People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

University of 8 May 1945-Guelma-

Faculty of Mathematics, Computer Science and Science of Matter
Department of Computer Science

Master Thesis

Specialty: Computer Science

Option:

Informatics Systems

Theme

Développement d'un modèle intelligent pour la prédiction de la volatilité financière

Presented by:

Ousseini Beidou Habibou

Jury Members:

N	Full Name	Quality
1	Dr DERDAR Salah	Chairman
2	Dr FERKOUS Chokri	Supervisor
3	Dr ABDELMOUMENE Hiba	Examiner

June 2024.

Résumé

L'évaluation de l'incertitude liée aux fluctuations des prix des actifs est cruciale en finance, la volatilité étant une mesure essentielle. La capacité à prédire cette volatilité est vitale pour prendre des décisions d'investissement efficaces, gérer les risques et fixer le prix des produits. Cette recherche vise à expliquer le concept de volatilité et son importance pour les institutions financières et les entreprises mondiales. Elle fournit un aperçu complet des techniques d'apprentissage automatique, en mettant l'accent sur les réseaux de neurones artificiels, dans le contexte de problèmes de classification tels que la prévision des hausses et des baisses de l'indice BIST 100. La conception de modèles d'apprentissage automatique et de réseaux de neurones artificiels est présentée en détail, y compris le processus de collecte de données, de préparation des données, de sélection d'algorithmes et d'optimisation de modèles. Chaque étape est essentielle pour développer un modèle fiable et efficace. Les modèles sont ensuite mis en œuvre et les résultats analysés, comparant les performances de différents modèles afin de déterminer le meilleur modèle pour prédire les mouvements du marché. Les résultats de l'étude ont montré que le modèle de forêt aléatoire et les arbres de décision obtenaient les meilleurs résultats en matière de prévision des fluctuations des marchés financiers par rapport aux autres modèles testés. Les résultats de la recherche apportent une contribution précieuse au domaine de la prévision des fluctuations des marchés financiers à l'aide de l'apprentissage automatique et de la technologie des réseaux neuronaux artificiels. L'étude met en évidence le potentiel de ces technologies pour améliorer considérablement la précision des prévisions, analyser les risques et prendre des décisions d'investissement éclairées. En démontrant l'efficacité de l'apprentissage automatique et des modèles de réseaux neuronaux artificiels pour prédire les fluctuations des marchés financiers, cette recherche ouvre d'énormes possibilités pour améliorer les performances des investissements financiers et la gestion des risques. Les résultats indiquent également que des recherches supplémentaires sont nécessaires pour développer des modèles plus complexes et plus précis, prenant en compte l'évolution du paysage économique et financier.

Mots-clefs : volatilité, volatilité financière, apprentissage automatique, prédiction de la volatilité

ملخص

يعد تقييم حالة عدم اليقين المرتبطة بتقلبات أسعار الأصول أمرًا بالغ الأهمية في مجال التمويل، حيث يمثل التقلب مقياسًا رئيسيًا. تعد القدرة على التنبؤ بهذا التقلب أمرًا حيويًا لاتخاذ قرارات استثمارية فعالة وإدارة المخاطر وتسعير المنتجات. يهدف هذا البحث إلى شرح مفهوم التقلب وأهميته بالنسبة للمؤسسات المالية والشركات العالمية. ويقدم نظرة شاملة لتقنيات التعلم الآلي، مع التركيز على الشبكات العصبية الاصطناعية، في سياق مشاكل التصنيف مثل التنبؤ بالارتفاعات والانخفاضات في مؤشر BIST 100، كما يتم تفصيل نماذج التعلم الآلي والشبكات العصبية الاصطناعية، بما في ذلك عملية البيانات جمع وإعداد البيانات واختيار الخوارزمية وتحسين النموذج. كل خطوة ضرورية لتطوير نماذج موثوقة وفعالة. يتم بعد ذلك تنفيذ النماذج وتحليل النتائج، ومقارنة أداء النماذج المختلفة لتحديد النموذج الأفضل للتنبؤ بتقلبات السوق. وأظهرت نتائج البحث أن نماذج الغابات العشوائية وأشجار القرار تحقق أفضل النتائج في التنبؤ بتقلبات الأسواق المالية مقارنة بالنماذج الأخرى التي تم اختبارها. تشكل نتائج البحث مساهمة قيمة في مجال التنبؤ بتقلبات الأسواق المالية باستخدام تقنيات التعلم الآلي والشبكات العصبية الاصطناعية. وتسلط الدراسة الضوء على إمكانات هذه التقنيات لتحسين دقة التنبؤ بشكل كبير، وتحليل المخاطر واتخاذ قرارات استثمارية سليمة. من خلال إظهار فعالية التعلم الآلي ونماذج الشبكات العصبية الاصطناعية في التنبؤ بتقلبات السوق المالية، تفتح هذه الدراسة إمكانيات واسعة لتحسين أداء الاستثمار وإدارة المخاطر في مجال التمويل. وتشير النتائج أيضًا إلى الحاجة إلى مزيد من البحث لتطوير نماذج أكثر تعقيدًا ودقة، مع الأخذ في الاعتبار المشهد الاقتصادي والمالي المتغير باستمرار.

الكلمات المفتاحية: التقلبات، التقلبات المالية، التعلم الآلي، التنبؤ بالتقلبات.

Abstract

Assessing the uncertainty associated with asset price fluctuations is crucial in finance, with volatility being a key measure. The ability to predict this volatility is vital for making effective investment decisions, managing risk and pricing products. This research aims to explain the concept of volatility and its importance for financial institutions and global businesses. It provides a comprehensive overview of machine learning techniques, with emphasis on artificial neural networks, in the context of classification problems such as predicting rises and falls in the BIST 100 index. machine learning models and artificial neural networks is detailed, including the process of data collection, data preparation, algorithm selection and model optimization. Each step is essential to develop reliable and efficient models. The models are then implemented and the results analyzed, comparing the performance of different models to identify the best model for predicting market volatility. The research results show that random forest models and decision trees achieve the best results in predicting financial market volatility compared to other models tested. The research results constitute a valuable contribution to the field of forecasting financial market volatility using machine learning and artificial neural network techniques. The study highlights the potential of these techniques to significantly improve forecast accuracy, analyze risks and make sound investment decisions. By demonstrating the effectiveness of machine learning and artificial neural network models in predicting financial market volatility, this study opens vast possibilities for improving investment performance and risk management in finance. The results also suggest the need for further research to develop more complex and accurate models, taking into account the ever-changing economic and financial landscape.

Keywords: volatility, financial volatility, machine learning, volatility prediction.

REMERCIEMENTS

Tout d'abord, nous remercions Allah qui, par sa bonté infinie, nous a permis d'élaborer ce travail.

Je tiens à exprimer ma profonde gratitude à toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce travail :

- À mon encadrant, Dr. FERKOUS Chokri, pour sa guidance précieuse, son expertise inestimable et ses conseils éclairés qui ont grandement enrichi ce projet.
- J'adresse mes remerciements les plus sincères aux membres du jury qui ont bien voulu examiner ce modeste travail.
- Mes remerciements s'adressent également à tous les professeurs de notre cher département informatique pour leur générosité et la patience dont ils ont su faire preuve malgré leurs charges professionnelles.
- -À mes collègues et amis, pour leurs encouragements constants, leur support technique et leur amitié sincère tout au long de cette aventure académique.
- À toutes les personnes qui ont participé à l'étude en fournissant des données et en partageant leur expertise, votre contribution a été essentielle à la réussite de ce travail.
- Enfin, à ma famille, pour leur soutien inconditionnel, leur patience et leur compréhension pendant les moments où j'ai dû me consacrer à cette recherche.

DEDICACE

À mes parents, en reconnaissance de leur amour inconditionnel, de leur soutien inébranlable et les sacrifices qu'ils ont consentis pour mon éducation. Votre soutien constant a été mon moteur tout au long de cette trajectoire. À mes frères et sœurs, je tiens à exprimer ma gratitude pour leur soutien sans faille, leurs encouragements et leur compréhension pendant les périodes difficiles de ce projet.

Je tiens à exprimer ma gratitude envers mes professeurs et encadrants pour leur précieux enseignement, leur orientation éclairée et leur disponibilité à répondre à mes interrogations et à me guider tout au long de mes études.

À mes proches amis, pour leur amitié authentique, leurs précieux conseils et les instants de détente qui ont apaisé le fardeau des obligations académiques.

À tous ceux qui ont joué un rôle de près ou de loin dans mon parcours scolaire, que ce soit mes camarades de classe, des chercheurs, des mentors ou des professionnels, votre influence positive a été bénéfique pour enrichir mon expérience et favoriser mon développement personnel et professionnel.

Enfin, je tiens à exprimer ma gratitude envers toutes les personnes bienveillantes qui croient en moi et qui ont partagé leur sagesse, leur expérience et leur passion pour l'apprentissage. Cela m'a permis de tirer des leçons précieuses de nos échanges.

Cette dédicace est un témoignage de ma reconnaissance éternelle envers chacune de ces personnes qui ont rendu possible la réalisation de ce projet.

Table des matières

Liste des Figures

Liste des tableaux

Abréviations et Acronymes

GARCH	Hétéroscédasticité conditionnelle autorégressive généralisée
ARCH	Hétéroscédasticité conditionnelle autorégressive
LLF	Loglikelihood function
FX	Foreign Exchange
SEMIFARMA	Semi-parametric fractional autoregressive Moving Average
EGARCH	Exponential Generalized Autoregressive Conditional Heteroskedasticity
NYSE	New York Stock Exchange
TSE-300	Toronto Stock Exchange 300
CBOE	Chicago Board Options Exchange
NASDAQ	National Association of Securities Dealers Automated Quotations
ML	Machine learning
IA	intelligence artificielle
SVM	Support Vector Machine
RBF	Radial Basis Function
KNN	K Nearest Neighbor
ANN	Artificial Neural Network
RELU	unité linéaire rectifiée
MSE	Mean Squared Error
BIST-100	Borsa İstanbul 100 Index
S&P 500	Standard & Poor's 500
Small Cap 2000	Russell 2000 Small Cap Index
DT	Decision Tree
RF	RandomForestClassifier
LR	Regression logistique

INTRODUCTION GENERALE

Les conséquences de la crise des subprimes de 2007 sur l'économie américaine et d'autres régions du monde, tout comme l'évènement de la pandémie COVID 19 ont donné une meilleure compréhension de l'importance des variations économiques nationales et internationales. De même, Les variations des marchés financiers attirent l'attention des autorités publiques, des investisseurs, des chercheurs et d'autres acteurs. Malgré des études et des recherches approfondies sur le concept de volatilité et ses prévisions, il demeure inconnu les causes des fluctuations spectaculaires des cours boursiers. En fait, mesurer et prévoir les changements importants dans la volatilité du marché peut permettre aux investisseurs et aux autres acteurs de mieux appréhender la corrélation entre les rendements prévus, les primes de risque et la volatilité, les acteurs du marché peuvent ainsi optimiser leurs choix d'investissement financier. C'est pourquoi nous avons choisi le sujet « Développement d'un modèle intelligent pour la prédiction de la volatilité financière » comme projet de fin d'étude.

Face à ces défis, l'apprentissage automatique (machine learning) se présente comme une solution innovante et puissante. La machine learning est une discipline de l'intelligence artificielle qui permet aux systèmes d'apprendre à partir de données, au lieu d'être programmés explicitement. Cette technologie est omniprésente dans divers secteurs et joue un rôle crucial pour prise des décisions satisfaisantes aux clients. Son objectif est de former des algorithmes qui soient capables de traiter de manière efficace et pertinente les informations disponibles. Il se décompose en deux étapes principales : l'apprentissage et la prédiction. Ce projet examine les méthodes d'apprentissage automatique dans le domaine de la finance, notamment pour prédire la volatilité.

Afin de répondre à nos interrogations de recherche, notre travail est structuré en quatre chapitres : :

Dans le premier chapitre, nous examinerons la littérature concernant le concept de volatilité et son développement.

Dans le deuxième chapitre nous abordons l'apprentissage automatique tout en mettant l'accès sur les différents types d'apprentissage automatique. Ensuite, nous décrivons brièvement plusieurs de ses algorithmes, notamment ceux de l'apprentissage supervisé, afin de mieux

appréhender les points forts et les points faibles de chacun de ces algorithmes, afin de prédire de manière plus précise la hausse et la baisse du prix de l'indice boursier BIST 100.

Dans le troisième chapitre nous passons à la conception de notre étude est méthodiquement organisé. Nous décrivons les différents étapes de travail. Chaque étape est cruciale pour la création d'un modèle efficace et fiable. Cette partie inclut une analyse détaillée des différentes phases de notre étude, soulignant les défis rencontrés et les solutions apportées pour surmonter ces obstacles.

Le quatrieme se focalise sur l'implementation et les résultats obtenus après l'application des différents modèles étudiés.

Chapitre 1

La volatilité financière

I.1. Introduction

En finance, la volatilité est une variable latente et inobservable qui doit être calculée pour indiquer dans quelle mesure la valeur d'un actif a changé sur une période donnée. La volatilité d'un actif peut donc s'expliquer par son niveau d'incertitude. Il s'agit donc d'une variable extrêmement importante pour les décisions d'investissement et la détermination des prix des actifs, en particulier des produits dérivés. Pour les institutions financières mondiales, il est nécessaire de prévoir la volatilité future afin de déterminer leurs vulnérabilités actuelles.

Sur les marchés des changes, la volatilité est un facteur très important pour la grande majorité des entreprises locales et internationales, ainsi que pour les économies des pays où la valeur des devises est déterminée par le marché. Les entreprises gèrent les risques associés aux fluctuations du marché des changes, car il s'agit d'un risque supplémentaire puisqu'elles doivent souvent posséder plusieurs devises ou plusieurs actifs eux-mêmes rattachés à plusieurs devises différentes pour pouvoir exercer leurs activités (Beauchamp et al., 2017).Si les entreprises internationales ne se protègent pas des risques de volatilité, plusieurs valeurs comptables telles que les actifs, les passifs et les bénéfices nets peuvent être directement affectées, soit positivement, soit négativement.

I.2. Contexte historique

Bachelier (Bachelier, 1900) affirmait : « La dynamique boursière ne peut jamais être une science exacte. » Cette affirmation a conduit certains auteurs à l'associer à la notion d'incertitude dans les fluctuations boursières. Après Bachelier, Markowitz (Markowitz, 1952) a introduit le concept de variance dans le domaine financier et a développé la théorie moderne du portefeuille. Il a démontré qu'en possédant différents actifs sous-jacents, les investisseurs pouvaient atténuer le risque financier de leur portefeuille, réduisant ainsi la volatilité de ce portefeuille. À la suite des travaux de Markowitz(Markowitz, 1952), Sharpe(Sharpe, 1964) a développé le modèle théorique d'évaluation des actifs financiers (MEDAF), qui utiliserait le ratio de Sharpe pour évaluer la performance d'un portefeuille d'actifs. par rapport au financier en termes de risque (volatilité). Même si les théories de Markowitz et Sharp étaient révolutionnaires pour leur époque, elles continueront à être la cible de nombreuses critiques. En fait, leur théorie repose sur le principe selon lequel les cours des actions sont des phénomènes continus, alors qu'en réalité les marchés financiers sont soumis à des degrés divers de changements forts, ce qui implique des phénomènes discontinus. La plupart des critiques associent ces phénomènes aux dangers que représentent des événements incertains.

De la théorie financière de Bachelier, Markowitz et Sharpe ont émergé de nouvelles théories, comme la théorie de la valorisation des options de Black et Scholes(Black & Scholes, 1973; Kouaga, n.d.), dont le concept de volatilité inconditionnelle était très controversé en raison de sa variance constante. Une faille dans l'approche utilisée par ces différents auteurs a été identifiée par Engle(Engle, 1982), qui soutenait que les séries financières sont toutes normalement distribuées avec une variance constante. La plupart des critiques associent ces phénomènes aux dangers que représentent des événements incertains. À partir de la théorie financière de Bachelier, Markowitz et Sharpe, de nouvelles théories ont émergé, comme la théorie de l'évaluation des options de Black et Scholes, dont le concept de volatilité inconditionnelle était controversé en raison de sa variance constante. Une faille dans les méthodes utilisées par ces différents auteurs a été découverte par Engle (Engle, 1982), qui soutenait que les séries financières sont toutes normalement distribuées avec une variance constante. Elle apportera donc une contribution significative au concept de volatilité en développant la famille de modèles ARCH, qui prend en compte certains facteurs tels que : l'information, les événements à faible probabilité et le temps. Les travaux de Bollerslev(Bollerslev, 1986) permettront de généraliser le modèle ARCH(p) d'Engle en développant des modèles GARCH (p, q) plus efficaces. Kermiche (Kermiche, 2008) a démontré que la volatilité implicite d'une option est la volatilité à laquelle la formule de Black-Scholes est égale à la valeur marchande de l'option, en supposant que les autres paramètres de la formule sont connus à l'avance. L'inversion de la formule de Black-Scholes permet d'obtenir une volatilité implicite largement utilisée en finance. Afin d'apporter des solutions aux lacunes observées dans le modèle de Black et Scholes, Hull et White (Hull & White, 1987) mettront l'accent sur les modèles stochastiques avec effets d'aplatissement, c'est-à-dire des données financières distribuées avec des effets positifs. Aplatissement excessif avec de grandes variations. Selon Taylor(Taylor, 1986), l'hypothèse selon laquelle la volatilité dans le modèle de Black et Scholes a toujours une valeur constante peut être résolue par l'une des méthodes des modèles stochastiques, en considérant dans ce modèle un certain nombre d'imprévus. des événements peuvent survenir. Impact sur les marchés financiers. Contrairement aux modèles de volatilité implicite et de volatilité historique, les modèles de volatilité stochastique utilisent la volatilité comme une variable inobservable dont les propriétés stochastiques sont difficiles à prouver. La volatilité est une variable aléatoire qui évolue dans le temps et sa propagation peut provoquer divers chocs économiques. Par exemple, la crise des prêts hypothécaires à risque aux États-Unis en 2007 a été déclenchée par

de violentes fluctuations du marché immobilier. Les travaux de Bensafta et Gervasio sur la transmission et la contagion de la volatilité entre les marchés boursiers démontrent l'importance de la volatilité dans la crise des subprimes et dans la mondialisation des marchés. Ils croient que les marchés financiers du monde entier sont interdépendants. Par ailleurs, Ross(Ross, 1989) et Gharbi (Gharbi, 2013) conviennent que la dynamique des fluctuations sera toujours affectée par certaines incertitudes : macroéconomiques, géopolitiques, structurelles et cycliques.

I.3. Théorie et mesure de la volatilité

Les marchés financiers ont un concept clé en finance qui est la volatilite reflétant l'évolution des prix des actifs. Elle est étudiée à travers diverses théories économiques et mesurée par différentes méthodes. Cette section explore les principales théories et évolutions de la volatilité.

Théorie économique sur la volatilité

En finance tout comme en économie, les fluctuations des marchés financiers occupent une place centrale. Différentes théories économiques ont été élaborées afin de comprendre et donner des explications la nature et le fonctionnement de la variation des prix des actifs.

Theorie du marche efficient

Quand le terme « marché efficient » a été introduit dans la littérature économique il y a trente ans, il a été défini comme un marché qui « s'ajuste rapidement à de nouvelles informations » (E. F. . Fama et al., 1969; Malkiel, 2003).

Il est rapidement apparu, toutefois, que si un marché performant dépend d'un ajustement rapide aux nouvelles informations, ce n'est pas le seul. Selon Fama(E. F. Fama, 1991; Malkiel, 2003), une définition plus contemporaine est que les prix des actifs sur un marché performant "représentent pleinement toutes les informations disponibles". Cela signifie que les informations sont traitées de manière rationnelle par le marché, ce qui signifie que les informations pertinentes ne sont pas négligées et que les erreurs systématiques ne sont pas commises. En conséquence, les prix restent toujours en accord avec les "essentiels".

Les mots de cette définition ont été choisis avec soin, mais ils masquent néanmoins certaines des subtilités inhérentes à la définition d'un marché d'actifs efficient.

D'une part, c'est une version affirmée de l'hypothèse qui ne pourrait être vraie que si « toutes les informations disponibles » étaient obtenues gratuitement. Si l'acquisition de l'information était au contraire coûteuse, il serait nécessaire de l'encourager financièrement. Cependant, si l'information était déjà "intégralement reflétée" dans les prix des actifs, il n'y aurait pas d'encouragement financier (Grossman, S.; Stiglitz, 1980; Malkiel, 2003). Selon Jensen(Jensen, 1978; Malkiel, 2003), une hypothèse plus faible, mais économiquement plus réaliste, est que les prix reflètent l'information jusqu'à ce que les bénéfices marginaux d'agir sur l'information (les profits attendus) ne dépassent pas les coûts marginaux de sa collecte.

