Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

République Algérienne Démocratique et Populaire

Université Abou Bekr Belkaid - Tlemcen

Faculté des Sciences

Département d'informatique

Thème

Essai de prévision dans l'univers des blockchains à l'aide de l'intelligence artificielle

Réalisé par :

Djelloul Imad ALLAL

Présenté le 24 Juin 2024 devant le jury composé de MM.

Président	Sidahmed BERRABAH, Dr. (Université de Tlemcen)
Examinateur	Amine BELHOCINE, Dr. (Université de Tlemcen)
Encadrant	Salim ZIANI CHERIF, Dr. (Université de Tlemcen)

Année Universitaire : 2023 - 2024

Dédicaces

"Je dédie ce modeste travail à mes deux parents, en reconnaissance de leur soutien infaillible, de leurs encouragements constants et de leur amour inconditionnel. Leur présence et leurs sacrifices ont été les fondements solides qui ont illuminé chaque aspect de mon chemin.

Je souhaite également dédier ce travail à mon oncle Amine, pour ses précieux conseils, son soutien indéfectible et sa source inépuisable de motivation. Sa présence a été une inspiration."

بسم الله الرحمن الرحيم

قُل لَّا يَعْلَمُ مَن فِي السَّمَاوَاتِ وَالْأَرْضِ الْغَيْبَ إِلَّا اللَّهُ
وَمَا يَشْعُرُونَ أَيَّانَ يُبْعَثُونَ

النمل - ٦٥

Remerciements

Je tiens à exprimer toute ma gratitude envers mon encadrant, Dr. Ziani Cherif Salim, pour m'avoir accordé sa confiance dès le début de cette année académique et pour avoir accepté de collaborer avec moi. Sa bienveillance et son soutien ont été des piliers essentiels tout au long de ce projet. Je suis reconnaissant de pouvoir bénéficier de ses conseils avisés et de sa grande expertise, qui ont enrichi mon expérience et m'ont permis de progresser significativement.

Je profite de cette occasion pour remercier chaleureusement les membres du jury : Dr. Berrabah Sidahmed, maître de conférences à l'université Aboubekr Belkaid, ainsi que Dr. Belhocine Amine, maître de conférences au département d'informatique de l'université Aboubekr Belkaid.

Je souhaite exprimer ma sincère gratitude envers ma famille pour leur soutien constant. Mes parents, Kamal et Rachida, ont toujours été présents pour moi. Ma grand-mère, Nouria, ainsi que mes frères et sœurs, Anas, Lilia et Chahine, ont joué des rôles essentiels à mes côtés. Un grand merci également à mes oncles, Chafik et Amine, ainsi qu'à mes tantes Chahida et Meriam, pour leurs encouragements précieux. Amine en particulier est une source inestimable de motivation pour moi.

Un grand merci à tous mes amis, Rayan Mehdaoui, Moncef Ghellai, Issam Ziani, Khalil Sayah, Ibrahim Chaabane Sari, Rania Merad, Neila Mered, et Meryem Sebaa, et ma cousine Yasmine Allal, pour leur amitié précieuse et leur soutien constant.

Pour conclure, merci à toutes les personnes qui m'ont soutenu. Vos gestes, paroles et encouragements ont été cruciaux pour mon parcours et ma réussite.

Résumé

Le Bitcoin, en raison de sa volatilité extrême et de ses fluctuations imprévisibles sur le marché financier, présente un défi significatif pour la prévision précise de son prix futur. Ce projet se concentre sur le développement et l'évaluation de modèles avancés d'intelligence artificielle (IA), tels que la régression linéaire, les LSTM, les SVM et les Random Forest, dans le but de fournir des prévisions fiables malgré ces conditions volatiles. En parallèle, nous étudions l'influence potentielle des sentiments des articles de presse et d'autres variables externes comme le marché du pétrole et l'indice S&P 500 sur les tendances du marché du BTC. L'objectif est de mieux comprendre et d'anticiper les dynamiques sous-jacentes de cette cryptomonnaie en rapide évolution.

Ce rapport présente une approche enrichie visant à mieux comprendre les dynamiques sous-jacentes influençant les fluctuations des prix du Bitcoin, tout en explorant l'intersection entre la blockchain et l'intelligence artificielle.

Mots clés : Bitcoin, blockchain, time series forecasting, intelligence artificielle, régression linéaire, LSTM, SVM, Random Forest, analyse de sentiment.

Table des matières

Remerciements **III**

Résumé **IV**

Table des matières **V**

Liste des figures **VII**

Liste des tableaux **IX**

Introduction générale **I**

I Cadre conceptuel et historique de la blockchain **4**
I.1 Introduction 5
I.2 Origines et évolution de la Blockchain 5
I.3 Structure et sécurité du Bitcoin 8
I.3.1 Transactions électroniques pair-à-pair (P2P) 8
I.3.2 Double dépense 9
I.3.3 Mécanisme de la preuve de travail (Proof of Work) 10
I.4 Conclusions 13

II Intelligence Artificielle : Concepts de base et méthodes **14**
II.1 Introduction 16
II.2 Découverte de l'intelligence artificielle 16
II.2.1 Définition et historique 16
II.2.2 Applications de l'intelligence artificielle 17
II.3 Concepts de base de l'intelligence artificielle 17
II.3.1 Apprentissage automatique (Machine Learning) 17
II.3.2 Réseaux de neurones artificiels 18
II.3.3 Apprentissage profond (Deep Learning) 18
II.4 Modèles d'intelligence artificielle 19
II.4.1 Régression linéaire 19
II.4.2 Long Short-Term Memory (LSTM) 20
II.4.3 Support Vector Machine (SVM) 21

II.4.4 Random Forest 24
II.5 Méthodes du Machine Learning 26
II.5.1 Apprentissage supervisé 26
II.5.2 Apprentissage non supervisé 29
II.5.3 Apprentissage par renforcement 31
II.6 Processus de fonctionnement d'un réseau de neurones artificiels 31
II.6.1 Propagation avant 31
II.6.2 Rétropropagation 32
II.7 Deep Learning 33
II.7.1 Réseaux de neurones profonds (DNN) 34
II.7.2 Architectures des DNN en Deep Learning 34
II.8 Conclusions 37

III Essais de prévisions sur le Bitcoin : expérimentations 38
III.1 Introduction 39
III.2 Approche détaillée 39
III.3 Étapes de réalisation du projet 40
III.3.1 Environnement et outils de travail 40
III.3.2 Ensembles de données utilisés 40
III.3.3 Méthodes de prévision et métriques de performance 45
III.3.4 Implémentation des modèles de prévision 47
III.4 Résultats et observations 49
III.4.1 Régression 49
III.4.2 Classification 51
III.5 Discussions 53
III.6 Conclusions 54

Conclusions & Perspectives 56

Bibliographie 57

Liste des figures

I.1 Arbre de Merkle 6
I.2 Illustration de la Structure d'une Blockchain. 9
I.3 La difficulté moyenne de minage du Bitcoin de janvier 2009 à janvier 2024 (terahash/seconde). 11

II.1 Réseau de neurone artificiel vs humain: illustration. 18
II.2 Exemple de régression linéaire. 19
II.3 Illustration d'un modèle LSTM. 21
II.4 Exemple d'un hyperplan optimal du SVM. 22
II.5 Illustration montrant comment le Kernel facilite la séparation de données non-separables linéairement. 23
II.6 Exemple du Random Forest. 25
II.7 Comparaison entre les arbres de décisions et la Random Forest. 25
II.8 Illustration de la descente de gradient. 27
II.9 Illustration montrant la régression 28
II.10 Illustration montrant la classification. 29
II.11 Exemple du Clustering 30
II.12 Exemple d'anomalie. 30
II.13 Chaîne de neurones biologiques naturelle vs réseau de neurones artificiels profonds. 34
II.14 Illustration d'un réseau de neurones récurrents. 35

III.1 Fichiers contenant l'historique des transactions de chaque journée dans la blockchain Bitcoin. 42
III.2 Courbe du prix du Bitcoin en dollar (USD) entre la période de (2014 - 2024). 44
III.3 Courbe du prix du Bitcoin en dollar (USD) entre la période de (2021 - 2024). 44
III.4 Comparaison entre les prévisions de la régression linéaire et le LSTM pour l'ensemble de données $D1$. 50
III.5 Comparaison entre les prévisions de la régression linéaire et le LSTM pour l'ensemble de données $D2$. 51
III.6 Comparaison entre les prévisions des SVM pour les trois ensembles de données; $D1$, $D3$, et $D4$ en utilisant la matrice de confusion. 52

III.7 Comparaison entre les prévisions du Random Forest pour les trois ensembles de données; $D1$, $D3$, et $D4$ en utilisant la matrice de confusion. . 53

Liste des tableaux

III.1 Dénotation et classification des ensembles de données 41
III.2 Dataset, méthode et fréquence 41
III.3 Table démontrant l'équilibre entre les différentes phases du marché du Bitcoin dans la période de 2021 - 2024 45
III.4 Récapitulatif des versions finales de nos jeux de données 45
III.5 Comparaison entre les prévisions de la régression linéaire et le LSTM pour les ensembles de données $D1$ et $D2$ en utilisant la MAPE comme métrique de performance. 51
III.6 Comparaison entre les prévisions des SVM et du Random Forest pour les ensembles de données $D1$, $D3$ et $D4$ en utilisant l'exactitude comme métrique de performance. 53

Introduction générale

Exacerbés par la crise financière survenue en 2008, le monde a été confronté à une série de problèmes dans le domaine financier. Une crise qui avait sérieusement ébranlé la confiance du public et des professionnels dans les institutions financières traditionnelles, mettant en lumière les pratiques opaques et les abus de pouvoir qui étaient monnaie courante [29]. De plus, des millions de personnes dans le monde étaient exclues du système financier traditionnel à cette époque-là, limitant leur accès aux services financiers essentiels [46]. En outre, les transactions en ligne étaient souvent sujettes à la fraude et à la manipulation, soulignant le besoin d'un système de paiement plus sécurisé et infalsifiable [22].

En 2009, face aux défis posés par la crise financière mondiale et les autres facteurs, la technologie de la blockchain a émergé avec l'introduction du Bitcoin (BTC), une monnaie numérique décentralisée contrôlée par un réseau de participants, chacun contribuant à la validation et à la sécurisation des transactions. Cette innovation a permis de créer un système financier plus transparent et décentralisé, à la fois en offrant un registre public et immuable accessible à tous, mais également permettant des transactions sécurisées et vérifiables grâce à une cryptographie robuste et à un système de consensus distribué [36]. Le Bitcoin et la technologie de la blockchain ont suscité l'espoir d'une inclusion financière accrue, notamment en facilitant les paiements transfrontaliers et en réduisant les coûts de transaction. En effet, une étude a montré que les transactions Bitcoin pour les transferts transfrontaliers sont en moyenne 267 fois moins chères et 350 fois plus rapides que les services de transfert d'argent traditionnels [51]. De plus, la blockchain permet de réduire les frais de transaction en éliminant les intermédiaires, ce qui est particulièrement bénéfique pour les populations non bancarisées [42], et répond également au besoin de transparence et de traçabilité des transactions dans divers secteurs, permettant de suivre le parcours des biens et services de manière transparente et immuable [36].

Le Bitcoin a connu une ascension fulgurante depuis son lancement en 2009, atteignant des sommets historiques de presque 65 000 dollars en novembre 2021, puis de plus de 75 000 dollars en mars 2024, alimenté par une adoption croissante en tant que réserve de valeur et moyen de paiement alternatif, ainsi que par l'intérêt grandissant des investisseurs institutionnels [12, 3]. Cependant, ces montées spectaculaires ont été accompagnées d'une volatilité extrême, avec des fluctuations de prix pouvant atteindre plusieurs milliers de dollars en une seule journée. Cette volatilité est principalement attribuée à la nature spéculative du marché du Bitcoin, à son manque de réglementation et de maturité par rapport aux marchés financiers traditionnels, ainsi qu'à sa sensibilité aux événements médiatiques, aux annonces de célébrités et à l'intérêt institutionnel [45]. Prenons le cas de 2022, où le prix du Bitcoin a chuté brusquement, tombant sous les 20

000 dollars pour la première fois depuis 2020, en raison de l'effondrement de projets de cryptomonnaies de grande envergure — tel que TerraUSD (UST) et de son token associé, Luna (LUNA) [28] — de problèmes de liquidité et de faillites [23], mais aussi de guerres externes [2].

Il est donc nécessaire de noter que, bien que souvent vanté pour son potentiel de rendement élevé, investir dans le bitcoin présente des risques significatifs pour les investisseurs, et il est impératif pour eux de bien comprendre et de gérer les risques inhérents à cet actif numérique. Cependant, grâce aux avancées technologiques et notamment à l'avènement de l'intelligence artificielle, il est possible d'implémenter des techniques permettant de plus ou moins prévoir le prix du BTC, facilitant ainsi l'analyse aux personnes souhaitant y investir.

Ce projet vise donc à développer et d'évaluer des modèles de prévision avancés pour estimer la valeur future du Bitcoin en utilisant des techniques d'intelligence artificielle (IA). En intégrant des algorithmes de Machine Learning et de Deep Learning, nous visons à fournir des prévisions aussi fiables que possible. Cependant, il est crucial de noter que la prédiction du cours du Bitcoin reste un défi majeur en raison de la volatilité extrême et des fluctuations imprévisibles du marché, ce qui limite la précision et la fiabilité des résultats obtenus.

Dans un premier temps, nous tenterons de prévoir le prix du Bitcoin. Deux méthodes seront utilisées : la régression linéaire et le Long Short Term Memory (LSTM), et ce pour deux fréquences différentes : quotidienne et horaire.

Dans un second temps, nous tenterons de prévoir le mouvement du marché du BTC en termes de "hausse" ou de "baisse". Pour cette étape, nous utiliserons deux approches distinctes : les machines à vecteurs de support (SVM) et les forêts aléatoires (Random Forest). Nous prendrons également en compte différents facteurs pouvant influencer le comportement du Bitcoin, tels que les sentiments des articles de presse parlant du Bitcoin, ainsi que les marchés extérieurs pouvant avoir une corrélation avec cette cryptomonnaie, comme le pétrole et le S&P 500. Cette approche enrichie vise à fournir des prévisions plus robustes et à mieux comprendre les dynamiques sous-jacentes influençant les fluctuations des prix du Bitcoin.

Ce rapport présente deux domaines technologiques distincts : la blockchain et l'intelligence artificielle, et explore le potentiel impact de l'IA sur le Bitcoin.

Le premier chapitre se concentre sur la blockchain, définissant ses caractéristiques, son mode de fonctionnement et ses diverses applications. En particulier, le Bitcoin, la première et la plus connue des cryptomonnaies basées sur la blockchain, est examiné en détail, couvrant en détail son fonctionnement, son adoption

Le deuxième chapitre quant à lui se focalise sur l'intelligence artificielle ainsi que ses méthodes et modèles, et aborde en détail ceux utilisés pour la prévision du prix ainsi

que de la tendance du Bitcoin.

Le troisième chapitre, centre de cette étude, décrit les ensembles de données utilisés, expose les méthodologies de prévision choisies et présente leur implémentation. Les résultats obtenus y sont ensuite discutés, la performance des modèles est évaluée et les facteurs influençant la précision des prévisions sont identifiés.

Enfin, le rapport se conclut par une synthèse des conclusions et explore les perspectives futures, suggérant des pistes d'amélioration et des directions pour les recherches ultérieures.

I Cadre conceptuel et historique de la blockchain

Sommaire du chapitre

I.1 Introduction 5

I.2 Origines et évolution de la Blockchain 5

I.3 Structure et sécurité du Bitcoin 8

I.3.1 Transactions électroniques pair-à-pair (P2P) 8

I.3.2 Double dépense 9

I.3.3 Mécanisme de la preuve de travail (Proof of Work) 10

I.4 Conclusions 13

I.1 Introduction

La blockchain est une technologie de stockage et de transmission d'informations centrée sur un registre décentralisé, sécurisé et transparent. Introduite en 2008 par Satoshi Nakamoto pour soutenir le Bitcoin, cette technologie a révolutionné la manière dont les transactions numériques sont effectuées.

Sa capacité à assurer la transparence, la sécurité et l'immuabilité des données offre de nombreux avantages, tels que l'amélioration de l'efficacité opérationnelle, la réduction des coûts et l'augmentation de la confiance entre les parties prenantes.

La blockchain trouve des applications dans divers domaines tels que la finance, la santé, la gestion des chaînes d'approvisionnement, et bien d'autres.

I.2 Origines et évolution de la Blockchain

Bien que devenue connue seulement récemment, l'histoire de la blockchain remonte à quelques décennies.

1991: Le système de Timestamping

En 1991, deux chercheurs, Stuart Haber et W. Scott Stornetta [31], avaient publié un article intitulé "How to Time-Stamp a Digital Document", ces derniers étant préoccupés par la possibilité de falsification et d'anti datation des documents électroniques —ce qui pouvait compromettre la confiance dans les transactions et les enregistrements numériques ; ils ont proposé un système pour horodater ces documents de manière sécurisée, empêchant ainsi toute falsification ou antidatation.

En utilisant des techniques cryptographiques, ils ont cherché à établir un système où les documents pourraient être vérifiés de manière fiable sans nécessiter une autorité centrale, assurant ainsi une plus grande transparence et sécurité dans le stockage et la gestion des données numériques. Le fonctionnement du processus de timestamping (horodatage) décrit dans l'article de Haber et Stornetta est le suivant :

Lorsqu'un utilisateur envoie un document à un serveur de timestamping, le serveur ajoute un timestamp (marque temporelle) au document. Ce timestamp est une séquence de caractères ou d'informations encodées qui identifie précisément le moment où le document a été reçu. L'objectif de cela est de créer une empreinte digitale temporelle pour le document.

Ensuite, le serveur relie ce document au document précédent en incluant le timestamp du document précédent dans le nouveau document, ce qui crée par la suite une chaîne

continue de documents, chacun étant lié au précédent par son timestamp. Cette liaison de timestamps assure que chaque document soit enregistré dans un ordre chronologique précis. Si quelqu'un tente de modifier un document, le timestamp du document modifié ne correspondra plus au timestamp suivant dans la chaîne. Cette incohérence rendra l'entrée suivante invalide, ce qui signifie que toute tentative de manipulation des données sera immédiatement détectée. [31]

1992: Introduction aux arbres de Merkle

Un an plus tard, en 1992, et afin d'améliorer l'efficacité et la sécurité des enregistrements numériques, Stuart Haber, W. Scott Stornetta et Dave Bayer [4] ont introduit les arbres de Merkle dans leur conception de systèmes de timestamping.

