
Rapport Pôle Projet

Projet 04 : Modèles Graphiques Parcimonieux pour la Modélisation Financière
Partie 02 : Optimisation Financière par la Data

ADIB AYMANE CHAOUI | BHEDDAR ZAKARIA
BOUMOUSSOU YOUNES | EL ACHKAR SALMA
PILORGET MAXIME | ELYAZIDI NABIL

DEUXIÈME ANNÉE CYCLE D'INGÉNIEUR
PÔLE DATA SCIENCE
SEMESTRE 7

Janvier 2025

Résumé

Dans un contexte financier marqué par des incertitudes croissantes, l'optimisation des portefeuilles d'investissement et la gestion des risques représentent des défis cruciaux. Ce projet explore des approches innovantes mêlant la théorie moderne du portefeuille, le machine learning, et l'analyse des sentiments financiers. En intégrant des techniques telles que les modèles graphiques parcimonieux, l'optimisation basée sur le ratio de Sharpe, et l'utilisation de FinBERT pour l'analyse des sentiments, nous avons développé une méthodologie robuste et pratique pour une meilleure prise de décision en investissement. Les résultats incluent des simulations de Monte Carlo, une frontière efficiente et des stratégies testées à travers des périodes de crises majeures, illustrant la résilience et l'efficacité des approches proposées. Cette recherche met également en avant un outil web interactif, rendant accessible l'analyse financière avancée à un public varié.

Table des matières

1	Introduction	5
2	Travail Réalisé en S6 - Récapitulatif	7
2.1	Introduction aux Modèles Graphiques Gaussiens	7
2.2	Analyse des Séries Temporelles	7
2.3	Optimisation de la Parcimonie : Estimateur de Lasso	7
2.4	Applications Pratiques	7
2.5	Perspectives	8
3	Fondement Théorique	9
3.1	Portefeuille : Espérance, risque et variance	9
3.1.1	Paramètres individuels des actifs	9
3.1.2	Interactions entre les actifs	10
3.1.3	Construction du portefeuille	10
3.2	Ratio de Sharpe	11
3.2.1	Définition mathématique	11
3.2.2	Interprétation du ratio	12
3.2.3	Limites du ratio de Sharpe	12
3.2.4	Application pratique	12
3.3	Conditions de Karush-Kuhn-Tucker	12
3.3.1	Formulation du problème	13
3.3.2	Théorème	13
3.3.3	Interprétation géométrique	14
3.3.4	Hypothèses de régularité	14
3.3.5	Application pratique	14
3.4	Indicateurs Financiers	15
3.4.1	Ratio P/E (Price-to-Earnings Ratio)	15
3.4.2	Debt/Equity Ratio	16
3.4.3	ROE and ROA	16
3.4.4	Dividend Yield	17
3.4.5	Le ratio de Calmar	17
3.4.6	Bêta	17
3.5	Analyse de Sentiment : FinBERT	18
3.5.1	Contexte et Motivation	18
3.5.2	De BERT à FinBERT	19
3.5.3	Spécialisation Financière de FinBERT	19
3.5.4	Processus d'Inférence	20
3.5.5	Évaluation et Métriques	21
3.5.6	Limites et Considérations	21
3.6	Algorithme de la Forêt Aléatoire	21
3.6.1	Introduction	21

3.6.2	Qu'est-ce qu'un arbre de décision ?	21
3.6.3	Fonctionnement de la Forêt Aléatoire	23
3.6.4	Construction des arbres individuels	23
3.6.5	Propriétés statistiques	24
3.6.6	Importance des variables	24
3.6.7	Hyperparamètres clés	25
3.6.8	Avantages et limitations	25
3.6.9	Mise en œuvre pratique	25
3.7	PCA et Affinity Propagation : Projet S6	26
3.7.1	Analyse en Composantes Principales (PCA)	27
3.7.2	Affinity Propagation	27
3.7.3	Synergie dans le Projet	27
4	Mise en œuvre pratique	28
4.1	Optimisation des portefeuilles	28
4.2	Techniques étudiées	28
4.2.1	Maximisation du ratio de Sharpe	28
4.2.2	Minimisation du risque	28
4.2.3	Optimisation pour un rendement cible	29
4.3	Résultats de l'Étude	29
4.3.1	Sélection des Données et Paramètres Initiaux	29
4.3.2	Analyse des Corrélations	31
4.3.3	Calcul des Rendements Annualisés	32
4.3.4	Simulation Monte Carlo et Frontière Efficiente	32
4.4	Optimisation et Amélioration des Résultats	34
4.4.1	Intégration du Taux Sans Risque	34
4.5	Backtesting des stratégies	35
4.5.1	Test en période de crise (2007-2011)	37
4.5.2	Test en période de pandémie (2021-2023)	39
4.6	Analyse des sentiments	42
4.6.1	Introduction	42
4.6.2	Architecture du pipeline	42
4.6.3	Méthodologie détaillée	43
4.6.4	Validation du modèle	43
4.6.5	Étude de cas : Analyse de Tesla (TSLA)	43
4.6.6	Gestion des limitations techniques	45
5	Résultats du Projet	47
5.1	Motivation	47
5.2	Tech Stack de la Solution	47
5.2.1	Backend	47
5.2.2	Frontend	48
5.3	Présentation de la Solution et guide d'utilisation	48

5.3.1	Démarrage de l'application	48
5.3.2	Architecture et fonctionnalités	49
5.3.3	Critiques et Perspectives d'Amélioration	52
5.4	Résultats du modèle de Forêt Aléatoire	54
5.4.1	Création des variables explicatives et de la cible	54
5.4.2	Optimisation et sélection du modèle	55
5.4.3	Calcul du profit et pertes (P&L)	55
5.4.4	Résumé des résultats	55
6	Répertoire GitHub	56
7	Conclusion	57

1 Introduction

Dans un contexte financier en constante évolution, caractérisé par une augmentation exponentielle des données disponibles et une complexité accrue des marchés, l'optimisation des portefeuilles d'investissement et la gestion des risques représentent des défis stratégiques. Les investisseurs, qu'ils soient particuliers ou institutionnels, doivent faire face à des décisions complexes impliquant des incertitudes croissantes et des facteurs macroéconomiques fluctuants.



FIGURE 2 – Turbulences sur les marchés financiers.

L'émergence des technologies numériques, couplée à l'évolution des capacités de calcul et à la disponibilité de techniques avancées d'apprentissage automatique, offre aujourd'hui des outils puissants pour dépasser les limites des modèles traditionnels. La théorie moderne du portefeuille, développée par Harry Markowitz en 1952, reste une pierre angulaire en matière de diversification et d'optimisation rendement-risque. Cependant, ses hypothèses fondamentales, telles que la normalité des rendements et la stabilité des corrélations entre actifs, se heurtent souvent aux réalités des marchés financiers, particulièrement en période de crise où les interactions entre actifs deviennent hautement instables.

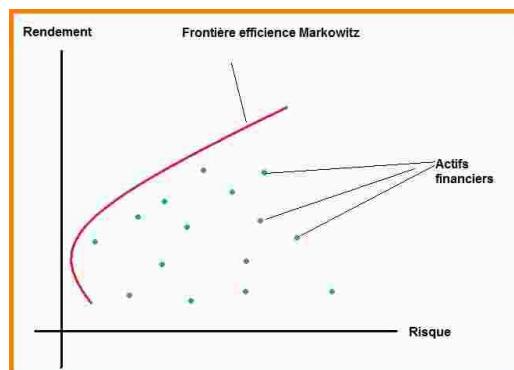


FIGURE 3 – Illustration de la frontière efficiente selon la théorie moderne du portefeuille¹.

1. Ce graphe illustre la frontière efficiente de Markowitz, qui regroupe les portefeuilles offrant le meilleur

Dans ce contexte, l'intégration des techniques modernes de machine learning et de big data s'avère cruciale. Ces approches permettent non seulement une meilleure estimation des paramètres fondamentaux tels que les rendements espérés et les matrices de covariance, mais aussi l'identification de structures complexes au sein des données financières. En parallèle, l'analyse des sentiments financiers, alimentée par des modèles comme FinBERT, ouvre de nouvelles perspectives pour inclure des informations qualitatives issues de l'actualité économique et des tendances du marché.

Le présent rapport s'inscrit dans cette dynamique en explorant des méthodologies innovantes qui combinent théorie moderne du portefeuille, apprentissage automatique et analyse des sentiments. Plus précisément, nous nous appuyons sur des modèles graphiques parcimonieux pour capturer les relations entre actifs, sur l'optimisation basée sur le ratio de Sharpe pour maximiser les performances ajustées au risque, et sur des simulations de Monte Carlo pour évaluer la résilience des stratégies proposées.

Les résultats de ce travail incluent des applications pratiques sur des données réelles, une frontière efficiente adaptée aux conditions de marché, ainsi que des tests robustes en périodes de crise et de forte volatilité. Par ailleurs, un outil web interactif a été développé pour rendre accessibles ces analyses complexes à un large public, facilitant ainsi la prise de décision en investissement. Ce projet vise ainsi à offrir une méthodologie solide et polyvalente pour relever les défis actuels de la finance moderne.

rendement pour un niveau de risque donné. Les points représentent des actifs financiers, et ceux sous la courbe sont considérés comme sous-optimaux.

2 Travail Réalisé en S6 - Récapitulatif

Au cours du semestre S6, notre projet intitulé “*Modèles Graphiques Parcimonieux pour la Modélisation Financière*” a abordé plusieurs aspects théoriques et pratiques de la modélisation des données financières à l'aide de modèles graphiques gaussiens. Voici une synthèse des activités principales menées :

2.1 Introduction aux Modèles Graphiques Gaussiens

Nous avons exploré les notions fondamentales des modèles graphiques gaussiens, qui permettent de représenter les relations statistiques entre variables sous forme de graphes. Les principaux points abordés incluent :

- La définition des graphes non orientés et leur utilité pour encoder les relations d'indépendance conditionnelle.
- L'estimation de la matrice de covariance inverse (matrice de précision) à l'aide de différentes méthodes : Maximum de Vraisemblance, Graphical Lasso et rétrécissement Ledoit-Wolf.
- Une comparaison des performances des différents estimateurs.

2.2 Analyse des Séries Temporelles

L'étude des séries temporelles a été un aspect central de notre travail. Nous avons traité :

- La notion de stationnarité et son importance dans la modélisation des données temporelles.
- Les méthodes pour estimer la tendance d'une série à l'aide de moyennes mobiles, tant empiriques que paramétriques.
- Des applications pratiques comprenant des tests de stationnarité (statistiques roulantes, test de Dickey-Fuller augmenté) et la stationnarisation des séries non stationnaires.

2.3 Optimisation de la Parcimonie : Estimateur de Lasso

Nous avons optimisé l'hyperparamètre λ pour le Graphical Lasso, à travers :

- La validation croisée (cross-validation) pour évaluer la performance du modèle sur des données de test.
- La technique de division Train-Test pour mesurer les performances sur des échantillons disjoints.

Ces étapes nous ont permis de développer une matrice de précision parcimonieuse représentant fidèlement les corrélations significatives.

2.4 Applications Pratiques

- **Clustering avec Affinity Propagation** : Nous avons regroupé les résultats du Graphical Lasso à l'aide de l'algorithme Affinity Propagation, permettant une classification

automatique et robuste des données.

- **Optimisation de Portefeuille** : Nous avons appliqué la théorie de Markowitz pour optimiser un portefeuille financier. Les simulations ont été effectuées sur des actifs réels et ont mis en évidence les relations entre corrélations, risque et rendement.

2.5 Perspectives

En conclusion, ce travail a jeté les bases d'une modélisation efficace et parcimonieuse des données financières. Les pistes futures incluent l'utilisation de ces modèles pour la prédiction des prix d'actifs financiers, avec des approches telles que les modèles ARIMA ou d'autres techniques de prédiction avancées.

3 Fondement Théorique

3.1 Portefeuille : Espérance, risque et variance

Dans le cadre de la théorie moderne du portefeuille, la compréhension des paramètres fondamentaux qui régissent le comportement des actifs financiers est cruciale pour toute démarche d'optimisation. Cette section présente les concepts clés et leur interprétation économique. [6]

3.1.1 Paramètres individuels des actifs

Rendement espéré (μ_i)

Le rendement espéré d'un actif i est défini comme la valeur moyenne anticipée de ses performances futures :

$$\mu_i = \mathbb{E}[R_i],$$

où R_i représente le rendement aléatoire de l'actif i .

Les méthodes couramment utilisées pour estimer le rendement espéré incluent :

- **Moyennes historiques** : Calculées sur des horizons temporels spécifiques (mensuels, annuels).
- **Prévisions d'analystes financiers** : Basées sur des évaluations qualitatives et quantitatives des entreprises.
- **Modèles économétriques** : Par exemple, les modèles de séries temporelles tels qu'ARIMA ou GARCH.

Impact des conditions de marché : Les rendements espérés peuvent être significativement influencés par les phases de marché. Par exemple, pendant une crise, les prévisions de rendement sont souvent révisées à la baisse pour refléter l'augmentation de l'incertitude.

Volatilité (σ_i)

La volatilité mesure l'ampleur des fluctuations des rendements d'un actif et constitue une mesure clé du risque spécifique à cet actif :

$$\sigma_i = \sqrt{\text{Var}(R_i)},$$

où $\text{Var}(R_i)$ représente la variance des rendements.

Caractéristiques de la volatilité :

- **Stabilité relative** : La volatilité est généralement plus stable dans le temps que les rendements espérés.
- **Augmentation en période de stress** : Les périodes de crise financière entraînent une hausse notable de la volatilité, rendant les actifs plus risqués.

3.1.2 Interactions entre les actifs

Les interactions entre les rendements des différents actifs jouent un rôle crucial dans l'optimisation des portefeuilles. Elles sont principalement mesurées à l'aide des covariances et corrélations entre les actifs.

Covariance (σ_{ij}) : La covariance entre les rendements de deux actifs i et j est définie par :

$$\sigma_{ij} = \mathbb{E}[(R_i - \mu_i)(R_j - \mu_j)],$$

où R_i et R_j sont les rendements des actifs i et j , et μ_i, μ_j leurs rendements espérés.

Caractéristiques de la covariance :

- **Diversification** : Une covariance faible ou négative entre deux actifs réduit le risque global du portefeuille.
- **Instabilité** : Les covariances varient en fonction des conditions de marché. Par exemple, elles augmentent souvent lors des crises financières (effet de contagion).

Matrice de variance-covariance (Σ) : Les covariances entre n actifs sont regroupées dans une matrice symétrique Σ :

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{bmatrix}.$$

Corrélation (ρ_{ij}) : La corrélation, normalisée entre -1 et 1 , est donnée par :

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}.$$

Une corrélation proche de 1 indique des rendements étroitement liés, tandis qu'une corrélation négative reflète une relation inverse.

3.1.3 Construction du portefeuille

L'objectif principal de la gestion de portefeuille est d'optimiser la répartition des investissements entre différents actifs pour atteindre un équilibre entre risque et rendement.

Vecteur des poids (w) : Un portefeuille est défini par un vecteur de poids $w = (w_1, w_2, \dots, w_n)^T$, où w_i représente la proportion du capital investi dans l'actif i .

