UCLouvain
FUCaM MONS

LOUVAIN
School of Management

Louvain School of Management

Etude des performances du Machine Learning dans la gestion de portefeuilles

Auteur : Arnaud De Paepe
Promoteur(s) : Frédéric Vrins
Co-promoteur : Mikael Petitjean
Année académique 2019-2020
Master 120 crédits en Ingénieur de Gestion, à finalité spécialisée

Remerciements

Je tiens tout d'abord à remercier mes parents qui m'ont permis de me lancer dans des études universitaires, qui m'ont motivé durant de ces cinq années et qui ont pris le temps de m'aider dans la réalisation de ce mémoire.

Je tiens ensuite à remercier mes amis et ma copine pour leur bienveillance et leur soutient durant les moments de doute, ainsi que pour leur disponibilité et contribution, de prêt ou de loin, dans ce travail.

Enfin, je remercie mon co-promoteur, Monsieur Mikael Petitjean, pour avoir accepté d'encadrer ce projet et pour sa disponibilité. Je souhaite finalement remercier mon promoteur, Monsieur Frédéric Vrins, pour m'avoir guidé et encadré tout au long de ce mémoire. Je lui exprime ma profonde gratitude, tant pour le temps consacré que pour la qualité de ses réponses.

Arnaud De Paepe

Résumé

Ce mémoire analyse les performances du Machine Learning dans la prédiction boursière, comparées à d'autres outils financiers dans trois secteurs distinctifs : la biotechnologie, l'immobilier et l'énergie. Nous allons dans un premier temps parcourir la littérature concernant la sélection de titres et les différentes méthodes couramment utilisées, pour ensuite s'attarder sur le Machine Learning et sur certains de ses algorithmes. De cette revue de littérature, nous observerons que trois grandes méthodes d'analyses financières existent : l'analyse technique, l'analyse fondamentale et l'analyse de sentiment. Nous retiendrons l'analyse technique pour la partie empirique, plus précisément les quatre indicateurs suivants : le Simple Moving Average (SMA), le Moving Average Convergence Divergence (MACD), le Relative Strength Index (RSI) et le Rate Of Change (ROC). Nous verrons également les algorithmes les plus cités de la littérature, qui seront ensuite utilisés dans la partie empirique : deux arbres de décision (CART et C5.0), une forêt aléatoire (RF), un réseau de neurones (ANN) et une machine à vecteurs de support (SVM). Le taux de précision (Accuracy) sera l'outil utilisé pour mesurer la capacité des modèles à prédire le cours du lendemain, à savoir s'il s'agira d'une tendance haussière ou baissière. Le Sharpe Ratio mesurera si chaque modèle propose une stratégie d'investissement intéressante ou non. En considérant la période du 1^{er} janvier 2011 au 1^{er} janvier 2018, et en utilisant un tracker par industrie, respectivement le XBI, le RWR et le XLE, on s'apercevra que le Machine Learning performe mieux que l'analyse technique pour les deux premiers domaines, tant sur le plan des rendements proposés qu'au niveau des taux de précision obtenus. Cependant, on ne peut pas affirmer avec certitude que les modèles ML battent les outils techniques dans le cas du secteur de l'énergie. De manière générale, on constate que les algorithmes ML performent bien dans le cas de cours boursier croissants sur la période considérée.

Table des matières

Table des figures

Liste des tableaux

1. Introduction

Un des plus grand challenges pour les gestionnaires de fonds et autres acteurs financiers est la prédiction des futures tendances boursières. Un marché étant une entité qui peut être extrêmement volatile, il est parfois difficile de prendre les bonnes décisions d'investissement. De plus, certaines théories, notamment celle de Random Walk qui est directement liée à la théorie de l'efficience des marchés, stipulent que le cours boursier fluctue de manière aléatoire, qu'aucune tendance ne régit son évolution, et qu'il est de ce fait impossible de prédire son cours (Sitte & Sitte, 2002). C'est dans ce cadre incertain que les investisseurs sont en constante recherche de nouvelles solutions d'investissement, capable de prédire l'évolution avec une grande précision, pour satisfaire leur objectif premier, à savoir la maximisation des gains et rendements.

Parallèlement, l'intelligence artificielle fait beaucoup parler d'elle, notamment le Machine Learning, qui se développe fortement dans d'autres domaines et qui semble proposer des résultats remarquables, ce qui explique sa prise d'ampleur. Les acteurs financiers vont donc commencer à s'y intéresser, en y voyant une potentielle solution aux problèmes de prédictions.

Dans ce mémoire, nous allons parcourir la littérature, comprendre l'investissement boursier, ainsi que les méthodes d'investissement courantes, pour après décrire le fonctionnement général du Machine Learning, son intérêt et les différents algorithmes qui le composent. Une fois tous les concepts compris et assimilés, nous allons

voir différents cas d'applications du Machine Learning à la finance, comparés à des modèles financiers plus traditionnelles. Ensuite, l'objectif est d'étudier le sujet et de comparer les performances, tant sur le plan des prédictions que des rendements proposés, de cinq modèles Machine Learning et de quatre outils d'analyse technique. Ainsi, l'objectif est de répondre à la question suivante :

"Comment performent les algorithmes de Machine Learning face aux méthodes d'analyse technique dans le cas des secteurs de la biotechnologie, de l'immobilier et de l'énergie ?"

Pour se faire, nous allons utiliser un tracker par domaine étudié, respectivement le XBI, le RWR et le XLE, où nos modèles vont donc tenter de prédire le cours boursier du lendemain (hausse ou baisse). Pour mesurer leur capacité de prédiction, nous utiliserons le taux de précision (accuracy). Le rendement cumulé par stratégie sera ensuite calculé et nous utiliserons le Sharpe Ratio pour définir le niveau de performance de nos stratégies. Toute cette analyse sera exécutée sur le logiciel R.

2. Théorie

2.1 Stock Picking

L'investissement boursier suscite beaucoup d'attention. Si on regarde en 2012, plus de 23.000 milliards de dollars en actions ont été échangés dans les carnets d'ordres électroniques sur les marchés boursiers américains. De plus, différentes institutions comme les sociétés cotées, les médias, ou encore les gouvernements rendent disponibles et publiques une grande quantité d'information, telle que des données macro-économiques ou encore des états financiers relatifs aux performances des sociétés cotées en bourse. Toutes ces données sont à la base du nouveau challenge pour les investisseurs, à savoir comprendre et pouvoir interpréter celles-ci pour pouvoir saisir une opportunité d'investissement plus rapidement que les autres investisseurs, dans un environnement de marché bruité et dynamique (Hu et al., 2015).

Le stock picking fait donc référence à une méthode d'investissement qui, basée sur une ensemble de critère, consiste à sélectionner un ensemble de titres financiers se rapportant à des entreprises. Il s'agira majoritairement de critères comptables ou relatifs au marché boursier.

Avant de parcourir différentes approches adoptées par les stock pickeurs, il est intéressant de se demander si de telles méthodes peuvent réellement avoir un caractère prédictif, dans un environnement où le marché est dit "efficient" (Malkiel & Fama, 1970). Ce terme désigne un marché où l'information est complètement et

instantanément reflétée dans le prix, avec l'hypothèse d'absence d'opportunité d'arbitrage (de profit sans risque). Fama distingue trois types d'efficience des marchés : l'efficience faible (weak hypothesis), l'efficience semi-forte (semi strong hypothesis) et l'efficience forte (strong hypothesis) (Cairn, 2007).

— L'efficience faible où le cours historique de l'action est directement reflété dans son prix

— L'efficience semi-forte où l'information publique est, avec le cours historique, reflétée dans le prix de l'action.

— L'efficience forte où l'information privée, en plus des informations citées précédemment, est reflétée dans le prix.

Certains pensent que les fonds d'investissements qui sur-performent sont seulement considérés comme chanceux, et ceux qui sous-performent malchanceux, et remettent donc en question les compétences des stock pickeurs. On constate que de nombreuses études ont été menées sur ce sujet, notamment celle Gray & Kern (2009) qui, en collaboration avec un club privé d'investissement composé majoritairement de gestionnaires d'hedge funds (VIC), montre des retours sur investissements positifs et croissants sur une échelle de 3 ans, et affirme donc que les stock pickeurs sont bien compétents dans leurs sélections de titres et remettent donc en question la théorie d'efficience des marchés.

Aussi, Duan, Hu & McLean (2009) nous révèlent que les marchés peuvent devenir "inefficient", et qu'il est donc possible de tirer profit de ceux-ci, si les coûts empêchent l'arbitrage d'être pleinement efficace. Les gestionnaires de fonds peuvent donc baser leur sélection sur ce critère et favoriser les actions avec un grand coût d'arbitrage car celles-ci peuvent être mal tarifées.

Un paradigme se consacre à l'étude des compétences des financiers. Il s'agit de la finance comportementale, qui cherche, de manière empirique, à analyser si les acteurs financiers adoptent un comportement rationnel lors de la sélection de titres. En effet, ces individus seraient sujet à certains biais, qui viennent donc impacter leur processus décisionnel. On soulignera ici les principaux biais comportementaux généralement adoptés par ces acteurs : le conservatisme, où l'individu ne modifie pas ses décisions après l'apparition de nouvelles informations ; la comptabilité mentale, où l'on considère des décisions séparément alors que celles-ci devraient être combi-

nées ; la représentativité, qui exprime le fait d'extrapoler à partir d'échantillons trop petits et la sur-confiance qui exprime le fait que les individus surestiment leurs compétences de prévision. Il existe bien d'autres biais qui influencent le comportement des gestionnaires de portefeuilles (Aftalion, 2005).

Deux types de gestions existent lorsque l'on parle de gestion de portefeuille. Il s'agira de la gestion active et la gestion passive. **La gestion active** désigne la situation où un investisseur arbitre parmi les actifs existants d'un marché. Cet investisseur construit un portefeuille qui maximise le rapport rendement/risque de sorte que celui-ci soit supérieur à celui du portefeuille de marché, appelé "benchmark". Ce style de gestion est généralement plus coûteux, car implique des coûts relatifs à la recherche d'information . Cette gestion s'oppose à **la gestion passive**, plus simple, où un gestionnaire tente de reproduire un portefeuille de marché, avec les mêmes rendements (Aaron, Bilon, Galanti & Tadjeddine, 2006). Cette gestion se base sur l'idée qu'il est impossible de battre l'indice de marché, comme dictée par la théorie Random Walk (Gorgulho, Neves & Horta, 2011).

En ce qui concerne la sélection des titres financiers, celle-ci peut s'effectuer selon différentes approches que nous allons parcourir à travers cette section. Nous distinguerons trois approches majeures différentes : l'analyse fondamentale, l'analyse technique et l'analyse de sentiment.

2.1.1 L'analyse fondamentale

L'analyse fondamentale étudie les caractéristiques comptables ou financières d'une entreprise. Les analystes qui se servent de cette méthode pour leurs prédictions vont donc s'attarder sur des données publiques, que l'on retrouve sur des bilans comptables ou encore des données macro-économiques, pour avoir une vue d'ensemble d'une société ou d'un marché, et les aider dans leurs décisions d'investissement. L'objectif est de déterminer la valeur intrinsèque d'une action grâce à des variables comme les dividendes, la croissance, les investissements, ... avec comme règle simple : "Si la valeur intrinsèque de l'actif est supérieure à la valeur qu'il a sur le marché, investissez dans cet actif. Sinon, considérez qu'il s'agit d'un mauvais investissement et évitez-le." (Dase, Pawar & Daspute, 2011). Cette analyse remet en cause la forme semi-forte de l'efficience des marchés, où l'information publique est directement re-

flétée dans le prix. L'analyse fondamentale se base sur trois aspects (Hu & al., 2015) :

— L'analyse macro-économique, qui va étudier l'impact de l'environnement macro-économique sur les futurs profits d'une société. On retrouvera comme indicateur le produit intérieur brut (PIB) ou encore l'indice de prix à la consommation (IPC).

— L'analyse de l'industrie, qui tente de prédire la valeur d'une société en fonction des prédictions d'opportunité d'un secteur, en analysant par exemple les profits des entreprises dans une industrie.

— L'analyse d'entreprise, qui examine l'état actuel d'une société pour en tirer sa valeur interne notamment par ses états et rapports financiers.

On peut examiner plus en détail cette dernière, l'analyse d'entreprise. En effet, différentes caractéristiques propres à une entreprise peuvent être utilisées par des investisseurs pour tenter d'en tirer des profits sur le marché. Il s'agit de :

— L'approche Value vs Growth

— Les caractéristiques propres aux investissements

— Les caractéristiques propres à la rentabilité

— Les caractéristiques intangibles d'une entreprise

♦ Approche Value vs Growth

On distingue ici deux styles de gestion différents.

- Les investisseurs adaptant une approche de croissance (Growth) vont chercher à investir dans les sociétés connaissant une augmentation de leurs bénéfices, de leurs ventes ou encore de leurs RoE (Return on Equity). Ces sociétés sont majoritairement associées à des P/E (Price-to-Earnings ratio) élevés, ainsi que des P/B (Price-to-Book ratio). Ces deux ratios peuvent être utilisés à des fins spéculatives car ils peuvent prédire une potentielle croissance des bénéfices.
- Le style de gestion valeur (Value) concernera principalement la sélection de sociétés dont les actions sont sous-évaluées par le marché, dont les actions sont peu populaires (les sociétés appartenant à des industries matures, laissant peu de place à l'innovation, et avec des attentes de faibles croissances). Il s'agit donc d'investir dans les actions n'ayant pas la cote, en attente d'une revalorisation de celles-ci. A l'inverse de la gestion Growth, cette approche se caractérise par de faibles P/E, P/B et P/CF (Price-to-Cash Flow ratio) (Ibbotson & Riepe, 1997).

Il existe un débat perpétuel pour tenter de déterminer lequel des deux styles est supérieur. Il semblerait tout de même que la combinaison de ces deux processus soit la meilleure décision envisageable (Fleischer, 2019).

♦ **Les caractéristiques propres aux investissements**

Ces caractéristiques capturent le niveau d'investissement d'une entreprise. Une société qui investit beaucoup dans son capital est le signe d'une stabilité et d'une force financière, et indique donc une opportunité d'investissement rentable. Des exemples de caractéristiques sont donc l'investissement en capital, les variations des immobilisations corporelles par rapport aux ventes ou encore les variations de stocks.

♦ **Les caractéristiques propres à la rentabilité**

Il s'agit de ce qui définit le niveau de rentabilité d'une entreprise. En effet, des bénéfices élevés et une croissance de rendements sont des signes positifs et pourraient potentiellement justifier une croissance du prix de l'action. Cette rentabilité s'explique par plusieurs caractéristiques d'une entreprise. On y retrouve par exemple le RoA (Return on Asset), le Cash flow-to-Debt ratio, le rendement sur capital investi, la croissance des ventes, ...

♦ **Les caractéristiques intangibles d'une entreprise**

Certains actifs non-physiques d'une société peuvent aussi avoir une grande valeur économique, tels que les brevets, une bonne gouvernance d'entreprise, la marque ou encore une technologie propre à une société. Les caractéristiques intangibles tentent de quantifier la valeur des actifs incorporels d'une entreprise. D'autres caractéristiques non-tangibles sont l'âge de la société ou le taux de financement des pensions (Fleischer, 2019).

2.1.2 L'analyse technique

Après avoir expliqué différentes méthodes d'investissement se basant sur des critères comptables et financiers d'une entreprise, nous allons maintenant nous attarder sur les critères propres aux marchés boursiers. Pour cela, il est important dans un premier temps de bien définir le marché et ce qui le compose. Un marché ou bourse

est donc un marché public, dominé par la loi de l'offre et la demande, où prend place l'échange de produits financiers, tels que des actions, obligations, ... entre des courtiers. Ces produits financiers sont premièrement issus par les entreprises ou entités gouvernementales (sur le marché primaire), et s'échangent ensuite sur le marché secondaire, où des traders se les vendent et se les rachètent. Ce marché a pour premier objectif de renforcer le capital des sociétés, où une action représente une part d'une entreprise. Mais il permet aussi la création d'opportunité d'investissement pour les particuliers, le partage des bénéfices notamment via les dividendes, mais aussi la participation aux décisions d'une entreprise, ... (Setty, Rangaswamy & Subramanya, 2010).

A partir de cela, nous allons maintenant expliquer la deuxième méthode d'analyse : l'analyse technique. Elle consiste en la prédiction du meilleur moment pour vendre ou acheter une action. Cette prédiction se base sur l'idée que le cours d'action suit des tendances, parfois cyclique, dictées par les investisseurs en réponses à différentes forces. Les analystes utilisent des graphiques qui vont les aider à observer ces tendances et à tenter de déterminer l'évolution du prix des actions. Ces graphiques se basent généralement sur différentes données telles que les prix les plus bas, les prix les plus élevés, le volume, ... Cette analyse technique, appelée aussi le "charting", a été largement utilisée, mais aussi largement critiquée. En effet, le sujet des critiques est que l'observation des tendances, à partir des graphiques, est très subjective et que de ce fait, différentes tendances peuvent être retirées de l'observation d'un même graphe (Dase et al., 2011). Cette analyse vient donc contredire la forme faible de l'efficience des marchés (où le cours historique est directement reflété dans le prix de l'action).

Différents outils, indicateurs sont utilisés pour tenter de prédire le cours des actions, en se basant sur les prix historiques et le volume des actions échangées. Nous allons en mentionner quelques-uns dans cette partie.

♦ **La moyenne mobile simple**

La moyenne mobile simple (SMA) est un de ces outils d'analyse technique. Cet indicateur va permettre la visualisation des tendances des prix des actions en calculant une moyenne des prix d'actions sur une période spécifique et va permettre d'assouplir cette tendance. Cette moyenne se calcule comme suit :

$$SMA_n = \sum_{i=1}^{n} \frac{x_i}{n} \tag{1}$$

où n désigne la période couverte par la moyenne, x_i le prix de clôture de l'action à l'instant i. En général, cette mesure est utilisée par les investisseurs sur des périodes de 5, 10, 20, 50, 100 et 200 jours. Une tendance peut être déterminée en calculant deux moyennes mobiles, une sur une courte période (par exemple 20 jours) et une autre sur une longue période (par exemple 200 jours). Le croisement de ces deux moyennes donne un signal aux investisseurs. Si la valeur de la moyenne mobile sur la courte période est supérieure à celle de la longue période, cela indique que la tendance du prix va augmenter. La tendance sera baissière dans le cas contraire (Lauren & Harlili, 2014).

♦ **La moyenne mobile exponentielle et la MACD**

La moyenne mobile exponentielle (EMA) est une autre moyenne mobile, qui attribue un poids plus élevé aux données récentes pour leur accorder plus d'importance. Son expression mathématique est :

$$EMA_i = EMA_{i-1} * (1 - \frac{2}{n+1}) + X_i * (\frac{2}{n+1}) \tag{2}$$

où n est la période couverte et X_i est le prix de l'action à la période i. Ici aussi, la tendance sera assouplie ce qui simplifie son analyse. Cette mesure est utile pour obtenir **la MACD (Moving average convergence divergence)** qui constitue l'un des indicateurs les plus fiables. Elle met en scène deux lignes : la première est la ligne MACD, qui exprime la relation entre deux moyennes mobiles exponentielles (EMA), à savoir une soustraction entre une moyenne mobile exponentielle court terme (de 12 périodes) à une long terme (de 26 périodes). La deuxième est la ligne de déclenchement (Trigger) qui correspond à une moyenne mobile exponentielle de

neuf jours d'une MACD. C'est la différence entre ces 2 lignes qui fournit des informations interprétables pour la prédiction de l'évolution future d'action, sous forme d'histogrammes (Gorgulho et al., 2011).

$$MACD(s,l) = EMA(s) - EMA(l) \tag{3}$$

$$Trigger(n) = EMA(n)\ d'\ une\ MACD(s,l) \tag{4}$$

$$Hist = MACD(s,l) - Trigger(n) \tag{5}$$

♦ **Le Rate of Change**

Le ROC (Rate of change) est une mesure qui présente la différence en pourcentage entre le prix de clôture actuel et le prix d'il y a n périodes. Cet instrument va nous permettre de mesurer la vitesse d'évolution du prix d'action. Ainsi, si le prix augmente (diminue) trop rapidement, cela indiquera probablement des conditions de surachat (de survente). On aura un signal d'achat si le ROC devient positif, et un signal de vente dans le cas inverse où il franchit la barre des valeurs négatives. Il se calcule ainsi :

$$ROC_i = \frac{X_i - X_{i-n}}{X_{i-n}} * 100 \tag{6}$$

où n est le nombre de périodes considérées et X_i est le prix d'action à la période i (Gorgulho et al., 2011).

♦ **Le Relative Strength Index**

Le RSI (Relative Strength Index) est un oscillateur dynamique, qui montre la solidité des prix en comparant les mouvements haussiers et de baisse des prix. Développé par Welles Wilder, le Relative Strength Index va donc comparer l'ampleur des gains et des pertes sur une période fixée pour en observer la variation des mouvements de prix d'une action. Au même titre que le ROC, il indiquera les conditions de surachat et de survente dans le cadre de négociation d'une action. De manière générale, le RSI se calcule sur 14 jours, et pour l'obtenir nous devons d'abord calculer la Relative Strength (RS) qui s'obtient en divisant une moyenne mobile exponentielle (EMA) de 14 périodes haussières par une moyenne mobile exponentielle de 14

périodes baissières :

$$RS = \frac{EMA(14)\ haussiere}{EMA(14)\ baissiere} \tag{7}$$

et est ensuite convertie en Relative Strength Index, qui prendra une valeur entre 0 et 100 :

$$RSI = 100 - 100 * \frac{1}{1 + RS} \tag{8}$$

Wilder soutient que les titres sont dits "surachetés" si la valeur de notre indicateur RSI indique une valeur supérieure ou égale à 70, et qu'il s'agisse donc d'un signal de vente pour les investisseurs. Dans le cas contraire, une valeur inférieure ou égale à 30 indique des conditions de survente de l'action, et indique à l'inverse un signal d'achat (Hari & Dewi, 2018).

2.1.3 L'analyse de sentiment

Cette analyse ne prend pas en compte les données historiques d'une action ou des données comptables et financières, mais considère plutôt des données non-quantifiables dans son approche. Il s'agira donc d'analyser le flux d'information propre à une société à travers la presse, des articles financiers, des publications, ... pour en sortir un sentiment global, soit positif, soit négatif. C'est sur base de ce sentiment, cette opinion publique qu'il sera possible de prédire la tendance boursière d'une entreprise, où un sentiment positif indique une potentielle tendance haussière et un sentiment négatif une tendance baissière (Kalyani, Bharathi & Jyothi, 2016).

Cette méthode, aussi appelée *opinion mining*, utilise donc de nombreuses sources textuelles pour en déduire un sentiment général, annonciateur d'une tendance. Pour se faire, l'analyse de texte, la linguistique informatique et le traitement de langage sont utilisés pour étudier ces nombreuses données. Par exemple, des algorithmes vont étudier les informations, qu'il s'agisse d'un article, un tweet, ... et vont assigner une valeur positive, négative ou neutre aux mots présents. Les mots comme "croissance", "bénéfice", "positif" se verront attribuer une valeur positive et des mots comme "faillite", "chute", "perte" recevront une valeur négative (Twinword, n.d). Il s'agit là d'un "lexique de sentiment". De là, il sera possible d'attribuer une valeur globale à l'information d'entrée. Cela peut se faire à différents niveaux. La tâche

peut être de classifier le sentiment exprimé d'un document complet, ou peut aussi se faire au niveau de la phrase, en examinant les différentes phrases d'un document et en leur attribuant une valeur positive ou négative. Cette analyse de sentiment est donc souvent associée à des algorithmes Machine Learning, utile pour l'étape de prédiction (Liu, 2012).