Les arbres de Merkle, nommés d'après leur inventeur Ralph Merkle, sont des structures de données cryptographiques qui permettent de regrouper plusieurs enregistrements de données dans une séquence sécurisée de blocs [33]. Chaque feuille de l'arbre représente un hash d'un bloc de données, et chaque nœud non-feuille est un hash des nœuds enfants, culminant en une racine de Merkle unique qui résume l'intégrité de l'ensemble des données (figure I.1).

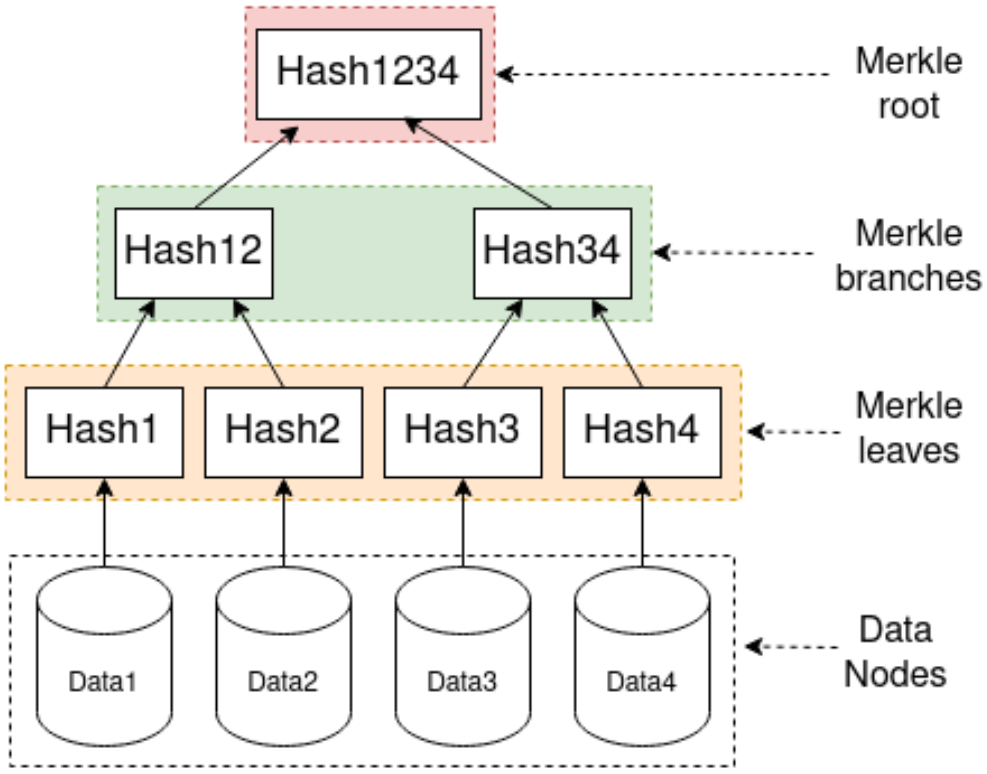

Fig. I.1 – Arbre de Merkle. [21]

Cette structure hiérarchique permet de vérifier efficacement l'intégrité des données sans avoir à stocker ou à vérifier chaque enregistrement individuellement. En effet, pour prouver qu'un enregistrement particulier fait partie de l'ensemble, il suffit de contrôler une petite quantité de données de preuve, par exemple, les hashs de quelques nœuds, sans avoir besoin de vérifier chaque élément de manière distincte, ce qui rend le processus de vérification extrêmement rapide et peu coûteux en termes de ressources.

Les arbres de Merkle jouent désormais un rôle crucial dans la technologie de la blockchain en assurant l'intégrité et l'immuabilité des données [54]. Toute tentative de modification d'un enregistrement entraînerait un changement dans le hash de la feuille correspondante, ce qui se propageait jusqu'à la racine de Merkle, rendant ainsi toute falsification immédiatement détectable. Cette propriété est essentielle pour maintenir la confiance et la sécurité dans les systèmes décentralisés, où il n'existe pas d'autorité centrale pour vérifier l'intégrité des données.

2008: Émergence du Bitcoin

En octobre 2008, un individu sous le pseudonyme de Satoshi Nakamoto a publié un livre blanc intitulé "Bitcoin : A Peer-to-Peer Electronic Cash System", introduisant ainsi un système de monnaie numérique qui permettrait des transactions directes entre pairs sans avoir besoin d'un intermédiaire de confiance [37], pour buts de :

- Garantir la sécurité,
- Éliminer les risques de fraude et d'annulation de transactions,
- Permettre des transactions à faible coût.

Celui-ci étant publié sur une liste de diffusion de cryptographie, il a rapidement attiré l'attention des chercheurs et des développeurs intéressés par les systèmes de paiement décentralisé (c'est-à-dire un système de paiement contrôlé par un réseau ouvert de participants) [38].

2009: Lancement de la toute première blockchain

Quelques mois plus tard, le 3 janvier 2009, Nakamoto a miné le premier bloc de la blockchain Bitcoin, connu sous le nom de "Genesis Block" ou "bloc 0" [38]. Ce bloc contenait un message encodé faisant référence à un article de presse sur le plan de sauvetage des banques britanniques, soulignant la motivation de Nakamoto à créer une alternative au système financier traditionnel [48].

2010 - 2015: Expansion et diversification

Après le lancement du Bitcoin, la technologie de la blockchain a commencé à attirer l'attention au-delà du domaine des cryptomonnaies. En 2013, Vitalik Buterin [10] a proposé la création d'une nouvelle blockchain appelée Ethereum.

Lancée en 2015, Ethereum a introduit le concept des "contrats intelligents", des programmes autonomes qui s'exécutent automatiquement lorsque certaines conditions sont remplies. Cette innovation a ouvert la voie à une multitude d'applications décentralisées (dApps) et a élargi l'utilisation de la blockchain à des domaines tels que la gestion des chaînes d'approvisionnements, les services financiers, la santé, et bien d'autres [38].

2016 - Présent : Adoption et réglementation

Plus tard, en 2016, la blockchain a continué à se développer et à se diversifier. De nombreuses entreprises et institutions financières ont commencé à explorer et à adopter cette technologie pour améliorer la transparence, la sécurité et l'efficacité de leurs opérations. On peut citer des géants comme IBM, Microsoft, et Amazon, qui ont lancé leurs propres plateformes de blockchain pour offrir des solutions adaptées à divers secteurs [34].

En parallèle, les gouvernements et les régulateurs du monde entier ont commencé à s'intéresser à la blockchain et aux cryptomonnaies, cherchant à établir des cadres réglementaires pour encadrer leur utilisation et prévenir les abus [58].

I.3 Structure et sécurité du Bitcoin

I.3.1 Transactions électroniques pair-à-pair (P2P)

Dans son document, Nakamoto introduit le concept de transactions électroniques pair-à-pair (P2P), où les utilisateurs peuvent échanger des actifs numériques directement entre eux, sans l'intervention d'un intermédiaire centralisé, telle que la banque [37].

Dans un réseau P2P, chaque participant (appelé nœud ou "mineur") possède une copie complète de la blockchain, assurant ainsi la décentralisation et la résilience du système en permettant une distribution équitable des données et des responsabilités. Il est à noter que toute personne souhaitant devenir participant peut le faire en utilisant le logiciel Bitcoin Core.

Cette architecture distribuée garantit que chaque nœud ait accès à toutes les transactions et peut vérifier leur validité de manière indépendante. En cas de panne ou d'attaque sur un nœud particulier, les autres nœuds continuent de fonctionner normalement, préservant ainsi l'intégrité et la disponibilité des données à travers le réseau.

I.3.2 Double dépense

I.3.2.1 Problème de la double dépense

Le problème de la double dépense est un défi majeur pour les systèmes de monnaie numérique. Ce dernier pourrait se produire lorsqu'un utilisateur tente de dépenser la même unité de monnaie numérique plus d'une fois.

Dans un système de paiement traditionnel, les institutions financières agissent comme des tiers de confiance pour vérifier que les fonds ne sont pas dépensés plus d'une fois.

Cependant, dans un système décentralisé sans intermédiaire, il est nécessaire de trouver une autre solution pour garantir que chaque unité de monnaie numérique ne puisse être dépensée qu'une seule fois [54].

I.3.2.2 Solution proposée par Nakamoto

Pour résoudre le problème de la double dépense dans son système de paiement, Nakamoto propose l'utilisation de la blockchain, afin d'enregistrer toutes les transactions effectuées dans un registre public distribué.

Les nœuds du réseau valident les transactions et les regroupent dans des blocs, qui sont ensuite liés de manière cryptographique, chaque bloc contient un hash cryptographique du bloc précédent, formant ainsi une chaîne continue et immuable de blocs (blockchain) (figure I.2).

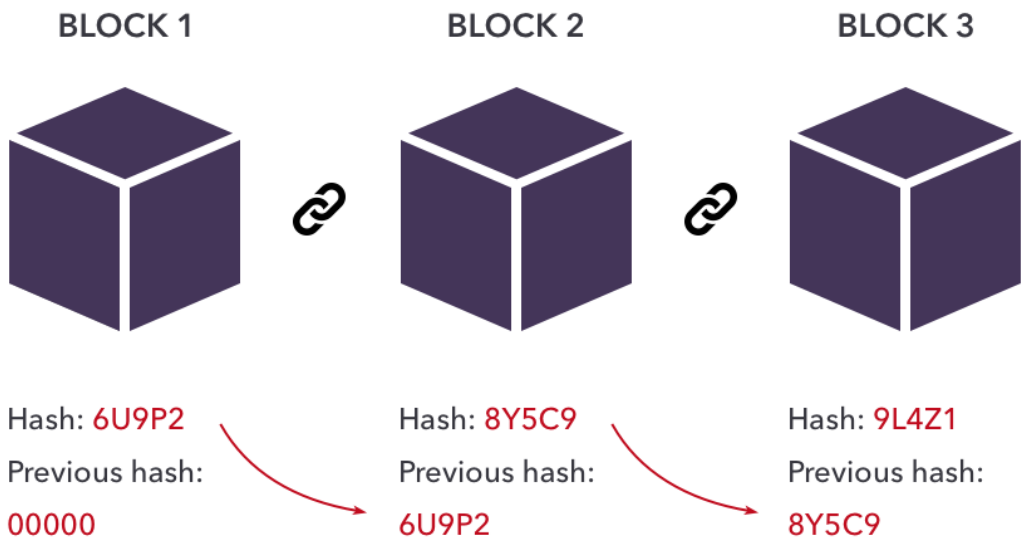

Fig. I.2 – Illustration de la Structure d'une Blockchain.

Cette structure garantit que toute tentative de modification d'un bloc entraînerait un changement dans tous les blocs suivants, rendant la falsification facilement détectable [1].

I.3.3 Mécanisme de la preuve de travail (Proof of Work)

Pour empêcher les attaques et les modifications non autorisées de la blockchain, Nakamoto a introduit une innovation clé appelée "preuve de travail" (Proof of Work, PoW).

I.3.3.1 Concept de Base

Le PoW est un mécanisme de consensus qui nécessite que les mineurs résolvent des puzzles cryptographiques complexes pour valider les transactions et ajouter de nouveaux blocs à la blockchain (un processus appelé également minage). Ce processus est conçu pour être difficile et coûteux en termes de puissance de calcul, mais facile à vérifier pour les autres nœuds du réseau [38].

I.3.3.2 Processus de Minage

Le processus de minage implique les étapes suivantes :

— **Sélection des Transactions :** Les mineurs sélectionnent les transactions non confirmées dans le mempool (pool de mémoire) et les regroupent dans un bloc candidat.

— **Calcul du Hash :** Les mineurs doivent trouver un nonce (un nombre arbitraire) qui, lorsqu'il est combiné avec les données du bloc et passé à travers une fonction de hachage cryptographique (SHA-256 dans le cas de Bitcoin), produit un hash qui respecte une certaine condition de difficulté (voir I.3.3.3). Cette condition est généralement que le hash doit commencer par un certain nombre de zéros.

— **Validation et Propagation :** Une fois qu'un mineur trouve un nonce valide, il diffuse le nouveau bloc au reste du réseau. Les autres nœuds vérifient la validité du bloc et l'ajoutent à leur copie de la blockchain si toutes les vérifications sont réussies [1].

I.3.3.3 Difficulté et Ajustement

La difficulté du puzzle cryptographique est ajustée périodiquement (tous les 2016 blocs pour le Bitcoin) pour garantir que les nouveaux blocs sont ajoutés à la blockchain

à un rythme constant, environ toutes les 10 minutes. Cet ajustement est basé sur la puissance de calcul totale du réseau. Si les blocs sont minés plus rapidement que prévu, la difficulté augmente, et si les blocs sont minés plus lentement, la difficulté diminue [1].

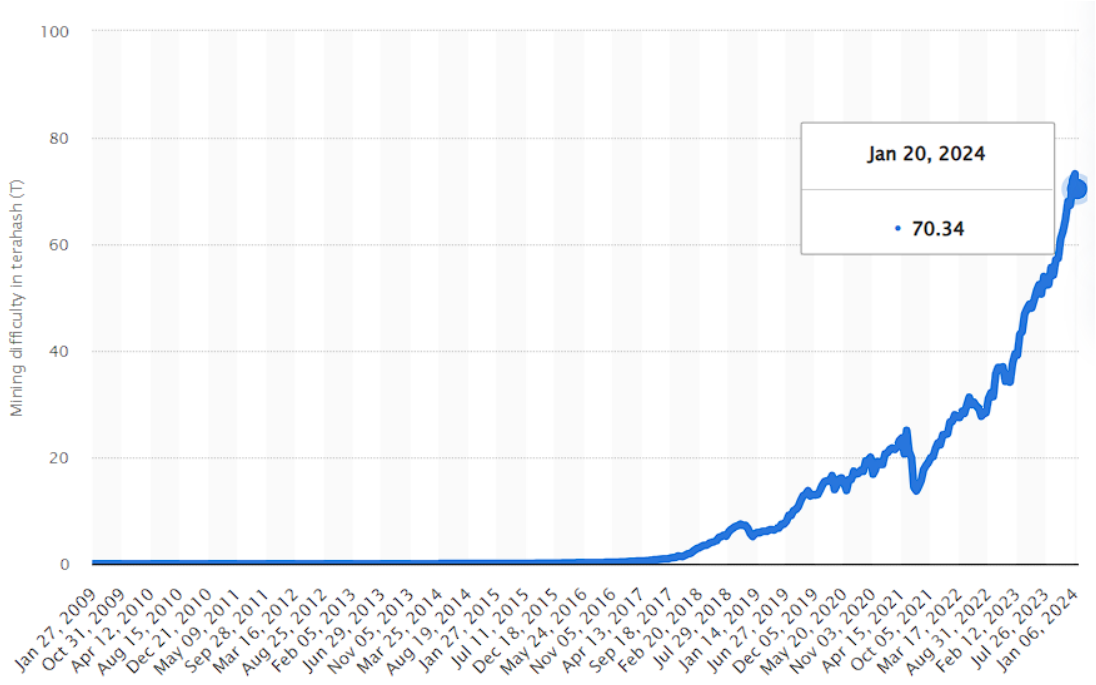

Fig. I.3 – La difficulté moyenne de minage du Bitcoin de janvier 2009 à janvier 2024 (terahash/seconde). [47]

La figure ci-dessus (figure I.3) est une représentation de la difficulté moyenne —en terahash par seconde (TH/s)— de minage du Bitcoin, de janvier 2009 à janvier 2024.

Dans le réseau du Bitcoin, le terme ”terahash” (TH/s) est une unité de mesure de la puissance de calcul utilisée pour le minage. Notons que :

$$1 \text{ terahash (TH)} = 10^{12} \text{ hachages} \tag{I.1}$$

D'après la figure *(figure 1.3)*, on remarque que la difficulté moyenne en janvier 2024 était de 70.34 TH/s.

$$70.34 \text{ terahash (TH)} = 70.34 \times 10^{12} \tag{I.2}$$

Soit, 70,340,000,000,000, ou 70.34 trillions de hachages par seconde

Cela signifie donc que le mineur ou le réseau doit être capable de calculer 70,34 trillions de hachages par seconde. Plus la puissance de calcul totale du réseau est élevée, plus il est difficile pour un attaquant de réaliser une attaque de 51%, où il contrôlerait plus de 50% de la puissance de calcul du réseau [7].

I.3.3.4 Sécurité et Résistance aux Attaques

La preuve de travail rend les attaques coûteuses et difficiles à réaliser. Pour réussir une attaque de 51%, où un attaquant contrôle plus de 50% de la puissance de calcul du réseau, l'attaquant doit dépenser une quantité énorme de ressources pour surpasser le reste du réseau. Cela rend les attaques économiquement non viables pour la plupart des attaquants potentiels [7].

I.3.3.5 Récompenses et Incitations

Le premier mineur qui réussit à ajouter un nouveau bloc à la blockchain reçoit une récompense en Bitcoins nouvellement créés, ainsi que les frais de transaction inclus dans le bloc. Cette récompense sert d'incitation économique pour les mineurs à continuer de sécuriser le réseau. La récompense en Bitcoins est divisée par deux environ tous les quatre ans, cet événement est connu sous le nom de "Halving" (voir sous-section I.3.3.6) [1].

I.3.3.6 L'impact du Cycle de Halving sur le Réseau Bitcoin

Se produisant environ tous les quatre ans, le Halving est un phénomène du Bitcoin qui réduit de moitié la récompense de minage afin de contrôler l'inflation et limiter l'offre totale à 21 millions de bitcoins.

Ce mécanisme maintient une incitation à long terme pour les mineurs, car les frais de transaction deviendront leur principale source de revenus une fois que tous les Bitcoins seront émis. En augmentant la rareté de la cryptomonnaie, le Halving peut potentiellement accroître la valeur des bitcoins existants.

Cet événement prévisible et transparent permet aux participants du marché d'anticiper les changements d'offres. En intégrant le halving directement dans le protocole, Nakamoto a assuré la décentralisation et la résistance à la censure du système du Bitcoin [38].

I.4 Conclusions

En conclusion de ce chapitre, nous avons exploré en détail ce qu'est la blockchain, en mettant l'accent sur ses mécanismes fondamentaux, ses avantages et ses différentes applications.

Dans le chapitre à venir, nous changerons complètement de domaine, et examinerons les principes de base de l'intelligence artificielle, ses diverses applications et comment elle peut être intégrée avec la blockchain pour créer de potentielles solutions.