Deuxièmement, que signifie dire que les prix sont cohérents avec les fondamentaux ? Il nous faut un modèle pour établir un lien entre les fondamentaux économiques et les prix des actifs. Bien qu'il existe des modèles candidats dans tous les marchés d'actifs qui fournissent ce lien, personne n'est sûr que ces modèles capturent pleinement ce lien de manière empiriquement convaincante. Cela est important car les tests empiriques de l'efficience du marché en particulier ceux qui analysent les rendements des prix des actifs sur de longs horizons sont inévitablement des évaluations conjointes de l'efficacité du marché et d'un modèle particulier de prix des actifs. Lorsque l'hypothèse conjointe est rejetée, comme c'est souvent le cas, il est logiquement possible que cela soit dû à des difficultés dans le modèle spécifique de prix des actifs plutôt que dans l'hypothèse du marché efficient. C'est le problème du « mauvais modèle » (E. F. Fama, 1991; Malkiel, 2003).

Enfin, un commentaire sur le mot « efficient ». Il semble que le terme ait été choisi initialement en partie parce qu'il établit un lien avec le concept économique plus large de l'efficience dans la réservation des ressources Fama a donc entamé son analyse de l'hypothèse en 1970 de marché efficient (spécifiquement appliquée au marché boursier) par :

La principale fonction du marché des capitaux consiste à répartir la propriété du stock de capital de l'économie. Sur le plan général, l'idéal est un marché où les prix fournissent des indications précises sur l'allocation des ressources : c'est-à-dire un marché où les entreprises peuvent prendre des décisions de production et d'investissement, et les investisseurs peuvent choisir parmi les titres qui représentent la propriété des activités des entreprises en supposant que les prix des titres à tout moment « reflètent pleinement » toutes les informations actuelles.

Il peut sembler assez naturel que le lien entre un marché d'actifs qui reflète l'information disponible de manière efficace et sa fonction dans l'optimisation de la répartition des ressources

Cependant, une étude plus approfondie a révélé qu'un marché d'actifs efficace sur le plan informatif n'a pas forcément besoin de générer une efficacité allocative ou productive dans l'économie de manière plus générale. Selon Stiglitz(Malkiel, 2003; Stiglitz, 1981), les deux concepts diffèrent en raison de l'incomplétude des marchés et du rôle révélateur des prix lorsque l'information est coûteuse.

Mesure de la volatilité

Il existe deux concepts de volatilité. Premièrement, les mouvements inconditionnels ou historiques ne sont que des observations rétrospectives des mouvements passés des taux de change. Cette fluctuation historique est en réalité la somme de fluctuations inattendues dues à des événements de type « actualité », une autre composante étant les fluctuations conditionnelles. La volatilité est un phénomène continu et le modèle de Garch postule un processus de comportement volatil, permettant d'isoler les composantes conditionnelles de la volatilité auxquelles on pourrait s'attendre. Elle peut donc être comparée à la volatilité implicite des options de change, qui sont essentiellement des prévisions de marché prospectives.

Volatilité historique ou non conditionnelle

Dans des études antérieures, la volatilité historique au moment t est souvent définie comme la volatilité réalisée au moment t-1. Si la technique de mesure mentionnée ci-dessus est suivie, alors l'information contenue dans l'écart entre les deux contrats consécutifs serait ignorée(Hansen, 2001; Padhi & Shaikh, 2014). Ainsi, pour la présente étude, une définition différente de la volatilité historique(Christensen, B. J.; Hansen, 2002; Hansen, 2001; Li, S.; Yang, 2009; Padhi & Shaikh, 2014) est utilisée pour un contrat donné avec T jours jusqu'à l'échéance au moment t. La volatilité historique correspondante est calculée en utilisant le rendement quotidien logarithmique de la période remontant à T jours à partir du moment t comme suit :

t-1 =, Équation 11

Ici = désigne la moyenne du rendement quotidien logarithmique de l'indice au moment t-1.

Volatilité implicite

Selon Padhi et Shaikh (Padhi & Shaikh, 2014), le marché considère la volatilité implicite comme une estimation de la volatilité future d'un actif financier. Les prix actuels des options sur cet actif sont utilisés pour calculer la volatilité implicite, contrairement à la volatilité historique, qui repose sur les fluctuations passées des prix. Elle se calcule par des modèles

d'évaluation d'options, comme le modèle de Black-Scholes tient compte de divers facteurs, tels que le prix de l'actif sous-jacent, le prix de l'option, le temps restant jusqu'à la date d'échéance de l'option, le taux d'intérêt sans risque et le prix d'exercice de l'option en elle-même.

On utilise fréquemment la volatilité implicite pour évaluer la perception du marché concernant l'incertitude future des prix de l'actif sous-jacent.

La volatilité implicite désigne la fluctuation actuelle d'une action telle qu'elle est représentée par le prix de ses options financières. Il est difficile d'inverser les modèles d'évaluation des options, ce qui rend la volatilité implicite numériquement calculée. En utilisant le modèle d'évaluation des options BS (Black-Scholes), la volatilité implicite est évaluée à l'aide de la méthode de la dichotomie, de la manière suivante.

:

Volatilité estimée= $_{L+}$ Équation I2

Ici $_{et}$ sont respectivement les valeurs de volatilité basse et haute, C_L et C_H sont les valeurs d'options correspondantes et C est le prix du marché de l'option.

Ecart-type

En règle générale, on évalue la volatilité en utilisant l'écart-type ( ) ou la variance ( ) des rendements logarithmiques.

= Équation I3

Où : P_t est valeur du prix au temps t.

Différentes méthodes sont disponibles pour mesurer la volatilité. Prenons l'exemple de la différence entre les rendements ou la valeur minimale et maximale obtenue pendant une période spécifique. Le modèle GARCH et ses différentes versions sont des modèles de prédiction de la volatilité, font appel au carré des rendements logarithmiques. Davidian et Caroll (Davidian, M.; Carroll, 1987) ont estimé que la méthode basée sur les rendements absolus est l'une des plus solides. Cette approche peut aussi parfois être plus efficace pour prédire la volatilité(Beauchamp et al., 2019; Forsberg, L.; Ghysels, 2007; Taylor, 2007).

Le modèle ARCH

Une nouvelle catégorie de modèles autorégressifs conditionnellement hétéroscédastiques (ARCH) est proposée par Engle (Engle, 1982) qui intégrent cette caractéristique dans le calcul de la volatilité, c'est-à-dire la variance conditionnelleLa notion de "variance conditionnelle"

signifie que la prédiction actuelle de la variance est en relation avec la variance précédente. Prenons l'exemple du processus suivant pour améliorer le rendement d'un actif. :

Équation I4

L'innovation peut être décomposée par :

Ou est volatilité conditionnelle, qui correspond à la prédiction du rendement, calculée en utilisant toutes les informations disponibles au temps t-1. Cette moyenne conditionnelle est représentée par un bruit blanc gaussien qui suit une loi normale avec une moyenne nulle et une variance réduite à 1. Le modèle ARCH(q) de la variance conditionnelle est comme suit :

Équation I5

Où et sont des paramètres de valeurs positives pour éviter une variance négative.

Le modèle GARCH

Le modèle ARCH ne peut pas être utilisé pour identifier les propriétés liées à la volatilité d'un nombre maximal de données précédentes. On constate que lorsque nous souhaitons incorporer davantage d'innovations antérieures dans notre modèle, le nombre de paramètres augmente de manière proportionnelle. Plus précisément, un modèle ARCH présente une grande similitude avec un modèle de moyenne mobile. D'après Bollerslev (1986), le modèle GARCH (Generalized Autoregressive Conditional Heteroskedasticity) a été présenté. Ce modèle considère à la fois les avancées antérieures et les variations conditionnelles antérieures, ce qui permet de collecter un maximum d'informations sur la volatilité.Le modèle GARCH(p,q) de la variance conditionnelle est tel que :

Équation I6

Où

La contrainte initiale consiste à garantir un processus stationnaire, tandis que les autres exigent une variance conditionnelle positive. Les paramètres du modèle sont déterminés en considérant la distribution normale des données et en maximisant la somme des valeurs des fonctions de vraisemblance logarithmique (Loglikelihood function, ou LLF en anglais).

Équation 17

Où représente la variance conditionnelle et rt le rendement logarithmique de l'actif. Les modèles de type GARCH sont couramment utilisés avec des données quotidiennes, intra journalières et parfois hebdomadaires. Toutefois, lorsque les données ont un intervalle plus élevé, comme mensuel(Baillie & Bollerslev, 2002), il est prouvé que le phénomène de regroupement de la volatilité disparaît, ce qui entraîne une diminution de l'efficacité d'un modèle GARCH.

On a examiné plusieurs autres versions du modèle GARCH. Selon Nelson(Beauchamp et al., 2019; Nelson, 1991), pour prendre en compte l'effet d'asymétrie des variations, c'est-à-dire l'observation selon laquelle la volatilité semble être davantage influencée par des variations négatives que par des variations positives de mêmes valeurs, le modèle EGARCH (Exponential GARCH) a été développé. Selon Hamilton et Susmel(Hamilton & Susmel, 1994), il est également proposé de modéliser la volatilité en utilisant différents régimes et états. La volatilité actuelle est généralement déterminée par le niveau de volatilité actuel. Les conditions optimales d'un modèle peuvent varier d'un état à l'autre. Par exemple, lorsqu'il y a une forte volatilité, les paramètres optimaux seront différents de ceux d'un état où la volatilité est faible. Il est essentiel de considérer ces différents ensembles de paramètres optimaux, tout en assignant de nouveaux paramètres qui indiquent les probabilités d'être dans un état ou un autre, en tenant compte de toutes nos données antérieures. Selon Marcucci (Marcucci, 2005), il a été démontré que ce type de modèle GARCH qui peut changer de régime (Regime Switching) pouvait potentiellement être bien plus efficace que les modèles GARCH traditionnels tels que GARCH(1,1) et EGARCH(1,1) pour prédire la volatilité sur une période inférieure à dix jours. Il convient de souligner qu'aucune variante d'un modèle GARCH ne peut être considérée comme la meilleure dans son ensemble. Il est essentiel de procéder à une analyse préliminaire des données étudiées pour déterminer quel type de modèle le mieux approprie.

I.4. Impact sur la liquidité

Un marché est qualifié de liquide lorsque des transactions peuvent être effectuées rapidement et des quantités importantes d'actifs peuvent être achetées ou vendues sur celui-ci sans avoir de frais de transactions élevés. Le spread, également connu sous le nom d'écart offre-demande, est une des méthodes utilisées pour évaluer la liquidité. C'est la différence entre le prix d'achat le plus bas et le prix de vente le plus élevé pour un actif spécifique. Le spread est un frais de transaction qui se manifeste par une perte pour le participant et un bénéfice pour le courtier, et qui est payé immédiatement après avoir effectué une transaction d'achat ou de vente. À mesure que celui-ci diminue, le marché est perçu comme plus liquide.

Un marché plus fluctuant peut attirer certains investisseurs vers celui-ci en raison du potentiel d'y réaliser des bénéfices. Un marché devient plus attrayant pour un courtier lorsque la demande augmente, c'est-à-dire qu'il y a un plus grand nombre de participants et de transactions, dans le but de réaliser des bénéfices, notamment grâce au spread. La base du marché FX est constitué d'un grand nombre de participants et de transactions quotidiennes pour répondre aux besoins réels d'échange de devises, ce qui entraîne une grande volatilité. Ce niveau de volatilité attire également certaines institutions et investisseurs individuels qui souhaitent en bénéficier, ce qui accroît encore plus le volume et la volatilité. La présence d'un grand nombre de courtiers sur ce marché, les participants des écarts offre-demande généralement très faibles en raison de la compétition entre les courtiers. La plupart des produits disponibles sur ce marché sont donc considérés comme étant extrêmement liquides.

En revanche, lorsque la volatilité est extrêmement élevée, un courtier est exposé à un risque accru(Zhang et al., 2008). Par exemple, lorsque de nouvelles économies importantes sont annoncées, un courtier a la possibilité d'ouvrir automatiquement les positions de certains clients en raison de pertes subies et d'un manque de capital dans le compte. Il est possible de fermer à un prix qui établit la perte à un montant supérieur au capital disponible dans le compte du client, ce qui entraîne une perte pour le courtier d'un montant égal à la différence entre le capital disponible dans le compte du client et le montant total de la perte(Stoll, 1978). Afin de compenser le risque supplémentaire lié à un actif spécifique, les courtiers vont augmenter le spread, ce qui entraîne une baisse de la liquidité sur le marché(Frankel et al., 1996). Bien que cela puisse être préjudiciable lorsque la volatilité atteint un certain niveau, celle-ci apporte au marché plus d'avantages que d'inconvénients..

I.5. Revue empirique de la prédiction de la volatilité

D'apres Nelson(Nelson, 1991) et Rabemananjara et Zakoian(Kouaga, n.d.; Rabemananjara & Zakoian, 1993), l'hypothèse de non-linéarité demeure essentielle pour prédire la fluctuation de certaines bourses. Selon Engle (Engle, 1982) et Bollerslev (Bollerslev, 1986), les modèles de prédiction ARCH et GARCH supposent une évolution asymétrique de la variance et non symétrique. La variation des marchés financiers ces dernières années laisse entendre que ces modèles ne sont pas suffisamment capables de prédire les variables qui présentent des phénomènes d'asymétries. Cappiello(Cappiello, 2006) et Elizabeth (Elizabeth, 2012) attribuent ces changements à la mondialisation des marchés et aux divers chocs économiques, qui sont responsables des fortes variations et des phénomènes d'asymétrie. Dans leur étude de 2012, Chikhi et ses collègues ont démontré que les modèles de prédiction SEMIFARMA (Semi-parametric Fractional Autoregressive Moving Average) surpassent les modèles EGARCH en termes de prédiction de la volatilité des indices boursiers à mémoire très longue, tels que le Dow Jones. Outre les phénomènes de déséquilibres, ils démontrent également que les modèles de prédiction SEMIFARMA tiennent compte des phénomènes de persistance des variations de volatilité provoquées par les chocs informationnels. La combinaison des modèles utilisés permettra de prédire la fluctuation des rendements des taux de change EUR/USD. Ces modèles se distinguent par leur capacité à intégrer les facteurs non linéaires à court et à long terme d'une série financière. Cela les rend encore plus efficaces par rapport aux modèles de prédictions autorégressifs GARCH et EGARCH. Toutefois, Hagen et Yu(Bluhm & Yu, 2001) emploieront des méthodes distinctes telles que les séries chronologiques univariées et celles de la volatilité implicite des prix des options afin de prédire la volatilité sur les marchés financiers en Allemagne. Ils concluent que ces modèles de prévision sont sensibles aux erreurs de mesure et aux différentes approches prédictives. Dans son analyse visant à prédire la volatilité quotidienne du rendement des actions négociées à la NYSE (New York Stock Exchange), Brooks(Brooks, 1998) examine divers modèles statistiques de prédiction et constate que l'amélioration des prévisions de performance des marchés avec des mesures de volume retardées est très limitée. Selon une étude menée par Martens et Jason(Martens & Zein, 2002), il constate que les modèles de prévisions GARCH ne contiennent pas au moins autant d'informations que ceux des prévisions à volatilité implicite. Le modèle EGARCH (1,1) présente des performances supérieures pour prédire la volatilité des indices boursiers. Le modèle GARCH (1,1) se révèle plutôt efficace pour prédire la volatilité des devises.

Il est difficile de déterminer avec précision l'horizon de prévisibilité de la volatilité, car il semble plus court que long. Il semble que leurs diverses recherches montrent que la volatilité peut être anticipée sur des périodes à long terme (30 à 60 mois). Selon West et Cho(West & Cho, 1995), les modèles de prédiction GARCH montrent une certaine précision par rapport aux modèles de prédiction homoscédastiques et aux modèles non paramétriques univariés pour prédire la variance conditionnelle de la volatilité des marchés sur une période d'une semaine. De même, Brailsford et Faff (1996) examinent différents modèles de prédiction de la volatilité et reconnaissent que le modèle GARCH est plus efficace pour prédire la volatilité à court terme (1 mois). Toutefois, le modèle GARCH présente des performances moindres dans la prédiction à long terme (6 mois). Il est apparu que le modèle de prédiction GARCH (1,1) semble plus efficace pour prédire la volatilité de l'indice TSE-300 à court terme (1 mois). Ils montrent, d'autre part, que cette prédiction diffère en présence d'autres modèles. Selon Farés(Farès, 2008), il a été démontré que le modèle EWMA (moyenne mobile exponentielle pondérée) est plus performant que le modèle GARCH (1,1) pour prédire la volatilité de l'indice S&P500 à long terme. Selon Black et Scholes(Black & Scholes, 1973), Racicot et Théoret(Racicot & Théoret, 2005), il est prouvé que la volatilité historique peut être considérée comme un indicateur précis de la volatilité future (anticipée). Cependant, Ross (1989) souligne l'importance de considérer, en plus des données historiques, les phénomènes aléatoires liés à la volatilité afin de prédire plus efficacement les volatilités à venir. La prédiction de la volatilité future des options de l'indice S&P 100 reste inefficace grâce au modèle de volatilité implicite . Le modèle de volatilité implicite semble plus pertinent que le modèle de volatilité historique et le modèle GARCH pour prédire la volatilité future, à la fois sur le marché de gré à gré et sur les marchés boursiers de Hong Kong, selon Yu(Bluhm & Yu, 2001). Le modèle de la marche du hasard de Bachelier(Bachelier, 1900) est en accord.

I.6. Conclusion

En résumé, on peut dire que le modèle GARCH est mieux adapté pour prédire la volatilité à court terme que à long terme, tandis que le modèle EGARCH est plus approprié pour prédire la volatilité de certaines séries financières asymétriques. Il en va de même pour la volatilité implicite, qui est plus performante pour anticiper la volatilité future que la volatilité historique. Cependant, nous notons que les prévisions de fluctuations importantes sur les marchés financiers peuvent varier considérablement selon les différents modèles économétriques, ce qui explique le caractère hautement improbable de la volatilité.

Chapitre 2

Apprentissage automatique

II.1. Introduction

L'apprentissage automatique, également connu sous le nom de machine learning (ML) en anglais, est une forme d'intelligence artificielle (IA) qui permet au système d'apprendre à partir de données plutôt que de les programmer séparément. Au quotidien, l'apprentissage automatique est particulièrement répandu et est mis en pratique dans divers secteurs. Différents exemples nous permettent de mettre en évidence l'importance et les conséquences de cette technologie dans notre quotidien. Cette technologie semble être un élément indispensable pour les entreprises qui veulent améliorer la compréhension de leurs clients et ainsi satisfaire leurs besoins. Le but de l'apprentissage automatique est donc de former un algorithme afin de traiter de manière pertinente et efficace les diverses informations disponibles. L'apprentissage automatique se subdivise en deux étapes principales : l'apprentissage et la prédiction.

Ce chapitre a abordé l'apprentissage automatique et ses différentes utilisations.

II.1.1. Types d'apprentissage automatique

Différents types d'apprentissage automatique existent, chacun possédant des caractéristiques distinctes et répondant à des besoins particuliers. Il est crucial de saisir ces diverses catégories afin de sélectionner la meilleure méthode en fonction du problème à résoudre. La figure II3 illustre les differents types d'apprentissage automatique.

