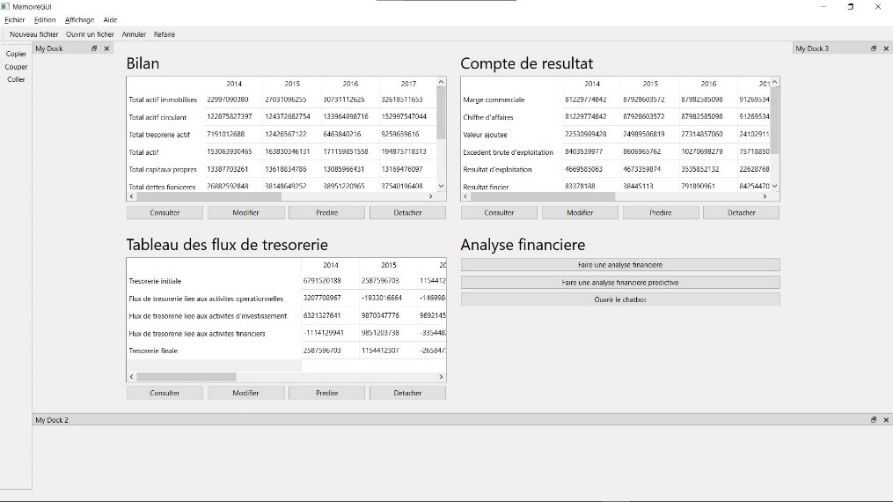
Nous pouvons donner l’exemple d’un cas d’utilisateur du chatbot. L’utilisateur pose sa question depuis l’application Qt, une requête REST va être invoquée avec la question comme paramètre, le web server va le prendre le traiter et renvoyer la réponse sous format JSON ou XML.

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

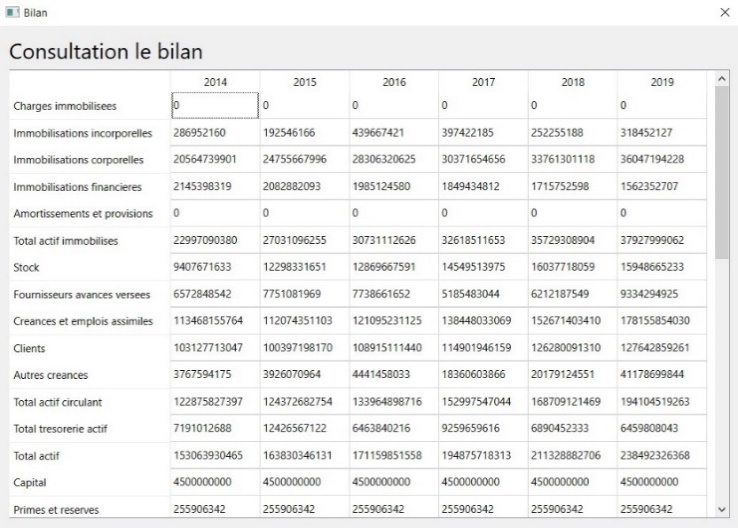
#### Présentation de l’application

Toute l’explication a été faites 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, celle que l’utilisateur va ouvrir très souvent.

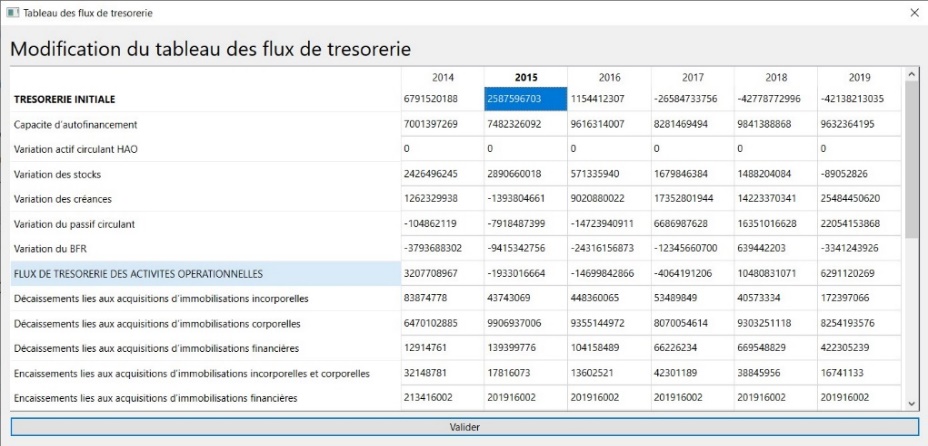
* La page d’accueil



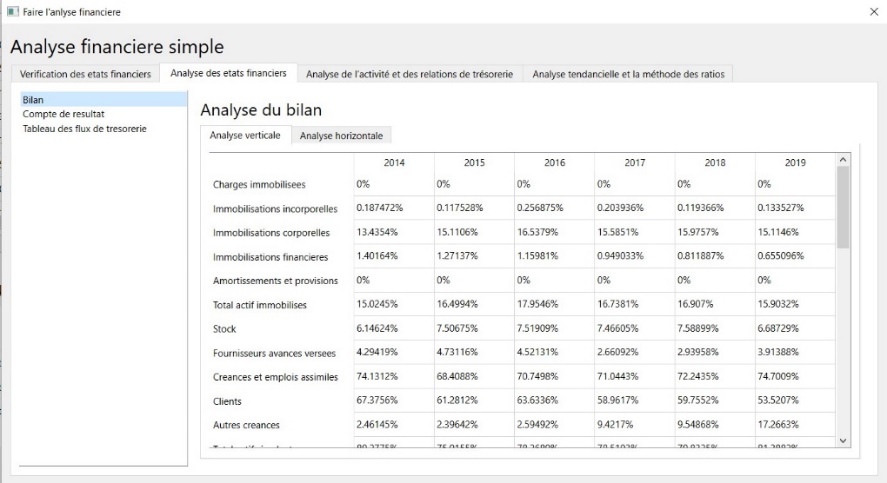
* La page de consultation du bilan



* La page de modification du tableau des flux de trésorerie



* La partie analyse des états financiers



* La partie analyse prédictive de la rentabilité sur les 3 prochaines années

### 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 financier prédictive. Mais avant tout tentative de développement de modèles, nous aurons besoins de données. Donc nous allons voir dans un premier temps comment va se faire la collecte, l’analyse et de traitement de données, ensuite va être procéder au développement des modèle notamment de régression.

#### 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 Economique et Monétaire Ouest-Africain (UEMOA). C’est ici que l’on peut échanger des actions et obligation pour le compte des entreprises et Etats de la zone UEMOA. La BRVM regroupe les 46 entreprises et banque les plus performantes de la zone.

Eu égard de cela, tout entreprise cotée à la BRVM a le devoir chaque année de publiée ces états financiers pour le compte des actionnaires, des Etats, des investisseurs, de potentiels investisseurs ou n’importe quelles personnes physiques ou morales. Leurs états financiers sont aussi publiés dans le [site de la BRVM](https://www.brvm.org/) 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 ses 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 autres 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.