II Intelligence Artificielle : Concepts de base et méthodes

Sommaire du chapitre

II.1 Introduction 16

II.2 Découverte de l'intelligence artificielle 16

II.2.1 Définition et historique 16

II.2.2 Applications de l'intelligence artificielle 17

II.3 Concepts de base de l'intelligence artificielle 17

II.3.1 Apprentissage automatique (Machine Learning) 17

II.3.2 Réseaux de neurones artificiels 18

II.3.3 Apprentissage profond (Deep Learning) 18

II.4 Modèles d'intelligence artificielle 19

II.4.1 Régression linéaire 19

II.4.2 Long Short-Term Memory (LSTM) 20

II.4.3 Support Vector Machine (SVM) 21

II.4.4 Random Forest 24

II.5 Méthodes du Machine Learning 26

II.5.1 Apprentissage supervisé 26

II.5.2 Apprentissage non supervisé 29
II.5.3 Apprentissage par renforcement 31
II.6 Processus de fonctionnement d'un réseau de neurones artificiels 31
II.6.1 Propagation avant 31
II.6.2 Rétropropagation 32
II.7 Deep Learning 33
II.7.1 Réseaux de neurones profonds (DNN) 34
II.7.2 Architectures des DNN en Deep Learning 34
II.8 Conclusions 37

II.1 Introduction

Depuis ses débuts dans les années 1950, l'intelligence artificielle (IA) a révolutionné notre monde de manière significative. Perçu comme de la science-fiction auparavant, ce domaine est désormais omniprésent dans notre quotidien.

De la détection de fraudes aux véhicules autonomes, l'IA ne cesse de nous impressionner en transformant divers aspects de notre vie, et ce grâce à ses capacités d'apprentissage automatique et de traitement des données en grande quantité, permettant ainsi de prédire et d'anticiper des comportements complexes.

Dans le domaine de la finance par exemple, l'IA pourrait être utilisée pour prévoir et anticiper les tendances du marché en temps réel, et offrir des informations cruciales aux investisseurs et aux décideurs financiers.

Ce chapitre est consacré au domaine de l'intelligence artificielle. Dans celui-ci, nous allons voir ce qu'est l'IA, ainsi que ses composantes fondamentales, à savoir l'apprentissage automatique, les réseaux de neurones, et l'apprentissage profond, et à la fin, nous détaillerons les différentes architectures des modèles employés au cours de notre projet, mettant en œuvre ainsi leurs spécificités et leurs avantages. En particulier, nous explorerons comment des modèles tels que la régression linéaire, le LSTM, les SVM ou même le Random Forest, peuvent être utilisés pour faire des prévisions sur le marché du Bitcoin.

II.2 Découverte de l'intelligence artificielle

II.2.1 Définition et historique

L'intelligence artificielle est un domaine en informatique visant à créer des systèmes capables d'imiter des fonctions humaines telles que l'apprentissage, le raisonnement, et la résolution de problèmes. Elle utilise diverses techniques comme les réseaux de neurones ou l'apprentissage automatique afin de permettre aux machines de comprendre leur environnement et de résoudre des problèmes de manière autonome [9].

L'histoire de l'IA a officiellement débuté en 1956 lors de la conférence de Dartmouth [19], durant laquelle des pionniers comme John McCarthy et Marvin Minsky se sont réunis pour discuter de la possibilité de créer des machines intelligentes. Cette conférence a jeté les bases théoriques et conceptuelles de l'IA, en introduisant des idées et des objectifs qui ont guidé la recherche et le développement du domaine pendant des décennies, et c'est lors de cette dernière que le terme "intelligence artificielle" a été introduit [40].

Quelques décennies plus tard, dans les années 2000, l'apprentissage automatique, les réseaux de neurones, et l'apprentissage profond ont vu le jour, marquant ainsi une

révolution dans le domaine [30]. Ces avancées technologiques ont permis aux machines d'apprendre à partir de données, ouvrant ainsi la voie à de nombreuses applications.

II.2.2 Applications de l'intelligence artificielle

L'intelligence artificielle peut être utilisée dans différents domaines, on peut notamment la retrouver dans :

— **La finance :** L'intelligence artificielle joue un rôle crucial dans le secteur financier, en apportant des améliorations significatives dans divers secteurs. Par exemple, elle peut être utilisée pour prédire les ventes quotidiennes et classer les résultats de recherche de manière à maximiser les revenus des entreprises. Par ailleurs, l'IA peut également être appliquée à la prédiction des cours boursiers, l'une des séries temporelles les plus complexes, fournissant ainsi des analyses précieuses pour les investisseurs et les entreprises financières [20].

— **La médecine :** L'IA permet parfois de réaliser des diagnostics plus rapides et plus sûrs, notamment pour détecter les cancers sur les lames histologiques ou les radiographies [53]. De plus, elle aide et corrige les interventions chirurgicales assistées par robot, augmentant ainsi la précision et la sécurité des opérations. Cette technologie peut également être utilisée dans de nombreuses applications pour gérer les maladies chroniques, comme l'enregistrement continu de la glycémie, offrant ainsi une gestion plus efficace et proactive des conditions de santé à long terme.

— **Le transport :** Dans le transport urbain, l'intelligence artificielle peut améliorer considérablement la sécurité des travailleurs en analysant les causes des accidents [52]. En outre, elle est utilisée pour détecter les comportements suspects dans les transports grâce à la vidéosurveillance.

Ces applications —Comme évoqué dans la section précédente—, ont été rendues possibles grâce à l'avènement des techniques de l'apprentissage automatique, des réseaux de neurone et de l'apprentissage profond. Les section à venir explorent plus en profondeur l'univers de l'IA, en découvrant les aspects cachés de ce monstre technologique.

II.3 Concepts de base de l'intelligence artificielle

II.3.1 Apprentissage automatique (Machine Learning)

L'apprentissage automatique, ou "Machine Learning (ML)", est une sous-discipline de l'intelligence artificielle qui repose sur la création de modèles capables d'apprendre

à partir de données. Plutôt que d'être explicitement programmés pour accomplir une tâche spécifiée préalablement, ces modèles utilisent des algorithmes qui permettent de les entraîner à identifier des motifs dans les données, et d'améliorer leurs performances au fil du temps.

Le Machine Learning est appliqué dans de nombreux domaines tels que la médecine, la finance, ou même la biologie, et englobe trois approches différentes : l'apprentissage supervisé, l'apprentissage non supervisé, et l'apprentissage par renforcement [11].

II.3.2 Réseaux de neurones artificiels

Les réseaux de neurones artificiels (ANN) sont des systèmes informatiques inspirés par la structure et le fonctionnement du cerveau humain (figure II.1). Ils sont majoritairement utilisés dans l'apprentissage automatique.

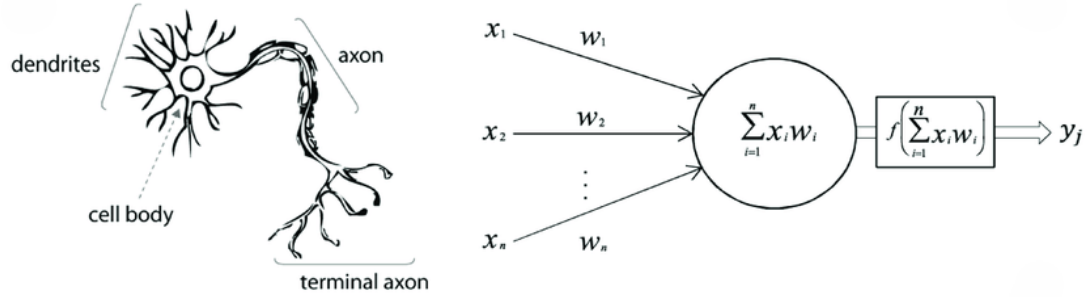

Fig. II.1 – Réseau de neurone artificiel vs humain : illustration. [32]

II.3.3 Apprentissage profond (Deep Learning)

L'apprentissage profond, ou "Deep Learning (DL)", est une sous-discipline de l'apprentissage automatique qui se concentre sur l'utilisation de réseaux de neurones artificiels profonds pour modéliser des données complexes.

À la différence des méthodes d'apprentissage automatique traditionnelles, le DL permet de traiter des volumes massifs de données non structurées, comme des images, des sons, et des textes, en extrayant automatiquement des caractéristiques pertinentes. Cette capacité est rendue possible grâce à des architectures de réseaux de neurones multicouches (voir sous-section II.7.1), qui permettent de capturer des représentations hiérarchiques des données [27].

II.4 Modèles d'intelligence artificielle

Divers modèles d'intelligence artificielle contribuent à son développement. Cette section présente ceux qui ont été utilisés pour la réalisation de notre projet.

II.4.1 Régression linéaire

La régression linéaire est un algorithme de régression (voir sous-sous-section II.5.1.2) qui cherche à ajuster une ligne droite aux données de manière à minimiser la somme des carrés des écarts entre les valeurs observées et les valeurs prédites [11] (figure II.2).

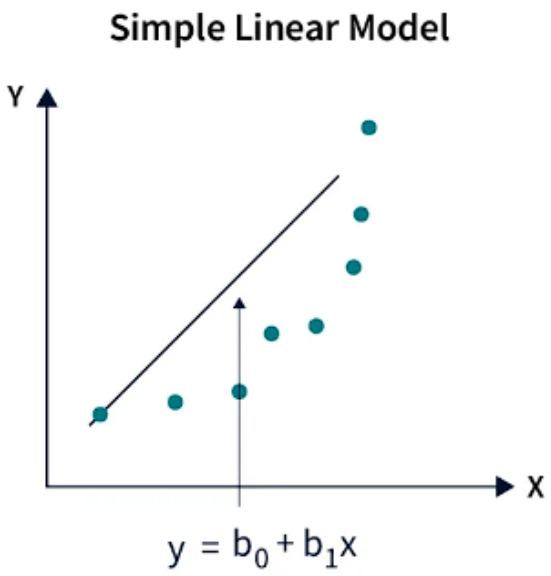

Fig. II.2 – Exemple de régression linéaire. [50]

La formule de base est :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \tag{II.1}$$

où :

— y est la variable dépendante,
— $x_1, x_2, \ldots, x_n$ sont les variables indépendantes,
— $\beta_0, \beta_1, \ldots, \beta_n$ sont les coefficients à estimer,
— ϵ est l'erreur.

L'objectif est de trouver les valeurs des paramètres $\beta_0, \beta_1, \ldots, \beta_n$ qui minimisent l'erreur quadratique moyenne (MSE) entre les valeurs observées et les valeurs prédites par le modèle.

La formule de la MSE est donnée par :

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

où :

- $\mathbf{y}_i$ représente les valeurs observées de la variable dépendante,
- $\hat{\mathbf{y}}_i$ représente les valeurs prédites par le modèle de régression,
- $\mathbf{n}$ est le nombre total d'observations.

II.4.2 Long Short-Term Memory (LSTM)

Le Long Short-Term Memory (LSTM) est un type de réseau de neurones récurrents (RNN) (voir sous-sous-section II.7.2.2) conçu pour mieux capturer les dépendances à long terme dans les séquences de données.

Contrairement aux RNN traditionnels, les LSTM sont capables de conserver des informations sur de longues périodes grâce à une structure de mémoire intéressante [18].

Un LSTM est composé de plusieurs cellules de mémoire, chacune ayant trois portes principales : la porte d'entrée (input gate), la porte d'oubli (forget gate), et la porte de sortie (output gate). Ces portes régulent le flux d'informations à travers la cellule de mémoire (state candidate gate) (figure II.3).

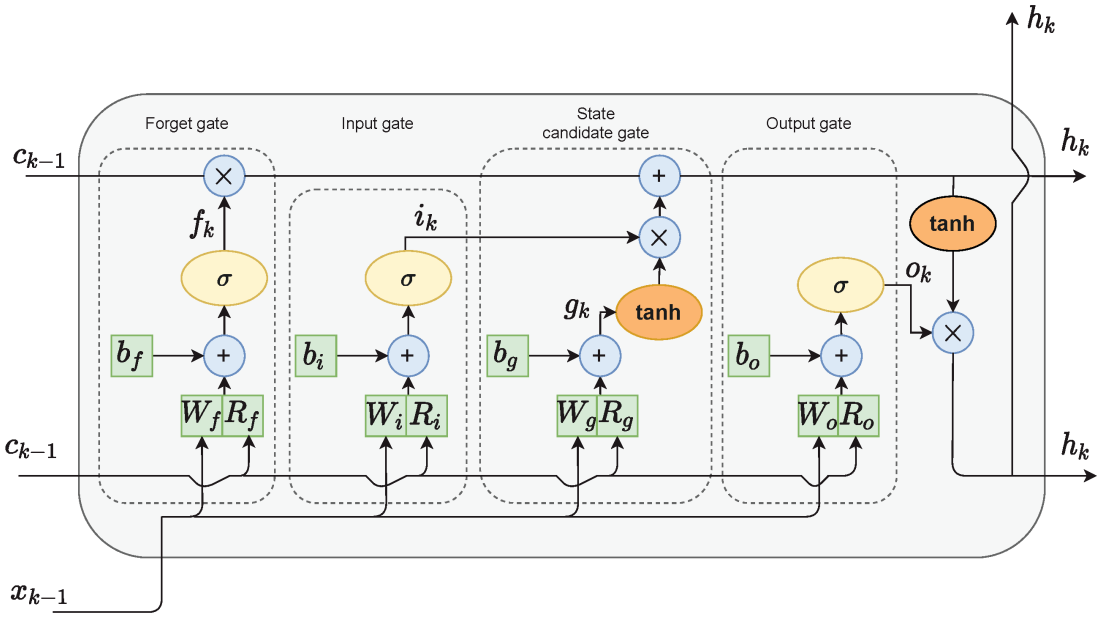

FIG. II.3 – Illustration d'un modèle LSTM. [55]

1. **Porte d'oubli :** Cette porte décide quelles informations de l'état de la cellule précédente doivent être oubliées. La fonction sigmoïde produit des valeurs entre 0 et 1, où 0 signifie "oublier complètement" et 1 signifie "conserver complètement" [18].

2. **Porte d'entrée :** Cette porte contrôle quelles nouvelles informations doivent être ajoutées à l'état de la cellule. La fonction sigmoïde détermine l'importance des nouvelles informations [18].

3. **Porte de sortie :** Cette porte décide quelles informations de l'état de la cellule doivent être sorties et utilisées comme état caché actuel. La fonction sigmoïde régule la sortie, et la fonction tangente hyperbolique applique une transformation non linéaire à l'état de la cellule [18].

II.4.3 Support Vector Machine (SVM) [13].

Les SVM sont une classe d'algorithmes d'apprentissage supervisé utilisés pour la classification (voir sous-sous-section II.5.1.3) et sont particulièrement efficaces pour les problèmes de classification binaire, mais ils peuvent également être utilisés pour la régression.

L'objectif des SVM est de trouver un hyperplan séparateur optimal entre les différentes classes d'un ensemble de données. Cet hyperplan est celui qui maximise la marge,

c'est-à-dire la distance aux points des données les plus proches de chaque classe. Ces points particuliers sont appelés les vecteurs de support (figure II.4).

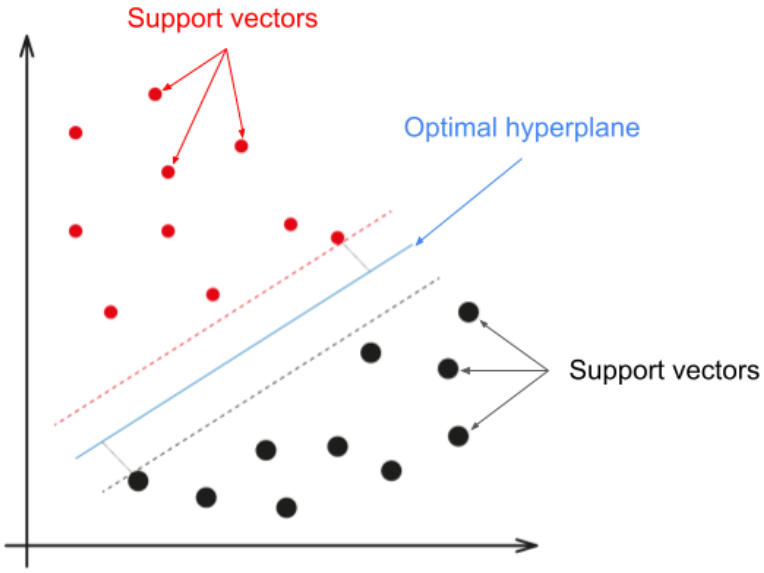

FIG. II.4 – Exemple d'un hyperplan optimal du SVM. [13]

II.4.3.1 Données séparables linéairement

Mathématiquement, pour des données linéairement séparables, l'hyperplan optimal est défini par :

$$\mathbf{w}^\top \mathbf{x} + b = 0 \tag{II.3}$$

Avec :

- $\mathbf{w}$ est le vecteur normal à l'hyperplan (vecteur de poids),
- $\mathbf{x}$ est le vecteur d'entrée (caractéristiques ou données),
- b est le terme de biais.

Lorsque nous évaluons un nouvel échantillon $\mathbf{x}_i$, nous calculons $\mathbf{w}^\top \mathbf{x}_i + b$.

Si $\mathbf{w}^\top \mathbf{x}_i + b > 0$, alors $\mathbf{x}_i$ est classé comme appartenant à la classe positive.

Si $\mathbf{w}^\top \mathbf{x}_i + b < 0$, alors $\mathbf{x}_i$ est classé comme appartenant à la classe négative.

II.4.3.2 Données non-séparables linéairement

Cependant, pour des données non séparables linéairement, il est nécessaire d'introduire l'utilisation des noyaux (*Kernels*).

Un *Kernel* est une fonction K qui prend deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ en entrée et retourne leur produit scalaire dans un espace de redescription de dimension potentiellement infinie (figure II.5). (Un espace de redescription est un espace abstrait où les données peuvent être représentées de manière différente, souvent pour faciliter l'apprentissage ou l'interprétation. Dans le contexte des noyaux, cela implique de projeter les données d'un espace de dimension finie $\mathbb{R}^n$ vers un espace potentiellement de dimension infinie.)