Dans la prochaine section, nous allons décrire les fondements du Machine Learning, le fonctionnement de celui-ci et les distinctions qui se font au sein du Machine Learning. Nous allons ensuite décrire différents algorithmes de Machine Learning qui seront utilisés dans ce projet.

2.2 Machine Learning

Le Machine Learning (ML) a connu un essor de son utilisation et de son application à des problèmes d'automatisation dans divers domaines. Récemment, il a connu un regain d'intérêt. En effet, les récents progrès ont permis d'adapter les algorithmes et de les rendre applicables à divers scénarios de la vie réelle. Son impact est considérable par exemple dans le domaine des soins de santé, ou les algorithmes ont permis de faciliter et d'améliorer les décisions prises en imagerie médicale, mais aussi pour les diagnostics assistés par ordinateur. Son impact est aussi fortement notable dans d'autres domaines, comme pour les voitures autonomes ou dans la domotique (Boutaba et al., 2018).

Le Machine Learning (l'apprentissage automatique) est une sous-branche de l'intelligence artificielle désignant l'ensemble des méthodes et algorithmes permettant à des machines de découvrir et d'apprendre des modèles, sans que ces dernières ne reçoivent d'instructions de programmation (Rasekhschaffe & Jones, 2019). Pour que cet auto-apprentissage puisse avoir lieu, les machines ont besoin de larges bases de données sur lesquelles elles vont s'exercer et apprendre de manière autonome. Le Machine Learning repose de ce fait sur le Big Data, qui désigne le stockage volumineux de données. Le Machine Learning va pouvoir tirer des relations, des liens de ces masses de données sans avoir besoin d'un support humain. Plus les données sont nombreuses, plus le Machine Learning peut être exploité efficacement (Lebigdata, 2018).

Le Machine Learning étant une science moderne qui peut sembler abstraite et difficile à percevoir, celle-ci ne date cependant pas d'hier. En effet, le premier algorithme fut inventé par Frank Rosenblatt en 1957. Il s'agit du perceptron, un outil permettant de classer des images et basé sur les réseaux de neurones (algorithme qui sera détaillé dans la section 2.3.1).

En dehors du secteur de la finance, le Machine Learning a déjà fait ses preuves comparé à des modèles statistiques traditionnels, notamment dans la reconnaissance des images (voitures autonomes) et vocale (Siri et Alexa). C'est avec le Machine Learning qu'on arrive à obtenir des robots capable de dépasser l'intelligence humaine. AlphaZero en est le parfait exemple. Ce robot est capable de battre le champion

du monde des échecs grâce à une méthode de reconnaissance. Aucune expertise ou connaissance ne lui a été apportée et cependant, il est devenu un maître dans l'art des échecs en seulement 4 heures en auto-apprentissage, c'est-à-dire en jouant contre lui-même. Cette possibilité d'offrir de telles performances s'explique aussi par le développement de certains facteurs (Rasekhschaffe & Jones, 2019) :

— La puissance de calcul des ordinateurs a fortement augmenté depuis 1970 si on se réfère à la loi de Moore

— La disponibilité des données, indispensable pour l'utilisation du Machine Learning, a considérablement augmenté, parallèlement aux coûts de stockage qui ont diminué

— La naissance de puissants nouveaux algorithmes basés sur la combinaison des deux points précédents, et de nouvelles méthodes informatiques/statistiques.

2.2.1 L'intérêt pour le Machine Learning en finance

Récemment, on constate la présence de nombreuses études analysant les comportements et prédictions du prix des actions et les variations d'indices, impliquant des algorithmes de Machine Learning. De nos jours, les traders utilisent des systèmes de trading intelligent pour mieux prédire l'évolution des prix face à différentes situations et conditions, ce qui facilitent la prise de décision des investisseurs. Les cours des actions fluctuent de manière très dynamique et sont susceptibles de changer rapidement, provoqués par la combinaison de paramètres connus (tels que les ratios P/E, le cours des jours précédents, ...) et de paramètres inconnus (les rumeurs, le résultat des prochaines élections, ...).

Normalement, un bon trader prédit le prix des actions et achète une action si son prix va augmenter dans un avenir proche, et revend une action si son prix s'apprête à diminuer. Cependant, la volatilité expliquée ci-dessus rend la tâche des traders bien difficile, car il est impossible de prédire avec 100% d'exactitude l'évolution du prix d'une action. C'est dans ce cadre-là que les investisseurs et gestionnaires de portefeuille ont commencé à se pencher vers le Machine Learning. Bien qu'il soit difficile de remplacer les compétences acquises par un trader expérimenté, un algorithme de prédiction précis peut également engendrer des bénéfices pour des entreprises d'investissements, mettant en évidence une relation entre la précision de l'algorithme et les profits de son utilisation (Shah, 2007).

Selon De Prado (2018), dans son livre intitulé "Advances in Financial Machine Learning", sa grande utilité s'explique notamment par le fait que de nombreuses opérations financières s'appuient sur des règles pré-définis, par exemple lorsqu'il s'agit de fixer un prix sur une option ou encore dans des cas de surveillance de risque. L'essentiel de l'automatisation en finance s'appliquait à ces cas, faisant des marchés financiers des entités super connectés où l'échange d'information s'effectue rapidement. Il était donc demandé aux machines d'exécuter ses tâches en suivant les règles aussi vite que possible. Cependant, la prochaine vague d'automatisation demandera à l'humain de baser sa décision sur son propre instinct, plutôt que de suivre les règles. Et étant un être sujet à de nombreux biais, tels que ses émotions, ses craintes et ses espoirs, l'humain n'est pas spécialement bon pour prendre des décisions fondées sur des faits. Pour empêcher tout biais de fausser les résultats, il est préférable que les investisseurs s'appuient sur les machines qui proposent une décision basée sur les faits appris de données concrètes. De plus, les marchés financiers sont sujet à de nombreuses lois. Une machine se conformera toujours aux lois si elle a été programmée à cet effet. Si une décision douteuse a été prise ou si une partie du processus venait à échouer, il est toujours possible pour les investisseurs de revenir sur les registres pour comprendre ce qu'il s'est passé.

2.2.2 Apprentissage supervisé vs. apprentissage non-supervisé

On parlera d'apprentissage supervisé lorsque les données ainsi que les solutions associées sont fournies à l'algorithme. Pendant l'apprentissage, l'algorithme va tenter de chercher, de déduire des relations entre les variables explicatives et les variables expliquées (étiquette). L'algorithme va s'exercer sur une base d'apprentissage contenant des exemples déjà traités. Les étiquettes peuvent correspondre à des valeurs continues (dans le cas de problèmes de régression) ou des catégories discrètes (dans le cas de problème de classification). Certains algorithmes, comme les arbres de décisions, peuvent être utilisés pour résoudre des problèmes de classification et de régression, tandis que les régressions linéaires ne permettent que la résolution de problèmes de régression.

En ce qui concerne les algorithmes d'apprentissage non-supervisé, ils sont conçus pour découvrir des structures cachées dans un ensemble de données sans étiquettes.

Dans ce cas-ci, le résultat souhaité est inconnu. On y retrouve notamment les algorithmes de regroupement (clustering) et de réduction de dimensions (Fleischer, 2019).

2.2.3 Algorithme de classification vs algorithme de régression

Cette section couvre une autre distinction fondamentale du Machine Learning et se différencie au niveau de l'output. Dans un algorithme de classification, l'objectif est de trouver une fonction de relation entre les données d'entrées et la variable de sortie discrète. Dans le cas de la régression, les variables de sorties seront continues. Prenons l'exemple où il faut classer les futurs mouvements du prix d'une action. L'algorithme de classification pourrait avoir deux variables de sorties discrètes, 1 si le prix de l'action augmente et -1 si celui-ci diminue. Alors que l'algorithme de régression va proposer en sortie des évolutions du prix de l'action en pourcentages continus (Rasekhschaffe & Jones, 2019)

2.2.4 Évaluer la performance des algorithmes de classification

Évaluer les performances d'un algorithme de classification est souvent plus délicat que d'évaluer celles d'un algorithme de régression. Plusieurs méthodes permettent cependant de mesurer les performances de l'algorithme. La première concerne *la matrice de confusion*. L'idée derrière cette matrice est de faciliter la visualisation des performances et de mesurer le pourcentage des individus mal classés. Plus ce pourcentage sera élevé, plus on peut remettre en question l'efficacité et les performances de l'algorithme. Sur les lignes de cette matrice se trouveront les classes prédites, tandis que chaque colonne représente une classe réelle (Géron, 2019). Une matrice de confusion pour un classificateur binaire est illustrée sur la figure 1 (Openclassrooms, 2020). Sur la diagonale principale de cette matrice se trouvent les cas correctement classés (TN et TP), toutes les autres cellules de la matrice contiennent des exemples mal classés. A partir de cette matrice, il nous est possible de calculer différentes mesures. La plus connue est *l'Accuracy* et mesure le nombre d'individus correctement classés sur le nombre total d'individus. Il s'agit donc de la somme des éléments de la diagonale de la matrice que l'on divise par le total de la matrice.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

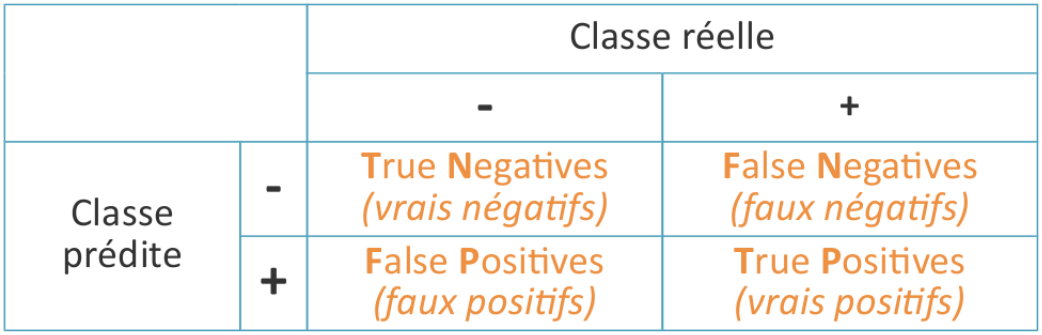


FIGURE 1 – Illustration d'une matrice de confusion

Cette matrice donne beaucoup d'informations, mais parfois nous préférons une mesure plus brève. On introduit donc ici la *Precision* d'un classificateur qui se mesure comme suit :

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

où TP désigne le nombre de True Positives et FP le nombre de False Positives. La *Precision* sera donc égale à 1, ce qui est un score parfait, si le classificateur prédit que des True Positives. Cette *Precision* est souvent utilisée en combinaison avec une autre mesure :

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

où FN désigne le nombre de False Negatives. *Recall* mesure le ratio des cas positifs qui ont été correctement détectés par l'algorithme (Géron, 2019). Dans certains cas, il peut être préférable de parvenir à détecter les cas indésirables que de détecter les cas souhaités. Dans une situation où l'on construit un algorithme de prédiction des prix des actions, il est jugé préférable de maximiser la *Precision* plutôt que le *Recall* car il semble plus judicieux d'atténuer les pertes de transactions plutôt que de sauter sur chaque opportunité (Fleischer, 2019).

Precision et *Recall* sont ainsi combinés dans une même mesure appelée F_1 score, qui est la moyenne harmonique des deux outils de mesures. La moyenne harmonique, à l'inverse de la moyenne arithmétique, va accorder un poids plus élevé aux faibles valeurs. De ce fait, le F_1 score d'un classificateur sera élevé si les mesures *Precision* et *Recall* sont toutes les deux élevées. Le F_1 score se calcule avec la formule 12. Malheureusement, augmenter la *Precision* diminue le *Recall*, et vice-versa ce qui

rend compliqué d'obtenir un F_1 score élevé. Ce phénomène s'appelle le compromis *Precision/Recall*.

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2 * \frac{precision * recall}{precision + recall} = \frac{TP}{TP + \frac{FN + FP}{2}} \tag{12}$$

Dans la suite de ce travail, nous parlerons et utiliserons les matrices de confusion et *l'Accuracy*, que nous désignerons par le terme de "précision".

2.2.5 Évaluer la performance des algorithmes de régression

En ce qui concerne les problèmes de régression, il existe également plusieurs méthodes d'évaluation. Nous retiendrons les trois mesures les plus citées dans la littérature : le Root Mean Square Error (RMSE), le Mean Absolute Error (MAE) et le R-squared (R^2).

Le Root Mean Square Error (RMSE), appliqué à un système, va permettre de définir quel est le niveau d'erreur que le système émettra dans ses prédictions. Un poids plus élevé est accordé aux larges erreurs. Cette mesure s'exprime sous la forme mathématique suivante :

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2} \tag{13}$$

où m désigne le nombre d'individus dans la base de données, $\hat{y}_i$ est la valeur prédite de la variable dépendante pour l'individu i et y_i est la valeur observée, attendue pour l'élément i. Une valeur de notre RMSE égale à 0 indiquerait que le modèle de régression s'adapte parfaitement aux données, et donc une petite valeur est préférable. Bien qu'il s'agisse là de l'outil de mesure le plus utilisé pour les problèmes de régression, il existe une autre mesure semblable, mais plus adaptée aux situations contenant de nombreux cas particuliers. On préférera dans ce cas appliquer *le Mean Absolute Error (MAE)*(Géron, 2019).

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |\hat{y}_i - y_i| \tag{14}$$

Un dernier outil souvent utilisé est *le R-squared (coefficient de détermination)* ou *l'adjusted R-squared*. Selon Miles (2014), ces deux mesures sont des statistiques dérivées d'analyses basées sur le modèle linéaire général (régression, Anova, ...). Il représente la proportion de la variance expliquée par les variables de prédiction dans l'échantillon (R-squared) et un estimateur de la population (adjusted R-squarred). Il s'agit donc de la proportion de la variance dans la variable dépendante qui est prévisible à partir des variables indépendantes. De manière générale, plus la valeur sera proche de 1, mieux notre modèle sera en mesure de prédire notre variable dépendante, même si dans certaines situations un faible R^2 n'indique pas forcément une mauvaise prédiction. En effet, certains domaines d'études, comme ceux sur le comportement humain, présenteront toujours des variations inexpliquées. Si le modèle présente une variance élevée et si les résidus sont fortement dispersés, notre valeur de R^2 sera petite, mais la ligne de régression peut cependant toujours correspondre à la meilleure prédiction. (Medium, 2019). R^2 s'exprime mathématiquement selon la formule suivante :

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i^m (\hat{y}_i - y_i)^2}{\sum_i^m (y_i - \bar{y})^2} \tag{15}$$

où $\bar{y}$ est la moyenne des valeurs de la variable dépendante. Un autre potentiel problème avec le R^2 est que sa valeur peut augmenter avec l'ajout de variables, sans même que celles-ci ne possèdent de qualité prédictive. Pour cette raison, nous utiliserons une autre mesure qui tient compte du nombre de variables indépendantes et qui aide à indiquer si une variable est "inutile" ou non : l'adjusted R-squared. Au fur et à mesure que l'on ajoute des variables sans qualité prédictive, la valeur de notre adjusted R^2 sera faible et ceci se remarque dans sa formule mathématique ou k est le nombre de variables indépendantes (Miles, 2014) :

$$Adjusted\ R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1} \tag{16}$$

où n désigne la taille de l'échantillon. Sa valeur sera également comprise entre 0 et 1 et sera toujours égale ou inférieure à celle de R^2 (Medium, 2019).

2.3 Algorithmes de Machine Learning

Dans cette section, nous allons parcourir et détailler différents algorithmes, qui seront par la suite testés dans le cadre de ce projet. Nous verrons donc les réseaux de neurones artificiels, les machines à vecteurs de support, deux arbres de décision (CART et C5.0) ainsi que les algorithmes de forêt aléatoire.

2.3.1 Les réseaux de neurones artificiels

Ces algorithmes s'inspirent du réseau de neurones biologique, et permettent de résoudre divers problèmes de prédiction, d'optimisation ou encore de reconnaissance d'images. Des approches conventionnelles tentaient de résoudre ces problèmes, mais elles n'étaient performantes que dans leurs propres environnements et ne proposaient aucune flexibilité. Les réseaux de neurones artificiels (artificial neural network ou ANN) sont des alternatives intéressantes car ils s'adaptent aux différents domaines (Jain, Mao & Mohiuddin, 1996). Les ANN, qui sont des algorithmes d'apprentissage supervisé, peuvent être définis comme des structures inter-connectées composées d'éléments de traitements simples (appelés neurones ou noeuds artificiels) qui permettent le calcul massif, en parallèle, de données. Les réseaux de neurones sont attractifs pour certaines de leurs caractéristiques remarquables comme leur parallélisme élevé, leur apprentissage et adaptivité ou encore la non-linéarité qui permet donc de mieux s'adapter aux données (Basheer & Hajmeer, 2000). Un réseau se présente comme sur la figure 2, où un élément sphérique représente un neurone et où le réseau est divisé en différentes couches. On retrouvera d'abord la couche d'entrée, suivi d'une ou plusieurs couches intermédiaires que l'on appelle "hidden layer" ou donc couche cachée (une seule dans ce cas-ci), pour finalement arrivé sur la couche de sortie (Wang & Raj, 2015).

Le fonctionnement au niveau du neurone peut être illustré comme sur la figure 3 (Calas, 2009). Chaque neurone fait une somme pondérée de ses entrées (appelées aussi synapses), en attribuant donc un poids (w_i) à chaque valeur (x_i), et aura comme résultat de sortie une valeur qui dépendra de la fonction d'activation.

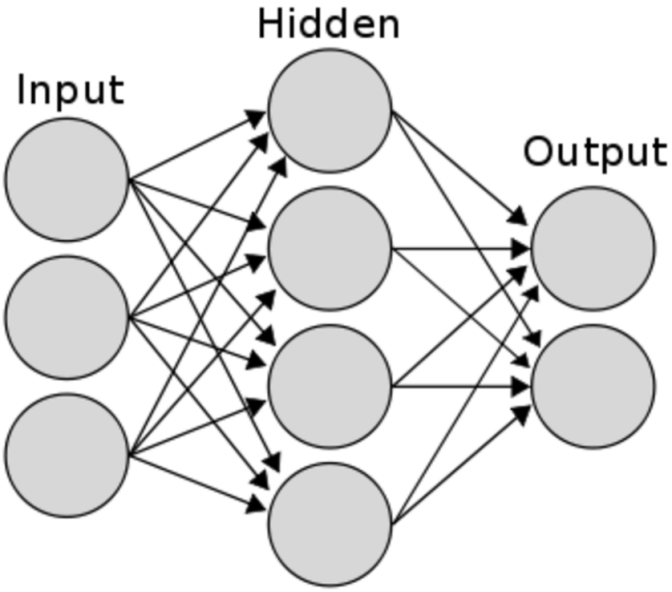


FIGURE 2 – Illustration d'un réseau de neurones artificiels avec une couche cachée

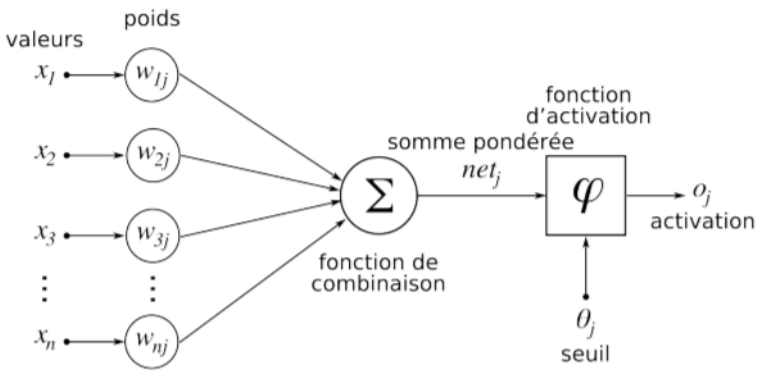


FIGURE 3 – Illustration du fonctionnement à l'échelle du neurone

Mathématiquement, l'activation du noeud j peut s'exprimer comme suit (Jain et al., 1996) :

$$o_j = \varphi\left[\sum_{i=1}^{n} w_i x_i - \theta_j\right] \tag{17}$$

où θ_j est un seuil propre au noeud j. Le résultat obtenu dépendra donc de la fonction d'activation utilisée φ. La valeur en sortie peut être utilisée soit comme un input d'une nouvelle couche de neurones, soit comme un résultat final. Les principaux types de fonctions sont (Wikistat, n.d) :

- *linéaire* : φ est la fonction d'identité
- *ReLU (Rectified linear unit)* : $\varphi(x) = \max(0,x)$
- *sigmoïde* : $\varphi(x) = \dfrac{1}{1+e^x}$
- *radicale* : $\varphi(x) = \sqrt{1/2\pi}e^{-x^2/2}$
- ...

2.3.2 Les machines à vecteurs de support

Les machines à vecteurs de support (support vector machine ou SVM) sont des algorithmes d'apprentissage supervisé, utiles tant pour les problèmes de classification que de régression, et dont l'objectif est de séparer les données en classes à l'aide d'un séparateur que l'on appellera "frontière" et qui va maximiser la distance entre ces classes, appelée "marge". Cette frontière, qui maximise la marge, s'appelle l'hyperplan de séparation optimale. La figure 4 en est une illustration, dans un modèle de classification binaire où on retrouve en vert notre hyperplan (Gunn, 1998). Cependant, cette "frontière" suppose donc que les données sont linéairement séparables, ce qui est rarement le cas. Les SVM vont donc transférer les données dans un espace vectoriel à plus grande dimension pour permettre une meilleure séparation de celles-ci. Cette fonction de maximisation de marge permet de proposer un modèle plus généralisable car offre une robustesse face aux données bruitées (DataAnalyticsPost, n.d.).

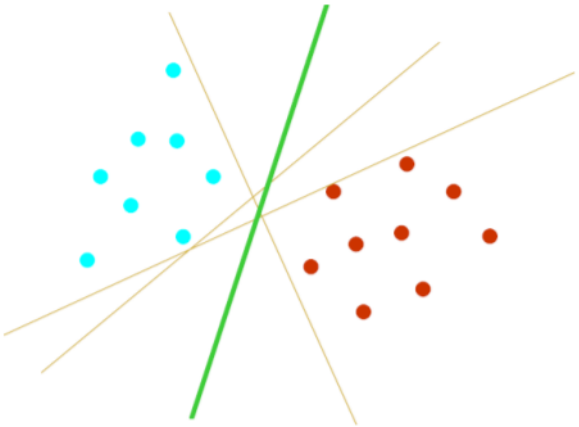


FIGURE 4 – Illustration d'un hyperplan de séparation optimale

2.3.3 Les arbres de décisions

Un arbre de décision (decision tree) est un algorithme d'apprentissage supervisé qui va permettre la prise de décision en prenant en entrée une population, un échantillon pour ensuite procéder à une catégorisation basée sur des facteurs discriminants. Cet outil va donc répartir les individus en groupes homogènes et va émettre des prédictions à partir de données connues. Un arbre de décision se présente comme sur la figure 5 (Dev, 2019).