#### La prédiction des valeurs

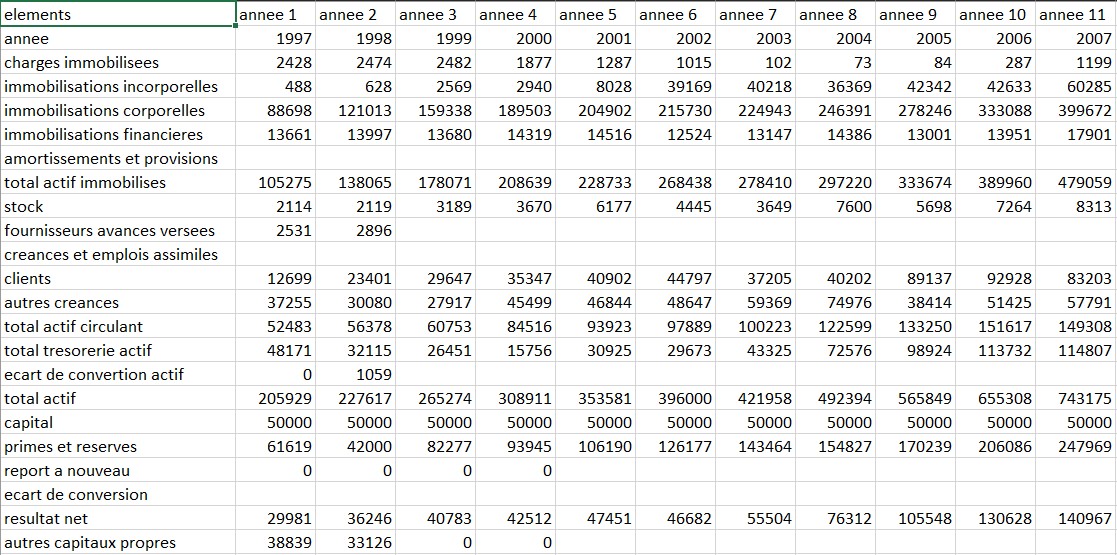
La prédiction est un vaste domaine qui ne relève pas seulement de l’IA. En effet beaucoup de domaines scientifiques essaient de faire des prédictions sur les valeurs avec lesquelles ils travaillent, c’est notamment le cas de l’économie, la météo, la bourse etc.

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.

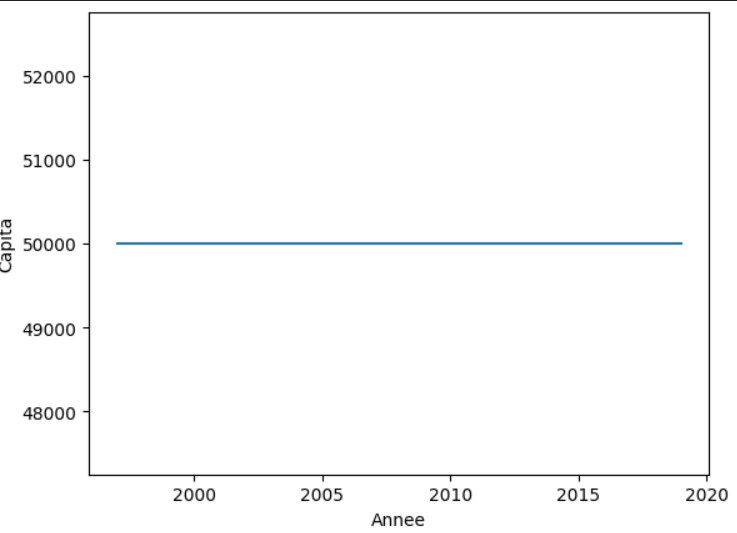
D’abord il nous faut des données, elles sont collectées, et nous avons vu comment dans la précédente partie. Nous allons faire la prédiction de chaque élément de chaque état financier. Ce qui nous fait une centaine de prédiction à faire.

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.



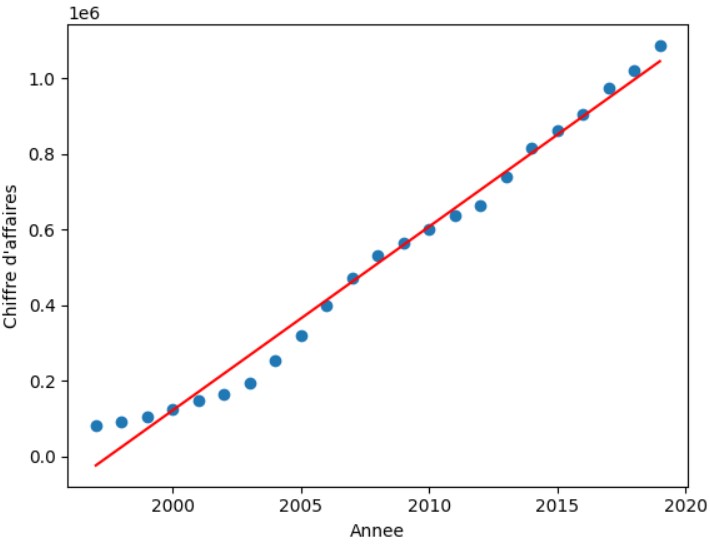
Du fait que nous avons plusieurs prédictions à faire, nous allons seulement en présenter quatre (4) du plus simple au plus intéressant.

* 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 de 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

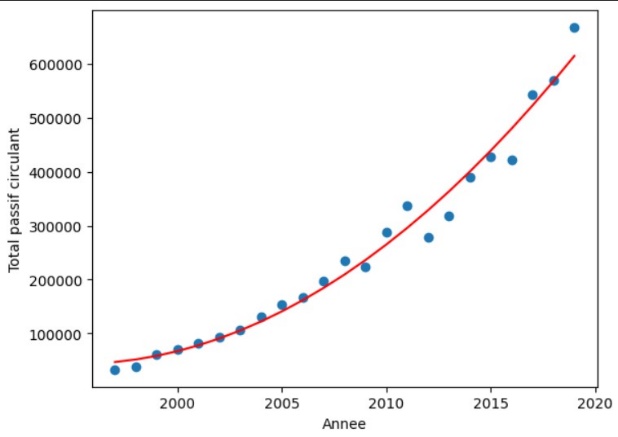


Il y a deux choses à voir ici, les points en bleu qui représentent les valeurs réelles du chiffre d’affaires en fonction des années et la droite linéaire en rouge qui est la droite de régression. La régression linéaire a bien fonctionné ici puisqu’elle épouse à la presque perfection les données. Cette droite va renvoyer le coefficient et la constante qui représentent nos w0 et w1 que nous avons déjà expliqué pour faire des prédictions pour les années à venir.

