UCLouvain

LOUVAIN
School of Management

Louvain School of Management

Construction d'un portefeuille d'actions sur base d'un modèle de Machine Learning

Auteur : Jérémy BRANCART
Promoteur(s) : Frédéric VRINS
Lecteur : Nathan LASSANCE
Année académique 2019 - 2020

Remerciements

Tout au long de la réalisation de ce mémoire ainsi que de mes années d'études, j'ai eu la chance d'être soutenu, conseillé et aidé par plusieurs personnes.

Tout d'abord, j'aimerais remercier mon promoteur, le Professeur Frédéric Vrins, pour sa patience ainsi que pour ses remarques et conseils qui m'ont permis de mener à bien ce mémoire.

Ensuite, je tiens à remercier tout particulièrement mes parents, Brigitte et Michel Brancart, de m'avoir permis de réaliser des études universitaires et surtout pour leur soutien sans faille, quelque soit la situation, ainsi que pour leurs innombrables encouragements reçus tout au long de ces années d'études.

J'aimerais également remercier mes grands parents maternels pour leur acceuil plus que chaleureux durant mes premiers blocus universitaires. J'espère les rendre fiers en obtenant mon diplôme.

Il est également important pour moi de remercier ma marraine, Claudine Brancart, ainsi que son mari, Axel Cooman, pour le temps passé à la relecture de ce mémoire.

Finalement, j'aimerais remercier mes amis et en particulier, Denis Martin et Vincent Carpentier pour leur aide et leur soutien reçu durant la rédaction de ce mémoire.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Les investisseurs ont toujours été de farouches adeptes de l'utilisation de nouvelles technologies pour obtenir un avantage informationnel. Au XVème siècle, les "traders" Vénitiens ont très vite commencé à utiliser les téléscopes afin d'inspecter les drapeaux des bateaux entrants dans le port et de cette manière, obtenir des indices sur leur cargaison pour acheter et vendre des marchandises en conséquence.[Wigglesworth, 2020]

De nos jours, les "traders" ne dérogent pas aux habitudes du passé. Cependant, les nouvelles technologies ne sont plus des objets tels que le téléscope mais plutôt des technologies dérivées de la science informatique et de la science des données telles que le Machine Learning, l'utilisation de données alternatives ou encore du Big Data. D'ailleurs un des quotidiens économiques le plus lus au monde, le *Financial Times* [Wigglesworth, 2020] a titré un de ses grands reportages ; "*Stockpickers turn to big data to arrest decline*", c'est-à-dire que les gestionnaires de fonds, les professionnels du monde de l'investissement font appel au Big Data afin de contrer le déclin actuel auquel leur industrie fait face.

L'adoption de ces nouvelles technologies, par les professionnels est un fait, et plus précisément encore, l'adoption du Machine Learning. Le graphique présenté à L'Annexe 1 de ce travail montre que les groupes d'investissement adoptent de façon de plus en plus importante le Machine Learning. Dans une interview accordéé au numéro supllémentaire "*Fonds*" du journal *L'Echo* [Van Maldegem, 2020] , Nicola Horlick, une pionnière de la gestion de fonds de la City de Londres, a mentionné que : "*L'avenir est aux mains des mathématiciens qui développent des algorithmes traitant des énormes quantités de données disponibles [...] Je suis convaincue que des gestionnaires "quantitatifs" (c'est-à-dire des gestionnaires qui composent des fonds sur base*

d'algorithmes ou de modèles mathématiques) l'emporteront face aux gestionnaires traditionnels".

C'est donc au vu de cette tendance qu'il nous est paru intéressant de réaliser un mémoire de fin d'étude sur l'application de ces outils quantitatifs au monde de la finance. Notre travail se concentrera sur l'application d'un de ces outils, le **Machine Learning**, à la finance et plus précisément à la gestion de portefeuille. En effet, *Harry Markowitz*, avec son article paru en 1952 [Markowitz, 1952], a posé les fondations de la sélection de portefeuille en utilisant des méthodes mathématiques. le montrendant, assez vite, les professionnels ont émis certaines critiques à l'égard des techniques de *Markowitz* et ont commencé à développer des améliorations qui surperforme, en terme de rendement, le portefeuille de *Markowitz*.

Au regard de l'importance croissante accordée au Machine Learning dans le monde de la finance, il nous paraissait intéressant de construire un portefeuille sur base du Machine Learning et d'en observer ses performances.

Après la présentation du contexte de ce mémoire, une **question de recherche** principale a été identifiée :

Comment construit-on un portefeuille d'actions sur base d'un algorithme de Machine Learning (plus précisément d'un algorithme de classification avec aspect probabiliste) et quel en est la performance ?

Une autre question de recherche, celle-ci d'ordre secondaire et se rapportant aux données avec lesquelles ce travail a été réalisé, a également été identifiée :

Est-ce que les portefeuilles construits en période dite de crise vont se révéler être sous-performants par rapport aux portefeuilles construits en période dite de "post-crise" ?

Afin de répondre à ces 2 questions et de mettre en pratique la théorie, ce mémoire se structure comme suit :

Le **Chapitre 2** décrira les grandes théories de gestions de portefeuille, leurs limites ainsi que les améliorations qui y ont été apportées.

Le **Chapitre 3** définit le Machine Learning et décrit son fonctionnement. Différentes applications de Machine Learning développées par de grandes institutions financières sont également exposée.

Le **Chapitre 4** décrit la base de données et explique les différentes étapes nécessaires à la construction d'un portefeuille sur base d'un modèle de Machine Learning basé sur un algorithme de classification avec aspect probabiliste.

Le **Chapitre 5** expose les différents résultats obtenus.

Le **Chapitre 6** conclu ce mémoire, présente les limites rencontrées durant l'étude et propose des pistes pour des éventuelles recherches futures.

Première partie

Partie Théorique

Chapitre 2

La sélection de portefeuilles

Dans ce chapitre, nous débuterons par un bref rappel de ce qu'est la théorie sur l'efficience des marchés et ensuite, nous exposerons les différents moments statistiques de la fonction de distribution du "return" d'un portefeuille et leurs caractéristiques respectives. Dans un deuxième temps, nous présenterons la théorie moderne de gestion de portefeuilles selon Harry Markowitz ainsi que ses limites et désavantages et les améliorations qui y ont été apportées. Dans ce travail, nous nous concentrerons sur une amélioration développée par Ledoit &Wolff [Ledoit and Wolf, 2004a] et qui consiste en une méthode dite de "shrinkage" permettant d'estimer de manière robuste les paramètres de la matrice des covariances. De plus, une autre technique de gestion de portefeuille, à savoir le portefeuille ayant la variance minimum (*minimum variance portfolio*) sera également expliquée. Même si ce n'est pas le sujet principal de ce mémoire, nous jugeons nécessaire d'expliquer en profondeur ces différentes théories relatives à la gestion de portefeuille.

2.1 L'efficience des marchés selon Eugene F. Fama

La base de la théorie sur les marchés dits **efficients** est fondée sur l'hypothèse que les conditions d'équilibre de marché peuvent être énoncées en terme de rendements attendus/espérés. Dans son article, *Efficient Capital Markets : A review of theory and empirical work* [Fama, 1970], Fama définit un marché **efficient** comme un marché dans lequel les prix reflètent toujours, à tout moments, parfaitement toute l'information disponible.

L'article distingue 3 types d'efficience de marché :

— L'éfficience **faible** : Avec ce type d'efficience, l'ensemble de l'informa-

tion disponible est uniquement formé des prix historiques ainsi que des séquences de rendements.
Les résultats de tests empiriques montrent qu'il y a une preuve constante d'une dépendance positive dans les variations quotidiennnes des prix et des rendements des actions ordinaires. De plus, cette forme de dépendance peut être utilisée comme une base de règles de trading rentables.

— L'éfficience **semi-forte** : Ici, l'ensemble de l'information inclut toute l'information disponible publiquement (publication des *annual earnings*, des fractions d'actions, ...)
Les résultats empiriques supportent également l'hypothèse des marchés **efficients**

— L'éfficience **forte** : Dans ce type d'efficience, un investisseur individuel ou un groupe d'investisseurs ont un accès monopolistique à toutes informations pertinentes dans le cadre de la formation des prix.
Cette dernière forme est plutôt vue comme un benchmark des déviations à l'efficience de marché.

Pour le détail des résulats des tests réalisés à propos des différentes formes d'efficience de marché, le lecteur pourra se référer à l'article cité ci-dessus ainsi qu'à [Fama, 1991]. Il est également important de noter ici, que dans ce mémoire, nous considérons le marché étudié comme efficient au sens de Fama.

Enfin, dans ce mémoire, si nous parlons de *l'efficience de marché*, nous ferons référence à la forme *faible* de cette dernière. En effet notre base de données ainsi que notre analyse sont toutes deux basées sur les séries de prix historiques et les rendements propres à une action. [*cfr.* Section 3.1]

2.2 Les moments statistiques d'un portefeuille

Considérons un vecteur de rendement aléatoire (le rendement d'un titre est une variable aléatoire) $X = (x_1, x_2, ..., x_N) \in \mathbb{R}^{\mathbb{N}}$ où N est le nombre de titres présents dans notre univers d'investissement. Définissons maintenant le portefeuille P comme le vecteur $P = w'X$ où w représente le vecteur des poids attribués à chacun des N titres. Les 4 premiers moments[1] d'un portefeuille peuvent donc être définis mathématiquement comme suit :

$$m_1(P) = w' \, \mu_X \, w \tag{2.1}$$

$$m_2(P) = w' \sum_X w \tag{2.2}$$

1. Une définition des moments statistiques se trouve à **l'Annexe 2**

$$m_3(P) = w' \, \Phi_X \, (w \otimes w) \tag{2.3}$$

$$m_4(P) = w' \, \Psi_X \, (w \otimes w \otimes w) \tag{2.4}$$

où $\sum_X$ est la matrice de covariance, $\Phi_X \in \mathbb{R}^{NxN^2}$ représente les co-skewness de X et $\Psi_X \in \mathbb{R}^{NxN^3}$ est la notation pour la matrice des co-kurtosis de X. Ces 2 derniers co-moments sont des mesures de la variation simultanée de deux variables aléatoires. $\otimes$ est quant à lui le symbôle du produit de Kronecker, qui est un cas particulier du produit matriciel [2].

2.3 Le portefeuille équipondéré

Le portefeuille équipondéré ou equally-weighted portfolio, est une stratégie de gestion de portefeuille naïve mais très connue dans le monde de l'investissement. Un portefeuille découlant de cette stratégie dispose d'une proportion de **chaque** titre ou action présent dans l'univers d'investissement. D'après [DeMiguel et al., 2009] cette stratégie peut surperformer certaines stratégies dites optimales. Si nous considérons, comme noté ci-dessus, w comme un vecteur de poids et N, le nombre total de titres présents dans un univers d'investissement donné, nous pouvons définir mathématiquement l'élément w_i du vecteur w comme suit :

$$w_i = \frac{1}{N} \tag{2.5}$$

2.4 La théorie moderne de gestion de portefeuille, selon Harry Markowitz

En 1952, par son article *Portfolio Selection* [Markowitz, 1952] paru dans *The Journal of Finance*, **Harry Markowitz**, lauréat en 1990 du prix de la banque de Suède en Sciences Economiques en mémoire d'Alfred NOBEL, a fourni une remarquable avancée en terme d'optimisation et de diversification de portefeuilles. En effet, cette théorie connue du public sous le nom de *modern portfolio theory* a permis de répondre à une question que tout investisseur s'est déja posée, *"comment pourrais-je répartir mes fonds disponibles, de manière optimale, entre tous les choix possibles d'investissements"*. Markowitz a répondu à cette question de 2 façons. La première en quantifiant le *return* et le *risque* associé à un titre, par des mesures statistiques, respectivement le return ou le rendement attendu et la variance, et deuxièmement

2. Une définition du produit de Kronecker est exposée à l'**Annexe 3**

en prônant le fait que les investisseurs devraient considérer le rendement et le risque **ensemble** et non plus séparément et déterminer la répartition de leurs fonds , parmi les alternatives possibles en se basant sur une sorte de compromis entre rendement et risque. Dans son article, *Markowitz* définit un investisseur rationnel comme un investisseur qui doit considérer le rendement attendu (le return ou le rendement d'un titre) comme une chose désirable et la variance (le risque associé à un titre) comme une chose indésirable. [Kolm et al., 2014]. Selon Markowitz, *le portefeuille optimal* est donc celui qui **maximise** le return ou le rendement pour un niveau de risque défini. Plus précisément, la solution est donc un ensemble de portefeuilles efficients pour un certain niveau de risque donné. Illsustrons cela à la *figure* 2.1, présentée ci-dessous.

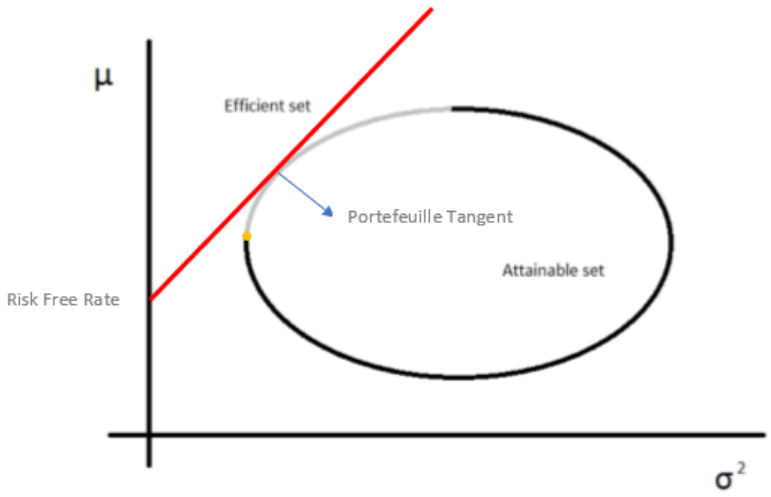

FIGURE 2.1 – Les différents portefeuilles selon Markowitz

Par ce graphique, il est clairement observable que, dans le cadre de la *modern portfolio theory* de **Markowitz**, certains portefeuilles sont considérés comme étant *efficients* s'ils disposent d'un niveau de risque minimal pour un certain niveau de rendement ou encore s'ils disposent d'un rendement maximal pour un certain niveau de risque donné. En d'autres termes, la frontière efficiente, ici "efficient set", regroupe l'ensemble des portefeuilles optimaux pour tout niveau de risque. A noter que dans ce mémoire, nous considérerons uniquement la première situation.

Une autre révolution qu'a apportée *H. Markowitz* au monde de la finance, lors de la publication de son article en 1952, *Portfolio Selection* [Markowitz, 1952]

est le principe de *diversification d'un portefeuille*. Ce principe repose sur l'idée que le niveau du risque d'un portefeuille dépend des corrélations entre les titres le constituant et non pas du taux de risque moyen entre tous les titres, pris individuellement. Avant la publication de l'article, la norme était plutôt d'investir,séparément, dans des titres qui offraient la future plus haute valeur au regard de leur prix actuel. [Kolm et al., 2014]

2.4.1 Le concept du compromis rendement/risque d'un point de vue mathématique

Considérons un monde d'investissement dans lequel on retrouve un vecteur $\vec{n} = (A_1, A_2, ..., A_N)$ de N différents titres (*remarque :* dans ce mémoire, nous allons exclusivement nous concentrer sur les actions, mais un portefeuille peut tout aussi bien être constitué d'autres types d'actifs tels que des obligations, des options, etc), donc un portefeuille pourrait être constitué de tout actif N. Une stratégie de portefeuille ou simplement un portefeuille $P(w)$ sera représenté par un vecteur de N-dimensions $\vec{w} = [w_1, w_2, ..., w_N]$. En fait, chaque actif constituant le portefeuille est pourvu d'un poids w_i représentant une proportion d'une certaine richesse allouée à l'actif a_i. Il est également important de savoir que $\sum_{i=1}^{N} w_i = 1$

Considérant le fait qu'un portefeuille peut être constitué de N différents titres, le return/rendement réel qu'un portefeuille est capable d'octroyer dépend du poids w_i alloué au titre a_i ainsi que du return/rendement de chaque titre. Supposons maintenant que le vecteur $\vec{R} = (R_1, R_2, ..., R_N)$ représente le rendement des différents titres et que ces différents rendements sont pourvus d'un taux de rendement prévu (traduction de l'expression anglaise *expected rate of return*), $\mu = (\mu_1, \mu_2, ..., \mu_N)$. Nous sommes maintenant capables de montrer que le taux de rendement prévu/attendu du portefeuille $P(w)$ est la moyenne pondérée des rendements prévus/attendus des titres composant le portefeuille $P(w)$. Le rendement incertain du portefeuille $P(w)$, dénoté R_P, dépend linéairement des poids w_i alloués aux différents titres.