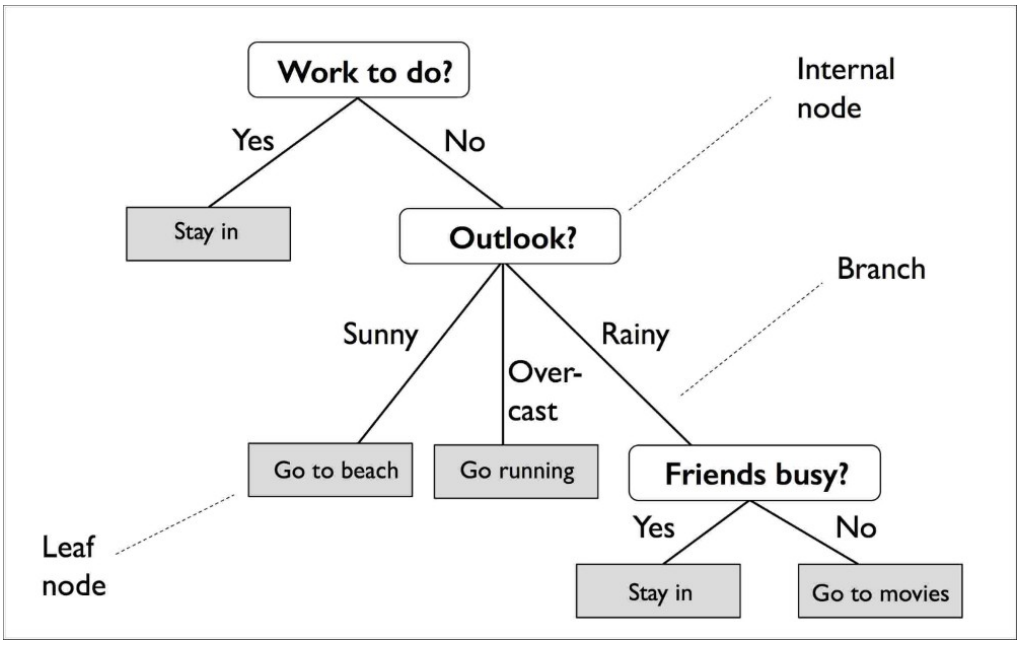


FIGURE 5 – Exemple d'arbre de décision

L'arbre commence par une "racine" (Work to do ?) où se trouve chacune de nos observations, et débouche après sur des intersections que nous appellerons "noeuds".

Les encadrés comme "Outlook ?" ou "Friends busy ?" correspondent à ces noeuds. Chaque noeud d'un arbre de décision porte sur un attribut discriminant des individus à classer, qui permet la classification de ces éléments de manière homogène entre les fils de chaque noeud. Ce qui lie un noeud et ses fils sont les "branches" qui contiennent les valeurs discriminantes de l'attribut du noeud. Sur notre illustration, les valeurs "Sunny" ou "Rainy", par exemple, appartiennent à des branches. Finalement, après avoir parcouru plusieurs noeuds, nous arrivons aux "feuilles" de l'arbre qui représentent les prédictions concernant les individus, les éléments à classifier. Sur la figure 5, il s'agit de "Stay in", "Go to movies", "Go to beach" ou encore "Go to running" (Calas, 2009).

2.3.3.1 L'algorithme CART

CART (Classification and regression tree) est un arbre de décision, comme étudié ci-dessus, strictement binaire et présentant deux branches pour chaque noeuds de l'arbre. Cet algorithme utilise l'indice de Gini pour déterminer dans quel attribut la branche devrait être générée. Cet indice mesure donc le degré d'inégalité pour une variable, sur un échantillon donné (Insee, n.d.). Le but est de choisir l'attribut qui va rendre l'indice de Gini minimal après la division du noeud (Patil, Lathi & Chitre, 2012). Cet indice se mesure comme suit :

$$Gini(S) = 1 - \sum_{i=1}^{m} P_i^2 \tag{18}$$

où S désigne l'ensemble, le noeud contenant m classes et P_i est la fréquence relative de la classe i dans l'ensemble S. Si S est pur, alors $Gini(S) = 0$. De ce fait, plus l'indice est faible, plus le noeud est considéré comme pur. Cependant, CART n'est pas approprié à des très grandes bases de données et même si ses performances ont déjà été prouvées, il est sujet au sur-apprentissage (overfitting, section 2.3.4.1) (Calas, 2009).

2.3.3.2 L'algorithme C5.0

Comme CART, C5.0 est un algorithme dont les fondements se basent sur les arbres de décisions. Cette amélioration du modèle C4.5, qui lui-même est une amélioration du modèle ID3, construit donc un arbre de manière récursive et divise l'échantillon en sélectionnant l'attribut qui maximise le gain d'information (Patil et

al., 2012). Le gain d'information peut s'obtenir comme suit :

$$Gain(S, A) = E(S) - \sum^{v} \frac{|S_v|}{|S|} * E(S_v) \tag{19}$$

où S est l'ensemble d'entraînement, A est l'attribut cible, S_v est le sous-ensemble de S qui contient les éléments dont la valeur de l'attribut est v, $|S|$ est le nombre d'éléments de S et $|S_v|$ le nombre d'éléments de S_v.
$E(S)$ correspond à l'entropie de Shannon. Cette entropie détermine la quantité d'information apportée par un événement et peut se mesurer comme suit :

$$E(S) = -\sum_{i=1}^{|S|} p(i) log_2 p(i) \tag{20}$$

où $p(i)$ est la probabilité d'avoir un individu de caractéristique i dans l'ensemble S (Calas, 2009). Comme pour l'indice de Gini, une faible entropie est préférable. La vitesse de l'algorithme C5.0 est considérablement plus élevée que celle de CART. De plus, CART ne résout que les problèmes par des arbres à tests binaires, tandis que les tests C5.0 permettent deux ou plusieurs résultats (Fakir & Ezzikouri, n.d.).

2.3.4 Les forêts aléatoires

Les algorithmes de forêts aléatoires (Random Forest ou RF) sont connus pour être des outils très efficaces de classification dans de nombreux domaines, notamment en finance. Il s'agit d'une méthode de classification d'ensemble, qui établit un ensemble de classificateurs, contrairement aux arbres de décisions CART et C5.0 qui ne construisent qu'un classificateur. Les RF sont donc une combinaison d'arbres de décisions, où la prédiction sera une moyenne pondérée des prédictions de chaque arbre (Fleischer, 2019). L'objectif de ces méthodes d'ensemble est donc de combiner des "weak learners", soit par des prévisions à pondération égale (bagging), ou par des prévisions à pondération précise (boosting) pour obtenir un "strong learner". (Rasekhschaffe & Jones, 2019).

♦ Bootstrap aggregating (bagging) vs. boosting

Parmi les algorithmes "bagging" (ou "bootstrap aggregating") dans lesquels on retrouve les Random Forest, des ensembles de formation sont formés en générant aléatoirement N exemples de l'ensemble de formation initiale, de taille N.

Ces exemples sont tirés avec replacement, c'est-à-dire que les exemples originaux peuvent être répétés plusieurs fois au sein d'un même ensemble de formation, tandis que d'autres seront laissés de côté. Chacun de ces ensembles est ensuite utilisé pour entraîner un différent modèle. La valeur en sortie de chaque ensemble sera après utilisée pour obtenir la valeur en sortie du modèle général (Maclin & Opitz, 1997).

Une autre méthode d'ensemble est le boosting dont l'objectif est aussi de construire un ensemble de classificateurs. Dans ce cas-ci, les "weak learners" sont formés séquentiellement et au fur et à mesure des itérations, les poids accordés aux exemples incorrectement prédits augmentent de manière à les rendre plus importants dans la prochaine itération. Et l'algorithme continue jusqu'à ce que tous les exemples se retrouvent correctement classés. Dans la majorité des cas, boosting performe mieux que bagging et donne des résultats plus précis, et permet de réduire la variance et le biais de la classification. Il peut donc être appliqué, par exemple, aux arbres de décisions CART où la variance est élevée (Akar & Güngör, 2012). Cependant, les algorithmes boosting prennent généralement plus de temps pour s'entraîner que les méthodes bagging parce qu'ils doivent être appliqués séquentiellement, tandis que les algorithmes bagging peuvent être exécutés en parallèle. (Rasekhschaffe & Jones, 2019).

2.3.4.1 Les dangers du sur-apprentissage

Lorsqu'un algorithme n'arrive pas à généraliser correctement, les prévisions sur les données hors échantillons ne seront pas précises et cela provient de deux phénomènes : le sur-apprentissage ou le sous-apprentissage. Le sur-apprentissage (ou overfitting) désigne le processus dans lequel un modèle s'adapte tellement aux données historiques qu'il en devient inefficace pour des prédictions futures. L'algorithme trouvera donc des relations dans les données d'entraînement qui au final ne s'appliquent pas dans le cas des données étudiées (Rasekhschaffe & Jones, 2019). Le sur-apprentissage est généralement caractérisé par une haute variance et un faible biais des estimateurs, ce qui est généralement le cas des algorithmes des arbres CART. Le sous-apprentissage (underfitting) est l'inverse et désigne le cas où l'algorithme n'apprend pas assez de relations que pour faire des prédictions précises, et se caractérise par une faible variance, mais un biais élevé (Jabbar & Khan, 2015).

De manière générale, les données financières posent problème lorsqu'il faut leur appliquer des algorithmes de Machine Learning. En effet, en finance, les relations entre les facteurs et les rendements sont souvent bruitées (noise) et entraînent donc de faibles signal-to-noise ratios. Un bruit (noise) est une information non-significative ou fausse tandis qu'un signal peut être défini comme une donnée utile. C'est parce que les signal-to-noise ratios sont faibles qu'il est important d'éviter le cas de sur-apprentissage car l'algorithme s'adaptera trop à des données non-pertinentes. Les méthodes d'ensemble (voir bagging et boosting dans la section 2.3.4) ainsi que les algorithmes Random Forest permettent d'atténuer les dangers du sur-apprentissage en combinant les prévisions des "weak learners" pour former un "strong learner" (Rasekhschaffe & Jones, 2019).

2.4 Revue de littérature

Dans cette section seront rassemblées et comparées un ensemble d'études ayant cherché à analyser les performances d'algorithmes Machine Learning. L'objectif est de pouvoir tirer une conclusion générale sur l'efficacité du ML, de voir s'ils surperforment les méthodes d'analyses financières classiques, sur base des études récoltées et de confronter les résultats à ceux obtenus dans notre partie empirique. Aujourd'hui, et plus que jamais, les investisseurs sont en constante recherche de supports pour les aider dans leurs décisions d'investissement dans le but de maximiser leur profit. Les recherches s'étendent par conséquent, impliquant ainsi d'autres domaines à la finance et l'utilisation de nouvelles méthodes qui peuvent les aider à atteindre cet objectif. La prédiction du cours boursier devient un domaine de plus en plus populaire, en particulier chez les spécialistes de l'intelligence artificielle qui tentent de combiner l'apprentissage, des éléments d'évolution et d'adaptation pour créer des systèmes de trading intelligents (Gorgulho, Neves & Horta, 2009).

La première étude sur laquelle nous allons nous intéresser est celle de Macchiarulo, réalisée en 2018. Son analyse se base sur le S&P 500, où les algorithmes se sont entraînés sur des données historiques mensuelles allant de janvier 1995 à décembre 2005, pour ensuite les appliquer sur une période de trading de 10 ans également, jusque décembre 2016. Dans cet article, l'auteur a d'abord utilisé les machines à vecteurs de support pour déterminer la direction du marché, et les réseaux de neurones pour déterminer les prix et les rendements réels des actions. Les deux algorithmes seront ensuite combinés par apprentissage ensembliste pour obtenir une prédiction. Une fois l'algorithme obtenu, celui-ci a été comparé à une vingtaine d'autres stratégies. On y retrouve notamment différents outils d'analyses techniques, de l'analyse fondamentale, des stratégies de Buy & Hold, ... Sur toute la période considérée, à savoir les 120 mois de la période teste, c'est bien l'algorithme de Machine Learning qui a obtenu la plus haute moyenne de rendement mensuel avec 1.19% et il s'agit de la seule stratégie proposant une moyenne supérieure à un pourcent. Les résultats pour chaque stratégie se retrouvent dans l'annexe 6. L'auteur a, ultérieurement, divisé ses données sur base des rendements en fonction que le marché était en hausse ou en baisse, et a testé les stratégies sur ces deux groupes de données. Pour les périodes haussières, le modèle Machine Learning a, de nouveau, sur-performé les autres stratégies avec une moyenne de 4.12%. Il est suivi de la stratégie basée sur l'analyse

fondamentale, puis de la stratégie Buy & Hold, avec des rendements avoisinants les 3%. L'analyse technique semble peu efficace dans un cas de marché haussier. Alors que dans le cas des périodes baissières, ces trois stratégies se retrouvent tout en bas du classement en proposant chacune un rendement moyen approximatif de -3%, alors que l'analyse technique propose les meilleures stratégies.

Une autre étude qui compare les performances d'algorithmes ML et d'indicateurs financiers est celle d'Yan, Sewell & Clack, de 2008. Ils se sont intéressés à l'optimisation de portefeuille, utilisant le Machine Learning, l'analyse technique et la programmation génétique (une sous-branche de l'intelligence artificielle) et ont porté leur étude sur le marché malaisien. Une comparaison a donc été effectuée entre les machines à vecteurs de support (SVM), le SGP (Standard Genetic Programming) et l'indicateur MACD où l'objectif était de visualiser quelle méthode offrait le meilleur retour sur investissement sur une période allant du 31 juillet 1997 au 31 décembre 1998. Dans ce cas-ci, le SGP sur-performe les deux autres stratégies, en proposant des résultats plus profitables (1%). Le MACD, bien qu'offrant un ROI négatif, sur-performe néanmoins le modèle Machine Learning SVM. Qu'il s'agisse du SVM ou de MACD, les stratégies présentent un Sharpe Ratio négatif, impliquant donc que les rendements sont trop faibles au regard du risque considéré.

Des études tentent de vérifier si des algorithmes Machine Learning peuvent "battre le marché", c'est-à-dire s'ils peuvent contredire l'hypothèse d'efficience des marchés, et donc sur-performer une stratégie de Buy & Hold, qui semblerait être la meilleure stratégie dans l'hypothèse où les prix suivent la théorie Random Walk. C'est notamment le cas de l'étude de Skabar & Cloete, datant de 2002, où un réseau de neurones est entraîné, à partir d'un algorithme génétique pour tenter de prédire les signaux d'achat et de vente. La fenêtre de temps couvre cinq ans, du 1^{er} juillet 1996 au 30 juin 2001. L'auteur a comparé les rendements de quatre séries de prix financiers avec les rendements obtenus sur des données de marché aléatoire dérivées de chacune de ces séries en utilisant une procédure bootstrapping. Les quatre indices considérés sont le Dow Jones Industrial Average, l'Australian All Ordinaries, le S&P500 et le NASDAQ. Ces ensembles bootstrap contiennent la même distribution de rendements journaliers que la série originale. Les résultats montrent que certaines séries temporelles financières ne sont pas entièrement aléatoires, et qu'une stratégie de trading basée seulement sur le prix historique peut obtenir des meilleurs rende-

ments qu'une stratégie Buy & Hold.

De nombreux modèles ML incluent des indicateurs techniques et fondamentaux dans leur analyse. L'étude de Emir, Dinçer & Timor, réalisée en 2012, mesure et compare la performance de modèles SVM et MLP (Multi Layer Perceptron) lorsque ceux-ci contiennent soit des indicateurs techniques, soit des variables fondamentales. L'étude a été menée sur l'ISE 30 Index (indice Turc) sur la période de temps allant de 2002 à 2010. 12 indicateurs techniques ont servi comme variables aux modèles, ainsi que 14 indicateurs fondamentaux. Les résultats obtenus ont mis en avant la performance du modèle SVM. En effet, qu'il s'agisse des variables techniques ou fondamentales, ce modèle performe mieux que le Multi Layer Perceptron. Les auteurs ont aussi comparé l'apport des variables, qu'elles soient fondamentales ou techniques. Dans ce cas-ci, il semblerait que les deux modèles avec analyse fondamentale offrent des résultats plus prometteurs que lorsque ces modèles incluent les outils techniques. Une double comparaison a donc été menée dans cette étude, au niveau des modèles et des variables à considérer, et mettent en avant l'algorithme SVM, et les indicateurs fondamentaux.

Il est également intéressant d'analyser des études comparant des indicateurs techniques et fondamentaux. Beyaz, Tekiner, Zeng & Keane (2018), comparent la performance de l'analyse fondamentale, technique et la combinaison des deux pour prévoir le prix futur des actions, appliqué à des modèles Machine Learning. Ils ont assemblé les données historiques de 140 entreprises du S&P 500. Il s'agit dans cette étude de modèles de régression et ils ont décidé d'utiliser le RMSE pour mesurer le degré d'erreur. Pour le modèle avec analyse technique, 10 indicateurs ont été calculés grâce au package "TTR" sur R, notamment le MACD, ROC ou encore le RSI. En ce qui concerne les indicateurs fondamentaux, les auteurs ont rassemblé des indicateurs mesurant la performance de l'entreprise, ceux sur la compétition, d'autres sur l'industrie et des indicateurs macro-économiques. Ces indicateurs ont après été modélisés dans deux algorithmes ML : un réseau de neurones (NN) et une régression à vecteurs de support (SVR), et ont été testés sur deux horizons de temps, un de 252 jours (une année boursière) et l'autre de 126 jours (6 mois boursier). Peu importe la période de temps considérée, l'analyse fondamentale sur-performe l'analyse technique car elle propose un RMSE plus petit. Cependant, la combinaison des deux semble encore réduire cette mesure du niveau d'erreur. Si nous comparons

maintenant la performance des deux algorithmes, peu importe la période ou quels indicateurs ont été utilisés, c'est le modèle SVR qui propose les meilleures prévisions.

Une étude similaire comparant la performance des deux méthodes d'analyses a été conduite sur le marché égyptien EGX. Wafi, Hassan & Mabrouk (2015) , ont conduit une étude empirique sur 37 entreprises non-financières, allant de 1998 à 2009, avec la période de prévision couvrant 2007 à 2009 pour ensuite comparer les prévisions obtenues avec les valeurs de marché réelles. Plusieurs outils sont utilisés pour évaluer la performance des prévisions des méthodes, à savoir le RMSE ou le MAE. Il s'agit dans ce cas-ci de l'analyse technique qui prédit le mieux le prix d'actions, car les deux mesures précédemment citées sont plus petites, et le biais et la variance sont également inférieures.

On constate donc que d'une étude à l'autre, les résultats obtenus divergent. La première étude soulignait que le Machine Learning permettait d'obtenir de meilleurs rendements que l'analyse technique et fondamentale, pour toute la période considérée et pour les périodes haussières. Tandis que la deuxième étude indique des faibles performances du ML avec un modèle SVM face à un indicateur technique MACD et à la programmation génétique. Deux autres études s'intéressent à la comparaison entre indicateurs techniques et fondamentaux. La première modélise les indicateurs dans des algorithmes SVR et NN et c'est les modèles d'analyse fondamentale qui sur-performent les modèles d'analyse technique. Alors que dans la dernière étude, c'est bel et bien l'analyse technique qui propose les meilleures prévisions car elle permet d'obtenir un niveau d'erreur moins important.

Finalement, nous constatons qu'il existe peu d'études comparant les performances de modèles Machine Learning face aux trois méthodes d'analyses, précédemment mentionnées (comme les deux premières études citées dans cette section). En effet, dans la grande majorité, les études utilisent des modèles ML qui, soit prennent comme variables des indicateurs techniques, soit des indicateurs macro-économiques ou comptables. L'analyse technique, fondamentale ou sentimentale sont donc généralement combinées à des algorithmes ML pour créer des modèles plus robustes, et ces différents outils sont donc rarement comparés. C'est pourquoi nous allons axer notre recherche sur la comparaison des résultats que nous offrent des algorithmes ML, face à des indicateurs d'analyse technique.

3. Etude empirique

3.1 L'objectif de ce projet

Comme mentionné dans la section précédente, peu d'études comparent les performances du ML face à l'analyse technique. C'est pourquoi, dans cette étude, nous allons nous intéresser à cette comparaison, et l'étudierons dans différents secteurs, à savoir la biotechnologie, l'immobilier et l'énergie. Respectivement, nous utiliserons différents trackers pour cette analyse à savoir le **SPDR S&P Biotech ETF** pour le secteur de la biotechnologie, aussi appelé **XBI**. Ce trackeur se compose de sociétés américaines, avec des capitalisations boursières qui sont majoritairement petites ou moyennes et contient 123 entreprises. Pour le secteur de l'immobilier, nous allons prendre le **SPDR Dow Jones REIT ETF**, ou **RWR** qui se compose de 117 sociétés américaines. Enfin, nous analyserons le **Energy Select Sector SPDR Fund**, nommé aussi **XLE** et qui contient 27 entreprises américaines dans le secteur de l'énergie. On peut retrouver leurs composantes en annexe 1 (SSGA, 2020 ; Barchart, 2020). L'objectif premier est de comparer la performance des algorithmes Machine Learning face aux méthodes traditionnelles financières d'analyse technique pour tenter de prouver que les algorithmes ML sur-performent les autres et proposent de meilleures stratégies. Pour y parvenir, nous allons dans cette étude considérer cinq algorithmes de Machine Learning (CART, C5.0, RF, ANN et SVM) et quatre méthodes d'analyse technique (SMA, MACD, ROC et RSI). Dans un premier temps, nous déterminerons les futurs signaux d'achat ou de vente, pour au final tenter de prédire un rendement global par stratégie. Toute cette analyse sera

exécutée sur R. Cet objectif nous amène à formuler la question suivante :

Question **: Comment performent les algorithmes de Machine Learning face aux méthodes d'analyse technique dans le cas des secteurs de la biotechnologie, de l'immobilier et de l'énergie ?**

Dans la section suivante, nous allons développer toute la méthodologie, basée sur une approche quantitative, mise en place pour tenter de répondre à cette question. Nous la développerons avec des illustrations du XBI, néanmoins le travail est identique pour les deux autres trackers. Cette méthodologie se divisera en 9 étapes :

- **3.2.1 : Importation de la base de données**, où nous verrons comment les données ont été importées de Yahoo Finance sur R
- **3.2.2 : Définition des variables indépendantes**, où nous détaillerons les trois variables choisies et leurs constructions
- **3.2.3 : Définition de la variable dépendante**, il s'agira ici de définir la variable Y, avec les possibles étiquettes de nos modèles (1 ou 0)
- **3.2.4 : Séparation du jeu de données**, étape très importante où l'on divise la base de données entre données d'entraînement et données test
- **3.2.5 : Implémentation des modèles Machine Learning**, où nous détaillerons l'implémentation de chacun de nos modèles sur R, avec le choix des paramètres
- **3.2.6 : Développement des outils d'analyse technique**, où nous développerons chacun des indicateurs qui serviront de comparaison aux modèles ML
- **3.2.7 : Prédiction et évaluation des modèles ML**, où nous verrons comment les prédictions sont faites, ainsi que les métriques utilisées pour évaluer les modèles
- **3.2.8 : Calcul des rendements et performances obtenus par stratégie**, où ici nous calculerons un rendement général par stratégie, et où nous allons détailler le Sharpe Ratio, outil utilisé pour évaluer les performances de chaque stratégie.
- **3.2.9 : Présentation des résultats**, où nous observerons les résultats obtenus.