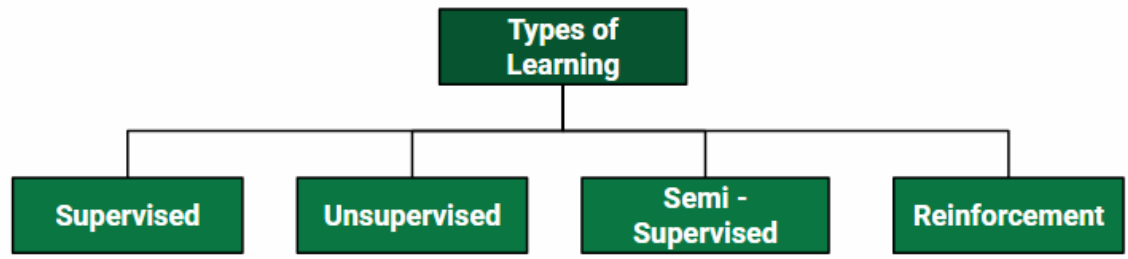

Figure II3: Les types de Machine Learning

II.1.1.1. L'apprentissage supervise

L'apprentissage supervisé est une technique d'apprentissage qui vise à prédire la relation entre les entrées et les sorties, en essayant de trouver des modèles dans ensembles de données étiquetés. Enseignement supervisé est également appelée fonction de mappage, car elle définit des modèles cachés dans la donnée. Dans l'apprentissage supervisé, il existe un ensemble de données étiquetées composé de données d'entrée paires et une sortie cible. La sortie cible est une valeur associée au des données d'entrée. Un algorithme d'apprentissage supervisé traite la formation étiquetée ensemble de données pour prédire de nouveaux ensembles de données

similaires ou la fonction de cartographi(Ayyıldız, 2023). Supervisé l'apprentissage peut accomplir deux tâches principales :

Classification : Le filtre anti-spam est un exemple classique : il est élaboré à partir de nombreux exemples d'e-mails classés en spam ou non-spam (ham) et doit s'adapter à la classification adéquate des nouveaux e-mails, comme illustré dans la figure ci-dessous (Géron, 2019).

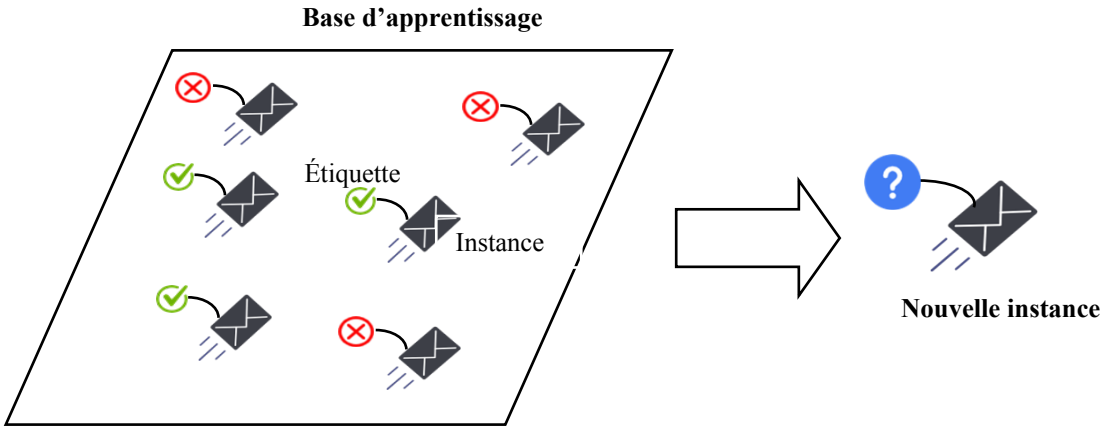

Figure II4: Schéma d'apprentissage supervisé "supervised learning" (exemple : la classification de spam)

Régression : Une autre activité courante est de prédire une valeur numérique cible, comme le prix d'une voiture, en se basant sur un ensemble de paramètres (kilométrage, âge, marque, etc.) connu sous le nom de prédicteurs. Cette forme de travail est appelée régression. Il est nécessaire de fournir au système de nombreux exemples de voitures avec leurs prédicteurs et leurs étiquettes (c'est-à-dire leurs prix) afin de l'entraîner. La figure ci-dessous représente une opération de régression(Géron, 2019).

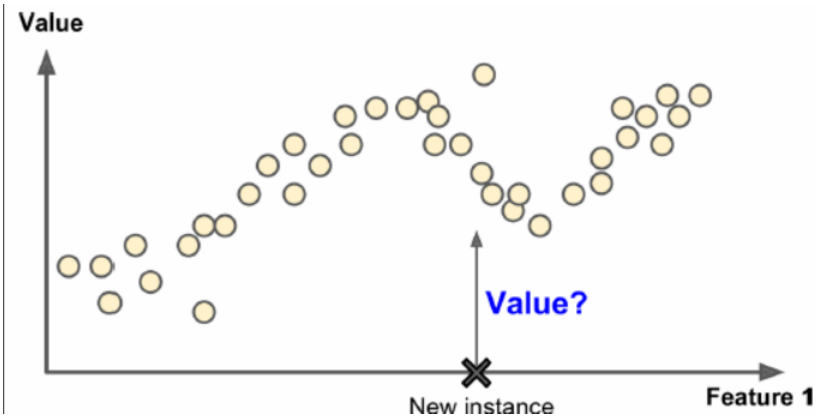

Figure II5: schéma de la régression

II.1.1.2. L'apprentissage non supervise

Dans les cas où les données ne sont pas correctement classées et ne sont pas étiquetées observations, il n'est pas possible d'effectuer un apprentissage supervisé(Ayyıldız, 2023). Sans surveillance les modèles d'apprentissage sont souvent utilisés dans les applications d'exploration de données qui impliquent de grands volumes de données d'entrée non structurées. L'apprentissage est une technique d'apprentissage qui se concentre sur les points communs plutôt que sur répondre aux commentaires. La méthode vise à déterminer la probabilité de points communs dans un ensemble de données spécifique et utilise ces points communs pour développer un modèle. L'algorithme, conformément à son objectif, apprend à accomplir une tâche sans fournir une approche logique pour faire quelque chose. Donc, l'approche non supervisée est plus complexe que le processus supervisé. L'apprentissage non supervisé peut également être décrit comme utilisant une méthode basée sur la récompense. Approche pour confirmer le succès de la réalisation d'objectifs spécifiques sans fournissant des instructions explicites sur la manière d'atteindre ces objectifs. Dans l'apprentissage non supervisé, il n'y a pas de concepts des sorties correctes ou incorrectes ; au lieu de cela, des similitudes, des différences et des modèles sont représentés mathématiquement. Les algorithmes d'apprentissage non supervisé sont fréquemment utilisés pour des applications telles que le clustering, l'association et les anomalies détection.

II.1.1.3. Apprentissage semi-supervisé

Les algorithmes d'apprentissage semi-supervise peuvent gérer des données d'entraînement partiellement étiquetées, généralement avec une grande quantité de données non étiquetées et une faible quantité de données étiquetées, et prédire tous les points non vus. Dans les cas où les données non étiquetées sont facilement accessibles mais où l'acquisition des étiquettes est coûteuse, l'apprentissage semi-supervisé est souvent utilisé. Il est possible de présenter plusieurs types de problèmes rencontrés dans les applications, tels que les tâches de classification, de régression ou de classement. Selon l'espoir, la répartition de données non étiquetées accessibles à l'algorithme pourrait lui permettre d'obtenir des performances supérieures à celles de l'apprentissage supervisé. De nombreuses recherches théoriques et appliquées contemporaines sur l'apprentissage automatique se concentrent sur l'analyse des conditions dans lesquelles cela peut être accompli(P-.Chen, 2019).

II.1.1.4. Apprentissage par renforcement

Le scénario global de l'apprentissage par renforcement est illustré dans la Figure II 6. Contrairement à l'enseignement supervisé, l'apprenant ne dispose pas passivement d'un ensemble de données étiquetées. Au lieu de cela, il recueillit des informations grâce à un processus d'action en interagissant avec le milieu. En réponse à une action, l'apprenant ou l'agent reçoit deux informations : son état actuel dans l'environnement et une récompense à valeur réelle, spécifique à la tâche et à son objectif correspondant. Le but de l'agent est d'améliorer sa récompense et donc de trouver le plan d'action ou la politique la plus efficace pour atteindre cet objectif. Cependant, les données qu'il obtient de l'environnement ne sont que la récompense immédiate de l'action qu'il vient de réaliser. La prise en compte de récompenses ou de pénalités différées est un élément essentiel de l'apprentissage par renforcement. Un dilemme se pose à l'agent : explorer des états et des actions inconnus pour obtenir davantage d'informations sur l'environnement et les récompenses, ou exploiter les informations déjà collectées pour optimiser sa récompense. Il s'agit du compromis entre l'exploration et l'exploitation, qui est associé à l'apprentissage par renforcement(P-.Chen, 2019).

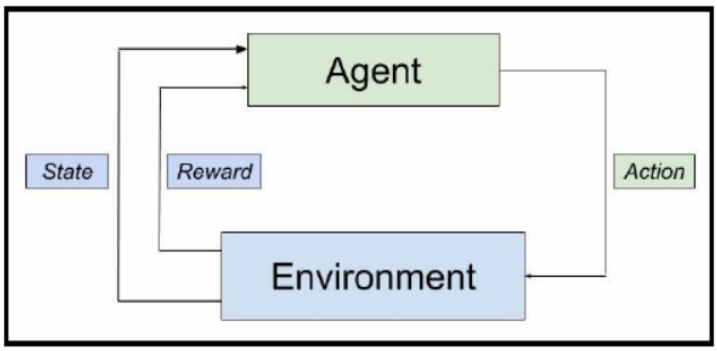

Figure II6:Apprentissage par renforcement

II.1.2. Algorithmes des machines Learning

Les algorithmes d'apprentissage automatique sont généralement divisés en plusieurs catégories en fonction du type de tâche qu'ils sont censés résoudre : classification, régression, clustering**.** Dans notre étude, nous avons sélectionné une gamme d'algorithmes qui représentent une tache de classifications car nous sommes confrontés a un problème de classification, chacun offrant une manière unique de modéliser la relation entre les données d'entrée et les sorties souhaitées. Voici une description détaillée des principaux algorithmes que nous utilisons :

II.1.2.1. Support Vector Machine (SVM)

L'algorithme Support Vector Machine (SVM) est un modèle d'apprentissage supervisé et l'algorithme d'apprentissage associé utilisé dans l'apprentissage automatique. Il est utilisé pour l'analyse de classification et de régression en analysant des données de grande dimension. Le SVM a été développé aux laboratoires AT&T Bell par Vladimir Vapnik et ses collègues (Niu, 2020; Vapnik & Cortes, 1995). Le SVM a été utilisé dans de nombreuses études pour prédire les cours boursiers(Henrique, 2018; Kim, 2003; Niu, 2020; Sapankevych, 2009). Il s'agit d'un type d'algorithme d'apprentissage supervisé qui peut être utilisé, entre autres tâches, pour des problèmes de classification binaire. Dans SVM, les points de données sont représentés sous forme de vecteurs dans un espace multidimensionnel et l'objectif est de déterminer si les points de données peuvent être séparés par un hyperplan. L'objectif est de trouver l'hyperplan qui maximise la marge, qui est la distance entre l'hyperplan et les points de données les plus proches de chaque classe. Le SVM minimise simultanément l'erreur de classification empirique et maximise la marge géométrique. Pendant la formation, SVM recherche un hyperplan séparateur qui sépare les points des deux classes. On suppose que plus la marge ou la distance entre les hyperplans est grande, meilleure est l'erreur de généralisation du classificateur. Si nous représentons les données d'entraînement D comme suit :

$$D:=\{(x1,y1),(x2,y2),\ldots,(xn,yn)\} \qquad \text{Équation II2}$$

où $yi\in\{1,-1\},i\in 1\ldots n$ est une constante représentant la classe à laquelle appartient le point xi et n désigne le nombre d'échantillons. Chaque xi est un vecteur réel de dimension p. La mise à l'échelle est importante pour protéger les fonctionnalités (attributs) avec des variances plus élevées. Ces données d'apprentissage sont obtenues à l'aide de l'hyperplan séparateur donné par $w\cdot x+b=0$, où b est un scalaire et w est un vecteur de dimension p. Le vecteur w est perpendiculaire à l'hyperplan séparateur. L'inclusion du paramètre b permet d'augmenter la marge. Sans b, l'hyperplan devrait passer par l'origine, limitant ainsi la solution. Les hyperplans parallèles peuvent être définis comme suit :

$$w\cdot x+b=1$$

Et $$\text{Équation II3}$$

$$w\cdot x+b=-1$$

Le résultat de l'entraînement T est un sous-ensemble des données d'entraînement ($T \subset D$). Les échantillons $xi \in T$ se trouvent le long des marges et sont appelés vecteurs de support. Comme les vecteurs de support représentent les échantillons le long de l'hyperplan de séparation, l'hyperplan peut être exprimé comme la superposition des vecteurs de support.

W= Équation II4

L'hyperplan avec la plus grande marge, défini par $M=2|w|$, satisfera l'équation suivante :

$$(w \cdot x_i + b) = 1, \forall i$$ Équation II5

où $x_i \in T$ est un vecteur support, et pour un hyperplan optimal

$$(w \cdot x_j + b) \geq 1, \forall j \in 1..n$$ Équation II6

où n est le nombre de points de données d'entraînement. Pour trouver l'hyperplan optimal avec une marge maximale, pendant le processus de formation, le modèle doit minimiser $\|w\|2$ sous réserve de contraintes d'inégalité (4). Ce problème d'optimisation peut être résolu en utilisant la méthode d'optimisation primale-duale de Lagrange. Dans de nombreux cas, cependant, il n'existe pas d'hyperplan dimensionnel $n-1$ optimal séparant les points d'échantillonnage des données d'apprentissage. Pour séparer les échantillons, nous pouvons utiliser une « astuce du noyau ». L'« astuce » consiste à remplacer l'opérateur produit scalaire par une fonction produit scalaire K d'un espace de dimension supérieure. En effet, l'espace du noyau peut même être de dimension infinie. L'équation II-6 devient :

Yj≥1,∀j∈1..n Équation II7

Nous pouvons imaginer la méthode comme intégrant les vecteurs d'échantillon dans l'espace de dimension supérieure. Après l'intégration, nous recherchons l'hyperplan optimal dans l'espace plus large. La fonction de noyau la plus utilisée est la fonction de noyau Radial Basis Function (RBF) :

$$(xi, xj) = e^{-\gamma \|xi - xj\|^2}$$ Équation II8

Où $\gamma > 0$ est le paramètre du noyau. La caractéristique du noyau RBF est qu'il mappe de manière non linéaire les échantillons dans un espace de dimension supérieure. Son utilisation est largement répandue dans divers scénarios problématiques (Durgesh, 2010). Dans notre cas, le modèle SVM est utilisé comme classificateur binaire qui classe le mouvement des prix comme favorable (à la hausse) ou défavorable (pas à la hausse). La formation « tente

d'identifier » des modèles dans les données qui indiqueraient une hausse soudaine du prix d'un instrument particulier.

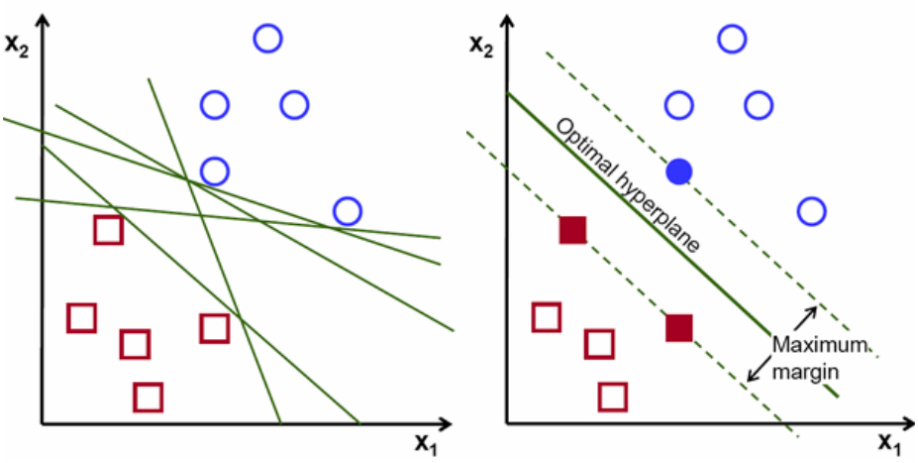

Figure II7:SVM

II.1.2.2. K plus proches voisins (KNN)

La méthode de classification K Nearest Neighbor (PPV Plus Proches Voisins) est spécialement conçue pour la classification et peut être appliquée aux tâches d'estimation. La méthode PPV consiste en un raisonnement par situation. L'idée de prendre une décision commence par l'identification d'un ou plusieurs cas similaires qui ont été résolus en mémoire. Un processus de formation n'est pas nécessaire pour développer un modèle à partir d'échantillons de formation. Afin de créer un modèle, l'échantillon d'apprentissage est associé à la fonction de distance et à la fonction de sélection de classe en tant que fonction de classe voisine la plus proche.

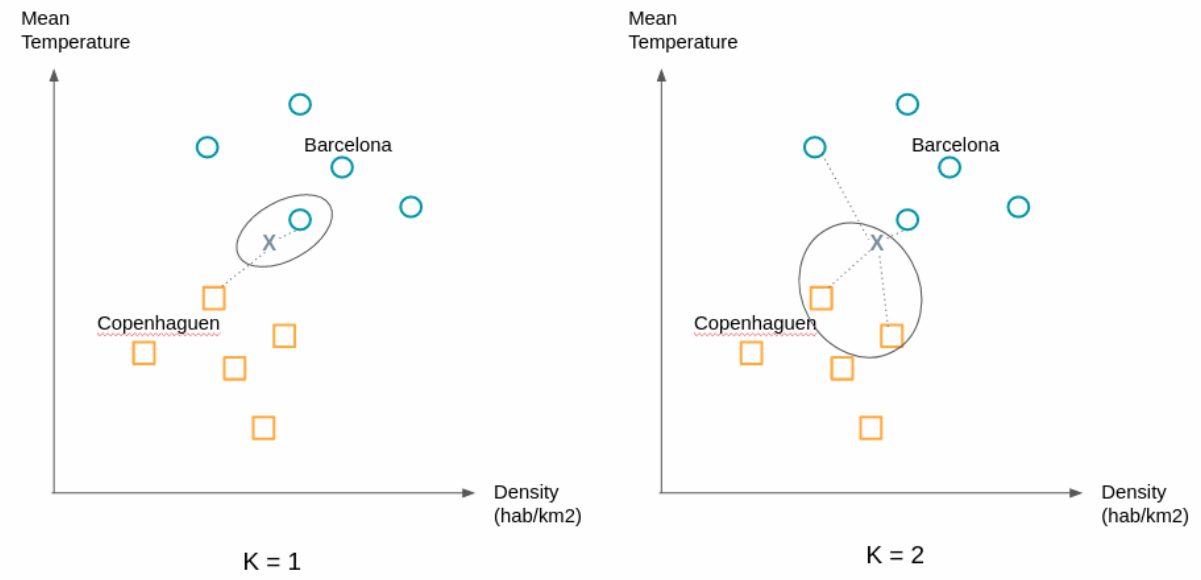

Figure II8: K nearest neighbours

Une observation est attribuée à la classe de ses K les plus proches voisins. « C'est tout ?! » vous me demanderez. Oui c'est tout, mais comme le montre l'exemple ci-dessous, le choix de K peut avoir un impact considérable sur de nombreuses choses. La séparation la plus satisfaisante sera donc recherchée en essayant différentes valeurs de K (DataKeen, 2024.).

II.1.2.3. La régression logistique

La régression logistique est la méthode statistique couramment utilisée dans les études empiriques impliquant des variables dépendantes catégorielles. Comme le démontre Allison(D.Allison, 1999), une variable dépendante dichotomique viole les hypothèses d'homoscedasticité et de normalité du terme d'erreur pour le modèle de régression linéaire. En conséquence, les estimations de l'erreur standard ne seront pas des estimations cohérentes des vraies erreurs standards, et les estimations des coefficients ne seront plus efficaces. De plus, estimer un modèle de probabilité linéaire avec la technique des moindres carrés ordinaires conduira à des valeurs prédites qui sont en dehors de la plage plausible de probabilité (0,1). Pour ces raisons, le modèle de régression logistique est utilisé lorsque la variable dépendante est dichotomique. Ce modèle transforme la probabilité en cote puis prend le logarithme des cotes. Ce faisant, la limite inférieure et supérieure de la probabilité est supprimée. Le modèle de régression logistique prend la forme suivante (D.Allison, 1999; Niu, 2020) :

$$\log=\alpha+\beta_1 x_{i1} +\beta_2 x_{i2} +\ldots+\beta_k x_{ik} \qquad \text{Équation II9}$$

où *i* désigne l'individu, p_i représente la probabilité que l'événement se produise, $1-p_i$ représente la probabilité que l'événement ne se produise pas, le rapport des deux représente les odds de l'événement, et l'expression à gauche de l'équation représente les log-odds, ou le logit. À droite de l'équation, α\alphaα représente l'ordonnée à l'origine, β\betaβ représente le coefficient de régression, et X représente la variable indépendante (pour plus de détails sur la régression logistique, voir aussi Kleinbaum et al. 1998; McCulloch et Searle 2001; Menard 2010; Pedhazur 1997; Rencher 2000; Tabachnick et Fidell 2001).