Contraintes budgétaires : Les poids doivent satisfaire les contraintes suivantes :

$$\sum_{i=1}^n w_i = 1 \quad (\text{budget total}).$$

Selon le contexte, d'autres contraintes peuvent s'ajouter :

- **Effet de levier** : Les poids peuvent excéder 1.
- **Ventes à découvert** : Les poids peuvent être négatifs.

Rendement du portefeuille ($E[R_p]$) : Le rendement espéré du portefeuille est une combinaison pondérée des rendements espérés des actifs :

$$E[R_p] = \sum_{i=1}^n w_i \mu_i = w^T \mu,$$

où $\mu = (\mu_1, \mu_2, \dots, \mu_n)^T$ est le vecteur des rendements espérés.

Risque du portefeuille ($\text{Var}(R_p)$) : Le risque (variance) du portefeuille est donné par :

$$\text{Var}(R_p) = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_{ij} = w^T \Sigma w,$$

où Σ est la matrice de variance-covariance.

La construction optimale d'un portefeuille repose sur l'équilibre entre le rendement attendu et le risque mesuré par la volatilité, en tenant compte des interactions entre les actifs.

3.2 Ratio de Sharpe

Le ratio de Sharpe, introduit par William F. Sharpe en 1966, est une mesure standardisée utilisée pour évaluer la performance d'un portefeuille ajustée au risque. Il compare le rendement excédentaire d'un portefeuille (par rapport à un actif sans risque) à sa volatilité. [7]

3.2.1 Définition mathématique :

Le ratio de Sharpe est défini par la formule suivante :

$$S = \frac{R_p - R_f}{\sigma_p},$$

où :

- R_p : Rendement moyen du portefeuille.
- R_f : Taux sans risque, représentant le rendement d'un actif sans risque (par exemple, les obligations d'État).
- σ_p : Écart-type des rendements du portefeuille, mesurant sa volatilité.

3.2.2 Interprétation du ratio :

- $S > 1$: Performance ajustée au risque considérée comme acceptable.
- $S > 2$: Performance très bonne.
- $S > 3$: Performance excellente.
- $S < 0$: Rendement inférieur au taux sans risque.

3.2.3 Limites du ratio de Sharpe :

- **Normalité des rendements** : Le ratio suppose que les rendements suivent une distribution normale, ce qui peut être incorrect dans des marchés financiers volatils.
- **Volatilité comme seule mesure de risque** : Le ratio ne prend pas en compte d'autres mesures de risque comme le maximum drawdown.
- **Non-linéarité des stratégies** : Peu adapté aux portefeuilles utilisant des instruments financiers complexes ou des options.

3.2.4 Application pratique :

Le ratio de Sharpe est largement utilisé dans :

- **Comparaison de portefeuilles** : Pour identifier le portefeuille offrant le meilleur rendement ajusté au risque.
- **Optimisation de portefeuille** : Maximiser S est un objectif central dans la théorie moderne du portefeuille.
- **Évaluation des gestionnaires** : Permet de mesurer la qualité d'un gestionnaire de portefeuille en termes de rendement ajusté au risque.

Extensions du ratio : Plusieurs variantes du ratio de Sharpe ont été développées pour pallier ses limitations :

- **Ratio de Sortino** : Utilise uniquement la volatilité des rendements négatifs comme mesure du risque.
- **Ratio de Treynor** : Remplace la volatilité totale par le bêta, mesurant le risque systématique.
- **Information Ratio** : Compare la surperformance du portefeuille par rapport à un indice de référence.

Malgré ces limitations, le ratio de Sharpe reste un outil fondamental dans l'analyse de portefeuille, permettant de comparer efficacement différentes stratégies d'investissement sur une base standardisée.

3.3 Conditions de Karush-Kuhn-Tucker

Les conditions de Karush-Kuhn-Tucker (KKT) sont une généralisation des multiplicateurs de Lagrange permettant de résoudre des problèmes d'optimisation avec des contraintes d'inégalités non linéaires. [1]

3.3.1 Formulation du problème

Soit $f : \mathbb{R}^n \rightarrow \mathbb{R}$ une fonction appelée fonction objectif, $g_j : \mathbb{R}^n \rightarrow \mathbb{R}, 1 \leq j \leq m$ et $h_j : \mathbb{R}^n \rightarrow \mathbb{R}, 1 \leq j \leq p$ des fonctions appelées contraintes. On suppose que f , les g_j et les h_j sont des fonctions de classe C^1 .

Le problème à résoudre est le suivant :

Trouver $x^* \in \mathbb{R}^n$ qui minimise f sous les contraintes $g_i(x^*) \leq 0$ et $h_j(x^*) = 0$ pour tout i, j

- $f : \mathbb{R}^n \rightarrow \mathbb{R}$ est la fonction objectif
- $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ sont les contraintes d'inégalité
- $h_j : \mathbb{R}^n \rightarrow \mathbb{R}$ sont les contraintes d'égalité

3.3.2 Théorème

Si f admet un minimum en x^* sous les contraintes $g_j(x^*) \leq 0$ pour tout j , alors il existe $(\lambda_j)_{1 \leq j \leq m} \in \mathbb{R}^m$ satisfaisant les conditions suivantes, connues sous le nom de conditions de Karush-Kuhn-Tucker. Les λ_j sont appelés multiplicateurs de Lagrange associés à la j -ième contrainte.

Conditions du premier ordre (Stationnarité)

Le point x^* est un point critique de

$$L_\lambda : x \mapsto f(x) + \sum_{j=1}^m \lambda_j g_j(x) + \sum_{j=1}^p \mu_j h_j(x),$$

le Lagrangien du problème. En d'autres termes, le gradient du Lagrangien s'annule en ce point :

$$\nabla L_\lambda(x^*) = 0,$$

où ∇ désigne le gradient, ou plus explicitement en termes de dérivées partielles,

$$\frac{\partial f}{\partial x_k}(x^*) + \sum_{j=1}^m \lambda_j \frac{\partial g_j}{\partial x_k}(x^*) + \sum_{j=1}^p \mu_j \frac{\partial h_j}{\partial x_k}(x^*) = 0, \quad 1 \leq k \leq n.$$

Conditions de complémentarité

Non-négativité des multiplicateurs

Pour tout $1 \leq j \leq m$,

$$\lambda_j \geq 0,$$

Complémentarité

Pour tout $1 \leq j \leq m$,

$$\lambda_j = 0 \quad \text{ou} \quad g_j(x^*) = 0.$$

On peut également exprimer cela de manière plus compacte comme suit :

$$\min[\lambda_j, |g_j(x^*)|] = 0.$$

Faisabilité primale

$$\begin{aligned} g_i(x^*) &\leq 0, \quad \forall i = 1, \dots, m \\ h_j(x^*) &= 0, \quad \forall j = 1, \dots, p \end{aligned}$$

3.3.3 Interprétation géométrique

Signification géométrique :

- Les gradients des contraintes actives forment un cône tangent à l'ensemble faisable au point x^* .
- Le gradient de la fonction objectif est orthogonal à ce cône.

Illustration graphique : Dans un problème à deux dimensions, cette condition peut être visualisée comme le point où les courbes de niveau de $f(x)$ sont tangentes à l'ensemble des contraintes faisables.

3.3.4 Hypothèses de régularité

Pour que les conditions KKT soient nécessaires et suffisantes, certaines hypothèses de régularité doivent être satisfaites. Ces hypothèses garantissent l'existence et la validité des multiplicateurs de Lagrange.

Conditions principales :

- **Continuité** : Les fonctions $f(x)$, $g_i(x)$, et $h_j(x)$ doivent être continûment différentiables.
- **Qualification des contraintes (LICQ)** : Les gradients des contraintes actives doivent être linéairement indépendants :

$\{\nabla g_i(x^*), \nabla h_j(x^*)\}$ sont linéairement indépendants.

- **Convexité** : Si le problème est convexe (fonction objectif convexe et contraintes convexes), les conditions KKT sont suffisantes pour garantir l'optimalité globale.

Impact des violations des hypothèses : En cas de non-satisfaction des hypothèses (par exemple, si les gradients des contraintes sont dépendants), les conditions KKT peuvent ne pas identifier toutes les solutions optimales.

3.3.5 Application pratique

Les conditions KKT sont largement utilisées dans la résolution de problèmes d'optimisation non linéaire avec contraintes. Voici quelques exemples pratiques de leur application :

1. Optimisation de portefeuille : Dans le contexte financier, les conditions KKT permettent de résoudre des problèmes comme la minimisation du risque pour un rendement cible donné :

$$\begin{aligned} \min \quad & w^T \Sigma w \\ \text{s.c.} \quad & w^T \mu = R_{\text{target}}, \\ & \sum_{i=1}^n w_i = 1, \\ & w_i \geq 0, \forall i. \end{aligned}$$

Les conditions KKT aident à identifier les poids optimaux w^* .

2. Algorithmes numériques : De nombreux algorithmes, comme la méthode des points intérieurs ou SLSQP (Sequential Least Squares Programming), utilisent les conditions KKT pour résoudre des problèmes d'optimisation sous contraintes.

3. Vérification de l'optimalité : Les conditions KKT permettent de vérifier si une solution candidate x^* est effectivement un optimum local. Par exemple, en finance, elles peuvent être utilisées pour valider les solutions d'optimisation de portefeuille obtenues à l'aide de méthodes numériques.

Les conditions KKT fournissent un cadre robuste pour résoudre des problèmes complexes d'optimisation. Elles sont au cœur de nombreuses applications en finance, en ingénierie et en science des données.

3.4 Indicateurs Financiers

Les indicateurs financiers représentent des outils essentiels pour effectuer une analyse approfondie des performances économiques et évaluer la santé financière des entreprises. Ils fournissent aux investisseurs, analystes et gestionnaires de portefeuille des informations quantitatives nécessaires à une prise de décision éclairée. Cette section propose une présentation détaillée des principaux ratios financiers utilisés dans le secteur.

3.4.1 Ratio P/E (Price-to-Earnings Ratio)

Le ratio cours/bénéfice, couramment appelé ratio P/E, est un indicateur clé qui évalue la valorisation d'une entreprise en fonction de sa capacité à générer des bénéfices.

$$P/E = \frac{\text{Prix par action}}{\text{Bénéfice par action}} \tag{1}$$

— Interprétation :

- Un ratio P/E élevé (supérieur à 20) reflète des anticipations de forte croissance future de la part des investisseurs.

- Un ratio P/E faible (inférieur à 10) peut indiquer une sous-évaluation de l'entreprise ou révéler des difficultés structurelles.
- La moyenne historique du SP 500 oscille généralement entre 15 et 17.

— **Limitations :**

- Ne prend pas en compte le niveau d'endettement de l'entreprise.
- S'avère peu pertinent pour évaluer les entreprises déficitaires.
- Peut varier considérablement selon les secteurs d'activité.

3.4.2 Debt/Equity Ratio

Le ratio Dette/Capitaux Propres mesure le niveau d'endettement d'une entreprise par rapport à ses fonds propres, offrant ainsi un aperçu de sa structure financière.

$$\text{Debt/Equity} = \frac{\text{Total des Dettes}}{\text{Capitaux Propres}} \quad (2)$$

— **Composantes :**

- Dette totale = Dette à court terme + Dette à long terme
- Capitaux propres = Actions + Bénéfices non distribués

— **Analyse :**

- Ratio < 1 : Structure financière conservatrice
- Ratio > 2 : Niveau d'endettement potentiellement risqué
- L'interprétation dépend fortement du secteur d'activité

3.4.3 ROE and ROA

Ces deux indicateurs mesurent l'efficacité avec laquelle une entreprise utilise ses ressources pour générer des profits.

$$\text{ROE (Return on Equity)} = \frac{\text{Bénéfice Net}}{\text{Capitaux Propres}} \times 100\% \quad (3)$$

$$\text{ROA (Return on Assets)} = \frac{\text{Bénéfice Net}}{\text{Total des Actifs}} \times 100\% \quad (4)$$

— **ROE :**

- Mesure la rentabilité des fonds propres
- ROE $> 15\%$ généralement considéré comme excellent
- Peut être décomposé selon la formule DuPont :

$$\text{ROE} = \text{Marge Nette} \times \text{Rotation des Actifs} \times \text{Levier Financier} \quad (5)$$

— **ROA :**

- Évalue l'efficacité de l'utilisation des actifs
- Particulièrement pertinent pour comparer des entreprises du même secteur
- ROA $> 5\%$ généralement considéré comme bon

3.4.4 Dividend Yield

Le rendement du dividende mesure le rapport entre le dividende versé et le cours de l'action, important pour les investisseurs recherchant des revenus réguliers.

$$\text{Dividend Yield} = \frac{\text{Dividende Annuel par Action}}{\text{Prix de l'Action}} \times 100\% \quad (6)$$

— **Caractéristiques :**

- Indicateur clé pour les stratégies de revenus
- Varie selon la maturité de l'entreprise et le secteur
- Doit être analysé avec le taux de distribution (Payout Ratio)

— **Ratio de Distribution :**

$$\text{Payout Ratio} = \frac{\text{Dividendes Totaux}}{\text{Bénéfice Net}} \times 100\% \quad (7)$$

Points Clés à Retenir

- Les ratios financiers doivent être analysés dans leur ensemble et non isolément
- La comparaison sectorielle est essentielle pour une interprétation pertinente
- L'évolution temporelle des ratios est aussi importante que leur valeur absolue
- Les spécificités du secteur d'activité influencent l'interprétation des ratios

3.4.5 Le ratio de Calmar

Le ratio de Calmar est un indicateur de performance qui mesure le rendement ajusté au risque d'un investissement, en considérant le drawdown maximum comme mesure du risque.

$$\text{Ratio de Calmar} = \frac{\text{Rendement annualisé moyen}}{\text{Maximum Drawdown absolu}} \quad (8)$$

— **Interprétation :**

- Un ratio supérieur à 1 indique une bonne performance ajustée au risque
- Plus le ratio est élevé, meilleure est la performance par unité de risque
- Particulièrement utile pour évaluer les fonds spéculatifs

— **Limitations :**

- Ne considère que le drawdown maximum sur la période
- Sensible à la période d'observation choisie
- Peut ne pas refléter la volatilité quotidienne

3.4.6 Bêta

Le Bêta (β) est un coefficient fondamental en finance qui mesure la volatilité relative d'un actif par rapport à son marché de référence. Il représente la sensibilité du rendement d'un titre aux variations du marché dans son ensemble.

$$\beta = \frac{\text{Cov}(r_i, r_m)}{\text{Var}(r_m)} = \rho_{i,m} \frac{\sigma_i}{\sigma_m} \quad (9)$$

où :