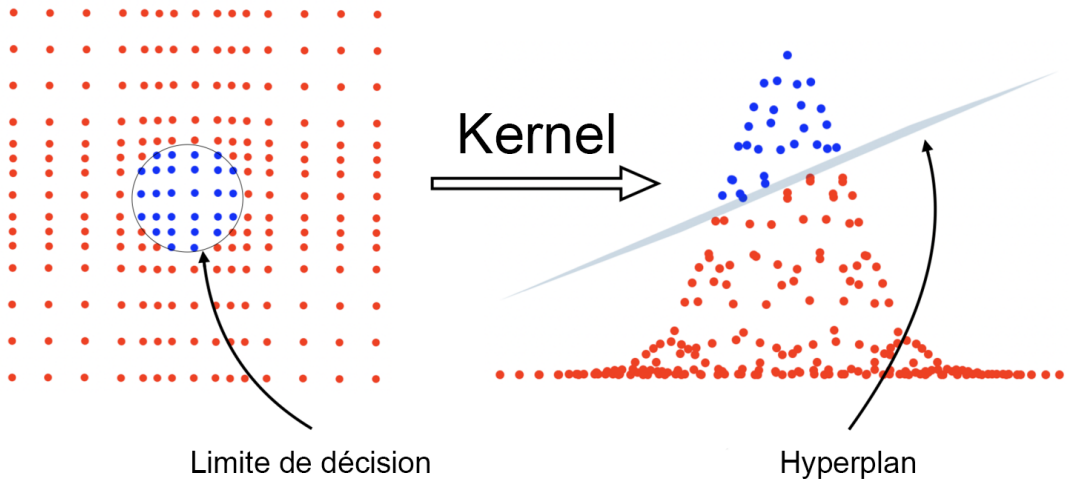

FIG. II.5 – Illustration montrant comment le Kernel facilite la séparation de données non-separables linéairement. [17]

Quelques noyaux couramment utilisés sont :

1. **Noyau linéaire** : Calcule le produit scalaire entre les vecteurs x et y.

$$K(x, y) = x^\top y \tag{II.4}$$

où :

— $K(x, y)$ est le noyau linéaire qui calcule le produit scalaire entre les vecteurs x et y.
— x et y sont des vecteurs de données.

2. **Noyau polynomial** : Calcule un polynôme de degré d en fonction du produit scalaire entre x et y, pondéré par γ et r.

$$K(x, y) = (\gamma x^\top y + r)^d \tag{II.5}$$

où :

— $K(x, y)$ est le noyau polynomial avec des paramètres γ, r, et un degré d.
— γ est un paramètre de mise à l'échelle du produit scalaire entre les vecteurs x et y.
— r est un terme de décalage.
— d est le degré du polynôme.

3. **Noyau gaussien RBF** : Calcule une fonction exponentielle décroissante de la distance euclidienne entre x et y, pondérée par γ.

$$K(x, y) = \exp(-\gamma \|x - y\|^2) \tag{II.6}$$

où :

— $K(x, y)$ est le noyau gaussien RBF avec paramètre γ.
— γ est un paramètre de mise à l'échelle de la distance euclidienne entre x et y.
— $\|x - y\|$ représente la norme euclidienne ou la distance euclidienne entre les vecteurs x et y.

Le choix du noyau dépend du problème et des propriétés souhaitées (lissage, invariances, etc.).

En appliquant celui-ci aux SVM, il est possible de transformer les données dans un espace de caractéristiques où elles deviennent linéairement séparables, permettant ainsi au modèle de capturer des relations complexes et non linéaires dans les données d'origine. Cette approche améliore la flexibilité et la puissance des SVM, rendant possible la classification de données qui ne pourraient pas être correctement séparées dans leur espace d'origine.

II.4.4 Random Forest

Le Random Forest est un autre algorithme d'apprentissage supervisé utilisé généralement pour la classification de données (sous-sous-section II.5.1.3). Cet algorithme est constitué de multiples arbres de décision (Decision Trees), chacun étant construit à partir d'un sous-ensemble aléatoire des données d'entraînement (bootstrap). Le RF fonctionne en combinant les prédictions de ses arbres pour obtenir une prédiction finale plus précise (figure II.6) [8].

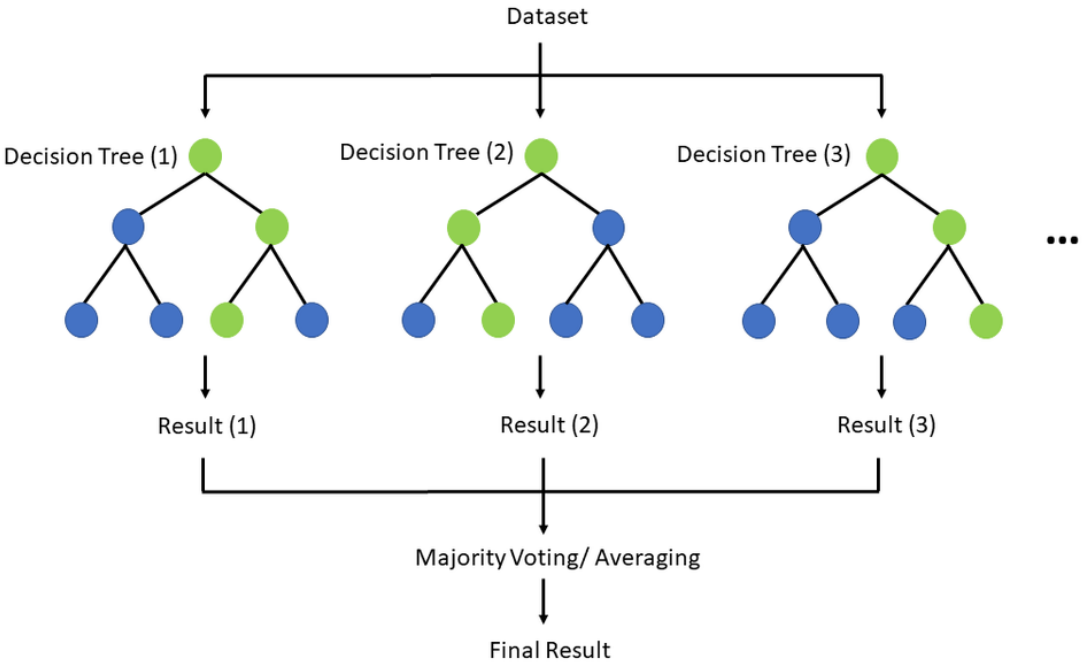

Fig. II.6 – Exemple du Random Forest. [49]

Bien que le RF soit lui-même composé d'arbres de décisions, ces deux concepts présentent de nombreuses différences dans leur comportement. La figure (figure II.7) montre une comparaison entre les arbres de décisions et la Random Forest.

Decision trees	**Random Forest**
1. Les arbres de décision souffrent généralement du problème de surapprentissage s'ils sont autorisés à croître sans aucun contrôle.	1. Les forêts aléatoires utilisent des sous-ensembles de données pour produire une sortie basée sur la moyenne ou le classement majoritaire, évitant ainsi le surapprentissage.
2. Un seul arbre de décision est plus rapide en termes de calcul.	2. Elles sont comparativement plus lentes.
3. Lorsqu'un ensemble de données est pris en entrée par un arbre de décision, il formule des règles pour faire des prédictions.	3. La forêt aléatoire sélectionne aléatoirement des observations, construit un arbre de décision, et prend la moyenne des résultats sans utiliser de formules.

Fig. II.7 – Comparaison entre les arbres de décisions et la Random Forest.

II.4.4.1 Fonctionnement de la Random Forest [8]

Construction des arbres : Afin d'introduire la diversité parmi les arbres, chaque arbre dans une forêt aléatoire est construit à partir d'un échantillon bootstrap de l'ensemble de données d'entraînement.

1. **Feature Bagging :** La Feature Bagging est une technique permettant de sélectionner à chaque nœud de décision, lors de la construction de chaque arbre, un sous-ensemble aléatoire de variables. Parmi ce sous-ensemble, la variable qui offre la meilleure séparation des données est choisie pour diviser le nœud.

2. **Prédiction :** Pour faire une prédiction avec une forêt aléatoire, chaque arbre de la forêt donne une prédiction individuelle. Pour les problèmes de classification, la prédiction finale est déterminée par un vote majoritaire parmi les arbres. Pour les problèmes de régression, la prédiction finale est la moyenne des prédictions individuelles des arbres.

3. **Importance des variables :** Les Random Forests fournissent également une mesure de l'importance des variables. Cette importance est calculée en observant la diminution de l'exactitude de la prédiction lorsque les valeurs d'une variable sont permutées de manière aléatoire. Une grande diminution de l'exactitude indique que la variable est importante pour le modèle.

II.5 Méthodes du Machine Learning

Le ML englobe trois approches différentes : l'apprentissage supervisé, l'apprentissage non supervisé, et l'apprentissage par renforcement .

II.5.1 Apprentissage supervisé

L'apprentissage supervisé est une méthode où le modèle est formé à partir de données étiquetées, c'est-à-dire que chaque entrée est associée à une sortie correspondante.

Dans cette approche, le modèle apprend à partir de ces paires entrée-sortie, en ajustant ses paramètres pour minimiser l'erreur entre ses prédictions et les réponses correctes. Ce processus d'ajustement est souvent réalisé à l'aide d'algorithmes d'optimisation tels que la descente de gradient [11].

II.5.1.1 Définition et algorithme de la descente de gradient [35]

La descente de gradient est une méthode itérative qui vise à trouver le minimum d'une fonction de coût en ajustant ses paramètres dans la direction opposée à son gradient (figure II.8). Le gradient d'une fonction en un point donné est un vecteur qui indique la direction de la plus grande augmentation de cette fonction. En suivant la direction opposée, on se déplace vers le minimum de cette dernière.

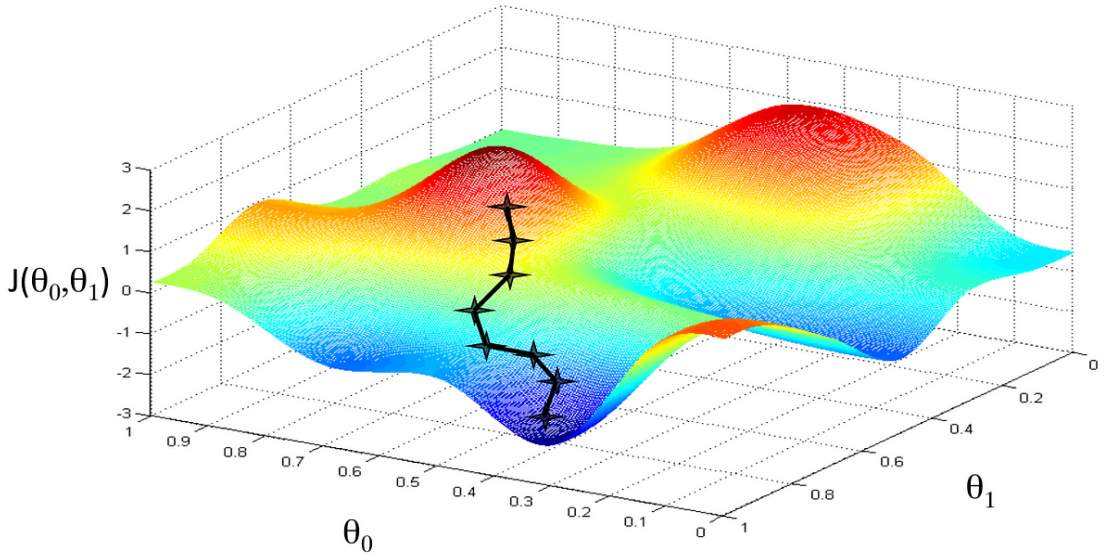

FIG. II.8 – Illustration de la descente de gradient.

Ci-dessous l'algorithme de la descente de gradient :

1. **Initialisation :** Choisir un point de départ initial pour les paramètres que l'on souhaite optimiser.

2. **Calcul du Gradient :** Calculer le gradient de la fonction objectif par rapport aux paramètres au point actuel.

3. **Mise à jour des paramètres :** Mettre à jour les paramètres en se déplaçant dans la direction opposée du gradient. La mise à jour est généralement effectuée selon la formule :

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t) \tag{II.7}$$

où

— θ_t représente les paramètres à l'itération t,
— η est le taux d'apprentissage (Learning Rate),
— $\nabla f(\theta_t)$ est le gradient de la fonction objectif f au point θ_t.

4. **Répétition :** Répéter les étapes 2 et 3 jusqu'à ce que la convergence soit atteinte, c'est-à-dire jusqu'à ce que les changements dans la fonction objectif deviennent négligeables ou qu'un nombre maximal d'itérations soit atteint.

La descente de gradient est largement utilisée en ML pour entraîner des modèles des deux techniques clés de l'apprentissage supervisé, à savoir :

II.5.1.2 La Régression

La régression est un algorithme d'apprentissage supervisé utilisé pour prédire une valeur continue. Son but est de modéliser la relation entre une variable dépendante (la sortie) et une ou plusieurs variables indépendantes (les entrées) (figure II.9).

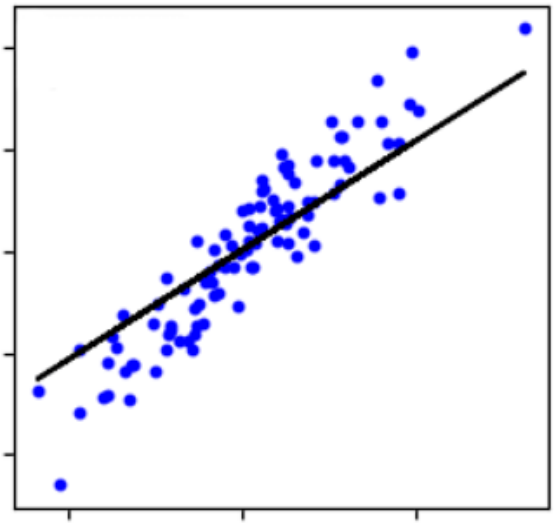

Fig. II.9 – Illustration montrant la régression

Parmi les modèles de régression les plus importants, on retrouve la régression linéaire, la régression polynomiale, ainsi que la régression logistique [11].

II.5.1.3 La Classification

La classification est une méthode d'apprentissage supervisé utilisée pour prédire des catégories auxquelles des données peuvent appartenir, en fonction de caractéristiques

observées (figure II.10). L'objectif est de construire un modèle qui peut prédire correctement les étiquettes des nouvelles données non étiquetées.

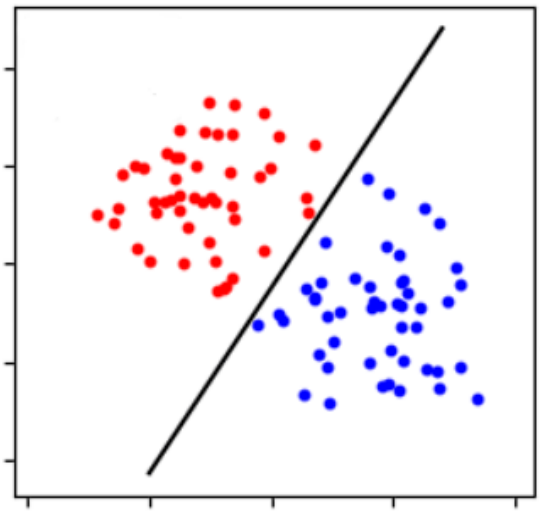

FIG. II.10 – Illustration montrant la classification.

Il existe différents modèles de classification en apprentissage supervisé, les plus connus sont : les arbres de décision, les forêts aléatoires, les machines à vecteurs de support, ainsi que les réseaux de neurones artificiels [11].

II.5.2 Apprentissage non supervisé

L'apprentissage non supervisé est une méthode d'apprentissage automatique où le modèle apprend à partir de données non étiquetées, cela veut dire que, contrairement à l'apprentissage supervisé, il n'y a pas de réponses correctes fournies pour chaque exemple d'entraînement [11]. Deux des méthodes populaires utilisées dans l'apprentissage non supervisé sont :

II.5.2.1 Clustering

Le clustering est une technique d'apprentissage non supervisé qui consiste à regrouper des données en groupes (clusters), de sorte que les objets dans le même cluster soient plus similaires entre eux qu'avec ceux des autres clusters (figure II.11). Cette technique est particulièrement utile pour explorer des données et identifier des structures cachées sans utiliser de labels [11].

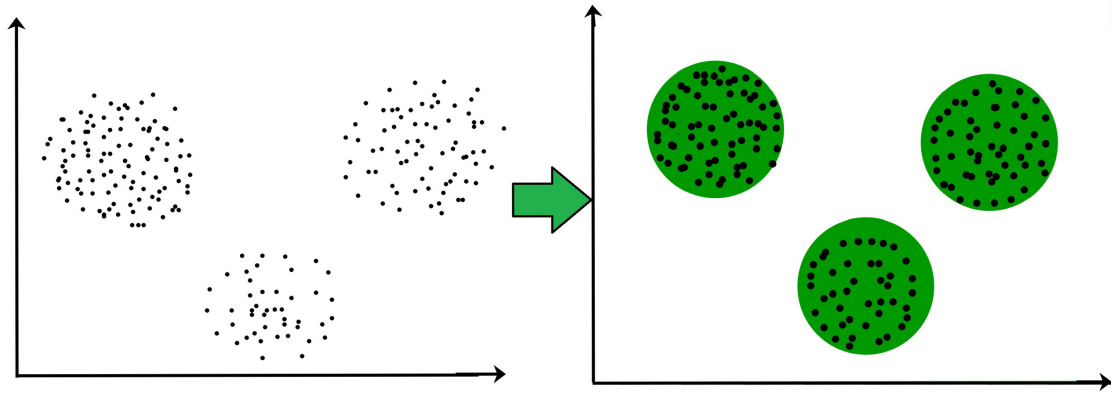

FIG. II.11 – Exemple du Clustering. [15]

II.5.2.2 Détection d'anomalies

La détection d'anomalies, une autre technique d'apprentissage non supervisé, est cruciale pour identifier les comportements qui s'écartent du comportement normal dans un ensemble de données (figure II.12). En effet, elle permet de détecter des événements rares mais significatifs, tels que les fraudes, les défaillances de systèmes ou les intrusions réseau [11][24].

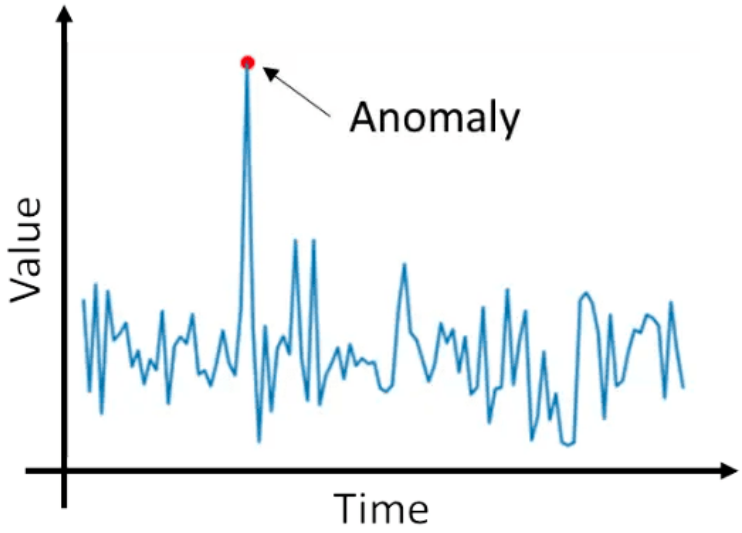

FIG. II.12 – Exemple d'anomalie. [25]

II.5.3 Apprentissage par renforcement

L'apprentissage par renforcement (RL) est une méthode d'apprentissage automatique où un agent apprend à prendre des décisions en interagissant avec un environnement. L'objectif ici est de maximiser une récompense cumulative en choisissant des actions qui influencent l'état de l'environnement.