Comme on peut le voir sur le côté droit de l'équation II-9, la spécification du modèle de régression logistique est très similaire à celle du modèle de régression linéaire en termes de variables indépendantes. Comme la régression linéaire, la régression logistique peut gérer à la fois des variables indépendantes continues et catégorielles. Les principaux logiciels statistiques incluent des procédures faciles à utiliser pour les modèles de régression logistique. Bien qu'il soit relativement simple de spécifier et d'estimer des modèles de régression

logistique, l'interprétation des résultats est plus compliquée et moins intuitive par rapport à la régression linéaire. Cela est dû au fait que dans le modèle de régression logistique, la relation entre les probabilités et l'ensemble des variables indépendantes n'est pas linéaire ; au contraire, c'est la relation entre le logit et l'ensemble des variables indépendantes qui est supposée être linéaire. En conséquence, les estimations des coefficients représentent le changement dans le log des cotes correspondant à un certain changement dans la variable indépendante. Reconnaissant la difficulté d'interpréter le log des côtes, la pratique courante est d'exponentier les estimations des coefficients pour obtenir le rapport de côtes. Cette étape élimine la complication de l'interprétation du log, et le rapport de cotes devient la taille d'effet standard produite par les logiciels statistiques. Néanmoins, le concept et l'interprétation appropriée du rapport de cotes laissent encore de nombreux chercheurs perplexes. Une recherche dans la littérature actuelle en recherche éducative montre que différentes approches d'interprétation des résultats de régression logistique ont été adoptées. Par exemple, le rapport de cotes est interprété comme un risque relatif dans certaines études (M.Chiang, 2012; Niu, 2020; Sullivan & Cosden, 2015) mais pas dans d'autres (J. Jaeger & K. Eagan, 2011; Niu, 2020). Certaines études n'interprètent pas du tout le rapport de cotes mais rapportent et interprètent plutôt les effets marginaux et les probabilités prévues (King, 2015; Ko, 2015; Niu, 2020). Quelle approche est la plus couramment adoptée dans la recherche éducative ? Est-ce que l'approche la plus courante est la meilleure ? Quelles sont les implications de l'adoption de différentes approches ?

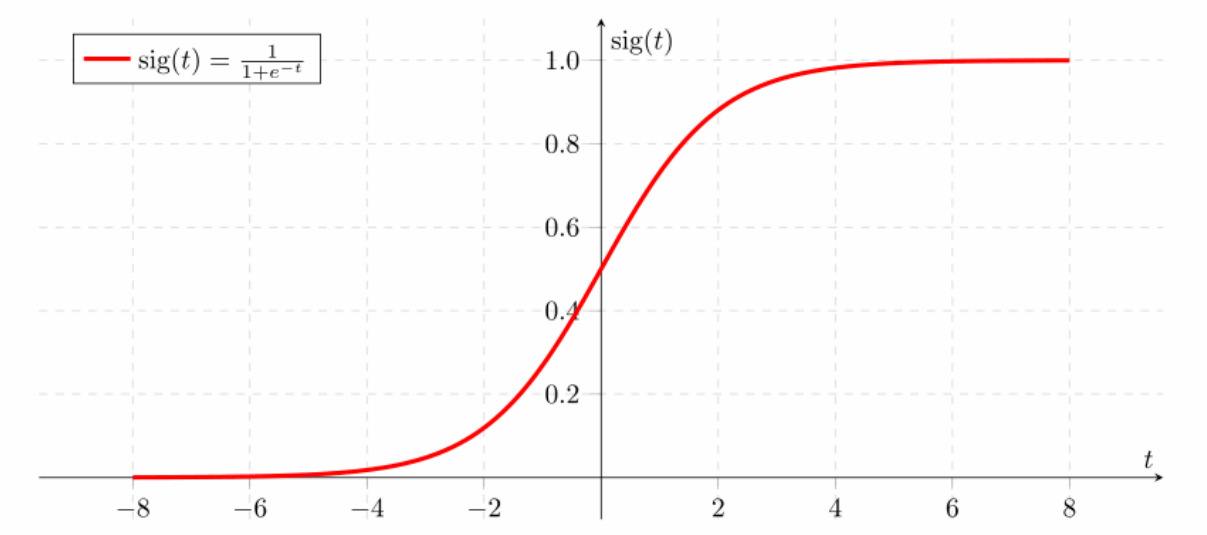

Figure II9:Graphe et expression de la fonction sigmoïde

II.1.2.4. Arbres de décisions

De la même manière que les SVM, les arbres de décision sont des algorithmes d'apprentissage automatique polyvalents qui peuvent effectuer des tâches de classification et de régression, voire des tâches à sorties multiples. Ceux-ci sont des algorithmes très puissants qui peuvent s'adapter à des ensembles de données spécifiques. Les algorithmes d'apprentissage automatique les plus performants actuellement disponibles sont les arbres de décision, qui jouent également un rôle crucial dans les forêts aléatoires (Géron, 2019). Ils développent généralement une structure de questions « si/autre », ce qui les conduit à prendre une décision. Ces interrogations ressemblent à celles que l'on peut poser dans un jeu de 20 interrogations. Supposons qu'on souhaite distinguer les quatre animaux suivants : ours, faucons, pingouins et dauphins. Il est primordial de minimiser le nombre de questions afin d'obtenir la réponse adéquate. On peut d'abord se demander si l'animal a des plumes, question qui restreint vos choix à deux animaux. Dans le cas où la réponse serait « oui », une autre question pourrait nous permettre de distinguer les faucons des pingouins. Par exemple, nous pouvons interroger sur la capacité de vol de l'animal. Si vous n'avez pas de plumes, vous pouvez choisir les dauphins et les ours, et nous devrons poser une question pour distinguer ces deux animaux : par exemple, demander si l'animal a des nageoires.

Cette série de questions peut être présentée sous la forme d'un arbre de décision, comme illustré dans la figure suivante :

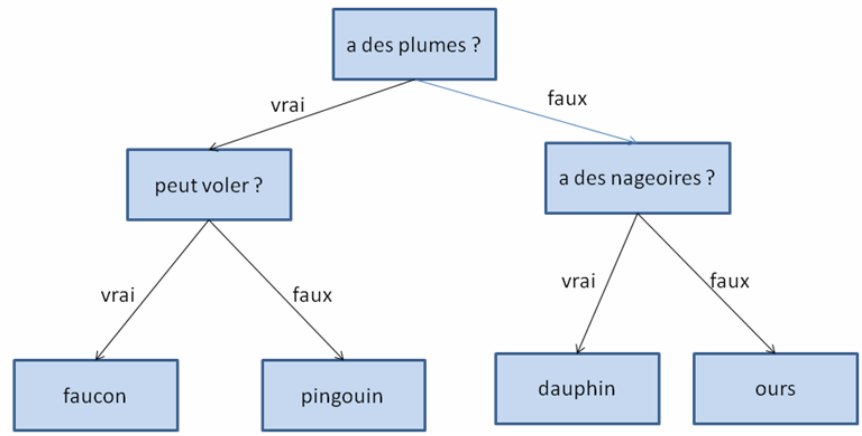

Figure II10:Un arbre de décision pour distinguer entre plusieurs animaux

Dans ce schéma, chaque point de l'arbre représente soit une question, soit un point final (également appelé feuille) qui contient la réponse. Les bords relient les réponses à une question à la question suivante que nous poserions.

En langage d'apprentissage automatique, nous avons créé un modèle permettant de distinguer quatre catégories d'animaux (faucons, pingouins, dauphins et ours) en se référant aux trois caractéristiques « des plumes », « peut voler » et « a des nageoires ». Au lieu de fabriquer ces

modèles de manière manuelle, il est envisageable de les obtenir à partir de données grâce à un apprentissage supervisé Les arbres de décision présentent deux bénéfices par rapport à de nombreux algorithmes : Des non-experts peuvent facilement visualiser et comprendre le modèle obtenu (du moins pour les petits arbres), et les algorithmes sont totalement invariants en ce qui concerne la dimension des données. Chez les algorithmes d'arbre de décision, chaque caractéristique est traitée de manière individuelle et les divisions possibles des données ne sont pas déterminées par l'échelle, il n'est pas nécessaire de réaliser un prétraitement comme la normalisation ou la standardisation des caractéristiques. Les arbres de décision se révèlent particulièrement efficaces lorsqu'il y a des caractéristiques à des échelles totalement différentes, ou lorsqu'il y a un mélange de caractéristiques binaires et continues.

II.1.2.5. Les forêts aléatoires

Comme nous venons de le souligner, l'un des inconvénients majeurs des arbres de décision est leur tendance à surajuster les données éducatives. Random Forest est la solution à ce problème. Fondamentalement, une forêt aléatoire est une série d'arbres sélectionnés où chaque arbre est légèrement différent des autres. Les forêts aléatoires reposent sur l'idée que chaque arbre peut faire des prédictions raisonnablement bonnes, mais peut surapprendre certaines données. Lorsque nous construisons plusieurs arbres, tous bons et s'ajoutant les uns aux autres de différentes manières, nous pouvons réduire le nombre d'ajouts en faisant la moyenne de leurs résultats. Grâce à des calculs mathématiques rigoureux, la réduction du surajustement peut être démontrée tout en conservant le pouvoir prédictif de l'arbre. Pour mettre en œuvre cette stratégie, plusieurs arbres de décision doivent être créés. Chaque arbre doit pouvoir prévoir l'objectif de façon acceptable et être aussi différent des autres arbres. Les forêts aléatoires sont appelées ainsi parce que les arbres sont construits au hasard pour assurer la diversité de chaque arbre. Le choix des points de données utilisés pour la construction d'un arbre et les caractéristiques de chaque test de division permettent de randomiser les arbres d'une forêt aléatoire. On utilise principalement la technique d'appentissage ensembliste appelée « bagging » ou « bootstrap aggregating » pour générer des forêts aléatoires. Dans cette méthode, il est nécessaire de générer différentes entités d'un même modèle (plusieurs arbres de décisions) et de les entrainer sur une partie aléatoire d'une collection de données. Une fois que chaque arbre a été entrainer, nous pouvons regrouper les résultats de chaque arbre pour effectuer la prédiction(Géron, 2019).

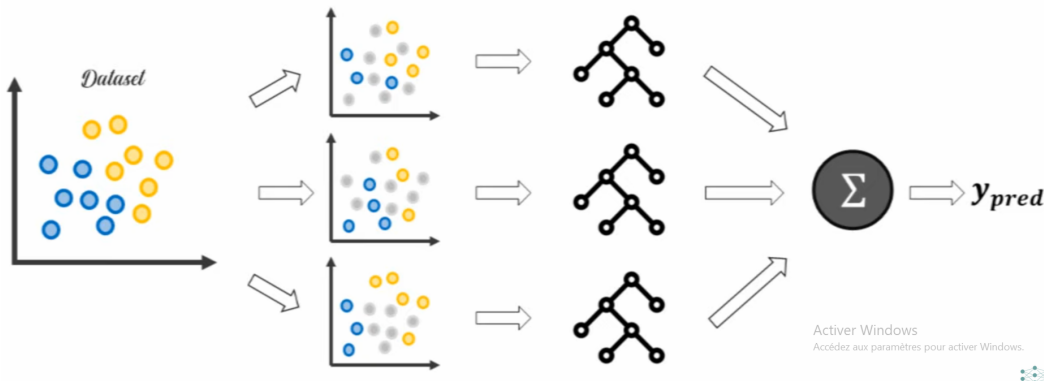

Figure II11: Les forêts aléatoires

II.1.2.6. Réseaux de neurones

Les réseaux de neurones artificiels (ANN) présentent une significative dans le domaine de l'intelligence artificielle et du machine learning. Ces réseaux sont élaborés en se basant sur le fonctionnement du cerveau humain afin de reproduire la façon dont les neurones biologiques gèrent l'information. Ils sont extrêmement performants dans la détection de motifs complexes.

Les réseaux de neurones artificiels sont constitués de diverses couches de neurones artificiels qui sont interconnectées. Les signaux d'entrée sont recueillis par chaque neurone, transformés par une fonction d'activation, puis transmis à la couche suivante. Grâce à cette procédure, répétée à travers différentes couches, le réseau peut prendre et représenter des relations non linéaires entre les variables d'entrée.

II.1.2.6.1. Des neurones biologiques aux neurones artificiels

Des réseaux de neurones ont été développés pour reproduire le système nerveux humain pour les opérations d'apprentissage automatique en traitant les unités de calcul du modèle d'apprentissage d'une manière similaire aux neurones humains. La vision centrale des réseaux de neurones consiste à concevoir une intelligence artificielle en développant des machines dont l'architecture permet de reproduire les calculs effectués par les neurones humains. La consommation excessive de données et le calcul intensif des réseaux de neurones ont été considérés comme des obstacles à leur utilisation. Enfin, la croissance significative de la puissance de calcul et la disponibilité accrue des données depuis les années 1990 ont conduit à une augmentation des taux de réussite des réseaux neuronaux, ce qui a conduit à ce que ce domaine soit surnommé « apprentissage profond ». En théorie, un réseau de neurones peut apprendre n'importe quelle fonction mathématique en utilisant suffisamment de données d'entraînement. Ces techniques d'apprentissage automatique sont devenues très populaires,

simulant le processus d'apprentissage dans les organismes biologiques. Les neurones sont des cellules présentes dans le cerveau humain. Les axones et les dendrites des neurones sont relies les uns aux autres, et les synapses sont les zones de connexion entre les axones et les dendrites pour les relier(C. Aggarwal, 2018). On peut observer ces connexions dans la figure 12.

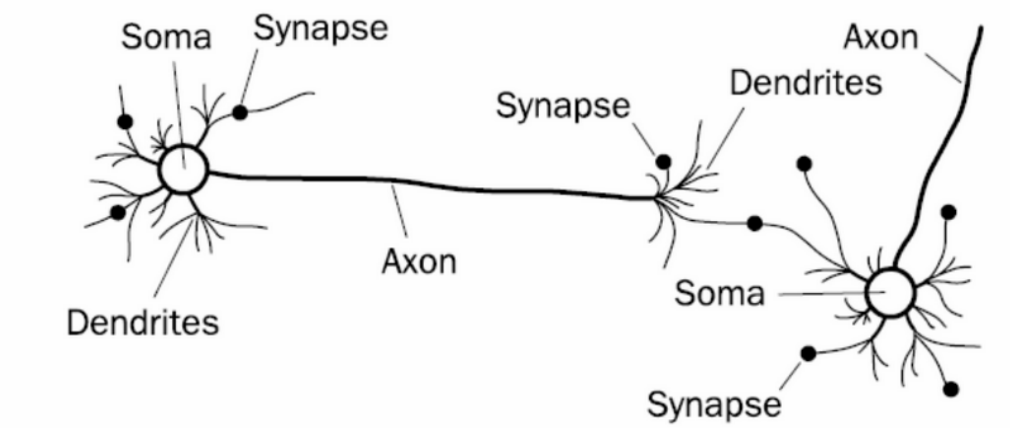

Figure II12: Réseau de neurones biologique

En réponse à des stimuli externes, la force des connexions synaptiques change souvent. Ce changement correspond à la manière dont la biologie apprend. Les réseaux de neurones artificiels utilisent des unités informatiques appelées neurones pour simuler ce mécanisme biologique. Les poids relient les unités de calcul les unes aux autres, de la même manière que la force des connexions synaptiques dans les organismes biologiques. Un poids est attribué à chaque entrée du neurone, ce qui affecte la fonction calculée par l'unité(C. Aggarwal, 2018).

Cette structure est présentée dans la figure suivante.

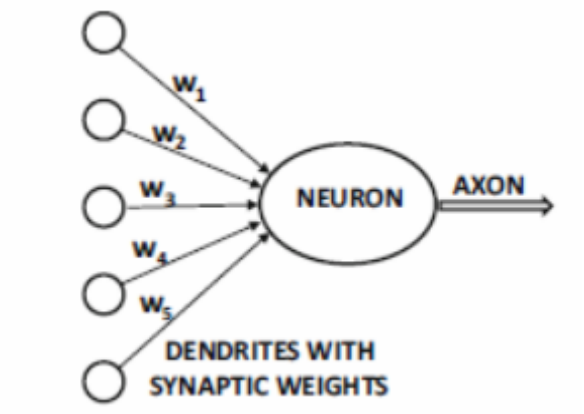

Figure II13: Réseau de neurones artificiel

II.1.2.6.2. Couches d'un réseau de neurones

Un réseau de neurones comprend un nombre infini de neurones, organisés en couches interconnectées. La couche d'entrée représente toutes les données et les conditions initiales, tandis que la couche de sortie peut inclure plusieurs neurones. Cela se révèle particulièrement bénéfique dans le domaine de la classification, où chaque neurone de sortie incarne une catégorie. Les couches intermédiaires (entre l'entrée et la sortie) sont appelées couches cachées, car les calculs effectués ne sont pas visibles pour l'utilisateur. L'architecture spécifique des réseaux neuronaux multicouches est appelée réseaux à propagation avant, car les couches successives se nourrissent mutuellement dans le sens de l'entrée vers la sortie. Dans l'architecture des réseaux à propagation avant, il est par défaut que tous les nœuds d'une couche soient connectés à ceux de la couche de référence.

La figure suivante montre les différentes couches d'un réseau de neurones :

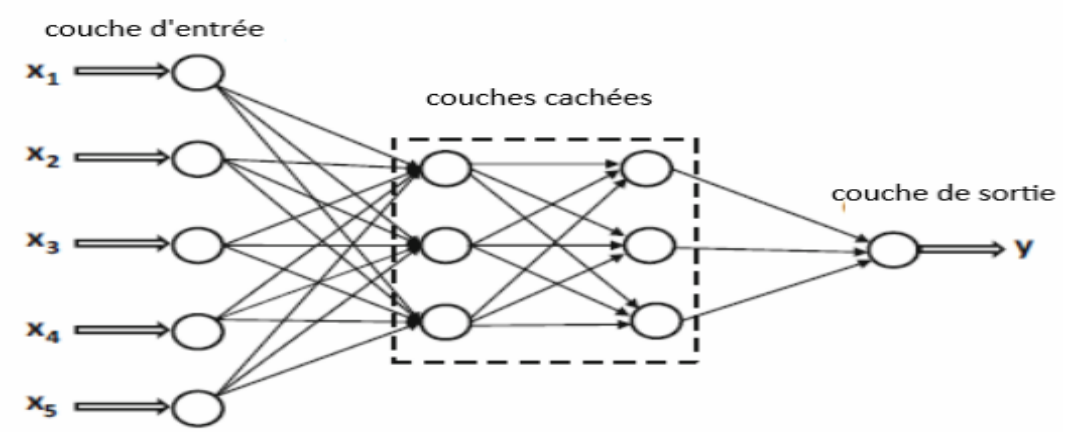

Figure II14: Réseaux de neurones multicouches.

II.1.2.6.3. Fonctions d'activations

Dans le but d'imiter la fonction d'activation et de désactivation, la fonction d'activation implique de prendre l'entrée combinée z présentée dans la figure II-15, de lui appliquer une fonction et de lui transmettre la valeur de sortie. La fonction d'activation détermine donc l'état d'un neurone. En l'absence d'activation des neurones, leur sortie correspondrait à la somme des poids des entrées. En outre, tout le réseau de neurones, c'est-à-dire une combinaison de neurones, serait une combinaison de fonctions linéaires, ce qui est aussi une fonction linéaire. Cela signifie que, malgré l'ajout de couches cachées, le réseau demeure toujours équivalent à un simple modèle de régression linéaire, avec toutes ses limites. Pour transformer le réseau en une fonction non linéaire, nous utiliserons des fonctions d'activation non linéaires pour les

neurones. La fonction d'activation des neurones d'une même couche est habituellement la même, mais il est envisageable que différentes couches aient des fonctions d'activation distinctes(I. Vasilev, D. Slater, G. Spacagna, P. Roelants, 2019).