Nous pouvons donc écrire que :

$$R_P(w) = w_1R_1 + w_2R_2 + ... + w_NR_N = \sum_{i=1}^{N} w_iR_i \tag{2.6}$$

Le rendement espéré du portefeuille $P(w)$, dénoté μ_P est donc :

$$\mu_P = \sum_{i=1}^{N} w_i \, \mu_i \tag{2.7}$$

Dans le reste de ce mémoire, nous utiliserons la notation matricielle, qui dans le cas des équations 2.6 et 2.7 s'écrivent respectivement :

$$R_P = w^{'} R \tag{2.8}$$

$$\mu_P = w^{'} \mu \tag{2.9}$$

Comme mentionné plus haut, un investisseur, dans un monde d'investissement suivant parfaitement la théorie de *H.Markowitz* doit s'intéresser au rendement espéré d'un portefeuille mais également au risque associé à ce dernier (c'est pourquoi on parle d'un univers de compromis entre rendements et risques, *[risk-return trade-off en anglais]). Markowitz* considère la variance d'une variable aléatoire (à savoir le rendement d'un portefeuille est une variable aléatoire) comme une mesure du risque associé à un portefeuille $P_(w)$, que nous noterons σ^2. le montrendant, il est impératif de noter qu'un portefeuille est très souvent constitué de deux ou plusieurs titres. Dans ce cas, dans le cadre de l'élaboration d'un portefeuille suivant la théorie de *H.Markowitz*, il nous sera nécessaire de calculer la dépendance mutuelle des différents titres. Cette dernière peut être obtenue grâce à la **covariance**. La covariance entre le rendement du titre i, R_i, et le rendement du titre j, R_j peut s'écrire sous la forme :

$$\sigma_{i,j} = \rho_{i,j}\, \sigma_i\, \sigma_j \tag{2.10}$$

où $\rho_{i,j}$ représente le coefficient de corrélation entre R_i et R_j. Dans le cas d'un monde d'investissement constitué de N titres, la matrice des covariances (ou la matrice variance-covariance) entre les rendements des N titres sera donc représentée par une matrice de NxN dimensions que nous pouvons écrire comme suit :

$$\sum = \begin{bmatrix} \sigma_{11} & ... & \sigma_{1N} \\ ... & ... & ... \\ \sigma_{N1} & ... & \sigma_{NN} \end{bmatrix} \tag{2.11}$$

Nous pouvons donc écrire l'équation de la variance d'un portefeuille $P(w)$ telle que :

$$\sigma^2_{P(w)} = w' \sum w \tag{2.12}$$

Nous sommes maintenant en mesure de formuler un problème d'optimisation

dans lequel un investisseur désire minimiser le risque tout en maximisant le rendement du portefeuille dans lequel il a investi. Ce problème peut se formuler de la façon suivante :

$$\begin{aligned} \min_{w} \quad & \sigma^2_{P(w)} = w' \sum w, \\ \text{s.c.} \quad & w' \mu = r \\ & w' \mathbf{1} = 1. \end{aligned} \tag{2.13}$$

où $\mathbf{1}$ représente le $(Nx1)$ vecteur de 1 et r est un certain niveau de risque.

Dans ce mémoire, nous utiliserons uniquement cette formulation de minimisation et non pas la formulation qui consisterait en une maximisation du rendement, même si ces deux conceptions d'un problème d'optimisation sont équivalentes. Dans le cas qui nous intéresse, la fonction objective est linéaire tandis que les contraintes sont quadratiques.

2.4.2 Les limites et désavantages

Malgré le fait que l'article de **Markowitz**, *Portfolio Selection* [Markowitz, 1952] ait indéniablement eu un impact majeur, tant dans le monde académique que dans celui de la finance dans son entièreté et que la technique de construction de portefeuilles découlant de ladite *modern portfolio theory* démontre une certaine simplicité et un attrait intuitif, beaucoup de professionnels de la gestion de portefeuille sont restés indifférents pendant de nombreuses années à la mise en pratique du type d'optimisation de portefeuille proposé par **Markowitz**, et ce pour différentes raisons.

Selon [Michaud, 1989], la première raison est , d'ordre "politique". En effet, l'utilisation d'un outil d'optimisation par une société (qui peut être par exemple, un hedge fund, un gestionnaire d'actifs, une banque, une compagnie d'assurances , etc) requiert des changements dans la structure de l'organisation et du management du processus d'investissement habituel. Avec ce genre de technologies, la comité d'investissement de la société ne prend plus à lui seul les décisions clés et finales concernant la politique d'investissement. Cela enlève de la valeur ajoutée au métier des *stocks pickers*[3].De plus l'introduction d'un outil d'optimisation peut également mener au développement de stratégies d'investissements basées sur plus d'élements quantitatifs sans tenir compte de l'aspect humain, ce qui peut entrainer des ajustements non-souhaités dans la politique d'investissement de la société. En particulier car

3. Un personne du style de Warren Buffet, qui par expérience acquise sait par un feeling quelles actions sont à acheter, à vendre ou encore à conserver

la mise en place d'un algorithme d'optimisation de portefeuille nécessite une aptitude à comprendre la signification financière des caractéristiques statistiques d'un portefeuille, ce qui mène de fait, les personnes ayant un profil plus quantitatif a joué un rôle de plus en plus central au sein de la politique d'investissement. Ce qui ne réjouit pas les investment managers habituels qui sont plus enclins à garder un grand pouvoir de décisions dans la stratégie d'investissement plutôt que de voir des "quants" ou des modèles mathématiques usurper leur rôle et surtout leurs pouvoirs.

La deuxième raison est plutôt d'ordre "technique". Les professionnels du milieu ont en effet observé que les outils d'optimisation de portefeuille basés directement sur la *Modern Portfolio Theory*, sans avoir recours à diverses adaptations, ont tendance à être des outils qui se révèlent peu fiables dans la pratique [Kolm et al., 2014]. Ce manque de fiabilité est le résultat du mélange de plusieurs failles dans la théorie proposée par **Markowitz**. En effet le concept d'optimisation *"rendement - risque"*, (*mean-variance optimization*), est doté de plusieurs défaillances qui mènent souvent à des portefeuilles optimaux totalement dénués de sens financièrement. Parmi ses défaillances, retenons principalement les points suivants :

- *L'optimisation rendement - risque est très sensible aux erreurs d'estimation des données d'entrée* μ_i, σ_i *et* $\sum_i$ [Chopra and Ziemba, 1993] , ont effectivement démontré qu'un petit changement dans les paramètres d'entrée peut mener à de grands changements dans la composition du portefeuille optimal et donc dans les poids w_i accordés à chaque actif i composant l'univers d'investissement pris en compte. Effectivement, selon [Kolm et al., 2014] beaucoup de professionnels considèrent ces données de sorties résultant de l'optimisation *rendement-risque* comme étant opaques et non-intuitives.
- *L'optimisation rendement - risque a tendance à maximiser les erreurs :* Selon, [Michaud, 1989], les outils d'optimisation *rendement - risque* sont perçus comme des *maximisateurs d'erreurs d'estimation*. Le concept d'optimisation *rendement-risque* tend à surpondérer les titres i ayant de gros rendements estimés et une faible variance (un faible risque donc) comparé à la pondération du benchmark. Une conséquence de ce fait est que l'on peut se retrouver face à des erreurs d'allocations de titres. En effet, les portefeuilles optimaux issus des problèmes d'optimisation *rendement - risque* ont fortement tendance à être constitués de poids w_i fortement positifs et négatifs, ce qui signifie que c'est un portefeuille constitué à la fois de positions extrêmement "long" et de positions extrêmement "short". Ce qui s'avère être loin d'être optimal.

— *Ce type de portefeuille est instable :* Une des raisons de cette instabilité est le mauvais conditionnement de la matrice de covariance.
— Ce concept d'optimisation de portefeuille ne prend pas en compte les moments statistiques d'ordre supérieur, qui sont *le skewness* et le *kurtosis*.

Pour plus d'informations et de détails au sujet des limites du portefeuille de **Markowitz**, il est conseillé aux lecteurs de consulter [Michaud, 1989].

2.4.3 Les améliorations

Les limitations exposées dans la sous-section précédente ne doivent pas être vues comme un signe que l'optimisation de portefeuille *rendement - risque* est défectueuse mais plutôt comme le fait que l'approche originale de **Markowitz** doit être considérée uniquement comme un point de départ et que cette approche nécessite certaines extensions afin d'obtenir des portefeuilles optimaux stables, robustes et fiables. Ces extensions peuvent inclure, parmi d'autres :

— L'introduction des *coûts de transaction* dans le problème d'optimisation. [Litterman, 2003]
— L'addition de plusieurs nouvelles contraintes dans le problème d'optimisation de base, telles que par exemple des contraintes liées au *short-selling*. Le lecteur intéressé trouvera des informations détaillées à ce sujet dans [Jagannathan and Ma, 2003]
— La modélisation et la quantification de l'impact des erreurs d'estimation des rendements espérés caractérisés par le moment statistique μ_i, ainsi du risque caractérisé à son tour par $\sum_i$ et σ_i. Dans ce mémoire nous nous concentrerons sur ce type d'extensions et en particulier sur celle apportée par **Ledoit & Wolff**. Celle-ci sera abordée et développée dans la sous-section suivante. Voir [Ledoit and Wolf, 2004a] ainsi que [Ledoit and Wolf, 2003] et [Ledoit and Wolf, 2004b].

2.4.4 L'estimateur de Ledoit & Wolff - Shrinkage

Concept général

Dans la théorie moderne de gestion de portefeuille, développée par **H. Markowitz**, c'est à dire dans le contexte d'un problème d'optimisation dit *"rendement-risque"*, une estimation des rendements espérés ainsi que des covariances de tous les titres composant l'univers d'investissement dans lequel nous travaillons est requise. le montrendant, dans l'industrie de la gestion de portefeuille, la plupart des portfolio managers ont leur propre expertise

dans un secteur ou une industrie précise. Il est donc irréaliste de penser que ces professionnels d'un secteur seraient capables de fournir des estimations convenables dans tous les domaines, ce qui augmente, de fait,les erreurs d'estimation. Pour rappel, le lecteur a vu dans la sous-section 2.4.2, ci-dessus, une des raisons pour lesquelles l'optimisation de portefeuille *rendement - risque* a été si peu adoptée, dans sa forme initiale, au sein de l'industrie financière.

Afin de remédier à cela et au fait que le concept d'optimisation de portefeuille *rendement-risque* est un "maximisateur d'erreurs d'estimations", plusieurs techniques ont été mises au point pour estimer correctement les données d'entrée d'un problème d'optimisation de portefeuille.

Dans ce mémoire, nous développerons une technique d'estimation de la matrice de covariance établie par les chercheurs **Olivier Ledoit** et **Michael Wolf**. Ceux-ci ont proposé une formule pour estimer la matrice de covariance des rendements de titres qui peut remplacer de manière très bénéfique la matrice de covariance, estimée de façon standard, dans tout problème d'optimisation du type *rendement - risque*. L'estimateur découlant de cette technique est appelé : *shrinkage estimator* de la matrice de covariance (Par choix, nous garderons sa dénomination anglaise dans ce mémoire).

L'idée sous-jacente de cet estimateur est que les coefficients de la matrice de covariance, estimés de façon standard, qui sont extrêmement hauts, ont tendance à contenir beaucoup d'erreurs positives et ont donc besoin d'être réduits, tirés vers le bas, pour compenser. A l'inverse, les coefficients extrêmement bas ont tendance à contenir des erreurs négatives, ils doivent donc quant à eux être augmentés, tirés vers le haut. **Ledoit** et **Wolf** ont appelé cela le"*shrinkage* des valeurs extrêmes vers le centre".

Toute construction de *skrinkage estimators* doit impérativement être pourvue de 3 éléments :

1. Un estimateur sans structures
2. Un estimateur avec beaucoup de structures (que nous dénommerons dans ce mémoire *shrinkage target*). Ce type d'estimateur ne contient pas beaucoup d'erreurs d'estimations mais à tendance à être mal défini et très biaisé
3. Une constante de "*shrinkage*"

La principale force d'un *shrinkage estimator* est qu'il va combiner 2 estimateurs extrêmes pour obtenir un estimateur, qui sera une sorte de "compromis" et qui fonctionnera mieux que l'un ou l'autre extrême.

Le principe de construction de l'estimateur de **Ledoit** and **Wolf** peut être défini comme suit : il faut considérer d'une part la matrice de covariance, estimée de façon standard (*sample covariance matrix*) que nous dénommerons ici S. Cette matrice est facile à calculer et a la propriété d'être non-biaisée, ce qui signifie que sa valeur estimée est égale à la vraie matrice des covariances. le montrendant elle comprend beaucoup d'erreurs d'estimation lorsque le nombre de points de données est comparable ou même inférieur au nombre de titres, ce qui est une situation habituelle dans les applications financières. D'autre part, il faut également prendre en compte une *shrinkage target*, que nous dénommerons ici F. Le but de la construction est de trouver une combinaison linéaire convexe telle que $\delta F + (1-\delta)S$ où δ est un nombre $\in \{0,1\}$ et est la constante de *shrinkage*. Cette dernière mesure le poids donné à l'estimateur pourvu de beaucoup de structures. Le choix de cette valeur peut s'avérer être un problème car le choix de cette constante doit donner lieu à un compromis entre S et F. le montrendant il existe une constante optimale, δ^*, qui est celle qui minimise la distance entre le *shrinkage estimator* et la vraie matrice de covariance. Le plus compliqué,dans une situation mettant en action des *shrinkage estimators*, est de choisir la *shrinkage target* et la *constante de shrinkage*. La sous-section suivante est consacrée à la manière de choisir ces deux mesures.

Le choix de la *shrinkage target*

Une "*shrinkage target*" doit impérativement répondre à 2 conditions en même temps. La target n'implique qu'un nombre restreint de paramètres libres mais reflète également des caractéristiques importantes de la quantité inconnue qui doit être estimée. Dans ce mémoire nous utiliserons, comme *shrinkage target* le **constant correlation model**.[4] En utilisant ce dernier, l'estimation du modèle est claire. En effet, la moyenne de toutes les corrélations estimées est un estimateur de la corrélation commune.

Les éléments de la *shrinkage target*, F, qui est ici la matrice des corrélations estimées constantes peuvent être définis comme suit :[5]. Soit un actif (dans notre cas une action) i et un actif j

4. Ledoit & Wolff ont pour habitude d'utiliser comme estimateur le "single-factor matrix de Sharpe. L'estimateur que nous utiliserons a des performances similaires mais est plus facile à implémenter

5. Pour le développement complet, se référer de la page 12 à la page 15 de [Ledoit and Wolf, 2003]

$$f_{ii} = s_{ii} \; et \; f_{ij} = \overline{r}\sqrt{s_{ii}s_{jj}} \tag{2.14}$$

où

s_{ij} représente les éléments d'entrée de la matrice S,

$\overline{r}$ représente la moyenne des corrélations estimées

r_{ij}, la corrélation estimée entre l'action i et l'action j

Le choix de la *constante de shrinkage*

Comme mentionné, cette *constante de shrinkage optimale*, δ^*, sera un nombre $\in \{0, 1\}$ qui minimisera la distance attendue entre le *shrinkage estimator* et la vraie matrice de covariance. Cette *constante de shrinkage* ou *shrinkage intensity* sera obtenue par une minimisation d'une fonction de perte qui n'implique pas l'inverse de la matrice de covariance. Afin d'obtenir cette valeur optimale et sous les hypothèses que N (le nombre d'actions avec lequel nous travaillons) est fixe et que T, l'horizon temporel tend vers l'infini, **Ledoit & Wolff** ont développé la formule suivante : [6]

$$\widehat{\delta^*} = max\left\{0,\; min\{\frac{\widehat{\kappa}}{T},\; 1\}\right\}. \tag{2.15}$$

Sous les hypothèses exposées ci-dessus, **Ledoit & Wolff** ont prouvé que δ^* se comporte asymptotiquement comme une constante sur T jusqu'aux termes d'ordre supérieur. Cette constante, symbolisée par κ peut s'écrire mathématiquement comme :

$$\kappa = \frac{\pi - \rho}{\gamma} \tag{2.16}$$

où

π représente la somme des variances asymptotiques des inputs de la matrice de covariance estimée échelonnée par $\sqrt{T}$

ρ représente la somme des covariances asymptotiques des inputs de la *shrinkage target* échelonné par $\sqrt{T}$

6. Un développement complet de cette formule est consultable à la page 12 de [Ledoit and Wolf, 2003]

γ représente l'ensemble des spécifications erronnées de la *shrinkage target.*[7]

La formule du *shrinkage estimator* de Ledoit & Wolff

Sur base des différentes mesures exposées ci-dessus, le *shrinkage estimator* de **Ledoit & Wolff** peut être représenté mathématiquement de la façon suivante :

$$\widehat{\sum}_{Shrink} = \widehat{\delta}^* F + (1 - \widehat{\delta}^*) S \tag{2.17}$$

2.5 Le portefeuille disposant de la variance minimale globale

Ce portefeuille, pouvant être localisé sur le point le plus bas de la frontière efficiente (voire Figure 2.1, et plus précisément le point orange), représente comme son nom l'indique le portefeuille avec la plus petite variance parmi tous, (en anglais, *global minimum variance portfolio*).