- r_i représente le rendement de l'actif
- r_m représente le rendement du marché
- $\text{Cov}(r_i, r_m)$ est la covariance entre les rendements de l'actif et du marché
- $\text{Var}(r_m)$ est la variance des rendements du marché
- $\rho_{i,m}$ est le coefficient de corrélation entre l'actif et le marché
- σ_i et σ_m sont respectivement les écarts-types des rendements de l'actif et du marché
- **Interprétation détaillée :**
 - $\beta = 1$: L'actif a la même volatilité que le marché (ex : un ETF suivant le S&P 500)
 - $\beta > 1$: L'actif est plus volatile que le marché
 - Ex : $\beta = 1.5$ signifie qu'une hausse de 1% du marché entraîne en moyenne une hausse de 1.5% de l'actif
 - Typique des valeurs technologiques et des small caps
 - $\beta < 1$: L'actif est moins volatile que le marché
 - Ex : $\beta = 0.5$ signifie qu'une baisse de 1% du marché entraîne en moyenne une baisse de 0.5% de l'actif
 - Caractéristique des valeurs défensives (utilities, biens de consommation de base)
 - $\beta < 0$: L'actif évolue en sens inverse du marché (rare, cas des valeurs refuges comme l'or)
- **Applications pratiques :**
 - Construction de portefeuille : ajustement du risque systématique
 - Évaluation d'actifs : composante essentielle du modèle CAPM
 - Stratégies de couverture : calibration des positions de couverture
- **Limitations et considérations :**
 - Instabilité temporelle : le bêta peut varier significativement dans le temps
 - Dépendance à l'historique : calculé sur des données passées, peut ne pas être prédictif
 - Sensibilité à la période d'estimation : varie selon la fenêtre temporelle choisie
 - Hypothèse de normalité : suppose une distribution normale des rendements
 - Choix de l'indice : le bêta dépend de l'indice de marché choisi comme référence

3.5 Analyse de Sentiment : FinBERT

3.5.1 Contexte et Motivation

L'analyse de sentiment dans le domaine financier présente des défis uniques que les modèles NLP généraux ne parviennent souvent pas à traiter efficacement. Par exemple, une phrase comme "Les bénéfices ont baissé de 10%" peut sembler négative au premier abord, mais pourrait être interprétée positivement dans un contexte où une baisse de 20% était attendue. FinBERT a été spécifiquement conçu pour capturer de telles nuances.[5] [3]

3.5.2 De BERT à FinBERT

BERT (Bidirectional Encoder Representations from Transformers) a introduit une approche révolutionnaire en traitement du langage naturel grâce à sa compréhension véritablement bidirectionnelle. Pour comprendre le fonctionnement de FinBERT, il est essentiel de saisir les innovations fondamentales de BERT :

Bidirectionnalité Contrairement aux modèles traditionnels qui lisent le texte dans une seule direction (de gauche à droite ou vice versa), BERT capture simultanément les relations contextuelles dans les deux directions. Par exemple, dans la phrase "La fusion avec [MASK] a augmenté la valeur de l'action", le modèle utilise les contextes précédents et suivants pour prédire le mot "l'entreprise concurrente".

Architecture Transformer L'architecture Transformer, qui forme la base de FinBERT, repose sur plusieurs composants clés :

1. **Embeddings positionnels** : Chaque mot se voit attribuer deux types d'embeddings :

$$E_{\text{final}} = E_{\text{token}} + E_{\text{position}} \quad (10)$$

où E_{position} encode la position relative du mot dans la séquence.

2. **Mécanisme d'attention multi-tête** : Le cœur de l'architecture évalue les relations entre tous les mots d'une séquence. Pour chaque tête d'attention :

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (11)$$

Cette formule peut être décomposée comme suit :

- QK^T calcule la similarité entre chaque paire de mots
- $\sqrt{d_k}$ normalise les valeurs pour éviter des gradients excessivement élevés
- La fonction softmax convertit les scores en probabilités
- La multiplication avec V génère la représentation pondérée finale

3.5.3 Spécialisation Financière de FinBERT

FinBERT se distingue par trois caractéristiques principales :

1. Pré-entraînement spécialisé Le modèle est initialement entraîné sur un large corpus de documents financiers, incluant :

- Rapports financiers comme les formulaires 10-K et 10-Q
- Articles provenant de médias financiers
- Transcriptions d'appels liés aux résultats financiers

La fonction de perte utilisée pour l'entraînement avec le Masquage de Modèles de Langage (MLM) s'exprime comme suit :

$$\mathcal{L}_{\text{MLM}} = - \sum_{i \in \mathcal{M}} \log P(x_i | \hat{x}) \quad (12)$$

Où :

- \mathcal{M} représente l'ensemble des tokens masqués
- x_i désigne le token réel
- \hat{x} correspond à la séquence contenant des masques

2. Adaptation du Vocabulaire Le vocabulaire du modèle intègre des termes spécifiques au domaine financier :

- Indicateurs financiers comme le ratio P/E ou l'EBITDA
- Expressions de marché telles que bull market et bear market
- Acronymes spécialisés comme IPO ou M&A

3. Ajustement pour l'Analyse de Sentiment Financière Le modèle est ajusté sur des ensembles de données annotées pour l'analyse de sentiment dans des contextes financiers. La prédiction finale repose sur une couche softmax :

$$P(c|h) = \frac{e^{W_c h}}{\sum_{c' \in C} e^{W_{c'} h}} \quad (13)$$

Où :

- h correspond à la représentation finale du token [CLS]
- W_c représente les poids associés à la classe c
- C est l'ensemble des catégories de sentiment

3.5.4 Processus d'Inférence

L'analyse d'un texte donné suit un processus structuré en plusieurs étapes :

1. **Prétraitement** : Transformation du texte brut en une séquence prête pour le modèle :

$$\text{Input} = [\text{CLS}] + \text{TokenizeFinancial}(\text{text}) + [\text{SEP}] \quad (14)$$

2. **Encodage Contextuel** : Passage de la séquence à travers des couches Transformer successives

3. **Classification** : Utilisation de la représentation du token [CLS] pour la prédiction finale

3.5.5 Évaluation et Métriques

Différentes métriques sont utilisées pour évaluer FinBERT :

$$\text{Précision}_c = \frac{\text{VP}_c}{\text{VP}_c + \text{FP}_c} \quad (15)$$

$$\text{Rappel}_c = \frac{\text{VP}_c}{\text{VP}_c + \text{FN}_c} \quad (16)$$

$$F1_c = 2 \times \frac{\text{Précision}_c \times \text{Rappel}_c}{\text{Précision}_c + \text{Rappel}_c} \quad (17)$$

Le sous-script c fait référence à une classe de sentiment (positive, négative, neutre).

Moyenne Macro La performance globale est mesurée par :

$$\text{Macro-}F1 = \frac{1}{|C|} \sum_{c \in C} F1_c \quad (18)$$

3.5.6 Limites et Considérations

- **Biais des Données** : Les biais présents dans les textes financiers utilisés pour l'entraînement peuvent se refléter dans le modèle
- **Évolution Temporelle** : La relation entre les événements et les sentiments peut évoluer au fil du temps
- **Granularité** : Certaines nuances spécifiques à des sous-domaines de la finance peuvent être difficiles à modéliser

3.6 Algorithme de la Forêt Aléatoire

3.6.1 Introduction

La forêt aléatoire est un algorithme d'apprentissage automatique très utilisé qui appartient à la classe des méthodes d'apprentissage ensembliste. Les méthodes d'ensembles combinent les prédictions de plusieurs modèles individuels pour améliorer les performances prédictives globales et la robustesse. La forêt aléatoire est particulièrement appréciée pour sa polyvalence, sa facilité d'utilisation et son efficacité dans les tâches de classification et de régression. Elle exploite la puissance des arbres de décision tout en réduisant leur tendance au surapprentissage. [2]

3.6.2 Qu'est-ce qu'un arbre de décision ?

Un arbre de décision est un modèle d'apprentissage automatique qui partitionne de manière itérative les données en fonction de règles de décision basées sur les valeurs des attributs. À chaque nœud de l'arbre, une condition est appliquée pour diviser les données en sous-ensembles.

L'objectif est de maximiser la séparation des classes (dans les problèmes de classification) ou de minimiser l'erreur (dans les problèmes de régression).

Pour déterminer la meilleure division à chaque nœud, des indicateurs de qualité de scission sont utilisés, notamment :

Indice de Gini : L'indice de Gini mesure l'impureté d'un nœud, c'est-à-dire la probabilité qu'un échantillon soit mal classé si une classe est attribuée au hasard.

$$G(t) = \sum_{k=1}^K p_k(1 - p_k),$$

où p_k représente la proportion d'échantillons appartenant à la classe k dans le nœud t .

Entropie : L'entropie quantifie le niveau d'incertitude ou de mélange des classes dans un nœud.

$$H(t) = - \sum_{k=1}^K p_k \log_2(p_k),$$

où p_k est encore la proportion d'échantillons appartenant à la classe k .

Erreur Quadratique Moyenne (MSE) : Utilisée dans les tâches de régression, la MSE mesure la différence moyenne au carré entre les prédictions et les valeurs réelles :

$$\text{MSE} = \frac{1}{N_t} \sum_{i \in \text{nœud } t} (y_i - \bar{y})^2,$$

où \bar{y} est la moyenne des valeurs cibles des échantillons dans le nœud t .

L'arbre est généré de manière récursive jusqu'à atteindre un critère d'arrêt, comme une profondeur maximale ou un nombre minimal d'échantillons par feuille.

Les principes fondamentaux de la forêt aléatoire peuvent être résumés ainsi :

Bagging (Bootstrap Aggregating) : Chaque arbre de la forêt est entraîné sur un échantillon bootstrap du jeu de données original, c'est-à-dire que le jeu de données est échantillonné aléatoirement avec remplacement. Cela garantit une diversité entre les arbres.

Sélection Aléatoire de Caractéristiques : Lors de la construction de chaque arbre, un sous-ensemble aléatoire de caractéristiques est considéré pour la division à chaque noeud. Cela décourage les arbres et améliore la généralisation.

3.6.3 Fonctionnement de la Forêt Aléatoire

Phase d'Entraînement :

- **Bootstrap Sampling** : À partir du jeu de données d'entraînement contenant N échantillons, des échantillons bootstrap sont créés en échantillonnant aléatoirement N échantillons avec remplacement. Certains échantillons peuvent apparaître plusieurs fois dans un échantillon bootstrap, tandis que d'autres peuvent être exclus.
 - **Construction des Arbres** : Pour chaque échantillon bootstrap, un arbre de décision est construit. À chaque nœud de l'arbre :
 - Un sous-ensemble aléatoire de m caractéristiques est sélectionné (où $m < M$, avec M étant le nombre total de caractéristiques).
 - La meilleure caractéristique et le seuil sont choisis à partir de ce sous-ensemble en fonction d'un critère de division tel que l'impureté de Gini (pour la classification) ou l'erreur quadratique moyenne (pour la régression).
- L'arbre est construit jusqu'à atteindre un critère d'arrêt spécifié (par exemple, profondeur maximale, nombre minimal d'échantillons par feuille).
- Chaque arbre est entraîné indépendamment, ce qui rend le processus hautement parallélisable.

Phase de Prédiction :

- Pour les tâches de classification, chaque arbre de la forêt produit une classe prédite. La prédiction finale est déterminée par vote majoritaire parmi les arbres :

$$\hat{y} = \operatorname{argmax}_k \sum_{t=1}^T \mathbb{1}(T_t(x) = k),$$

où $T_t(x)$ est la classe prédite pour l'entrée x par l'arbre t , k représente une classe, et $\mathbb{1}$ est la fonction indicatrice.

- Pour les tâches de régression, chaque arbre produit une prédiction numérique, et la prédiction finale est la moyenne des sorties de tous les arbres :

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T T_t(x),$$

où $T_t(x)$ est la prédiction numérique par l'arbre t .

3.6.4 Construction des arbres individuels

Sélection des variables À chaque nœud, un sous-ensemble aléatoire de m variables est considéré pour la division. Typiquement :

- Pour la régression : $m \approx \frac{p}{3}$
- Pour la classification : $m \approx \sqrt{p}$

où p est le nombre total de variables prédictives.

Critère de division Pour la classification, le critère de Gini est souvent utilisé :

$$\text{Gini}(t) = 1 - \sum_{k=1}^K p_{k|t}^2 \quad (19)$$

où :

- t est le nœud considéré
 - K est le nombre de classes
 - $p_{k|t}$ est la proportion d'observations de classe k au nœud t
- Pour la régression, on utilise la réduction de la variance :

$$\text{Var}_{\text{reduction}} = \text{Var}_{\text{parent}} - \left(\frac{n_{\text{gauche}}}{n} \text{Var}_{\text{gauche}} + \frac{n_{\text{droite}}}{n} \text{Var}_{\text{droite}} \right) \quad (20)$$

3.6.5 Propriétés statistiques

Variance de la forêt La variance des prédictions de la forêt est réduite par rapport à celle d'un arbre unique :

$$\text{Var}(F) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \quad (21)$$

où :

- ρ est la corrélation entre les arbres
- σ^2 est la variance d'un arbre individuel
- B est le nombre d'arbres

Out-of-Bag Error L'erreur OOB est estimée sur les observations non incluses dans l'échantillon bootstrap :

$$\text{OOB}_{\text{error}} = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}_{\text{OOB}}(x_i)) \quad (22)$$

où $\hat{f}_{\text{OOB}}(x_i)$ est la prédiction moyenne des arbres n'ayant pas utilisé x_i dans leur échantillon bootstrap.

3.6.6 Importance des variables

Deux mesures principales sont utilisées :

1. Diminution moyenne de l'impureté

$$\text{MDI}_j = \frac{1}{B} \sum_{b=1}^B \sum_{t \in T_b} p(t) \Delta i(s_t, j) \quad (23)$$

où :

- $p(t)$ est la proportion d'observations atteignant le nœud t
- $\Delta i(s_t, j)$ est la réduction d'impureté due à la division s_t sur la variable j

2. Permutation importance

$$\text{PI}_j = \frac{1}{B} \sum_{b=1}^B (\text{OOB}_{\text{error}}^{j,b} - \text{OOB}_{\text{error}}^b) \quad (24)$$

où $\text{OOB}_{\text{error}}^{j,b}$ est l'erreur OOB après permutation aléatoire de la variable j .

3.6.7 Hyperparamètres clés

Les principaux hyperparamètres à optimiser sont :

- B : nombre d'arbres dans la forêt
- m : nombre de variables considérées à chaque division
- n_{\min} : nombre minimal d'observations par feuille
- d_{\max} : profondeur maximale des arbres

3.6.8 Avantages et limitations

Avantages

- **Robustesse** : Résistant au sur-apprentissage grâce au bagging
- **Non-linéarité** : Capture automatiquement les interactions complexes
- **Parallélisation** : Construction des arbres facilement parallélisable
- **Peu de paramètres** : Relativement simple à optimiser

Limitations

- **Interprétabilité** : Moins interprétable qu'un arbre unique
- **Mémoire** : Peut nécessiter beaucoup de mémoire pour de grands ensembles de données
- **Extrapolation** : Difficulté à extrapoler hors de la plage des données d'entraînement

3.6.9 Mise en œuvre pratique

Validation croisée Pour l'optimisation des hyperparamètres, nous avons utilisé une validation croisée *GridSearchCV* avec $K = 3$ plis. La validation croisée permet d'évaluer les performances moyennes sur plusieurs sous-ensembles des données. La formule utilisée pour calculer le score moyen est :

$$\text{CV}_{\text{score}} = \frac{1}{K} \sum_{k=1}^K \text{score}_k \quad (25)$$

Ici, K est le nombre de plis de validation croisée, et score_k correspond au score obtenu sur le k -ième pli. Cette approche garantit une évaluation robuste des performances tout en limitant le surapprentissage.