Contrairement aux apprentissages supervisés et non supervisés, le RL repose sur un processus d'essais et d'erreurs, ou l'agent reçoit des récompenses ou des punitions en fonction des actions qu'il entreprend, ce qui l'aide à ajuster ses stratégies pour améliorer ses performances au fil du temps [11].

Bien que les méthodes précédemment abordées soient efficaces, elles présentent certaines limitations lorsqu'il s'agit de traiter des cas complexes, tels que la reconnaissance d'images, la reconnaissance vocale, et le traitement du langage naturel. C'est à ce niveau que les réseaux de neurones interviennent, offrant des solutions avancées grâce à leur capacité à apprendre à partir de données complexes.

II.6 Processus de fonctionnement d'un réseau de neurones artificiels

Un ANN est composé d'unités (neurones artificiels), organisées en couches : une couche d'entrée, une couche cachée, et une couche de sortie. Pour fonctionner correctement, chaque neurone doit recevoir des signaux d'entrée, les pondérer, appliquer une fonction d'activation non linéaire, et transmettre le résultat aux neurones de la couche suivante. Cette première étape est appelée la propagation avant [27].

II.6.1 Propagation avant [41]

1. **Réception des signaux d'entrée :** Les signaux d'entrée sont en général des valeurs numériques représentant diverses caractéristiques des données d'entrée. Par exemple, dans un réseau de neurones utilisé pour la reconnaissance d'images, les signaux d'entrée pourraient être les valeurs des pixels de l'image.

2. **Pondération des signaux d'entrée :** Les signaux d'entrée reçus par un neurone sont ajustés en fonction de leur importance grâce à des poids synaptiques. Mathématiquement, si un neurone reçoit des signaux d'entrée $x_1, x_2, \ldots, x_n$, et que les poids correspondants sont $w_1, w_2, \ldots, w_n$, alors le neurone calcule une somme pondérée de ces entrées :

$$z = \sum_{i=1}^{n} w_i x_i \tag{II.8}$$

Où b est un biais, un autre paramètre ajustable qui permet de décaler la fonction d'activation.

3. **Application de la fonction d'activation non linéaire :** Une fois que le neurone a calculé la somme pondérée des entrées, il applique une fonction d'activation non linéaire à cette somme. La fonction d'activation introduit de la non-linéarité dans le modèle, ce qui permet au réseau de neurones de modéliser des relations complexes dans les données. Les fonctions d'activation couramment utilisées incluent la fonction sigmoïde, la fonction tanh (tangente hyperbolique), et la fonction ReLU (Rectified Linear Unit). Par exemple, pour la fonction ReLU, l'activation est définie comme :

$$a = \max(0, z) \tag{II.9}$$

4. **Transmission du résultat aux neurones de la couche suivante :** Le résultat de la fonction d'activation, appelé l'activation du neurone a, est par la suite transmis aux neurones de la couche suivante.

Ce processus se répète pour chaque neurone dans chaque couche du réseau, permettant ainsi au réseau de transformer petit à petit les données d'entrée de manière hiérarchique et de plus en plus abstraite à chaque couche [27].

Maintenant que nous savons comment les signaux d'entrée sont transformés dans chaque neurone, il est crucial de comprendre comment les poids et les biais du réseau de neurones sont ajustés afin d'optimiser les performances du réseau. C'est là que la rétro-propagation entre en jeu. En analysant les erreurs entre les sorties attendues et les sorties réelles du réseau, la rétro-propagation calcule les ajustements nécessaires pour les poids et les biais à travers les différentes couches du réseau. Ce processus itératif permet au réseau de s'adapter et d'améliorer ses prédictions au fil du temps, rendant ainsi son apprentissage plus efficace et précis. Examinons celà de plus près [27].

II.6.2 Rétropropagation [41]

Calcul de l'erreur : Après la propagation avant, la sortie du réseau est comparée à la valeur cible pour calculer l'erreur (ou la perte).

Propagation de l'erreur en arrière : L'erreur est propagée en arrière à travers le réseau,

en commençant par la couche de sortie et en remontant jusqu'à la couche d'entrée. Pour chaque neurone, on calcule sa contribution à l'erreur totale.

Mise à jour des poids et des biais : Les poids et les biais sont ajustés en fonction de l'erreur calculée. Cela se fait en utilisant l'algorithme de descente de gradient, qui met à jour les paramètres pour minimiser l'erreur. La mise à jour des poids wi et des biais b se fait selon les formules :

$$w_i \leftarrow w_i - \eta \frac{\partial L}{\partial w_i} \tag{II.10}$$

$$b \leftarrow b - \eta \frac{\partial L}{\partial b} \tag{II.11}$$

Ici,

— η est le taux d'apprentissage (Learning Rate). Il sert à contrôler la taille des pas que l'algorithme d'optimisation effectue lors de la mise à jour des poids du réseau de neurones pendant l'entraînement.

— L est la fonction de perte. Elle est définie sous la formule (équation II.12) pour la régression et (équation II.13) pour la classification.

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2 \tag{II.12}$$

$$L(y, p) = -(y \log(p) + (1 - y) \log(1 - p)) \tag{II.13}$$

Où,

— y : Valeur réelle de la cible
— $\hat{y}$: Prédiction du modèle pour la variable cible
— p : Probabilité prédite par le modèle pour une classe spécifique

II.7 Deep Learning

Comme mentionné précédemment, le deep learning (DL) est une branche de l'intelligence artificielle (IA) qui se concentre sur l'apprentissage et l'entraînement de réseaux

de neurones artificiels profonds (DNN). Contrairement aux réseaux de neurones traditionnels, les DNN sont caractérisés par leur architecture multicouche, d'où le terme 'profond'.

II.7.1 Réseaux de neurones profonds (DNN)

Les réseaux de neurones profonds (DNN) sont des ANN avec plusieurs couches cachées entre l'entrée et la sortie (figure II.13). Ces couches supplémentaires permettent au réseau de modéliser des relations plus complexes dans les données. Grâce à des ar-

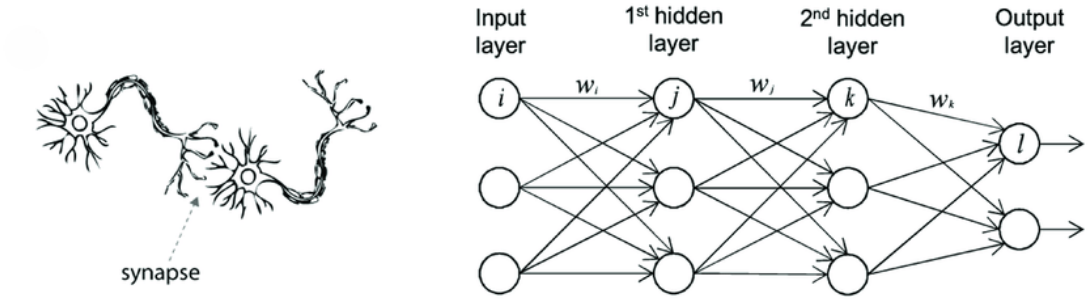

Fig. II.13 – Chaîne de neurones biologiques naturelle vs réseau de neurones artificiels profonds. [32]

chitectures de réseaux de neurones multicouches, qui permettent de capturer des représentations hiérarchiques des données, le DL permet de traiter des volumes massifs de données non structurées, comme des images, des sons, et des textes, en extrayant automatiquement des caractéristiques pertinentes [27].

II.7.2 Architectures des DNN en Deep Learning

II.7.2.1 Réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutifs (CNN) sont une architecture de DNN conçue spécialement pour traiter des données structurées en matrice, comme les images.

Ils utilisent des couches de convolution pour extraire des caractéristiques locales des données en appliquant des filtres de convolution qui viennent s'appliquer sur l'entrée. Ces filtres permettent de détecter des motifs tels que des bords, des textures et des formes.

Les CNN sont largement utilisés dans la vision par ordinateur pour des tâches comme la classification d'images, la détection d'objets et la segmentation d'images [5].

II.7.2.2 Réseaux de neurones récurrents (RNN) [44]

Les réseaux de neurones récurrents (RNN) sont une classe de réseaux de neurones profonds particulièrement adaptés pour traiter des données séquentielles ou temporelles. Contrairement aux réseaux de neurones traditionnels où toutes les entrées et sorties sont indépendantes les unes des autres, les RNN possèdent des connexions récurrentes qui leur permettent de maintenir une mémoire interne de l'information passée (figure II.14), ce qui est crucial pour des tâches telles que la reconnaissance vocale, la traduction automatique, et la modélisation de séries temporelles.

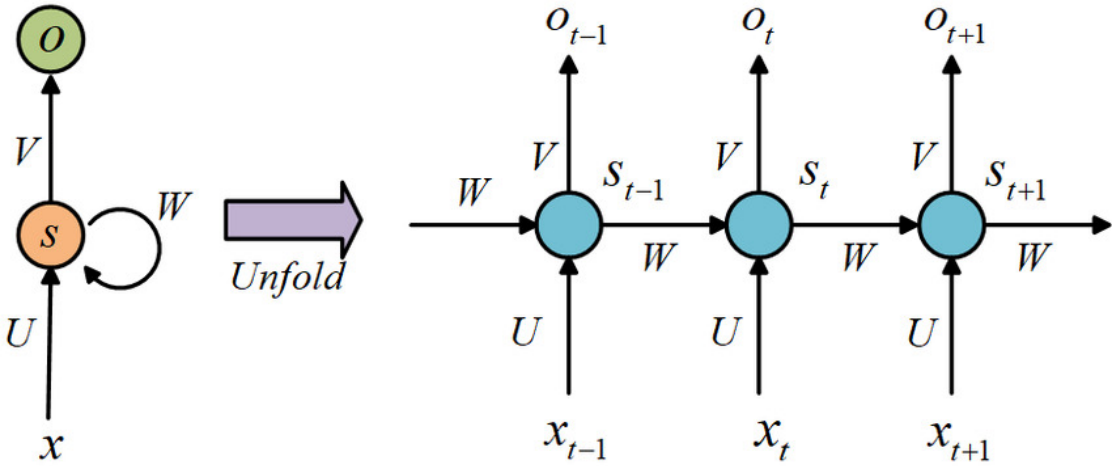

Fig. II.14 – Illustration d'un réseau de neurones récurrents. [57]

L'algorithme des RNN est le suivant :

a) Entrée actuelle (x_t) : La donnée d'entrée à l'étape temporelle t. Par exemple, dans une séquence de données financières, x_t pourrait être le prix d'une action à l'étape t (actuel).

b) État caché précédent (h_{t-1}) : L'état caché du réseau à l'étape temporelle $t-1$. Cet état inclut les informations apprises par le réseau jusqu'à l'étape $t-1$.

c) État caché (h_t) : Similaire à l'état caché précédent, mais à l'instant t.

d) Sortie prédite (y_t) : La sortie générée par le réseau à l'étape temporelle t.

e) Matrices de poids (W, U, V) :

- W : Matrice de poids qui connecte l'entrée actuelle x_t à l'état caché actuel h_t.

- U : Matrice de poids qui connecte l'état caché précédent h_{t-1} à l'état caché actuel h_t.
- V : Matrice de poids de la couche de sortie.

f) Biais :

- b : Vecteur de biais ajouté pour ajuster les valeurs de l'état caché.
- c : Vecteur de biais pour la sortie.

g) Fonctions d'activation :

- f et g : Des fonctions d'activation (généralement non linéaires, afin de pouvoir modéliser des relations complexes).

Dans un RNN, à chaque étape temporelle t, l'état caché h_t est mis à jour en fonction de l'entrée actuelle x_t et de l'état caché précédent h_{t-1}. La mise à jour se fait généralement via une fonction non linéaire telle que la tangente hyperbolique (tanh) ou la fonction sigmoïde (σ).

La formule de mise à jour est la suivante :

$$h_t = f(W \cdot x_t + U \cdot h_{t-1} + b) \tag{II.14}$$

avec

$$f = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad \text{ou bien} \quad f = \sigma(x) = \frac{1}{1 + e^{-x}} \tag{II.15}$$

D'après la formule *(2.14)*, on peut directement distinguer que l'état caché h_t dépend à la fois de l'entrée actuelle x_t (via $W \cdot x_t$) et de l'état précédent h_{t-1} (via $U \cdot h_{t-1}$), permettant ainsi de capturer les dépendances à long terme.

La sortie y_t est ensuite calculée à partir de l'état caché h_t :

$$y_t = g(V \cdot h_t + c) \tag{II.16}$$

La formule *(2.16)* permet au RNN de calculer la valeur de sortie en transformant l'état caché h_t via une combinaison linéaire $V \cdot h_t$ et une fonction d'activation g, produisant ainsi la prédiction du réseau à l'instant t.

II.8 Conclusions

Dans ce chapitre, nous avons vu ce qu'est l'intelligence artificielle, en explorant ses origines et son évolution jusqu'à l'émergence de l'apprentissage automatique.

Nous avons examiné les différents modèles et les différentes approches de l'apprentissage, y compris les techniques supervisées telles que la régression et la classification, les méthodes non supervisées comme le clustering et la détection d'anomalies, ainsi que l'apprentissage par renforcement. Nous avons également plongé dans le domaine captivant de l'apprentissage profond, en expliquant les concepts de base des réseaux de neurones et en explorant des architectures avancées telles que les CNN et les RNN.

Dans le prochain chapitre, nous développerons l'objet principal de ce projet, à savoir l'utilisation de l'intelligence artificielle dans le monde de la blockchain afin d'essayer de prédire le marché du Bitcoin.

III Essais de prévisions sur le Bitcoin : expérimentations

Sommaire du chapitre

III.1 Introduction 39

III.2 Approche détaillée 39

III.3 Étapes de réalisation du projet 40

III.3.1 Environnement et outils de travail 40

III.3.2 Ensembles de données utilisés 40

III.3.3 Méthodes de prévision et métriques de performance 45

III.3.4 Implémentation des modèles de prévision 47

III.4 Résultats et observations 49

III.4.1 Régression 49

III.4.2 Classification 51

III.5 Discussions 53

III.6 Conclusions 54

III.1 Introduction

Le marché des cryptomonnaies, en particulier celui du Bitcoin, est connu pour sa volatilité et sa complexité, le rendant donc difficile à prévoir avec précision. Les investisseurs peuvent se retrouver avec de grosses pertes d'argent en raison de cette volatilité imprévisible, faisant de leurs investissements un véritable jeu de hasard, où les mouvements de prix peuvent sembler aléatoires et incontrôlables. Même les experts du domaine ne peuvent garantir des gains, et il est fréquent de les voir perdre des sommes considérables en très peu de temps.

Dans ce chapitre, nous expliquons comment nous pouvons utiliser l'intelligence artificielle pour essayer de prévoir le marché du BTC, et proposer des solutions aux investisseurs.

III.2 Approche détaillée

Notre étude se concentre sur l'application de divers modèles d'IA. Etant donné la complexité et la rapidité des fluctuations des prix des cryptomonnaies, nous avons exploré différentes approches pour tester les prévisions. En premier lieu, nous avons utilisé deux techniques de régression, la régression linéaire et le LSTM, pour la prévision des valeurs futures. Ensuite, nous avons appliqué indépendamment de la régression, des méthodes de classification, telles les SVM et les forêts aléatoires, pour catégoriser les mouvements de prix futurs en termes de ”hausse” ou ”baisse”.

Nous avons également pris en compte la diversité des jeux de données afin de mieux comprendre les facteurs influençant le phénomène étudié, car en réalité, différentes variables extérieures peuvent influencer les fluctuations du marché du Bitcoin. En effet, dans l'article “Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting” [14], les auteurs ont examiné comment les sentiments des publications exprimés sur les réseaux sociaux peuvent être intégrés dans les modèles de prévision pour améliorer leur précision. En utilisant des techniques de traitement du langage naturel pour analyser les sentiments, l'étude démontre que les signaux émotionnels peuvent avoir un impact significatif sur les mouvements de prix, ajoutant ainsi une dimension précieuse aux approches traditionnelles basées uniquement sur les données historiques de marché. Dans un autre article intitulé “Explainable artificial intelligence modeling to forecast bitcoin prices” [16], les auteurs (John W. Goodell, et al.) ont essayé de démontrer qu'il existe une corrélation entre le prix du Bitcoin et les prix des différents autres marchés, tel que la valeur du S&P 500 (Standard & Poor's 500).

Notre objectif principal est donc la détermination des combinaisons de modèles et de données offrant les meilleures prévisions.

III.3 Étapes de réalisation du projet

Notre projet se déroule en 4 parties :

1. Le choix de l'environnement et de l'outil de travail,
2. Le choix et la collecte des données,
3. Le choix des approches de prévision et des métriques de performance,
4. L'implémentation des méthodes de prévision.

III.3.1 Environnement et outils de travail

Comme langage de programmation, nous avons utilisé Python, notamment à cause de sa puissance et sa polyvalence. Python est largement utilisé dans la science des données et l'apprentissage automatique. Il dispose d'une riche bibliothèque de modules, tels que Pandas pour la manipulation des données, NumPy pour les calculs numériques, et Scikit-learn pour le machine learning, ce qui en fait un choix parfait pour ce type de projet.

Pour les modèles du Machine Learning, nous avons utilisé la bibliothèque open-source scikit-learn, réputée pour sa robustesse et sa documentation complète. Elle offre une large gamme d'algorithmes de machine learning pour des tâches telles que la régression et la classification. Elle permet également de prétraiter facilement les données, d'évaluer les modèles et d'automatiser les flux de travail via des pipelines. Son intégration fluide avec d'autres outils Python comme NumPy, Pandas et Matplotlib, combinée à une vaste communauté de soutien, en fait un choix idéal pour développer des solutions de machine learning de manière efficace et rapide.

Tandis que pour les modèles du Deep Learning, nous avons opté pour TensorFlow en raison de sa capacité à gérer des calculs à grande échelle et de sa puissance dans l'entraînement de réseaux de neurones profonds. TensorFlow se distingue par sa flexibilité et son architecture modulaire, qui permettent de concevoir et de déployer facilement des modèles complexes. De plus, sa compatibilité avec les GPU améliore considérablement les performances, rendant ainsi cet outil idéal pour des applications d'apprentissage profond nécessitant une haute performance et une optimisation maximale.