Le tableau ci-dessous nous donne pour la droite de régression du chiffre d’affaires son coefficient, sa constante et la prédiction de quelques valeurs futures.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficient | Constante | 2025 | 2030 | 2035 |
| 48578.23616 | -97034585.22134 | 1336343.0148221 | 1579234.1956521 | 1822125.376482 |

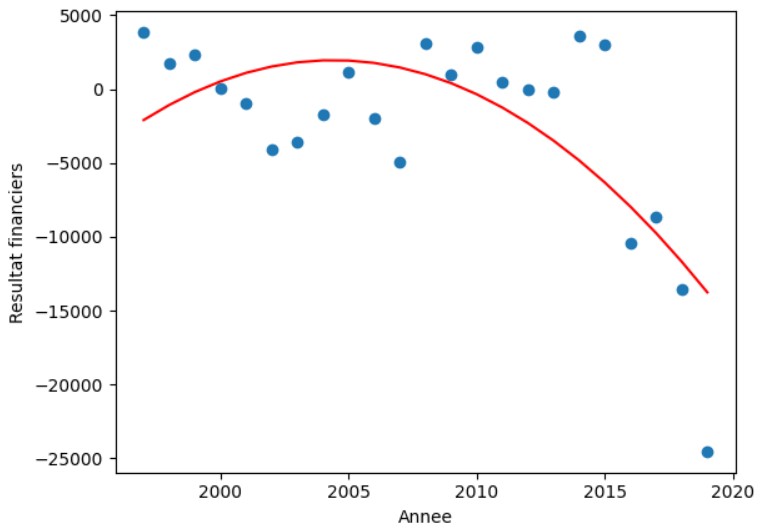
* Le passif circulant



Ce cas est un tantinet plus intéressant que le précédant 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 w0 et w1 mais aussi w2 et w3 (le biais), puisque le degré de notre polynôme est égal à 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| W0 | W1 | W2 | W3 | Année A |
| 0 | -4.00030285e+06 | 1.00251237e+03 | 3.99062354e+09 | W0+W1\*A + W2A2+W3 |

* 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 décroisent 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 compliquée que les autres, mais il n’en est rien. Nous allons simplement le faire passer dans une régression polynomiale et la magie va opérée. Le résultat pour ce modèle est la suivante.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| W0 | W1 | W2 | W3 | Année A |
| 0 | 2.95998776e+05 | -7.3836900e+01 | -2.9664942e+08 | W0+W1\*A + W2A2+W3 |

Il faut noter que pour tous les modèles, nous allons suivre exactement les mêmes procédures que 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 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.

#### Validation des modèles

La validation intervient dans le plusieurs étapes dans le processus de développent, son but est de permettre de choisir le meilleurs modèles avec les meilleurs paramètres. Dépendamment des modèles, il y aura différentes couches de validation. Il en va de soit que les modèles de régression n’auront pas autant de niveaux de validation que les ANN, RNN ou le *Reinforcement Learinn*g par exemple. Pour nos modèles de 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’entrainement et de test

Généralement, la première des choses que l’on va faire c’est de diviser le dataset en données d’entrainements 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 que tous *data scientists* a plusieurs étés confrontés, c’est : il me faut assez de pour faire l’entrainement donc le maximum des données doivent retrouver a l’entrainement, mais il me faut aussi assez de données pour faire le test. Dans ce cas, il faut des arbitrages et aussi tester plusieurs diviser, en règles générale on prend 80% des données pour l’entrainement et 20% pour le test. Il y a aussi une autre type division c’est 60% pour l’entrainement, 20% pour le test et 20% pour la validation. Cette dernière technique est utilisée pour les données de très grandes tailles.

Pour ce qu’il s’agit de notre travail, nous avons adaptée la division 80%, 20% des données car les nous avions seulement des données de 1997 à 2019. Dans certains cas, il est fait ce qu’on appelle la *cross validation* où le dataset va être diviser *folds*, disons 5 sous ensemble et chaque sous-ensemble va être entrainer séparément, cela reprend un peu l’idée du *random forest*. Cependant nous n’avons pas assez de donnes pour tester cette technique.

* Métrique de la validation

A ce moment, nous allons présenter les métriques utilisées pour la comparaison des différents modèles. Ces métriques peuvent être considéré comme des fonctions d’erreurs qui nous servent à calculer la distance entre les valeurs prédites et la réalité des valeurs, et tout ceci s’applique bien évidemment au données tests. Pour ce qu’il s’agit de la régression il y a trois metriques assez populaire 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 (R2) : la variance entre les valeurs prédites et les valeurs réelles

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | MAE | R2 |
| Régression linéaire | 60299.543411 | 44766.908198 | 0.945807 |
| Régression polynomiale | 37793.955849 | 30096.976848 | 0.978711 |
| Random forest | 39616.84609 | 27628.576000 | 0.970016 |

Ce tableau ci-dessus résume bien les résultats dont nous disposons et nous clairement savoir que la régression polynomiale est le modèle l’adaptée à nos chiffres d’affaires. Cette dernière présente un RMSE et un MAI plus faible et R2 qui se rapproche le plus 1. Donc c’est la raison pour laquelle c’est la régression polynomiale qui a été choisi pour la prédiction du chiffre d’affaires. Il sera abordé les résultats des aux autres modèles dans ce qui suit.

### Résultats et discussion

* **Résultats**

Pour ce qu’il s’agit de ce travail, nous avons étudier publieurs modèles de régression et quelques fois de classification. Nous avons commencé par la collecte des données puis passer au Fearture Engereening c’est-à-dire (l’analyse et 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 satisfaisant pour les uns les laisse a désiré pour les autres. Quoi qu’il en soit voici un les résultats de certains des modèles que nous avons développés.

|  |  |  |  |
| --- | --- | --- | --- |
|  | L’excèdent brute d'exploitation | | |
|  | RMSE | MAE | R2 |
| Régression linéaire | 20590.648902 | 17408.292125 | 0.975251 |
| Régression polynomiale | 26508.796376 | 22186.520542 | 0.958980 |
| Random forest | 22382.350571 | 17977.178000 | 0.970757 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Le résultat net | | |
|  | RMSE | MAE | R2 |
| Régression linéaire | 21191.289213 | 16698.198544 | 0.886268 |
| Régression polynomiale | 22892.405651 | 18745.964893 | 0.867276 |
| Random forest | 19647.888037 | 16322.178000 | 0.902232 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total actifs | | |
| Régression linéaire | 151944.961321 | 137918.854288 | 0.771658 |
| Régression polynomiale | 94550.704900 | 64463.893510 | 0.911582 |
| Random forest | 93707.317637 | 96959.800000 | 0.901471 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total des capitaux propres | | |
| Régression linéaire | 19838.776782 | 19273.393123 | 0.988315 |
| Régression polynomiale | 46235.199281 | 30806.219956 | 0.936534 |
| Random forest | 25752.790772 | 20621.528000 | 0.980310 |

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

* **Discussion**

Maintenant que nous développé nos modèles et obtenus les résultats, il est enfin possible de faire une causerie de nos résultats comme les analyses de matchs après un tournoi de basket. J’espere que cette analogie passera.

Pour ce qu’il s’agit des résultats, nous avons constaté qu’il y a principale trois modèles qui ont été testé que la régression linéaire, la régression polynomiale et le *random forest*. Et comme il est de coutume dans l’IA, ce sont les qui données qui définissent les modèles, il n’est pas pour le moment d’actualité l’outils qui permettre de trouver le modèle parfait du premier, donc il faut en tester plusieurs. Le modèle de régression linéaire bien que simple, nous permit de représenter plusieurs données, c’est le cas du total des capitaux propres, du report a nouveau, des impôts diffères 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 sortie avec par endroit les métriques qui avaient plus valeurs que les modèles de régression. C’est notamment du résultat net et des primes réserves. De là, puisqu’on avait pris l’analogie du basketball, le MVP désigné c’est nulle autre que la régression polynomiale qui a été le plus performent pour le maximum des données et des plus importants à savoir le CA, la VA, le EBE et j’en passe.