3.2 La méthodologie

3.2.1 Importation de la base de données

Comme mentionnée ci-dessus, notre étude porte sur trois trackers propre à trois secteurs différents : le XBI, le RWR et le XLE. Les données historiques journalières disponibles des trackers ont dans un premier temps été récoltées sur Yahoo Finance. Le jeu de données s'étend sur sept ans, allant du 1^{er} janvier 2011 au 1^{er} janvier 2018 et contient six variables :

— **Open** : Le prix à l'ouverture d'une date précise

— **High** : Le prix le plus haut auquel il a été échangé durant une journée précise

— **Low** : Le prix le plus bas auquel il a été échangé durant une journée précise

— **Close** : Le prix à la fermeture d'une journée précise

— **Volume** : La quantité d'action échangée à une date précise

— **Adj.Close** : Le prix ajusté à la fermeture d'une date, en tenant compte de la distribution des dividendes.

Etant donné qu'il s'agit de données financières, il est important de convertir la base de données en série temporelle pour pouvoir par la suite en analyser les comportements. Nous allons pour cela utiliser la fonction "xts" sur R.

	Open	High	Low	Close	Adj.Close	Volume
2011-01-03	21.18000	21.42667	21.18000	21.30000	20.67854	197700
2011-01-04	21.31000	21.31000	21.05000	21.14000	20.52321	130200
2011-01-05	21.00000	21.38333	21.00000	21.36667	20.74327	210000
2011-01-06	21.41667	21.41667	21.23000	21.29333	20.67206	213600
2011-01-07	21.29667	21.29667	21.10333	21.24333	20.62353	51300

FIGURE 6 – Importation du jeu de données sur R

3.2.2 Définition des variables indépendantes

Comme dans un modèle de régression, nous devons définir les variables indépendantes sur lesquelles nos modèles de Machine Learning vont s'appuyer pour faire leurs prédictions. A l'aide des données déjà récoltées, nous avons construit trois variables : la première correspond à la différence entre le prix à l'ouverture et le prix à la clôture à la même période. La deuxième représente la différence entre le prix le plus haut et le prix le plus bas sur une même date. Enfin, la dernière variable

mesure la différence entre le volume échangé du lendemain et le volume échangé du jour même. Cette dernière variable ne peut être calculée pour la dernière date de notre échantillon. Cette ligne sera donc supprimée.

```
#adding the independent variables

n<-1761
i=1
Open.close<-numeric(1761)
for(i in 1:n){
Open.close[i]<- Open[i] - Close[i]
i<-i+1}
data<-cbind(data,Open.close)

j=1
High.low<-numeric(1761)
for(j in 1:1761){
  High.low[j]<- High[j]-Low[j]
  j<-j+1}
data<-cbind(data, High.low)

k=1
Diff.Volume<-numeric(1761)
for(k in 1:1761){
  Diff.Volume[k]<-Volume[k+1]-Volume[k]
  k<-k+1}
data<-cbind(data, Diff.Volume)
```

	Open	High	Low	Close	Adj.Close	Volume	Open.close	High.low	Diff.Volume
2011-01-03	21.18000	21.42667	21.18000	21.30000	20.67854	197700	-0.119999	0.246666	-67500
2011-01-04	21.31000	21.31000	21.05000	21.14000	20.52321	130200	0.170000	0.260000	79800
2011-01-05	21.00000	21.38333	21.00000	21.36667	20.74327	210000	-0.366667	0.383333	3600
2011-01-06	21.41667	21.41667	21.23000	21.29333	20.67206	213600	0.123333	0.186666	-162300
2011-01-07	21.29667	21.29667	21.10333	21.24333	20.62353	51300	0.053333	0.193334	224400

FIGURE 7 – Construction de nos variables indépendantes

3.2.3 Définition de la variable dépendante

La variable dépendante, que nous avons appelé Y dans ce projet, correspond à l'étiquette attachée à notre donnée, et à ce que le modèle va tenter de prédire sur base des variables indépendantes. Dans notre cas, nous avons utilisé des algorithmes de classification. Cette variable définit si le cours de bourse du jour suivant se clôturera à la hausse ou à la baisse et prend soit la valeur 1, soit la valeur 0. Une valeur de 1 signifie un signal d'achat à la période concernée, tandis qu'une valeur de 0 signifie qu'il faut la vendre. Pour se faire, nous avons calculé les rendements, en pourcent, en nous basant sur l'Adjusted closing price. Une fois nos rendements obtenus, nous avons pu créer la variable, égale à 1 si le rendement est positif et 0 si le rendement est négatif. La dernière ligne de la base de données a été supprimée, car il n'était pas possible de calculer le rendement à cette date, ni la différence de volume.

```
#Returns calculation

l=1
Returns<-numeric(1761)
for(l in 1:1761){
  Returns[l]<-100*(Adj.Close[l+1]-Adj.Close[l])/Adj.Close[l]
  l<-l+1
}
data<-cbind(data, Returns)
View(data)

#conversion in factor

data<-data[-1761,]
View(data)
Y<-ifelse(data$Returns<0,-1, 1)
data<-cbind(data,Y)
names(data)[11]<-"Y"

A.V<-as.factor(Y)
A.V<-ifelse(Y==-1, "Vendre", "Acheter")
data<-cbind.data.frame(data,A.V)
View(data)
names(data)[12]<-"A.V"
```

	Open	High	Low	Close	Adj.Close	Volume	Open.close	High.low	Diff.Volume	Returns	Y	A.V
2011-01-03	21.18000	21.42667	21.18000	21.30000	20.67854	197700	-0.119999	0.246666	-67500	-0.75116035	-1	Vendre
2011-01-04	21.31000	21.31000	21.05000	21.14000	20.52321	130200	0.170000	0.260000	79800	1.07222495	1	Acheter
2011-01-05	21.00000	21.38333	21.00000	21.36667	20.74327	210000	-0.366667	0.383333	3600	-0.34325837	-1	Vendre
2011-01-06	21.41667	21.41667	21.23000	21.29333	20.67206	213600	0.123333	0.186666	-162300	-0.23479029	-1	Vendre
2011-01-07	21.29667	21.29667	21.10333	21.24333	20.62353	51300	0.053333	0.193334	224400	-0.03138163	-1	Vendre

FIGURE 8 – Construction de notre variable dépendante Y

3.2.4 Séparation du jeu de données

La séparation de l'ensemble de données est une étape indispensable dans la création d'un modèle de Machine Learning. L'objectif est de diviser la base de données en deux parties : la première partie où l'algorithme de classification va s'entraîner et tenter de découvrir des liens entre nos trois variables explicatives et notre variable à expliquée, et la deuxième partie où l'algorithme va tenter de prédire la variable dépendante, à savoir l'étiquette, sur ces nouvelles données. Au plus l'algorithme a de données pour s'entraîner, au plus celui-ci proposera un modèle précis. Nous avons donc décidé d'utiliser 80% des données comme données d'entraînement, et 20% pour tester le modèle.

```
#partition into 80% and 20%
train<-data[1:1419,]
validate<-data[1420:1760,]
```

FIGURE 9 – Division de notre jeu de données

3.2.5 Implémentation des modèles Machine Learning

Dans cette sous-section sera détaillé chacun des modèles développés sur base des données financières du XBI. Pour rappel, le procédé est identique pour les deux autres secteurs. Comme mentionnés au-dessus, cinq algorithmes différents ont été testés. Il ne s'agit que d'algorithmes de classification et supervisé, c'est-à-dire que les différentes classes possibles que la variable Y peut prendre sont fournies à l'algorithme, à savoir qu'elle peut être égale à 1 ou 0.

3.2.5.1 CART

Le premier modèle développé est un algorithme d'arbre de décision, qui a la particularité d'être strictement binaire et qui présente deux branches pour chaque noeuds. Un arbre de décision commence toujours par une racine, prolongée de branches désignant les règles de décision. On débouche ensuite sur des noeuds internes, qui représentent des caractéristiques, pour arriver aux feuilles qui correspondent aux résultats. Il est facilement interprétable et offre une vue d'ensemble sous forme d'organigramme. CART s'inspire de l'indice de Gini (formule 18, p.24) qui est une mesure du degré d'inégalité de la répartition d'une variable. Un attribut avec un petit indice sera donc préférable. Pour son implémentation sur R, nous avons d'abord dû installer le package "party" qui est un package prévu pour les arbres de décision. Ce package nous a permis d'utiliser la fonction "ctree", pour obtenir notre modèle (figure 10). Le résultat obtenu est un arbre à deux couches seulement, qui ne tient compte que de la variable exprimant la différence de volume. Si cette différence est inférieure ou égale à 174.900, notre modèle prédit une probabilité supérieure pour un signal d'achat, et à contrario une probabilité plus grande pour un signal de vente si la différence est supérieure à 174.900.

3.2.5.2 C5.0

Dans le même principe que CART, C5.0 est un arbre de décision mais qui est construit non pas sur l'indice de Gini, mais sur le gain d'information (formule 19, p.25). Il divisera l'échantillon en se basant sur l'attribut qui maximise ce gain. Le package "C50" a été installé pour la création du modèle, et permet l'utilisation de la fonction "C5.0" qui, appliquée à nos données d'entraînement, permet d'obtenir notre modèle (figure 11). A nouveau, nous obtenons un arbre à deux couches, avec une racine portant sur la différence de volume, et deux feuilles, similairement inter-

prétable à l'arbre obtenu dans l'algorithme de CART. Ici, notre modèle indique une probabilité de signal d'achat supérieur pour une différence de volume inférieure ou égale à 57.900, et un signal de vente dans le cas inverse.

```
#decision tree model
library(party)
tree<-ctree(A.V~Open.close+High.low+Diff.Volume, data = train)
tree
plot(tree)
```

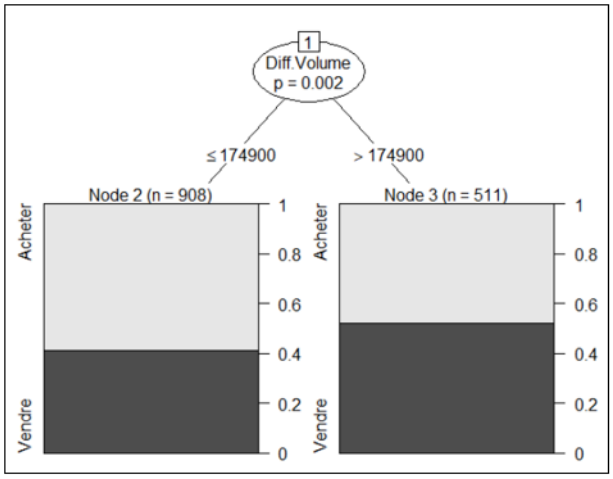


FIGURE 10 – Modèle CART

```
#C5.0 model

m2<- C5.0(A.V~Diff.Volume + Open.close + High.low, data = train)
summary(m2)
plot(m2)
```

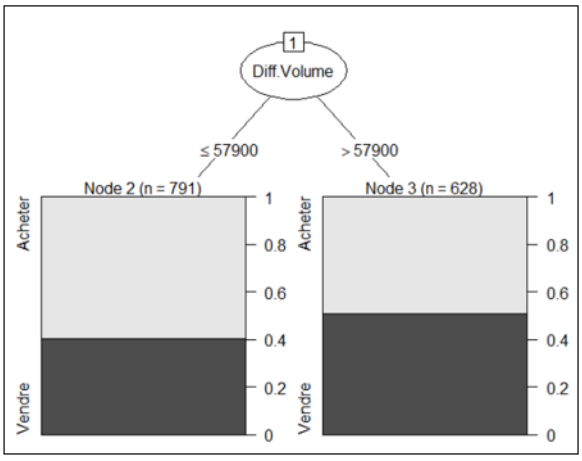


FIGURE 11 – Modèle C5.0

3.2.5.3 Random Forest

L'algorithme Random Forest désigne une méthode d'ensemble, qui va combiner différents "weak learners" (des arbres de décision) en un "strong learner". Il

s'agira donc d'une combinaison d'arbres de décision, qui va prendre la moyenne de toutes les prédictions. L'avantage de cet algorithme est qu'il permet d'éviter le sur-apprentissage, mais il est cependant plus lent à générer et plus difficilement interprétable qu'un seul arbre. Sur R, nous avons d'abord installé le package "randomForest", ainsi que le package "caret" qui nous sera utile dans l'étape de prédiction. Pour créer notre modèle, nous avons utilisé la fonction "randomForest" disponible dans le premier package. Un paramètre qu'il est important de déterminer est le nombre d'arbres du modèle. Pour se faire, nous avons représenté graphiquement les "Out-of bag errors" (OOB) du modèle en fonction du nombre d'arbres. Cette OOB est une mesure d'erreur spécifique aux méthodes d'ensemble bootstrap. Selon le graphique, un modèle formé d'approximativement 50 arbres semble minimiser l'erreur. Nous avons donc fait varier ce paramètre et nous sommes arrivés, en considérant 51 arbres, à obtenir la meilleure précision. Aussi, dans le cadre de ce projet, il s'agira d'un algorithme basé sur des arbres CART.

```
#Model Random Forest
library(randomForest)
set.seed(222)
rf<-randomForest(A.V~Diff.Volume + Open.close + High.low, data=train)


#Error rate of RF
plot(rf)

#avec 51 arbres sur les test data
set.seed(222)
rf1<-randomForest(A.V~Diff.Volume + Open.close + High.low, data=train,
                  ntree = 51,
                  mtry = 3,
                  importance = TRUE,
                  proximity= TRUE)
```

FIGURE 12 – Modèle Random Forest

3.2.5.4 Artificial Neural Network

Un réseau de neurones artificiels (ANN) est un algorithme d'apprentissage supervisé, inspiré du réseau de neurones biologique, impliquant des noeuds artificiels permettant le calcul de données en parallèle. Le réseau dans sa globalité se divise en trois partie : la couche d'entrée, les couches cachées pour enfin arriver sur la couche de sortie. De nombreux paramètres définissent un réseau de neurones. Dans ce projet, nous nous sommes contentés de faire varier le nombre de neurones par couche et le nombre de couches cachées présentes dans le modèle. Sur R, il nous fallait dans un premier temps normaliser les variables pour n'avoir que des valeurs entre 0 et 1. Ensuite, le package "neuralnet" nous a permis d'obtenir notre modèle (annexe 2). Ce package permet l'utilisation de la fonction "neuralnet", qui a été utilisée dans

cette étude et qui a pour fonction d'activation la fonction sigmoïde. Cependant, en faisant varier ces paramètres et lorsque le modèle était trop complexe (plusieurs couches cachées et beaucoup de neurones par couche), ce dernier n'arrivait pas à calculer les poids et le logiciel faisait apparaître un message d'erreur. Dans la limite du calculable, le modèle avec la meilleure précision contient qu'une couche cachée avec deux neurones sur celle-ci.

```
#MIN-MAX normalization

data$Diff.Volume<-(data$Diff.Volume - min(data$Diff.Volume))/(max(data$Diff.Volume)-min(data$Diff.Volume))
data$Open.close<-(data$Open.close - min(data$Open.close))/(max(data$Open.close)-min(data$Open.close))
data$High.low<-(data$High.low - min(data$High.low))/(max(data$High.low)-min(data$High.low))

ANN<-ifelse(data$Returns<0,0, 1)
data<-cbind(data,ANN)


#partition
training<-data[1:1419,]
testing<-data[1420:1760,]

#Neural Network
library(neuralnet)
set.seed(333)
n<-neuralnet(ANN~Diff.Volume+Open.close+High.low,
             data=training,
             hidden = 2,
             err.fct = "ce",
             linear.output = FALSE)
```

FIGURE 13 – Modèle Neural Network

3.2.5.5 Support Vector Machine

Les machines à vecteurs de support sont des algorithmes qui vont classer les éléments grâce à l'hyperplan de séparation optimale. Il s'agit d'un séparateur qui va maximiser la marge, c'est-à-dire la distance entre les classes. Cet algorithme est très utile dans les cas d'espaces à grande dimension. Pour pouvoir utiliser la fonction "svm" qui permet la création du modèle, il faut installer le package "e1071". Nous nous sommes intéressés à différentes fonctions que le noyau (kernel) peut prendre, ainsi qu'au paramètre "cost" qui peut être défini comme le poids qui pénalise la marge. Concernant le noyau, trois fonctions ont été testées : la fonction radiale, linéaire et polynomiale. Rapidement, un problème persistait avec la fonction linéaire et le modèle ne parvenait pas à prédire de signaux de vente. En comparant donc la fonction radiale et polynomiale, il s'avère que la fonction radiale proposait la meilleure précision. Par la suite, nous avons recherché le facteur "best.model" qui a retenu les paramètres qui rendent le modèle plus précis. Sur une gamme de valeurs proposées, un coût de 128 semblait optimiser le modèle.

```
########Tuning
library(e1071)
set.seed(1234)
tmodel<-tune(svm, A.V ~ Diff.Volume + Open.close + High.low, data = train,
     kernel = "radial",
     ranges = list(epsilon = seq(0,1,0.1), cost = 2^(2:9)))

#best model
mymodel2<-tmodel$best.model
summary(mymodel2)
```

FIGURE 14 – Modèle Support Vector Machine

3.2.6 Développement des outils d'analyse technique

Dans cette sous-section, nous détaillerons les indicateurs propres à l'analyse technique, utilisés pour déterminer les signaux d'achat et de vente. L'analyse technique est l'étude des tendances boursières, se basant sur les prix historiques et les volumes, qui permettent de déterminer les meilleurs moments pour vendre ou acheter une action. Nous avons appliqué quatre méthodes d'analyse différentes : Simple moving average (SMA), Moving Average Convergence Divergence (MACD), Rate of Change (ROC) et Relative Strength Index (RSI).

3.2.6.1 Simple Moving Average

Il s'agit ici d'un indicateur de moyenne mobile qui va mesurer la moyenne du prix de clôture (closing price) sur une période de temps fixée. Pour déterminer une tendance, nous considérons deux moyennes mobiles sur deux périodes différentes, une à court-terme et une à long-terme. Lorsque la moyenne de la période courte est supérieure à la moyenne de la période longue, une tendance haussière est attendue. On aura une tendance baissière dans le cas contraire. Pour pouvoir appliquer cette méthode, nous avons installé le package "quantmod". Celui-ci nous a permis d'utiliser la fonction "SMA". Dans le cadre de ce projet, la période de la moyenne court-terme sera de 5 jours et celle de la moyenne long-terme de 20 jours. Nous avons représenté graphiquement l'évolution du prix, ainsi que les moyennes mobiles avec la fonction "lineChart" (voir annexe 3) où nous retrouvons en rouge la moyenne sur 5 jours et en bleu la moyenne sur 20 jours. Comme expliqué dans la théorie, nous constatons en effet que ces moyennes assouplissent la courbe du prix. Nous avons un signal d'achat lorsque la ligne rouge est au-dessus de la ligne bleue. Mathématiquement, nous avons donc simplement calculé la différence entre les deux moyennes, qui nous a permis ainsi d'assigner une valeur de 1 ou 0 en fonction de la tendance.

```r
#packages
library(quantmod)

#Representation of SMA

sma<-cbind(xbi_yf$Close, SMA(xbi_yf$Close, 5), SMA(xbi_yf$Close, 20))
colnames(sma)<-c('Close', 'SMA_S', 'SMA_L')
View(sma)
lineChart(xbi_yf$Close, theme=chartTheme('white'))
addSMA(n=5)
addSMA(n=20, col='blue')

#calculation of difference

diff<-numeric(1760)
diff<-sma$SMA_S-sma$SMA_L
sma<-cbind(sma,diff)

names(sma)[4]<-"diff"
```

FIGURE 15 – Implémentation de nos moyennes mobiles simples

3.2.6.2 Moving Average Convergence Divergence

Il s'agit là aussi d'un indicateur de tendance. La ligne MACD s'obtient par la différence entre une moyenne mobile exponentielle courte (12 périodes) sur une moyenne mobile exponentielle longue (26 périodes). Il est intéressant de comparer cette ligne avec une moyenne exponentielle de neuf périodes de cette MACD, qui nous donnera une deuxième ligne appelée la ligne de signal. La différence entre ces deux lignes nous donnera un signal soit d'achat (si la différence est positive), soit de vente (si la différence est négative). Pour les calculs sur R, nous avons installé deux packages, "quantmod" et "TTR"(Technical Trading Rules). Ce deuxième nous a permis d'utiliser la fonction MACD, avec laquelle nous avons fixé les périodes de nos moyennes exponentielles. Ensuite, la fonction "barChart" du package "quantmod" nous a permis de représenter les deux lignes, la ligne de signal et la ligne MACD (annexe 4).

```r
####MACD
library(quantmod)
library(TTR)
macd<-MACD(xbi_yf$Close, nFast=12, nSlow=26, nSig=9)
barChart(xbi_yf$Close, theme=chartTheme("white"))
addMACD(fast=12, slow=26, signal=9)

#calculation of difference
diff<-macd$macd-macd$signal
Y<-ifelse(diff<0, 0,1)

macd<-cbind(macd,Y)
names(macd)[3]<-"Y"
test<-macd[1420:1760,]
test<-merge(test,testing$Returns, all=TRUE)
View(test)
```

FIGURE 16 – Implémentation de notre indicateur MACD

3.2.6.3 Relative Strength Index

Il s'agit là d'un indicateur technique oscillateur qui sert à mesurer l'importance de récents changements de prix pour en déterminer des conditions de survente ou de surachat. Le RSI se calcule généralement sur 14 jours et s'obtiendra en calculant le Relative Strength, divisant une EMA(14) haussière par une EMA(14) baissière, qu'il faut indexer après. Une valeur inférieure à 30 indique des conditions de survente et donc un signal d'achat. Une valeur supérieure à 70 désigne des conditions de surachat et donc un signal de vente. Sur R, nous avons utilisé le package "TTR", comme dans le modèle MACD, qui nous a permis d'utiliser la fonction "RSI" que nous avons défini sur une période de 14 jours.

```
#####RSI
library(TTR)
rsi<-RSI(xbi_yf$Close, n=14)
Y<-ifelse(rsi<30,1,0)
rsi<-cbind(rsi, Y)
names(rsi)[2]<-"Y"

test<-rsi[1420:1760,]
test<-cbind(test, testing$Returns)
```

FIGURE 17 – Implémentation de notre indicateur RSI

3.2.6.4 Rate of Change

Cet outil va mesurer la vitesse de variation de prix sur une période de temps. Il est exprimé en pourcentage et représente donc le dynamisme d'une variable. Dans le cadre de ce projet, nous avons considéré une période 20 jours, c'est-à-dire que l'outil va calculer le ratio entre la valeur actuelle et la valeur d'il y a 20 jours. Le ROC permet donc de distinguer des tendances. En effet, un ROC positif impliquerait donc qu'un titre sur-performe le marché dans le court terme et représente donc un signal d'achat. Dans le cas contraire, un titre avec un ROC négatif va potentiellement perdre en valeur et représente donc un signal de vente. Sur R, nous avons appliqué la fonction "ROC" du package "TTR".

```
#######ROC
library(TTR)
donnee<-xbi_yf

roc<-ROC(donnee$Close, n = 20)
View(roc)
y<-ifelse(roc<0, 0, 1)
roc<-cbind(roc,y)
names(roc)[2]<-"y"

test<-roc[1420:1760,]
testing<-XBI1[1420:1760,]
test<-merge(test, testing$Returns, all =TRUE)
View(test)
```

FIGURE 18 – Implémentation de notre indicateur ROC

3.2.7 Prédictions et évaluations des modèles ML

Cette étape est primordiale pour la suite de nos modèles Machine Learning. En effet, lorsque nous avons créé nos modèles, ceux-ci s'adaptaient aux données d'entraînement. Cependant, nous voulons obtenir nos signaux de vente et d'achat pour nos données test. Pour se faire, nous avons utilisé la fonction "predict", appliquée à nos modèles ML, qui va donner une étiquette pour les nouvelles données étudiées et tenter de prédire les signaux avec la meilleure précision. Dans certains cas, l'installation d'un package supplémentaire était nécessaire, notamment le package "caret" pour notre modèle Random Forest. Une fois les prédictions faites, nous avons représenté les résultats sous des matrices de confusion et calculons la précision (accuracy) de nos modèles pour voir s'ils ont été capables de correctement prédire les signaux.

```
#misclassification on test data

testPred<-predict(tree,newdata=validate)
tab<-table(testPred,validate$A.V)
print(tab)
sum(diag(tab))/sum(tab)
```

FIGURE 19 – Prédictions, matrice de confusion et précision pour CART

Notre matrice de confusion offre une visualisation générale de nos prédictions. Cette mesure nous informe du nombre d'éléments qui auront été correctement classés, ainsi que ceux pour lesquels l'algorithme aura attribué un mauvais signal. Prenons le cas de notre modèle CART, on retrouve sa matrice de confusion en annexe 5, ainsi que celles des autres modèles. Sur les 341 éléments de nos données test, nous constatons que le modèle a correctement prédit 110 signaux d'achat et 77 signaux de vente. Il a également prédit à tort 86 signaux d'achat qui sont en réalité des signaux de vente, et 68 signaux de vente qui sont en réalité des signaux d'achat.