Enfin, pour l'environnement de travail, nous avons choisi Jupyter Notebook. Ce dernier est idéal pour le développement interactif, car il permet d'écrire et d'exécuter du code par cellules, facilitant ainsi les essais et les ajustements en temps réel.

III.3.2 Ensembles de données utilisés

Quatres ensembles de données ont été utilisés. Afin de permettre une meilleure compréhension, nous allons les dénoter comme suit (tableau III.1) :

Ensembles de données	Dénotation
Ensemble n°1	D1
Ensemble n°2	D2
Ensemble n°3	D3
Ensemble n°4	D4

Tab. III.1 – Dénotation et classification des ensembles de données

III.3.2.1 Données utilisées pour la régression

Pour la régression, nous avons utilisé deux jeux de données différents ($D1$ et $D2$).

$D1$ est un ensemble de données contenant les informations du marché du Bitcoin quotidiennes, comprenant les valeurs d'ouverture et de clôture ainsi que les valeurs minimales et maximales, pour chaque journée. $D2$ quant à lui ne contient que les prix de clôture, et ce pour chaque heure. Ici, nous avons intentionnellement choisi une fréquence horaire afin de tester des prévisions à court terme (tableau III.2).

III.3.2.2 Données utilisées pour la classification

Pour la classification, trois datasets ont été utilisés ($D1$, $D3$ et $D4$).

La première est la même que celle utilisée pour la régression ($D1$). Dans la seconde ($D3$), nous avons intégré une variable extérieure à l'ensemble $D1$: le score des sentiments des articles parlant du Bitcoin. La troisième ($D4$), quant à elle, contient le prix des autres marchés (pétrole, gaz, Nasdaq, S&P 500) en plus de celui du BTC. Toutes ces données représentent des valeurs quotidiennes (tableau III.2).

Ensemble de données	Régression	Classification	Fréquence
D1	x	x	Quotidienne
D2	x	-	Horaire
D3	-	x	Quotidienne
D4	-	x	Quotidienne

Tab. III.2 – Dataset, méthode et fréquence

III.3.2.3 Collectes et prétraitement de données

III.3.2.3.1 Régression

Le premier jeu de données ($D1$) est disponible dans différents sites de trading et est téléchargeable gratuitement. Dans notre cas, nous avons utilisé Yahoo Finance, une plateforme en ligne qui offre une gamme complète d'informations financières, d'outils et de ressources pour aider les utilisateurs à gérer leurs investissements. Ce jeu de données s'étend sur une période allant de 2014 à 2024.

Le second ensemble de données ($D2$) diffère du premier. D'après nos recherches, aucune source ne propose les prix horaires du Bitcoin de 2024, nous avons donc utilisé la plateforme Bitget. Bitget est une plateforme d'échange de crypto-monnaies qui permet aux utilisateurs d'acheter, de vendre, et de trader divers actifs numériques. Cette plateforme offre également la possibilité de télécharger l'historique des transactions pour chacun de ces actifs, y compris celle du Bitcoin [6]. L'historique en soi contient le prix du BTC au moment de chaque transaction, ainsi que le volume de celle-ci. Il est important de noter ici que, des milliers de transactions se déroulent chaque seconde, ce qui nous permettra d'avoir une actualisation fréquente du prix du Bitcoin.

Cependant, une contrainte majeure se présente lors du téléchargement de ces fichiers d'historique. En effet, ce dernier nécessite une acquisition manuelle de chaque fichier pour chaque journée (figure III.1), ce qui entraîne une perte de temps très sérieuse. Afin de remédier à cela, nous avons automatisé le processus en utilisant la bibliothèque Selenium, conçue pour l'extraction de données ainsi que l'automatisation des interactions avec les navigateurs web [43].

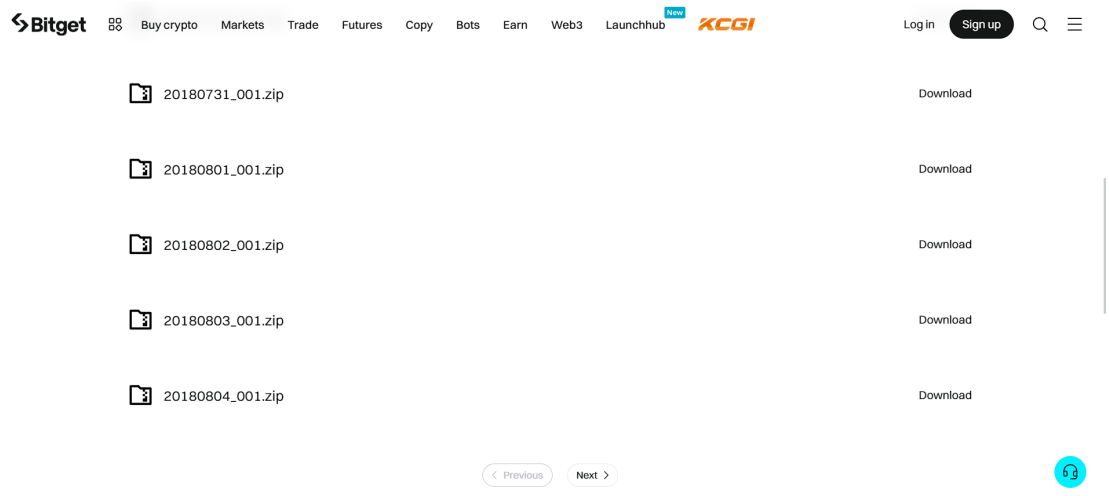

Fig. III.1 – Fichiers contenant l'historique des transactions de chaque journée dans la blockchain Bitcoin. [6]

Enfin, nous avons fusionné tous les fichiers en un seul jeu de données, et pour chaque fréquence horaire, nous avons pris la valeur la plus récente du BTC. Ces étapes ont été réalisées via la bibliothèque Pandas, conçue pour la manipulation de données.

III.3.2.3.2 Classification

Le premier ensemble est le même que celui de la régression ($D1$).

Pour le second ($D3$), nous avons collecté les descriptions des articles de presse concernant le Bitcoin. L'objectif est d'attribuer un score de sentiment à ces descriptions, puis d'intégrer ce score avec le prix du Bitcoin. Lors de nos recherches, nous avons trouvé un ensemble de données dans le site web Kaggle contenant les descriptions des articles de presse parlant du Bitcoin, ainsi que le score de sentiments de ces derniers [39], cependant, nous avons rencontré deux problèmes :

La méthode qu'ils ont utilisé pour l'analyse des sentiments n'est pas assez fiable ; cette méthode, appelée "TextBlob", repose sur des techniques de traitement du langage naturel qui peuvent, bien que souvent efficaces pour des analyses basiques, manquer de précision lorsqu'il s'agit d'interpréter les sentiments de domaines complexes, telle que la finance. Pour y remédier, nous avons utilisé un autre modèle de traitement du langage naturel, ce modèle est basé sur l'architecture BERT (Bidirectional Encoder Representations from Transformers), et est entraîné sur une large quantité de données du domaine de la finance. BERT est un modèle de transformateur pré-entraîné sur de grandes quantités de texte et adapté pour des tâches spécifiques comme l'analyse des sentiments.

Le jeu de données contient les descriptions d'articles couvrant la période de 2021 à 2023 et n'inclut donc pas ceux de 2024. Pour intégrer ces derniers, nous avons choisi comme source Yahoo Finance et Binance, notamment à cause de leur réputation, et utilisé de nouveau l'outil Selenium pour extraire les descriptions.

Nous avons par la suite attribué pour chaque valeur quotidienne du BTC, le score des sentiments des articles. Sachant qu'une journée pourrait avoir plusieurs articles, le score final qui lui sera attribué n'est rien d'autre que le Mean Average de tous les scores dans cette même journée.

La dernière dataset ($D4$) contient, en plus des valeurs du Bitcoin, celles des autres marchés, notamment celles du Nasdaq, du S&P 500, du cours du pétrole brut, du cours du gaz, ainsi que celles de la cryptomonnaie Ethereum. Cette dernière est déjà disponible sur le site web Kaggle [26].

III.3.2.4 Choix de l'intervalle de temps

En analysant la courbe du Bitcoin de 2014 à 2024 (figure III.2), nous avons constaté que, malgré les fluctuations au fil des années, il existe une période d'équilibre entre

2021 et 2024, qui reflète également la diversité des conditions du marché (figure III.3 et tableau III.3). Ainsi, pour améliorer la capacité de notre modèle à faire des prédictions précises sur de futures conditions de marché, nous avons choisi cet intervalle de temps.

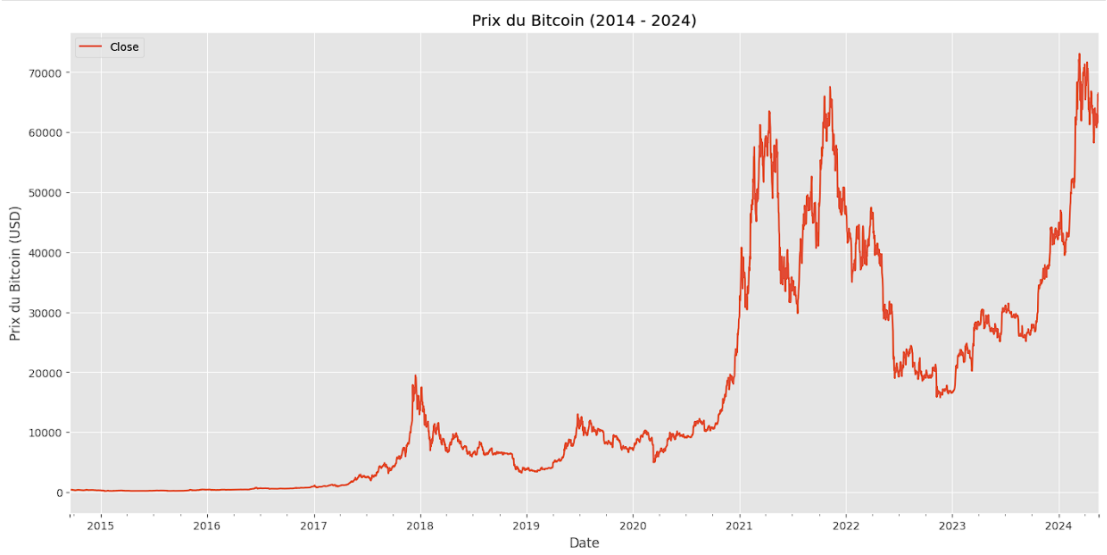

Fig. III.2 – Courbe du prix du Bitcoin en dollar (USD) entre la période de (2014 - 2024).

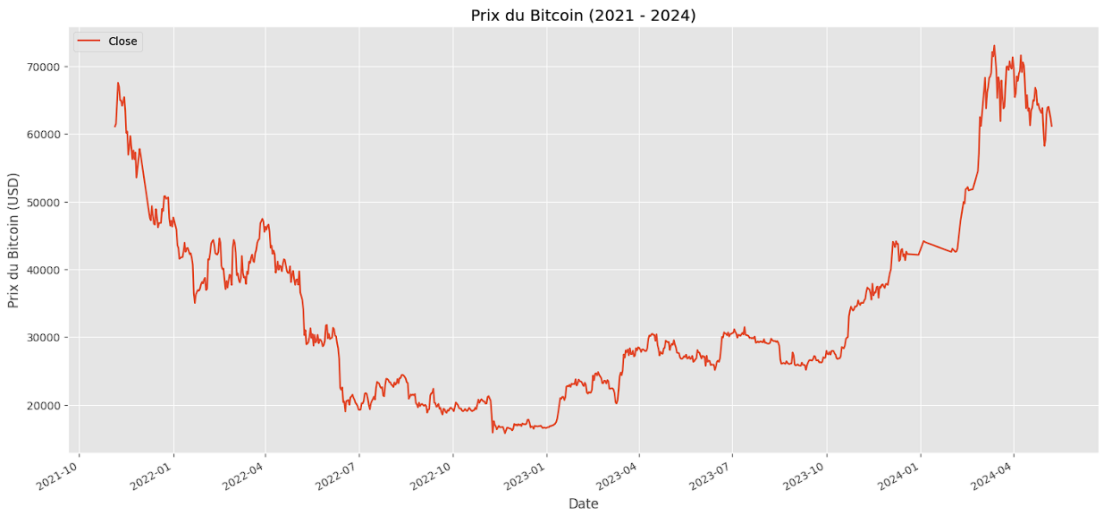

Fig. III.3 – Courbe du prix du Bitcoin en dollar (USD) entre la période de (2021 - 2024).

Nombre de hausses	Nombre de baisses
411	435

TAB. III.3 – Table démontrant l'équilibre entre les différentes phases du marché du Bitcoin dans la période de 2021 - 2024

III.3.2.5 Données finales

La table ci-dessous (tableau III.4) présente un tableau récapitulatif décrivant les versions finales des différents ensembles de données ($D1$, $D2$, $D3$, $D4$).

Ensemble de données	Variables	Fréquence	Taille (lignes, colonnes)
D1	Ouverture, Clôture, Minimal, Maximal	Quotidienne	(846, 4)
D2	Clôture	Horaire	(25980, 1)
D3	Ouverture, Clôture, Minimal, Maximal, Score des sentiments	Quotidienne	(846, 5)
D4	Clôture, Nasdaq, S&P 500, Pétrole, Gaz, Ethereum	Quotidienne	(846, 6)

TAB. III.4 – Récapitulatif des versions finales de nos jeux de données

III.3.3 Méthodes de prévision et métriques de performance

Après avoir collecté divers ensembles de données variés pour couvrir différentes conditions du marché, nous avons choisi d'utiliser plusieurs modèles de prévision. Cela nous permettra de voir les choses sous différents angles et de renforcer la fiabilité de nos analyses.

III.3.3.1 Choix des modèles de prévision

Quatre modèles ont été utilisés : la régression linéaire et le LSTM pour la régression, le Random Forest et les SVM pour la classification.

III.3.3.1.1 Régression linéaire et LSTM

La régression linéaire a été choisie pour sa simplicité et sa capacité à modéliser des relations directes entre les variables, ce qui est souvent efficace pour interpréter et analyser des données dans divers domaines.

Le choix du LSTM quant à lui revient à sa capacité à analyser les dépendances temporelles complexes dans les séries chronologiques, en utilisant des réseaux de neurones récurrents.

Utiliser deux méthodes complètement différentes nous permettra de savoir si la complexité des LSTM pourra apporter des avantages significatifs par rapport à la simplicité de la régression linéaire dans la précision et la capacité à s'adapter aux variations du marché du Bitcoin.

III.3.3.1.2 SVM et Random Forest

Les SVM sont connus pour leur capacité à trouver l'hyperplan optimal qui sépare les classes de manière maximale dans un espace de caractéristiques, ce qui est fondamental pour des tâches de classification binaire comme la prédiction de la hausse ou de la baisse des prix.

Quant aux forêts aléatoires, ces dernières sont réputées pour offrir une grande robustesse et une capacité à gérer des données complexes et bruitées. Elles sont particulièrement efficaces pour éviter le sur-apprentissage grâce à leur mécanisme de bagging et de sélection aléatoire de sous-ensembles de données, ce qui permet d'obtenir des prédictions plus stables et fiables [56].

III.3.3.2 Evaluation des modèles

Afin d'évaluer nos modèles, nous avons divisé chaque jeu de données en trois parties : entraînement, validation et test, selon le format suivant : 80% pour l'entraînement et la validation, et 20% pour le test. Les données d'entraînement sont utilisées par les modèles pour apprendre à faire des prédictions, tandis que les données de validation servent à ajuster et améliorer ces modèles. Enfin, les données de test, séparées des deux autres ensembles, permettent d'évaluer la précision des modèles sur des données qu'ils n'ont jamais vues.

Pour mesurer la précision des prédictions des prix, nous avons utilisé la moyenne de l'erreur absolue en pourcentage (MAPE) comme métrique de performance, cette dernière est assez intéressante car elle nous permet d'évaluer le pourcentage de l'erreur moyenne entre les valeurs prédites et les valeurs réelles indépendamment de l'échelle des données. Cette mesure est particulièrement utile lorsque les valeurs des données sont très élevées, ce qui est notre cas.

Tandis que pour mesurer la prévision du mouvement en termes de "hausse" ou "baisse", nous avons utilisé la métrique d'exactitude.

III.3.3.2.1 Moyenne de l'erreur absolue en pourcentage

La MAPE est calculée en prenant la moyenne des erreurs absolues en pourcentage entre les valeurs prédites et les valeurs réelles. Une MAPE plus faible indique une meilleure précision du modèle de prévision. Sa formule est la suivante :

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{III.1}$$

où :

— n représente le nombre d'observations,
— y_i désigne la valeur réelle de l'observation i,
— $\hat{y}_i$ représente la valeur prédite de l'observation i.

III.3.3.2.3 Valeur de l'exactitude

L'exactitude est une mesure simple mais puissante pour évaluer la performance globale d'un modèle de classification. Elle représente la proportion des prédictions correctes (vrais positifs et vrais négatifs) parmi toutes les prédictions effectuées. Une exactitude élevée indique que le modèle est généralement bon pour prédire correctement les résultats. Elle est définie comme suit :

$$\text{Exactitude} = \frac{TP + TN}{TP + TN + FP + FN} \tag{III.2}$$

où :

— TP (True Positive) : Nombre de vrais positifs, c'est-à-dire les cas où le modèle a correctement prédit la présence de la condition.
— TN (True Negative) : Nombre de vrais négatifs, c'est-à-dire les cas où le modèle a correctement prédit l'absence de la condition.
— FP (False Positive) : Nombre de faux positifs, c'est-à-dire les cas où le modèle a incorrectement prédit la présence de la condition.
— FN (False Negative) : Nombre de faux négatifs, c'est-à-dire les cas où le modèle a incorrectement prédit l'absence de la condition.

III.3.4 Implémentation des modèles de prévision

Après avoir sélectionné nos outils de travail, collecté les données nécessaires et distingués nos ensembles de données , puis choisi nos modèles de prévision ainsi que les métriques de performance appropriées, il nous ne reste plus qu'à implémenter ces modèles. Pour une meilleure hiérarchie, nous divisons cette section en 2 parties : une pour la régression, et une autre pour la classification.

III.3.4.1 Implémentation de la régression

Dans cette section, nous cherchons à déterminer laquelle des deux techniques ; simple ou complexe, donne les meilleurs résultats.

Pour la régression linéaire, nous avons adopté une approche directe en ajustant les paramètres sans recourir à un optimiseur spécifique, utilisant plutôt les données disponibles pour modéliser les relations linéaires entre les variables explicatives et la variable cible.