D'après l'article *Estimating the Global Minimum Variance Portfolio*, rédigé par [Memmel and Kempf, 2006], de nombreuses études empiriques permettent de démontrer qu'investir dans un portefeuille ayant *la variance globale minimale* permet d'obtenir de meilleurs rendements, hors échantillon, qu'un investissement dans le portefeuille dit *tangent* (voire *figure* 2.1). Ce postulat est lié au risque élevé d'estimations associé au rendement espéré (*cfr sous-section les limites et désavantages*). Plusieurs articles, dont [Ledoit and Wolf, 2003], suggère donc d'investir dans le portefeuille disposant de *la variance minimale globale* plutôt que dans le portfolio dit *tangent.*

Le portefeuille disposant de *la variance minimale globale* est en fait la solution du problème d'optimisation, ou plutôt de minimisation suivant :

$$\begin{aligned} \min_{w} \quad & \sigma^2_{P(w)} = w' \sum w, \\ \text{s.c.} \quad & w' \mathbf{1} = 1. \end{aligned} \tag{2.18}$$

7. Le lecteur intéressé par les formules de ces 3 valeurs (π, ρ, γ) pourra se référer à la page 13 de [Ledoit and Wolf, 2003]

Les poids $w_{MV} = (w_{MV,1}, ..., w_{MV,N})$ du portfeuille disposant de *la variance minimale globale* peuvent être obtenus grâce à la formule suivante :

$$w_{MV} = \frac{\sum^{-1}}{\mathbf{1}' \sum^{-1} \mathbf{1}} \tag{2.19}$$

Nous sommes également en mesure de formuler l'équation définissant le rendement espéré de ce portefeuille μ_{MV} ainsi que celle de la variance de ce dernier. Ces équations prennent donc respectivement la forme suivante :

$$\mu_{MV} = \mu' \, w_{MV} = \frac{\mu' \sum^{-1} \mathbf{1}}{\mathbf{1}' \sum^{-1} \mathbf{1}} \tag{2.20}$$

$$\sigma^2 = w'_{MV} \sum w_{MV} = \frac{1}{\mathbf{1} \sum^{-1} \mathbf{1}} \tag{2.21}$$

Il est ici important de noter que la variance **minimale globale** requiert la connaissance parfaite de la matrice de covariance des rendements $\sum$. Or, nous ne connaissons pas à l'avance cette matrice et nous devons donc l'estimer. Une présentation d'un estimateur de la matrice de covariance performant et reconnu est présenté dans la sous-section 2.4.4.

Chapitre 3

Le Machine Learning

Selon Denis Panel, chief investment officer de BNP Paribas Asset Management dans [Lovell and Kempf, 2019]. "Le futur de l'industrie de l'asset management résidera dans la combinaison entre les approches fondamentales et les techniques quantitatives". En effet, par l'avènement du *big data* et la puissance de calcul toujours plus grande des ordinateurs, on observe une importance grandissante donnée aux algorithmes et à l'intelligence artificielle dans la totalité de l'industrie financière. Une étude de *Morningstar*, relayée dans le *Trends-Tendance* du 28 Novembre 2019 [Thomas, 2019] peut témoigner de cette évolution. Celle-ci nous renseigne sur le fait que depuis l'automne 2019, et pour la première fois dans l'histoire, les fonds gérés par des algorithmes, c'est à dire des fonds pour lesquels la stratégie n'est plus déterminée par l'homme, ont dépassé, en terme de sommes gérées, les fonds traditionnels, où le gestionnaire choisit lui-même les titres dans lesquels il investit. Aux Etats-Unis, plus de la moitié des fonds sont désormais gérés de la sorte, ce qui représente une masse de 4.300 milliards de dollars.
Le lecteur aura donc souligné l'importance capitale des outils quantitatifs (dont le machine learning fait partie) dans le monde financier actuel et plus particulièrement dans l'asset management.

3.1 Définition du Machine Learning

Le Machine Learning (qui se traduit, en français, par apprentissage automatique ou apprentissage de machine) est un sous-domaine d'une technologie très en vogue de nos jours, *l'Intelligence Artificielle - IA* [1]. le montrendant,

1. L'intelligence artificielle est définie dans le Larousse comme étant l'ensemble des techniques mises en oeuvre en vue de réaliser des machines capables de simuler l'intelligence humaine.

là où le focus de l'*IA* est de rendre les machines "intelligentes", qu'elles soient capables de raisonner de façon rationnelle comme les humains et de résoudre des problèmes, le Machine Learning quant à lui, se préoccupe plutôt de créer des systèmes informatiques ainsi que des algorithmes, de sorte que les machines apprennent, s'adaptent et changent leur comportement par rapport aux données, à leurs expériences précédentes, accumulées dans le but d'améliorer au fil du temps la performance des machines [Izenman, 2013]. C'est avec des applications, telles qu'évoquées ci-dessus, que l'on peut se rendre compte de l'importance du *big data*, données existant en très grand volume, dans de très nombreuses variétés différentes et qui doivent être traitées à une vitesse très élevée (Velocity) [2]. Il convient également d'exposer le fait que le Machine Learning est usuellement divisé en 3 catégories d'apprentissage :

L'apprentissage supervisé : Dans cette catégorie, l'algorithme, que l'on appelle un algorithme d'apprentissage, reçoit un ensemble de données d'entrée ou inputs labellisés/étiquettés (variables pouvant être continues ou catégorielles) ainsi qu'une variable de sortie ou output correct. Les différentes catégories sont également définies au préalable. Le but de ce type d'apprentissage est de tenter de trouver une fonction des variables d'entrée pour approximer la variable de sortie connue. On pourrait donc illustrer ces dires par une fonction qui prendra la forme qui suit : $Y = f(X)$ où X représente l'ensemble des données d'entrée et Y les variables de sortie. Si la variable de sortie est continue, cela donnera un problème de régression alors que si elle est catégorielle, cela donnera un problème de classification.

Attardons-nous un peu plus en profondeur sur ce type de problème, dit de **classification**, étant donné que c'est avec celui-ci que nous réaliserons nos analyses. Dans un problème de Machine Learning de ce type, le but principal est de trouver une façon systématique de classer une nouvelle donnée, qui appartient à l'ensemble de données d'entrées X, dans une catégorie définie au préalable. La classification de cette nouvelle donnée entrante se basera sur les connaissances que le modèle a pu acquérir à partir d'un échantillon d'apprentissage formé de données similaires. Dans cette échantillon d'apprentissage, les différentes classes sont évidemment déterminées et connues et le nombre de ces dernières est fini et connu [Collard, 2019]. La figure 7.4 montre les différents types d'algorithmes utilisés dans ce type de problème d'apprentissage.

2. Le *big data* respecte la règle dite des "5V's"- Voire l'**Annexe 4**

L'apprentissage non-supervisé : Dans ce type de problème d'apprentissage, il y a des données d'entrée mais il n'y a pas d'informations disponibles pour définir une variable de sortie appropriée. Dans cette catégorie seront explorées différentes caractéristiques des variables d'entrée telles que la localisation de valeurs aberrantes ou encore une estimation de la densité de probabilité jointe. Une fois que le modèle d'apprentissage non-supervisé aura repéré les similitudes (grâce aux différentes caractéristiques des variables d'entrée), il pourra les regrouper par groupe ou classe. Il est noté ici que les modèles dit de *clustering* font partie de cette catégorie. Une liste des différents algorithmes utilisés dans cette catégorie est présentée dans la figure 7.4, se trouvant dans l'Annexe 5.

L'apprentissage par renforcement - Reinforcement Learning : Il s'agit ici d'un domaine d'apprentissage qui s'intéresse à la façon dont des agents devraient agir dans un environnement, en fonction de leur état courant, afin de maximiser une certaine notion de récompenses (qui peuvent être soit positives, soit négatives). L'agent cherche, au travers d'expériences itérées, un comportement décisionnel (appelé politique ou stratégie), qui est une fonction associant à l'état courant l'action à exécuter, optimal au sens où il maximise la somme des récompenses au cours du temps. Il s'agit donc de trouver un équilibre entre l'exploration de territoires inexplorés et l'exploitation des connaissances actuelles. Il est noté que pour ce type d'apprentissage, *Les processus de décision de Markov*[3] sont fortement utilisés. On peut par exemple retrouver ces types de problèmes d'apprentissage dans le contrôle des robots ou encore dans le jeu de Go.[4]

Notons pour terminer cette section, que pour rappel dans le cadre de ce mémoire, nous ne travaillerons qu'en apprentissage supervisé et plus précisément avec des problèmes de classification. Les autres catégories d'apprentissage et problèmes tels que la régression ou le clustering n'ont été abordés dans ce travail qu'à titre informatif.

3. La définition d'un processus de décision de Markov est donné à la sous-section 1.2.1 dans [PDMIA, 2008]

4. AlphaGo, modèle qui a battu le meilleur joueur de Go du monde est une application du reinforcement learning.

3.2 L'utilisation du Machine Learning en Finance

Machine Learning in finance - 15 applications for data science aspirants[5]

3.3 Focus sur quelques applications du Machine Learning développées par des acteurs majeurs de l'industrie financière

Ce mémoire se limitera à l'approche des applications dans le secteur de l'asset management et de la gestion de portefeuille. De plus, la liste d'exemples choisie ci-après est non-exhaustive, de nombreuses institutions ont développé des outils de Machine Learning pour leur division s'occupant de l'asset management. Les exemples qui suivent sont exposés afin que le lecteur ait une idée d'applications de Machine Learning dans ce secteur ainsi que de la complexité et l'importance grandissante de cette technologie en croissance dans l'industrie financière et plus précisément dans l'asset management.

3.3.1 L'utilisation du Machine Learning chez Goldman Sachs

Osman Ali, portfolio manager pour les *Quantitative Investment Strategies* chez Goldman Sachs Asset Management (GSAM), pense que le Machine Learning, dans le secteur de l'investissement, est un outil puissant qui aide la branche de gestion d'actifs de la banque d'affaire à analyser de très grands, déstructurés et complexes ensembles de données, afin d'en extraire des informations précieuses [Ali, 2017]. Son institution, Goldman Sachs, utilise le Machine Learning, et plus précisément une forme de Machine Learning appelé *Natural Language Processing (NLP)* avec des données dites alternatives[6] pour remplir 3 objectifs principaux :

— **Mieux comprendre les sentiments qui prévalent au sein des marchés :** Goldman Sachs a développé des algorithmes de Machine Learning capables de lire et d'analyser des ensembles de données alternatives tels que des articles et des communiqués de presse, des publications de résultats trimestriels ou encore des transcripts d'*earnings*

5. https ://data-flair.training/blogs/machine-learning-in-finance/
6. telles que des données de transactions de cartes de crédit, des images satellites, ...

calls.[7] Le but ici est de vraiment comprendre les nuances du phrasé des analystes par rapport à leurs habitudes. Par exemple, si un analyste complimente le management d'une société plus qu'à son habitude, cela veut dire, selon Goldman Sachs, que l'analyste révèle inconsciemment son optimisme par rapport à cette société. Il est important de noter ici, que la banque d'affaire réalise ces opérations pour des centaines d'analystes qui suivent des dizaines de milliers d'entreprises, tout au long d'une année. Tout cela est mis en place afin de détecter le moindre changement de sentiment et de prendre l'avance sur le marché. [Staff, 2019a]

— **Identifier les liens intersociétés :** Toujours en analysant des articles et communiqués de presse, des recherches d'analystes mais aussi des transcripts de journaux télévisés ou des dépôts de brevets, les algorithmes développés par la banque d'affaire sont capables de détecter des liens inter-entreprises en examinant la fréquence à laquelle certaines entreprises sont citées ensemble dans les documents analysés. Il existe en effet des milliers de possibilités de liens différents entre entreprises, au-delà du plus familier, à savoir la relation client-fournisseur. La détection de ces liens est importante, étant donné que les cours de bourse de sociétés liées peuvent avoir une influence sur l'une ou sur l'autre. [Ali, 2017]

— **Extraction d'un thème d'investissement :** Avec les mêmes données alternatives que celles citées précédemment, les algorithmes peuvent également détecter quels sont les sujets actuels d'investissement pertinents.

Mentionnons que Goldman Sachs Asset Management a reçu l'award de *"Asset Manager of the year 2019*, décerné par le Risk Magazine[8], pour ses avancées innovantes en terme de Machine Learning dans l'asset management.

3.3.2 L'utilisation du Machine Learning chez Bank of America

Avec 11 milliards de budget consacré à l'IT annuellement, il n'est pas étonnant que Bank of America (BofA) prévoit une partie de ce dernier à la mise en oeuvre d'algorithmes de Machine Learning dédiés à la gestion de portefeuille. Parmi les premiers produits émanant de techniques de Machine

7. **Earnings Calls** : téléconférence ou webcast dans lequel une entreprise cotée en bourse discute des ses résultats financiers d'une certaine période avec des analystes. Source : https ://en.wikipedia.org/wiki/Earnings$_c all$

8. le lecteur intéressé pourra consulter [Staff, 2019a]

Learning utilisé, nous pouvons en mettre un en lumière, qui dans le cadre de ce mémoire, se révèle des plus intéressant.

En effet, le *digital innovation group* de la BofA a développé, à l'aide du Machine Learning, un outil capable de contrer un problème bien connu du monde de l'asset management, à savoir la "malédiction" de Markowitz [9]. Cet outil, appellé chez BofA *Dynamically Diversified Momentum (DDM) index*, s'enorgueillit de résoudre ce probléme en utilisant des techniques de clustering, visant à diviser l'univers d'investissement en groupes porteurs de niveaux de risque similaires, de relâcher la dépendance aux estimations de corrélation ainsi que de découvrir l'alpha [10] caché. La stratégie de gestion de portefeuille découlant du *DDM* montre un rendement annualisé de 4.6 %.

Alors que d'autres institutions ont vu beaucoup de décaissements de leurs stratégies quantitatives (Quantitative Investment Strategies - QIS), une innovation comme exposée ci-dessus a permis à la Bank of America d'augmenter ses actifs sous gestion dans ce type de business basé sur des méthodes quantitatives ainsi que pour ses activités de dérivés d'actions sur indices investissables. Un client de la banque a d'ailleurs jugé ce produit "extraordinaire". [Staff, 2019b]

9. cfr la sous-section 2.4.2 de ce mémoire

10. **L'Alpha** est un indicateur de performance pour un fond ou un actif par rapport à un indice de référence. Cette mesure permet de voir si un fond/un acif bat son indice ou se fait battre par celui-ci. Par exemple, si l'alpha est de 2, cela voudrait dire que le fond bat son indice de 2%

Deuxième partie

Partie empirique

Chapitre 4

Données et Méthodologie

Dans ce chapitre, nous décrirons la base de données ainsi que la méthodologie utilisée afin de réaliser notre étude, qui consiste, pour rappel, à établir une méthodologie de construction de portefeuilles sur base d'un modèle de Machine Learning et plus précisément d'un algorithme de classification avec aspect probabiliste et d'en analyser ses performances.

4.1 La base de données utilisées

Dans ce mémoire et afin de tester et de comparer les performances de nos différents portefeuilles construits, il a été décidé de baser notre analyse sur les valeurs boursières des titres composant le **CAC40**. Ce dernier est un indice qui reflète la performance des 40 actions [1] d'entreprises les plus importantes et les plus activement négociées sur Euronext Paris. [2] le montrendant, dans ce mémoire et au regard des périodes temporelles étudiées [*cfr.* point suivant], nous n'exploiterons pas les rendements des 40 actions les plus importantes de la bourse de Paris mais celui des **39** plus importantes. En effet, l'entreprise *TechnipFMC*, issue de la fusion entre l'entreprise française *Technip* et l'américaine *FMC Technologies*, n'a débuté sa cotation sur Euronext Paris que le 17 janvier 2017 [3], ce qui dans le cadre de notre étude, ne nous permet pas de collecter les données nécessaires. Comme conséquence, nous avons donc chosi de limiter le nombre d'entreprises étudiées à **39**.