Évaluation des performances Dans cette tâche de classification, nous avons choisi d'utiliser la F1-score comme métrique principale pour l'évaluation des performances, plutôt que l'accuracy. Cette décision repose sur le fait que la F1-score est mieux adaptée dans les cas où les classes sont déséquilibrées ou lorsque la précision et le rappel doivent être équilibrés. La F1-score est définie comme suit :

$$\text{F1-score} = 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}} \quad (26)$$

Où :

- **Précision (Precision)** : Proportion des prédictions positives correctes parmi toutes les prédictions positives. Elle est définie comme :

$$\text{Précision} = \frac{\text{VP}}{\text{VP} + \text{FP}} \quad (27)$$

où VP représente les vrais positifs et FP les faux positifs.

- **Rappel (Recall)** : Proportion des échantillons positifs correctement identifiés parmi tous les échantillons réellement positifs. Elle est définie comme :

$$\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad (28)$$

où FN représente les faux négatifs.

En utilisant la F1-score comme métrique, nous avons évalué différentes combinaisons d'hyperparamètres pour identifier ceux qui offrent le meilleur compromis entre précision et rappel.

Comparaison avec d'autres métriques Bien que l'accuracy soit une métrique couramment utilisée pour la classification, elle peut être trompeuse en cas de déséquilibre des classes. Pour résumer :

- **Accuracy** :

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(y_i = \hat{y}_i) \quad (29)$$

Cette métrique calcule la proportion d'échantillons correctement classés, mais ne reflète pas les performances si une classe domine largement les données.

- **F1-score** : Comme mentionné, cette métrique est plus adaptée pour des données déséquilibrées et lorsque l'équilibre entre précision et rappel est essentiel.

En conclusion, la F1-score a été choisie pour optimiser les hyperparamètres et évaluer les performances du modèle en raison de sa pertinence pour cette tâche de classification spécifique.

3.7 PCA et Affinity Propagation : Projet S6

L'analyse en composantes principales (PCA) et l'Affinity Propagation constituent deux techniques fondamentales en analyse de données et apprentissage non supervisé, particulièrement pertinentes dans le cadre de ce projet. [4]

3.7.1 Analyse en Composantes Principales (PCA)

L'analyse en composantes principales est une méthode statistique multivariée qui permet de transformer des variables corrélées en nouvelles variables décorrélées appelées composantes principales. Cette technique vise à :

- Réduire la dimensionnalité des données tout en préservant le maximum de variance
- Identifier les directions principales de variabilité dans les données
- Faciliter la visualisation des données multidimensionnelles

Mathématiquement, la PCA recherche les vecteurs propres de la matrice de covariance des données :

$$C = \frac{1}{n-1} X^T X \quad (30)$$

où X représente la matrice de données centrées et n le nombre d'observations.

3.7.2 Affinity Propagation

L'Affinity Propagation est un algorithme de clustering qui identifie des exemplars (points représentatifs) parmi les données en échangeant des messages entre les points. Ses caractéristiques principales sont :

- La détermination automatique du nombre de clusters
- La sélection de points de données réels comme centres de clusters
- L'utilisation d'une matrice de similarité S où $s(i, k)$ représente l'aptitude du point k à servir d'exemplar pour le point i

L'algorithme fonctionne en mettant à jour itérativement deux types de messages :

$$r(i, k) = s(i, k) - \max_{k' \neq k} \{a(i, k') + s(i, k')\} \quad (31)$$

$$a(i, k) = \min \{0, r(k, k) + \sum_{i' \notin \{i, k\}} \max \{0, r(i', k)\}\} \quad (32)$$

où $r(i, k)$ est la responsabilité et $a(i, k)$ est la disponibilité.

3.7.3 Synergie dans le Projet

La combinaison de ces deux techniques permet :

- Une réduction efficace de la dimensionnalité des données par PCA
- Un clustering robuste des données réduites via Affinity Propagation
- Une meilleure compréhension de la structure sous-jacente des données

4 Mise en œuvre pratique

4.1 Optimisation des portefeuilles

4.2 Techniques étudiées

Dans cette section, nous présentons trois approches différentes pour l'optimisation de portefeuille : la maximisation du ratio de Sharpe, la minimisation du risque, et l'optimisation pour un rendement cible.

4.2.1 Maximisation du ratio de Sharpe

La première approche consiste à maximiser le ratio de Sharpe, défini comme :

$$\max_w \frac{R_p - R_f}{\sigma_p} \quad (33)$$

sous les contraintes :

$$\begin{aligned} \sum_{i=1}^n w_i &= 1 \\ 0 \leq w_i &\leq 1, \quad \forall i \in \{1, \dots, n\} \end{aligned}$$

où :

- $w = (w_1, \dots, w_n)$ représente les poids du portefeuille
- R_p est le rendement attendu du portefeuille
- R_f est le taux sans risque
- σ_p est l'écart-type du portefeuille

4.2.2 Minimisation du risque

La deuxième approche vise à minimiser la variance du portefeuille :

$$\min_w \sigma_p^2 = w^\top \Sigma w \quad (34)$$

sous les contraintes :

$$\begin{aligned} \sum_{i=1}^n w_i &= 1 \\ 0 \leq w_i &\leq 1, \quad \forall i \in \{1, \dots, n\} \end{aligned}$$

où Σ représente la matrice de covariance des rendements.

4.2.3 Optimisation pour un rendement cible

La troisième approche consiste à minimiser le risque pour un niveau de rendement cible donné :

$$\min_w \sigma_p^2 = w^\top \Sigma w \quad (35)$$

sous les contraintes :

$$\begin{aligned} w^\top \mu &= R_{target} \\ \sum_{i=1}^n w_i &= 1 \\ 0 \leq w_i &\leq 1, \quad \forall i \in \{1, \dots, n\} \end{aligned}$$

où :

- μ représente le vecteur des rendements espérés
- R_{target} est le rendement cible souhaité

Ces trois problèmes d'optimisation sont résolus numériquement en utilisant la méthode SLSQP (Sequential Least Squares Programming), qui est particulièrement adaptée aux problèmes d'optimisation sous contraintes non linéaires avec des bornes sur les variables.

4.3 Résultats de l'Étude

4.3.1 Sélection des Données et Paramètres Initiaux

Dans cette étude, réalisée à l'aide des notebooks, nous avons sélectionné les tickers suivants :

- AAPL
- MSFT
- GOOGL
- AMZN
- TSLA
- META
- NVDA
- NFLX
- JPM
- V
- SNAP

La période d'analyse s'étend du 1^{er} janvier 2018 au 31 décembre 2022. Par ailleurs, le taux sans risque (`RISK_FREE_RATE`) utilisé dans cette étude est de 0,0619, soit 6,19%. Les prix de clôture (closing prices) de ces actions ont été collectés pour la période spécifiée.

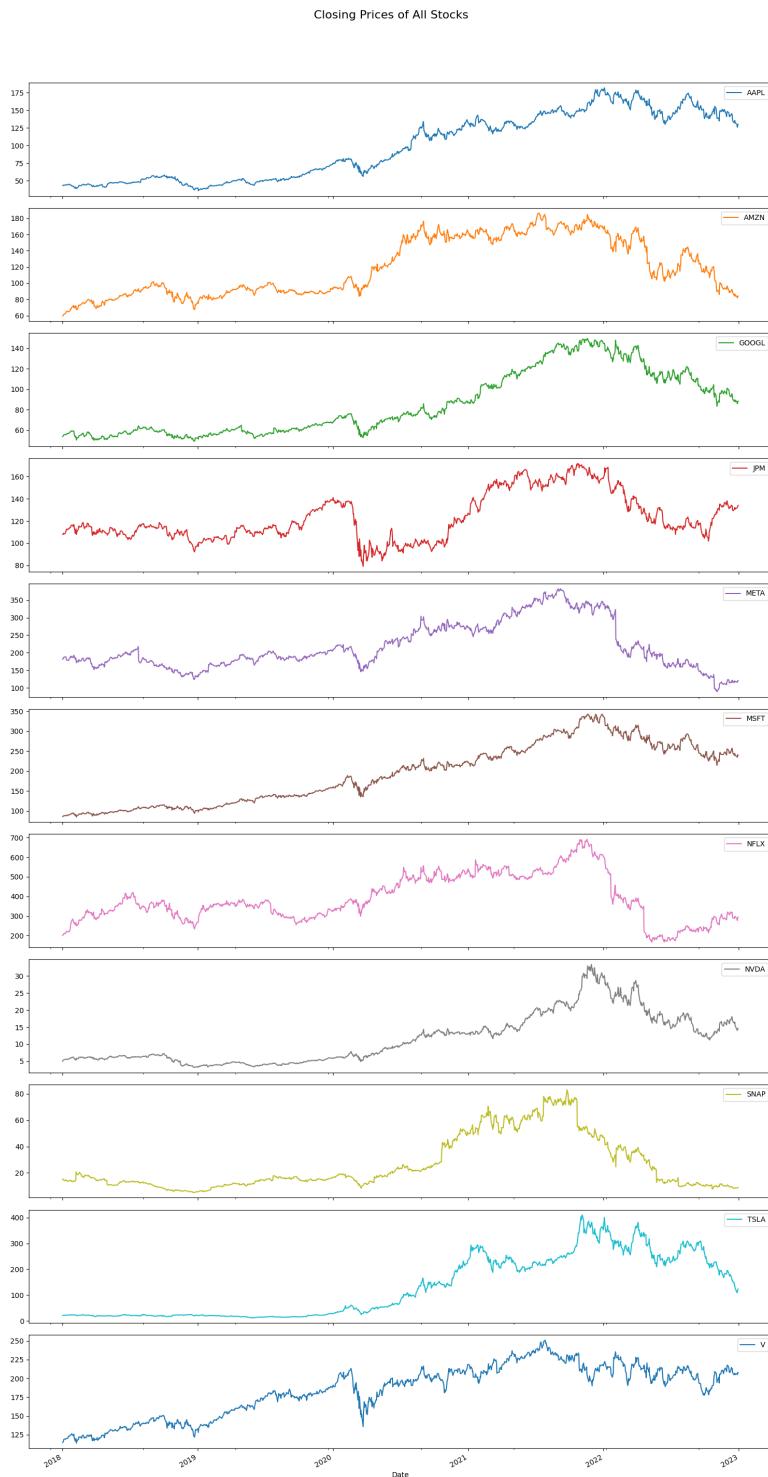


FIGURE 4 – Closing Prices

4.3.2 Analyse des Corrélations

Après avoir transformé ces données en rendements quotidiens (*daily returns*) pour des raisons d'analyse, nous pouvons visualiser les corrélations de deux manières. À droite, nous avons le diagramme de corrélation réel, tandis qu'à gauche, il est filtré en binaire pour ne considérer que les corrélations significatives, où $|\text{correlation}| > 0.68$.



FIGURE 5 – Diagramme de corrélation réel

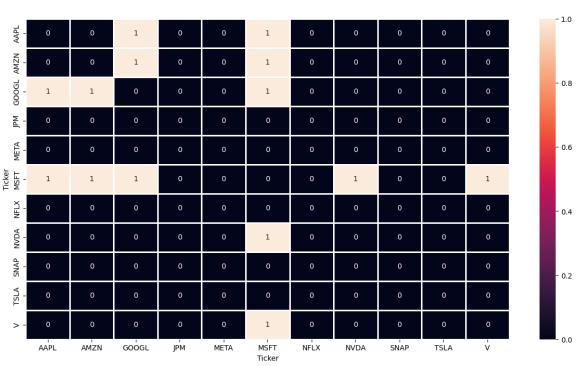


FIGURE 6 – Diagramme de corrélation filtré en binaire $|\text{correlation}| > 0.68$

Les observations montrent que certains tickers classiques, comme MSFT et GOOGL, ainsi qu'AAPL, présentent des corrélations fortes. Cette relation peut également être prédite et visualisée à l'aide d'un diagramme de type *pair plot* des rendements.

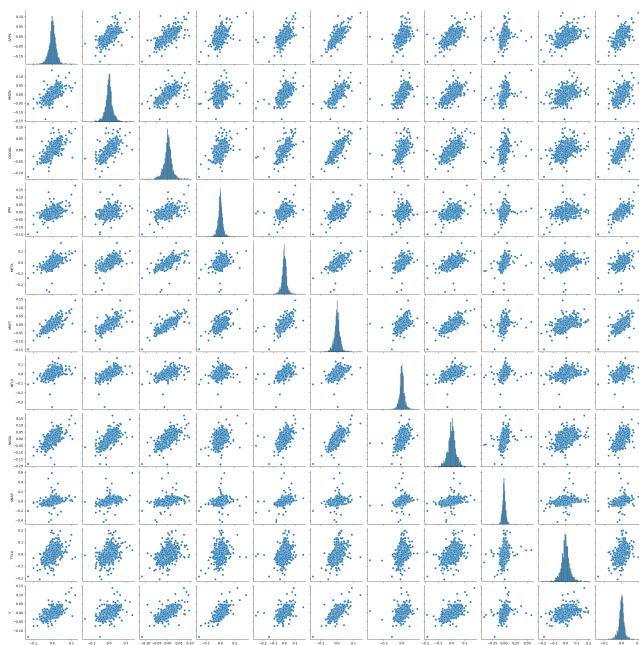


FIGURE 7 – Diagramme pair plot des rendements des tickers sélectionnés

4.3.3 Calcul des Rendements Annualisés

Une fois cette étape réalisée, nous calculons les rendements annualisés pour les intégrer dans notre modèle. Les rendements annualisés obtenus sont les suivants :

Ticker	Rendement Annualisé
AAPL	0,319492
AMZN	0,142585
GOOGL	0,160564
JPM	0,099518
META	0,014319
MSFT	0,288837
NFLX	0,214898
NVDA	0,420153
SNAP	0,237906
TSLA	0,758733
V	0,176419

TABLE 1 – Rendements annualisés par titre

Les formules utilisées pour ces calculs sont :

$$\text{Rendement Annualisé} = (1 + \text{Rendement quotidien moyen})^{\text{TRADING_DAYS_PER_YEAR}} - 1$$

$$\text{Covariance Annualisée} = \text{Covariance des rendements quotidiens} \times \text{TRADING_DAYS_PER_YEAR}$$

4.3.4 Simulation Monte Carlo et Frontière Efficiente

La simulation est réalisée en générant des poids aléatoires pour chaque actif dans le portefeuille, puis en calculant le rendement et le risque (l'écart-type) associés à chaque combinaison de portefeuilles.