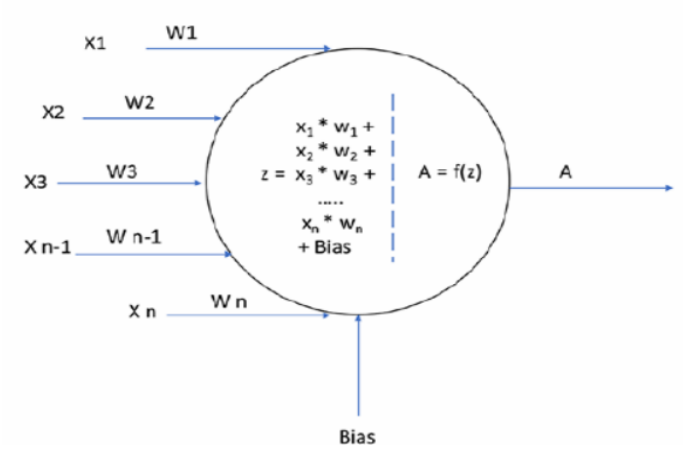

Figure II15: Fonction d'activation sur un neurone unique

II.1.2.6.4. Fonction d'activation linéaire

La fonction d'activation la plus utilisee est l'activation identitaire ou linéaire,

$$f(x) = x$$

On utilise souvent la fonction d'activation linéaire dans le nœud de sortie, lorsque la valeur cible est une valeur réelle. On l'utilise même pour les sorties discrètes lorsqu'il est nécessaire de mettre en place une fonction de perte de substitution lissée. L'illustration suivante, connue sous le nom de « figure 16 », présente le schéma d'une fonction d'activation linéaire.

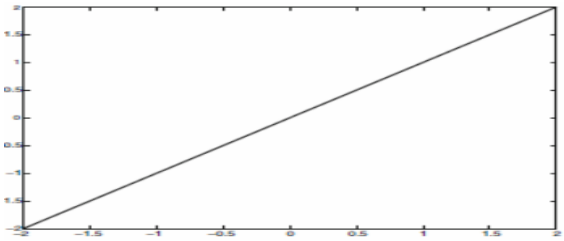

Figure II16: Fonction d'activation linéaire

II.1.2.6.5. Fonction d'activation signe

La fonction d'activation représentée par la figure 17 peut être employée afin de créer une correspondance avec des sorties binaires lors de la prédiction. Cependant, sa non différentiable empêche son utilisation pour générer la fonction de perte lors de l'entraînement. Par exemple, tandis que le perceptron utilise la fonction de signe pour prédire, le critère du perceptron à l'entraînement ne requiert qu'une activation linéaire.

f (x) = sign(x) Équation II10

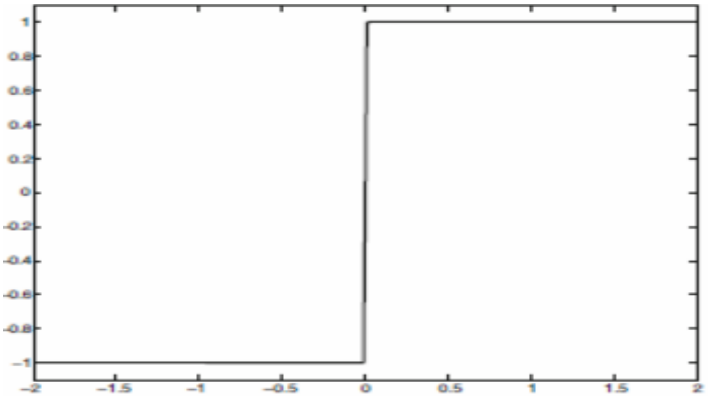

Figure II17: Fonction d'activation signe

II.1.2.6.6. Fonction sigmoïde

Une fonction sigmoïde « figure 18 » se présente comme suit :

f (x) = Équation II11

La sortie est comprise entre 0 et 1, comme le montre l'image ci-dessous. La sortie non linéaire (en forme de S comme mentionné ci-dessus) améliore considérablement le processus d'apprentissage car elle ressemble étroitement aux principes suivants : faible impact : faible rendement et impact élevé : rendement élevé - et elle limite également la plage de sortie de 0 à 1.

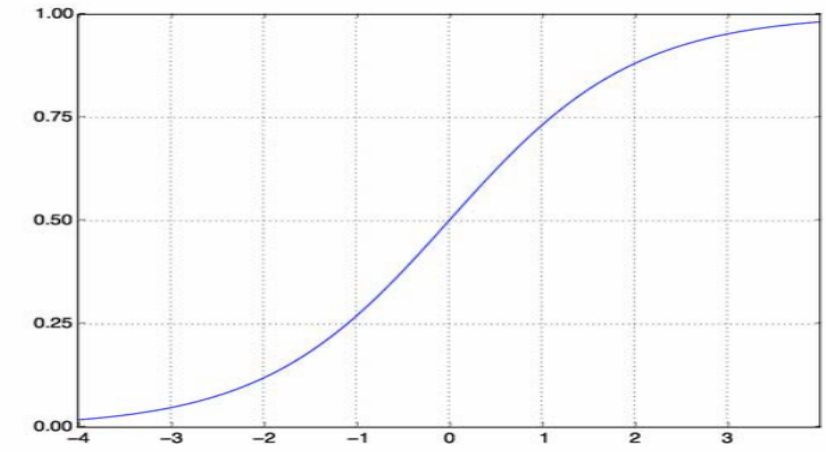

Figure II18:Fonction sigmoïde

II.1.2.6.7. Tangente hyperbolique

La fonction tangente hyperbolique « figure 19 » tanh a une forme similaire à celle de la fonction sigmoïde, sauf qu'elle est rééchelonnée horizontalement et traduite/rééchelonnée verticalement à [-1,1], Les fonctions tanh et sigmoïde sont liées comme suit :

$$f(x) = 2 * \text{sigmoid}(2x)-1 \qquad \text{Équation II12}$$

Lorsque les calculs doivent être à la fois positifs et négatifs, la fonction tanh est préférée à la fonction sigmoïde. En outre, sa position moyenne et son gradient plus élevé (du fait de l'étirement) par rapport au sigmoïde rendent l'entraînement plus facile. L'utilisation des fonctions sigmoïde et tanh a toujours été privilégiée pour intégrer la non-linéarité dans le réseau de neurones.

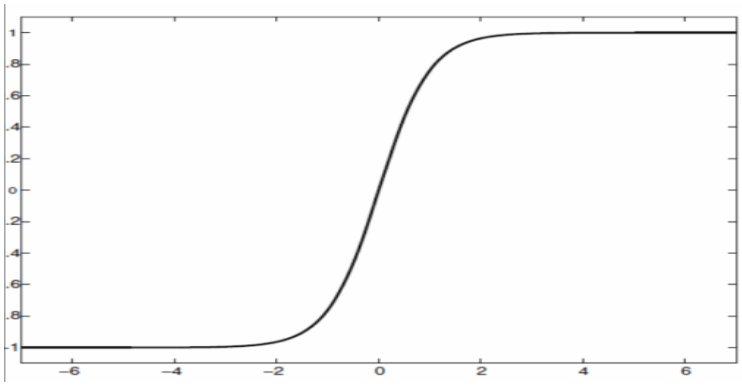

Figure II19: Tangente hyperbolique

II.1.2.6.8. ReLU

La fonction f(z) = max(0,z) est utilisée par la ReLU « rectified linear unit » « figure 20 », ce qui implique que si la sortie est positive, elle donnera la même valeur, sinon elle donnera 0. Le diagramme suivant présente la plage de sortie de la fonction :

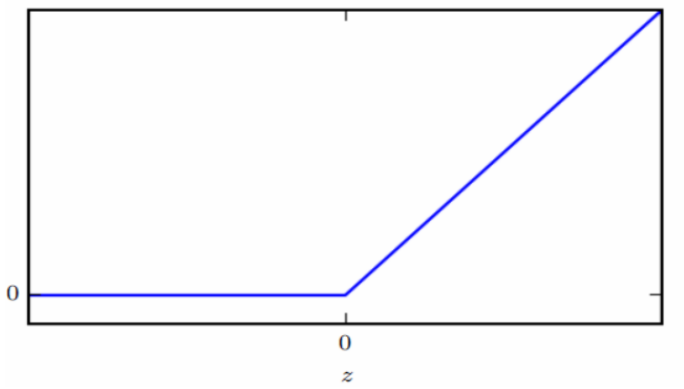

Figure II20: ReLU

II.1.2.6.9. Fonction de perte

Elle évalue la disparité entre les réelles valeurs et celles prédictives d'un réseau neuronal artificiel. Les fonctions de perte les plus importantes sont l'erreur quadratique moyenne (MSE), l'erreur absolue moyenne et l'erreur quadratique moyenne (MSE). logarithmique, entropie binomiale. Dans notre cas, nous utilisons le MSE (la variabilité expliquée par le modèle par rapport aux données après avoir considéré une relation de régression).

II.1.2.6.10. Optimiser un réseau neuronal artificiel

Il existe plusieurs méthodes pour optimiser un modèle. L'optimiseur le plus courant est l'estimation adaptative du mouvement (Adam), et il fonctionne mieux pour minimiser la fonction de coût lors de l'entraînement(Nokeri, 2021).

II.1.3. Avantages et inconvénients

De nombreux domaines ont été profondément transformés par les algorithmes d'apprentissage automatique, allant de l'analyse de données à l'intelligence artificielle. Ces algorithmes, capables d'apprendre à partir de données et de faire des prédictions ou des classifications, présentent de nombreux bénéfices significatifs. Cependant, ils comportent des inconvénients et des obstacles. Il est primordial de comprendre les atouts et les faiblesses des algorithmes d'apprentissage automatique pour les utiliser de manière efficace et optimiser leur efficacité dans diverses situations. Le tableau suivant (Jain & Kulkarni, 2020) nous donne un aperçu les avantages et désavantages de ces algorithmes.

Algorithme	Advantages	inconvenients
Support Vector Machines (SVM)	- Efficace pour les espaces de grande dimension. -Performant pour la classification et la régression.	- Complexe à implémenter et à interpréter. -Sensible au choix du noyau et des hyperparamètres.
K-Nearest Neighbors (K-NN)	- Facile à comprendre et à implémenter. -Aucune phase d'entraînement nécessaire.	- Lenteur en prédiction pour de grands ensembles de données. -Sensible aux caractéristiques non pertinentes.
Arbres de Décision	- Facile à visualiser et à interpréter. - Gère bien les données catégorielles et continues.	- Sujet au surapprentissage (overfitting). -Peut devenir très complexe.
Forêts Aléatoires (Random Forests)	- Réduit le risque de surapprentissage. - Performant pour des données	- Moins interprétable que les arbres de décision individuels. - Gourmand en ressources

	de grande taille et complexes.	computationnelles.
Gradient Boosting Machines (GBM)	- Haute précision et performance. - Flexibilité pour les différents types de données.	- Long temps d'entraînement. - Nécessite un réglage précis des hyperparamètres.

Tableau II1:Avantages et inconvénients les algorithmes de machine learning

II.1.4. Conclusion

En conclusion, ce chapitre a permis une exploration approfondie des différents algorithmes d'apprentissage automatique utilisés dans cette étude. À travers une analyse détaillée, nous avons examiné les principes fondamentaux de ces algorithmes, leur fonctionnement, ainsi que leurs avantages et limitations. La création d'un modèle d'apprentissage automatique performant et fiable nécessite la réalisation de diverses étapes. Grâce à la compréhension de ces concepts essentiels, nous sommes davantage préparés à aborder la conception et l'utilisation des modèles dans notre projet. En outre, cette étude a mis en évidence l'importance de choisir les algorithmes les plus adaptés en fonction des caractéristiques particulières de nos données et des objectifs de notre recherche. En incorporant ces connaissances dans nos méthodes fondamentales, allant de la collecte de données à l'évaluation des performances des modèles, qui font tous partie de notre architecture globale, nous sommes confiants dans notre capacité à élaborer des modèles d'apprentissage automatique efficaces et précis pour répondre à nos questions de recherche.

Chapitre 3

Conception d'un système intelligent pour la prédiction de la volatilité

I.1. Introduction

Ce chapitre examine en détail la conception et la mise en place des modèles d'apprentissage automatique utilisés pour anticiper la volatilité des marchés financiers. Plusieurs étapes cruciales sont nécessaires pour mettre en œuvre un modèle d'apprentissage automatique efficace et fiable, allant de la collecte des données à l'évaluation des performances des modèles, qui font tous partie de notre architecture globale.

I.2. Architecture générale du travail

Nous présenterons dans cette partie l'organisation générale du travail effectué dans le cadre de notre étude. Cela comprend une étude approfondie du processus de travail, depuis la collecte des données jusqu'à l'évaluation des projets. Chaque étape revêt une importance capitale pour concevoir un modèle efficace et fiable. Voici une description approfondie des éléments clés de cette structure:

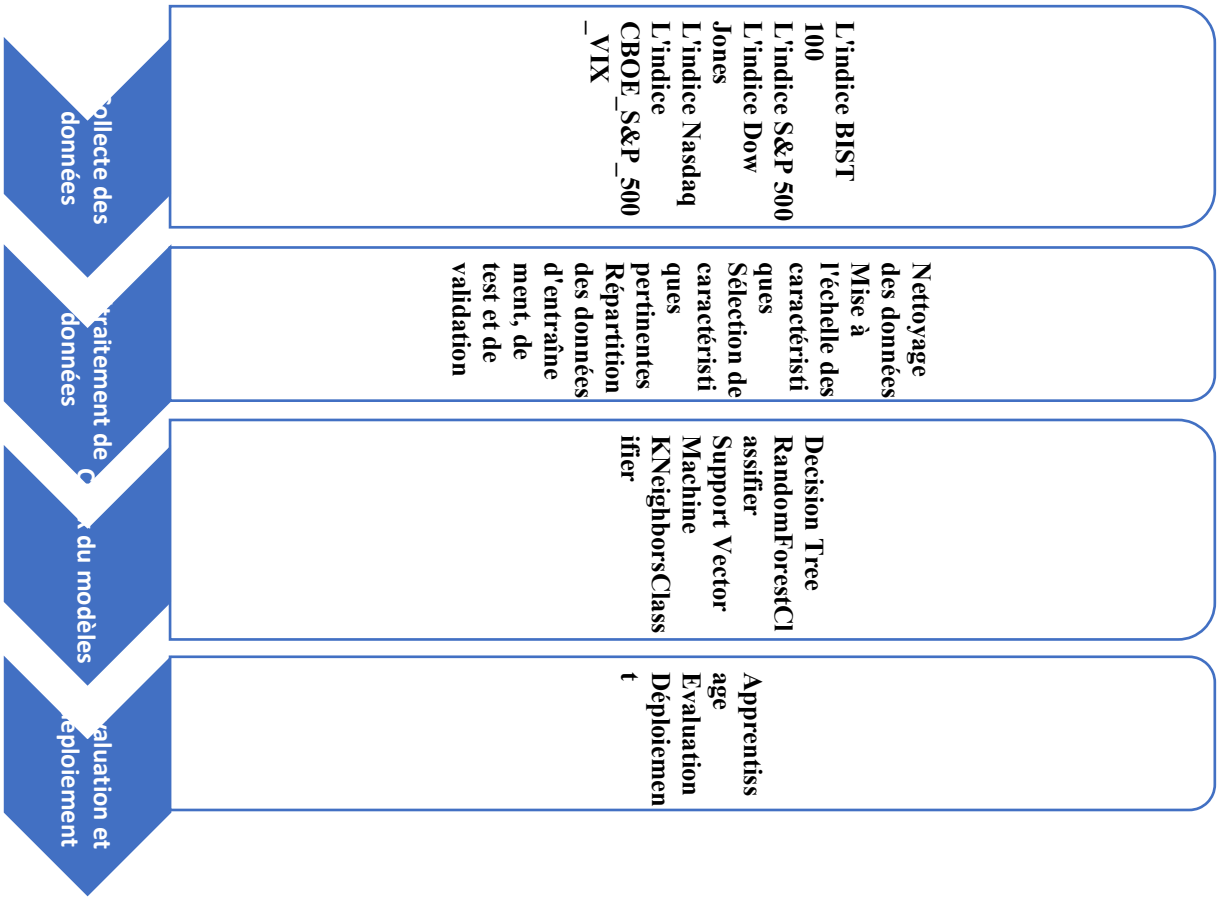

Figure III1: l'architecture générale de notre travail

I.2.1. Collecte des données

Dans cette partie, nous examinerons en détail les sources et les techniques employées afin de recueillir les informations requises pour notre étude. Cela peut englober des détails sur les ensembles de données employés, leur origine, leur taille, leur format, ainsi que les techniques de collecte utilisées afin d'assurer la qualité et la fiabilité des informations.

Les informations employées pour cette étude proviennent de la série des cours de l'indice de Bist 100 de la Turquie. Notre objectif est de réaliser cette tâche en utilisant un échantillon qui couvre la période allant du 24 janvier 2001 au 3 mars 2024.Ces informations proviennent de cotations quotidiennes ajustées à la fermeture des marchés financiers et sont issues de la base de données d'Investing.com.Une fois que les données ont été collectées, nous utilisons la bibliothèque pandas préinstallée dans Python pour pouvoir lire nos données. L'analyse des séries temporelles est adaptée à l'estimation d'une variable continue qui dépend du temps. Dans ce chapitre, nous l'utilisons pour identifier la structure des données séquentielles. C'est une méthode transparente pour identifier des motifs et prévoir les prix futurs des paires de

devises. Les données de marché sont souvent séquentielles et comportent certains éléments stochastiques, ce qui signifie qu'il existe un processus aléatoire sous-jacent. Nous sommes intéressés à découvrir les motifs dans le prix de clôture ajusté de la paire de devises au fil du temps, puis à faire des prédictions fiables sur les mouvements de prix.

Notre dataset contient 8434 lignes et 32 colonnes, les colonnes représentent le prix d'un indice boursier d'une entreprise de chaque pays dont entre autre :

- L'indice BIST 100 (Borsa Istanbul 100 Index) est un indice boursier turc qui représente les 100 entreprises cotées en bourse les plus liquides et les plus cotées sur la Borsa Istanbul, la principale bourse du pays. Le marché boursier turc est largement utilisé comme indicateur de la performance globale de cet indice. Il englobe de nombreux domaines économiques, tels que la finance, les télécommunications, l'industrie, les biens de consommation, et bien d'autres encore. Les investisseurs, les analystes financiers et les médias surveillent fréquemment la BIST 100 afin d'évaluer la santé et la tendance du marché boursier turc.
- Le S&P 500 est considéré comme l'un des indices boursiers les plus suivis et employés à travers le monde. Il regroupe les 500 entreprises les plus importantes cotées en bourse aux États-Unis, choisies en fonction de leurs montants boursiers, de leur liquidité, de leur domaine d'activité et d'autres critères. Cet indice représente le marché boursier américain dans son ensemble, car il englobe de nombreux secteurs économiques tels que la technologie, la santé, l'énergie, les services financiers, les biens de consommation, etc.
    - Le Dow Jones Industrial Average (également connu sous le nom de Dow Jones) est l'un des indices boursiers les plus anciens et les plus suivis à l'échelle mondiale. L'indice Dow Jones a été créé en 1896 par Charles Dow et regroupe 30 des plus grandes entreprises cotées en bourse aux États-Unis. Ces sociétés sont choisies afin de représenter différents domaines économiques tels que la technologie, la finance, les biens de consommation, l'industrie, etc.
    - Nasdaq, ou National Association of Securities Dealers Automated Quotations, est une bourse électronique américaine spécialisée dans la cotation des actions de sociétés technologiques, de haute technologie et de biotechnologie. Le Nasdaq, créé en 1971, est devenu l'une des principales bourses américaines, rivalisant avec le New York Stock Exchange (NYSE).
    - Le S&P 500 VIX, également connu sous le nom de VIX, est un indicateur de volatilité qui évalue la volatilité implicite du marché boursier américain, notamment du S&P

500. Le VIX est un indice de peur ou de crainte sur le marché, créé par le Chicago Board Options Exchange (CBOE) en 1993. Les prix des options d'achat et de vente sur l'indice S&P 500 sont utilisés pour calculer le VIX. Il témoigne des attentes des investisseurs concernant la volatilité du marché dans les 30 jours à venir. Une valeur élevée du VIX témoigne d'une prévision de volatilité plus élevée, tandis qu'une valeur basse témoigne d'une prévision de volatilité faible. Le VIX est fréquemment employé en tant qu'indicateur de l'opinion du marché. En général, une hausse du VIX est liée à un moment de stress ou d'incertitude sur les marchés financiers, tandis qu'une baisse du VIX peut témoigner d'un sentiment de confiance et de stabilité.

- L'indice boursier Small Cap 2000 évalue la performance des petites capitalisations aux États-Unis. Il est administré par Russell Investments, fournisseur d'indices financiers, et fait partie de la série des indices Russell.