Afin de réaliser au mieux une étude de performance des portefeuilles, il a été décidé de travailler avec 2 périodes temporelles distinctes, à savoir une période

1. la composition de CAC40 se trouve à l'Annexe 6 de ce travail.
2. La Bourse de Paris fait partie de groupe boursier paneuropéen Euronext
3. Voir [zonebourse, 2017]

dite de "crise" (de **2007 à 2012**) et une période dite de "post-crise" (de **2013 à 2018**). Pour chacune de ces périodes distinctes, les données réparties entre les dates du **02 janvier** au **31 décembre** ont été collectées. Comme mentionné dans [Jagannathan and Ma, 2003], travailler avec les rendements quotidiens d'une action permet d'obtenir de meilleures estimations que si l'on utilisait comme données les rendements hebdomadaires ou mensuels. Dans ce mémoire, nous nous attacherons à suivre les recommandations des 2 auteurs précédemment cités. Seuls **les derniers prix** de chaque action pour chaque séance quotidienne durant les années précitées ont été téléchargés. Quant au rendement relatif d'un titre, nous pouvons le calculer de la manière suivante :

Soit un titre i et une journée n. Nous calculons donc le rendement relatif d'un titre i en divisant le prix de clôture (PC) en n par le prix de clôture en $n-1$ et nous soustrayons 1 du résultat obtenu, ce qui nous donne la formule suivante :

$$Rendement\ relatif\ en\ n\ = (\frac{PC(n)}{PC(n-1)}) - 1 \tag{4.1}$$

De plus, étant donné que l'un des objectifs de notre travail est d'appliquer des techniques de *Machine Learning* et plus précisément des techniques de *classification*, il nous a également été nécessaire de créer nous-même des classes (des catégories) dans lesquelles nous pouvons placer nos différents points de données. Pour ce faire, nous nous sommes référés à l'article de [Tilakaratne, 2004] et du mémoire de [Collard, 2019] , dans lequel il est précisé que l'on peut classer des titres par rapport à leur rendement. Pour cette étude, 3 classes ont été créées :

- La classe **"Vendre"** ; Sera attribuée aux titres dont le rendement relatif est inférieur à **-0.005**.
- La classe **"Conserver"** regroupera les titres dont le rendement relatif est situé entre **-0.005** et **0.005**.
- La classe **"Acheter"** est quant à elle destinée aux titres dont le rendement relatif est supérieur à **0.005**.

Les données des prix **quotidiens** des actions composant le CAC40 ont donc été réunies dans un fichier *Excel*, ce qui représente donc **59943** observations des derniers prix pour la période dite de "crise" et **59748** pour la période dites de "post-crise". Au total, l'étude réalisée concerne **119691** données relatives au prix quotidien des titres composant l'indice CAC40. Ces différents

chiffres sont à multiplier par 3 étant donné qu'une colonne des données du rendement relatif et une de la classe (la catégorie) du titre ont été créées. Nous travaillerons donc avec **179829** données, pour la période dite de crise et **179244** pour la période dite de "post-crise". Dans un souci de clarté, il a également été décidé de travailler avec 2 bases de données distinctes suivant la période étudiée plutôt qu'avec une seule reprenant l'ensemble de ces dernières. Enfin, pour terminer, mentionnons ici que ces données ont été collectées grâce aux terminaux Bloomberg. [4]

4.1.1 Observations préliminaires

Tendances de l'indice CAC40 de 2007 à 2018

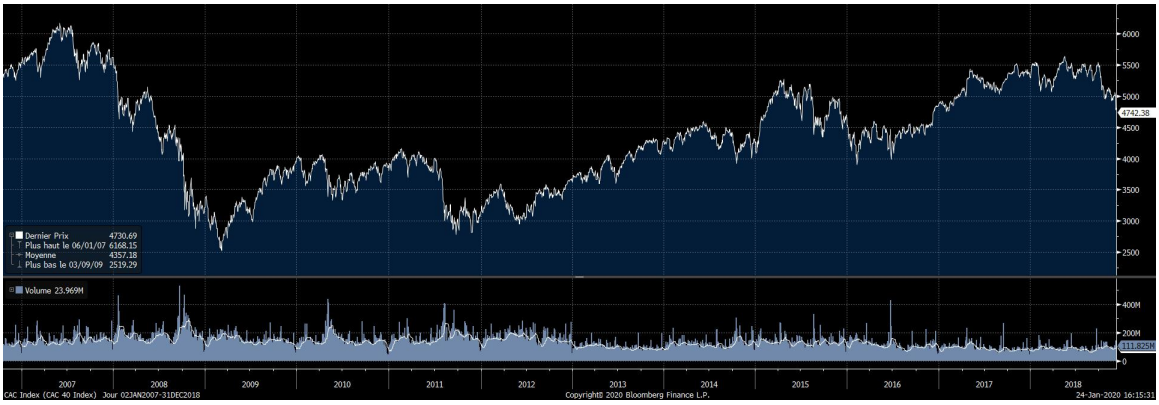

FIGURE 4.1 – Tendance du CAC40, 2007-2018

Dans le graphique représenté à la figure 4.1, on peut dans un premier temps observer une réelle chute de l'indice à partir de début 2008 jusqu'au premier quadrimestre de 2009. Cette période correspond à la crise financière mondiale de 2008. Deuxièmement, on peut également constater une deuxième chute du CAC40 aux environs de la fin du second quadrimestre de 2011, ce qui correspond à la période de crise de la dette de la zone euro. A partir de 2012, une remontée croissante de l'indice s'est amorcée, sans le montrendant atteindre les niveaux d'avant crise.

Une partie intéressante de cette étude sera justement de comprendre comment les différentes stratégies de gestion de portefeuille se comportent en ces périodes de crise. Finalement, une attention particulière sera portée au

4. accessibles à la Bibliothèque de la faculté des sciences économiques, sociales, politiques et de communication (BSPO) de l'Université Catholique de Louvain.

comportement de la stratégie de portefeuille issue du Machine Learning, afin d'observer si cette dernière est capable de s'adapter en période difficile.

4.2 La méthodologie

Dans cette section, nous nous attèlerons à expliciter les différents outils tant théoriques que de programmation [5], qui nous ont permis d'effectuer notre étude et donc de répondre à nos questions de recherche exposées dans l'introduction.

La méthodologie proposée dans ce mémoire est basée sur les différentes étapes que nous avons suivi afin de construire une stratégie de gestion de portefeuille basée sur un modèle de Machine Learning utilisant un algorithme de classification avec aspect probabiliste [*cfr.* sous-section - L'algorithme c5.0]. Chaque étape est scrupuleusement décrite et illustrée par des screenshots de code R afin que le lecteur comprenne réellement tant la réflexion présente derrière cette étape que la logique de programmation informatique. De plus, certains exemples d'outputs obtenus grâce aux algorithmes implémentants notre méthodologie sont également présentés afin que le lecteur ait une vision claire du résultat recherché par les différentes étapes. Pour rappel, le but de ce mémoire est de comprendre comment un portefeuille d'action est construit sur base d'un modèle de Machine Learning permettant de travailler avec des probabilités et quelles en sont les performances en période dites de "*Crise*" et de "*PostCrise*.

4.2.1 L' indicateur de performance utilisé

Dans ce mémoire, nous avons choisi d'utiliser un indicateur de performance qui permet de connaître, comme son nom l'indique, la performance du portefeuille construit. Pour ce faire, nous avons choisi d'utiliser le **Sharpe Ratio**. La théorie relative à ce dernier est développée dans la sous-section 4.2.2.

Dénomination	**Abréviation**
Ratio de Sharpe	SR

TABLE 4.1 – Indicateur de performance utilisé

5. le code relatif à l'analyse réalisée durant ce mémoire a été réalisé à l'aide de **RStudio**

4.2.2 Le ratio de Sharpe - "Sharpe Ratio" - SR

Développé par **William F. Sharpe** [Sharpe, 1994], cette mesure peut être vue comme une manière d'aider les investisseurs à observer le rendement d'un portefeuille (représenté par la moyenne des rendements des titres composant le portefeuille, μ_P) comparé au risque associé (σ_P). Plus simplement, c'est également une façon de résumer 2 mesures (μ_P et σ_P) en une seule, à savoir le **ratio de Sharpe**. Ce dernier peut être défini mathématiquement comme suit :

$$SR = \frac{\mu_P - R_f}{\sigma_P} \tag{4.2}$$

où R_f représente le taux sans risque (en anglais, risk-free rate), qui peut être défini comme le rendement théorique associé à un investissement garantissant un rendement avec un niveau de risque associé de 0. Afin d'illustrer cela, nous pouvons prendre comme exemple le cas des Etats-Unis, où le rendement associé aux *T-Bills* ou *treasury bills* est considéré comme un taux sans risque. Un ratio de Sharpe peut s'interpréter de la façon suivante :

- S'il est **négatif**, cela signifie que le taux sans risque est supérieur au rendement du portefeuille. Il s'agit clairement d'une mauvaise situation.
- S'il est compris **entre 0 et 1** cela signifie que l'excédent du rendement par rapport au taux sans risque est plus faible que le risque pris [Sharpe, 1994]
- S'il est **supérieur à 1**, cela signifie que le rendement du portefeuille est supérieur au taux sans risque. C'est une situation enviable étant donné que le portefeuille génère une rentabilité plus forte qu'un investissement sans risque.

le montrendant, il est également important de mentionner certaines limites associées à ce ratio. Dans un premier temps, ce dernier repose sur l'hypothèse que les rendements des titres composants un portefeuille sont normalement distribués. Deuxièmement, le **ratio de Sharpe** peut faire l'objet de manipulation de la part de portfolio managers désirant améliorer leurs propres historiques de rendements. Enfin, le ratio dont il est question ici ne prend en compte que les 2 premiers moments statistiques d'un portefeuille, ce qui signifie qu'il ne tient pas compte des autres différences possibles entre portefeuilles provenant des autres moments. [Sharpe, 1994]

4.2.3 Focus sur la construction de stratégies sur base du Machine Learning

Le but de notre étude, en plus de construire une stratégie de portefeuille sur base d'un modèle de Machine Learning, est d'exploiter, par un algorithme de classification, le niveau de probabilité dont est muni un point de données classé dans une classe définie au préalable (pour rappel, 3 classes ont été créées). Nous allons donc utiliser et implémenter un algorithme de classification [*cfr* section 2.2] qui peut nous fournir des niveaux de probabilités qu'un point de données soit classé dans une des 3 classes définies au préalable.

Afin de répondre à ce besoin, des recherches ont été effectuées et il s'est révélé le plus pertinent d'utiliser l'algorithme de classification appelé **C5.0** [6]. Celui-ci, en plus d'être en mesure de nous fournir un output probabiliste, est devenu la norme pour produire des arbres de décision car il fonctionne très bien pour la plupart des problèmes de classification et a des performances quasi similaires aux algorithmes de Machine Learning plus élaborés [Lantz, 2013].

Implémentation de l'algorithme de classification avec RStudio et ses outputs

Dans cette étude, chaque action a été traitée séparément, c'est-à-dire que nous avons implémenté 39 fois la méthodologie qui sera décrite ci-après. Nous obtenons donc 39 outputs différents, chacun étant attribué à une action composant le CAC40.

La méthodologie employée, afin de créer un modèle de Machine Learning nous permettant d'obtenir un output avec aspect probabiliste, est composée de 5 étapes qui seront expliquées ci-après.

6. Pour une définition plus formelle de cette algorithme, le lecteur est invité à se rendre à la page 124 de [Lantz, 2013]

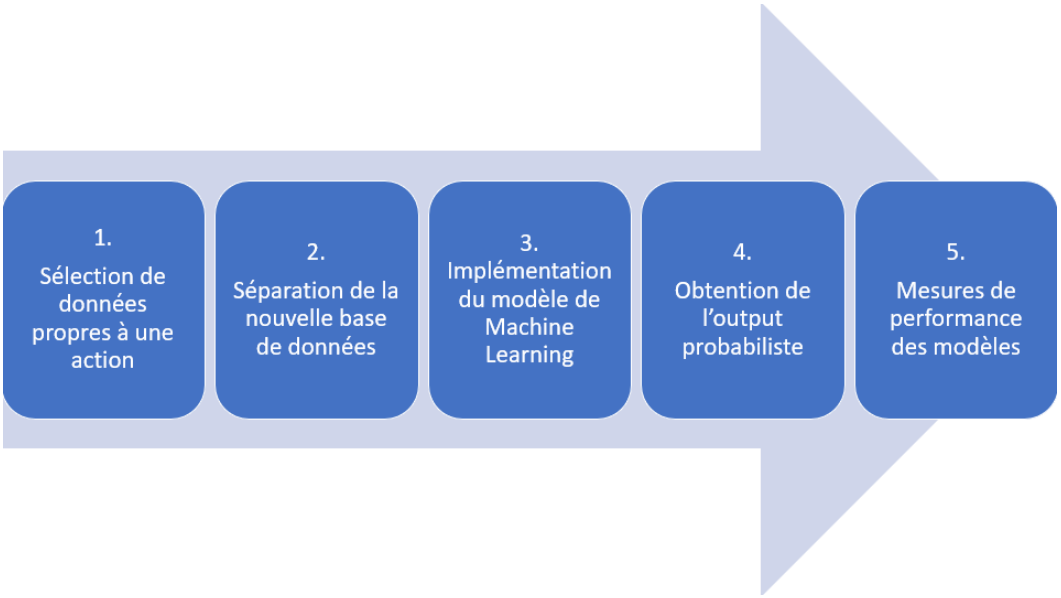

FIGURE 4.2 – Résumé des différentes étapes liées à l'implémentation d'un modèle de Machine Learning

1. **La sélection des données propres à une action :** Etant donné que nous travaillons action par action, nous devons sélectionner dans notre base de données globale les 3 colonnes relatives à une action.

   Les 2 premières colonnes (regroupées dans un même objet au sens du language de programmation R) représentent les 2 variables *explicatives* utilisées dans ce modèle. Ces dernières sont respectivement *le prix de clôture* de l'action considérée au jour i et *le rendement relatif* de l'action au jour i [*cfr.* équation 4.1]. La troisième colonne quant à elle, se rapporte aux *classes* prédéfinies au jour i [*cfr.* Section 3.1]. Cette colonne peut également être considérée comme la colonne des *variables cibles*, étant donné que nous chercherons à prédire, en fonction de son niveau de probabilité, dans quelle classe une action se trouvera au jour i.

2. **Séparation de la base de données :** Pour une facilité d'utilisation du modèle de Machine Learning, la base de données ainsi créée doit être "éclatée".

   Afin d'éviter que le modèle de Machine Learning ne souffre de problèmes

de sur-apprentissage ("overfitting"), il est nécessaire de séparer les données en deux ensembles distincts. Le premier ensemble, que nous dénommerons ici *TrainingSet*, servira à l'entrainement du modèle tandis que le deuxième ensemble, dénommé dans ce mémoire *Testing* permettra de tester le modèle entrainé sur de nouvelles données.

Le "split" de la base de données a été réalisé par l'entremise de la fonction, du language de programmation R, **SORT** [7] et de la fonction **SAMPLE** [8] [Hainaut, 2020].

Enfin, mentionnons également que nous avons décidé que le *TrainingSet* regrouperait 60% de la base de données et donc que l'ensemble *TESTINg* serait formé des 40% restants.

L'implémentation de cette étape a été réalisée, en utilisant le langage de programmation R, comme suit [9] :

```
Tri = DataCrisis_Tris[,1:2]
Tri

Tri2 = DataCrisis_Tris[,3]

Essai = sort(sample(nrow(Tri), nrow(Tri)*0.6))

TrainingSet <- Tri[Essai,]
TrainingSet
TRAININg = preProcess(TrainingSet)


Testing = Tri[-Essai,]
Testing
TESTINg = preProcess(Testing)
dim(Testing)

TrainingClass = Tri2[Essai,]
TrainingClass14 = as.matrix(TrainingClass)
TrainingClassFit = as.factor(TrainingClass14)
TrainingClassFit
class(TrainingClassFit)
length(TrainingClassFit)

TestingClass = Tri2[-Essai,]
TestingClass
dim(TestingClass)
TestingClass14 = as.matrix(TestingClass)
TestingClassFit = as.factor(TestingClass14)
class(TestingClassFit)
```

FIGURE 4.3 – Implémentation de l'étape 2. dans RStudio

7. Explications : https ://www.rdocumentation.org/packages/base/versions/3.6.2/topics/sort
8. Explications : https ://www.rdocumentation.org/packages/base/versions/3.6.2/topics/sample
9. A noter que le screen de la figure 4.3 est l'implémentaion de la méthode pour une seule action, ici l'action de l'entreprise Hermès. Cette implémentation a été réalisée 39 fois avec à chaque fois des objets propres.