Grâce à cette matrice, il nous était possible de calculer la précision de notre algorithme. Celle-ci représente le ratio entre les individus correctement classés et le nombre total d'individus. Dans le cas de notre modèle CART pour le secteur biotechnologique (XBI), la précision (accuracy) est donc : (110+77)/(110+77+68+86) = 0.5483871. On constate donc que la majorité des étiquettes prédites sont correctes et par conséquent, notre modèle nous permet de remettre en cause la théorie d'efficience faible des marchés qui dit que le cours historique est directement reflété dans le prix, car il semble possible de profiter d'opportunité d'arbitrage. Les résultats obtenus des indicateurs techniques sont également représentés sous la forme de matrice de confusion, et les taux de précision ont été calculés pour pouvoir correctement comparer les modèles ML à ces indicateurs (annexe 5).

3.2.8 Calcul des rendements et performances obtenus par stratégie

Après avoir obtenu les signaux de nos données test, il était possible de calculer le rendement global de chaque stratégie, couvrant donc la période du 23 août 2016 jusqu'au 1 janvier 2018 (341 périodes). Pour rappel, un signal d'achat (1) correspond donc à une hausse du prix pour le lendemain, et un signal de vente (0) à une tendance baissière pour le prix du lendemain. L'objectif est donc, à l'aide de ces signaux et des rendements calculés sur base de l'Adjusted closing price, de calculer ce rendement général. Pour se faire, nous avons multiplié les rendements de chaque période avec le signal. Le principe est que l'on profite de ce rendement lorsqu'on achète le titre. Ensuite, nous avons calculé les rendements cumulatifs de chaque stratégie. Le rendement cumulatif se calcule comme suit :

$$Rendement = (1 + R_1)(1 + R_2)...(1 + R_N) - 1 \tag{21}$$

où R_1 est le rendement de la première période, R_2 celui de la deuxième période ainsi de suite pour nos 341 (N) périodes.

Il est intéressant à présent de mesurer la performance de notre stratégie. Pour se faire, nous avons utilisé le Sharpe Ratio. Ce ratio, aussi appelé le ratio récompense/variabilité, est une mesure introduite en 1966 par William Sharpe qui permet la mesure de la performance d'un portefeuille en ajustant son risque. Un ratio élevé

```r
#return calculation

w=341
i=1
accumulated=integer(w)
for(i in 1:w){
  accumulated[i] <- vect[i]*validate$Returns[i]
  i<-i+1
}
accumulated<-accumulated/100

#total return
### multi-period returns
a=1
tot=1
for(a in 1:341){
  tot<-tot*(1+accumulated[a])
  a<-a+1
}
tot-1
```

FIGURE 20 – Calcul de notre rendement général

s'interprète comme un retour sur investissement élevé par rapport au risque pris, et donc qu'il s'agit d'un bon investissement. Il s'exprime mathématiquement comme suit (CorporateFinanceInstitute, n.d.) :

$$Sharpe\ Ratio = \frac{R_x - R_f}{\sigma_x} \tag{22}$$

où R_x représente le rendement du portefeuille, R_f le risk-free rate et σ_x l'écart-type du rendement du portefeuille, aussi appelé la volatilité. Le risk-free rate désigne le taux d'intérêt qu'un investisseur peut espérer gagner sur un investissement sans risque. Dans ce cas-ci, nous avons désigné la stratégie de Buy & Hold appliquée au S&P 500 comme étant l'investissement sans risque. Nous avons considéré le S&P 500 car l'indice couvre un nombre multiples de secteurs, à l'inverse du XBI, du RWR et du XLE qui ne proposent aucune diversification. En ce qui concerne la volatilité, elle se calcule sur base des rendements journaliers, en prenant donc une fenêtre de temps de 341 périodes. Un Sharpe Ratio positif implique que la stratégie considérée sur-performe la stratégie de Buy & Hold. Et un Sharpe Ratio négatif implique donc que les rendements sur investissement sont trop faibles pour le risque considéré.

```r
###Sharpe ratio
SR<-(mean(accumulated)-mean(SP$Returns))/sd(accumulated)
SR
```

FIGURE 21 – Calcul du Sharpe Ratio

3.2.9 Présentation des résultats

Dans cette section, nous verrons les résultats obtenus et tenterons de répondre à la question sur laquelle se base notre projet, à savoir *"Comment performent les algorithmes de Machine Learning face aux méthodes d'analyse technique dans le cas des secteurs de la biotechnologie, de l'immobilier et de l'énergie ?"*. La table 1 reprend l'ensemble des résultats obtenus pour nos neuf stratégies différentes appliquées au XBI (la biotechnologie). La table 2 reprend les résultats du RWR (l'immobilier) et la table 3 représente les résultats obtenus pour le XLE (l'énergie). Chacun des tableaux reprend :

— les taux de précision (accuracy) obtenus pour les données d'entraînement

— les taux de précision (accuracy) obtenus pour les données test

— les rendements cumulés (du 23 août 2016 au 1^{er} janvier 2018) par stratégie

— les Sharpe Ratio qui définissent l'attractivité de la stratégie.

Ce sont les trois dernières colonnes qui nous aident à répondre à la question posée.

3.2.9.1 Secteur biotechnologique

En ce qui concerne les prédictions sur les données test pour le secteur biotechnologique (table 1), c'est-à-dire le pourcentage des signaux d'achat et de vente correctement prédits sur notre période, nous constatons que le modèle Random Forest est le plus précis avec 57.48% de précision, suivi de l'arbre de décision C5.0 avec 55.42% et puis de l'arbre CART avec 54.83%. De manière général, nous constatons que les modèles Machine Learning sur-performent les modèles d'analyse technique car les cinq modèles avec les meilleures précisions sont les cinq modèles ML étudiés. Nous constatons également un taux de précision identique pour les indicateurs SMA et MACD. Ceci n'est qu'une pure coïncidence, car après avoir observé les matrices de confusion (annexe 5), nous constatons que ces deux matrices ne sont pas identiques. Finalement, chacun de nos modèles possède un taux de précision supérieur à 50%, à l'exception du modèle RSI, et ils permettent donc de remettre en cause la forme faible de l'efficience des marchés.

En ce qui concerne les rendements cumulés obtenus, c'est l'algorithme C5.0 qui se hisse en tête en proposant un rendement de 52.68%, suivi par CART avec 49.05% et le modèle Random Forest qui vient se glisser sur la troisième marche du podium

avec un rendement de 48.73%. On retrouve donc les trois mêmes algorithmes mais dans un ordre différent. Nous constatons également que certains résultats sont interpellants, surtout ceux du RSI et ROC. Pour le RSI, après analyse de la matrice de confusion, nous remarquons que le modèle définit un signal d'achat (1) seulement à quatre reprises. Il est donc normal que le rendement soit faible car pour les 337 autres périodes, un signal de 0 a été assigné. Pour le ROC, notre modèle prédit 216 signaux d'achat. Or, seulement 115 sont réellement des signaux d'achat. Pour les 101 autres périodes, notre modèle achète donc le titre alors que les rendements sont négatifs, ce qui explique donc ce faible rendement.

A partir des deux colonnes déjà discutées, nous pouvons déjà répondre en partie à la question du projet. Étant donné que les cinq algorithmes Machine Learning proposent des taux de précision et des rendements supérieurs à ceux des outils d'analyse technique, nous pouvons à priori affirmer que les modèles ML sur-performent les indicateurs techniques dans le cas d'une analyse sur le tracker XBI.

Pour aller au bout de cette analyse, nous allons aborder la méthode d'évaluation de ces modèles. Pour se faire, nous avons utilisé le Sharpe Ratio pour mesurer le niveau de rendement face au risque pris, en comparant nos stratégies ML et d'analyse technique à la stratégie de Buy & Hold du S&P 500. Pour nos modèles Machine Learning, nous constatons que nos Sharpe ratio sont tous positifs. Ceci s'explique majoritairement parce que les rendements cumulés de chaque stratégie ML sont supérieurs aux rendements cumulés de la startégie Buy & Hold du S&P 500. En ce qui concerne la stratégie Buy & Hold, celle-ci offre un rendement cumulé de 22.25%. Pour l'analyse technique, seulement l'outil SMA et MACD proposent des stratégies plus efficaces que la stratégie Buy & Hold, et donc le RSI et le ROC ne sont pas de bonnes stratégies d'investissement. Nous constatons aussi que le Sharpe Ratio est positif pour chaque modèle qui prédit avec plus de 50% les signaux, à l'exception de l'indicateur ROC.

Stratégie	Taux de Précision		Performance	Evaluation
	Entraînement	Test	Rendements	Sharpe Ratio
CART	56.37%	54.83%	49.05%	0.066
C5.0	55.74%	55.42%	52.68%	0.077
RF	100%	57.48%	48.73%	0.056
ANN	57.71%	54.25%	39.21%	0.039
SVM	63.28%	53.95%	32.31%	0.023
SMA	51.14%	52.78%	24.69%	0.009
MACD	51.80%	52.78%	31.14%	0.022
RSI	45.76%	48.38%	2.66%	-0.173
ROC	49.60%	51.90%	1.41%	-0.038

TABLE 1 – Tableau récapitulatif des résultats obtenus pour le XBI

3.2.9.2 Secteur de l'immobilier

Nous allons maintenant nous intéresser au secteur de l'immobilier en utilisant le tracker RWR. Comme dit précédemment, les résultats sont affichés dans la table 2. Au niveau de la précision sur les données test, on constate que les modèles Machine Learning proposent tous un taux supérieur à 50%, alors qu'aucun indicateur technique n'arrive à dépasser ce seuil. On retrouve ici en tête de classement le réseau de neurones, suivi des deux arbres de décision CART et C5.0. Ces deux derniers proposent les mêmes taux, et ont la même matrice de confusion car les deux modèles ont généré les mêmes arbres. Ceux-ci ne se composent que d'une feuille, où la probabilité d'un signal d'achat est supérieure à celle d'un signal de vente. C'est pourquoi les modèles ne prédisent que des signaux d'achat, ce qui se remarque dans les matrices de confusion.

Au niveau des rendements cumulés, une fois de plus les algorithmes ML semblent mieux performer, ce qui est logique car le taux de précision et la performance sont étroitement liés. Cependant, si nous comparons les résultats à ceux du secteur précédent, nous constatons une énorme différence. En effet, seulement trois stratégies offrent des rendements cumulés positifs pour l'immobilier, à savoir le réseau de neurones, le RSI et les machines à vecteurs de support. Tandis que chaque stratégie dans le secteur biotechnologique propose des rendements positifs. Il est dès lors intéressant de chercher à comprendre l'origine de cette différence. Celle-ci sera expliquée par après, dans la comparaison intersectorielle. On constate également un taux de -10% pour les moyennes mobiles simples. Celui-ci s'explique par le faible taux de précision des prédictions sur les données test. Si nous regardons maintenant les Sharpe

Ratio, ceux-ci sont tous négatifs. Ce phénomène s'explique car aucune des stratégies n'arrive à égaler le benchmark, à savoir la stratégie Buy & Hold du S&P 500. On peut donc affirmer qu'investir dans ce tracker, pendant la période teste, était une mauvaise décision car les rendements sont trop faibles par rapport au risque considéré. Cependant, les ratios des algorithmes ML restent supérieurs que ceux de la finance technique.

Stratégie	Taux de Précision		Performance	Evaluation
	Entraînement	Test	Rendements	Sharpe Ratio
CART	54.61%	51.90%	-1.13%	-0.074
C5.0	54.61%	51.90%	-1.13%	-0.074
RF	100%	51.61%	-2.92%	-0.106
ANN	56.02%	53.95%	3.89%	-0.061
SVM	58.70%	51.61%	1.01%	-0.074
SMA	52.07%	45.45%	-10.08%	-0.173
MACD	50.57%	49.85%	-6.11%	-0.140
RSI	45.40%	49.85%	3.57%	-0.466
ROC	50.89%	44.86%	-8.41%	-0.171

TABLE 2 – Tableau récapitulatif des résultats obtenus pour le RWR

3.2.9.3 Secteur de l'énergie

Comme pour les industries précédentes, intéressons-nous premièrement aux taux de précision obtenus pour les données test. Les résultats sont affichés dans la table 3. La première constatation flagrante est qu'une seule stratégie arrive à prédire avec plus de 50% de précision les signaux d'achat et de vente, il s'agit de l'indicateur RSI. Si nous regardons sa matrice de confusion en annexe 5, nous constatons que l'outil ne prédit que quatre signaux d'achat (deux seulement sont exactement prédits). Les 337 autres signaux prédits sont donc des signaux de vente. Or, la majorité des signaux sur la période considérée sont actuellement des signaux de vente, à savoir 174. C'est pour cette raison que le taux de précision est supérieur à 50%. Nous constatons aussi à nouveau des taux similaires pour CART et C5.0 car une fois de plus, les deux modèles ont généré exactement les mêmes arbres. Il n'est pas évident de conclure dans ce cas si le ML performent mieux que l'analyse technique. Car même si les taux proposés par les algorithmes sont constants et avoisinent les 49%, c'est bel et bien l'indicateur RSI qui prédit au mieux les signaux.

Au niveau des rendements, c'est l'indicateur ROC qui donne le meilleur rendement cumulé, qui a atteint les 8.87%, suivi des deux arbres de décision avec 6.1%. La majorité des stratégies donnent des rendements positifs, 4 sur 5 pour le ML et 2 sur 4 pour l'analyse technique. On reste cependant très loin des résultats obtenus pour l'industrie biotechnologique. Au niveau des Sharpe Ratio, ceux-ci sont comme pour le secteur de l'immobilier, tous négatifs, certains se rapprochant fortement de -1. C'est le cas du RSI, qui pourtant avait le meilleur taux de prédiction. Ceci peut s'expliquer par une faible volatilité, étant donné que la majorité des signaux prédits sont des signaux de vente (valeur de 0). Aucune des stratégies ne propose donc de bonnes solutions d'investissement, au vu du risque considéré.

Stratégie	Taux de Précision		Performance	Evaluation
	Entraînement	Test	Rendements	Sharpe Ratio
CART	56.02%	49.56%	6.10%	-0.069
C5.0	56.02%	49.56%	6.10%	-0.069
RF	100%	49.85%	2.46%	-0.082
ANN	56.09%	49.85%	3.02%	-0.076
SVM	64.48%	49.26%	-2.27%	-0.098
SMA	48.14%	46.92%	-2.99%	-0.100
MACD	50.64%	48.68%	0.23%	-0.078
RSI	48.61%	51.02%	-0.77%	-0.669
ROC	49.89%	49.26%	8.87%	-0.050

TABLE 3 – Tableau récapitulatif des résultats obtenus pour le XLE

3.2.9.4 Comparaison intersectorielle

Afin de pouvoir répondre à la question, il est intéressant de faire une comparaison intersectorielle des résultats obtenus. Concernant les taux de précisions sur les données test, nous constatons directement que les modèles Machine Learning proposent de meilleures précisions que l'analyse technique pour le secteur de la biotechnologie et de l'immobilier. Cependant pour le secteur de l'énergie, la réponse semble moins évidente car même si les taux sont généralement plus élevés que ceux proposés par les indicateurs, c'est bel et bien l'outil RSI qui offre la meilleure précision. Toujours concernant la précision, le Machine Learning et l'analyse technique sont beaucoup plus précis dans leurs prédictions pour le secteur biotechnologique que pour les deux autres secteurs. Alors que huit stratégies sur neuf ont une précision supérieure à 50% pour le 1^{er} secteur, seulement une stratégie dépasse ce taux pour le 3^{ieme} secteur, et cinq pour le 2^{ieme}. Il semblerait donc que le secteur ait une influence sur ces

données. Si nous regardons le cours de bourse de ces trois trackers, on remarque que seulement le tracker XBI (biotech) connaît une vraie croissance boursière sur la période teste, équivalente à 36%. Alors que le XLE affiche une croissance de seulement 8% et le RWR une décroissance de 1.06%. Il semblerait donc que les algorithmes Machine Learning prédisent mieux les signaux dans le cas d'un cours boursier en hausse. Cette analyse rejoint donc l'étude de Macchiarulo (2018), qui observe que le ML sur-performe les autres stratégies en situation de périodes haussières.

Regardons maintenant les rendements obtenus et les Sharpe Ratio. Il est évident que le ML sur-performe les autres méthodes dans le cas de l'industrie biotechnologique et de l'immobilier. Dans ces deux secteurs, les Sharpe Ratio des outils techniques sont tous inférieurs à ceux des algorithmes ML. Et les rendements cumulés de ces outils sont majoritairement plus petits. Cependant la réponse pour le domaine de l'énergie, c'est-à-dire un secteur en faible croissance sur la période teste, reste toujours ambiguë et on ne peut pas affirmer avec certitude que le ML bat les outils techniques.

Nous pouvons donc en conclure que les algorithmes ML permettent d'obtenir une meilleure précision, et offrent de meilleures stratégies, face à l'analyse technique dans le cas du secteur biotechnologique et de l'immobilier. Mais que cette réponse ne peut s'étendre au domaine de l'énergie pour les raisons précédemment citées.

3.2.10 Limites

Tout au long de notre analyse, des choix ont été faits et qui ont pu potentiellement impacter les résultats de notre étude. Tout d'abord, le choix des variables dépendantes et indépendantes, la division en pourcentage du jeu de données en données test et d'entraînement, la période considérée, ... sont toutes des variables qui lorsqu'on les fait varier, peuvent modifier ou non les résultats obtenus. Pour obtenir des modèles plus précis, il serait sans doute judicieux de considérer une plus grande base de données, avec de nombreux arguments pour déterminer les signaux.

Concernant les modèles, le choix des paramètres peut également modifier leur performance et on était limité dans le choix de ceux-ci. On aurait pu obtenir des modèles avec une meilleure précision et qui offraient des rendements plus importants s'il

était possible de modifier chacun des paramètres. Pour éviter le sur-apprentissage, on aurait pu aussi appliquer la validation croisée (cross-validation), qui consiste à appliquer les liens faits lors de la phase d'apprentissage à différents ensembles de validation pour avoir des estimations plus robustes.

De plus, l'objectif premier consiste à comparer les performances de cinq modèles Machine Learning à ceux de quatre méthodes d'analyse technique pour voir si le ML sur-performe les indicateurs techniques. Or, il existe une abondance d'algorithmes ML, ainsi que de nombreux outils de finance technique. Pour pouvoir répondre avec certitude à la question de recherche, il faudrait donc considérer une plus vaste gamme de modèles. De plus, cette étude se limite à la comparaison entre ML et analyse technique. Or, il aurait été intéressant de considérer des modèles d'analyse fondamentale, mais aussi d'analyse de sentiment, ou encore la combinaison de différents modèles.

Toutes les limites citées ci-dessus peuvent être prises en compte dans une étude future pour tenter de mieux répondre à la question posée.

4. Conclusion

Ce mémoire a pour but d'analyser l'utilité du Machine Learning dans la gestion de portefeuilles, et de mesurer la performance des algorithmes face à des méthodes financières plus conventionnelles.

Selon la littérature, le Machine Learning pourrait concurrencer, voir battre les autres méthodes d'analyse, notamment si on regarde le modèle combiné SVM-ANN de l'étude de Macchiarulo (2018), qui sur-performe l'analyse technique, l'analyse fondamentale et la stratégie de Buy & Hold sur toute la période considérée, et sur les périodes haussières. L'utilisation du modèle SVM combiné à des indicateurs techniques ou fondamentaux, qui permet d'obtenir des taux de précision supérieurs à 90% sur certaines périodes (Emir et al., 2012), montre également que le Machine Learning a sa place dans le domaine de la finance de marché et de la gestion de portefeuilles. L'étude de Skabar & Cloete (2002), partage aussi ce sentiment où un modèle ANN, basé uniquement sur les prix historiques, permet de remettre en question la théorie d'efficience des marchés et de proposer des meilleurs rendements qu'une stratégie Buy & Hold.

Afin de mener cette analyse pour notre partie empirique, nous nous sommes intéressés à trois secteurs distinctifs : le tracker XBI (la biotechnologie), le tracker RWR (l'immobilier) et le tracker XLE (l'énergie). La fenêtre de temps considérée allait du 1^{er} janvier 2011 au 1^{er} janvier 2018. Nous avons défini cinq modèles Machine Learning : deux arbres de décisions (CART et C5.0), un modèle de forêt aléatoire (RF),

un réseau de neurones (ANN) et une machine à vecteurs de support (SVM). Nous avons décidé de comparer ces algorithmes avec des outils d'analyse technique. Pour se faire, nous avons considéré quatre indicateurs : Simple Moving Average (SMA), Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI) et le Rate of Change (ROC), ce qui nous donne un total de neuf stratégies. Chaque stratégie avait pour objectif, à l'aide de trois variables indépendantes, de prédire le cours du lendemain (Y), prenant la valeur de 1 si celui-ci augmentait, ou 0 si celui-ci baissait, et de mesurer le rendement cumulé avec ces décisions d'investissements. Nous avons observé que dans le cas des industries de la biotechnologie et de l'immobilier, le Machine Learning battait l'analyse technique, tant au niveau des prédictions que des rendements obtenus. Cependant pour le domaine de l'énergie, il est difficile d'affirmer sur base des résultats, que les modèles ML battent les outils techniques.

Il est possible d'aller plus loin dans l'étude des performances d'algorithmes ML appliquées à la prédiction boursière. Notre étude ne porte que sur la comparaison entre Machine Learning et analyse technique. Des recommandations pour de prochaines études seraient de s'intéresser également à l'analyse fondamentale, l'analyse de sentiment et des combinaisons de différentes méthodes. Il serait également intéressant d'élargir le champ des secteurs étudiés, pour voir si le secteur a réellement un impact sur l'efficacité des modèles. Concernant la robustesse des modèles, une analyse couvrant une plus large période, avec une augmentation du nombre de variables en entrée, offrirait de meilleurs modèles et permettrait d'obtenir de meilleurs résultats. Une dernière indication serait d'élargir le nombre de modèles étudiés. En effet, il existe une abondance d'algorithmes qui n'ont pas été vus au cours de ce travail et dont il serait intéressant d'en examiner leur efficacité.