À l'inverse des indices tels que le S&P 500 qui se focalisent sur les grandes entreprises, le Small Cap 2000 surveille les résultats de 2000 entreprises américaines dont la capitalisation boursière est considérée comme étant parmi les plus petites de l'univers financer. Souvent, ces entreprises sont plus jeunes, ce qui leur confère un potentiel de croissance plus élevé, mais également un niveau de risque accru.

Tous ces indices sont collectés sur le site de investing.com puis on les a fusionnés pour former une seule base de données.

I.2.2. Prétraitement de données

Le nettoyage et la préparation des données brutes sont indispensables avant de pouvoir les utiliser. Dans le processus de prétraitement des données, on peut gérer les valeurs manquantes, normaliser, convertir les variables catégorielles en variables numériques et éliminer les doublons. L'objectif de cette méthode est d'améliorer la qualité des données afin d'obtenir des résultats précis et fiables.

I.2.2.1. Nettoyage des données

Le nettoyage des données consiste à de détecter et corriger ou supprimant l les valeurs manquantes ou de les remplacer 0 ou les valeurs plus proches de cette valeurs manquante pour améliorer la qualité de notre analyse.

I.2.2.2. Mise à l'échelle des caractéristiques

Dans leurs calculs, la majorité des algorithmes d'apprentissage automatique se servent de la distance euclidienne entre deux points de données. C'est pourquoi les éléments de grande magnitude seront plus importants dans les calculs de distance que les éléments de faible magnitude. Afin d'éviter cette situation, on utilise la normalisation des fonctionnalités ou la normalisation du score Z. Il s'agit de la classe "StandardScaler" de "sklearn.preprocessing".

I.2.2.3. Sélection de caractéristiques pertinentes

Il est procédé au nettoyage de ce jeu de données historiques sur divers indices boursiers collectés sur investing.com, en incluant la sélection de caractéristiques, la création de caractéristiques et la division en série temporelle pour l'entraînement et le test. Malgré la diversité des caractéristiques présentes dans le jeu de données, nous allons utiliser ces caractéristiques, à savoir le BIST 100, le S&P 500, Nasdaq, Dow Jones, le CBOE_S&P_500_VIX et le Small_Cap_2000, afin de créer des caractéristiques basées sur leurs performances.La conversion de la colonne de rendement prix de BIST 100 en catégorielle sera effectuée entre 0 et 1.pour pouvoir faire une prédiction du hausse et basse du prix de BIST 100. 0 est mis pour une baisse et 1 pour une hausse. Par la suite on ajouter 20 derniers caractérisés du rendement prix de BIST 100 comme caractérisés supplémentaires qui montrent la hausse et baisse du prix de BIST 100 des vingt (20) jours précédents dans le but de prédire une tendance de prix à plus long terme. En raison des données de séries temporelles.

Le calcul du rendement se fait de la manière suivante :

R= Équation III1

- R est le rendement a l'instant t
- p_t est le prix à l'instant t
- p_{t-1} est le prix à l'instant t-1

Nous allons essayer d'ajouter aussi d'autres nouvelles colonnes à partir de la date de notre dataset. Ces colonnes sont :

- **df1['Day'] = df1['Date'].dt.day**: Cette ligne crée une nouvelle colonne appelée "Day" dans le DataFrame **df1**. Cette colonne contiendra le jour correspondant à la date dans la colonne "Date" du DataFrame **df1**. La méthode **dt.day** extrait le jour de chaque date.

- **df1['Month'] = df1['Date'].dt.month**: Cette ligne crée une nouvelle colonne appelée "Month" dans le DataFrame df1. Cette colonne contiendra le mois correspondant à la date dans la colonne "Date" du DataFrame df1. La méthode dt.month extrait le mois de chaque date.
- df1['Year'] = df1['Date'].dt.year: Cette ligne crée une nouvelle colonne appelée "Year" dans le DataFrame df1. Cette colonne contiendra l'année correspondant à la date dans la colonne "Date" du DataFrame df1. La méthode dt.year extrait l'année de chaque date.

Après l'exécution de ces lignes de code, le DataFrame df1 contiendra les colonnes supplémentaires "Day", "Month" et "Year", chacune contenant respectivement le jour, le mois et l'année extraits de la colonne "Date".

	Date	BIST 100	S&P 500	Dow Jones	Nasdaq	CBOE_S&P_500_VIX	Small_Cap_2000	Log_BIST_S&P500	Log_BIST_All_Dow	Log_BIST_Nasdaq	...	
0	24/01/2001 00:00	108,83	1360,4	10649,81	2840,39	21,57	502,06	0,013031499	0,006765776	0,029906705	...	-0,0
1	25/01/2001 00:00	108,85	1364,3	10646,97	2859,15	22,03	502,25	0,002866804	-0,000266671	0,006604727	...	0,0
2	26/01/2001 00:00	109,8	1357,5	10729,52	2754,28	22,64	499	-0,004984241	0,00775338	-0,036678733	...	0,0
3	27/01/2001 00:00	107,43	1355	10659,98	2781,3	22,57	498,68	-0,001841621	-0,006481185	0,009810186	...	-0,0
4	28/01/2001 00:00	107,43	1355	10659,98	2781,3	22,57	498,68	0	0	0	...	0,0
...	...	...	...	...	...	...	...	...	...	...	...	
8430	27/02/2024 00:00	9334,13	5069,53	39069,3	15976,25	13,74	2028,97	-0,003786747	-0,001590278	-0,001285881	...	
8431	28/02/2024 00:00	9179,48	5078,2	38972,74	16035,3	13,43	2056,11	0,001710218	-0,002471506	0,003696111	...	-0,0
8432	29/02/2024 00:00	9062,36	5069,76	38949,29	15947,74	13,84	2040,31	-0,001662006	-0,000601703	-0,005460453	...	0,
8433	01/03/2024 00:00	9193,69	5096,27	38995,93	16091,92	13,4	2054,84	0,005229044	0,001197454	0,009040779	...	-0,
8434	02/03/2024 00:00	9097,15	5137,09	39087,38	16274,94	13,11	2076,39	0,00800978	0,002345117	0,01137341	...	0,0

Figure III2: la base de données après le prétraitement

I.2.3. Division des données d'entraînement, de test et de validation

La méthode utilisée pour distinguer les données : Dans le code, on utilise la méthode train_test_split de la bibliothèque scikit-learn afin de diviser les données en ensembles d'entraînement, de test et de mesure. Attention à la taille des ensembles d'entraînement et de test, ainsi qu'aux proportions utilisées : En apprentissage automatique, les données sont partagées par le code selon un ratio de 80% pour l'ensemble d'entraînement, 10% pour l'ensemble de test et 10% pour les données de la validation. En règle générale, cette répartition est choisie pour optimiser l'utilisation des données disponibles pour l'entraînement tout en conservant un ensemble de tests et de validation suffisamment étendu pour évaluer les valeurs du modèle. Le code suivant résume de ce qui précède :

Diviser les données en deux ensembles pour pouvoir équilibrer les données selon notre variable cible.

```
data_class_0 = fd[fd['H or B'] == 0]
data_class_1 = fd[fd['H or B'] == 1]
```

Pour l'ensemble d'entraînement avec la classe 0 :

Nombre de lignes dans l'ensemble d'entraînement avec la classe 0: 5309

Pour l'ensemble d'entraînement avec la classe 1 :

Nombre de lignes dans l'ensemble d'entraînement avec la classe 1: 3125

Divisons chaque ensemble en ensembles d'entraînement, de test et de validation

```
Train_class_0, temp_class_0 = train_test_split(data_class_0, test_size=0.2,
random_state=42)
test_class_0, val_class_0 = train_test_split(temp_class_0, test_size=0.5,
random_state=42)
train_class_1, temp_class_1 = train_test_split(data_class_1, test_size=0.2,
random_state=42)
test_class_1, val_class_1 = train_test_split(temp_class_1, test_size=0.5,
random_state=42)
```

Concaténer les ensembles d'entraînement, de test et de validation pour obtenir des ensembles équilibrés.

```
train_set = pd.concat([train_class_0, train_class_1])
test_set = pd.concat([test_class_0, test_class_1])
val_set = pd.concat([val_class_0, val_class_1])
```

Mélanger les ensembles

```
train_set = train_set.sample(frac=1, random_state=42).reset_index(drop=True)
test_set = test_set.sample(frac=1, random_state=42).reset_index(drop=True)
val_set = val_set.sample(frac=1, random_state=42).reset_index(drop=True)
```

I.2.4. Choix des modèles

Une fois que les données sont prêtes à être incorporées dans un algorithme, la phase de Machine Learning débute. Aujourd'hui, en raison de la diversité des bibliothèques de Machine Learning, cette étape de mise en œuvre n'est pas la plus difficile du projet. Ainsi, il est facile

(et conseillé) de mettre en place plusieurs modèles pour résoudre le problème initial. Nous allons sélectionner un certain nombre de modèles d'apprentissage automatique afin d'en trouver le mieux à notre problématique.

Le Tableau 1 nous donne un récapitulatif des modèles sélectionnés.

Algorithmes	Acronyme
Decision Tree	DT
RandomForestClassifier	RF
SupporT Vector Machine	SVM
KNeighborsClassifier	KNeighborsClassifier
Artificial Neural Network	ANN

Tableau III1: Les modèles sélectionnés

Le choix des hyperparamètres

Les hyperparamètres sont des paramètres réglables qui permettrons d'améliorer nos mesures de performances de nos modèles afin d'évaluer nos modèles d'apprentissage automatique par la technique de validation croisée. Pour chaque modèle nous allons chercher les meilleurs paramètres en explorant une grille d'hyperparamètres prédéfinie. Les paramètres utilisés pour implémenter le modèle d'ensemble proposé sont mentionnés dans le tableau suivant :

Modèles	Paramètres	description
RF	Min_samples_leaf	2
	N_estimators	200
	Min_samples_split	2
SVM	C	10
	Kernel	sigmoid
	Gamma	auto
	Degree	2
KNN	N_neighbors	9
	Weights	uniform
	Metric	manhattan
DT	min_samples_split	2
	max_depth	None
ANN	Dense Layers	64,64,32
	Dropout	0,2
	Activation	Relu,Sigmoid

	Batch size	32
	Optimizer	Adam
	Loss	binary_crossentropy
	Metric	accuracy
	Epochs	100

Tableau III2: Les hyperparametres choisis pour les modèles utilisés

I.2.5. Entrainement des modèles

Cette étape est la plus importantes de l'apprentissage automatique. Le modèle est entraîné en y intégrant les données collectées et analysées lors des étapes précédentes. Elles auront été divisées en variables d'entrée (« X ») et variables de sortie (« target y ») dans un modèle d'apprentissage supervisé. La phase d'apprentissage a donc pour objectif de renforcer et de manière itérative la capacité du modèle à réagir à une situation particulière, à résoudre un problème complexe ou à réaliser une tâche. En réduisant au minimum les risques d'erreur/de coût. Il sera primordial de considérer la qualité des données et leur représentativité de la situation à analyser afin d'éviter toute tendance biaisée dans les résultats tout en maintenant une précision optimale.

Pour éviter tout biais statistique, on utilise généralement la division des jeux de données en 2 ou 3 parties.

Dans notre cas 3 parties. La première étape implique l'entraînement du modèle ("train set"), la deuxième étape consiste à le valider ("validation set") et la troisième étape peut être le test du modèle ("test set"). Cette séparation doit être réalisée de manière aléatoire tout en maintenant dans chaque partie une représentativité des données équivalente (par exemple, par validation croisée ou « Cross Validation »). Le code ci-après nous donne un aperçu de tout ce qui a été dit.

```
X_train = train_set.drop(columns=['H or B'])  # Caractéristiques de l'ensemble d'entraînement
y_train = train_set['H or B']  # Variables cibles de l'ensemble d'entraînement

X_test = test_set.drop(columns=['H or B'])  # Caractéristiques de l'ensemble de test
y_test = test_set['H or B']  # Variables cibles de l'ensemble de test
X_val=val_set.drop(columns=['H or B'])
y_val=val_set['H or B']
```

I.2.6. Evaluation du model

Le processus d'évaluation du modèle joue un rôle essentiel dans la création d'un modèle d'apprentissage automatique. Après avoir entraîné le modèle sur les données d'entraînement, il est primordial de vérifier sa performance sur des données inédites afin de garantir sa généralisation efficace et sa capacité à être utilisé de manière efficace sur des données réelles. On effectue cette évaluation en se basant sur différentes métriques qui offrent des solutions supplémentaires sur la capacité du modèle à faire des prédictions précises et solides.

L'évaluation du modèle, dans le cadre d'une tâche de classification binaire, prend en compte différents aspects :

Précision : Son indicateur est simplement "le nombre d'éléments de données choisis sont pertinents". Autrement dit, combien d'observations parmi lesquelles un algorithme a prédit des choses positives sont-elles en réalité positives?. D'après la formule (1), la précision correspond au nombre de vrais positifs de la somme des vrais positifs et des faux positifs.

Precicion= Équation III2

Rappel(Recall) : Il présente "combien d'éléments de données pertinents sont sélectionnés". En fait, parmi les observations qui sont réellement positives, combien d'entre elles ont été prédites par l'algorithme. Selon la formule (2), le rappel est égal au nombre de vrais positifs divisé par la somme des vrais positifs et des faux négatifs :

Recall= Équation III3

Score F1 : Le f-score ou f-mesure est une métrique qui tient compte à la fois de la précision et du rappel pour évaluer la performance d'un algorithme. En mathématiques, elle correspond à la moyenne harmonique de la précision et du rappel, qui est représentée de la manière suivante :

F1-Score=Recall= Équation III4

Acuuracy (exactitude): C'est le critère le plus couramment employé et peut-être le premier à être retenu pour évaluer les performances d'un algorithme dans les problèmes de classification. On peut la décrire comme la relation entre les données correctement classées et le nombre total d'observations (formule (4)). Bien que très pratique, l'exactitude n'est pas la mesure de performance la plus adaptée dans certaines circonstances, notamment lorsque les classes de la variable cible dans le jeu de données sont déséquilibrées.

accuracy= Équation III5

Matrice de confusion : Cette matrice est l'une des métriques les plus intuitives et descriptives utilisées pour évaluer l'exactitude et la précision d'un algorithme d'apprentissage automatique.

Son utilisation principale est dans les problèmes de classification où la sortie peut contenir deux types de classes ou plus. Pour plus d'informations.

Score ROC-AUC : On calcule cette mesure en utilisant la courbe ROC (courbe des caractéristiques de fonctionnement du récepteur) qui illustre la corrélation entre le taux de vrais positifs (connus également sous le nom de sensibilité ou rappel) et le taux de faux positifs (1 - spécificité). La surface sous la courbe ROC, ou ROC-AUC, est utilisée pour la classification binaire et démontre la capacité d'un modèle à discriminer les classes cibles positives et négatives. En particulier, si l'importance des classes positives et négatives est égale pour nous, le score ROC-AUC peut être une métrique de performance utile.

I.3. Conclusion

Les étapes essentielles de notre conception de notre modèle a été abordé en détail dans ce chapitre. La collecte des données a été la première étape, qui est cruciale pour obtenir un ensemble de données fiable et représentatif de notre problématique. Par la suite, les données ont été prétraites et les caractéristiques ont été sélectionnées afin de garantir leur nettoyage, leur normalisation et leur pertinence pour les algorithmes sélectionnés. Nous avons établi l'architecture globale de notre travail, décrivant comment les différentes étapes du processus s'articulent, de la collecte des données à l'évaluation des modèles.

Chapitre 4

Implémentation, tests et résultats

IV.1. Introduction

Au cours de ce chapitre, nous allons examiner en détail les résultats obtenus après avoir utilisé divers modèles de machine learning sur notre base de données. Nous étudierons les résultats des modèles à travers différentes étapes essentielles. Ce chapitre est structuré en différentes parties principales.

IV.2. Les outils de développement

Python représente un langage de programmation à la fois général et avancé. C'est un logiciel gratuit, open source et multiplateforme. Il est aussi adapté aux différents types de données : valeurs numériques, chaînes, listes, n-uplets et dictionnaires. En outre, c'est un langage d'interprétation. L'interprète lit le code source ligne par ligne. Par conséquent, ce langage est plus lent que les langages compilés tels que C et C++. Ce langage possède une syntaxe simple et facile à comprendre. En outre, Python est compatible avec des bases de données comme MySQL et MSSQL. En somme, Python est un langage polyvalent qui offre la possibilité de concevoir différentes applications. Il est apprécié par les novices et les programmeurs(Wikipedia, 2024)

Figure IV1:logo du langage python

IV.2.1. Anaconda

La distribution libre et open source d'Anaconda est utilisée pour développer des applications axées sur la science des données et l'apprentissage automatique, telles que le traitement de données à grande échelle, l'analyse prédictive et le calcul scientifique. Son objectif est de faciliter la gestion des paquets et le déploiement(Wikipedia, 2023)

IV.2.2. Jupyter

Jupyter est un logiciel en ligne qui permet de programmer dans plus de 40 langages de programmation, tels que Python, Julia, Ruby, R et Scala respectivement. C'est un projet

collectif pour la conception de logiciels libres, de formats ouverts et de services pour l'informatique interactive. Jupyter utilise le projet IPython pour son développement. Jupyter permet de concevoir des calepins ou des notebooks, c'est-à-dire des programmes comprenant à la fois du texte en mode mark down et du code en Julia, Python ou R. Ces calepins sont utilisés en science des données pour explorer et analyser des données(Wikipedia, 2014).

IV.2.3. Django

Django est une plateforme web libre basée sur Python. L'objectif est de faciliter le développement d'applications web en utilisant du code réutilisable. Django est un logiciel développé en 2003 pour le journal local de Lawrence (État du Kansas, aux États-Unis) et publié sous licence BSD depuis juillet 2005. Depuis juin 2008, le framework est développé et promu par la Django Software Foundation. Outre cette promotion régulière, des rencontres entre développeurs et utilisateurs de Django sont organisées deux fois par an depuis 2008. Les DjangoCon sont des conférences qui se déroulent en Europe et aux États-Unis (Wikipédia, 2024).

IV.2.4. Bibliothèques essentielles

La bibliothèque de programmes ou librairie logicielle est un ensemble de fonctions utilitaires regroupées et disponibles pour être utilisées sans avoir à les réécrire. Les fonctions sont classées selon qu'elles relèvent d'un même domaine conceptuel (mathématique, graphique, tris, etc.). La bibliothèque classique de Python est très étendue et offre une multitude d'outils.Les bibliothèques utilisées dans notre mise en œuvre seront définies ci-dessous :

IV.2.4.1. Pandas

Pandas, une bibliothèque open source, offre des structures de données et des outils d'analyse de données performants et simples à utiliser pour le langage de programmation Python(Development, 2024).

.

IV.2.4.2. Matplotlib

Matplotlib est une bibliothèque Python complète qui offre la possibilité de créer des visualisations statiques., animées et interactives(Wikipedia, 2024).

IV.2.4.3. Numpy

NumPy constitue un élément indispensable pour les calculs scientifiques en Python. C'est une bibliothèque Python qui intègre un tableau multidimensionnel, divers objets dérivés (tableaux et matrices masqués) et un ensemble de routines pour réaliser des opérations rapides des tableaux. Selon les auteurs, il y a différentes opérations telles que des opérations mathématiques, logiques, de manipulation de formes, de tri, de sélection, d'entrées/sorties, des transformées de fourrier discrètes, des bases d'algèbre linéaire, des opérations statistiques, de simulation aléatoire et bien d'autres encore(Wikipedia, 2024).

IV.2.4.4. Tensorflow

TensorFlow est une bibliothèque logicielle ou un cadre créé par Google qui permet de mettre en œuvre rapidement des techniques d'apprentissage automatique (ML) et d'apprentissage profond (DL). De nombreuses équations mathématiques sont simplifiées à calculer grâce à une combinaison d'algèbre computationnelle et de techniques d'optimisation. Il s'agit d'une bibliothèque open-source pour l'apprentissage automatique à grande échelle et le calcul numérique. Des modèles d'apprentissage automatique sont développés et entraînés en l'utilisant. Il emploie l'apprentissage automatique et l'intelligence artificielle pour améliorer le moteur de recherche, le sous-titrage d'images, la détection de motifs et d'autres fonctions. Dans l'apprentissage profond, les tenseurs sont la manière habituelle d'encoder les données. Les programmeurs peuvent élaborer des graphes de flux de données, qui sont des structures qui décrivent la circulation des données à travers un graphe ou un ensemble de nœuds de traitement. Il est équipé de différentes API (interfaces de programmation d'application). Celles-ci peuvent être divisées en deux groupes : bas niveau et haut niveau. TF simplifie la création de modèles ML pour le cloud, les appareils mobiles, les ordinateurs de bureau et le web, tant pour les débutants que pour les développeurs expérimentés (Ramchandani et al., 2022).