3. **Implémentation du modèle de Machine Learning en tant que tel :** Pour rappel, notre modèle est basé sur un algorithme de classification et plus précisément l'algorithme *C5.0* . Son implémentation nécessite le téléchargement du package "**C50**" [10]. La fonction R, nommée C5.0, du package est prévue pour réaliser cette étape. Dans la Figure 4.4 le lecteur trouvera la manière dont cet algorithme a été implémenté avec RStudio.

```
Tree_Method = C5.0(TrainDescrFit, TrainingClassFit, trials = 1, rules = FALSE, costs = NULL, )
```

FIGURE 4.4 – Implémentation de l'algorithme C5.0

Avec cette fonction, nous nous attelons donc, dans un premier temps, à entrainer le modèle. L'objet *TrainDescrFit* reprend les *variables explicatives* de l'action étudiée, pour un jour i. L'objet *TrainingClassFit* regroupe quant à lui les classes, préalablement définies, attribuées à l'action étudiée pour une journée i, en fonction des *variables explicatives*.

Une fois la fonction implémentée, il est nécessaire d'opérer ce que l'on appelle des prédictions. Dans cette phase le modèle va proposer une classe dans laquelle un point de données non-connu va être placé. Pour ce faire, la fonction **PREDICT** [11] est utilisée. Le détail de cette fonction est présenté dans la Figure 4.5.

```
PREDICTION = predict (Tree_Method, newdata = TestDescrFit, trials = 1, type = "class", na.action = na.pass)
PREDICTION
length(PREDICTION)
summary(PREDICTION)

PREDICTIONbis = predict (Tree_Method, newdata = TestDescrFit, trials = 1, type = "prob", na.action = na.pass)
PREDICTIONbis
```

FIGURE 4.5 – Implémentation de la fonction **PREDICT**

Avec cette fonction, et grâce au modèle entrainé (dénommé dans la figure 4.5, *Tree_Method*), nous sommes en mesure de prédire dans quelle classe un point de données de l'ensemble non-connu (dénommé dans la figure 4.5, *TestDescrFit*) sera placé. Un des avantages de l'utilisation de l'algorithme **C5.0** avec R Studio est qu'il nous permet d'obtenir 2 types d'outputs. Le premier (dénommé dans la figure 4.5, *PREDICTIOn*) nous donne directement la classe dans laquelle il est prédit

10. Téléchargeable ici : https ://CRAN.R-project.org/package=C50
11. explications : https ://www.rdocumentation.org/packages/car/versions/3.0-8/topics/Predict

que le nouveau point de données soit placé. Le deuxième (dénommé dans la figure 4.5, *PREDICTIOnbis*) nous permet d'obtenir le niveau de probabilité qu'un point de données soit classé dans chaque classe. Une description de cet output sera donnée dans l'étape 4 ci-dessous.

4. **Obtention d'un output probabiliste :** Comme il a été mentionné ci-dessus dans l'étape 3, l'algorithme **C5.0**, une fois implémenté dans RStudio, nous permet d'obtenir un output avec aspects probabilistes. Etant donné que nous avons décidé de baser notre étude sur un modèle de Machine Learning et plus précisément de classification avec aspect probabiliste, l'obtention de cet output se révèle être des plus importants. Dans la Figure 4.6 le lecteur trouvera un exemple des premières lignes de cet output.

	Acheter	Conserver	Vendre
1	0.0007288782	0.001405445	0.997865677
2	0.3413159507	0.658135557	0.000548492
3	0.3413159507	0.658135557	0.000548492
4	0.3413159507	0.658135557	0.000548492
5	0.0007288782	0.001405445	0.997865677
6	0.3413159507	0.658135557	0.000548492
7	0.3413159507	0.658135557	0.000548492
8	0.0007288782	0.001405445	0.997865677
9	0.3413159507	0.658135557	0.000548492
10	0.3413159507	0.658135557	0.000548492
11	0.3413159507	0.658135557	0.000548492
12	0.0007288782	0.001405445	0.997865677
13	0.3413159507	0.658135557	0.000548492
14	0.3413159507	0.658135557	0.000548492
15	0.3413159507	0.658135557	0.000548492
16	0.3413159507	0.658135557	0.000548492
17	0.3413159507	0.658135557	0.000548492
18	0.3413159507	0.658135557	0.000548492
19	0.0007288782	0.001405445	0.997865677
20	0.0007288782	0.001405445	0.997865677

FIGURE 4.6 – Les premières lignes de l'output probabiliste obtenu

A la Figure 4.6, nous pouvons observer que pour chaque point de données (symbolisé ici par les nombres allant de 1 à 20), un niveau de probabilité est attribué à chacune des 3 classes. Il est également à noter que pour chaque ligne, la somme des 3 niveaux de probabilité est bien égale à 1.

5. **Mesure de la performance du modèle de Machine Learning :** Dans le cas d'un problème de classification, la mesure de la performance du modèle va se baser sur 2 métriques, *la matrice de confusion* et *le taux de précision*. L'avantage avec le language de programmation

R, est que la fonction **confusionMatrix** [12], permet d'obtenir tant *la matrice de confusion* que *le taux de précision*. Dans la Figure 4.7 le lecteur trouvera un exemple de l'output fourni par la fonction **confusionMatrix**.

```
          Reference
Prediction  Acheter Conserver Vendre
  Acheter         0         0      0
  Conserver     138       212      0
  Vendre          5        11    249

Overall Statistics

               Accuracy : 0.7496
                 95% CI : (0.7134, 0.7834)
    No Information Rate : 0.4049
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5956

 Mcnemar's Test P-Value : < 2.2e-16
```

FIGURE 4.7 – Matrice de confusion et taux de précision

Une explication détaillée de *la matrice de confusion* est fournie à l'Annexe 7.

Création d'un portefeuille basé sur les outputs d'un algorithme de classification.

Afin de créer un portefeuille sur base de notre modèle de Machine Learning et plus précisément de notre algorithme de classification (tel que défini dans la sous-section précédente), nous allons définir les données d'entrée sur lesquelles se basera notre construction de portefeuille.
Cet input est en réalité "l'ouput" probabiliste de notre algorithme de classification, c'est-à-dire la matrice des probabilités, dont les premières lignes sont représentées à la Figure 4.6.

Avant de pouvoir créer notre portefeuille, deux étapes préliminaires sont requises.

1. **Création d'un vecteur d'indicateurs :** Le but de cette étape est de savoir dans quelle classe, sur base des niveaux de probabilités, un point de données sera placé. Pour ce faire, un algorithme a été implémenté. Voir la Figure 4.8.

12. Détails : https ://www.rdocumentation.org/packages/caret/versions/3.45/topics/confusionMatrix

```
lengthM <- nrow(PREDICTIONbis)
lengthM

result5 <- c()
result5 <- NULL

for (r in 1:lengthM) {

  if (PREDICTIONbis[r,1] > 0.500000000){

    result5 <- c(result5,1)
  }else{
    if (PREDICTIONbis[r,2] > 0.500000000){

      result5 <- c(result5,2)
    }else{
      if (PREDICTIONbis[r,3] > 0.500000000){

        result5 <- c(result5,3)
      }else{
}
    }
  }
}
result5

NewMatrix = cbind(PREDICTIONbis,result5)
NewMatrix
```

FIGURE 4.8 – Algorithme plaçant un point de données dans une classe, en fonction de son niveau de probabilité

Il a été choisi, dans le cadre de ce mémoire, de placer un point définitivement dans une classe sur base d'un seuil de probabilité. Dans le cas ci-présent, si la probabilité que le prix de clôture d'une action au jour i se trouve dans une classe est **supérieure** à **50%**, alors la décision finale sera de placer le titre associé dans cette classe, et ce pour une journée i.

En d'autres termes, si la probabilité que le prix d'une action pour une journée i se trouve dans la classe **Acheter** est supérieure à 50%, le chiffre **1** (car la colonne représentant cette classe est la première de la matrice d'inputs) sera affiché dans notre vecteur d'indicateurs. Par contre, si la probabilité qu'un point de données, pour un jour i soit placé dans la classe **Conserver** est supérieur à 50%, alors notre vecteur d'indicateurs sera complété par le chiffre **2** (car la colonne représentant cette classe est la deuxième de la matrice d'inputs). Enfin, le même raisonnement est effectué pour la classe **Vendre** à la différence que le chiffre associé à cette classe est **3**. Un exemple est donné dans la Figure 4.9, où les premières lignes de la nouvelle matrice (la matrice des probabilités, représentée à la Figure 4.6, à laquelle nous avons ajouté le vecteur colonne d'indicateurs [*result5* dans la Figure 4.9], généré par l'algorithme illustré à la Figure 4.8) sont affichées.

```
> NewMatrix
        Acheter    Conserver      Vendre result5
1   0.998182904 0.0007513597 0.001065736       1
2   0.001097359 0.0007624091 0.998140232       3
3   0.998182904 0.0007513597 0.001065736       1
4   0.998182904 0.0007513597 0.001065736       1
5   0.001097359 0.0007624091 0.998140232       3
6   0.001554591 0.9969134128 0.001531996       2
7   0.001554591 0.9969134128 0.001531996       2
8   0.001554591 0.9969134128 0.001531996       2
9   0.998182904 0.0007513597 0.001065736       1
10  0.001097359 0.0007624091 0.998140232       3
11  0.998182904 0.0007513597 0.001065736       1
12  0.998182904 0.0007513597 0.001065736       1
13  0.001554591 0.9969134128 0.001531996       2
14  0.001554591 0.9969134128 0.001531996       2
15  0.001554591 0.9969134128 0.001531996       2
16  0.001097359 0.0007624091 0.998140232       3
17  0.001554591 0.9969134128 0.001531996       2
18  0.001554591 0.9969134128 0.001531996       2
19  0.998182904 0.0007513597 0.001065736       1
20  0.001554591 0.9969134128 0.001531996       2
```

FIGURE 4.9 – Premières lignes de la "nouvelle matrice"

2. **Création d'une matrice de décision :** Le but de cette étape est de savoir, pour un jour i, quelles actions seront placées dans un portefeuille.
   Cette décision sera prise sur base d'une matrice de décision binaire (avec pour indice **1** si l'action sera présente dans le portefeuille le jour i et **0** dans le cas contraire). Celle-ci sera générée par un algorithme qui implémente les différents scénarios présentés dans le tableau 4.2 ci-dessous.

Jour $i-1$	**Jour** i	**Scenario OK ?**
Acheter	Vendre	OK
Acheter	Acheter	NON
Conserver	Conserver	OK
Conserver	Vendre	OK
Conserver	Acheter	OK
Vendre	Acheter	OK
Vendre	Conserver	NON
Vendre	Vendre	NON

TABLE 4.2 – Différents scénarios pris en compte

Ces scénarios ont été élaborés sur base de ce que réalisent, tous les jours, les analystes dans le monde professionnel. Ces derniers émettent

une recommandation d'achat, de conservation ou de vente d'un titre qu'ils suivent aux investisseurs. Ils peuvent recommander de vendre un titre qui était, le jour précédent, pourvu d'une recommandation d'achat. le montrendant, dans ce mémoire, nous avons fait l'hypothèse qu'un analyste ne passera pas directement d'un conseil à la vente à un conseil à conserver. Nous considérons qu'il passera par la recommandation "Acheter" avant. Le même raisonnement est suivi pour les scénarios passant de la recommandation "Vendre" à "Vendre" ainsi que de "Acheter" à "Acheter"

L'algorithme prenant en compte ces scénarios afin de générer la matrice de décision sur laquelle la composition de notre portefeuille sera basée, a été implémenté, de la manière présentée dans les Figure 4.10 et les Figure 4.11

```
Finalresult_38 <-c();
Finalresult_38 <- NULL

length_38 <- nrow(NewMatrix_38)

###logique du 1er jour###

if( NewMatrix_38[1,4] == 1 ){

  Finalresult_38 <- c(Finalresult_38,1)

} else {

  Finalresult_38 <- c(Finalresult_38,0)
```

FIGURE 4.10 – Logique du 1er jour pour l'algorithme générant une matrice de décision

La Figure 4.10 nous montre que pour le premier jour ($i = 1$) de notre univers d'investissement, notre algorithme fonctionne autrement que pour les autres jours. L'idée est que l'algorithme, présenté à la Figure 4.11, compare toujours la recommandation de la veille (jour $= i - 1$) avec les recommandations du jour i. Vu que nous n'avons pas de données pour la veille du premier jour, nous mettons en place un algorithme suivant une logique différente que pour les jours $i = 1 + n$. Etant donné que le vecteur d'indicateurs (*cfr* Figure 4.9) ne contient pas d'information pour le jour $(i = 1) - 1$, l'algorithme observera si pour ce jour $i = 1$, le prix de clôture de l'action à ce jour est classé dans la classe "**Acheter**". Si tel est le cas, l'algorithme générera le chiffre **1**, ce qui signifie que l'action est à prendre en compte dans la composition de notre portefeuille au jour $i = 1$. Si par contre ce n'est pas le cas et que le point de données se voit classé dans une des 2 autres classes, l'algorithme générera le chiffre **0** pour donner le signal

que l'action considérée ne doit pas être prise en compte dans la composition du portefeuille au jour i. Nous prenons donc l'hypothèse dans ce mémoire, que lors du premier jour ($i = 1$) d'investissement, l'investisseur ne dispose d'aucun titre. La seule action qu'il peut effectuer est donc d'acheter un titre et nous avons pris comme hypothèse dans ce mémoire que seuls les titres étant classés dans la classe **Acheter**, avec une probabilité supérieure à 50% pouvaient l'être réellement.

```
for( r in 2:length_36) {

  if(NewMatrix_36[r,4] == 1 && NewMatrix_36[r-1,4] == 3 ){

    Finalresult_36 <- c(Finalresult_36,1)

  }else{

    if(NewMatrix_36[r,4] == 2 && (NewMatrix_36[r-1,4] == 1 || NewMatrix_36[r-1,4] == 2)){
      Finalresult_36 <- c(Finalresult_36,1)

    }else{

      if(NewMatrix_36[r,4] == 3 && (NewMatrix_36[r-1,4] == 1 || NewMatrix_36[r-1,4] == 2)){
        Finalresult_36 <- c(Finalresult_36,1)

      }else{
        Finalresult_36 <- c(Finalresult_36,0)

      }

    }

  }

}

result_Matrix_36 <- matrix(Finalresult_36, ncol=1)
result_Matrix_36
```

FIGURE 4.11 – Suite de l'algorithme générant une matrice de décision

La Figure 4.11, montre comment, sur base des scénarios exposés dans le tableau 4.2, l'algorithme génère la matrice de décision binaire pour une action concernée. A chaque itération, l'algorithme, à l'aide de la condition **IF**, va analyser pour chaque ligne si les scénarios "OK" sont suivis dans notre vecteur d'indicateurs. Si tel est le cas, l'algorithme générera le chiffre **1**. Sinon, il générera le chiffre **0**. Ces chiffres seront placés dans un vecteur qui formera par la suite une colonne de notre matrice de décision, présentée à la Figure 4.12.

Le but étant de travailler conjointement avec les 39 actions de notre base de données, nous avons regroupé toutes les matrices de décision en une seule, constituée donc de 39 colonnes. Dans la Figure 4.12, le lecteur trouvera une illustration des premières lignes et colonnes de cette matrice de décision. Notons également que les lignes corres-

pondent à un jour i.