L'objectif de ce travail est de s'intéresser uniquement à l'aspect lucratif de la finance. En effet, la performance des modèles et les rendements cumulés du ML sont au coeur de ce mémoire, tandis que son impact sur l'aspect social et éthique est purement négligé. Il peut donc être intéressant d'étudier ces potentielles conséquences dans de futures études similaires, car cette science révolutionne le rôle des différents acteurs financiers et est susceptible de modifier leur processus décisionnel. La place de l'homme, à côté de ces machines, est incertaine. Plus que cela, l'utilisation du Machine Learning remet en cause les théories financières, comme observé dans la revue

de littérature et dans la partie empirique avec la théorie d'efficience des marchés. Nous pouvons par contre être certain que le Machine Learning n'a pas fini de faire parler de lui, et qu'il n'en est pas au stade final de son exploitation et utilisation.

5. Bibliographie

Aaron, C., Bilon, I., Galanti, S., & Tadjeddine, Y. (2006). Les styles de gestion de portefeuille existent-ils ?. *Problèmes économiques, 2905*, 40.

Aftalion, F. (2005). Le MEDAF et la finance comportementale. *Revue française de gestion,* (4), 203-214.

Akar, Ö., & Güngör, O. (2012). Classification of multispectral images using Random Forest algorithm. *Journal of Geodesy and Geoinformation, 1*(2), 105-112.

Barchart. (2020). Consulté sur https ://www.barchart.com/

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks : fundamentals, computing, design, and application. *Journal of microbiological methods, 43*(1), 3-31.

Beyaz, E., Tekiner, F., Zeng, X. J., & Keane, J. (2018, June). Comparing Technical and Fundamental indicators in stock price forecasting. In *2018 IEEE 20th International Conference on High Performance Computing and Communications ; IEEE 16th International Conference on Smart City ; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity /DSS)*(pp. 1607-1613). IEEE.

Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., & Caicedo, O. M. (2018). A comprehensive survey on machine learning for networking : evolution, applications and research opportunities. *Journal of Internet Services and Applications, 9*(1), 16.

Cairn. (2007). *Prévisibilité des rentabilités boursières*. Consulté sur https ://www.cairn.info/revue-economie-et-prevision-2004-5-page-71.htm

Calas, G. (2009). Etudes des principaux algorithmes de data mining. *Spécialisation Sciences Cognitives et Informatique Avancée, France*.

CorporateFinanceInstitute. (n.d). *Sharpe Ratio*. Consulté sur https ://corporatefinanceinstitute.com/resources/knowledge/finance/sharpe-ratio-definition-formula/ ?fbclid=IwAR2TeTcfksT10wNDb3PmNcainXPL7xHZQS2DJT3MSwYsFLhMbaL6AMViX5w

Dase, R. K., Pawar, D. D., & Daspute, D. S. (2011). Method-ologies for Prediction of Stock Market : an Artificial Neural Network. *International Journal of Statistika and Mathematika, 1*(1, 2011), 08-15.

DataAnalyticsPost. (n.d). *SVM*. Consulté sur https ://dataanalyticspost.com/Lexique/svm/

De Prado, M. L. (2018). *Advances in financial machine learning*. John Wiley & Sons.

Dev. (2019). *Classification and Regression Analysis with Decision Trees*. Consulté sur https ://dev.to/nexttech/classification-and-regression-analysis-with-decision-trees-jgp

Duan, Y., Hu, G., & McLean, R. D. (2009). When Is Stock Picking Likely to Be Successful ? Evidence from Mutual Funds. *Financial Analysts Journal, 65*(2), 55–66.

Emir, S., Dinçer, H., & Timor, M. (2012). A stock selection model based on fundamental and technical analysis variables by using artificial neural networks

and support vector machines. *Review of Economics Finance, 2*, 106-122.

Fakir M. & Ezzikouri H. (n.d.). Algorithmes de classification : ID3 & C4.5. *Département informatique, Maroc.*

Fleischer, D. (2019). *Automated Trading System using Machine Learning*(thèse de doctorat, Univeristé de Stavanger, Norvège).

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

Gorgulho, A., Neves, R., & Horta, N. (2011). Applying a GA kernel on optimizing technical analysis rules for stock picking and portfolio composition. *Expert systems with Applications, 38*(11), 14072-14085.

Gorgulho, A., Neves, R., & Horta, N. (2009). Using GAs to balance technical indicators on stock picking for financial portfolio composition. In *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference : Late Breaking Papers* (pp. 2041-2046).

Gray, W., & Kern, A. (2009). Do Hedge Fund Managers Have Stock Picking Skills ?. *Social Science Research Network. http ://papers. ssrn. com/abstract, 1477586*.

Gunn, S. R. (1998). Support vector machines for classification and regression. *ISIS technical report, 14*(1), 5-16.

Hari, Y., & Dewi, L. P. (2018). Forecasting System Approach for Stock Trading with Relative Strength Index and Moving Average Indicator. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10*(2-3), 25-29.

Hu, Y., Liu, K., Zhang, X., Su, L., Ngai, E. W. T., & Liu, M. (2015). Application of evolutionary computation for rule discovery in stock algorithmic trading : A literature review. *Applied Soft Computing, 36*, 534-551.

Ibbotson, R. G., & Riepe, M. W. (1997). Growth vs. value investing : And the winner is... *Journal of Financial Planning, 10*(3), 64.

Insee. (n.d.). *Indice de Gini / Coefficient de Gini*. Consulté sur https ://www.insee.fr/fr/metadonnees/definition/c1551.

Jabbar, H., & Khan, R. Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks : A tutorial. *Computer, 29*(3), 31-44.

Kalyani, J., Bharathi, P., & Jyothi, P. (2016). Stock trend prediction using news sentiment analysis. *arXiv preprint arXiv :1607.01958*.

Lauren, S., & Harlili, S. D. (2014, August). Stock trend prediction using simple moving average supported by news classification. In *2014 International Conference of Advanced Informatics : Concept, Theory and Application (ICAICTA)* (pp. 135-139). IEEE.

Lebigdata. (2018). *Machine Learning et Big Data : définition et explications*. Consulté sur https ://www.lebigdata.fr/machine-learning-et-big-data

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*(1), 1-167.

Macchiarulo, A. (2018). Predicting and beating the stock market with machine learning and technical analysis. *Journal of Internet Banking and Commerce, 23*(1), 1-22.

Maclin, R., & Opitz, D. (1997). An empirical evaluation of bagging and boosting. *AAAI/IAAI, 1997*, 546-551.

Malkiel, B. G., & Fama, E. F. (1970). Efficient capital markets : A review of theory and empirical work. *The journal of Finance, 25*(2), 383-417.

Medium. (2019). *Looking at R-squared*. Consulté sur https ://medium.com/@erika.dauria/looking-at-r-squared-721252709098

Miles, J. (2014). R squared, adjusted R squared. *Wiley StatsRef : Statistics Reference Online.*

Openclassrooms. (2020). *Évaluez et améliorez les performances d'un modèle de machine learning*. Consulté sur https ://openclassrooms.com/fr/courses/4297211-evaluez-et-ameliorez-les-performances-dun-modele-de-machine-learning/4308256-evaluez-un-algorithme-de-classification-qui-retourne-des-valeurs-binaires

Patil, N., Lathi, R., & Chitre, V. (2012). Comparison of C5. 0 & CART classification algorithms using pruning technique. *Int. J. Eng. Res. Technol, 1*(4), 1-5.

Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine learning for stock selection. *Financial Analysts Journal*, 75(3), 70-88.

Setty, D. V., Rangaswamy, T. M., & Subramanya, K. N. (2010). A review on data mining applications to the performance of stock marketing. *International Journal of Computer Applications, 1*(3), 33-43.

Shah, V. H. (2007). Machine learning techniques for stock prediction. *Foundations of Machine Learning/ Spring, 1*(1), 6-12.

Sitte, R., & Sitte, J. (2002). Neural networks approach to the random walk dilemma of financial time series. *Applied Intelligence, 16*(3), 163-171.

Skabar, A., & Cloete, I. (2002, January). Neural networks, financial trading and the efficient markets hypothesis. In *ACSC* (Vol. 2, pp. 241-249).

SSGA. (2020). *SPDR S&P Biotech ETF*. Consulté sur

https ://www.ssga.com/us/en/individual/etfs/funds/spdr-sp-biotech-etf-xbi

Twinword. (n.d.). *Sentiment Analysis For Finance*. Consulté sur https ://www.twinword.com/blog/sentiment-analysis-for-the-financial-sector/

Wafi, A. S., Hassan, H., & Mabrouk, A. (2015). Fundamental analysis vs. technical analysis in the Egyptian stock exchange—Empirical study. *International Journal of Business and Management Study–IJBMS, 2*(2), 212-218.

Wang, H., & Raj, B. (2015). A survey : Time travel in deep learning space : An introduction to deep learning models and how deep learning models evolved from the initial ideas. *arXiv preprint arXiv :1510.04781.*

Wikistat. (n.d). *Réseaux de neurones*. Consulté sur http ://wikistat.fr/

Yahoo Finance. (2020). Consulté sur https ://finance.yahoo.com/

Yan, W., Sewell, M. V., & Clack, C. D. (2008, July). Learning to optimize profits beats predicting returns- comparing techniques for financial portfolio optimisation. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation* (pp. 1681-1688).

6. Annexes

Annexe 1 : Composantes des trackers

XBI

Name	Ticker	Identifier	SEDOL	Weight	Sector	Shares Held	Local Currency
Moderna Inc.	MRNA	60770K10	BGSXTS3	4.281735	Biotechnology	2980089.000	USD
Regeneron Pharmaceuticals Inc.	REGN	75886F10	2730190	2.522653	Biotechnology	190905.000	USD
Immunomedics Inc.	IMMU	45290710	2457961	2.308768	Biotechnology	3104353.000	USD
BioMarin Pharmaceutical Inc.	BMRN	09061G10	2437071	2.257253	Biotechnology	896421.000	USD
Seattle Genetics Inc.	SGEN	81257810	2738127	2.233645	Biotechnology	631117.000	USD
United Therapeutics Corporation	UTHR	91307C10	2430412	2.051333	Biotechnology	801824.000	USD
Vertex Pharmaceuticals Incorporated	VRTX	92532F10	2931034	2.027365	Biotechnology	328262.000	USD
Exelixis Inc.	EXEL	30161Q10	2576941	2.004826	Biotechnology	4033316.000	USD
Sarepta Therapeutics Inc.	SRPT	80360710	B8DPDT7	1.989392	Biotechnology	544310.000	USD
Alnylam Pharmaceuticals Inc	ALNY	02043Q10	B00FWN1	1.917709	Biotechnology	610539.000	USD
Gilead Sciences Inc.	GILD	37555810	2369174	1.789994	Biotechnology	1092883.000	USD
ACADIA Pharmaceuticals Inc.	ACAD	00422510	2713317	1.768902	Biotechnology	1621322.000	USD
Biohaven Pharmaceutical Holding Company Ltd.	BHVN	G1119610	BZ8FXC4	1.746521	Biotechnology	1063611.000	USD
AbbVie Inc.	ABBV	00287Y10	B92SR70	1.742152	Biotechnology	818397.000	USD
Neurocrine Biosciences Inc.	NBIX	64125C10	2623911	1.740551	Biotechnology	650103.000	USD
Amicus Therapeutics Inc.	FOLD	03152W10	B19FQ48	1.722930	Biotechnology	5559274.000	USD
Incyte Corporation	INCY	45337C10	2471950	1.692010	Biotechnology	785617.000	USD
Alexion Pharmaceuticals Inc.	ALXN	01535110	2036070	1.675085	Biotechnology	650985.000	USD
Ultragenyx Pharmaceutical Inc.	RARE	90400D10	BJ62Z18	1.672509	Biotechnology	1043131.000	USD
Ligand Pharmaceuticals Incorporated	LGND	53220K50	2501578	1.638726	Biotechnology	677882.000	USD
Invitae Corp.	NVTA	46185L10	BVVCNT1	1.630726	Biotechnology	4167176.000	USD
Exact Sciences Corporation	EXAS	30063P10	2719951	1.569904	Biotechnology	824345.000	USD
Amgen Inc.	AMGN	03116210	2023607	1.546976	Biotechnology	303627.000	USD
Mirati Therapeutics Inc.	MRTX	60468T10	BBPK0J0	1.542542	Biotechnology	640556.000	USD
Ionis Pharmaceuticals Inc.	IONS	46222210	BDJ0LS6	1.468128	Biotechnology	1140425.000	USD
Blueprint Medicines Corp.	BPMC	09627Y10	BWY52P3	1.408725	Biotechnology	834404.000	USD
Biogen Inc.	BIIB	09062X10	2455965	1.394517	Biotechnology	242180.000	USD
Global Blood Therapeutics Inc	GBT	37890U10	BZ05388	1.327576	Biotechnology	935370.000	USD
Iovance Biotherapeutics Inc	IOVA	46226010	BF0DMK7	1.250503	Biotechnology	1887770.000	USD
Natera Inc.	NTRA	63230710	BYQRG48	1.222101	Biotechnology	1201247.000	USD
PTC Therapeutics Inc.	PTCT	69366J20	B17VCN9	1.218293	Biotechnology	1093590.000	USD

Agios Pharmaceuticals Inc.	AGIO	00847X10	BCBVTX1	1.157243	Biotechnology	1089863.000	USD
CareDx Inc.	CDNA	14167L10	BP3YM74	1.143248	Biotechnology	1648509.000	USD
bluebird bio Inc.	BLUE	09609G10	BBFL7S1	1.136104	Biotechnology	801339.000	USD
Emergent BioSolutions Inc.	EBS	29089Q10	B1HJLW5	1.125089	Biotechnology	726051.000	USD
Arena Pharmaceuticals Inc.	ARNA	04004760	BF3N4P3	1.116819	Biotechnology	787737.000	USD
Acceleron Pharma Inc	XLRN	00434H10	BDGTXQ8	1.116071	Biotechnology	507721.000	USD
Adverum Biotechnologies Inc	ADVM	00773U10	BD6NXD7	1.058115	Biotechnology	1967598.000	USD
Alkermes Plc	ALKS	G0176710	B3P6D26	1.036919	Biotechnology	2640796.000	USD
Intercept Pharmaceuticals Inc.	ICPT	45845P10	B7N59F8	1.021312	Biotechnology	602566.000	USD
Insmed Incorporated	INSM	45766930	2614487	0.992514	Biotechnology	1578705.000	USD
Apellis Pharmaceuticals Inc.	APLS	03753U10	BYTQ6X1	0.903966	Biotechnology	1288678.000	USD
Arrowhead Pharmaceuticals Inc.	ARWR	04280A10	BYQBFJ8	0.879686	Biotechnology	1082963.000	USD
SAGE Therapeutics Inc.	SAGE	78667J10	BP4GNK9	0.876931	Biotechnology	987836.000	USD
FibroGen Inc.	FGEN	31572Q80	BSDRYR8	0.875501	Biotechnology	986225.000	USD
Allogene Therapeutics Inc.	ALLO	01977010	BFZNYB7	0.824852	Biotechnology	892048.000	USD
Coherus BioSciences Inc.	CHRS	19249H10	BRK0149	0.812359	Biotechnology	2162713.000	USD
Deciphera Pharmaceuticals Inc.	DCPH	24344T10	BZ1LFB2	0.811702	Biotechnology	609420.000	USD
Fate Therapeutics Inc.	FATE	31189P10	BCZS820	0.803955	Biotechnology	1094668.000	USD
Karyopharm Therapeutics Inc.	KPTI	48576U10	BG3FZW0	0.794179	Biotechnology	1927055.000	USD
Halozyme Therapeutics Inc.	HALO	40637H10	2975098	0.774476	Biotechnology	1487270.000	USD
REGENXBIO Inc.	RGNX	75901B10	BZ0G875	0.744640	Biotechnology	908300.000	USD
Turning Point Therapeutics Inc.	TPTX	90041T10	BJXBP30	0.726952	Biotechnology	499575.000	USD
Heron Therapeutics Inc	HRTX	42774610	BJ0XLZ3	0.714911	Biotechnology	1617503.000	USD
Madrigal Pharmaceuticals Inc.	MDGL	55886810	BD59BS7	0.702165	Biotechnology	286037.000	USD
TG Therapeutics Inc.	TGTX	88322Q10	B828K63	0.691409	Biotechnology	1660749.000	USD
Allakos Inc.	ALLK	01671P10	BDD19P8	0.653757	Biotechnology	380057.000	USD
Editas Medicine Inc.	EDIT	28106W10	BZ8FPH3	0.645001	Biotechnology	898533.000	USD
Momenta Pharmaceuticals Inc.	MNTA	60877T10	B018VB0	0.643145	Biotechnology	791547.000	USD
Portola Pharmaceuticals Inc.	PTLA	73701010	B93PNR2	0.640062	Biotechnology	1612829.000	USD
Puma Biotechnology Inc.	PBYI	74587V10	B7F2TY6	0.623893	Biotechnology	2781382.000	USD
AnaptysBio Inc.	ANAB	03272410	BDRW1L7	0.592659	Biotechnology	1391755.000	USD

Veracyte Inc	VCYT	92337F10	BFTWZY0	0.562021	Biotechnology	993598.000	USD
Esperion Therapeutics Inc.	ESPR	29664W10	BBNBTD2	0.551953	Biotechnology	571881.000	USD
Ironwood Pharmaceuticals Inc. Class A	IRWD	46333X10	B3MZ6K5	0.549510	Biotechnology	2489050.000	USD
OPKO Health Inc.	OPK	68375N10	2115902	0.533369	Biotechnology	9418391.000	USD
MacroGenics Inc.	MGNX	55609910	BFDV8K0	0.527073	Biotechnology	911145.000	USD
Aimmune Therapeutics Inc	AIMT	00900T10	BYVDTK8	0.525698	Biotechnology	1250752.000	USD
Catalyst Pharmaceuticals Inc.	CPRX	14888U10	B1G7Q03	0.524194	Biotechnology	5277572.000	USD
Sangamo Therapeutics Inc.	SGMO	80067710	2573083	0.495021	Biotechnology	2375536.000	USD
Cytokinetics Incorporated	CYTK	23282W60	BBBSBJ5	0.449872	Biotechnology	875444.000	USD
Myriad Genetics Inc.	MYGN	62855J10	2614153	0.447058	Biotechnology	1564193.000	USD
Clovis Oncology Inc.	CLVS	18946410	B6RS2B3	0.416894	Biotechnology	2610218.000	USD
Epizyme Inc.	EPZM	29428V10	B9Z1QZ7	0.388412	Biotechnology	858174.000	USD
Radius Health Inc	RDUS	75046920	BM68KK1	0.383964	Biotechnology	1300164.000	USD
Twist Bioscience Corp.	TWST	90184D10	BGKG6G7	0.380682	Biotechnology	424385.000	USD
Intellia Therapeutics Inc.	NTLA	45826J10	BYZM6C2	0.380268	Biotechnology	829637.000	USD
Viking Therapeutics Inc.	VKTX	92686J10	BQQG1V1	0.369377	Biotechnology	2220449.000	USD
Enanta Pharmaceuticals Inc.	ENTA	29251M10	B9L5200	0.368227	Biotechnology	336823.000	USD
Xencor Inc.	XNCR	98401F10	BGCYWN8	0.366019	Biotechnology	540718.000	USD
Principia Biopharma Inc.	PRNB	74257L10	BFZ4V53	0.337952	Biotechnology	243539.000	USD
G1 Therapeutics Inc.	GTHX	3621LQ10	BF0QXG9	0.333452	Biotechnology	686732.000	USD
Alector Inc.	ALEC	01444210	BJ4LDC4	0.331558	Biotechnology	481625.000	USD
Athenex Inc.	ATNX	04685N10	BF04FC3	0.320702	Biotechnology	1098288.000	USD
Y-mAbs Therapeutics Inc.	YMAB	98424110	BG31GH0	0.319628	Biotechnology	309397.000	USD
Denali Therapeutics Inc.	DNLI	24823R10	BD2B4V0	0.301185	Biotechnology	523257.000	USD
Eagle Pharmaceuticals Inc.	EGRX	26979610	BJH7VB4	0.301139	Biotechnology	294527.000	USD
Retrophin Inc.	RTRX	76129910	B95XCC2	0.292187	Biotechnology	846692.000	USD
Dicerna Pharmaceuticals Inc.	DRNA	25303110	BJ62Z07	0.290951	Biotechnology	535526.000	USD
Akebia Therapeutics Inc.	AKBA	00972D10	BKKMP44	0.290653	Biotechnology	1041954.000	USD
Chemocentryx Inc.	CCXI	16383L10	B6ZL968	0.284609	Biotechnology	213434.000	USD
Atara Biotherapeutics Inc	ATRA	04651310	BP4WT09	0.275336	Biotechnology	1152465.000	USD
Flexion Therapeutics Inc.	FLXN	33938J10	BJ36RM8	0.271872	Biotechnology	935314.000	USD

Kodiak Sciences Inc.	KOD	50015M10	BFXC933	0.259285	Biotechnology	182372.000	USD
Anika Therapeutics Inc.	ANIK	03525510	2035754	0.258416	Biotechnology	341372.000	USD
Vericel Corporation	VCEL	92346J10	BSBMN89	0.257554	Biotechnology	804615.000	USD
BridgeBio Pharma Inc.	BBIO	10806X10	BK1KWG8	0.252304	Biotechnology	351153.000	USD
Vanda Pharmaceuticals Inc.	VNDA	92165910	B12W3P6	0.244616	Biotechnology	964074.000	USD
BioCryst Pharmaceuticals Inc.	BCRX	09058V10	2100362	0.242251	Biotechnology	2664482.000	USD
Voyager Therapeutics Inc.	VYGR	92915B10	BY7RB53	0.231195	Biotechnology	800248.000	USD
Krystal Biotech Inc.	KRYS	50114710	BD6JX35	0.228932	Biotechnology	228403.000	USD
ImmunoGen Inc.	IMGN	45253H10	2457864	0.228720	Biotechnology	2267400.000	USD
Dynavax Technologies Corporation	DVAX	26815820	BRJZSK0	0.221398	Biotechnology	1632681.000	USD
Gossamer Bio Inc.	GOSS	38341P10	BJ0CK86	0.216429	Biotechnology	778958.000	USD
AMAG Pharmaceuticals Inc.	AMAG	00163U10	2008121	0.213207	Pharmaceuticals	1398846.000	USD
Precigen Inc	PGEN	74017N10	BKM5C84	0.192135	Biotechnology	2097943.000	USD
Rhythm Pharmaceuticals Inc.	RYTM	76243J10	BF2YWG4	0.183688	Biotechnology	380204.000	USD
Progenics Pharmaceuticals Inc.	PGNX	74318710	2152501	0.181861	Biotechnology	1816806.000	USD
ZIOPHARM Oncology Inc.	ZIOP	98973P10	B0HZZ46	0.181819	Biotechnology	2498230.000	USD
Kura Oncology Inc.	KURA	50127T10	BYZD465	0.156787	Biotechnology	393099.000	USD
Rocket Pharmaceuticals Inc.	RCKT	77313F10	BDFKQ07	0.133047	Biotechnology	287910.000	USD
Akcea Therapeutics Inc.	AKCA	00972L10	BF47GP8	0.120644	Biotechnology	389553.000	USD
Agenus Inc.	AGEN	00847G70	B58J3K4	0.115149	Biotechnology	1395527.000	USD
AVROBIO Inc	AVRO	05455M10	BF2G3Y5	0.113603	Biotechnology	233430.000	USD
Homology Medicines Inc.	FIXX	43808310	BFMMJ23	0.111138	Biotechnology	318379.000	USD
Spectrum Pharmaceuticals Inc.	SPPI	84763A10	2982924	0.105437	Biotechnology	1535038.000	USD
NextCure Inc.	NXTC	65343E10	BJL3CL5	0.099113	Biotechnology	184152.000	USD
Karuna Therapeutics Inc.	KRTX	48576A10	BJMLSD2	0.090477	Biotechnology	41003.000	USD
Kadmon Holdings Inc.	KDMN	48283N10	BD0R7M2	0.089186	Biotechnology	839936.000	USD
Prothena Corp. Plc	PRTA	G7280010	B91XRN2	0.051704	Biotechnology	227584.000	USD
STATE STREET INSTITUTIONAL LIQ STATE STR	70286227	70286227	Unassigned	0.043709	Unassigned	1975874.500	USD
U.S. Dollar	CASH_USD	CASH_USD	Unassigned	0.001161	Unassigned	52496.880	USD
ACHILLION PHARMACEUTICALS INC COMMON STO	004CVR03	004CVR03	Unassigned	0.000528	Unassigned	51865.000	USD