IV.2.4.5. Scikit-learn

Scikit-learn est une extension Python gratuite pour l'apprentissage automatique. De nombreux acteurs, notamment dans le domaine académique, contribuent à son développement, tels que des instituts français d'enseignement supérieur et de recherche tels qu'Inria.
Dans son framework, elle offre de nombreuses bibliothèques d'algorithmes à mettre en œuvre facilement. Ces bibliothèques sont disponibles, en particulier pour les chercheurs en données.
En particulier, elle inclut des fonctionnalités pour évaluer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, ainsi que des machines à vecteurs de support.

Son objectif est de s'adapter à d'autres bibliothèques libres Python, telles que NumPy et SciPy(Contributors, 2024).

IV.3. Présentation des résultats

Nous présentons les résultats obtenus pour chaque modèle en termes de précision, rappel, F1-score, exactitude, matrice de confusion et score ROC-AUC. Apres les étapes précédentes c'est à dire de la collecte des données à l'évaluation des modèles qu'on a abordé à la section conception du chapitre précédent. C'est le temps maintenant de fournir le fruit de notre travail. Apres avoir évalué nos différents modèles en se basant sur les critères de l'évaluation en occurrence les métriques prêt-cités, les modèles nous ont permis d'en avoir des bons résultats présentés dans le Tableau IV1. Pour plus de visualisation nous allons présenter ces résultats graphiquement. La Figure IV2 nous donne les résultats de métriques des différents modèles utiliser dans le cadre notre étude.

modèles	Precision	Recall	F1-Score	Accuracy	Score ROC-AUC
KNN	0.846794	0.845880	0.842139	0.84	0.8271
SVM	0.928303	0.921162	0.919008	0.92	0.9924
RF	1.00	1.00	1.00	1.00	1.0
DT	1.00	1.00	1.00	1.00	1.0
ANN	0.984892	0.984588	0.984523	0.98	0.9969

Tableau IV1: les performances des modèles choisis.

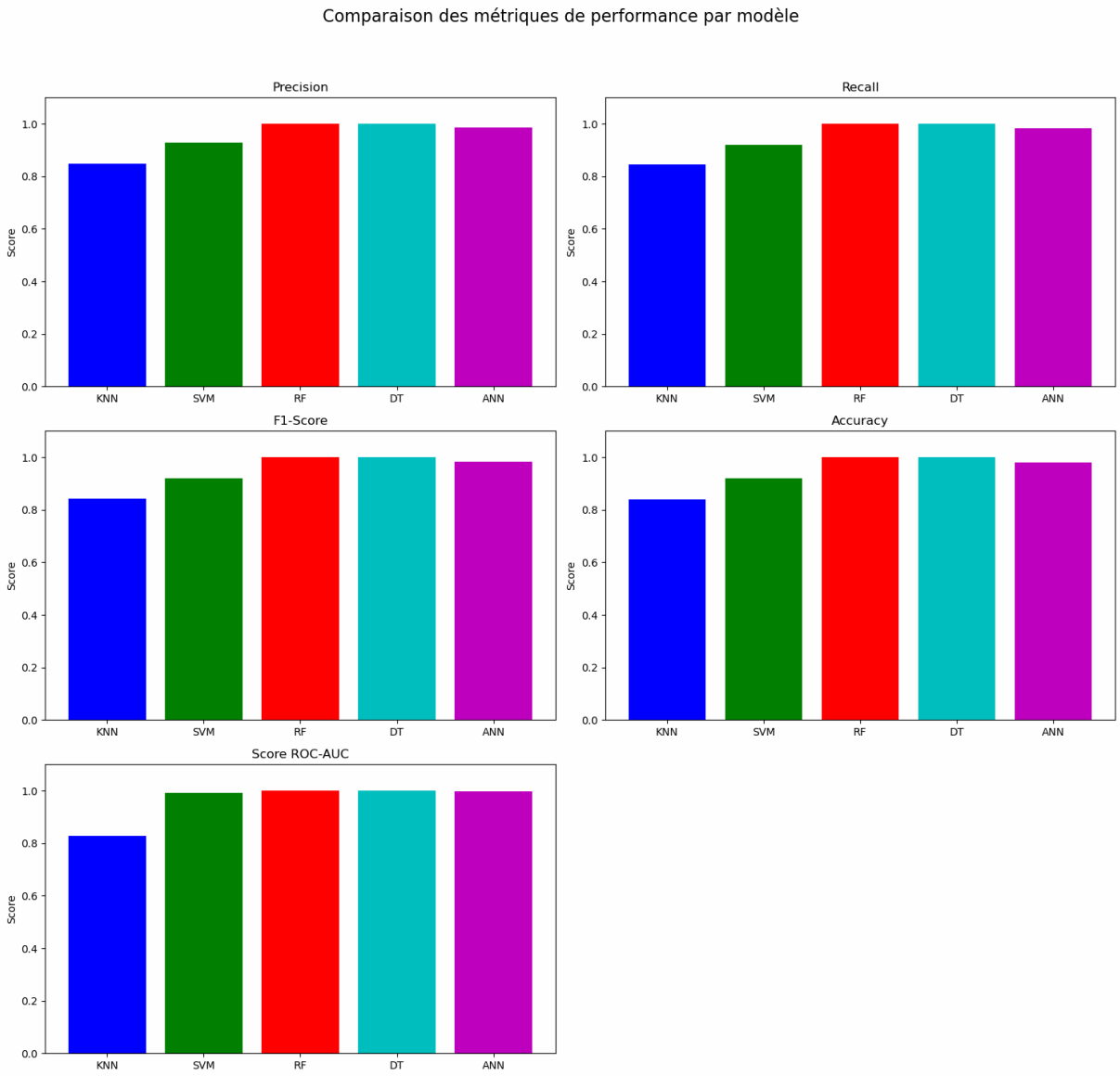

Figure IV2: Performances des modèles choisis.

Ce jeu de données est lié à un problème de classification binaire. Il convient de mentionner qu'il contient une caractéristique nommée "H or B" qui est notre variable cible par laquelle nous essayons de faire une prédiction sur une période donne. Nous allons soigneusement procéder à l'analyse des résultats donnés par les différents modèles en se basant sur les mesures de performances pour pouvoir les interpréter afin d'avoir une vision globale sur l'ensemble des modèles utilisés lequel de ces modèles est meilleur par rapport aux autres. Comme nous pouvons le voir dans la figure 1 tout comme dans le tableau 3 du chapitre 3, plusieurs algorithmes ont démontré des performances exceptionnelles pour ce jeu de données. Pour la précision, le score F1, le rappel (Recall), l'accuracy, et Score ROC-AUC pour les modèles RF,DT,LR,SVM et ANN ont obtenu les valeurs les plus élevées par rapport autres modèles.

Les résultats des diverses analyses des algorithmes de classification sur notre échantillon de données mettent en évidence des disparités significatives en ce qui concerne leurs performances. Le modèle K-Nearest Neighbors (KNN) affiche des résultats modestes avec des valeurs de précision, de rappel, de F1, et d'accuracy autour de 0.84, ainsi qu'un score ROC-AUC de 0.8271, indiquant une performance moindre pour ce problème de classification. Le Support Vector Machine (SVM) montre des résultats satisfaisants avec des scores élevés de précision (0.93), de rappel (0.92), de F1 (0.92), d'accuracy (0.92), et un score ROC-AUC de 0.9924, ce qui démontre une bonne capacité à distinguer correctement entre les classes. Les modèles Random Forest (RF) et Decision Tree (DT) présentent des résultats parfaits sur toutes les métriques (1.00), montrant une meilleure performance de ces modèles sur les données de test. Le réseau de neurones artificiel (ANN) a donné également une meilleure performances avec des scores élevés de précision (0.98), de rappel (0.98), de F1 (0.98), d'accuracy (0.98), et un score ROC-AUC de 0.9969, indiquant une capacité à bien classifier les données de test. En résumé, bien que les modèles Random Forest et Decision Tree montrent des mesures de performances remarquables , le réseau de neurones artificiel (ANN) et le Support Vector Machine (SVM) se révèlent être des options solides offrant des résultats satisfaisants.

IV.3.1. Analyse des performances des modèles

Une analyse critique des performances pour nos deux modèles choisi parmi tant d'autres sera effectuée. Nous examinerons les points forts et les points faibles de chaque algorithme, ainsi que leur comportement par rapport aux différentes métriques.

Avant d'afficher les résultats sous forme de graphiques, il est important de visualiser les performances des modèles Forest (RF) et Decision Tree (DT) et ANN tout au long de leur processus d'entraînement car ces trois modèles montrent des meilleures a l'égard des autres modèles. Les courbes de perte et d'exactitude pour l'entraînement et la validation pour le modèle ANN offrent des informations cruciales sur la convergence des modèles et leur capacité à généraliser sur les données non vues. Le graphique suivant montre les courbes de perte et d'exactitude pour le modèle ANN au fil des époques. Ces courbes permettent de mettre en relief visuellement la performance du modèle et d'évaluer son comportement pendant l'entraînement et la validation.

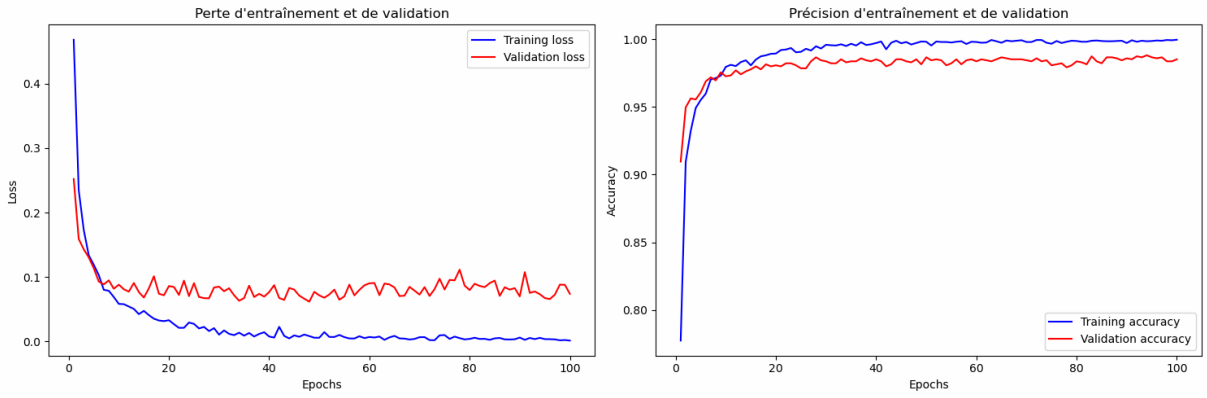

Figure IV3:Courbes d'entrainement et de validation du modèle ANN

IV.3.2. Comparaison de RF, DT vs ANN

Pour parvenir notre a but qui est le choix d'un meilleure modèle qui soit capable pour une meilleure prédiction ,il va falloir de faire une comparaison entre ces trois modèles à savoir le RandomForest (RF) et Decision Tree (DT) et ANN qui ont été choisi entres tant d'autres modèles comme les meilleures afin d'en trouver celui semble être meilleure .Une fois les résultats de RandomForest (RF) et Decision Tree (DT) et du Réseau de Neurones Artificiel (ANN) évalués sur notre jeu de données, nous remarquons des disparités notables en ce qui concerne la précision, le rappel et le score F1, exactitude et score ROC-AUC. Voici un résumé des résultats pour ces deux modèles :

métriques	**randomForest (RF)**	**Decision Tree (DT)**	**ANN**
Precision	1.00	1.00	0.98
Recall	1.00	1.00	0.98
F1-Score	1.00	1.00	0.98
Accuracy	1.00	1,0	0.98

Tableau IV2:comparaison entre Random Forest , Decision Tree et les réseaux de neurones

Les performances des modèles Random Forest (RF), Decision Tree (DT) et Artificial Neural Network (ANN) sur notre ensemble de données présentent des disparités significatives.

Le modèle Random Forest (RF) ainsi que le modèle Decision Tree (DT) offre des résultats parfaits sur toutes les mesures disponibles. Les résultats obtenus sont des scores de précision, de rappel (recal), de score F1 et d'exactitude (accuracy) de 1.00, ce qui démontre leur aptitude

à classer les données de test de manière pertinente. Ces résultats remarquables laissent penser que ces modèles sont très efficaces pour résoudre ce problème de classification.

Par contre, le réseau de neurones artificiel (ANN) offre des résultats légèrement moins bons mais tout de même remarquables. Il obtient une précision, un rappel et un F1 de 0.98, ainsi qu'une précision de 0.98. Il est évident que les résultats montrent que le modèle ANN est également extrêmement efficace, mais qu'il peut faire quelques erreurs par rapport aux modèles RF et DT.

Considérations Pratiques

En plus des performances métriques, Il est essentiel de prendre en compte les aspects concrets de la mise en œuvre et de la mise en œuvre des modèles:

- ✓ **Complexité Computationnelle :**

Les réseaux de neurones, tels que l'ANN, requièrent fréquemment des ressources computationnelles plus élevées pour leur entraînement, surtout lorsqu'il s'agit d'un nombre élevé d'epochs. Afin d'obtenir une performance optimale, l'ANN a été contraint de produire 100 épisodes, ce qui peut être chronophage et nécessiter beaucoup de temps et de CPU. Cela peut présenter des difficultés en ce qui concerne les ressources matérielles et le temps de traitement, en particulier si l'infrastructure disponible ne dispose pas de GPU ou de CPU puissants.

- ✓ **Facilité implémentation:**

Les modèles RF et DT présentent une facilité d'implémentation et d'exécution accrue. En comparaison avec un réseau de neurones, ils requièrent moins de configurations et de ressources computationnelles, ce qui le rendent plus adaptés à des environnements peu riches en ressources

Au vu de tout ce qui précède nous pouvons dire que les trois modèles, modèles Random Forest (RF), Decision Tree (DT) et Réseau de Neurones Artificiel (ANN), affichent des performances impressionnantes. Cependant, modèles Random Forest (RF), Decision Tree (DT) présentent une légère supériorité sur toutes les métriques. En particulier, modèles Random Forest (RF), Decision Tree (DT) on une précision, un rappel, un score F1 et une exactitude légèrement meilleure que l'ANN.

Étant donnée les résultats de performance et les aspects pratiques, il semble que les modèles Random Forest (RF), Decision Tree (DT) soient la solution la plus appropriée pour notre projet. Leur précision et leur précision sont légèrement plus élevées, et son utilisation est moins coûteuse en termes de ressources computationnelles. Ceci favorise une utilisation plus rapide et performante, ce qui revêt une importance capitale pour une application en production qui est soumise à des complexités de temps et de ressources matérielles.

IV.3.3. Comparaison avec les approches classiques

Notre problématique est la prédiction des hausses et baisses du prix de BIST 100. Cette tâche nous confronte à un problème de classification. Seuls les modèles d'apprentissage automatique dédiés à une tâche de classification nous permettront de faire une meilleure prédiction. Les méthodes traditionnelles comme ARIMA, SARIMA, GARCH, et VAR ont une vocation à prédire des variables continues.

Lorsqu'il s'agit de la prédiction des prix du BIST 100 en continu, ces méthodes traditionnelles sont les mieux placées. Cependant, tel n'est pas le cas dès lors que nous avons calculé le rendement du prix de BIST 100, puis transformé ce rendement en une variable catégorielle qui est notre variable cible. Dans cette situation, les méthodes traditionnelles ne sont plus appropriées.

IV.3.4. Implications et Limitations de l'Étude

Cette étude s'est concentrée sur l'application de modèles d'apprentissage automatique et d'apprentissage profond dans la prévision des actions. Il est à noter que les modèles d'apprentissage automatique ou d'apprentissage profond seuls ne sont pas suffisants. Les techniques d'ensemble sont capables de fournir des performances supérieures. Cependant, le simple développement d'un modèle ne suffit pas, et il est également important de mettre l'accent sur l'ajustement des hyperparamètres. La performance du modèle peut être améliorée en utilisant des hyperparamètres tels que la régularisation dans les cas d'apprentissage profond, le nombre de couches cachées, la profondeur maximale, le nombre d'estimateurs, et le taux d'apprentissage. Le choix des paramètres appropriés pour ces hyperparamètres peut considérablement améliorer la précision des prévisions boursières. Par exemple, un modèle avec un taux d'apprentissage plus élevé peut converger plus rapidement mais peut également être susceptible de surajustement, tandis qu'un modèle avec un taux d'apprentissage plus bas peut converger plus lentement mais mieux généraliser aux nouvelles données.

L'ajustement des hyperparamètres implique de choisir la combinaison optimale d'hyperparamètres pour un jeu de données et une architecture de modèle donnés. Cette procédure implique souvent de former plusieurs modèles avec des valeurs d'hyperparamètres distinctes et d'évaluer leurs performances sur un ensemble de validation. En général, la combinaison optimale d'hyperparamètres est déterminée par la capacité du modèle à réduire l'écart entre les prix des actions prédits et réels. L'analyse des modèles d'apprentissage automatique ainsi que les résultats comparatifs présentés dans ce projet de fin d'étude guideront les chercheurs dans le choix des algorithmes d'apprentissage automatique et d'apprentissage profond idéaux et préférés pour leurs travaux de recherche respectifs.

IV.3.5. Future recherche

Le projet de fin d'études exposé se focalise sur l'analyse des articles sur la prédiction, la prévision et la classification des prix des actions. Un défi majeur réside dans l'analyse des instruments financiers tels que les actions et les titres. Il est dit que le marché boursier évolue au fil du temps (Lim & Brooks, 2011; Sonkavde et al., 2023), et par conséquent, les approches développées pour traiter des problèmes spécifiques verront leurs performances diminuer tôt ou tard, même si leurs performances sont initialement appréciées.

Au fur et à mesure de l'évolution du marché boursier, influencé par différents éléments tels que les enjeux géopolitiques, le trading d'actions et les investissements, les défis sous-jacents évoluent également, tout comme les méthodologies pour aborder ces nouveaux défis (Shah, 2019; Sonkavde et al., 2023; Sprenger & Welpe, 2011). L'objectif principal de ce projet de fin d'étude est de présenter des recherches suffisantes sur la prédiction des prix des actions et la classification des actions.

D'après l'étude présentée dans ce travail, nous avons identifié certaines des principales zones où les chercheurs devraient concentrer leur attention et explorer de meilleures solutions. Dans cette section, une tentative est faite pour ouvrir de perspectives de recherche pour les chercheurs en recherche sur le marché boursier.

IV.4. Interface du système

Le modèle proposé a été complétée par une interface graphique développée à partir de django. Il est accessible via l'url du projet en activant le serveur. En premier lieu, l'utilisateur ouvre l'application et voici la fenêtre principale avec seulement quatre (4) boutons. Le

bouton(accueil) pour la page d'accueil la fenêtre principale, bouton de connexion et déconnexion, le bouton pour l'enregistrement et le bouton(prédiction) pour les chargement des données à prédire représentés dans (Figure IV4).

Figure IV4: page d'accueil de l'application

Le bouton connexion assure authentification pour la sécurité de notre application. Il est un devoir pour l'utilisateur de pouvoir se connecter ou déconnecter avec son compte (Figure IV5).