En parallèle pour le LSTM, nous avons utilisé la bibliothèque TensorFlow pour son implémentation. Pour affiner notre approche avec ce modèle, nous avons soigneusement ajusté sa configuration en modifiant plusieurs paramètres clés, tels que les tailles d'entrée (*input_size*) et de sortie (*output_size*), ainsi que la taille cachée (*hidden_size*) et le nombre de couches (*num_layers*). En optimisant ces paramètres, notre objectif était de tirer pleinement parti des dépendances temporelles présentes dans les données historiques du prix du Bitcoin. Nous avons également intégré l'optimiseur Adam pour ajuster efficacement les poids du réseau neuronal.

Cependant, il est nécessaire de normaliser les données avant de les utiliser dans le modèle LSTM. La normalisation est essentielle car elle permet de mettre toutes les variables sur une même échelle, ce qui améliore la convergence de l'optimiseur et la stabilité de l'entraînement. En utilisant la technique Min-Max, nous avons pu éviter les problèmes causés par les grandes variations de valeurs entre les différentes caractéristiques, ce qui a conduit à des performances de modèle plus cohérentes et précises.

Min-Max

La normalisation Min-Max est une technique de prétraitement des données utilisée pour redimensionner les valeurs d'un ensemble de données afin qu'elles se situent dans une plage spécifique, généralement entre 0 et 1. La normalisation Min-Max transforme les données en utilisant la formule suivante :

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{III.3}$$

où :

— X est la valeur d'origine,
— X_{min} est la valeur minimale de l'ensemble de données,
— X_{max} est la valeur maximale de l'ensemble de données,
— X_{norm} est la valeur normalisée.

III.3.4.2 Implémentation de la Classification

Ici, nous cherchons à voir s'il est possible d'obtenir des résultats de prévision du mouvement du marché du BTC, en utilisant deux des approches les plus réputées : les SVM et le Random Forest.

Pour les SVM, nous avons utilisé la bibliothèque Scikit-Learn pour leur implémentation. Notre approche a consisté à ajuster les paramètres clés comme le noyau, le coefficient de régularisation (C), et le paramètre gamma. Ces ajustements ont été cruciaux pour optimiser la capacité des SVM à séparer efficacement les données et à prévoir les variations du prix du Bitcoin en fonction des caractéristiques observées.

Quant au modèle Random Forest, également implémenté avec Scikit-Learn, nous avons ajusté des paramètres tels que le nombre d'arbres dans la forêt ($n\_estimators$), la profondeur maximale des arbres (max_depth), et le nombre minimum de points de données requis pour diviser un nœud ($min\_samples\_split$). Cette approche nous a permis de capturer les relations non linéaires et les interactions complexes entre les variables prédictives, améliorant ainsi la robustesse de nos prévisions sur différentes périodes temporelles du Bitcoin.

III.4 Résultats et observations

III.4.1 Régression

Les résultats obtenus sont intéressants, voire même prometteurs. La régression linéaire a démontré une capacité remarquable à capturer les tendances et les motifs sous-jacents des données, que ce soit pour les séries temporelles quotidiennes ou horaires (figure III.4). Les erreurs moyennes absolues (MAPE) obtenues ont indiqué une précision satisfaisante face aux fluctuations importantes du prix du Bitcoin (tableau III.5). En revanche, bien que le modèle LSTM soit théoriquement capable de modéliser des dépendances temporelles complexes, il n'a pas produit des résultats aussi précis sur nos ensembles de données que la régression linéaire (figure III.4).

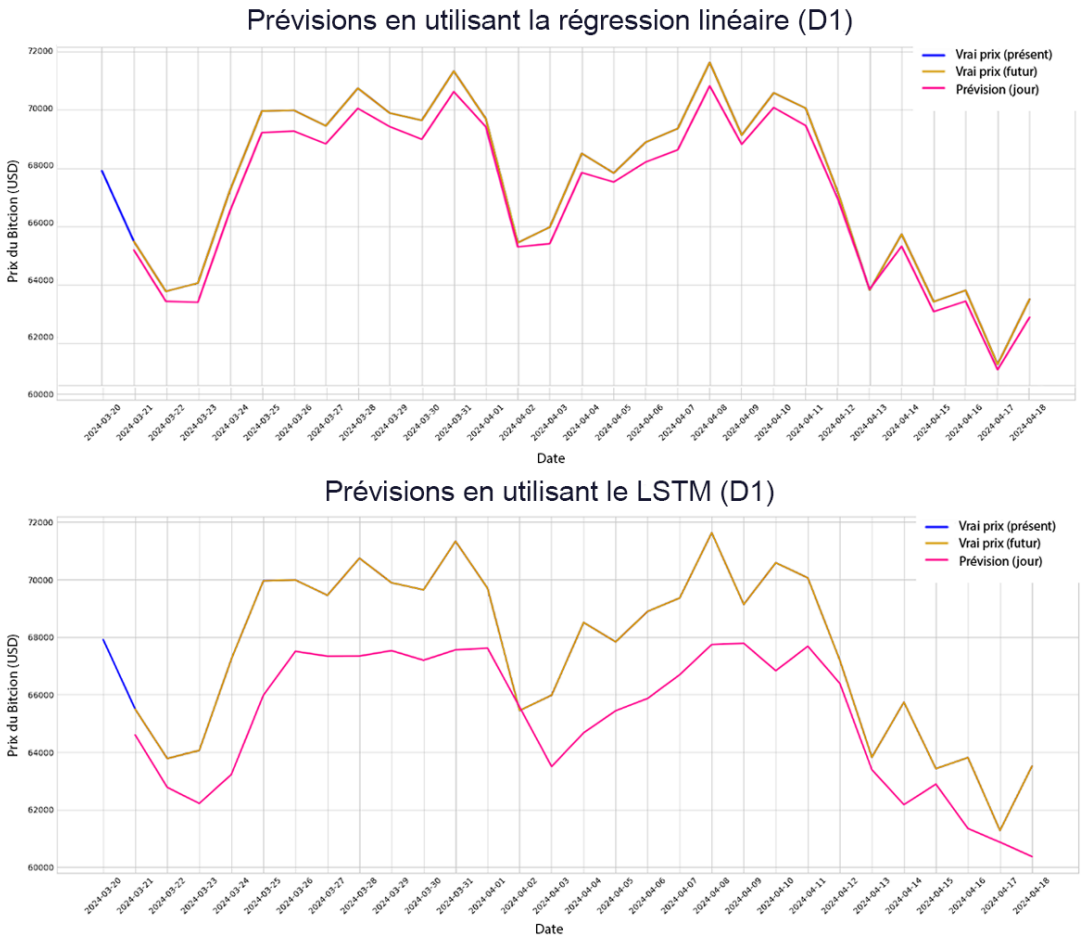

Fig. III.4 – Comparaison entre les prévisions de la régression linéaire et le LSTM pour l'ensemble de données $D1$.

Notons également que les performances des deux modèles sont meilleures sur l'ensemble de données $D2$ par rapport à $D1$ (figure III.5 et tableau III.5). Cette amélioration peut s'expliquer par la plus grande échelle et la fréquence plus fine des données dans $D2$. En effet, les valeurs horaires sont généralement moins volatiles que celles enregistrées quotidiennement.

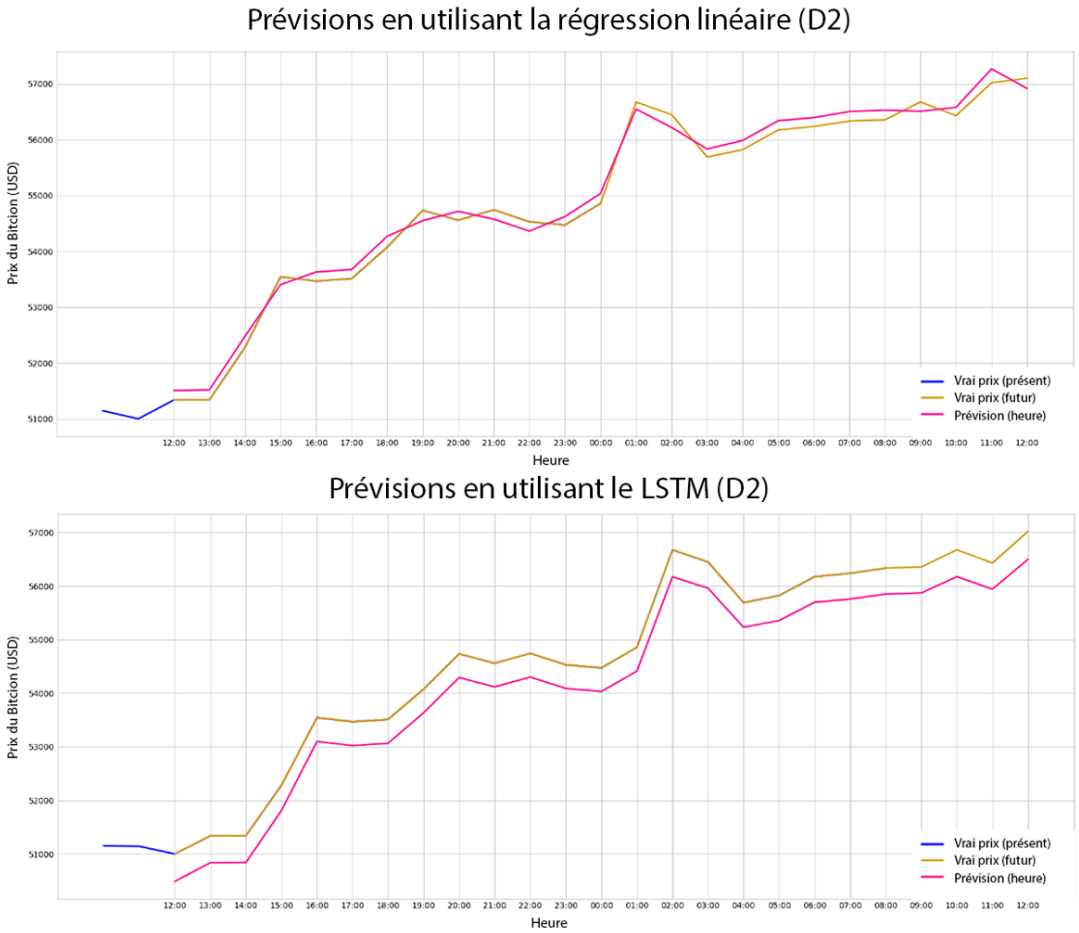

FIG. III.5 – Comparaison entre les prévisions de la régression linéaire et le LSTM pour l'ensemble de données $D2$.

Métrique de performance	Ensemble de données	Régression linéaire	LSTM
MAPE	D1	1.95%	3.36%
	D2	**0.2%**	0.63%

TAB. III.5 – Comparaison entre les prévisions de la régression linéaire et le LSTM pour les ensembles de données $D1$ et $D2$ en utilisant la MAPE comme métrique de performance.

III.4.2 Classification

Quant à la classification, les performances des modèles SVM et Random Forest, bien que moins robustes que celle de la régression, révèlent des résultats intéressants. Les SVM se sont montrés compétitifs, en réussissant plus ou moins à délimiter des frontières

de décision précises dans des espaces de données de haute dimension (figure III.6 et tableau III.6). Cela s'est traduit par des taux de précision robustes sur différents ensembles, bien que légèrement inférieurs à ceux de Random Forest dans la plupart des cas.

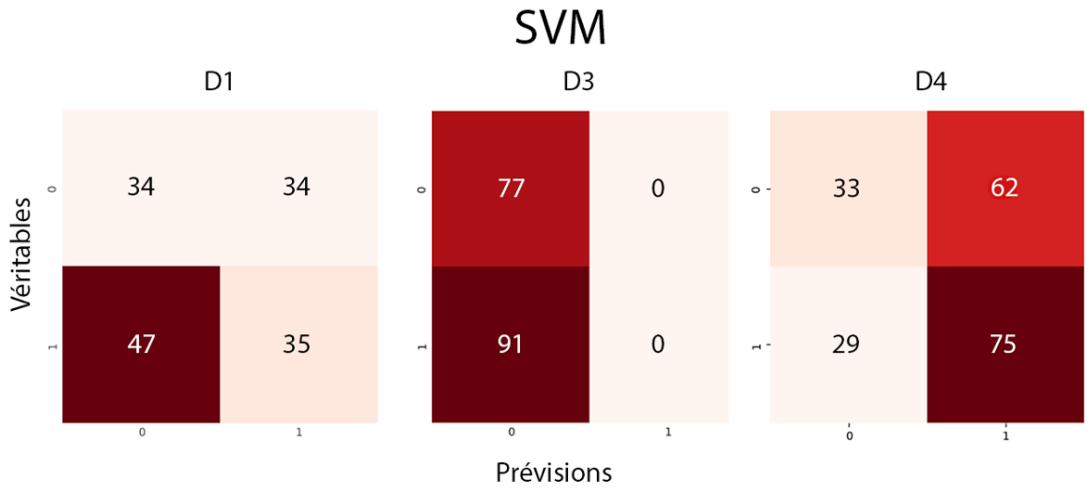

FIG. III.6 – Comparaison entre les prévisions des SVM pour les trois ensembles de données ; $D1$, $D3$, et $D4$ en utilisant la matrice de confusion.

D'autre part, le RF a démontré une capacité notable à gérer la complexité des données grâce à son approche ensembliste d'arbres de décision. Cette méthode s'est avérée particulièrement efficace pour capturer des relations non linéaires et des interactions complexes entre les variables, conduisant à des performances supérieures sur plusieurs ensembles de données (figure III.7 et tableau III.6).

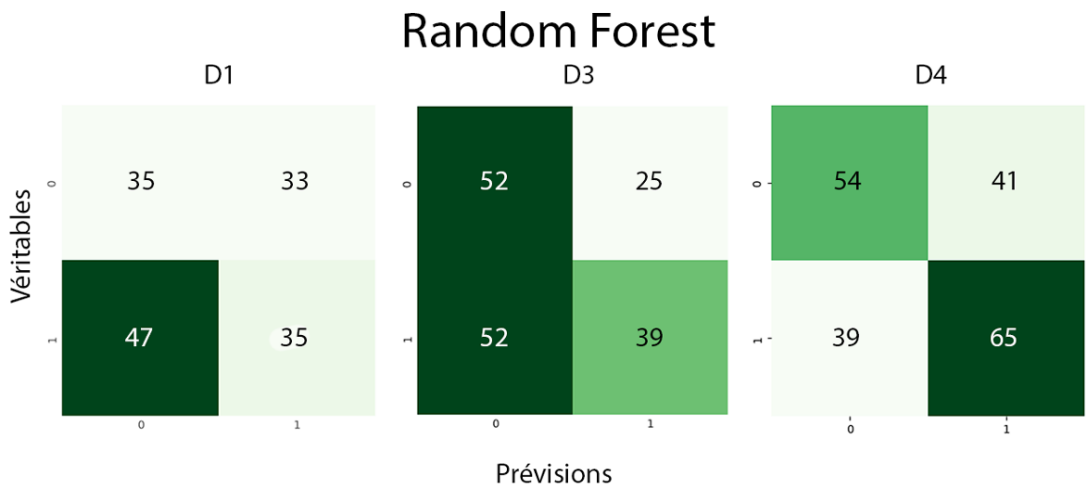

FIG. III.7 – Comparaison entre les prévisions du Random Forest pour les trois ensembles de données ; $D1$, $D3$, et $D4$ en utilisant la matrice de confusion.

Métrique de performance	**Ensemble de données**	**SVM**	**Random Forest**
Exactitude	D1	0.45%	0.46%
	D3	0.45%	0.54%
	D4	0.54%	**0.6%**

TAB. III.6 – Comparaison entre les prévisions des SVM et du Random Forest pour les ensembles de données $D1$, $D3$ et $D4$ en utilisant l'exactitude comme métrique de performance.

Ces observations ont pu mettre en lumière l'importance de sélectionner un modèle adapté aux caractéristiques spécifiques des données et des objectifs de prédiction, en tenant compte de leur échelle et de leur fréquence.

III.5 Discussions

En comparant les performances des modèles de régression linéaire et de LSTM sur les deux ensembles de données ($D1$ et $D2$), il apparaît clairement que la régression linéaire offre de meilleurs résultats (figure III.4 et figure III.5). Pour l'ensemble de données $D1$, basé sur une fréquence quotidienne, la régression linéaire a surpassé le LSTM, montrant une meilleure précision dans les prévisions des variations du prix du Bitcoin. De même, pour l'ensemble de données $D2$, qui contient des valeurs horaires, la régression linéaire a également démontré une performance supérieure au LSTM. Ces observations indiquent que, malgré la complexité et la capacité du LSTM à capturer les dépendances temporelles

à long terme, la régression linéaire reste plus efficace pour la prévision des prix du Bitcoin, tant sur des données quotidiennes qu'horaires.

Nous pouvons également nous apercevoir que pour les deux modèles, la précision du jeu de données de $D2$ est meilleure que celle de $D1$ (tableau III.5). Un facteur clé expliquant ces résultats est la fréquence des données et la taille de l'ensemble de données. En effet, l'ensemble de données $D2$, avec sa fréquence horaire, est plus détaillé par rapport à $D1$, qui est quotidien. Cette précision supplémentaire signifie que les valeurs horaires sont plus nombreuses et moins volatiles, ce qui permet aux modèles d'améliorer leur capacité à d'entraînement, ainsi que de capter plus facilement les tendances et les motifs sous-jacents.

Dans le second cas, en se servant de la matrice de confusion ainsi que de la métrique d'exactitude pour analyser les performances des modèles de classification SVM et Random Forest sur les trois ensembles de données ($D1$, $D3$ et $D4$), il est évident que le Random Forest surpasse les SVM dans la plupart des cas (figure III.6, figure III.7 et tableau III.6). Pour l'ensemble de données $D1$, ces modèles montrent des résultats similaires, avec une légère supériorité du RF. Cependant, pour les ensembles de données $D3$ et $D4$, le Random Forest démontre une performance nettement supérieure par rapport aux SVM, particulièrement sur $D3$ où l'écart de précision est le plus significatif.

Ces résultats indiquent que Random Forest est plus efficace que SVM pour la classification des données dans ce contexte particulier. La supériorité du Random Forest peut être attribuée à sa capacité à gérer les complexités et les variabilités des données de manière plus robuste. Contrairement à SVM, qui peut être sensible à la distribution et à l'échelle des données, Random Forest, avec son ensemble d'arbres de décision, offre une flexibilité et une résilience accrues face aux variations et aux anomalies des données.

Le fait que $D3$ et $D4$ présentent de meilleures performances pour les deux modèles par rapport à $D1$ prouve qu'il existe une réelle corrélation entre les faits extérieurs et le marché du BTC, et souligne l'importance de la qualité et de la structure de ces données utilisées pour l'entraînement. En outre, les résultats supérieurs du Random Forest sur $D4$ mettent en évidence l'influence que les autres marchés financiers peuvent avoir sur celui du Bitcoin, ainsi que l'importance de la combinaison du bon modèle avec les bonnes données.