Néanmoins tout ne s’est pas déroulé comme prévu, nos modèles comportent bien entendu les limites et des points d’amélioration. Le grand défi était d’avoir plus de données, car les données dont disposons ont comment de 1997 et se terminent en 2019 avec beaucoup de données manquantes. Mais nous ne pouvons pas en vouloir à la BRVM qui a été a commencé ces activités seulement en 1998. La qualité des données en finance est défis majeur de l’IA appliquée à la finance comme mentionnée dans la section des limites des travaux de recherches. L’autre difficultés personnel étaient de nombre de modèle que nous nous devions de créer, tous les éléments du compte de résultat et tous les bilans, nous arrivons facilement à 100 modèles.

Toutes ses difficultés en valaient la peine car nous avons attaqué un gros morceau qu’est la finance à travers l’analyse financières. Aux finales nous avons réussi autant que faire se peu de développer des modèles de prédictions acceptable. Pour toujours rester sur l’analogie du basketball, notre prochain adversaire, celui pour la finale est le Chatbot : un adversaire coriace.

## Chapitre 2 : Conception du Chatbot pour l’interrogation des états financiers

Nous voilâmes dans le chapitre le plus intéressant du mémoire en termes de développement de modèle. Beaucoup chose sur le NLP vont être apprise dans ce chapitre, nous allons suivre le pipeline qui va nous mener de la collecte jusqu’au Chatbot que nous allons utiliser pour interroger les états financiers. Ce chapitre va être plein d’enseignement car il va nous permettre de démystifier cette merveille de technologie que sont les Chatbot.

### 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 textes sont les plus faciles à collecter. Nous aurons besoin des données texte pour le développement du Chatbot. Notre principal objectif ici c’est de comprendre les questions que l’utilisateur du Chatbot peut poser. Bien sûr, le Chatbot que nous allons développer va être spécialisé dans nos états financiers, donc nous aurons besoin de texte qui traite 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 labà, 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 apparaitre sachant qu’un autre est apparu.

* Les sites d’informations

A côté des textes relatif à la finance, il va nous falloir évidement du texte généraliste qui traite de tout et de rien. Et les sites d’information représentent une bonne ressource pour trouver du texte de qualité. Ceci va être important pour comprendre les objets d’une question que l’utilisateur peut éventuellement posée au Chatbot.

* Les bases de données

Il y a beaucoup de *dataset* disponible dans le web, mais aussi dans les bases de données. C’est d’autre développeur qui crée ses *dataset* et qui les stockent dans des bases de données, il y a des site web comme [kaggle](https://www.kaggle.com/datasets) qu’on peut prendre comme exemple. Dans certaines organisations comme les GAFAM (Google, Amazon, Facebook, Apple, Microsoft), elles font la collecte de données et puis paient des personnes pour faire l’étiquetage.

* Les sites de génération de textes

Les sites de génération de textes vont nous être d’utilité importante pour notre texte, ils vont nous permettre de d’avoir des données qui sont spécifiques pour notre travail. Dans le cas d’un Chatbot spécialisé, il est très difficile de trouver dans données qui nous conviennent pour le Intent classification pour des raison évidente. Donc, l’un des meilleurs moyens d’avoir des données qui sont spécifique à notre travail, c’est de générer voire même créer nos propres données.

* Le Web Scraping

Le web scraping est une technique en informatique ou l’on développe des programmes qui vont aller recueillir des informations sur le web sans l’intervention direct d’une personne. Ces programmes ressemblent dans leur fonctionnement aux robots des moteur de rechercher même s’ils n’ont rien à voir l’un de l’autres, mais l’analogie était pertinente. Il est nécessaire de préciser que certaines grandes organisations comme Google n’aime vraiment pas le web scraping mais ce dernier n’est pas illégal.

Nous avons utilisé un condensé de tout cela pour avoir les données texte dont nous aurons besoin. Il y aura d’abord les textes qui vont nous permettre de faire le Intent classification, ensuite il va utiliser des techniques pour reconnaitre les entités d’une question posée, pour pouvoir répondre avec la plus grande des précisions. Si vous n’avez pas compris cette dernière phrase, ne vous inquiétez pas car nous allons revenir en détail sur tout cela dans la partie où l’on parle du développement du Chatbot.

### 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 de MIT. Ce Chatbot était capable de tenir une conversation en utilisant un algorithme de reconnaissance de mot, ce qui implique qu’il ne comprenait pas vraiment le texte mais rechercher 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 particulier. 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 faire avancer le NLP mais aujourd’hui les Chatbots sont là pour nous aider dans nos vies.

Comment marche les Chatbots ? Le fonctionnement des Chatbot est simple, c’est comme une conversation entre deux personnes, l’un parle l’autre écoute puis répond. Cependant dans le cas des Chatbot, très souvent c’est, je pose une question et le Chatbot réponds, mais il est possible d’avoir des banalités avec les Chatbot. Tout dépend de comment 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 l’on va introduire les deux (2) concepts de base de tous les Chatbot : les ***intents***et le ***entity detection.***

#### La modélisation des textes

Maintenant, il y a une phase qui précèdent le *intent classification* et le *entity detection*, c’est la représentation de texte. L’ordinateur ne comprend pas le texte et ne peut pas faire des calculs sur du texte, or nos données sont de type texte et devons les comprendre, comment faire ? il y a un certain nombre de méthodes qui nous permet de modéliser le texte en nombre, c’est-à-dire sous un format que l’ordinateur va comprendre.

* Label encoding

Le *label encoding* estune technique de représentation de texte où l’on va associer chaque mot de tout de le *dataset* a un nombre bien particulier. Ce qui implique que nous aurions autant de *label* que de mot.

Si nous prenons la phrase « L’argent fait le bonheur », nous aurons : 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

Pour ce qui est du *one hot encoding*, il est plus sophistiqué que le *label encoding* mais un peu compliqué. Ici nous allons créer un tableau a deux dimensions, autant de ligne que de mots. Pour ce qui est de la colonne, elle sera égale au nombre de mot dans le texte, pour un mot, il va être noté de 1 à la place qui lui est représenté dans la colonne et toutes les autres valeurs de 0. Prenons encore la phrase « L’argent fait le bonheur » comme exemple.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | L’ | Argent | Fait | Le | Bonheur |
| L’ | 1 | 0 | 0 | 0 | 0 |
| Argent | 0 | 1 | 0 | 0 | 0 |
| Fait | 0 | 0 | 1 | 0 | 0 |
| Le | 0 | 0 | 0 | 1 | 0 |
| Bonheur | 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 veut dire les dimensions vont facilement augmenter ci le texte augmente. Imaginons un texte de 5000 mots, nous aurions 5000 éléments dans la colonne.

* Bag of Word

Le *bag of word* s’inscrit dans le même cadre que le OHE mais pour apporter des améliorations a ce dernier. Ici la taille de la colonne ne varie pas c’est les lignes qui vont changer. La différence OHE c’est ici on ne modélise pas un mot mais une phrase. Le fonctionnement c’est qu’après avoir mis tous les mots en colonne, on compte le nombre de fois que ce mot apparait dans une phrase.