	RMS	MC	KER	SAF	OR	SU	AIR	DSY	LR	AI	HO	DG	SGO	BN	CA	BNP	STM	ML	AC	ATO	GLE	EN	SW	SAN	CS	UG	MT	VIE	ACA	FP	CAP	URW
[1,]	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[2,]	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0	1	1	0	0
[3,]	1	0	1	1	1	0	1	1	1	0	0	1	1	0	0	0	1	1	1	1	1	0	1	1	1	0	1	1	0	0	0	1
[4,]	0	1	0	0	1	1	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
[5,]	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	0	1	1	1	1
[6,]	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0	1	1	0	1	1	1	1	0	0	1	1
[7,]	1	1	1	1	1	1	1	1	0	0	1	1	1	1	0	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0
[8,]	1	1	1	1	1	1	0	1	1	0	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[9,]	0	1	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[10,]	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	0	1	0	1	1	1	0	0	1	0	0	0	1
[11,]	1	0	0	0	1	1	0	1	1	0	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1
[12,]	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	0	1	0	1	1	0	0	1	1
[13,]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	0	1	1	1	1	0	0	1	1
[14,]	1	1	1	1	1	1	0	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1	0	0	1	1	1	0	1	1	0	1
[15,]	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1
[16,]	1	1	0	0	0	0	1	0	1	1	0	1	1	1	1	0	0	1	1	1	0	1	1	0	1	0	1	0	0	0	0	1
[17,]	0	0	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	0	1	1	1	1	1	1	0	0	1	1	1	0	0
[18,]	1	0	0	0	1	0	0	1	0	1	1	0	1	1	0	1	1	1	1	0	1	1	1	0	0	1	0	1	0	0	1	1
[19,]	0	1	1	1	1	1	0	1	1	0	0	1	1	1	1	0	0	0	0	0	1	0	1	0	1	1	0	1	1	1	1	1
[20,]	1	1	0	0	0	1	0	0	1	1	1	0	0	1	0	1	1	0	0	1	0	0	1	1	0	0	0	1	0	0	1	0
[21,]	1	1	0	0	1	0	1	0	0	1	0	0	0	1	1	1	0	1	0	1	1	1	1	1	1	0	1	1	0	0	0	1

FIGURE 4.12 – Illustration des premières lignes et colonnes de notre matrice de décision

Une fois ces 2 étapes préliminaires effectuées, nous devons mettre en place une stratégie afin de créer un portefeuille pour un jour i. Dans ce but, nous avons choisi dans ce mémoire d'utiliser une stratégie de *portefeuilles équipondérés* (*cfr. Section 1.3*) pour tous les jours i composant notre matrice de décision. Pour rappel, d'après [DeMiguel et al., 2009], ce type de portefeuille peut surperformer d'autres stratégies dites optimales. De plus, étant donné que notre implémentation de modèle de Machine Learning ne prend pas en compte la corrélation entre les titres (un modèle a été implémenté par action), il se révèle plus opportun de travailler avec ce type de stratégie vu que ce dernier ne prend pas en compte la corrélation des titres dans la détermination de ses poids. Le but recherché ici est d'obtenir une matrice dont les éléments seront les poids des portefeuilles équipondérés et ce, toujours en se basant sur notre matrice de décision (Figure 4.12). Afin d'obtenir une matrice de poids, 3 étapes doivent être réalisées. Ces dernières sont expliquées ci-dessus.

1. **Faire correspondre notre matrice de décision avec la matrice de rendement associée :** L'objectif de cette étape est de remplacer, dans la matrice de décison (Figure 4.12), les chiffres 1 par les rendements qui y sont associés, en fonction du jour i et de l'action considérée. L'algorithme (dit *algorithme de correspondance*) présenté à la Figure 4.13 a été développé et implémenté.

```
tryingLast= sort(sample(nrow(C1), nrow(C1)*0.6))

MatriceRendements = C1[-tryingLast,]
MatriceRendements
##dim entre les 2 OK --> 615 x39 pour les 2##
#######Algo de correspondance######

PortfolioMatrix = matrix(nrow = nrow(FinalDecision), ncol = ncol(FinalDecision), byrow = FALSE)

    for(w in 1:nrow(FinalDecision)) {

      for (z in 1:ncol(FinalDecision)) {

        if (FinalDecision[w,z] == 1){

          PortfolioMatrix[w,z] <- MatriceRendements[w,z]
          }else{

            PortfolioMatrix[w,z] <- 0
  }
 }
}

colnames(PortfolioMatrix) = c('RMS', 'MC', 'KER', 'SAF', 'OR', 'SU', 'AIR', 'DSY', 'LR', 'AI', 'HO', 'DG', 'SGO', 'BN',
                              'CA', 'BNP', 'STM', 'ML', 'AC', 'ATO', 'GLE', 'EN', 'SW', 'SAN', 'CS', 'UG', 'MT', 'VIE', 'ACA',
                              'FP', 'CAP', 'URW', 'ENGI', 'PUB', 'ORA', 'VIV', 'EL', 'RNO', 'RI')
PortfolioMatrix
```

FIGURE 4.13 – Algorithme de correspondance

Dans l'algorithme présenté ci-avant, la première chose à faire est de créer une matrice de rendements dont le nombre de lignes correspond au nombre de lignes de notre matrice de décision (Figure 4.12).

A cette fin, nous avons choisi de décomposer notre matrice de rendement de la même manière que nous avons séparé notre base de données dans l'étape 2 de la sous-sous-section *Implémentation de l'algorithme de classification avec RStudio*. Une nouvelle matrice a également été créée. Cette dernière sera remplie au fur et à mesure de l'itération. Grâce à une double boucle **for** parcourant les lignes et les colonnes de la matrice de décision et à une condition **if** nous renseignant sur le fait que si un élément [w,z] de la matrice de décision est égal au chiffre 1, alors l'élément de la nouvelle matrice créée sera l'élément [w,z] de la matrice de rendement nouvellement créée. Si la condition n'est pas respectée, l'élément [w,z] de la nouvelle matrice équivaudra à un rendement de 0. La Figure 4.14 ci-dessous nous montre les premières lignes et colonnes de la nouvelle matrice.

```
> PortfolioMatrix
              RMS            MC           KER           SAF           OR            SU           AIR           D
 [1,]  0.0111706881  0.0000000000  0.000000e+00  0.0000000000  0.0000000000  0.0000000000  0.0035720361  0.00000000
 [2,]  0.0092633436  0.0061239193  7.142857e-03  0.0037111489  0.0014218009  0.0000000000 -0.0042002688  0.00261003
 [3,]  0.0000000000 -0.0085714286 -7.092199e-03  0.0000000000 -0.0014197823 -0.0165706052 -0.0078346391  0.00000000
 [4,]  0.0059510690  0.0000000000  0.000000e+00  0.0000000000  0.0164648910  0.0106440556  0.0000000000  0.01232519
 [5,] -0.0166883398  0.0069292487  0.000000e+00  0.0000000000 -0.0033783784  0.0000000000  0.0259361044  0.00201876
 [6,]  0.0026075619 -0.0058013053 -3.529827e-04 -0.0045984059  0.0000000000  0.0010921005  0.0065698479 -0.01150369
 [7,] -0.0097444781 -0.0057740888  5.665722e-03  0.0069284065 -0.0095238095 -0.0018433180 -0.0057803468  0.00000000
 [8,] -0.0157715260 -0.0103571429 -4.231312e-03  0.0051067781 -0.0131578947 -0.0062282469  0.0015324366 -0.00556521
 [9,]  0.0000000000  0.0025334781  0.000000e+00  0.0000000000  0.0042836744  0.0003705076  0.0000000000  0.00116822
[10,]  0.0000000000  0.0113345521  8.179232e-03 -0.0025208760  0.0038240918  0.0057674419  0.0000000000  0.00000000
[11,] -0.0016895459  0.0000000000  0.000000e+00  0.0000000000 -0.0038095238  0.0000000000  0.0246472248 -0.01380190
[12,]  0.0002112379  0.0022042616 -6.627136e-03 -0.0042600189  0.0043041607  0.0020377918  0.0049158631  0.00000000
[13,] -0.0037878788  0.0088954781 -3.486750e-04  0.0063512226  0.0000000000 -0.0011102887  0.0214368482  0.00000000
[14,]  0.0000000000  0.0000000000  7.022472e-03  0.0055883762 -0.0109004739  0.0000000000  0.0000000000  0.01068773
[15,] -0.0040598291  0.0057317539  1.008646e-02 -0.0001631055  0.0028804609  0.0000000000 -0.0073274985  0.00000000
[16,]  0.0004312204  0.0000000000  1.088929e-03  0.0044226044  0.0043902439  0.0076097561 -0.0031391014  0.01003991
[17,]  0.0000000000  0.0101801096  0.000000e+00  0.0000000000  0.0000000000  0.0096532703  0.0000000000  0.00449574
[18,]  0.0307934367  0.0323362975  2.791399e-02  0.0261617900  0.0297099980  0.0501706838  0.0162892332  0.01167793
```

FIGURE 4.14 – Premières lignes et colonnes de la nouvelle matrice

2. **Compter le nombre d'élements de la nouvelle matrice qui sont différents de 0 :** Etant donné que notre objectif est de construire un *portefeuille équipondéré*[13] pour chaque jour i, nous devons savoir exactement combien de titres composent notre portefeuille au jour i, et ce sur base de la matrice de décision ou, à ce stade, de la nouvelle matrice. Pour ce faire, l'algorithme (dénommé dans ce travail, algorithme de comptage) présenté à la Figure 4.15 a été implémenté.

```
N<-c()
for(x in 1:nrow(PortfolioMatrix)) {
  for (m in 1:ncol(PortfolioMatrix)) {
    if ( PortfolioMatrix[x,m] != 0) {
      N[x]<-sum(PortfolioMatrix[x,]!=0)
    }
  }
}
```

FIGURE 4.15 – Algorithme de comptage

Cette algorithme va compter le nombre d'éléments, composants la nouvelle matrice, différents de 0. Si l'élément [x,m] respecte cette condition, alors un vecteur par ligne, regroupant le nombre d'actions composant le portefeuille au jour i sera créé (il sera composé de 615 rendements dans notre cas). La Figure 4.16 reprend une illustration des premiers éléments de ce vecteur.

13. Pour un rappel de la définition, le lecteur est invité à se rendre à la sous-section 1.3 de ce mémoire

```
> N
  [1]  2 26 20 26 28 32 31 33 30 30 28 30 28 24 29 23 22 25 25 27 22 25 21 22 23 21 23 26 22 28 25 23 22 28 28 21 27
 [39] 27 30 28 28 26 28 26 28 26 24 28 26 22 25 28 23 24 24 23 27 22 23 22 25 25 28 24 21 23 21 25 24 21 23 17 15 19
 [77] 27 21 20 28 22 20 24 26 29 28 23 28 28 25 25 24 29 26 23 25 24 33 26 29 28 23 23 28 25 27 21 28 27 19 28 25 22
```

FIGURE 4.16 – Premiers éléments du vecteur comprenant le nombre d'actions qui composera un portefeuille au jour i

3. **Création de la matrice de poids :** Grâce à l'étape 2, nous savons de combien d'actions notre portefeuille au jour i sera composé. Il est maintenant nécessaire de connaître les poids qui seront attribués à chaque action composant un portefeuille au jour i. C'est donc dans ce but, que l'algorithme (dénommé algorithme de poids), représenté à la Figure 4.17 a été créé.

```
Weightsmatrix <- matrix(nrow = nrow(PortfolioMatrix), ncol = ncol(PortfolioMatrix), byrow = FALSE)

 for(x in 1:nrow(PortfolioMatrix)) {

   for (m in 1:ncol(PortfolioMatrix)) {

      if ( PortfolioMatrix[x,m] != 0) {

        Weightsmatrix[x,m] = round (1/N[x], digits = 4)

      } else {
        Weightsmatrix[x,m] = 0
   }
  }

 }

Weightsmatrix
```

FIGURE 4.17 – Algorithme de poids

Encore une fois, l'algorithme présenté à la Figure 4.17 fonctionne à l'aide de 2 boucles **for** et d'une condition **if-else**. Après avoir parcouru les boucles et avoir testé ou non le respect de la condition, l'algorithme permet de générer une matrice qui sera composée des poids de portefeuille. Ces poids sont calculés grâce à la formule exposée à la section 1.3 où $N = nombre\ d'actions\ composant\ le\ portefeuille$. Pour rappel, ce nombre a été déterminé grâce à l'algorithme représenté à la Figure 4.15.

Une fois toutes ces étapes réalisées, nous sommes en mesure de travailler avec des *portefeuilles équipondérés* pour chaque jour i [14].

14. La matrice de poids, déterminant nos portefeuilles, sera illustrée dans le Chapitre 4 : Résulats

Chapitre 5

Résultats

Dans ce chapitre, nous allons présenter 2 types de résultats. Les premiers exposés ci-dessous sont ceux à associer directement aux modèles de Machine Learning implémentés. Nous nous concentrerons sur les *taux de précision* associés à chaque modèle. Les seconds concernent les résultats relatifs aux différents portefeuilles créés sur base des outputs probabilistes provenant de nos modèles de Machine Learning implémentés. La performance de ces portefeuilles est quant à elle évaluée grâce au Sharpe Ratio [1]. Enfin, il est également à noter que tous les résultats présents ci-dessous sont présentés à la fois pour la base de données dite de "*Crise*" et celle dite de "*PostCrise*".

5.1 Performance des modèles de Machine Learning

La performance d'un modèle de Machine Learning découlant d'un algorithme de classification peut être mesurée grâce à 2 métriques, *la matrice de confusion* et le *taux de précision*. Dans cette section, nous présentons uniquement les résultats des *taux de précision*. Afin d'analyser au mieux ces résultats, nous avons décidé de calculer la moyenne des 39 taux de précision, par base de donnée. Nous avons également jugé, au regard du fait que nous travaillons avec la fonction **SAMPLE** (celle-ci prend à chaque fois des échantillons aléatoires de notre base de données globales), qu'il était intéressant de réaliser plusieurs itérations pour calculer les moyennes afin de voir si ces dernières ne sont pas trop éloignées les unes des autres. Le tableau ci-dessous nous montre ces moyennes en fonction des itérations.

1. Voir Sous-Section 4.2.3

Itérations	**Moyenne Crise**	**Moyenne PostCrise**
I	0,743479487	0,911053846
I + 1	0,839866667	0,910469231
I + 2	0,7398	0,909017949

TABLE 5.1 – Evolution des moyennes des taux de précision suivant les itérations pour les 2 bases de données

Dans le tableau 5.1, on observe que la moyenne des taux de précision, par bases de données distinctes, reste toujours dans le même ordre de grandeur. La seule variation a souligné est celle entre l'itération I et I +1. Dans celle-ci, on voit que le taux de précision gagne presque un dixième. Au vu de ces faibles variations, et malgré l'utilisation de la fonction **SAMPLE**, nous pouvons dire que les taux de précision de nos 39 modèles n'évolueront pas fortement.

Si nous interprétons les taux de précision en tant que tel, nous pouvons dire que ces derniers permettent de souligner le fait que nos modèles sont précis, surtout pour la période de *PostCrise*. Cette différence de précision entre les 2 bases de données vient peut-être du fait que les prix sont moins volatiles en période de *PostCrise*. Ces prix peuvent avoir une influence sur notre modèle, étant donné que pour rappel, nous utilisons *les prix de clôture* d'une action au jour i comme une de nos variables explicatives.

5.2 Portefeuilles équipondérés construits sur base d'un algorithme de classification avec aspect probabiliste

Pour les 2 bases de données, la méthodologie développée pour constuire un portefeuille équipondéré fonctionne étant donné que nous obtenons une matrice de poids une fois que les itérations des algorithmes mis en place sont terminées. Ci-dessous, le lecteur pourra observer un exemple [2] de ces matrices de poids pour les 2 bases de données.