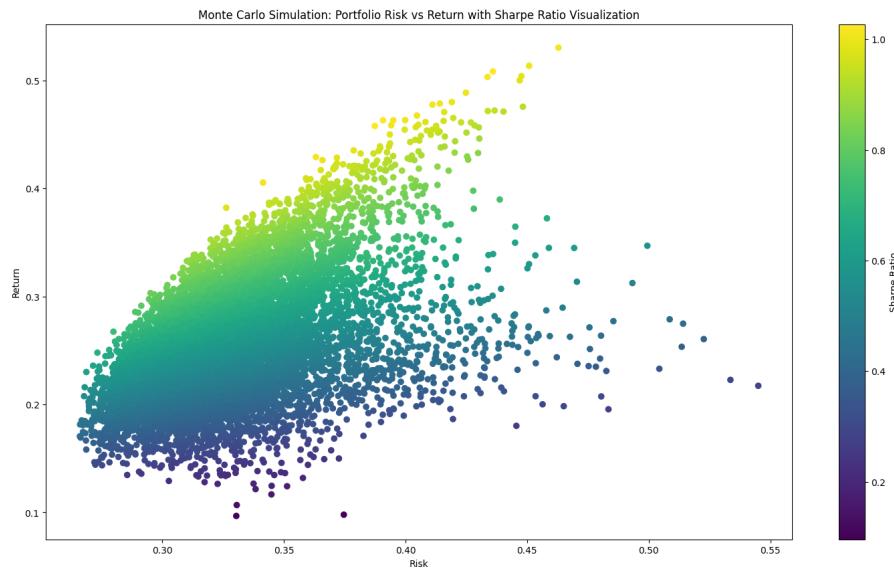


FIGURE 8 – Simulation de Monte Carlo : Portefeuilles générés avec le rendement et le risque associés

Cette frontière représente l'ensemble des portefeuilles qui offrent le rendement le plus élevé pour un niveau donné de risque. Pour la visualiser de manière plus claire, nous avons utilisé l'optimisation par rendement.

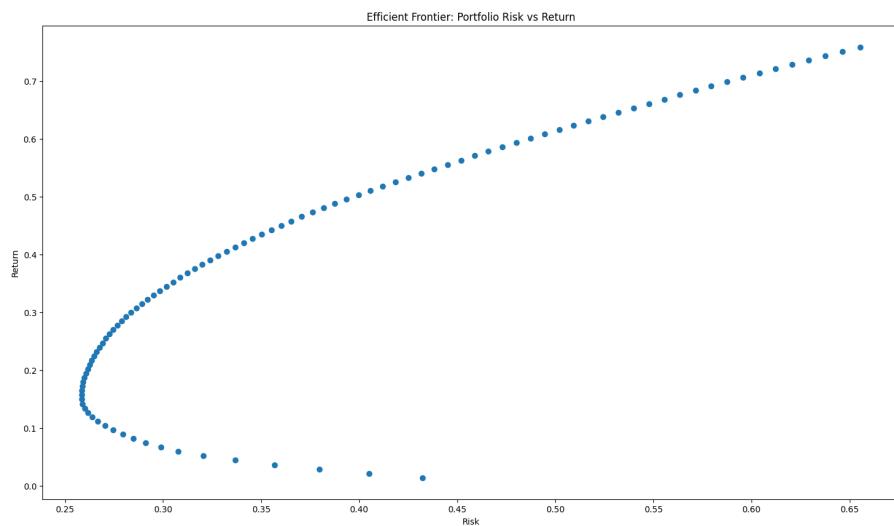


FIGURE 9 – Frontière efficiente obtenue par optimisation par rendement

Ainsi, pour récapituler, nous avons regroupé tous les résultats dans un graphique :

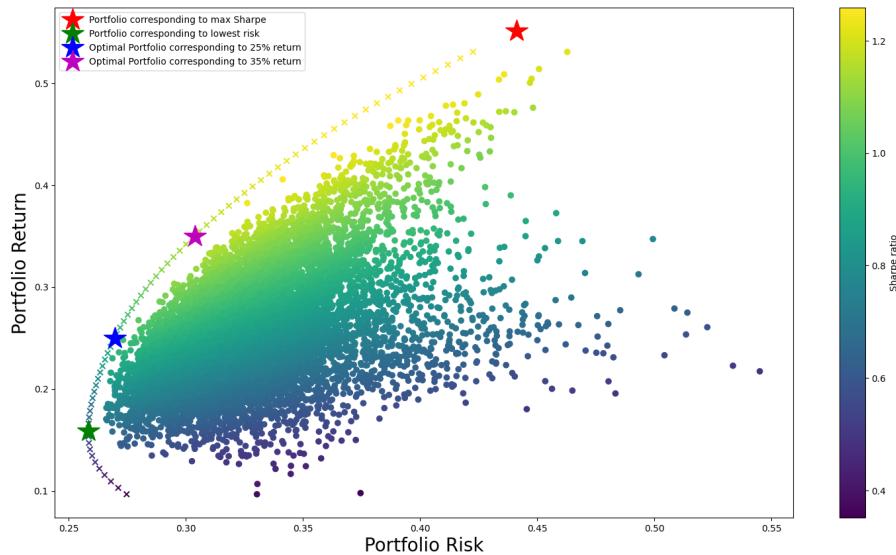


FIGURE 10 – Graphique récapitulatif des résultats

4.4 Optimisation et Amélioration des Résultats

4.4.1 Intégration du Taux Sans Risque

Importance du Taux Sans Risque Le taux sans risque (r_f) est utilisé comme référence dans plusieurs modèles financiers :

- Modèle CAPM : $E(r_i) = r_f + \beta_i(E(r_m) - r_f)$
- Modèle Black-Scholes : $C = S_0 N(d_1) - K e^{-r_f T} N(d_2)$

Implémentation Technique L'approche pour obtenir le taux sans risque se base sur les rendements des bons du Trésor américain à 10 ans, avec :

1. Extraction des données via l'API Yahoo Finance (symbole ^TNX)
2. Calcul de la moyenne des rendements
3. Conversion en décimal
4. Mécanisme de fallback

Analyse Comparative des Périodes Résultats sur différentes périodes :

- 2015-2020 : 2,27%
- 2007-2009 : 4,14%

Cette différence s'explique par :

$$\Delta \text{Politique} = r_{2007-2009} - r_{2015-2020} = 4,14\% - 2,27\% = 1,87\%$$

Implications Pratiques L’analyse démontre l’importance d’une approche dynamique dans l’estimation du taux sans risque, particulièrement pour :

- Les analyses rétrospectives et les backtests
- La calibration des modèles de risque
- L’évaluation d’instruments financiers sur différentes périodes historiques
- La compréhension de l’impact des cycles économiques sur les valorisations

4.5 Backtesting des stratégies

Le backtesting, ou test des stratégies, est une étape cruciale pour évaluer la performance d’une stratégie de trading. Bien que Quantconnect soit largement utilisé comme solution de backtesting, ses limitations dans la version gratuite nous ont poussés à concevoir notre propre framework de backtesting, utilisant les données de yfinance. La Figure 11 présente l’architecture de notre approche :

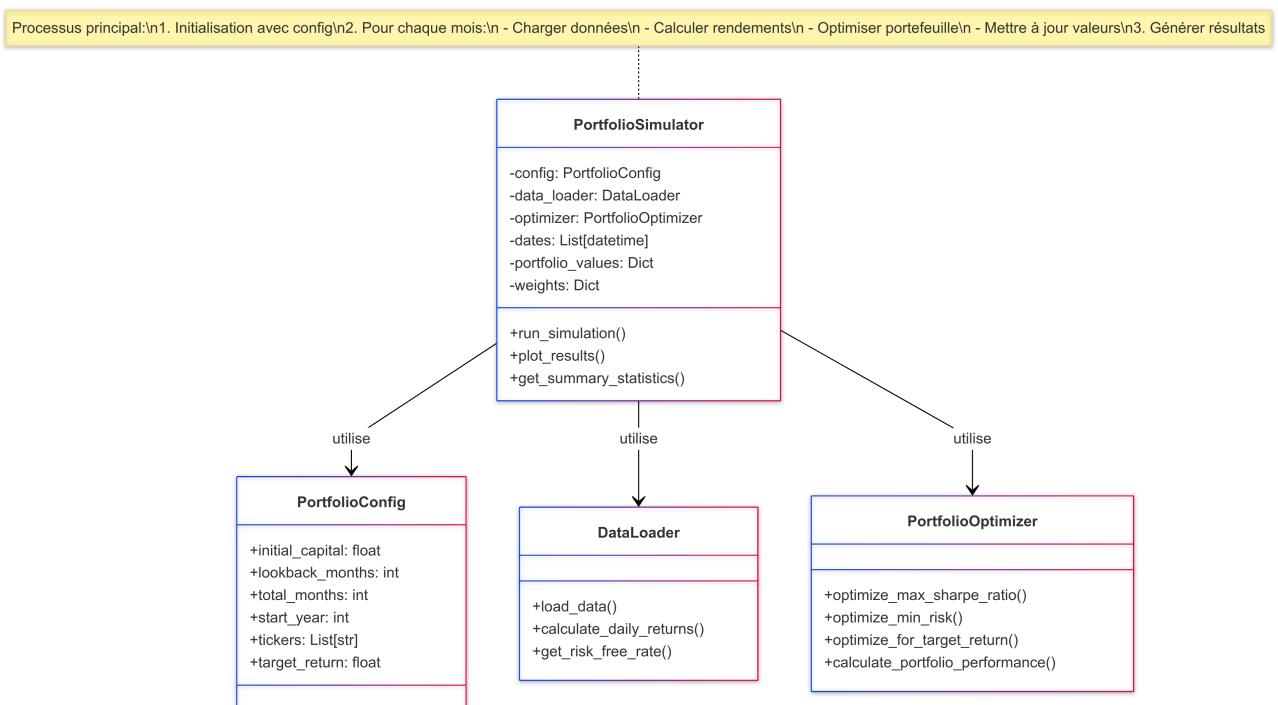


FIGURE 11 – Architecture du framework de backtesting reposant sur les données de yfinance pour l’évaluation des stratégies de trading.

La Figure 12 montre les interactions entre les différentes classes sous forme de diagramme de séquence :

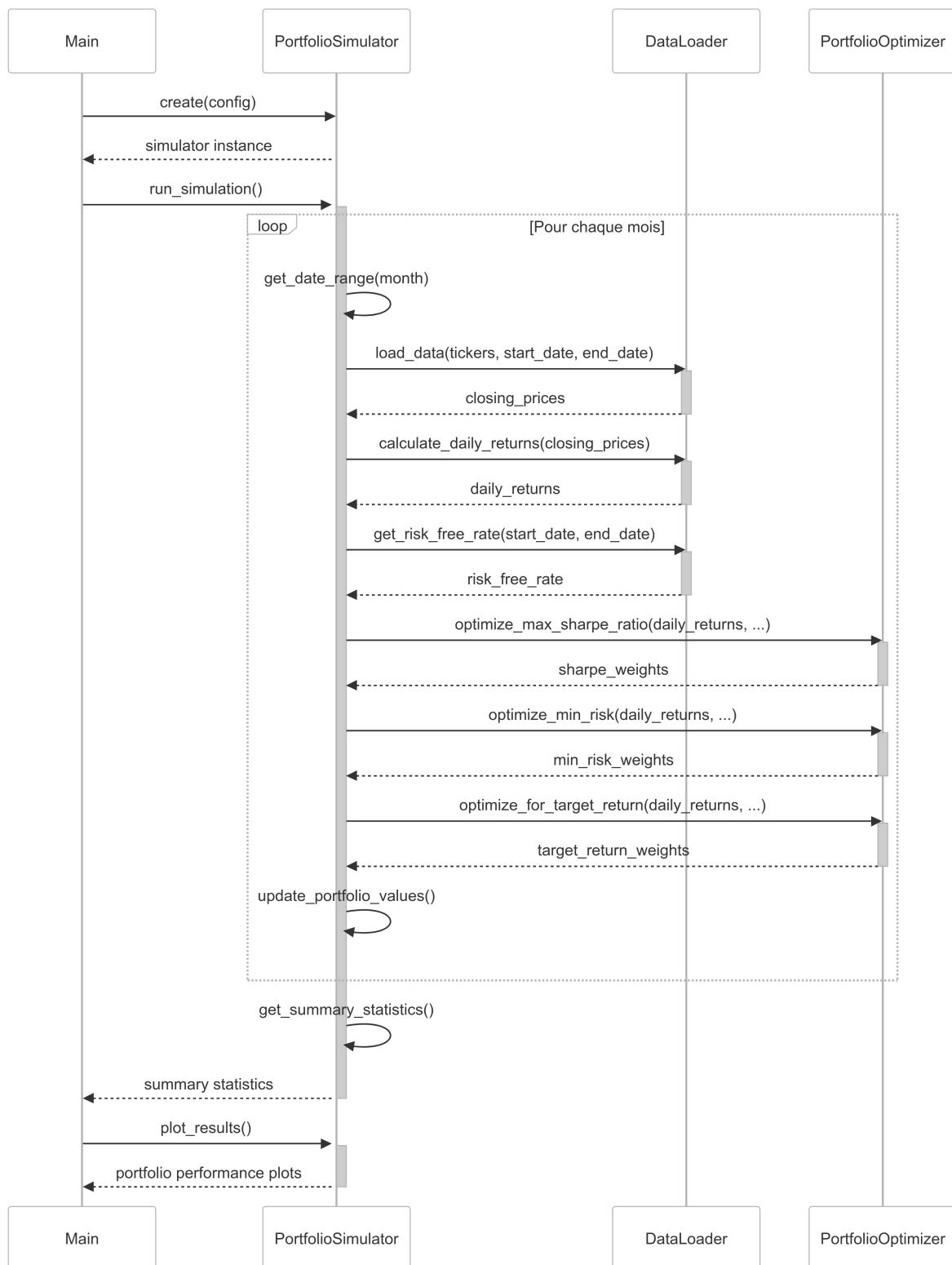


FIGURE 12 – Diagramme de séquence représentant les interactions entre les classes.

Afin de mieux appréhender le processus, voici un diagramme de flux illustrant le déroulement de notre approche :

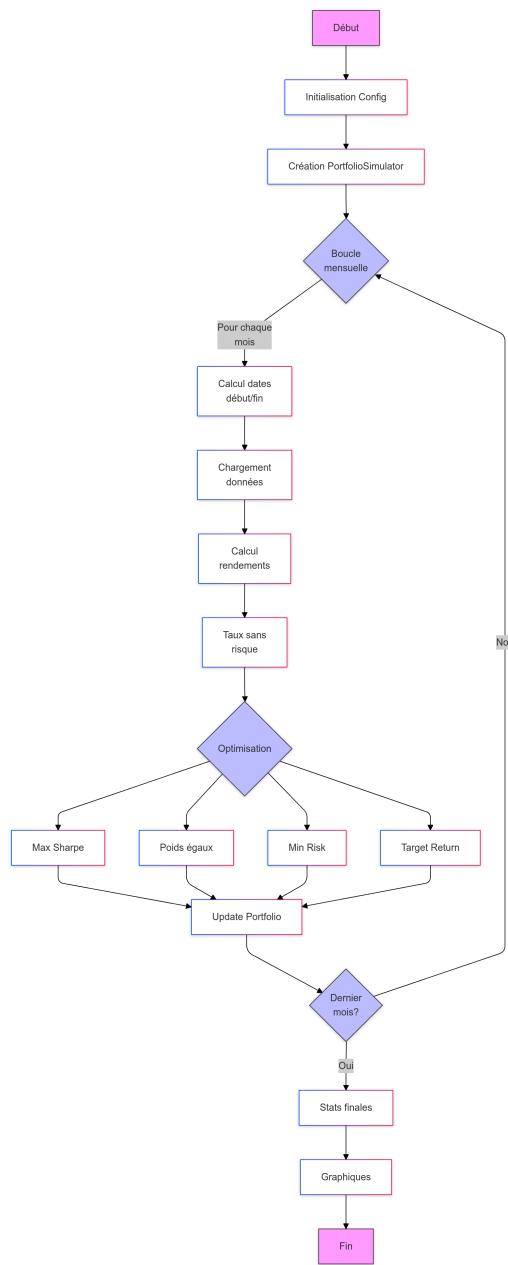


FIGURE 13 – Diagramme de flux représentant le processus du framework de backtesting.