RWR

Name	Ticker	Identifier	SEDOL	Weight	Sector	Shares Held	Local Currency
Prologis Inc.	PLD	74340W10	B44WZD7	10.402051	Industrial/Office	1403812.000	USD
Digital Realty Trust Inc.	DLR	25386810	B03GQS4	5.874758	Industrial/Office	510049.000	USD
Public Storage	PSA	74460D10	2852533	4.341713	Self Storage	285751.000	USD
AvalonBay Communities Inc.	AVB	05348410	2131179	3.193805	Residential	267575.000	USD
Welltower Inc.	WELL	95040Q10	BYVYHH4	3.135208	Healthcare	793661.000	USD
Alexandria Real Estate Equities Inc.	ARE	01527110	2009210	3.014312	Industrial/Office	239864.000	USD
Equity Residential	EQR	29476L10	2319157	2.946827	Residential	665061.000	USD
Realty Income Corporation	O	75610910	2724193	2.907886	Retail	652933.000	USD
Simon Property Group Inc.	SPG	82880610	2812452	2.906110	Retail	581243.000	USD
Invitation Homes Inc.	INVH	46187W10	BD81GW9	2.225891	Residential	1033885.000	USD
Essex Property Trust Inc.	ESS	29717810	2316619	2.190232	Residential	124383.000	USD
Healthpeak Properties Inc.	PEAK	42250P10	BJBLRK3	2.132394	Healthcare	1023522.000	USD
Ventas Inc.	VTR	92276F10	2927925	1.987998	Healthcare	709407.000	USD
Sun Communities Inc.	SUI	86667410	2860257	1.979461	Residential	186869.000	USD
Duke Realty Corporation	DRE	26441150	2284084	1.959043	Industrial/Office	700442.000	USD
Mid-America Apartment Communities Inc.	MAA	59522J10	2589132	1.930438	Residential	217454.000	USD
Boston Properties Inc.	BXP	10112110	2019479	1.876393	Industrial/Office	274759.000	USD
Extra Space Storage Inc.	EXR	30225T10	B02HWR9	1.846074	Self Storage	245436.000	USD
W. P. Carey Inc.	WPC	92936U10	B826YT8	1.698954	Diversified	327863.000	USD
Equity LifeStyle Properties Inc.	ELS	29472R10	2563125	1.694380	Residential	346359.000	USD
UDR Inc.	UDR	90265310	2727910	1.585447	Residential	560755.000	USD
Camden Property Trust	CPT	13313110	2166320	1.293786	Residential	185251.000	USD
Host Hotels & Resorts Inc.	HST	44107P10	2567503	1.135962	Hotels	1340987.000	USD
Americold Realty Trust	COLD	03064D10	B3SKZK7	1.094339	Industrial/Office	381052.000	USD
Regency Centers Corporation	REG	75884910	2726177	1.082625	Retail	322631.000	USD
Omega Healthcare Investors Inc.	OHI	68193610	2043274	1.031194	Healthcare	431653.000	USD
American Homes 4 Rent Class A	AMH	02665T30	BCF5RR9	1.007903	Residential	485602.000	USD
VEREIT Inc. Class A	VER	92339V10	BYVVTJ1	0.995084	Diversified	2050491.000	USD
Kilroy Realty Corporation	KRC	49427F10	2495529	0.910881	Industrial/Office	201383.000	USD
National Retail Properties Inc.	NNN	63741710	2211811	0.855065	Retail	327207.000	USD
Vornado Realty Trust	VNO	92904210	2933632	0.848923	Diversified	301756.000	USD
Federal Realty Investment Trust	FRT	31374720	2333931	0.836458	Retail	133880.000	USD

Healthcare Trust of America Inc. Class A	HTA	42225P50	BT9QF28	0.835041	Healthcare	415781.000	USD
Apartment Investment & Management Co Class A	AIV	03748R75	BJLNBG3	0.823257	Residential	283269.000	USD
CubeSmart	CUBE	22966310	B6SW913	0.788270	Self Storage	368410.000	USD
Kimco Realty Corporation	KIM	49446R10	2491594	0.778311	Retail	823110.000	USD
Rexford Industrial Realty Inc.	REXR	76169C10	BC9ZHL9	0.763307	Industrial/Office	233452.000	USD
Douglas Emmett Inc	DEI	25960P10	B1G3M58	0.734457	Industrial/Office	313700.000	USD
STORE Capital Corporation	STOR	86212110	BSKRKJ5	0.731916	Diversified	423415.000	USD
First Industrial Realty Trust Inc.	FR	32054K10	2360757	0.724900	Industrial/Office	242134.000	USD
EastGroup Properties Inc.	EGP	27727610	2455761	0.683990	Industrial/Office	74282.000	USD
Life Storage Inc.	LSI	53223X10	BDCSFJ6	0.672662	Self Storage	89310.000	USD
American Campus Communities Inc.	ACC	02483510	B02H871	0.664871	Residential	261865.000	USD
STAG Industrial Inc.	STAG	85254J10	B64BRQ5	0.648333	Industrial/Office	283056.000	USD
Cousins Properties Incorporated	CUZ	22279550	BJP0MF6	0.632111	Industrial/Office	282726.000	USD
QTS Realty Trust Inc. Class A	QTS	74736A10	BDSV8H8	0.578795	Industrial/Office	114812.000	USD
Equity Commonwealth	EQC	29462810	BPH3N63	0.562402	Industrial/Office	230296.000	USD
Healthcare Realty Trust Incorporated	HR	42194610	2417921	0.560765	Healthcare	255731.000	USD
Highwoods Properties Inc.	HIW	43128410	2420640	0.554632	Industrial/Office	196929.000	USD
SL Green Realty Corp.	SLG	78440X10	2096847	0.545365	Industrial/Office	145058.000	USD
Hudson Pacific Properties Inc.	HPP	44409710	B64B9P8	0.545029	Industrial/Office	290602.000	USD
Terreno Realty Corporation	TRNO	88146M10	B3N4753	0.534033	Industrial/Office	128357.000	USD
Brixmor Property Group Inc.	BRX	11120U10	BFTDJL8	0.531704	Retail	561844.000	USD
Agree Realty Corporation	ADC	00849210	2062161	0.509679	Retail	102115.000	USD
JBG SMITH Properties	JBGS	46590V10	BD3BX01	0.507839	Diversified	222858.000	USD
Spirit Realty Capital Inc.	SRC	84860W30	BHHZBZ8	0.493470	Diversified	195138.000	USD
Lexington Realty Trust	LXP	52904310	2139151	0.425586	Industrial/Office	523654.000	USD
Corporate Office Properties Trust	OFC	22002T10	2756152	0.419738	Industrial/Office	212635.000	USD
PS Business Parks Inc.	PSB	69360J10	2707956	0.392347	Industrial/Office	38067.000	USD
National Health Investors Inc.	NHI	63633D10	2626125	0.371757	Healthcare	84587.000	USD
EPR Properties	EPR	26884U10	B8XXZP1	0.365368	Retail	146841.000	USD
Taubman Centers Inc.	TCO	87666410	2872252	0.341781	Retail	116816.000	USD
Park Hotels & Resorts Inc.	PK	70051710	BYVMVV0	0.336732	Hotels	446526.000	USD
Weingarten Realty Investors	WRI	94874110	2946618	0.313614	Retail	228279.000	USD
Piedmont Office Realty Trust Inc. Class A	PDM	72019020	B3M3278	0.289989	Industrial/Office	238639.000	USD

Innovative Industrial Properties Inc	IIPR	45781V10	BD0NN55	0.286845	Industrial/Office	40434.000	USD
Apple Hospitality REIT Inc	APLE	03784Y20	BXRTX56	0.282487	Hotels	397671.000	USD
Washington Real Estate Investment Trust	WRE	93965310	2942304	0.269089	Diversified	156021.000	USD
Brandywine Realty Trust	BDN	10536820	2518954	0.263931	Industrial/Office	323194.000	USD
National Storage Affiliates Trust	NSA	63787010	BWWCK85	0.263633	Self Storage	117546.000	USD
Ryman Hospitality Properties Inc.	RHP	78377T10	B8QV5C9	0.256167	Hotels	104230.000	USD
Easterly Government Properties Inc.	DEA	27616P10	BVSS693	0.254542	Industrial/Office	142362.000	USD
Sunstone Hotel Investors Inc.	SHO	86789210	B034LG1	0.249108	Hotels	408680.000	USD
Four Corners Property Trust Inc.	FCPT	35086T10	BZ16HK0	0.241222	Retail	133281.000	USD
CareTrust REIT Inc	CTRE	14174T10	BMP8TL6	0.236417	Healthcare	181373.000	USD
Pebblebrook Hotel Trust	PEB	70509V10	B4XBDV9	0.221783	Hotels	247864.000	USD
LTC Properties Inc.	LTC	50217510	2498788	0.216250	Healthcare	74348.000	USD
Global Net Lease Inc	GNL	37937820	BZCFW78	0.214279	Diversified	169574.000	USD
Retail Properties of America Inc. Class A	RPAI	76131V20	B7QR337	0.211231	Retail	405857.000	USD
RLJ Lodging Trust	RLJ	74965L10	B3PY1N7	0.209851	Hotels	312978.000	USD
Columbia Property Trust Inc.	CXP	19828720	BFLY9C0	0.207450	Industrial/Office	216881.000	USD
Paramount Group Inc.	PGRE	69924R10	BSL7HC6	0.205377	Industrial/Office	361500.000	USD
Monmouth Real Estate Investment Corporation Class A	MNR	60972010	2504072	0.201834	Industrial/Office	185543.000	USD
Essential Properties Realty Trust Inc.	EPRT	29670E10	BFFK0X2	0.200792	Diversified	174317.000	USD
Industrial Logistics Properties Trust	ILPT	45623710	BFFK7S6	0.195742	Industrial/Office	123511.000	USD
Mack-Cali Realty Corporation	CLI	55448910	2192314	0.194270	Industrial/Office	171683.000	USD
American Assets Trust Inc.	AAT	02401310	B3NTLD4	0.193480	Diversified	91091.000	USD
Urban Edge Properties	UE	91704F10	BTPSGQ9	0.181330	Retail	209799.000	USD
Retail Opportunity Investments Corp.	ROIC	76131N10	B28YD08	0.176411	Retail	220561.000	USD
Office Properties Income Trust	OPI	67623C10	BYVLR75	0.175896	Industrial/Office	91312.000	USD
SITE Centers Corp.	SITC	82981J10	BGL0KF5	0.165634	Retail	281853.000	USD
Acadia Realty Trust	AKR	00423910	2566522	0.155966	Retail	163190.000	USD
Service Properties Trust	SVC	81761L10	BKRT1C8	0.154044	Hotels	311901.000	USD
DiamondRock Hospitality Company	DRH	25278430	B090B96	0.153768	Hotels	377675.000	USD
Independence Realty Trust Inc.	IRT	45378A10	BCRYTK1	0.153354	Residential	179541.000	USD
Getty Realty Corp.	GTY	37429710	2698146	0.146978	Retail	65147.000	USD
Xenia Hotels & Resorts Inc.	XHR	98401710	BVV6CY1	0.146366	Hotels	215036.000	USD
Macerich Company	MAC	55438210	2543967	0.145714	Retail	220405.000	USD

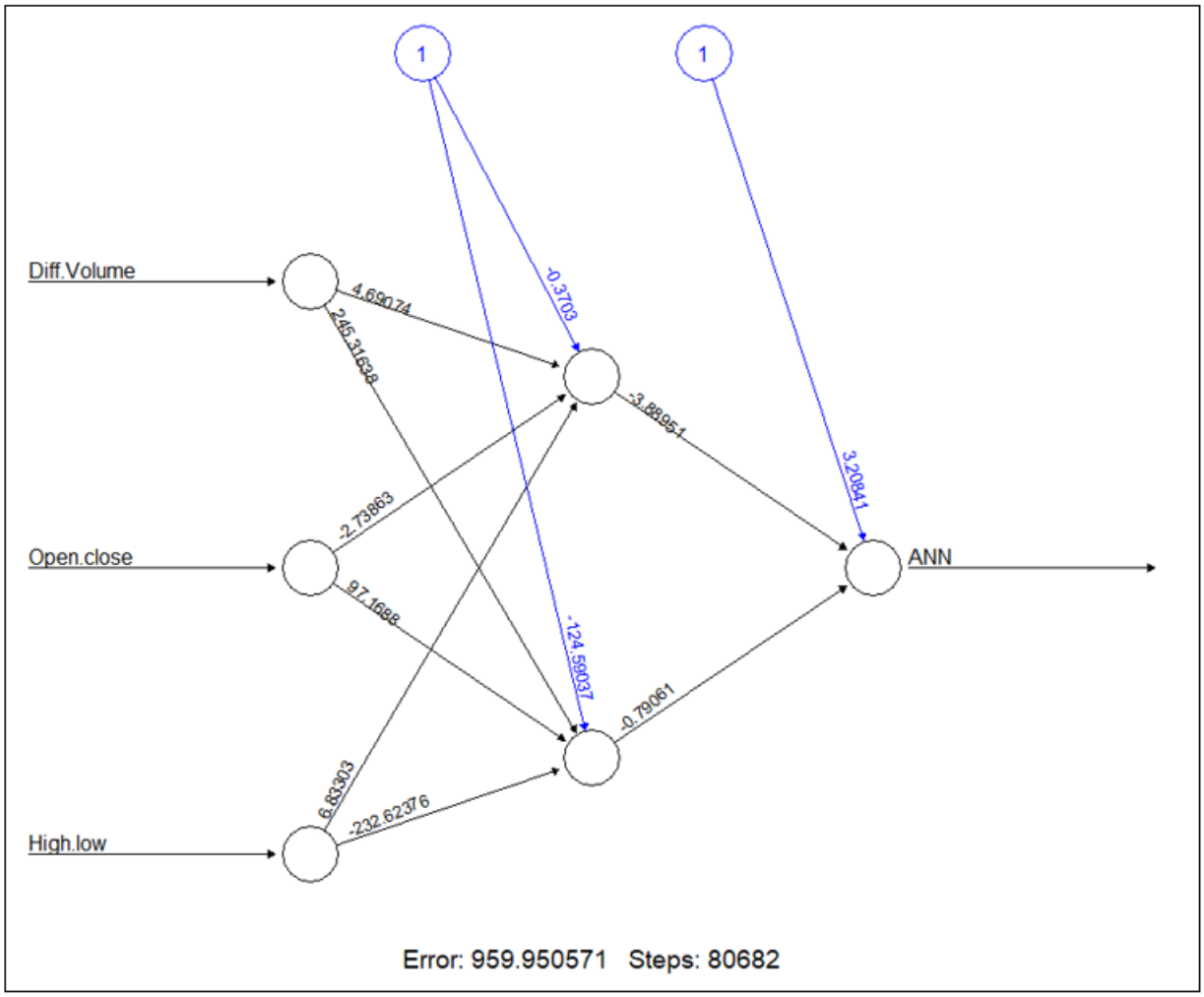
Empire State Realty Trust Inc. Class A	ESRT	29210410	BF321D7	0.143692	Diversified	274441.000	USD
Diversified Healthcare Trust	DHC	25525P10	BKRN595	0.134610	Healthcare	450926.000	USD
Kite Realty Group Trust	KRG	49803T30	BPBSZJ1	0.129687	Retail	159417.000	USD
Investors Real Estate Trust	IRET	46173050	BFWQ204	0.129580	Residential	23030.000	USD
Universal Health Realty Income Trust	UHT	91359E10	2927497	0.127828	Healthcare	23946.000	USD
Community Healthcare Trust Inc.	CHCT	20369C10	BXC87C3	0.127127	Healthcare	39002.000	USD
NexPoint Residential Trust Inc	NXRT	65341D10	BWC6PW6	0.109747	Residential	41439.000	USD
STATE STREET INSTITUTIONAL LIQ STATE STR	70286227	70286227	Unassigned	0.092234	Unassigned	1180270.100	USD
Tanger Factory Outlet Centers Inc.	SKT	87546510	2874582	0.091640	Retail	177141.000	USD
Summit Hotel Properties Inc.	INN	86608210	B3M7R64	0.086673	Hotels	200199.000	USD
Franklin Street Properties Corp.	FSP	35471R10	B02T2D1	0.078475	Industrial/Office	203281.000	USD
RPT Realty	RPT	74971D10	BG0YLC2	0.075547	Retail	152723.000	USD
City Office REIT Inc.	CIO	17858710	BL25F37	0.066901	Industrial/Office	90211.000	USD
Front Yard Residential Corp. Class B	RESI	35904G10	BDSHDS2	0.060240	Residential	93325.000	USD
Seritage Growth Properties Class A	SRG	81752R10	BZ0HC54	0.053034	Retail	66599.000	USD
Chatham Lodging Trust	CLDT	16208T10	B5LYMC1	0.039099	Hotels	89027.000	USD
Hersha Hospitality Trust Class A	HT	42782550	BYYTSK6	0.029444	Hotels	68257.000	USD
Washington Prime Group Inc.	WPG	93964W10	BD5JMM8	0.020391	Retail	355159.000	USD
Pennsylvania Real Estate Investment Trust	PEI	70910210	2680767	0.011519	Retail	115155.000	USD
U.S. Dollar	CASH_USD	CASH_USD	Unassigned	-0.022378	Unassigned	-286356.400	USD

XLE

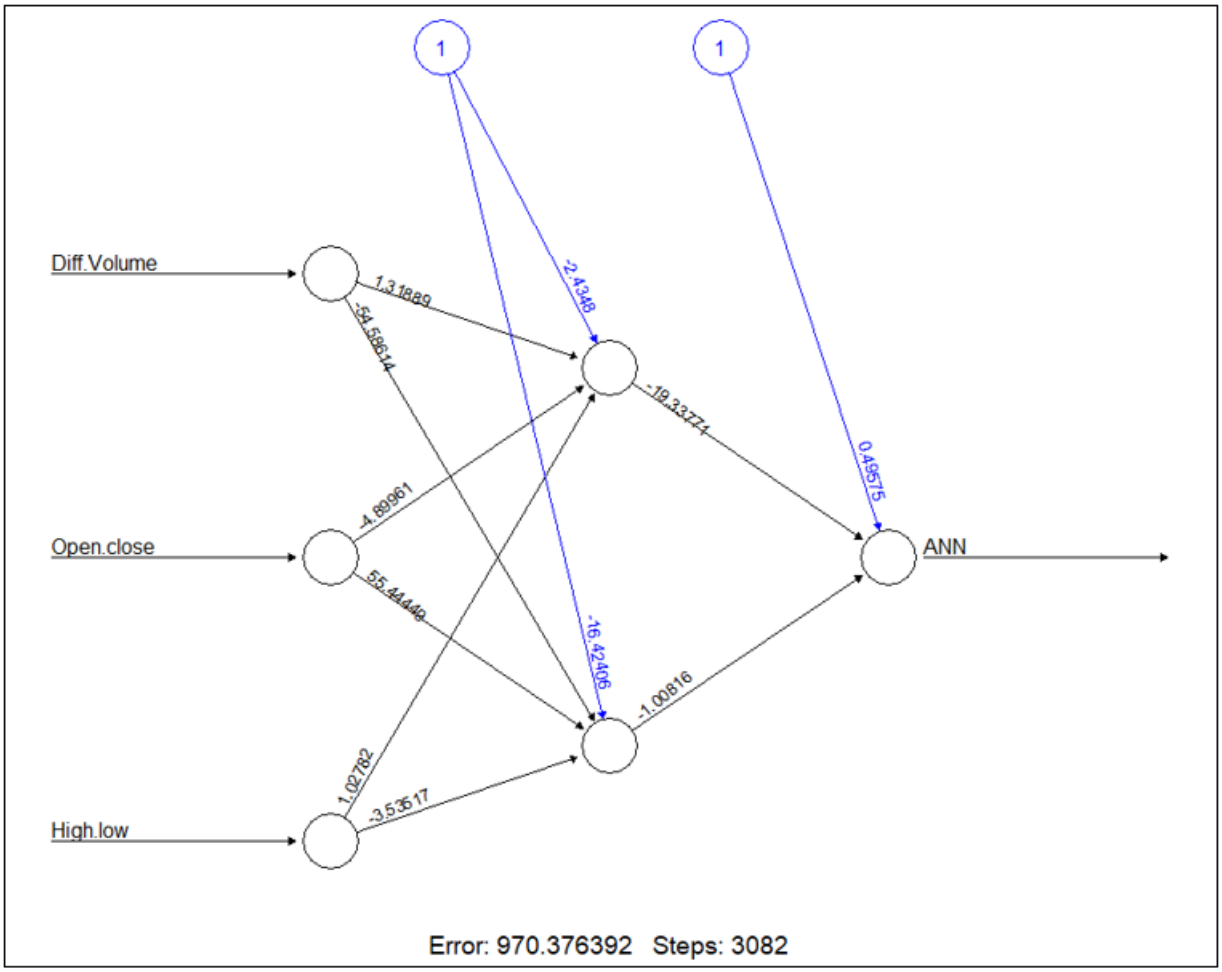
Symbol	Name	% Holding	Shares
CVX	Chevron Corporation	23.43%	27105704
XOM	EXXON MOBIL	22.96%	53090976
KMI	KINDER MORGAN	4.55%	31503250
COP	ConocoPhillips	4.51%	11204429
EOG	EOG RESOURCES	4.34%	9428035
PSX	Phillips 66	4.17%	6734373
SLB	Schlumberger Ltd.	4.06%	22479922
MPC	Marathon Petroleum Corp	3.87%	10532932
WMB	WILLIAMS COMPS	3.70%	19651470
VLO	VALERO ENERGY	3.59%	6603927
PXD	Pioneer Natural Resources Co	2.50%	2670412
OXY	OCCIDENTAL PET	2.47%	14578539
OKE	ONEOK Inc	1.93%	7123219
HES	Hess Corp	1.93%	4228890
HAL	Halliburton Co	1.74%	14210162
BKR	Baker Hughes a GE Co. Class A	1.62%	10610275
CXO	Concho Resources Inc	1.59%	3186197
COG	Cabot Oil & Gas Corp.	1.15%	6456189
FANG	DIAMOND BAK EN	1.01%	2556366
APA	Apache Corp	0.79%	6113604
NOV	National Oilwell Varco Inc	0.72%	6288492
NBL	NOBLE ENERGY	0.71%	7770208
MRO	MARATHON OIL	0.69%	12801570
HFC	HollyFrontier Corp	0.67%	2412523
DVN	DEVON ENERG	0.64%	6199031
FTI	TECHNIPFMC PLC	0.49%	6812941
	STATE STREET INSTITUTIONAL LIQ STATE STR	0.19%	18490594
XLE	The Energy Select Sector SPDR Fund	0.00%	0