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | L’ | Argent | Fait | Le | Bonheur | Est | Pouvoir |
| L’argent fait le bonheur | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| Le bonheur fait le bonheur | 0 | 0 | 1 | 1 | 2 | 0 | 0 |
| L’argent est le pouvoir | 1 | 0 | 1 | 0 | 0 | 1 | 1 |

Il y a moins d’entrée est plus de signification, le grand plus que l’on a ici c’est on modélise par phrase mais pas par mot. Le BOW connait aussi le *curse of dimensionality*, la taille de la colonne est proportionnelle aux données. En plus cela il y a aussi le problème du OOV (Out Of Vocabulary) pour des mots qui peut arriver dans les données testes et qui étaient absent dans les données d’entrainement.

* 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, ou trois mots par trois mots, enfin n mots par n mots : on parle de n grams. Chose qui nous permet de donner du sens au texte : « je suis » est plus compréhensible que « je » et « suis ».

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | L’argent | Argent fait | Fait le | Le bonheur | Argent est | Est le | Le pouvoir |
| L’argent fait le bonheur | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| L’argent est le pouvoir | 1 | 0 | 0 | 0 | 1 | 1 | 1 |

Nous voyons ici, on compte par un mot cela fait plus de sens que de compter par mots. 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 Frequnecy – Inverse Document Frequency* est une méthode qui calcule la fréquence d’un mot sur le document total. Pour ce faire, nous allons faire deux calcul le TF et le IDF, ensuite nous allons prendre le produit des deux.

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 méthode qui propose des moyens de transformer le texte en vecteur. Le plus populaire d’entre eux c’est le word2vec qui à partir d’un mot lui trouve son vecteur assimilé. Le bénéfice du werd2vec c’est que des mots qui se ressemble vont avoir des vecteurs qui se ressemble, il est possible de trouver la ressemblance entre deux 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 de mère se ressemble et ceux de maison et bâtiment se ressemblent.

Maintenant les valeurs de ces vecteurs ne tombent pas du ciel, ils sont trouvés avec le Machine Learning en faisant ce qu’on appelle le *Self Supervises Learning.*

* Fast text

Dans le *bag of n-grams*, on fait on comptait l’occurrence des mots dans une phrase, cela peut bien marcher mais souffre du OOV. Le *fast text* est venu pour palier à cela, ce qu’il fait par contre c’est de diviser le mot en *grams* et faire la représentation de chaque *gram.*

Si nous prenons le bonheur, avec 3 *grams*: {bo, bon, onh, nhe, heu, eur, ur} et le mot bonheur lui-même. Ce qui aura pour effet, si un mot qui ressemble à bonheur et qui n’était pas dans le *dataset*, nous pouvons travailler avec.

Il peut y arriver des fois qu’on est besoin de faire l’inverse, c’est-à-dire après avoir fait la représentation de texte et nos calculs, on obtient le résultat en nombre, de ce fait il faut faire l’opération inverse : transformer les nombres en texte.

#### 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 monnaie ? **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. Maintenant comment les machines font pour comprendre l’objet d’une question, il faut appeler 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 faisant ce que l’on appelle un *Intent classification.*

Nous allons faire un apprentissage supervisé, avec les données texte qui vont être étiquetées de leur objet. Les trois phases avec leur objet que nous avons donné représentent aussi des exemples d’à quoi va ressembler le *dataset.* Ce qui suit c’est la représentation de texte, transformer le texte en un format compréhensible par la machine. Apres avoir fait cela, nous le faisons passer dans un algorithme de Machine Learning, si nous avons une bonne *accuracy*, donc le modèle est bon, notre programme est capable maintenant de comprendre le texte. L’*intent classification* est un moyen très puissantqui nous permet de comprendre l’objet d’une question posée par un utilisateur.

#### L’entity detection

Maintenant que l’on comprend le sens des questions il faut bien répondre, mais pour répondre il nous faut des arguments, c’est 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 où nous pouvons répondre directement sans recherche d’arguments comme quand l’utilisateur salue ou remercie le Chatbot. 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, que va intervenir l’*entity detection.*

« Quel sera notre chiffre d’affaires de 2026 », quelle sont les éléments 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 sur lequel on veut faire une prédiction. C’est cela l’*entity detection*, maintenant une fois que nous avons relevé chiffre d’affaires et 2026 de la question, on appelle la fonction qui permet de faire les prédictions en lui passant les arguments, et après on envoie la réponse à l’utilisateur.

Il peut y arriver que l’utilisateur pose une question ou il n’y a pas tous les arguments mais le modèle parvient quand même à comprendre l’objet, exemple : « quelle est la prédiction de l’actif total », c’est une prédiction 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 prédiction si c’est ce dernier qui manque.

Parlons un peu des réponses, ce n’est vraiment pas la partie la plus compliquée dans le processus de développement. Une fois que l’on a passé avec brio les étapes d’***intents***et le ***entity detection*** et que nous avons aussi appelé les fonctions appropriées, on renvoie des réponses fixes préalablement définies. Maintenant il est possible d’avoir plusieurs réponses et renvoyer une au hasard pour avoir cet aspect moins robotique. Les réponses fixes sont appropriées pour notre programme sinon il faut utiliser des algorithmes générateurs de textes qui va rendre la chose beaucoup plus compliquée.

#### 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 intention, 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 la classification des intentions.

#### Test de fonctionnement du Chatbot

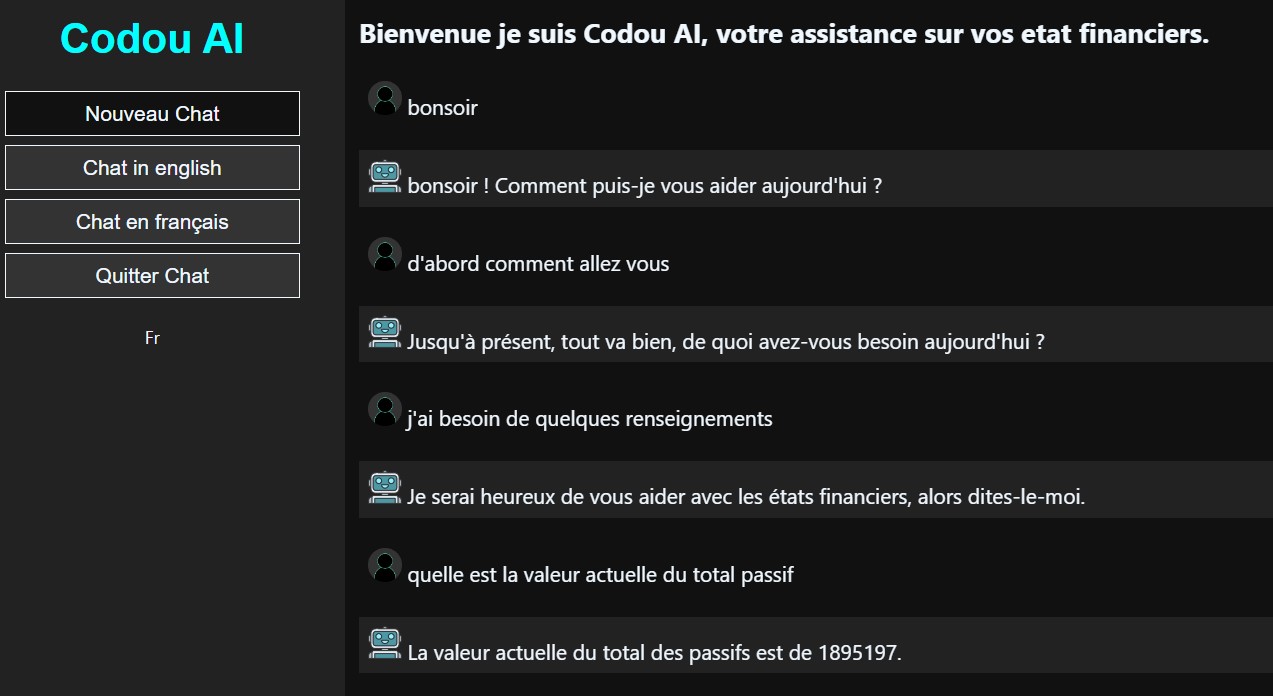
Maintenant donnons l’exemple de fonctionnement du Chatbot, ce que nous allons faire c’est d’entrer dans la "tête" du Chatbot qui reçoit une question, la traite et renvoie la réponse.