4.5.1 Test en période de crise (2007-2011)

Une fois ceci dit, nous allons essayer de faire quelques tests pour différents portefeuilles. Le premier test portera sur une période de crise pour voir si la stratégie *min risk* peut vraiment

sauver le portefeuille. Nous utiliserons la configuration suivante pour ce test, qui couvre la période avant, pendant et après la crise financière de 2007-2008 :

Paramètre	Valeur
Capital initial	100 000 \$
Période d'observation	12 mois de données historiques
Durée totale	4 ans (de 2007 à 2011)
Année de début	2007 (avant la crise)
Actifs choisis	Ford (F), Citigroup (C), Bank of America (BAC), AIG (AIG), Financial Select Sector SPDR Fund (XLF)
Objectif de rendement	30%

TABLE 2 – Configuration du portefeuille en période de crise

Ce portefeuille inclut des actions d'entreprises telles que Ford (*F*), Citigroup (*C*), Bank of America (*BAC*), AIG (*AIG*) et l'ETF sectoriel XLF, qui ont tous été fortement impactés pendant la crise financière.

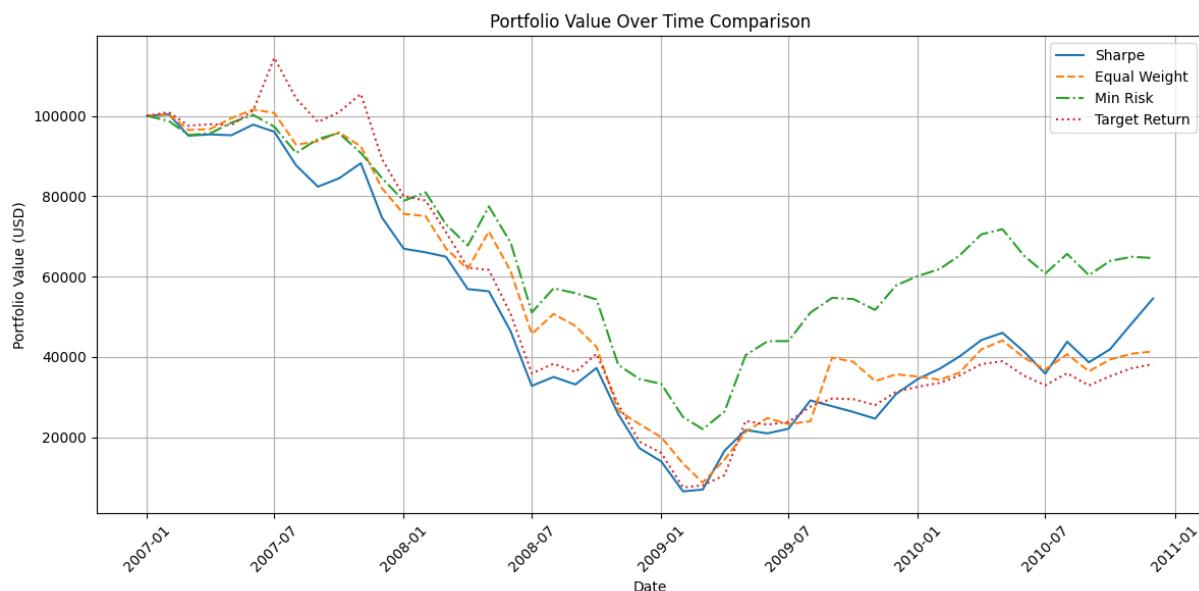


FIGURE 14 – Évolution du portefeuille pendant la crise financière de 2008

Stratégie	Valeur finale	Retour total
Sharpe Strategy	\$54,545.20	-45.45%
Equal Weight Strategy	\$41,377.96	-58.62%
Min Risk Strategy	\$64,621.13	-35.38%
Target Return Strategy	\$38,191.70	-61.81%

TABLE 3 – Résumé des résultats pour les différentes stratégies en période de crise

Stratégie	Sharpe Ratio	Calmar Ratio
Sharpe Strategy	-15.85	-15.04
Equal Weight Strategy	-29.19	-21.66
Min Risk Strategy	-23.71	-13.25
Target Return Strategy	-26.42	-22.89

TABLE 4 – Ratios Sharpe et Calmar des différentes stratégies en période de crise

Stratégie	Volatilité annualisée	Max Drawdown
Sharpe Strategy	88.74%	93.48%
Equal Weight Strategy	67.82%	91.39%
Min Risk Strategy	43.62%	78.02%
Target Return Strategy	80.97%	93.42%

TABLE 5 – Volatilité annualisée et Maximum Drawdown des différentes stratégies en période de crise

Après avoir observé les résultats du backtest, on constate que la stratégie *min risk* a légèrement surpassé les autres portefeuilles, même dans un environnement aussi volatil. Bien que les tickers sélectionnés aient connu des fluctuations importantes pendant la crise financière, la stratégie *min risk* a réussi à équilibrer le risque et à limiter les pertes potentielles grâce à une allocation optimale entre ces actifs.

4.5.2 Test en période de pandémie (2021-2023)

Pour notre second test, nous analysons un portefeuille constitué d'actifs d'entreprises technologiques ayant bien performé durant la pandémie de COVID-19 en 2021. Les entreprises comme Apple, Microsoft, Google, Amazon, et Tesla ont bénéficié d'une demande accrue pendant cette période de turbulences économiques mondiales.

Paramètre	Valeur
Capital initial	100 000 \$
Période d'observation	6 mois de données historiques
Durée totale	3 ans (de 2021 à 2023)
Année de début	2021 (début de la pandémie)
Actifs choisis	Apple (AAPL), Microsoft (MSFT), Google (GOOGL), Amazon (AMZN), NVIDIA (NVDA), Meta (META), Tesla (TSLA), AMD (AMD)
Objectif de rendement	15%

TABLE 6 – Configuration du portefeuille en période de pandémie

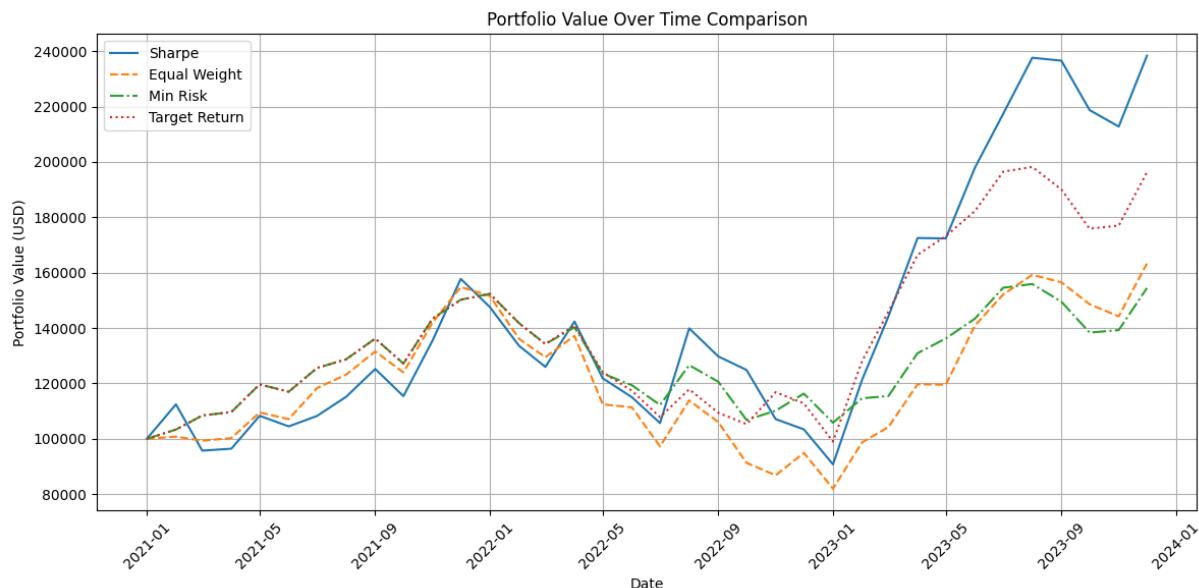


FIGURE 15 – Évolution du portefeuille pendant la période COVID-19

Stratégie	Valeur finale	Retour total
Sharpe Strategy	\$238,381.02	138.38%
Equal Weight Strategy	\$163,296.75	63.30%
Min Risk Strategy	\$154,455.30	54.46%
Target Return Strategy	\$196,293.15	96.29%

TABLE 7 – Résumé des résultats pour les différentes stratégies en période de pandémie

Stratégie	Sharpe Ratio	Calmar Ratio
Sharpe Strategy	77.06	79.08
Equal Weight Strategy	53.33	37.75
Min Risk Strategy	65.63	51.04
Target Return Strategy	84.59	72.03

TABLE 8 – Ratios Sharpe et Calmar des différentes stratégies en période de pandémie

Stratégie	Volatilité annualisée	Max Drawdown
Sharpe Strategy	43.58%	42.47%
Equal Weight Strategy	33.30%	47.05%
Min Risk Strategy	23.76%	30.55%
Target Return Strategy	29.80%	35.00%

TABLE 9 – Volatilité annualisée et Maximum Drawdown des différentes stratégies en période de pandémie

La stratégie Sharpe a donné les meilleurs résultats pendant la période de pandémie, avec un retour total de 138.38%. Cela peut être attribué à la forte concentration dans des actions telles que Tesla (TSLA) et NVIDIA (NVDA), qui ont vu une croissance exceptionnelle. Cependant, cette performance s'accompagne d'une volatilité élevée de 43.58%, engendrant des fluctuations significatives.

La stratégie visant un rendement cible de 15% a généré un retour solide de 96.29%, avec un ratio Sharpe de 84.59%, indiquant une performance relativement stable par rapport aux risques encourus. En dépit d'une volatilité modérée (29.80%), cette stratégie a bien équilibré le risque et le rendement.

La stratégie Min Risk a montré un retour total de 54.46%, ce qui est impressionnant étant donné la faible volatilité (23.76%) et le maximum drawdown limité à 30.55%. Cette performance démontre sa capacité à limiter les pertes tout en générant des rendements positifs, même en période de turbulence.

En comparant les deux périodes, il est intéressant de noter que la stratégie Min Risk s'est montrée particulièrement efficace en période de crise financière, limitant les pertes à -35.38% alors que les autres stratégies subissaient des pertes plus importantes. En période de pandémie, bien que générant des rendements plus modestes, elle a maintenu son profil défensif avec la plus faible volatilité et le plus faible drawdown maximum.

4.6 Analyse des sentiments

4.6.1 Introduction

L'analyse des sentiments des nouvelles financières est réalisée à l'aide du modèle FinBERT, un modèle de traitement du langage naturel spécialement conçu pour l'analyse de textes financiers. Cette approche combine des techniques avancées de NLP avec une méthodologie robuste pour extraire et analyser les sentiments des textes financiers.

4.6.2 Architecture du pipeline

Le pipeline de l'analyse des sentiments est composé de quatre étapes principales, comme illustré dans la Figure 16 :

1. **Configuration de l'environnement** : Initialisation des bibliothèques nécessaires et chargement du modèle pré-entraîné FinBERT.
2. **Collecte et prétraitement des données** : Extraction des nouvelles financières via une API, nettoyage des textes, et tokenisation.
3. **Analyse de sentiment** : Classification des textes en trois catégories (positif, négatif, neutre) avec attribution de scores de confiance.
4. **Visualisation et analyse** : Génération de graphiques et interprétation des résultats.

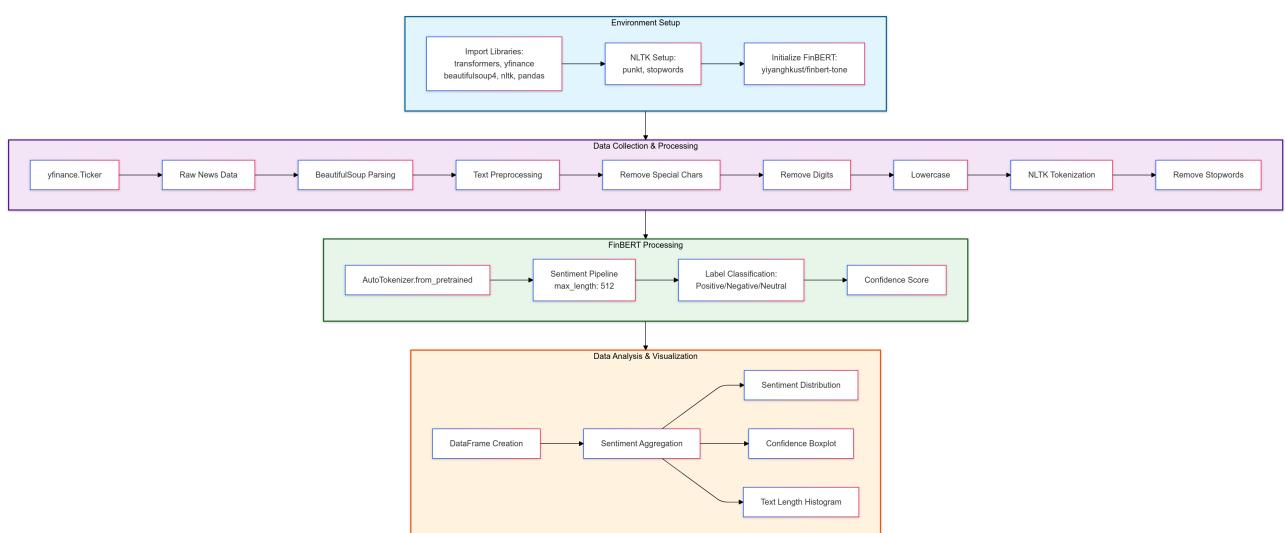


FIGURE 16 – Pipeline d'analyse de sentiment utilisant FinBERT.