2. On parle ici d'exemple car, étant donné que nous utilisons la fonction **SAMPLE**, ces matrices seront modifiées quand nous relancerons les algorithmes

```
> weightsmatrix
          RMS     MC    KER    SAF     OR     SU    AIR    DSY     LR     AI     HO     DG    SGO
 [1,] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
 [2,] 0.0000 0.0000 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0000 0.0000 0.0345
 [3,] 0.0000 0.0000 0.0385 0.0385 0.0385 0.0000 0.0000 0.0000 0.0385 0.0385 0.0000 0.0385 0.0385
 [4,] 0.0000 0.0385 0.0385 0.0385 0.0385 0.0000 0.0000 0.0385 0.0385 0.0000 0.0000 0.0385 0.0000
 [5,] 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294
 [6,] 0.0294 0.0000 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0000 0.0294 0.0294 0.0294
 [7,] 0.0000 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0000 0.0312 0.0312 0.0312 0.0312
 [8,] 0.0312 0.0000 0.0312 0.0312 0.0312 0.0000 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312
 [9,] 0.0000 0.0333 0.0333 0.0333 0.0000 0.0333 0.0333 0.0333 0.0333 0.0333 0.0000 0.0333 0.0333
[10,] 0.0323 0.0323 0.0323 0.0323 0.0323 0.0323 0.0323 0.0000 0.0323 0.0323 0.0323 0.0323 0.0323
[11,] 0.0345 0.0345 0.0000 0.0000 0.0000 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0000
[12,] 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0000 0.0345 0.0000 0.0000 0.0000 0.0345 0.0000
[13,] 0.0370 0.0370 0.0370 0.0000 0.0370 0.0370 0.0000 0.0000 0.0370 0.0370 0.0370 0.0370 0.0000
[14,] 0.0333 0.0333 0.0333 0.0333 0.0333 0.0333 0.0333 0.0333 0.0333 0.0000 0.0333 0.0333 0.0333
[15,] 0.0000 0.0000 0.0345 0.0345 0.0345 0.0345 0.0000 0.0345 0.0345 0.0345 0.0000 0.0345 0.0000
[16,] 0.0370 0.0370 0.0000 0.0000 0.0370 0.0370 0.0370 0.0370 0.0370 0.0370 0.0000 0.0370 0.0000
[17,] 0.0357 0.0000 0.0357 0.0357 0.0357 0.0357 0.0357 0.0357 0.0000 0.0357 0.0000 0.0000 0.0357
```

FIGURE 5.1 – Un exemple de matrice de poids pour la période de *Crise*

```
> weightsmatrix
          RMS     MC    KER    SAF     OR     SU    AIR    DSY     LR     AI     HO     DG    SGO
 [1,] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
 [2,] 0.0000 0.0000 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0000 0.0000 0.0345
 [3,] 0.0000 0.0000 0.0385 0.0385 0.0385 0.0000 0.0000 0.0000 0.0385 0.0385 0.0000 0.0385 0.0385
 [4,] 0.0000 0.0385 0.0385 0.0385 0.0385 0.0000 0.0000 0.0385 0.0385 0.0000 0.0000 0.0385 0.0000
 [5,] 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294
 [6,] 0.0294 0.0000 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0294 0.0000 0.0294 0.0294 0.0294
 [7,] 0.0000 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0000 0.0312 0.0312 0.0312 0.0312
 [8,] 0.0303 0.0000 0.0303 0.0303 0.0303 0.0000 0.0303 0.0303 0.0303 0.0303 0.0303 0.0303 0.0303
 [9,] 0.0000 0.0333 0.0333 0.0333 0.0000 0.0333 0.0333 0.0333 0.0333 0.0333 0.0000 0.0333 0.0333
[10,] 0.0323 0.0323 0.0323 0.0323 0.0323 0.0323 0.0323 0.0000 0.0323 0.0323 0.0323 0.0323 0.0323
[11,] 0.0345 0.0345 0.0000 0.0000 0.0000 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0000
[12,] 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0000 0.0345 0.0000 0.0000 0.0000 0.0345 0.0000
[13,] 0.0357 0.0357 0.0357 0.0357 0.0357 0.0357 0.0000 0.0357 0.0357 0.0357 0.0357 0.0357 0.0000
[14,] 0.0333 0.0333 0.0333 0.0333 0.0333 0.0333 0.0333 0.0333 0.0333 0.0000 0.0333 0.0333 0.0333
[15,] 0.0000 0.0000 0.0345 0.0345 0.0345 0.0345 0.0000 0.0345 0.0345 0.0345 0.0000 0.0345 0.0000
[16,] 0.0400 0.0400 0.0000 0.0000 0.0400 0.0400 0.0400 0.0400 0.0400 0.0400 0.0000 0.0400 0.0000
[17,] 0.0345 0.0000 0.0345 0.0345 0.0345 0.0345 0.0345 0.0345 0.0000 0.0345 0.0345 0.0000 0.0345
```

FIGURE 5.2 – Un exemple de matrice de poids pour la période de *PostCrise*

Même si dans les Figure 5.1 et 5.2, nous n'avons que les premières lignes et colonnes de cette matrice de poids, nous sommes quand même en mesure d'observer que pour chaque ligne illustrée, les poids sont égaux. Ce qui signifie donc bien que nous avons formé des *portefeuilles équipondérés* pour chaque jour i. Comme le lecteur l'aura compris, un poids de 0 au jour i signifie que nous ne prenons pas l'action en compte dans la composition de notre portefeuille pour ce jour.

5.3 Performance d'un portefeuille équipondéré au jour i

Comme mentionné au début de ce chapitre, la performance d'un portefeuille au jour i sera calculée à l'aide du *Sharpe Ratio*. Nous ne montrons pas

ici les différents *Sharpe Ratio* par jour i étant donné que comme déjà mentionné dans ce mémoire, il s'agirait d'exemples, étant donné l'utilisation de la fonction **SAMPLE**. le montrendant, après avoir exécuté plusieurs fois l'ensemble des algorithmes nécessaires à la création d'un *portefeuille équipondéré*, nous remarquons que la plupart des Sharpe Ratio sont négatifs. Comme dans la section 5.1, étant donné que nous utilisons la fonction **SAMPLE** pour séparer notre base de données, nous réalisons une moyenne des 615 Sharpe Ratios pour une itération afin d'avoir une idée claire des valeurs de ce ratio que peut nous donner notre méthodologie découlant des algorithmes créés. Une fois encore, ces moyennes ont été réalisées pour les 2 bases de données. Ces moyennes sont exposées dans le tableau 5.2

Itérations	**Moyenne Crise**	**Moyenne PostCrise**
I	-1.184627	-1.111307
I + 1	-1.160699	-1.168713
I + 2	-1.164012	-1.195395

TABLE 5.2 – Evolution des moyennes des taux des Sharpe Ratios suivant les itérations pour les 2 bases de données

Nous pouvons clairement observer que les moyennes de ces Sharpe Ratios sont très mauvaises. Comme expliqué dans la sous-section 4.2.3 de ce mémoire, un Sharpe Ratio négatif est une mauvaise situation. Nous sommes donc en mesure d'interpréter ces ratios en mentionnant le fait que notre stratégie de portefeuille basée sur un modèle de Machine Learning n'est pas efficace.

Un autre moyen de comprendre que notre stratégie de portefeuille mise en place n'est pas efficace correctement est d'avoir en fonction des itérations, le nombre de Sharpe Ratio positifs sur 615 jours. Le tableau 5.3 le montre.

Itérations	**Moyenne Crise**	**Moyenne PostCrise**
I	27	31
I + 1	29	28
I + 2	28	33

TABLE 5.3 – Nombre de Sharpe Ratios positifs pour une itération pour les 2 bases de données

Chapitre 6

Conclusion

La conclusion de ce mémoire débute par un bref résumé des résulats, ensuite les limites de l'étude sont décrites et enfin des suggestions pour des recherches futures sont proposées.

Le but de ce mémoire était de construire un portefeuille d'actions sur base d'un modèle de Machine Learning utilisant un algorithme de classification avec aspect probabiliste. Pour ce faire, nous avons développer notre propre méthodologie exposée à la sous-sous-section *Création d'un portefeuille basé sur les outputs d'un algorithme de classification* de ce travail. Cette dernière, après avoir été appliquée à nos deux bases de données, nous permet d'obtenir les poids d'un portefeuille (Voir les Figure 5.1 et Figure 5.2) et donc sa composition. Grâce à cette méthodologie et aux résultats qui en découlent, nous pouvons donc répondre à la première partie de notre première question de recherche qui était pour rappel "*Comment construit-on un portefeuille d'actions sur base d'un algorithme de Machine Learning (plus précisément d'un algorithme de classification avec aspect probabiliste) et quel en est la performance ?*". La réponse est la méthodologie développée.

Grâce à l'utilisation de Sharpe Ratio (Voir sous-section 4.2.1), nous obtenons des renseignements quant à la réponse à apporter au volet performance de la question de recherche. Ce dernier nous permet en effet d'obtenir les performances de nos portefeuilles créés par le biais de notre méthodologie développée dans la sous-sous-section *Création d'un portefeuille basé sur les outputs d'un algorithme de classification* et appliqué à nos 2 bases de données. Afin d'avoir une idéé plus précise des niveaux de performances que nos portefeuilles peuvent atteindre pour chaque jour i, nous avons créé nos propres métriques, à savoir la moyenne des Sharpe Ratio par itérations et par bases de donnée (présenté à la table 5.2) et le nombre de Sharpe Ratios positifs pour

615 jours, toujours par itérations de la méthodologie et par bases de données. Ces 2 métriques nous montrent clairement que les portefeuilles équipondérés, construits sur base de notre méthodologie nous donnent, tant pour la période de "*Crise*" que pour celle de *Postcrise*, des performances très décevantes et ce, malgré le fait que les performances propres aux prédictions par le modèle de Machine Learning en tant que tel sont bonnes (Voir section 5.1).

Avec ces performances, nous pouvons également répondre à notre deuxième question de recherche, qui est pour rappel "*Est-ce que les portefeuilles construits en période dite de crise vont se révéler être sous-performants par rapport aux portefeuilles construits en période dite de "post-crise" ?*. Nos Sharpe Ratios sont ici dans le même ordre de grandeur pour les 2 bases de données ce qui signifie donc qu'un portefeuille construit en période de "*Crise*" ne sous-performe pas un autre construit en période de *PostCrise*, tout du moins dans le cas de notre étude.

6.1 Les limites de l'étude

. Comme dans toute recherche, il est important de souligner les limites rencontrées lors de la réalisation de ce mémoire. Le lecteur prendra donc en compte les aspects suivants :

1. Les 2 bases de données dites de "*Crise*" et de "*PostCrise*" ont été créées entièrement par nos soins. Cela peut se révéler contraignant quant à la mise en place de modèles de Machine Learning, car la base de données peut se révéler inadaptée à ce type de problèmes. De plus, nous avons travaillé uniquement avec les actions composant l'indice CAC40. Les résultats obtenus sont donc spécifiques à la composition de cet indice et pourraient varier si nous appliquons notre modèle de décision à d'autres données.
2. La méthodologie employée dans ce mémoire pour un construire un portefeuille a été développée par nos soins. Elle ne se base donc pas sur un quelconque exemple provenant de la littérature.
3. Etant donné que dans ce mémoire, nous avons fait le choix d'appliquer un modèle de Machine Learning action par action (pour rappel, le modèle reposant sur l'algorithme de classification **C5.0** a été appliqué 39 fois, donc 1 modèle par action). La corrélation qu'il pourrait y avoir entre les différents titres n'est donc pas prise en compte. Cela peut se révéler problématique pour d'autres stratégies de portefeuille que celle du *portefeuille equipondéré*.

4. Vu la limite exposée au point 2 ci-dessus, seulement une stratégie de portefeuille (celle du *portefeuille équipondéré*) a été mise en place. Nous avons tenté de limiter notre analyse à la construction de portefeuille la plus pertinente au vu de l'implémentation de notre modèle de classification. le montrendant, étant donné le nombre important de stratégies de construction de portefeuille présentes dans la littérature, d'autres constructions de portefeuille pourraient être implémentées sur base d'un output de modèle de Machine Learning.
5. Notre modèle de classification se limite uniquement à l'utilisation de 2 variables *explicatives*, pour rappel, *le prix de clôture* d'une action au jour i et le rendement relatif de l'action au jour i tel que défini dans l'équation 4.1. D'autres variables, comme par exemple le volume d'échange d'une action, pourraient être également utilisées comme variables explicatives. De plus, dans cette étude, nous n'utilisons pas comme variables explicatives des indicateurs techniques comme par exemple la moyenne mobile.

6.2 Suggestions pour des études prochaines

Nous concluons ce mémoire en suggérant plusieurs propositions de recherches futures qui méritent d'être examinées.

1. Comme déja énoncé précédemment, les résultats de ce mémoire ont été obtenus sur base de 39 outputs provenant de 39 modèles de classification (ce sont toujours les mêmes itérations qui sont réalisées mais pour chaque action). Utiliser une approche permettant de réaliser une seule fois les itérations pour toutes les actions pourrait se révéler pertinent.
2. Les outputs desquels les résultats de ce mémoire résultent uniquement de l'implémentation d'un seul algorithme de classification avec aspect probabiliste (l'algorithme **C5.0**). Il serait donc intéressant d'implémenter d'autres algorithmes de classification, toujours avec aspect probabiliste et d'en comparer les résultats.
3. Dans ce mémoire, un point de données sera placé définitivement dans une classe si la probabilité qu'il y soit est **supérieure à 50%**. Il pourrait être intéressant de tenter de travailler avec des seuils plus élevés et de comparer les résultats obtenus.
4. La base de données a été divisée en 2 parties afin de mener à bien l'implémentation d'un algorithme de Machine Learning. Dans le cas de ce mémoire, 60% des données sont utilisées pour l'entrainement

du modèle et 40% sont utilisées comme données de test. Il serait intéressant de comparer les résultats pour des pourcentages différents.

5. Une seule stratégie de portefeuille a été créée sur base de notre modèle de classification. Une idée de recherche future serait d'en créer d'autres, comme par exemple le portefeuille minimum-variance, et d'en comparer les performances.

Chapitre 7

Annexes

7.1 Annexe 1 - Evolution de l'adoption du Machine Learning par les groupes d'investissement

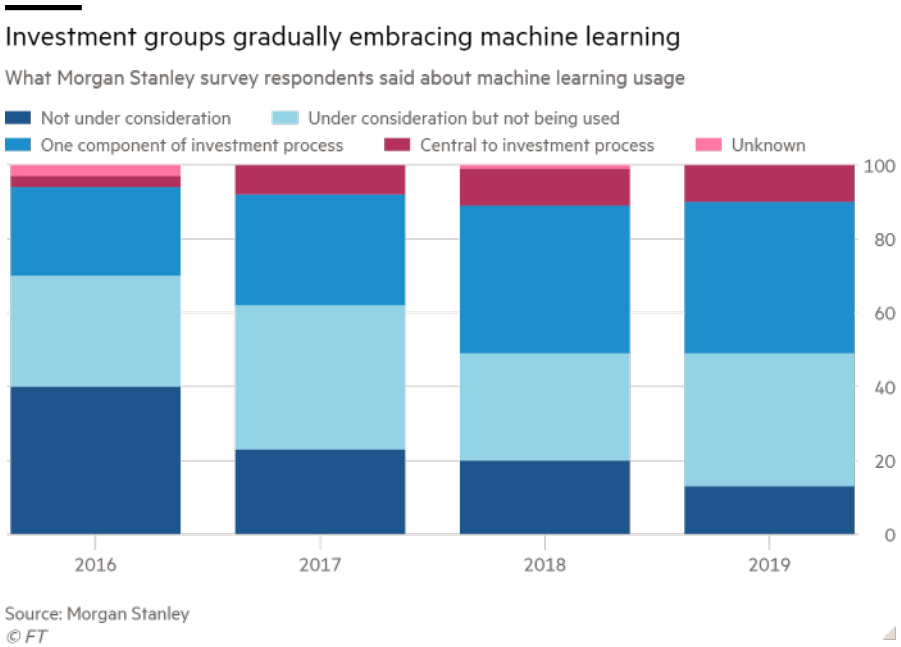

FIGURE 7.1 – Evolution de l'adoption du Machine Learning par les groupes d'investissement

7.2 Annexe 2 - Les moments statistiques

7.2.1 Les moments statistiques ; la variance, le skewness et le kurtosis.

Avant toutes choses, il est important de mentionner que nous travaillerons ici avec les moments *centrés*. Un moment centré, d'ordre $k \in \mathbb{N}$, d'une variable aléatoire, X, est un indicateur de dispersion de cette dernière. En d'autres termes, il s'agit d'une valeur attendue de l'écart de la variable aléatoire X par rapport à sa moyenne $\mu(X)$. Le *k-ieme* moment centré de la variable aléatoire X peut être défini comme suit :

$$m_k(X) = \mathbb{E}[(X - \mu(X))^k] \tag{7.1}$$

où $\mu(X)$ peut être définie comme la *moyenne* de la variable aléatoire X.

La variance de la variable aléatoire X est le deuxième moment ($k = 2$) de cette dernière. Mathématiquement, nous pouvons l'écrire de la façon suivante :

$$m_2(X) = \sigma^2(X) = \mathbb{E}[(X - \mu)^2] \tag{7.2}$$

Le troisème ($k = 3$) et le quatrième ($k = 4$) moments de la variable aléatoire X portent respectivement le nom de *skewness* ($m_3(X) = \gamma(X)$) et de *kurtosis* ($m_4 = \kappa(X)$). Leur définition mathématique sont les suivantes :

$$m_3(X) = \gamma(X) = \mathbb{E}[(\frac{X - \mu(X)}{\sigma(X)})^3] \tag{7.3}$$

$$m_4(X) = \kappa(X) = \mathbb{E}[(\frac{X - \mu(X)}{\sigma(X)})^4] \tag{7.4}$$

Le *skewness*, $\gamma(X)$, est la mesure du degré d'asymétrie de la distribution de la variable aléatoire X. Un *skewness* négatif verra la courbe représentant la distribution de la variable aléatoire X s'étaler sur le gauche alors que un *skewness* positif s'étalera sur la droite. Pour une meilleure compréhension, le lecteur est invité à observer la figure ci-dessous.