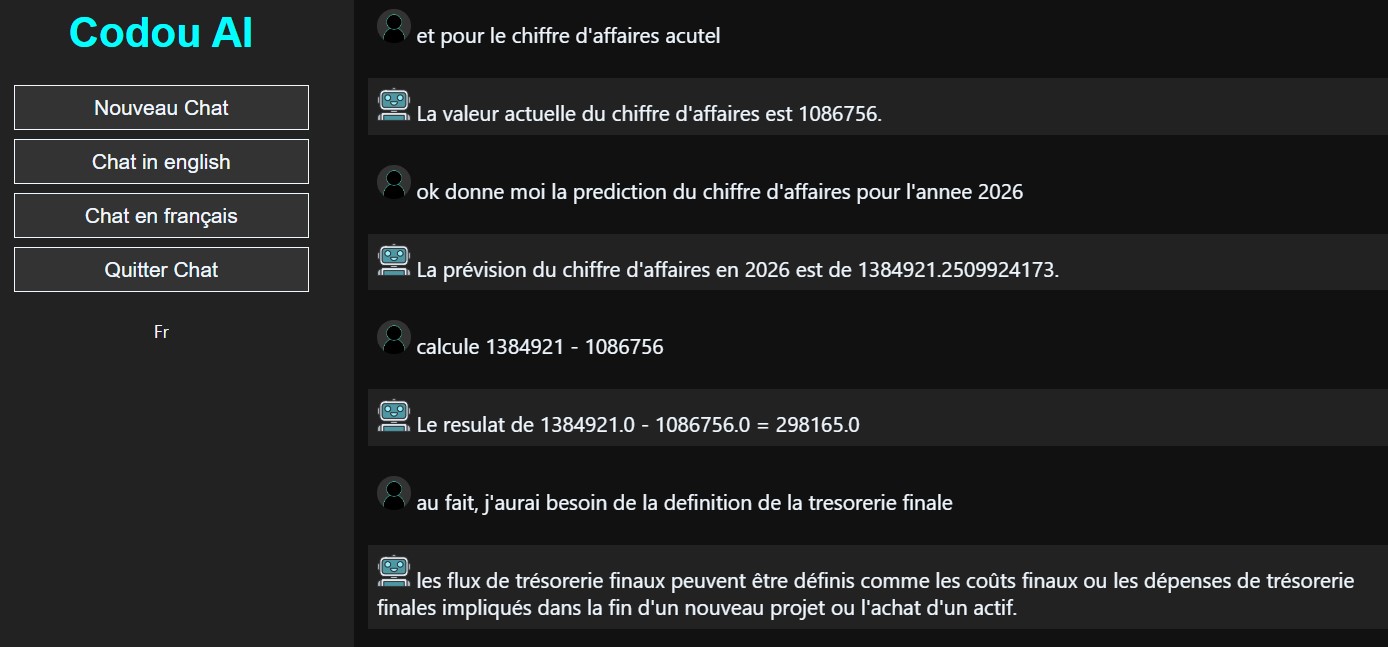
Utilisateur : quelle est la valeur présente du total actif ?

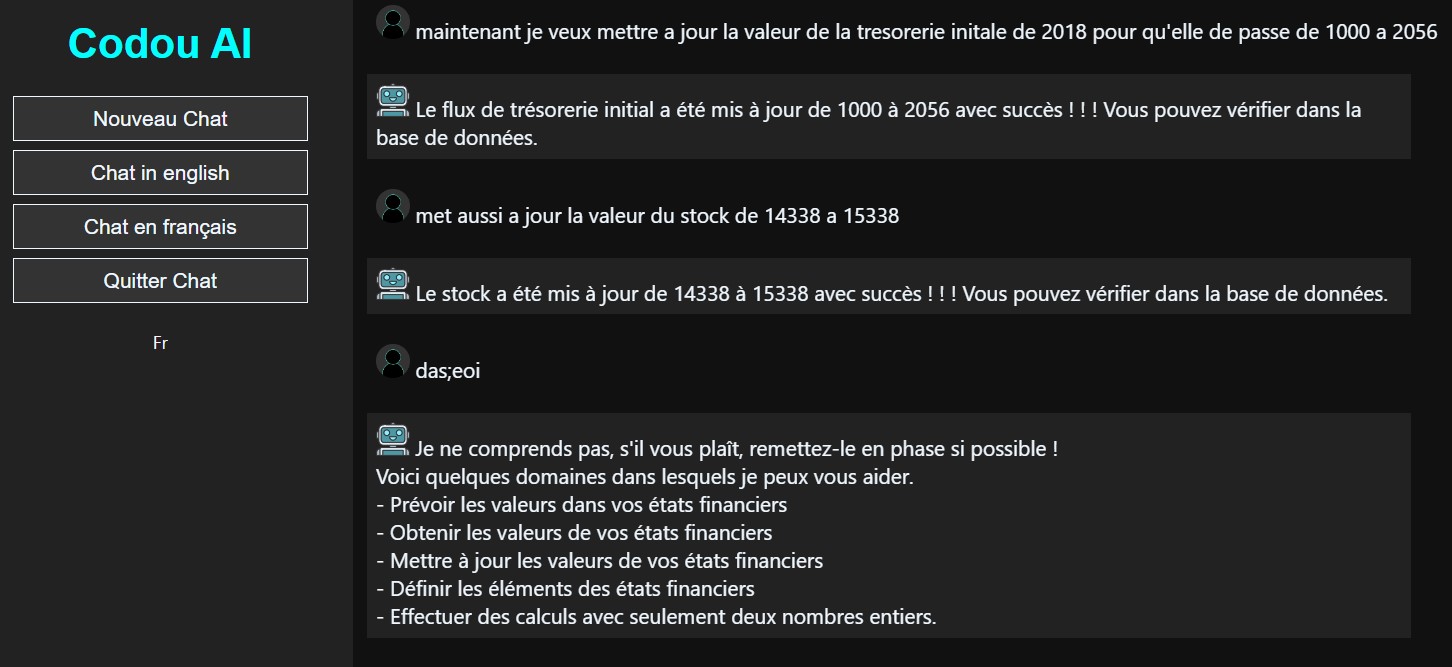
* La première chose à faire c’est la représentation de texte, utilisons pour cette phrase le BOW et nous aurons un vecteur qui compte les mots qui se trouve dans cette phrase.
* Apres avoir obtenu le vecteur nous allons le faire passer dans un algorithme de Machine Learning pour obtenir l’*intent*, le Naive Bayes en l’occurrence, et la sortie un « get » c’est recevoir une valeur dans la base de données.
* Apres l’*intent*, on passe au *entity detection*, ici nous comme entité l’actif total et « présente » qui signifie l’année actuelle, nous aurons (actif total, 2024), ce sera nos paramètres pour la fonction de « get ».
* Ensuite on appelle la fonction avec les paramètres qui nous renvoie la valeur dont on a besoin, disons 5000.
* Nous avons tout ce qui nous faut pour répondre, de toutes les réponses de « get » fixe on en choisi une par hasard, que l’on concatène avec la valeur et la réponse est prête.