4.6.3 Méthodologie détaillée

Notre approche se décompose en quatre phases principales :

1. Configuration de l'environnement

- Initialisation des bibliothèques essentielles (transformers, NLTK)
- Chargement du modèle FinBERT pré-entraîné

2. Collecte et prétraitement des données

- Extraction des nouvelles financières via l'API yfinance
- Nettoyage du texte : suppression des caractères spéciaux
- Tokenization et élimination des mots vides

3. Analyse de sentiment

- Classification des textes en trois catégories : positif, négatif, neutre
- Attribution d'un score de confiance pour chaque classification

4. Visualisation et analyse

- Génération de graphiques représentatifs
- Interprétation des tendances et patterns

4.6.4 Validation du modèle

Pour valider la performance du modèle, nous avons effectué plusieurs tests :

Test 1 : Phrase simple

"Company reports strong earnings growth"

```
Test sentiment: [{'label': 'Positive', 'score': 1.0}]
```

Test 2 : Phrase complexe avec nuances

"Despite Tesla's announcement of record-breaking profits this quarter, analysts remain skeptical about its long-term growth."

```
Test sentiment: [{'label': 'Negative', 'score': 0.9999996423721313}]
```

4.6.5 Étude de cas : Analyse de Tesla (TSLA)

Nous avons appliqué notre modèle aux actualités récentes de Tesla :

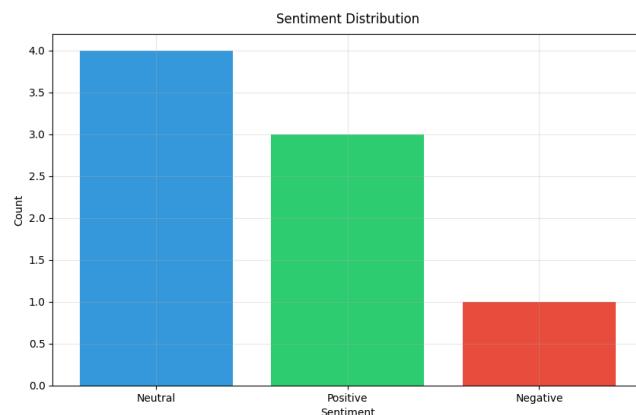


FIGURE 17 – Distribution des sentiments dans les articles récents sur TSLA

L’analyse de la Figure 17 montre que les sentiments positifs représentent une proportion significative, suggérant un optimisme général envers Tesla. Toutefois, une fraction notable de sentiments négatifs met en évidence certaines inquiétudes, probablement liées à des défis récents tels que des problèmes de production ou des fluctuations de marché. Les sentiments neutres indiquent que certains articles adoptent une approche équilibrée, soulignant les forces et les faiblesses de Tesla de manière impartiale.

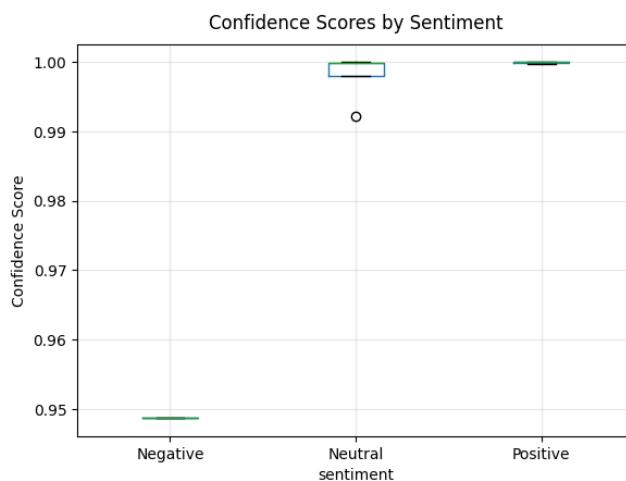


FIGURE 18 – Scores de confiance par catégorie de sentiment

La Figure 18 révèle que les scores de confiance pour les sentiments positifs et négatifs sont généralement élevés, ce qui renforce la fiabilité des prédictions de FinBERT. En revanche, les scores légèrement inférieurs pour les sentiments neutres pourraient refléter des ambiguïtés dans certains textes financiers, nécessitant une validation humaine ou un affinage des modèles.

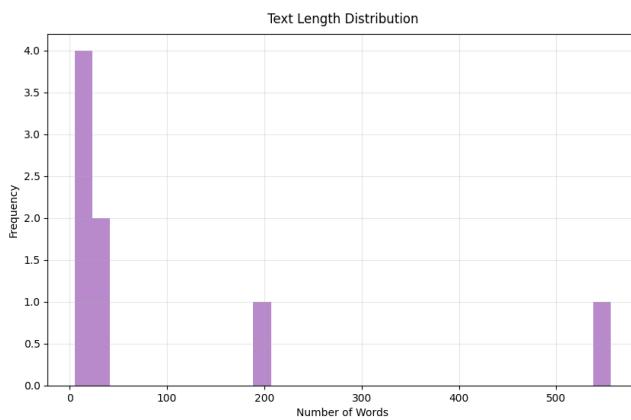


FIGURE 19 – Distribution des types de textes analysés

L’analyse des types de textes représentée dans la Figure 19 révèle des insights intéressants sur les sources d’information. Une majorité des textes proviennent d’articles de presse, ce qui montre une forte dépendance aux nouvelles officielles pour évaluer les sentiments. Les tweets et commentaires sur les réseaux sociaux représentent une proportion significative, suggérant que les opinions des utilisateurs individuels jouent également un rôle dans la perception globale. Cette répartition met en évidence la complémentarité entre les sources officielles et les données générées par les utilisateurs pour fournir une analyse plus complète.

Comparativement, les articles de presse offrent une analyse structurée et plus détaillée, généralement validée par des experts, ce qui confère un haut niveau de crédibilité. En revanche, les tweets et commentaires sur les réseaux sociaux apportent une vision immédiate et souvent émotionnelle des opinions du public. Ces deux perspectives sont essentielles pour obtenir une vue équilibrée.

Notons également que les commentaires des forums financiers et les posts sur les réseaux sociaux peuvent servir de précurseurs aux changements d’opinion, anticipant parfois les évolutions du sentiment général bien avant qu’elles ne soient reflétées dans les articles formels. Par exemple, une forte augmentation de mentions négatives sur les réseaux sociaux pourrait signaler un événement ou un problème latent que les analystes officiels n’ont pas encore traité. Ce type de complémentarité met en lumière l’importance de diversifier les sources pour une meilleure robustesse des analyses.

La prédominance des articles de presse peut indiquer une fiabilité et une profondeur accrues dans les informations, tandis que les réseaux sociaux apportent une perspective plus immédiate et émotionnelle. Cette combinaison enrichit la base de données d’analyse et peut influencer les décisions des investisseurs de manière différente selon la source.

4.6.6 Gestion des limitations techniques

Pour faire face aux défis liés à l’extraction de données via BeautifulSoup, nous avons mis en place une stratégie d’adaptation :

- **Analyse des titres uniquement** : En cas d’indisponibilité des articles complets

- **Analyse combinée** : Fusion du titre et de l'article lorsque disponibles
 - **Extraction ciblée** : Utilisation des balises <p> pour extraire le contenu pertinent
- Cette approche adaptative permet d'assurer une analyse robuste même en présence de données partielles ou incomplètes.

5 Résultats du Projet

5.1 Motivation

L'analyse et les résultats obtenus au cours de ce projet sont impressionnantes et reflètent la richesse des données traitées. Cependant, avec la diversité des théories, des modèles et des résultats générés, il devient crucial de structurer ces informations de manière accessible. L'objectif principal était de rendre toutes ces analyses compréhensibles et faciles à utiliser pour l'utilisateur final.

Pour répondre à ce besoin, nous avons développé une application web qui centralise toutes les fonctionnalités du projet dans une interface claire et interactive. Cette application permet de visualiser et d'explorer facilement les résultats des simulations de portefeuille, les différentes stratégies de gestion des risques, ainsi que les performances passées, tout en restant intuitive pour l'utilisateur.

Au lieu de naviguer à travers des notebooks ou des blocs de code complexes, l'utilisateur peut maintenant interagir directement avec une interface qui présente les informations essentielles de manière structurée. Cette approche permet à chacun de tirer pleinement parti des analyses sans être noyé dans les détails techniques, tout en offrant des options de personnalisation pour ajuster les paramètres du portefeuille selon les besoins spécifiques.

Ainsi, ce produit vise à rendre l'analyse financière accessible, tout en permettant une prise de décision plus rapide et plus éclairée grâce à une présentation simplifiée des résultats.

5.2 Tech Stack de la Solution

La solution développée repose sur une combinaison de technologies permettant de créer une application web interactive et robuste. Voici un aperçu de la stack technologique utilisée dans ce projet :

5.2.1 Backend

Le backend de l'application utilise **Flask**, un framework léger pour Python, pour gérer les routes et servir les requêtes de l'utilisateur. Flask est utilisé pour exposer les fonctionnalités via des API RESTful et pour gérer le rendu des pages HTML.

- **Flask** : Framework Python pour le backend, qui gère les routes HTTP et le rendu des templates.
- **Jinja2** : Moteur de templates utilisé par Flask pour générer dynamiquement les pages HTML.

et voici la structure de l'API créée en backend :

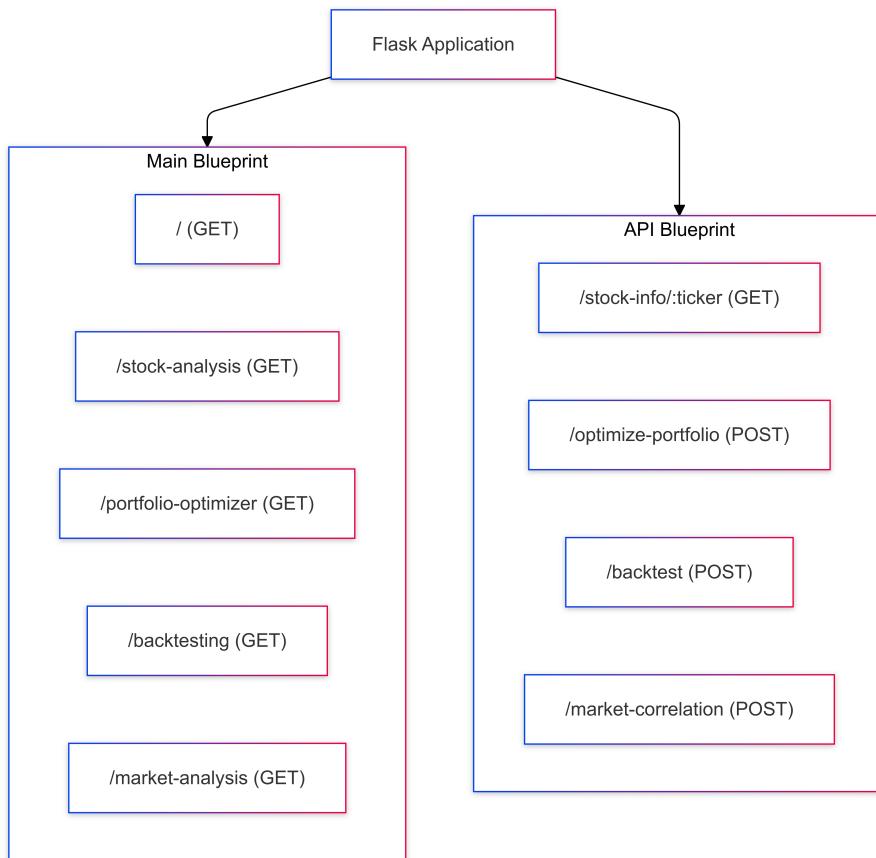


FIGURE 20 – Structure de l'API Backend

5.2.2 Frontend

Le frontend de l'application est construit avec les technologies suivantes :

- **HTML5** : Utilisé pour la structure de la page web.
- **CSS (Tailwind CSS)** : Framework CSS pour la mise en page réactive et la personnalisation rapide de l'apparence de l'interface utilisateur.
- **JavaScript** : Pour l'interactivité et la manipulation dynamique des éléments de la page.

5.3 Présentation de la Solution et guide d'utilisation

L'application web développée offre une suite complète d'outils d'analyse financière et d'optimisation de portefeuille. Cette section détaille le processus de démarrage et présente les différentes fonctionnalités disponibles.

5.3.1 Démarrage de l'application

Le lancement de l'application s'effectue via le fichier `app.py`. Lors du premier démarrage, un temps d'initialisation est nécessaire pour le chargement des différents modèles d'analyse. Si

certains modèles ne sont pas présents localement, ils seront automatiquement téléchargés.

5.3.2 Architecture et fonctionnalités

L'interface utilisateur est organisée en plusieurs pages distinctes, chacune dédiée à une fonction spécifique :

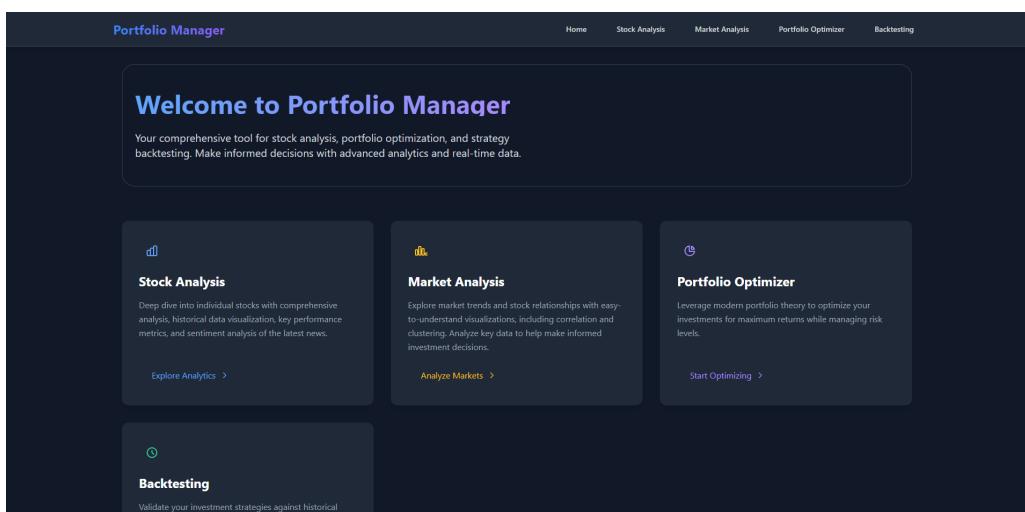


FIGURE 21 – Interface d'accueil de l'application

Page d'accueil La page d'accueil présente une interface moderne et épurée, dotée d'une barre de navigation intuitive permettant d'accéder aux différentes fonctionnalités de l'application.