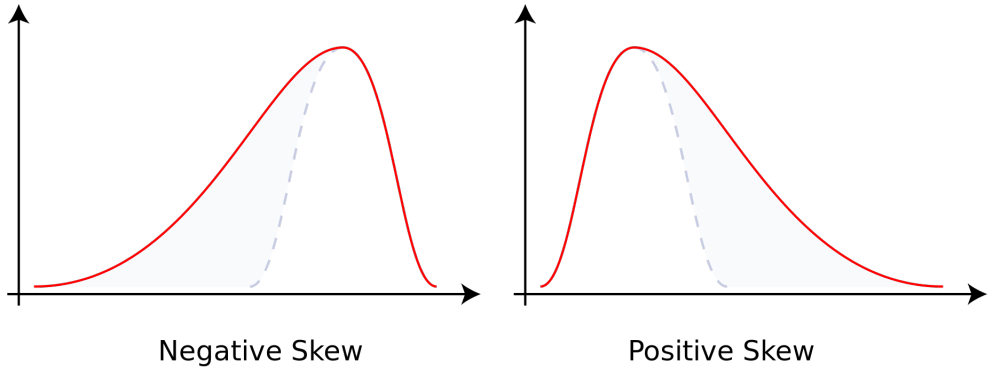

FIGURE 7.2 – Représentation d'un skewness positif et négatif.

Le *kurtosis*, $\kappa(X)$, quant à lui mesure l'acuité de la distribution de la variable aléatoire X. En d'autres termes, il mesure la répartition des masses de probabilité, leur concentrations à proximité du centre de probabilité qui n'est rien d'autre que la moyenne $\mu(X)$ de la variable aléatoire X. Il mesure les valeurs extrèmes dans les queues.

Dans un monde parfaitement gaussien, nous retouvons un *skewness* $\gamma(X) = 0$ et un *kurtosis* $\kappa(X) = 3$. Concernant la distribution de rendement de titres, le *skewness* s'étale beaucoup plus sur la gauche, ce qui signifie que des rendements négatifs sont plus susceptibles d'être observés que des rendements positifs [Cont, 2001]

7.3 Annexe 3 - Le produit de Kronecker

Soit une matrice A de dimension m lignes et n colonnes et une matrice B de dimension p lignes et q colonnes. **Le produit de Kronecker** $A \otimes B$ est la matrice définie par :

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & ... & a_{1n}B \\ a_{21}B & a_{22}B & ... & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & ... & a_{mn}B \end{bmatrix} \tag{7.5}$$

Si maintenant nous prenons un exemple chiffré ; soit

$$A = \begin{bmatrix} 1 & 3 & 2 \\ 1 & 0 & 0 \\ 1 & 2 & 2 \end{bmatrix} \tag{7.6}$$

et

$$B = \begin{bmatrix} 0 & 5 \\ 5 & 0 \\ 1 & 0 \end{bmatrix} \tag{7.7}$$

Nous obtenons donc que

$$A \otimes B = \begin{bmatrix} 0 & 5 & 0 & 15 & 0 & 10 \\ 5 & 0 & 15 & 0 & 10 & 0 \\ 1 & 1 & 3 & 3 & 2 & 2 \\ 0 & 5 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 10 & 0 & 10 \\ 5 & 0 & 10 & 0 & 10 & 0 \\ 1 & 1 & 2 & 2 & 2 & 2 \end{bmatrix} \tag{7.8}$$

7.4 Annexe 4 - Les 5V's du *big data*

Les 5V's du *big data* [Kolp, 2018] peuvent être vus comme des caractéristiques propres à ce type de données. Ces caractéristiques sont :

1. **Volume :** Une quantité de données beaucoup plus importante que celle des bases de données habituelles.
2. **Variété :** Beaucoup de différents types et formats de données.
3. **La Vélocité/ La Vitesse :** Les données arrivent à un rythme très rapide
4. **La Véracité :** La qualité des données. Les méthodes traditionnelles de qualité des données ne s'appliquent pas ; comment juger de l'exactitude et de la pertinence des données ?
5. **La Valeur :** Le *big data* est précieux pour favoriser de bonnes actions et décisions organisationelles.

7.5 Annexe 5 - Différents algorithmes de Machine Learning

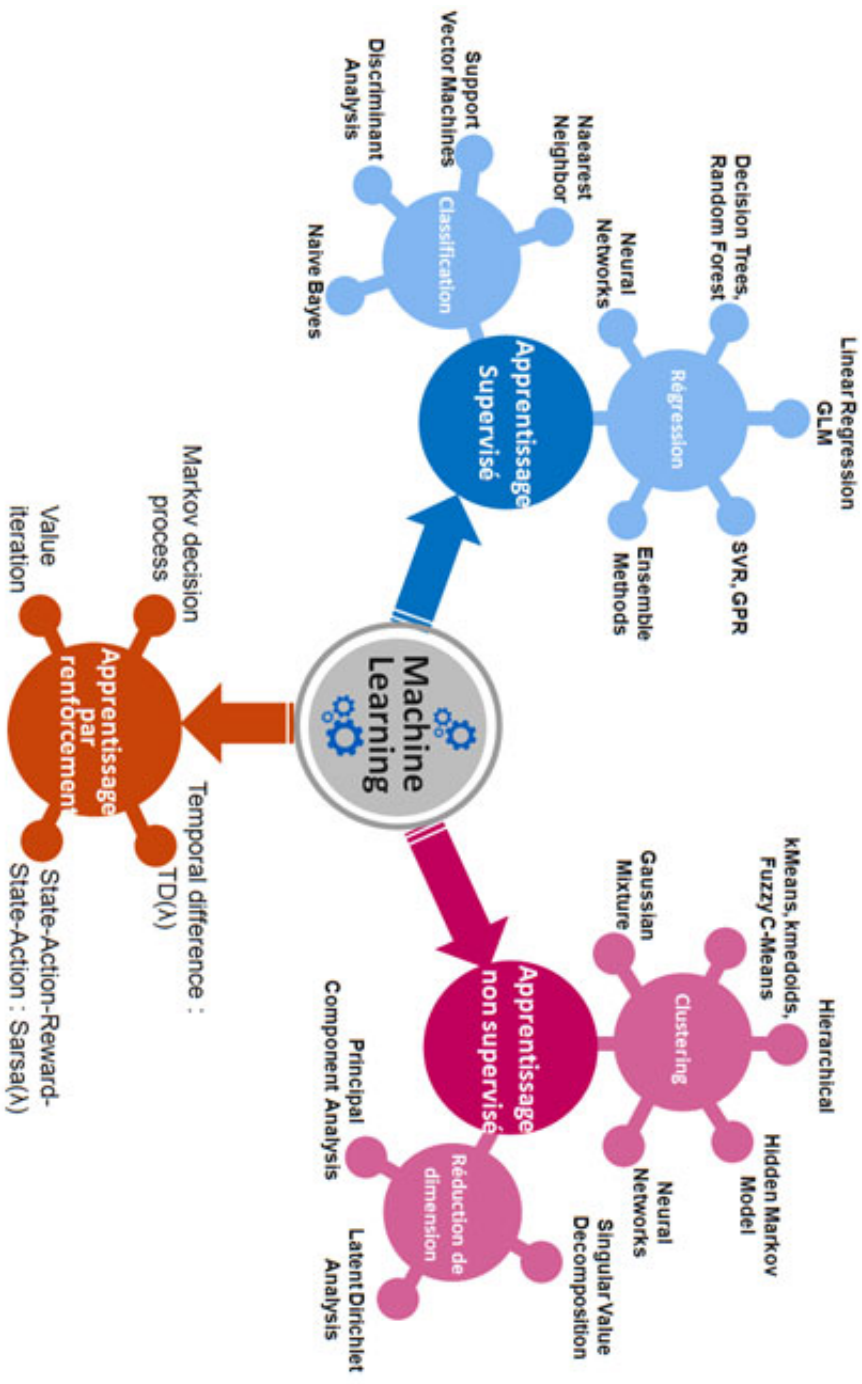

FIGURE 7.3 – Les différents types de Machine Learning et leurs applications.

7.6 Annexe 6 - Composition du CAC40

Dans cette Annexe, le lecteur trouvera les entreprises composant l'indice CAC40, par secteurs d'activités. Ces captures d'écran sont issues d'un terminal Bloomberg.

FIGURE 7.4 – Composition du CAC40 - 1

FIGURE 7.5 – Composition du CAC40 - 2

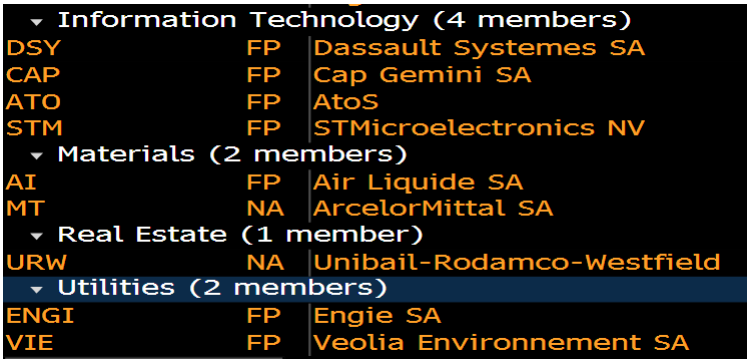

FIGURE 7.6 – Composition du CAC40 - 3

7.7 Annexe 7 - La matrice de confusion et le taux de précision

Dans [Saerens and Decaestecker, 2018], on trouve une définition de *la matrice de décision* ainsi que du *taux de précision*. Le lecteur les trouvera ci-après

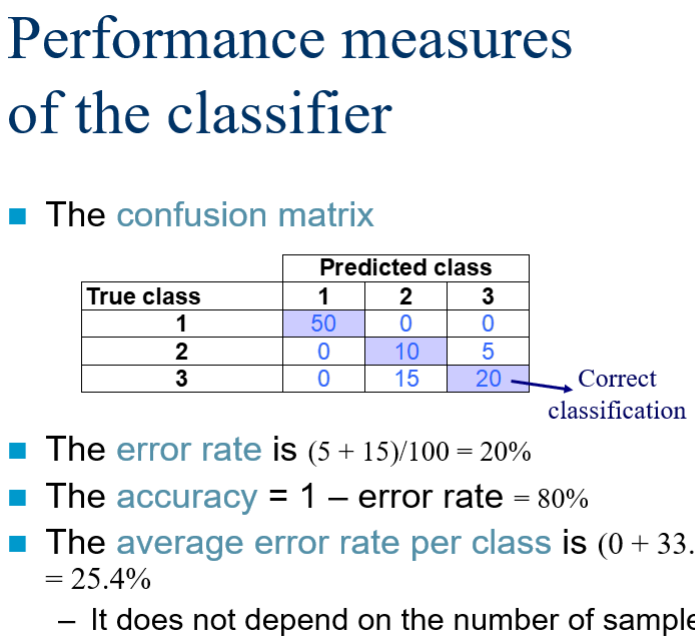

FIGURE 7.7 – Définition de la matrice de confusion et du taux de précision

Bibliographie

[Ali, 2017] Ali, O. (2017). Harnessing machine learning in your portfolio. Retrieved from : https ://www.gsam.com/content/gsam/global/en/market-insights/gsam-insights/2017/gsam-viewpoints.html.

[Chopra and Ziemba, 1993] Chopra, V. and Ziemba, W. (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. *The Journal of Portfolio Management*, 19(2) :6 – 11. 10.3905/jpm.1993.409440.

[Collard, 2019] Collard, B. (2019). L'impact de l'intelligence artificielle dans la gestion de portefeuille. Master's thesis, Louvain School of Management, Université Catholique de Louvain.

[Cont, 2001] Cont, R. (2001). Empirical properties of asset returns : stylized facts and statistical issues. *Quantitative Finance*, 1(3) :223 – 58. DOI : 101116.5992.

[DeMiguel et al., 2009] DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification : How inefficient is the 1/n portfolio strategy ? *Review of Financial Studies*, 22(5) :1915 – 1953. DOI = 10.1093/RFS/HHM075.

[Fama, 1970] Fama, E. (1970). Efficient capital market : A review of theory and empirical work. *The Journal of Finance*, 25(2) :383 – 417. DOI = 10.2307/2325486.

[Fama, 1991] Fama, E. (1991). Efficient capital market : Ii. *The Journal of Finance*, 46(5) :1575 – 1617. DOI = 10.2307/2328565.

[Hainaut, 2020] Hainaut, D. (2020). *LDAT2230 : Data Sciences in Insurance and Finance*. Université Catholique de Louvain, ISBA.

[Izenman, 2013] Izenman, A. J. (2013). *Modern Multivariate Statistical Techniques*. Springer, New York. DOI :10.1007/978-0-387-78189-1.

[Jagannathan and Ma, 2003] Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios : Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4) :1651 – 1683. 10.1111/1540-6261.00580.

[Kolm et al., 2014] Kolm, P., Tütüncü, R., and Fabozzi, F. (2014). 60 years of portfolio optimization : Practical challenge and current trends. *European Journal of Operational Research*, 234(2) :356–371. DOI : 10.1016/j.ejor.2013.10.060.

[Kolp, 2018] Kolp, M. (2018). *LLSMF2013 - LLSMF2014 : Database and Data Management Part*. Université Catholique de Louvain, Louvain School of Management.

[Lantz, 2013] Lantz, B. (2013). *Machine Learning with T*. Packt Publishing, Birmingham.

[Ledoit and Wolf, 2003] Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5) :603 – 621. DOI : 10.1016/S0927-5398(03)00007-0.

[Ledoit and Wolf, 2004a] Ledoit, O. and Wolf, M. (2004a). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4) :110 – 119. DOI = 10.2139/ssrn.433840.

[Ledoit and Wolf, 2004b] Ledoit, O. and Wolf, M. (2004b). A well conditionned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2) :365 – 411. DOI : 10.1016/S0047-259X(03)00096-4.

[Litterman, 2003] Litterman, R. (2003). *Modern Investment Management*. John Wiley & Sons Inc., Hoboken, New Jersey.

[Lovell and Kempf, 2019] Lovell, H. and Kempf, A. (2019). Bnpp am's multi asset quantitative solutions (maqs) - evolving "quantamental" and esg approaches. *Risk Magazine*, (145).

[Markowitz, 1952] Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1) :77 – 99.

[Memmel and Kempf, 2006] Memmel, C. and Kempf, A. (2006). Estimating the global minimum variance portfolio. *Schmalenbach Business Review*, 58(4) :332 – 348. DOI : 10.1007/BF03396737.

[Michaud, 1989] Michaud, R. (1989). The markowitz optimization enigma : Is "optimized" optimal ? *Financial Analysts Journal*, 45(1) :31 – 42. DOI : 10.2469/faj.v45.n1.31.

[PDMIA, 2008] PDMIA, G. (2008). Processus décisionnels de markov en intelligence artificielle. Retrieved from : http ://researchers.lille.inria.fr/ munos/papers/files/bouquinPDMIA.pdf.

[Saerens and Decaestecker, 2018] Saerens, M. and Decaestecker, C. (2018). *LLSMF2013-LLSMF 2014 : Lecture 01 - General Introduction*. Université Catholique de Louvain, Louvain School of Management.

[Sharpe, 1994] Sharpe, W. (1994). The sharpe ratio. *The Journal of Portfolio Management*, 21(1) :49 – 58. DOI : 10.3905/jpm.1994.409501.

[Staff, 2019a] Staff, R. (2019a). Asset manager of the year : Goldman sachs asset management. Retrieved from : https ://www.risk.net/awards/6148931/asset-manager-of-the-year-goldman-sachs-asset-management.

[Staff, 2019b] Staff, R. (2019b). Equity derivatives house of the year : Bank of america. Retrieved from : https ://www.risk.net/awards/7180756/equity-derivatives-house-of-the-year-bofa-securities.

[Thomas, 2019] Thomas, P.-H. (2019). Votre épargne bientôt gérée par les robots ? *Trends-Tendance*, (48) :42 – 44.

[Tilakaratne, 2004] Tilakaratne, C. (2004). A neural network approach for predicting the direction of the australian stock market index. Master's thesis, School of Information Technology and Mathematical Sciences, University of Ballarat, Australia.

[Van Maldegem, 2020] Van Maldegem, P. (2020). Interview de nicola horlick. *L'Echo*.

[Wigglesworth, 2020] Wigglesworth, R. (2020). Stockpickers turn to big data to arrest decline. *Financial Times*.

[zonebourse, 2017] zonebourse (2017). Technipfmc obtient le feu vert de l'amf pour son entrée en bourse. Retrieved from : https ://www.zonebourse.com/TECHNIP-33326591/actualite/TechnipFMC-obtient-le-feu-vert-de-l-AMF-pour-son-entree-en-Bourse-23696050/.

LOUVAIN-LA-NEUVE | BRUXELLES | MONS | TOURNAI | CHARLEROI | NAMUR
Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/lsm**