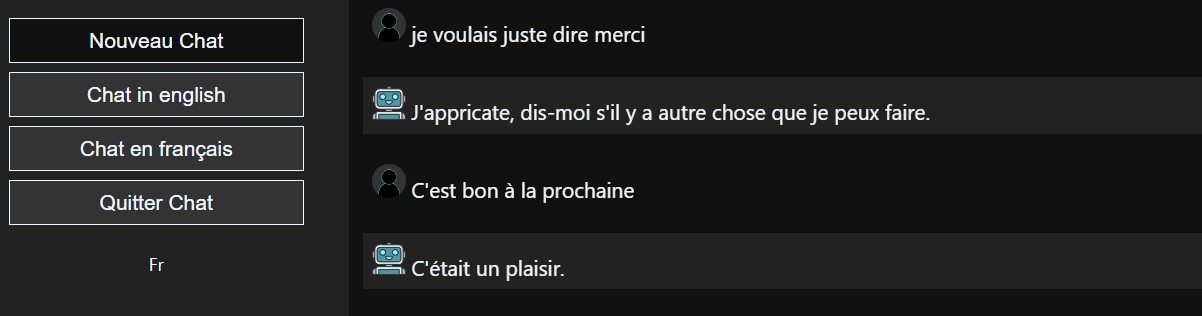
Chatbot : la valeur actuelle du total actif est de 5000.

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

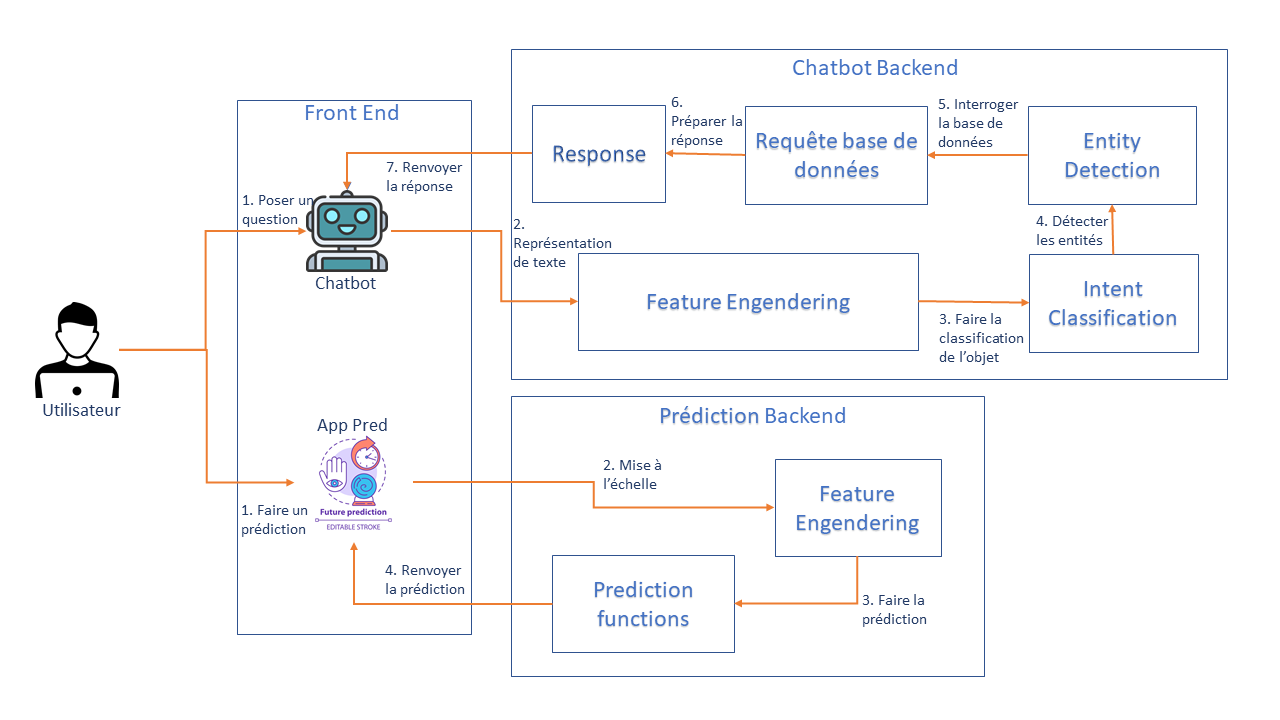








Nous pouvons conclure cette partie sur le Chatbot en disant pour son développement la question est plus importante que la réponse, si nous parvenons à comprendre parfaitement la question, nous parviendrons à donner la bonne réponse. Ci-après une image qui illustre l’architecture du processus global.



### Résultats et discussion

* **Résultats**

Pour dire vrai, développer ce Chatbot n’est une mince à faire, cela nous demander beaucoup compétence sur l’IA qui n’ont été toujours facile, mais la qualifie qu’il nous fallait le plus et dont nous nous sommes armés c’était la patience. C’est que nous allons présenter les résultats que nous avons obtenu 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 difficultés d’abord il y a la représentation de texte : ce qui veut dire transformer le texte en nombre dans le but qu’il passable dans un modèle. Le deuxième de niveau difficulté c’est le développement en tant que tel. La manière dont nous avons procéder est la suivante : pour chaque méthode de représentation, nous de faire passer sur tous modèles choisis. Il y aura trois méthodes de représentations à savoir le *bag of word*, le *bag of n-gram* et le *TF-IDF,* ajouté à cela six modèles que l’on a choisis.

Les métriques que nous utilisée c’est la précision, le *recall*, et le *f1-score* que nous avons déjà aborder 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 |

* + *Ramdom 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 |

* + *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 |

* + *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 |

* + *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 |

* + *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 |

Nous voyons bien comment différentes représentations de texte peuvent complètement les résultats des modèles. C’est toutes l’importance de choisir une technique de modélisation du. Enfin, il y a beaucoup de dire avec tous ses chiffres mais nous allons à les analyser et les interpréter mais aussi définir les limites de nos modèles dans la partie qui suit.