III.6 Conclusions

Dans cette section exploratoire, nous avons tracé un chemin méthodique vers notre objectif en adoptant une approche rigoureuse. Notre parcours a débuté par la sélection méticuleuse de l'environnement de travail et des outils appropriés, posant ainsi les fondations nécessaires pour nos analyses futures.

En parallèle, nous avons entrepris une collecte minutieuse des données, en accordant une attention particulière à leur qualité et à leur représentativité. Cette étape cruciale nous a permis de disposer d'un jeu de données robuste et pertinent pour nos expérimentations. Avec nos données en main, nous avons ensuite pris des décisions stratégiques quant aux approches de prévision à adopter et aux métriques de performance à utiliser. Ces choix ont été guidés par notre objectif de garantir une évaluation objective et efficace de nos modèles.

L'implémentation des méthodes de prévision a constitué le cœur de notre démarche. Nous avons mis en œuvre ces méthodes avec soin, en ajustant les paramètres pour optimiser leur capacité à générer des prévisions précises et fiables.

Chaque étape de notre processus a été documentée avec transparence, permettant une analyse approfondie des résultats obtenus. Cette approche nous a permis d'explorer et d'interpréter les implications de nos recherches, enrichissant ainsi notre compréhension des défis et des opportunités rencontrés.

En conclusion, cette section représente une avancée significative vers l'accomplissement de notre objectif, tout en mettant en lumière les choix stratégiques qui ont façonné notre parcours analytique.

Conclusions & Perspectives

Ce mémoire représente une exploration approfondie des capacités prédictives des modèles d'intelligence artificielle dans le domaine de la prédiction du prix du Bitcoin. Notre objectif principal était d'évaluer et de comparer différentes approches de modélisation pour anticiper les fluctuations du marché des cryptomonnaies.

Nous avons minutieusement analysé les performances de la régression linéaire, du LSTM, des SVM et du Random Forest sur des ensembles de données variés, incluant des fréquences quotidiennes et horaires. Nos résultats ont mis en lumière la supériorité de la régression linéaire pour prédire avec précision les prix du Bitcoin, surtout sur des données horaires plus détaillées.

Cependant, il est essentiel de considérer ces résultats avec prudence. L'efficacité des modèles peut varier considérablement en fonction de la qualité et de la granularité des données utilisées. Le marché du Bitcoin est intrinsèquement influencé par de nombreux facteurs externes imprévisibles tels que la réglementation gouvernementale, les événements géopolitiques, et les fluctuations macroéconomiques. Ces variables externes peuvent significativement affecter les prévisions des modèles, rendant difficile une prédictibilité totale.

L'exploration de différentes sources de données et l'adaptation des modèles en conséquence peuvent toutefois conduire à des améliorations significatives des performances prédictives. En perspective, cette étude ouvre la voie à de nouvelles recherches explorant des approches plus sophistiquées, telles que l'intégration de données économiques et financières additionnelles. L'incorporation de variables exogènes comme les indices boursiers, les taux d'intérêt ou les politiques monétaires pourrait potentiellement renforcer la fiabilité des prévisions à court et moyen terme dans le domaine dynamique des cryptomonnaies.

Plutôt qu'une conclusion définitive, ce mémoire représente le début d'une réflexion approfondie sur les possibilités et les défis liés à la prédiction des prix du Bitcoin à l'aide de l'intelligence artificielle. Il souligne la nécessité continue d'innover dans les méthodes de modélisation et d'analyse des données pour mieux saisir la volatilité et la complexité du marché des cryptomonnaies. Cette étude encourage également à envisager des approches interdisciplinaires, intégrant des données économiques et financières diversifiées, pour enrichir la robustesse des prédictions futures.

Bibliographie

[1] Andreas M. Antonopoulos. *Mastering Bitcoin : Unlocking Digital Cryptocurrencies*. O'Reilly Media, 2017.

[2] Isaac Appiah-Otoo. « The Impact of the Russia-Ukraine War on the Cryptocurrency Market ». In : *Asian Economics Letters* 4.1 (2023). doi : 10.46557/001c.53110.

[3] Dirk G. Baur et Thomas Dimpfl. « The volatility of Bitcoin and its role as a medium of exchange and a store of value ». In : *Empirical Economics* 61.5 (2021), p. 2663-2683. doi : 10.1007/s00181-020-01990-5.

[4] Dave Bayer, Stuart Haber et W. Scott Stornetta. *Improving the Efficiency and Reliability of Digital Time-Stamping*. Rapp. tech. Sequoia 2000 Technical Report, 1992.

[5] Y. Bi. « A Survey on Evolutionary Computation for Computer Vision and Image Analysis : Past, Present, and Future Trends ». In : (2022). eprint : arXiv:2209.06399.

[6] Bitget. *Transaction du Bitcoin - Historical Transaction Records*. https://www.bitget.com/data-download/spot-historical-transaction-record. Visited on 09/06/2024.

[7] Joseph Bonneau et al. « Sok : Research Perspectives and Challenges for Bitcoin and Cryptocurrencies ». In : *IEEE Symposium on Security and Privacy*. 2015, p. 104-121. doi : 10.1109/SP.2015.14. url : https://doi.org/10.1109/SP.2015.14.

[8] L. Breiman. « Random Forests ». In : *Machine Learning* 45 (2001), p. 5-32. doi : 10.1023/A:1010933404324.

[9] M. Brunn et W. Genieys. « L'intelligence artificielle va-t-elle bouleverser la profession médicale ? » In : *The Conversation* (2020).

[10] Vitalik Buterin. *Ethereum White Paper : A Next-Generation Smart Contract and Decentralized Application Platform*. 2013. url : https://ethereum.org/content/whitepaper/whitepaper-pdf/Ethereum_Whitepaper_-_Buterin_2014.pdf.

[11] R. Y. Choi et al. « Introduction to Machine Learning, Neural Networks, and Deep Learning ». In : *Translational vision science & technology* 9.2 (2020), p. 14. doi : 10.1167/tvst.9.2.14.

[12] John Edwards. *Bitcoin's Price History*. Investopedia. Consulté le (18/05/2024). 2024. url : https://www.investopedia.com/articles/forex/121815/bitcoins-price-history.asp#citation-38.

[13] Theodoros Evgeniou et Massimiliano Pontil. « Support vector machines : Theory and applications ». In : *Machine Learning and Its Applications*. T. 2049. Springer, 2001, p. 249-257. doi : 10.1007/3-540-44673-7_12.

[14] M. Frohmann et al. « Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting ». In : *Big Data and Cognitive Computing* 7.3 (2023), p. 137. doi : 10.3390/bdcc7030137.

[15] GeeksForGeeks. *Clustering in Machine Learning. GeeksForGeeks.* (Visité le 25/04/2024). 2024. url : https://www.geeksforgeeks.org/clustering-in-machine-learning/.

[16] John W. Goodell et al. « Explainable artificial intelligence modeling to forecast bitcoin prices ». In : *International Review of Financial Analysis* 88 (2023), p. 102702. issn : 1057-5219. doi : 10.1016/j.irfa.2023.102702.

[17] Hashpi. *The Intuition Behind Kernel Methods.* https://www.hashpi.com/the-intuition-behind-kernel-methods. (Visité le 2024-05-29). année non spécifiée.

[18] Sepp Hochreiter et Jürgen Schmidhuber. « Long Short-term Memory ». In : *Neural computation* 9 (1997), p. 1735-80. doi : 10.1162/neco.1997.9.8.1735.

[19] Marin IvezicMarch. *The 1956 Dartmouth Workshop : The Birthplace of Artificial Intelligence (AI).* 2017.

[20] Y. Jiao. « Applications of artificial intelligence in e-commerce and finance ». In : (2018).

[21] Teemu Kanstrén. *Merkle Trees : Concepts and Use Cases.* Fév. 2021.

[22] Mounia Khelifa, Abdelkader Khedaoui Mustapha et Ladjlat Brahim. « Typologie de fraude aux moyens de paiement électroniques et les exigences européennes de sécurité ». In : *Revue Afak des Sciences* (2021).

[23] Dietrich Knauth. « Crypto companies crash into bankruptcy ». In : *Reuters* (2022).

[24] M. H. Krishna et al. « Studies on Anomaly Detection Techniques ». In : *2023 7th International Conference on Computing Methodologies and Communication (ICCMC).* 2023, p. 813-817.

[25] Deepak Kumar. *Anomaly detection for cyber security via machine learning.* LinkedIn. Consulté le 25/04/2024. 2021. url : URL_de_Linkedin.

[26] Saket Kumar. *2019-2024 US Stock Market Data.* https://www.kaggle.com/datasets/saketk511/2019-2024-us-stock-market-data. Visited on 09/06/2024.

[27] Yann LeCun, Yoshua Bengio et Geoffrey Hinton. « Deep Learning ». In : *Nature* 521.7553 (2015), p. 436-444.

[28] Michelle Llamas. *Crypto Bankruptcies.* Consumer Notice. 2024. url : https://consumernotice.org/legal/crypto-bankruptcies.

[29] Anne-Laure Lombard. *La place de l'éthique dans les banques depuis la crise financière. Un simple effet d'annonce ?* KEDGE BUSINESS SCHOOL, 2019.

[30] H. Maâmatou. « Apprentissage semi-supervisé pour la détection multi-objets dans des séquences vidéos. Application à l'analyse de flux urbains ». Thèse de doct. 2017.

[31] Morgan P. McBee et Christopher Wilcox. « Blockchain Technology : Principles and Applications in Medical Imaging ». In : *Journal of Digital Imaging* 33.3 (2020), p. 726-734. DOI : 10.1007/s10278-019-00310-3. URL : https://doi.org/10.1007/s10278-019-00310-3.

[32] Zhenzhu Meng. *Using a Data Driven Approach to Predict Waves Generated by Gravity Driven Mass Flows - Figure Scientifique*. https://www.researchgate.net/publication/343858437_Using_a_Data_Driven_Approach_to_Predict_Waves_Generated_by_Gravity_Driven_Mass_Flows_-_Figure_Scientifique. 2020.

[33] Ralph C. Merkle. « A Digital Signature Based on a Conventional Encryption Function ». In : *Advances in Cryptology—CRYPTO'87*. 1987, p. 369-378. DOI : 10.1007/3-540-48184-2_32.

[34] William Mougayar. *The Business Blockchain : Promise, Practice, and the Application of the Next Internet Technology*. Wiley, 2016.

[35] Aatila Mustapha, Mohamed Lachgar et Kartit Ali. « An Overview of Gradient Descent Algorithm Optimization in Machine Learning : Application in the Ophthalmology Field ». In : 2020. DOI : 10.1007/978-3-030-45183-7_27.

[36] Satoshi Nakamoto. *Bitcoin : A Peer-to-Peer Electronic Cash System*. Rapp. tech. Bitcoin.org, 2008. URL : https://bitcoin.org/bitcoin.pdf.

[37] Satoshi Nakamoto. *Bitcoin : A Peer-to-Peer Electronic Cash System*. 2008. URL : https://bitcoin.org/bitcoin.pdf.

[38] Arvind Narayanan et al. *Bitcoin and Cryptocurrency Technologies*. Princeton University Press, 2016.

[39] Oliviervha. *Crypto News Dataset*. https://www.kaggle.com/datasets/oliviervha/crypto-news. Visited on 09/06/2024.

[40] J. Sanabria-Navarro et al. « Incidences of artificial intelligence in contemporary education ». In : *Comunicar* (2023).

[41] Nikhil Sardana. *Neural Networks : Forward and Backpropagation*. 2017.

[42] S. W. Schuetz et V. Venkatesh. « Blockchain, Adoption, and Financial Inclusion in India : Research Opportunities ». In : *International Journal of Information Management* 52 (2020), p. 101936. DOI : 10.1016/j.ijinfomgt.2019.04.009.

[43] Selenium. *Selenium Documentation*. https://www.selenium.dev/documentation/. Visited on 09/06/2024.

[44] Alex Sherstinsky. « Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network ». In : *Physica D : Nonlinear Phenomena* 404 (2020), p. 132306. ISSN : 0167-2789. DOI : 10.1016/j.physd.2019.132306. URL : https://doi.org/10.1016/j.physd.2019.132306.

[45] MacKenzie Sigalos. *Bitcoin's wild price moves stem from its design — you'll need strong nerves to trade it*. CNBC. 2021. URL : https://www.cnbc.com/2021/05/19/why-is-bitcoin-so-volatile.html.

[46] Théophile Sossa. « Le concept d'exclusion et sa signification en finance ». In : *La microfinance au Bénin*. Graduate Institute Publications, 2011. DOI : 10.4000/books.iheid.354.

[47] Statista. *Average mining difficulty of Bitcoin from January 2009 to January 20, 2024*. Visité le 2024-05-25. 2024. URL : https://www.statista.com/statistics/994097/bitcoin-mining-difficulty/.

[48] C. Tardi. *Genesis Block : Bitcoin Definition, Mysteries, and Secret Message*. Investopedia. visited on 23/05/2024. Mai 2024.

[49] TseKiChun. *Random forest explain. Wikimedia*. (Visité le 24/05/2024). 2021. URL : https://www.wikimedia.org.

[50] Utkarsh. *What is Regression in Data Mining ? Scaler*. (visité le 24/05/2024). 2023.

[51] Burak Uyduran. « The crypto effect on cross border transfers and future trends of cryptocurrencies ». In : *Financial Internet Quarterly* (2020). URL : https://hdl.handle.net/10419/266847.

[52] J. Wangata, M. Elenge et C. De Brouwer. « Les accidents du travail dans le transport urbain en commun de la ville province de Kinshasa, République Démocratique du Congo : une étude transversale descriptive ». In : *The Pan African medical journal* 19 (2014), p. 41. DOI : 10.11604/pamj.2014.19.41.4020.

[53] G. Wardeh. « Detection d'obstacles par vision par ordinateur en vue de la determination de la trajectoire d'un robot ». In : (1989).

[54] Dana Yaga et al. *Blockchain Technology Overview*. Rapp. tech. NISTIR 8202. National Institute of Standards et Technology, 2018. DOI : 10.6028/NIST.IR.8202. URL : https://doi.org/10.6028/NIST.IR.8202.

[55] Krzysztof Zarzycki et Maciej Ławryńczuk. « LSTM and GRU Neural Networks as Models of Dynamical Processes Used in Predictive Control : A Comparison of Models Developed for Two Chemical Reactors ». In : *Sensors* 21.16 (2021), p. 5625. DOI : 10.3390/s21165625. URL : https://doi.org/10.3390/s21165625.

[56] Yatao Zhang et al. « Comparing the Performance of Random Forest, SVM and Their Variants for ECG Quality Assessment Combined with Nonlinear Features ». In : *Journal of Medical and Biological Engineering* 39 (2018). DOI : 10.1007/s40846-018-0411-0.

[57] Juncheng Zhu et al. « Electric Vehicle Charging Load Forecasting : A Comparative Study of Deep Learning Approaches ». In : *Energies* 12 (2019), p. 2692. DOI : 10.3390/en12142692.

[58] Aviv Zohar. « Bitcoin : Under the Hood ». In : *Communications of the ACM* 58.9 (2015), p. 104-113. DOI : 10.1145/2701411. URL : https://doi.org/10.1145/2701411.

ملخص : تمثل البيتكوين، نظرًا لتقلباتها الشديدة وتقلباتها غير المتوقعة في السوق المالية، تحديًا كبيرًا للتنبؤ الدقيق بسعرها المستقبلي. يركز هذا المشروع على تطوير وتقييم نماذج الذكاء الاصطناعي المتقدمة، مثل نماذج الانحدار الخطي و LSTM و SVM و Forest Random بهدف توفير تنبؤات موثوقة على الرغم من هذه الظروف المتقلبة. وفي الوقت نفسه، نبحث في التأثير المحتمل لمعنويات المقالات الصحفية والمتغيرات الخارجية الأخرى مثل سوق النفط ومؤشر S&P 500 على اتجاهات سوق البيتكوين. والهدف من ذلك هو فهم وتوقع الديناميكيات الأساسية لهذه العملة المشفرة سريعة التطور بشكل أفضل.
يقدم هذا التقرير نهجًا ثريًا لفهم الديناميكيات الأساسية التي تؤثر على تقلبات أسعار البيتكوين بشكل أفضل، مع استكشاف التقاطع بين سلسلة الكتل والذكاء الاصطناعي.
الكلمات الرئيسية: بيتكوين، بلوكشين، توقعات السلاسل الزمنية، الذكاء الاصطناعي، الانحدار الخطي، LSTM، SVM، Random Forest، تحليل المشاعر.

Résumé : Le Bitcoin, en raison de sa volatilité extrême et de ses fluctuations imprévisibles sur le marché financier, présente un défi significatif pour la prévision précise de son prix futur. Ce projet se concentre sur le développement et l'évaluation de modèles avancés d'intelligence artificielle (IA), tels que la régression linéaire, les LSTM, les SVM et les Random Forest, dans le but de fournir des prévisions fiables malgré ces conditions volatiles. En parallèle, nous étudions l'influence potentielle des sentiments des articles de presse et d'autres variables externes comme le marché du pétrole et l'indice S&P 500 sur les tendances du marché du BTC. L'objectif est de mieux comprendre et d'anticiper les dynamiques sous-jacentes de cette cryptomonnaie en rapide évolution. Ce rapport présente une approche enrichie visant à mieux comprendre les dynamiques sous-jacentes influençant les fluctuations des prix du Bitcoin, tout en explorant l'intersection entre la blockchain et l'intelligence artificielle.
Mots clés : Bitcoin, blockchain, time series forecasting, intelligence artificielle, régression linéaire, LSTM, SVM, Random Forest, analyse de sentiment.

Abstract : Bitcoin, due to its extreme volatility and unpredictable fluctuations on the financial market, presents a significant challenge for the accurate prediction of its future price. This project focuses on the development and evaluation of advanced artificial intelligence (AI) models, such as linear regression, LSTM, SVM and Random Forest, with the aim of providing reliable forecasts despite these volatile conditions. At the same time, we are investigating the potential influence of press article sentiment and other external variables such as the oil market and the S&P 500 index on BTC market trends. The aim is to better understand and anticipate the underlying dynamics of this rapidly evolving cryptocurrency. This report presents an enriched approach to better understand the underlying dynamics influencing Bitcoin price fluctuations, while exploring the intersection between blockchain and artificial intelligence.
Keywords : Bitcoin, blockchain, time series forecasting, artificial intelligence, linear regression, LSTM, SVM, Random Forest, sentiment analysis