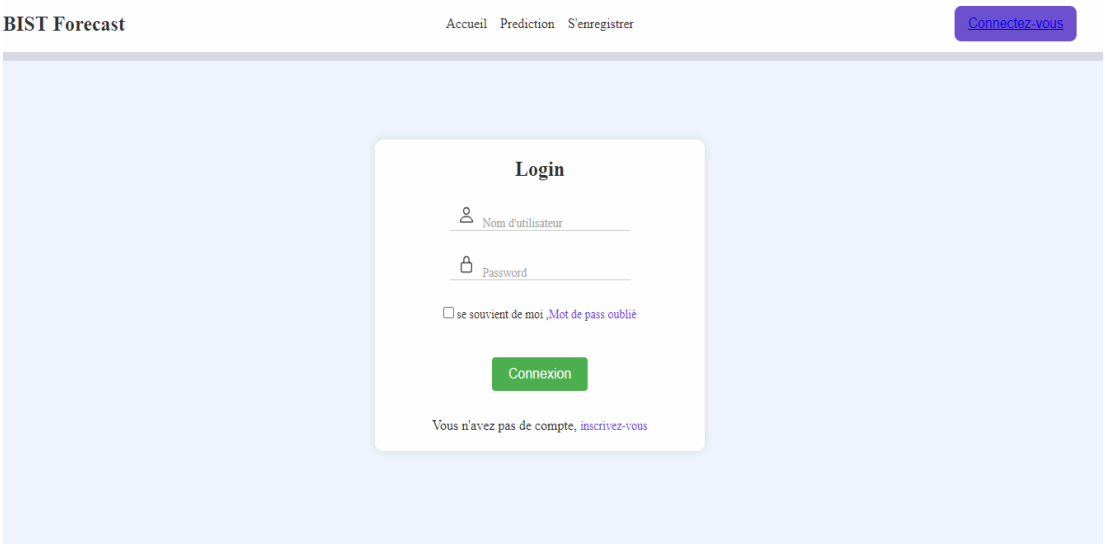

Figure IV5: login au système de prédiction de volatilité BIST 100

Avec en option la possibilité de pouvoir récupérer son mot de passe en d'oubli, il suffit juste de cliquer le lien mot de passe oublie, il se redirigera vers une autre page afin qu'il

puisse saisir son email de récupération. Apres avoir saisir l'email un message lui sera envoyé ou est inclut un lien qui va le diriger vers la page ou il va saisir un nouveau mot de passe.

Un des points très important est la prédiction, on ne peut pas permettre a n'import qui d'avoir un accès au bouton(Prediction) pour essayer de mener une prédiction sans qu'il soit connecté même si il clique sur ce bouton tant qu'il s'est pas connecté aucun accès ne lui sera accordé. Une fois connecte il aura un accès pour faire prédiction et pour faire une prédiction du hausse et baisse du prix il doit d'abord charger les données pour saisir la période de prédiction en cliquant sur le bouton faire prédiction (Figure IV6)

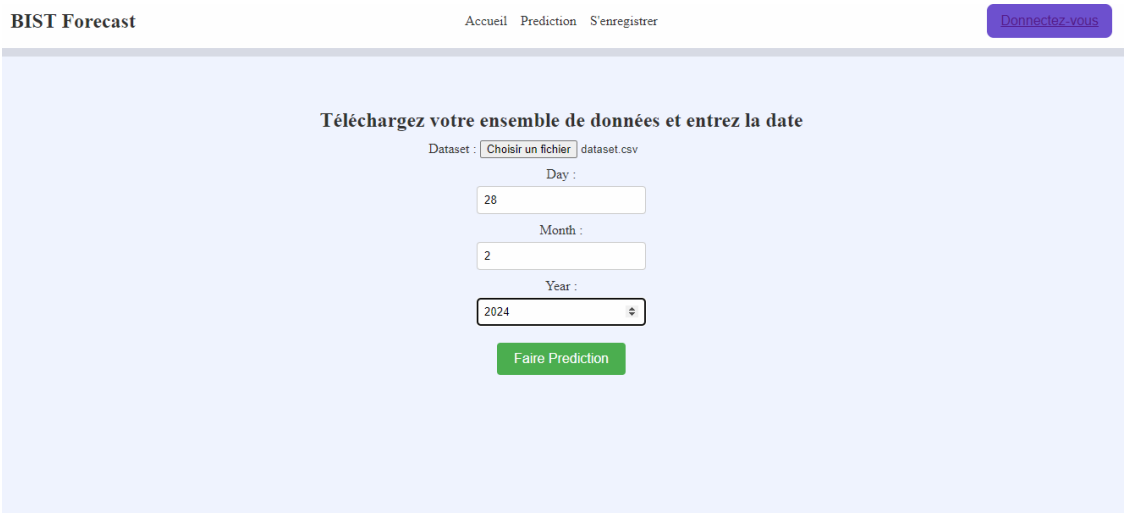

Figure IV6:page de prédiction

Lorsqu'il appuie sur le bouton de prédiction, il verra le résultat de la prédiction : 0 signifie une baisse de prédiction et 1 signifie une hausse de prédiction. (Figure IV7).

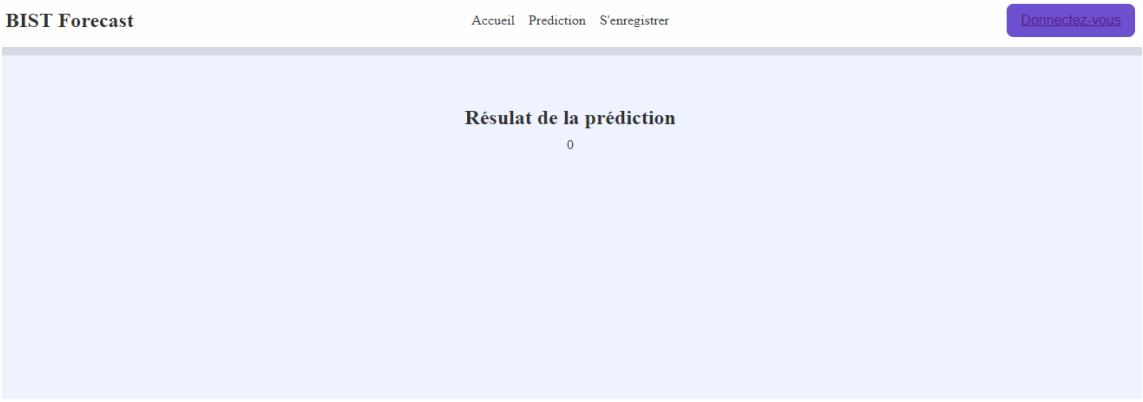

Figure IV7:Résultat de la prédiction

Si l'utilisateur saisi une date qui ne se trouve dans la base de données un message lui sera affiche pour lui informer qu'aucune donnée ne correspond aux valeurs saisies (Figure IV8).

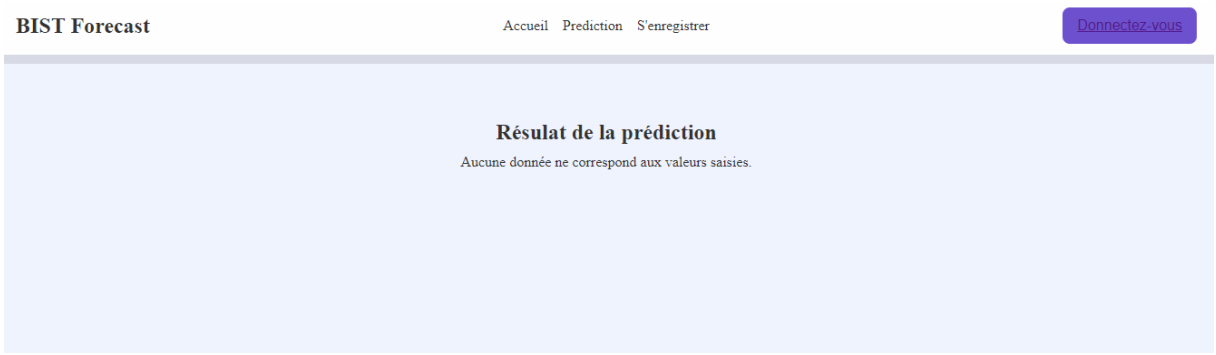

Figure IV8: affichage des résultats de prédiction

IV.5. Conclusion

En conclusion, après une analyse minutieuse de ces trois modèles, Random Forest (RF) et Decision Tree (DT) semblent être les choix le plus équilibrés pour notre projet en termes de performance et de simplicité, tandis que les modèles plus complexes comme l'ANN nécessitent une infrastructure informatique plus robuste pour être efficacement déployés.

CONCLUSION GENERALE

L'évaluation de l'incertitude liée aux fluctuations des prix des actifs est d'une importance capitale en finance, la volatilité étant une mesure essentielle. Il est crucial d'avoir la capacité de prédire cette volatilité afin de prendre des décisions d'investissement éclairées, de gérer les risques et de fixer les prix des produits financiers.

Cette étude met en lumière l'importance et le potentiel des modèles d'apprentissage automatique pour prédire la volatilité des marchés financiers. Selon les résultats obtenus, il est démontré que ces modèles peuvent offrir des prédictions précises et fiables, souvent supérieures aux techniques traditionnelles. Cependant, leur utilisation requiert une compréhension approfondie des concepts financiers, des compétences techniques solides en apprentissage automatique et une gestion minutieuse des données.

Les résultats de l'étude ont démontré que les modèles de forêts aléatoires et d'arbres de décision sont les plus efficaces pour prédire la volatilité des marchés financiers par rapport aux autres modèles étudiés.

Les perspectives d'avenir comprennent l'amélioration des modèles existants, l'étude de nouvelles architectures et méthodes d'apprentissage, ainsi que l'incorporation de sources de données supplémentaires pour améliorer la précision des prévisions. Ainsi, la disponibilité de ces technologies avancées peut fournir aux investisseurs et aux institutions financières des outils puissants pour naviguer dans des environnements économiques de plus en plus complexes et fluctuants.Grâce à cette expérience, notre compréhension de l'importance de l'apprentissage automatique s'est approfondie, non seulement dans le domaine de la finance et de l'économie, mais aussi dans d'autres secteurs.

Références bibliographiques

Ayyıldız, N. (2023). Prediction of Stock Market Index Movements with Machine Learning. In *Prediction of Stock Market Index Movements with Machine Learning* (Issue December). https://doi.org/10.58830/ozgur.pub354

Bachelier, L. (1900). Théorie de la spéculation. *Annales de l'École Normale Supérieure*, *17*, 21–86.

Baillie, R. T. ., & Bollerslev, T. (2002). The message in daily exchange rates: A conditional variance tale. *Journal of Business and Economic Statistics*, *20*(1), 60–68.

Beauchamp, W. D., Présenté, M., & Technologie, À. L. É. D. E. (2019). *Prédiction de la volatilité future dans le marché des devises à l ' aide de la volatilité implicite par*.

Black, F. ., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy*, *81*(3), 637–654.

Bluhm, H. H. W. ., & Yu, J. (2001). *Forecasting Volatility: Evidence from the German Stock Market*.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, *31*, 307–327.

Brooks, C. (1998). *Predicting Stock Index Volatility: Can Market Volume Help*. *17*, 59–80.

C. Aggarwal, C. et al. (2018). *Neural Networks and Deep Learning*. Springer.

Cappiello, L. et al. (2006). Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns. *Ournal of Financial Econometrics*, *4*, 537–572.

Christensen, B. J.; Hansen, C. S. (2002). New evidence on the implied-realized volatility relation. *The European Journal of Finance*, *8*(2), 187–205. https://doi.org/http://dx.doi.org/10.1080/13518470110071209

D.Allison, P. (1999). *Logistic Regression Using SAS: Theory and Application*. SAS Institute Inc.

DataKeen. (n.d.). *8 Machine Learning Algorithms Explained in Human Language*. Retrieved June 11, 2024, from https://datakeen.co/8-machine-learning-algorithms-explained-in-human-language/

Davidian, M.; Carroll, R. J. (1987). Variance Function Estimation. *Journal of the American Statistical Association*, *82*(400), 1079–1091.

Development, P. (2024). *Pandas Documentation*. https://pandas.pydata.org/docs/

Elizabeth, H. (2012). La volatilité selon les modèles GARCH, Focus sur l'asymétrie et la corrélation dynamique. *La Revue d'Opus Finance*, *1*.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with estimates of the variance of United. : *: Econometrica*, *50*(4), 987–1006.

Fama, E. F. (1991). Efficient Capital Markets: II. *Journal of Finance*, *46*(5), 1575–1617.

Fama, E. F. ., Fisher, L. ., Jensen, M. ., & Roll, R. (1969). The Adjustment of Stock Prices to New Information. *International Economic Review*, *10*(1), 1–21.

Farès, C. (2008). *Estimation et prévision de la volatilité de l'indice S&P 500*. Université du Québec à Montréal.

Forsberg, L.; Ghysels, E. (2007). Why Do Absolute Returns Predict Volatility So Well? *Journal of Financial Econometrics*, *7*(1), 31–67.

Frankel, J. A. ., Galli, G. ., & Giovannini, A. (1996). *The Microstructure of Foreign Exchange Markets*. University of Chicago Press.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Gharbi, S. (2013). *Réaction de la volatilité boursière aux annonces macro-économiques : cas de la Bourse de Paris*.

Grossman, S.; Stiglitz, J. (1980). On the Impossibility of Informationally Efficient Markets. *American Economic Review*, *70*(3).

Hamilton, J. D. ., & Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, *64*(2), 307–333.

Hansen, C. S. (2001). The relation between implied and realized volatility in the Danish option and equity markets. *Accounting and Finance*, *41*(3), 197–228. https://doi.org/http://dx.doi.org/10.1111/1467-629X.00059

Henrique, M. (2018). Stock price prediction using support vector regression on daily and up to the minute prices. *The Journal of Finance and Data Science*, *4*(3), 183–201. https://doi.org/10.1016/j.jfds.2018.04.003

Hull, J. ., & White, A. (1987). *The Pricing of Options on Assets with Stochastic Volatilities*. 281–230.

I. Vasilev, D. Slater, G. Spacagna, P. Roelants, V. Z. (2019). *Python Deep Learning: Exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow*. Packt Publishing Ltd.

J. Jaeger, A., & K. Eagan, M. (2011). Examining Retention and Contingent Faculty Use in a State System of Public Higher Education. *Educational Policy*, *25*(3), 507–537. https://doi.org/10.1177/0895904810361723

Jain, V., & Kulkarni, A. (2020). Survey on Various Algorithms of Machine Learning and its Applications. *International Research Journal of Engineering and Technology*. www.irjet.net

Jensen, M. C. (1978). Some Anomalous Evidence Regarding Market Efficiency. *Journal of Financial Economics*, *6*(2/3), 95–101.

Kermiche, L. (2008). Une modélisation de la surface de volatilité implicite par processus à sauts. *Finance*, *29*, 57–101.

Kim, J. (2003). *Financial time series forecasting using support vector machines*. *Neurocompu*(1–2), 307–319. https://doi.org/10.1016/S0925-2312(03)00372-2

King, B. (2015). Changing College Majors: Does it Happen More in STEM and Do Grades Matter? *Journal of College Science Teaching*, *44*(3). https://doi.org/10.2505/4/jcst15_044_03_44

Ko, K. (2015). Comparative Analysis of Job Motivation and Career Preference of Asian Undergraduate Students. *Public Personnel Management*, *44*(2), 192–213. https://doi.org/10.1177/0091026014559430

Kouaga, A. P. (n.d.). *Prédiction de la volatilité : cas de l'indice S&P/TSX*.

Li, S.; Yang, Q. (2009). The relationship between implied and realized volatility: evidence from the Australian. *Review of Quantitative Finance and Accounting*, *32*(4), 405–419. https://doi.org/http://dx.doi.org/10.1007/s11156-008-0099-2

Lim, K.-P., & Brooks, R. (2011). The evolution of stock market efficiency over time: A survey of the empirical literature. *Journal of Economic Surveys*, *25*, 69–108.

M.Chiang, H. et al. (2012). Predictive Factors of Participation in Postsecondary Education for High School Leavers with Autism. *Journal of Autism and Developmental Disorders*, *42*, 685–696. https://doi.org/10.1007/s10803-011-1297-7

Malkiel, B. G. (2003). The efficient market hypothesis and its. *Journal of Economic Perspectives*, *17*(1), 59–82. https://doi.org/10.1257/089533003321164958

Marcucci, J. (2005). Forecasting Stock Market Volatility with Regime-Switching GARCH Models. *Studies in Nonlinear Dynamics & Econometrics*, *9*(4), 1–55.

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, *7*, 77–91.

Martens, M. P. E. ., & Zein, J. (2002). *Predicting Financial Volatility: High-Frequency Time-Series Forecasts Vis-a-Vis Implied Volatility*.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica*, *59*(2), 347–370.

Niu, L. (2020). A review of the application of logistic regression in educational research: common issues, implications, and suggestions. *Educational Review*, *72*(1), 41–67. https://doi.org/10.1080/00131911.2018.1483892

Nokeri, T. C. (2021). Implementing machine learning for finance: A systematic approach to predictive risk and performance analysis for investment portfolios. In *Implementing Machine Learning for Finance: A Systematic Approach to Predictive Risk and Performance Analysis for Investment Portfolios*. https://doi.org/10.1007/978-1-4842-

7110-0

P-.Chen, L. (2019). *Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar: Foundations of Machine*.

Padhi, P., & Shaikh, I. (2014). On the relationship of implied, realized and historical volatility: evidence from NSE equity index options. *Journal of Business Economics and Management*, *15*(5), 915–934. https://doi.org/10.3846/16111699.2013.793605

Rabemananjara, R. ., & Zakoian, J. M. (1993). Threshold ARCH models and asymmetries in volatility. *Journal of Applied Econometrics*, *8*, 31–49.

Racicot, F.-É. ., & Théoret, R. (2005). *Quelques Applications du filtre de Kalman en Finance : Estimation et prévision de la volatilité stochastique et du rapport cours bénéfices*.

Ramchandani, M., Khandare, H., Singh, P., Rajak, P., Suryawanshi, N., Jangde, A. S., Arya, L., Kumar, P., & Sahu, M. (2022). Survey: Tensorflow in Machine Learning. *Journal of Physics: Conference Series*, *2273*(1). https://doi.org/10.1088/1742-6596/2273/1/012008

Ross, S. (1989). Information and Volatility: The No-Arbitrage Martingale Approach to Timing and. *Journal of Finance*, *44*, 1–17.

Sapankevych, I. (2009). ime series prediction using support vector machines: A survey. *IEEE Computational Intelligence Magazine*, *4*(2), 24–38. https://doi.org/10.1109/MCI.2009.932254

Shah, D. et al. (2019). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *International Journal of Financial Studies*, *7*, 26.

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *Journal of Finance*, *19*, 425–442.

Sonkavde, G., Dharrao, D. S., Bongale, A. M., Deokate, S. T., Doreswamy, D., & Bhat, S. K. (2023). Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications. *International Journal of Financial Studies*, *11*(3). https://doi.org/10.3390/ijfs11030094

Sprenger, T. O. ., & Welpe, I. M. (2011). News or noise? The stock market reaction to different types of company-specific news events. *SSRN Electronic Journal*.

Stiglitz, J. E. (1981). The Allocation Role of the Stock Market: Pareto Optimality and Competition. *The Journal of Finance*, *36*(2), 235–251.

Stoll, H. R. (1978). The Supply of Dealer Services in Securities Markets. *The Journal of Finance*, *33*(4), 1133–1151.

Sullivan, K., & Cosden, M. (2015). High School Risk Factors Associated with Alcohol Trajectories and College Alcohol Use. *Journal of Child & Adolescent Substance Abuse*, *24*, 19–27.

Taylor, S. J. (1986). *Modelling Financial Time Series*. John Wiley & Sons.

Taylor, S. J. (2007). *Modelling Financial Time Series* (2nd Editio). World Scientific.

Vapnik, V., & Cortes, C. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018

West, K. D. ., & Cho, D. (1995). The predictive ability of several models of exchange rate volatility. *Journal of Econometrics*, 367–391.

Zhang, M. Y. ., Russell, J. R. ., & Tsay, R. S. (2008). Determinants of bid and ask quotes and implications for the cost of trading. *Journal of Empirical Finance*, *15*(4), 656–678.

Webograhie

Wikipedia. (2014). *Jupyter*. https://fr.wikipedia.org/wiki/Jupyter#cite_note-kernels_community-2

Wikipedia. (2023). *Anaconda (distribution Python)*. https://fr.wikipedia.org/wiki/Anaconda_(distribution_Python)

Wikipedia. (2024a). *Matplotlib*. https://fr.wikipedia.org/wiki/Matplotlib

Wikipedia. (2024b). *NumPy*. https://fr.wikipedia.org/wiki/NumPy

Wikipedia. (2024c). *Python (langage)*. https://fr.wikipedia.org/wiki/Python_(langage)

Wikipedia. (2024d). *Scikit-learn*. https://fr.wikipedia.org/wiki/Scikit-learn

Wikipédia. (2024). *Django (framework)*. https://fr.wikipedia.org/wiki/Django_(framework)

DataKeen. (2024.). *8 Machine Learning Algorithms Explained in Human Language*. Retrieved June 11, 2024, from https://datakeen.co/8-machine-learning-algorithms-explained-in-human-language/

Wikipédia. (2024). Django (framework). https://fr.wikipedia.org/wiki/Django_(framework)