* **Discussion**

Puisque nous avons tous les résultats des modèles, il est maintenant de se poser, prendre du recul 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 alloue un *intent* en particulier. La première des choses a observée 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 améliore du BOW et le bag of n-gram. C’est pourquoi nous allons continuer cette discussion avec le résultat de méthode là. S’agissant des modèles, il y en a trois qui se démarque clairement des autres c'est le SVM, le *random forest* et le *decision tree*.

A ce moment il fallait choisis et cela n’allait pas se faire par hasard, donc il nous fallait un moyen de trancher. C’est nous ai venu l’idée de commencer à tester le Chatbot, en quoi faisant : et bien se mettre à la place de l’utilisateur et lui donner des textes que les trois modèles vont ensuite classer dans les différents *intents*. C’est un niveau supérieur de données de test. Apres avoir donner plusieurs phrases, il était devenu claire que c’est le *random forest* qui se démarquée le plus. Donc l’idée selon laquelle quand il s’agit de texte, il toujours par le *naive bayes* et ensuite tester les autres s’est avéré être vraie ici. Donc le modèle choisi c’est le *random forest* et la méthode de représentation c’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 BRVM mais ici il fallait faire la « création » de dataset soi-même. Ce qui en soi n’est pas du tout une tâche facile. Apres nous avions aussi rencontré beaucoup limites liées aux hardwares, même si les données ne sont pas gigantesques d’entrainement poser beaucoup surtout dans le cas de recherche en grille. Mais tout ceci n’a fait que rendre ce travail plus existent, dans quel monde serions-nous si le développement de genre de chose soit à la portée de n’importe qui.

Est-il possible d’améliorer ce travail ? La réponse est absolument. Dans chaque segment de développent 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 *intents* ou encore tester des models de LSTM ou GRU. Tout ceci serait intéressant mais l’objectif était de faire travail *from scratch* et comprendre se qu’il se passe dernier ses technologies que nous commençons a utilisé tous les jours.

Cette partie était captivante et on n’a pas envie de s’arrêter mais toute bonne chose a une fin, il est temps de conclure après ce gros morceau que constitue le développement des modèles, l’interface graphique et l’analyse financière.

## Conclusion de partie

Dans partie qui constitue celle pratique de notre travail, il a été fait la présentation financière de l’analyse, ensuite il a été explicité 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 simplisme.

La réponse que cette partie nous a permis de répondre c’est possible, l’IA ne doit forcément être une affaire des autres, une science tellement compliquée que je vais la laissée aux autres. Oui c’est complexe avec des prérequis complexe mais un beaucoup de volonté c’est à la portée des plus déterminés

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 risque qu’il n’a pas été fait le 1% du monde de l’IA, mais nous avons notre possible maintenant il temps de conclure ce travail fastidieux et chronophage.

# 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 dans quelle mesure l’intelligence artificielle va pouvoir aider les entreprises cotées à la Bourse régionale des Valeurs Mobilières (BRVM) à faciliter leur analyse financière et tout ce qui tourne au tour de leurs états financiers. 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 implémentant 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 performent
* 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

# Webographies

# Annexes

# Table des matières

[Remerciements I](#_Toc188130209)

[Sommaire II](#_Toc188130210)

[Liste des figures III](#_Toc188130211)

[Liste des tableaux III](#_Toc188130212)

[Liste des équations III](#_Toc188130213)

[Liste des sigles et des acronymes IV](#_Toc188130214)

[Introduction de générale 1](#_Toc188130215)

[Partie 1 : Fondements théoriques de l’intelligence artificielle appliquée à la finance 3](#_Toc188130216)

[Introduction de partie 3](#_Toc188130217)

[Chapitre 1 : Généralités et théories de l’intelligence artificielle 4](#_Toc188130218)

[Section 1 : L'intelligence artificielle : Définitions, origines et évolutions 4](#_Toc188130219)

[1. Définitions de l’intelligence artificielle 4](#_Toc188130220)

[2. Historique de l’intelligence artificielle 7](#_Toc188130221)

[2.1. Genèse de l’IA : le premier neurone artificiel 7](#_Toc188130222)

[2.2. Evolution 7](#_Toc188130223)

[2.3. Les sciences qui ont impulsé sa dynamique 8](#_Toc188130224)

[Section 2 : Fondements théoriques des algorithmes d’intelligence artificielle 10](#_Toc188130225)

[1. Les prérequis de l’intelligence artificielle 10](#_Toc188130226)

[1.1. Les mathématiques 11](#_Toc188130227)

[1.2. L’informatique 13](#_Toc188130228)

[2. Les algorithmes d’intelligence artificielle 15](#_Toc188130229)

[2.1. Machine Learning 15](#_Toc188130230)

[2.2. Deep Learning 26](#_Toc188130231)

[Chapitre 2 : Revue des travaux de recherche de l’intelligence artificielle appliquée à la finance 29](#_Toc188130232)

[Section 1 : Application générale de l’intelligence artificielle sur la finance 29](#_Toc188130233)

[1. Analyse prédictive 29](#_Toc188130234)

[2. Gestion des risques 30](#_Toc188130235)

[3. Services clients 30](#_Toc188130236)

[4. Détection de fraudes 31](#_Toc188130237)

[5. La bourse 32](#_Toc188130238)

[Section 2 : L’intelligence artificielle dans l’analyse des états financiers 32](#_Toc188130239)

[1. Les travaux de l’intelligence artificielle sur l’analyse financière 33](#_Toc188130240)

[2. Limites des travaux actuels 33](#_Toc188130241)

[Conclusion de partie 35](#_Toc188130242)

[Partie 2 : Conception et développement des outils d’IA appliquée à l’analyse des états financiers 36](#_Toc188130243)

[Introduction de partie 36](#_Toc188130244)

[Chapitre 1 : Analyse de développement de modèles prédictifs 37](#_Toc188130245)

[Section 1 : Mise en œuvre d’une application d’analyse financière 37](#_Toc188130246)

[1. Démarche d’une analyse financière 37](#_Toc188130247)

[2. Le frontend 42](#_Toc188130248)

[3. Le Backend 44](#_Toc188130249)

[4. Le Web server 45](#_Toc188130250)

[5. Présentation de l’application 45](#_Toc188130251)

[Section 2 : Développement de modèles prédictifs et l’analyse des données 47](#_Toc188130252)

[1. La collecte des données 47](#_Toc188130253)

[2. La prédiction des valeurs 48](#_Toc188130254)

[3. Validation des modèles 52](#_Toc188130255)

[Résultats et discussion 53](#_Toc188130256)

[Chapitre 2 : Conception du Chatbot pour l’interrogation des états financiers 56](#_Toc188130257)

[Section 1 : La collecte des données 56](#_Toc188130258)

[Section 2 : Le développement du Chatbot 57](#_Toc188130259)

[1. La modélisation des textes 58](#_Toc188130260)

[2. Les Intents 61](#_Toc188130261)

[3. L’entity detection 62](#_Toc188130262)

[4. La gestion des réponses 63](#_Toc188130263)

[5. Test de fonctionnement du Chatbot 64](#_Toc188130264)

[Résultats et discussion 67](#_Toc188130265)

[Conclusion de partie 70](#_Toc188130266)

[Conclusion générale et perspectives 71](#_Toc188130267)

[Bibliographies Webographies A](#_Toc188130268)

[Annexes A](#_Toc188130269)

[Table des matières B](#_Toc188130270)