Annexe 2 : Illustrations de nos réseaux de neurones

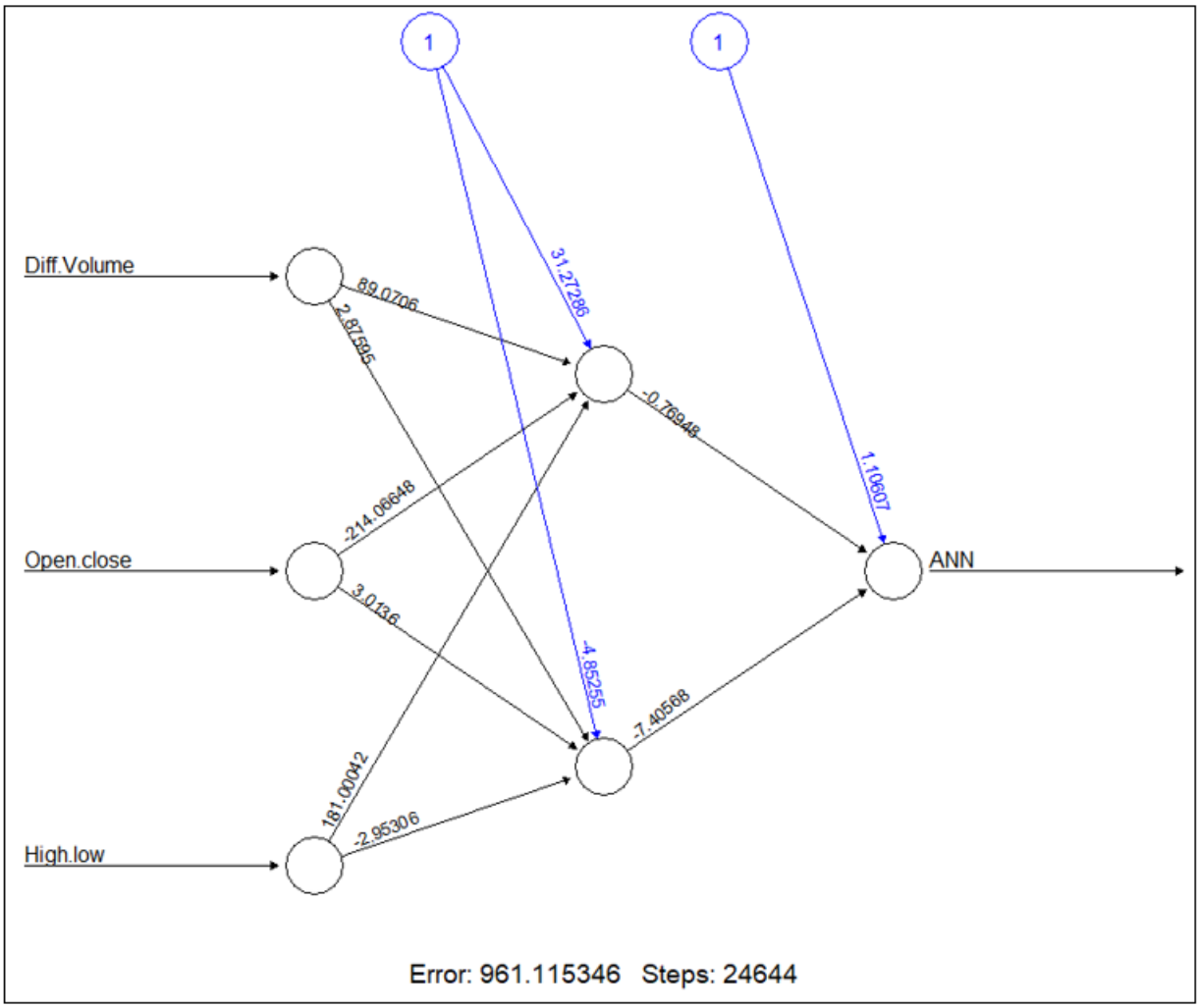
XBI



RWR


1
1
Diff.Volume
Open.close
High.low
ANN
1.3.1889
-54.59614
-4.89961
55.44449
1.02782
-3.53517
-2.41348
-16.42406
-19.33774
-1.00816
0.49575
Error: 970.376392 Steps: 3082

XLE

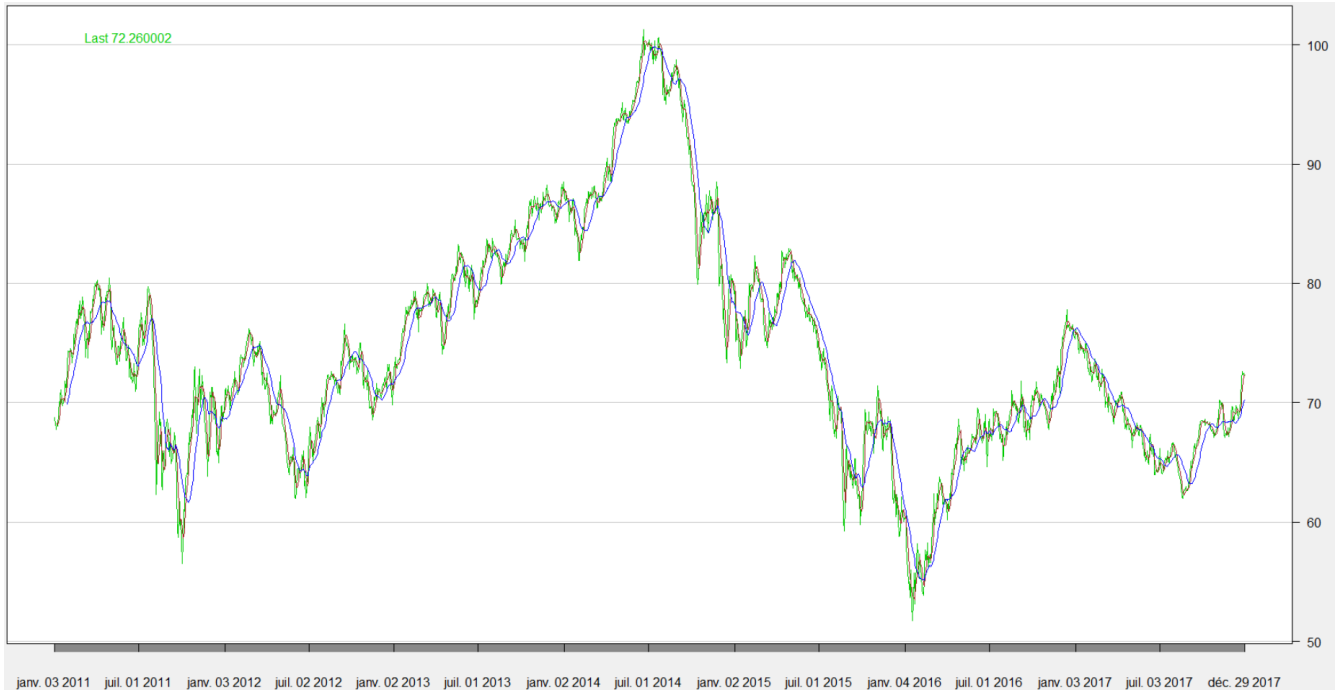

1
1
Diff.Volume
Open.close
High.low
89.0706
-2.87595
31.27286
-214.06648
3.0136
-4.85255
181.00042
-2.95306
-0.76948
1.10607
-7.40568
ANN
Error: 961.115346 Steps: 24644

Annexe 3 : Représentation des deux moyennes mobile simples

XBI

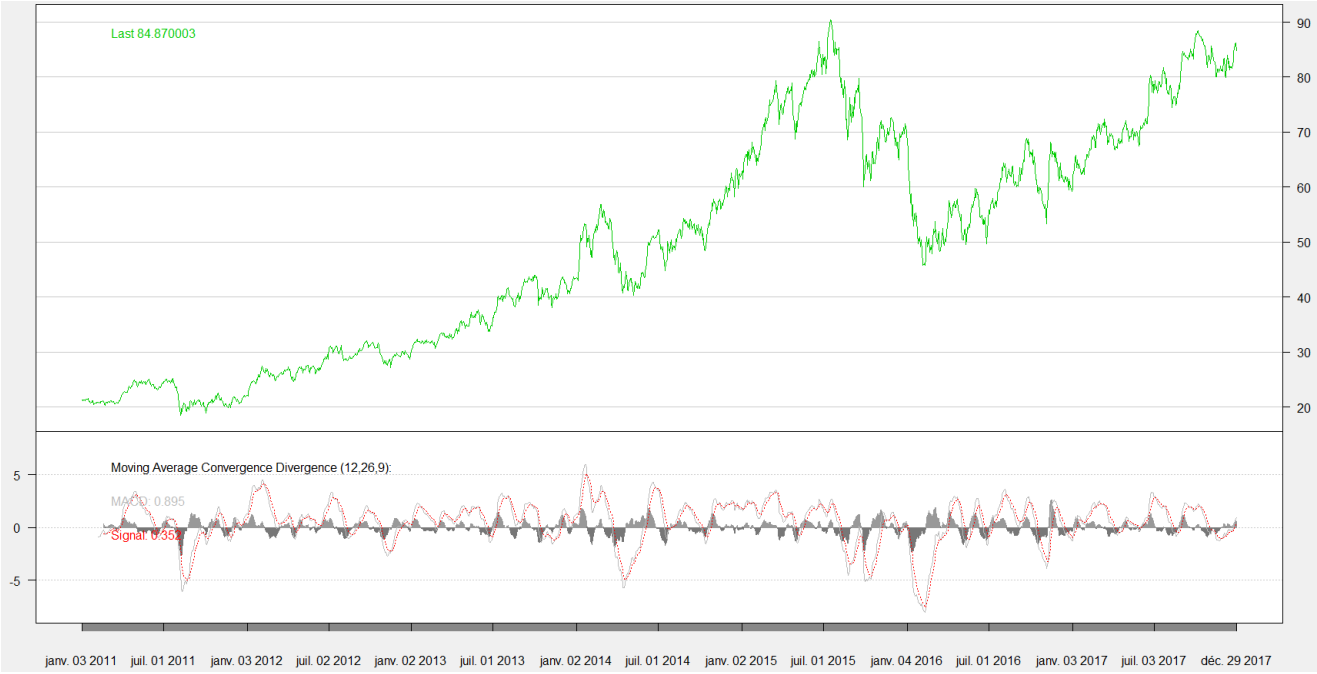
RWR

XLE

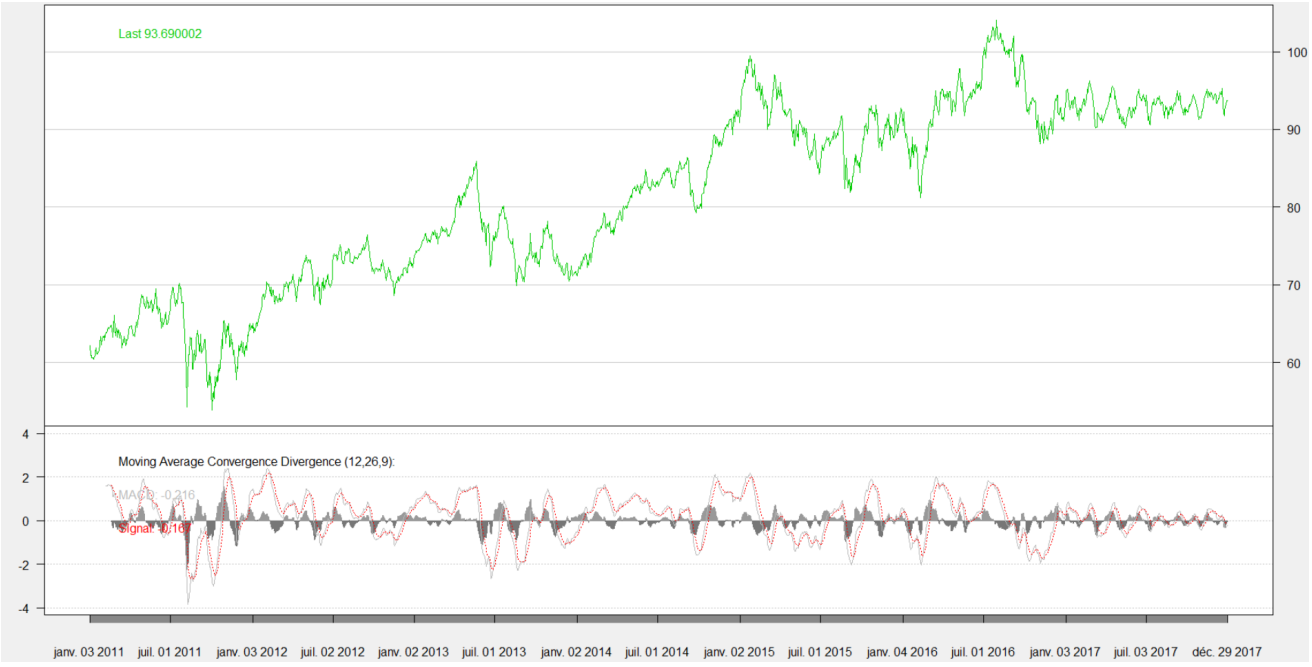

Last 72.260002
100
90
80
70
60
50
janv. 03 2011
juil. 01 2011
janv. 03 2012
juil. 02 2012
janv. 02 2013
juil. 01 2013
janv. 02 2014
juil. 01 2014
janv. 02 2015
juil. 01 2015
janv. 04 2016
juil. 01 2016
janv. 03 2017
juil. 03 2017
déc. 29 2017

Annexe 4 : Représentation de la ligne MACD et de la ligne de signal

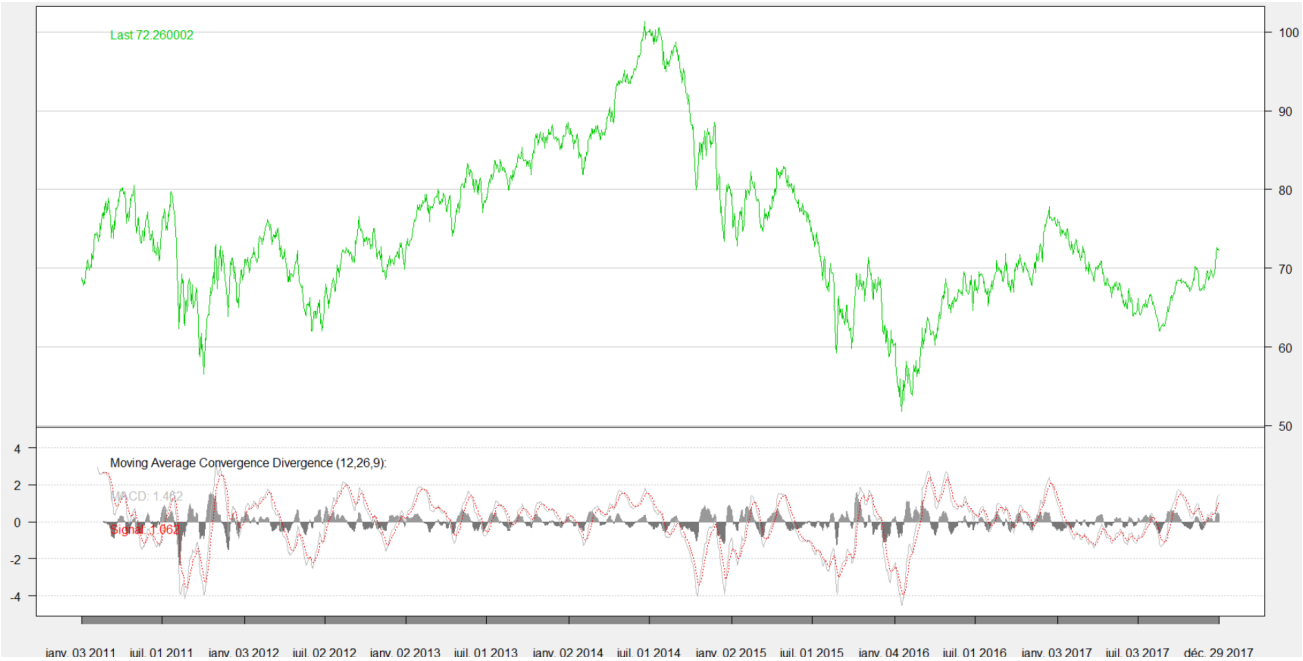
XBI



RWR



XLE


Last 72.260002
100
90
80
70
60
50
Moving Average Convergence Divergence (12,26,9):
MACD: 1.462
Signal: 1.062
4
2
0
-2
-4
janv. 03 2011
juil. 01 2011
janv. 03 2012
juil. 02 2012
janv. 02 2013
juil. 01 2013
janv. 02 2014
juil. 01 2014
janv. 02 2015
juil. 01 2015
janv. 04 2016
juil. 01 2016
janv. 03 2017
juil. 03 2017
déc. 29 2017

Annexe 5 : Matrices de confusion et précisions pour les données test

XBI

Machine Learning :

CART :

```
testPred  Acheter Vendre
  Acheter     110     86
  Vendre       68     77
> sum(diag(tab))/sum(tab)
[1] 0.5483871
```

C5.0 :

```
p1        Acheter Vendre
  Acheter     107     81
  Vendre       71     82
> sum(diag(confusion))/sum(confusion)
[1] 0.5542522
```

RF :

```
          Reference
Prediction Acheter Vendre
   Acheter     114     81
   Vendre       64     82

               Accuracy : 0.5748
```

ANN

```
pred2   0   1
    0  58  51
    1 105 127
> sum(diag(tab2))/sum(tab2)
[1] 0.542522
```

SVM

```
         Actual
Predicted Acheter Vendre
  Acheter     134    113
  Vendre       44     50
> sum(diag(matrix2))/sum(matrix2)
[1] 0.5395894
```

Analyse technique :

SMA :

```
     0   1
  0  78  76
  1  85 102
> sum(diag(tab))/sum(tab)
[1] 0.5278592
```

MACD :

```
    0  1
  0 98 96
  1 65 82
> sum(diag(tab))/sum(tab)
[1] 0.5278592
```

RSI :

```
     0   1
  0 162 175
  1   1   3
> sum(diag(tab))/sum(tab)
[1] 0.483871
```

ROC :

```
     0   1
  0  62  63
  1 101 115
> sum(diag(tab))/sum(tab)
[1] 0.5190616
```

RWR

Machine Learning :

CART :

```
testPred  Acheter Vendre
  Acheter     177    164
  Vendre        0      0
> sum(diag(tab))/sum(tab)
[1] 0.5190616
```

C5.0 :

```
p1        Acheter Vendre
  Acheter     177    164
  Vendre        0      0
> sum(diag(confusion))/sum(confusion)
[1] 0.5190616
```

RF :

```
          Reference
Prediction Acheter Vendre
   Acheter     102     90
   Vendre       75     74

               Accuracy : 0.5161
```

ANN :

```
pred2   0   1
    0  19  12
    1 145 165
> sum(diag(tab2))/sum(tab2)
[1] 0.5395894
```

SVM :

```
          Actual
Predicted Acheter Vendre
  Acheter     159    147
  Vendre       18     17
> sum(diag(matrix2))/sum(matrix2)
[1] 0.516129
```

Analyse technique :

SMA :

```
   val
      0   1
  0  78 100
  1  86  77
> sum(diag(tab))/sum(tab)
[1] 0.4545455
```

MACD :

```
     0  1
  0 83 90
  1 81 87
> sum(diag(tab))/sum(tab)
[1] 0.4985337
```

RSI :

```
> print(tab)

      0   1
  0 164 171
  1   0   6
> sum(diag(tab))/sum(tab)
[1] 0.4985337
```

ROC :

```
      0   1
  0  82 106
  1  82  71
> sum(diag(tab))/sum(tab)
[1] 0.4486804
```

XLE

Machine Learning :

CART :

```
testPred  Acheter Vendre
  Acheter     122    127
  Vendre       45     47
> sum(diag(tab))/sum(tab)
[1] 0.4956012
```

C5.0 :

```
p1        Acheter Vendre
  Acheter     122    127
  Vendre       45     47
> sum(diag(confusion))/sum(confusion)
[1] 0.4956012
```

RF :

```
          Reference
Prediction Acheter Vendre
   Acheter      90     94
   Vendre       77     80

               Accuracy : 0.4985
```

ANN :

```
pred2   0   1
    0  54  51
    1 120 116
> sum(diag(tab2))/sum(tab2)
[1] 0.4985337
```

SVM :

```
          Actual
Predicted Acheter Vendre
  Acheter     111    117
  Vendre       56     57
> sum(diag(matrix2))/sum(matrix2)
[1] 0.4926686
```

Analyse technique :

SMA :

```
    0  1
  0 88 95
  1 86 72
> sum(diag(tab))/sum(tab)
[1] 0.4692082
```

MACD :

```
    0  1
  0 87 88
  1 87 79
> sum(diag(tab))/sum(tab)
[1] 0.4868035
```

RSI :

```
     0   1
  0 172 165
  1   2   2
> sum(diag(tab))/sum(tab)
[1] 0.5102639
```

ROC :

```
    0  1
  0 94 93
  1 80 74
> sum(diag(tab))/sum(tab)
[1] 0.4926686
```

Annexe 6 : Résultats de l'étude de Macchiarulo A.

	N	Minimum	Maximum	Mean	Std. Deviation
Machine Learning	120	-20.4	23.5	1.192917	7.347258
Bollinger Bands	120	-13.129	19.71627	0.831313	3.758738
Trading Envelopes	120	-13.129	19.71627	0.831313	3.758738
KBand	120	-13.129	13.31265	0.76489	4.25558
Cmdty Channel Index	120	-16.5331	13.06419	0.538209	3.622277
Stochastics	120	-9.02627	11.50051	0.492497	3.374356
William's %R	120	-9.02627	12.64072	0.408589	3.396424
Buy and Hold	120	-16.5331	14.2041	0.403511	4.553905
Fundamental Analysis	120	-18.46	10.18	0.383	4.45161
MA Envelopes	120	-6.10397	13.94343	0.289602	2.477517
RSI	120	-4.61304	12.65053	0.276992	1.865896
MACD	120	-9.07901	8.290536	0.255896	3.016252
Ichimoku	120	-5.92016	7.675862	0.050554	2.08152
Triangular MA	120	-8.58764	6.438574	-0.1277	2.515717
DMI	120	-14.0814	8.636103	-0.19479	2.808664
Exponential MA	120	-8.7846	8.829373	-0.22599	2.59998
MA Oscillator	120	-10.4149	10.87842	-0.23029	3.917283
Fear and Greed	120	-13.9833	9.543643	-0.23482	3.406039
Simple MA	120	-16.8559	8.496324	-0.54573	3.442464
Weighted MA	120	-15.3914	6.827461	-0.55405	3.148252
Variable MA	120	-23.4974	6.569704	-0.67838	3.79582
Parabolic	120	-23.3883	11.8544	-0.69678	4.676691
Accum/Distrib Osc	120	-21.6401	16.67208	-1.01638	5.40478
Rex Oscillator	120	-18.6651	11.25333	-1.02282	4.683956
Rate of Change	120	-18.1407	14.31845	-1.20797	4.95222
Valid N (list wise)	120				

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Chaussée de Binche 151, 7000 Mons, Belgique | **www.uclouvain.be/lsm**