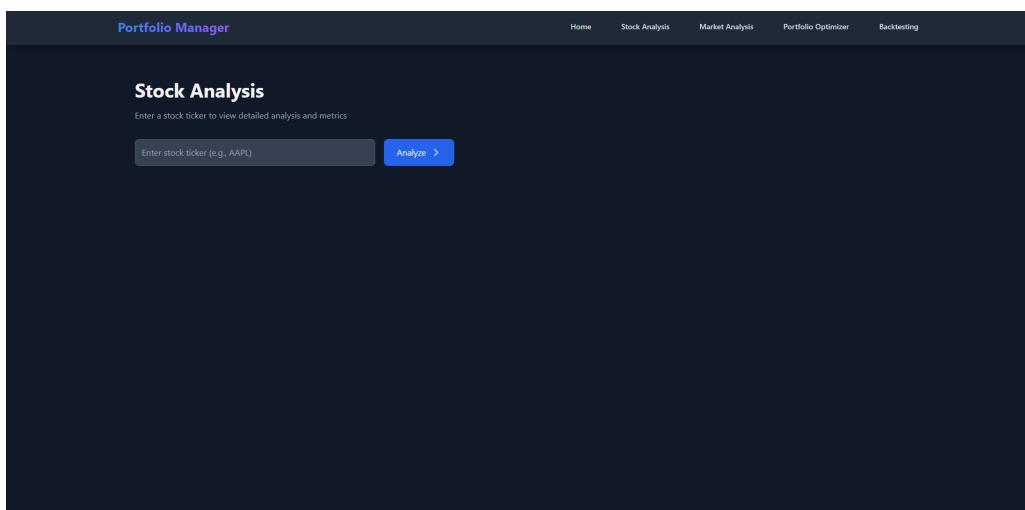


FIGURE 22 – Interface d'analyse individuelle des actions

Analyse individuelle des actions Cette section permet aux utilisateurs d'effectuer une analyse détaillée d'une action spécifique. Par exemple, en saisissant le symbole "INTC" (Intel Corporation), l'analyse comprend : Informations de base (Nom, Secteur, Capitalisation boursière), Indicateurs clés (Ratio P/E, Rendement du dividende, Béta, Dette/Équité, ROE, ROA), Graphique du cours de l'action et Dernières actualités :

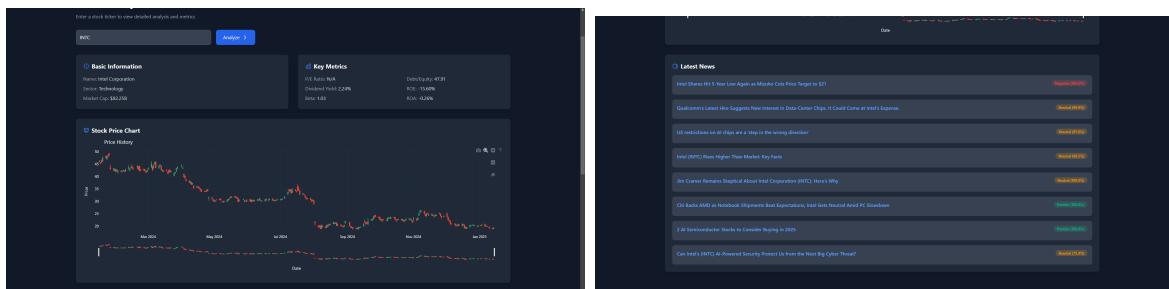


FIGURE 23 – Exemple d'analyse détaillée pour Intel (INTC)

Analyse de marché Cette fonctionnalité permet d'analyser les tendances et les relations entre différents titres sur une période définie. Les utilisateurs peuvent spécifier plusieurs symboles boursiers (par exemple, AAPL, MSFT, GOOGL) et une plage de dates. L'analyse se décompose en deux parties principales :

1. Analyse des corrélations

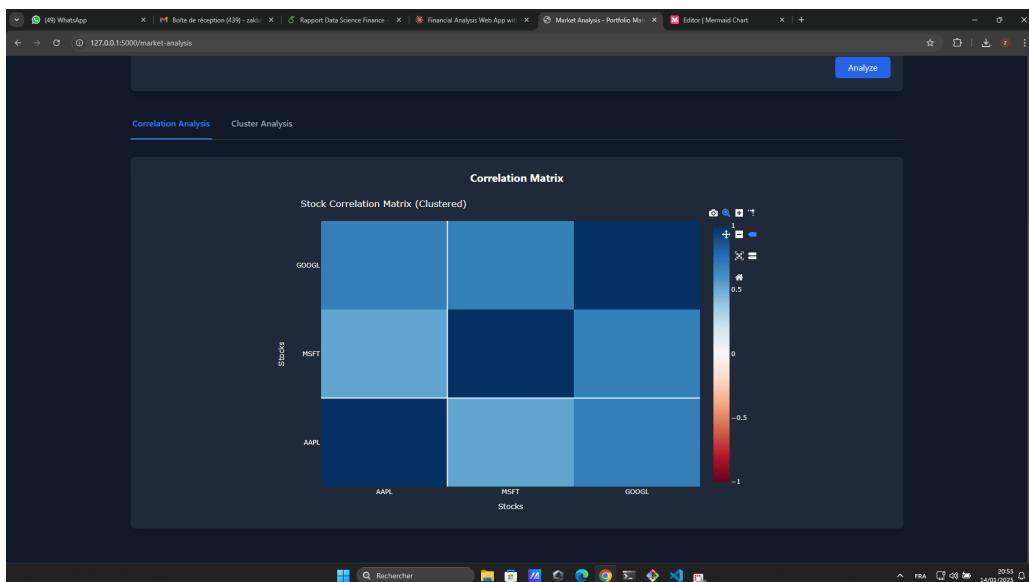


FIGURE 24 – Matrice de corrélation entre les actifs sélectionnés

2. Analyse par clusters

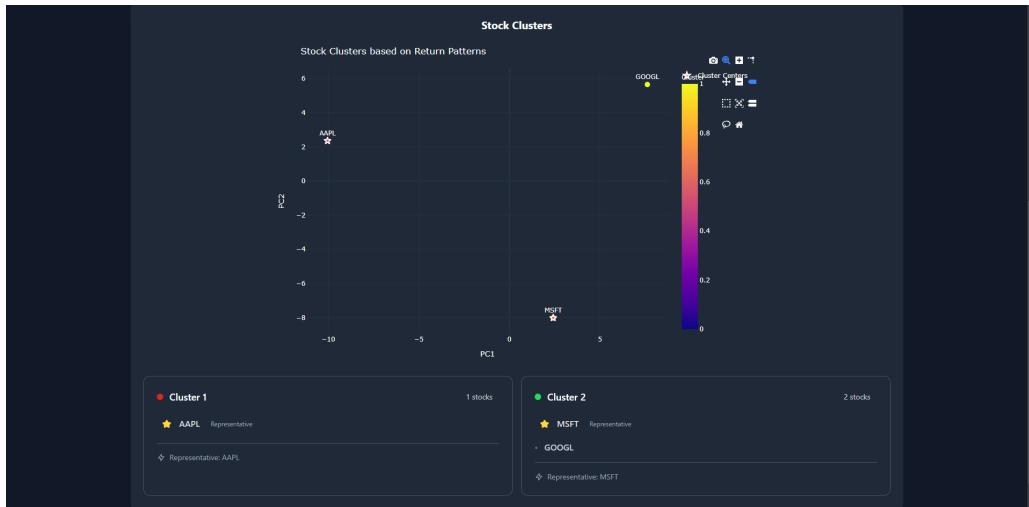


FIGURE 25 – Regroupement des actifs par clusters

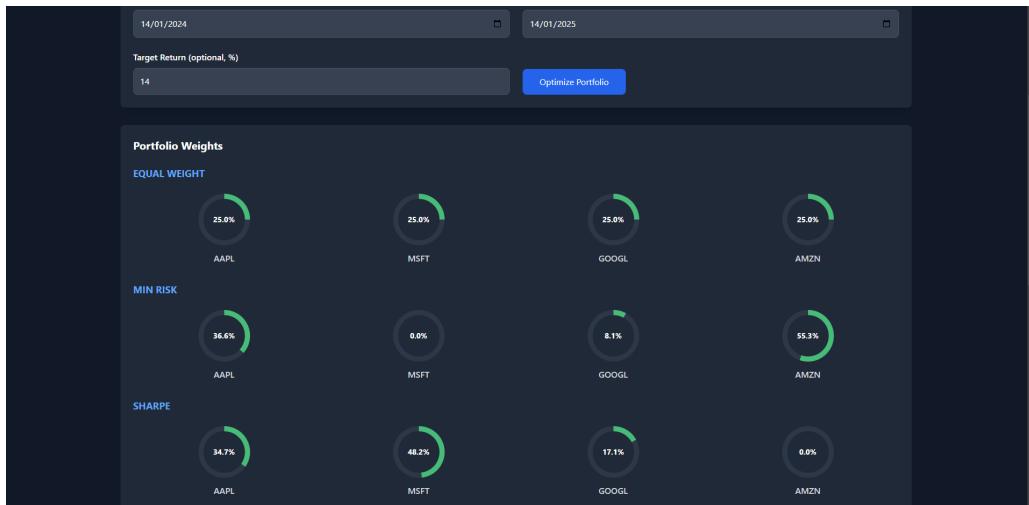


FIGURE 26 – Interface d'optimisation de portefeuille

Optimisation de portefeuille Cette section permet aux utilisateurs d'optimiser la composition de leur portefeuille selon différents critères de risque et de rendement, et fournit également les poids à considérer ainsi qu'un résumé de performance comprenant les métriques de rendement, de risque et de Sharpe.

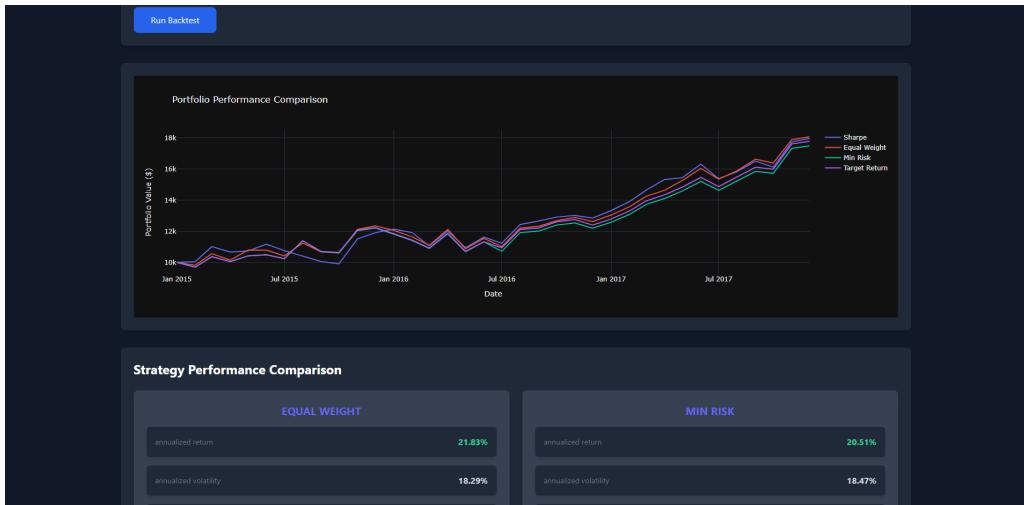


FIGURE 27 – Interface de backtesting

Backtesting Le module de backtesting permet de tester et valider les stratégies d’investissement sur des données historiques, permettant ainsi d’évaluer leur performance potentielle. Ceci comprend des métriques telles que le rendement annualisé, la volatilité annualisée, le ratio de Calmar, la valeur finale, la perte maximale (max drawdown), le ratio de Sharpe et le rendement total.

5.3.3 Critiques et Perspectives d’Amélioration

Bien que l’application réponde aux besoins fondamentaux d’analyse financière, plusieurs axes d’amélioration peuvent être identifiés :

Gestion des portefeuilles L’application pourrait bénéficier d’un système complet de gestion de portefeuilles permettant aux utilisateurs de :

- Créer et sauvegarder plusieurs portefeuilles personnalisés
- Suivre l’évolution des portefeuilles en temps réel
- Mettre en place des stratégies d’allocation dynamique automatisée
- Recevoir des alertes et notifications sur les mouvements significatifs

Architecture technique Des améliorations techniques pourraient enrichir l’expérience utilisateur :

- Migration vers un framework front-end moderne comme React.js pour une interface plus réactive et interactive
- Diversification des sources de données financières au-delà de Yahoo Finance pour une plus grande fiabilité et richesse d’information
- Implémentation d’un système d’authentification robuste permettant la personnalisation des services

Ces évolutions permettraient de transformer l'outil actuel en une plateforme plus complète et personnalisée, adaptée à un usage professionnel.

5.4 Résultats du modèle de Forêt Aléatoire

5.4.1 Crédation des variables explicatives et de la cible

Pour cette analyse, nous avons utilisé les données boursières d'Apple (AAPL) et avons calculé la cible (*Target*) ainsi que plusieurs variables explicatives (*features*) comme suit :

- **Cible** : La variable cible quotidienne (*Target_Daily*) est définie par :

$$\text{Target_Daily} = \mathbb{1}(\text{Close}_{t+1} > \text{Close}_t)$$

où **Close** représente le prix de clôture. Si le prix de clôture du jour suivant est supérieur à celui du jour courant, la cible est égale à 1 ; sinon, elle est égale à 0.

- **Variables explicatives** : Nous avons calculé les variables suivantes :

- **Volume** : Le volume total des transactions, fourni directement par les données brutes.
- **RSI (Relative Strength Index)** : Calculé comme suit :

$$\text{RSI} = 100 - \frac{100}{1 + \frac{\text{Moyenne des gains}}{\text{Moyenne des pertes}}}$$

où les gains et pertes sont calculés sur une période donnée (souvent 14 jours).

- **MACD, MACD_Signal, MACD_Hist** : Moyenne mobile convergence divergence et ses composantes, définies par :

$$\text{MACD} = \text{EMA}_{\text{rapide}} - \text{EMA}_{\text{lente}}$$

où **EMA** est la moyenne mobile exponentielle, et **MACD_Signal** est la moyenne mobile exponentielle de **MACD**. L'histogramme **MACD_Hist** est donné par :

$$\text{MACD_Hist} = \text{MACD} - \text{MACD_Signal}$$

- **BB_Lower** : Bande inférieure des bandes de Bollinger, définie comme :

$$\text{BB_Lower} = \text{SMA} - k \cdot \text{std}$$

où **SMA** est la moyenne mobile simple, **std** est l'écart type sur la période spécifiée, et **k** est un coefficient (souvent **k** = 2).

- **Volume_Change** : Variation relative du volume, calculée comme :

$$\text{Volume_Change} = \frac{\text{Volume}_t - \text{Volume}_{t-1}}{\text{Volume}_{t-1}}$$

- **RSI_Lag_1** et **RSI_Lag_2** : Valeurs décalées de l'indicateur RSI d'un et deux jours respectivement :

$$\text{RSI_Lag_1} = \text{RSI}_{t-1}, \quad \text{RSI_Lag_2} = \text{RSI}_{t-2}$$

- **RSI_Volume_Interaction** : Interaction entre le RSI et le volume, donnée par :

$$\text{RSI_Volume_Interaction} = \text{RSI} \times \text{Volume}$$

Pour éviter les redondances dans les variables explicatives, nous avons calculé une matrice de corrélation et éliminé les variables hautement corrélées. Après ce processus, nous avons retenu les variables suivantes :

- Volume,
- RSI,
- MACD,
- MACD_Signal,
- MACD_Hist,
- BB_Lower,
- Volume_Change,
- RSI_Lag_1,
- RSI_Lag_2,
- RSI_Volume_Interaction

5.4.2 Optimisation et sélection du modèle

Une recherche par grille (*GridSearchCV*) a été appliquée pour optimiser les hyperparamètres du modèle de Forêt Aléatoire. L'évaluation de chaque combinaison d'hyperparamètres a été effectuée en utilisant la F1-score comme métrique principale, compte tenu de la nature de la tâche de classification et du potentiel déséquilibre des classes.

5.4.3 Calcul du profit et pertes (P&L)

Après avoir sélectionné le meilleur modèle sur la base de la F1-score, nous avons appliqué le modèle sur des données de test pour prédire les mouvements journaliers du prix de l'action (*Target_Daily*). Le calcul des profits et pertes (*P&L*) est défini comme suit :

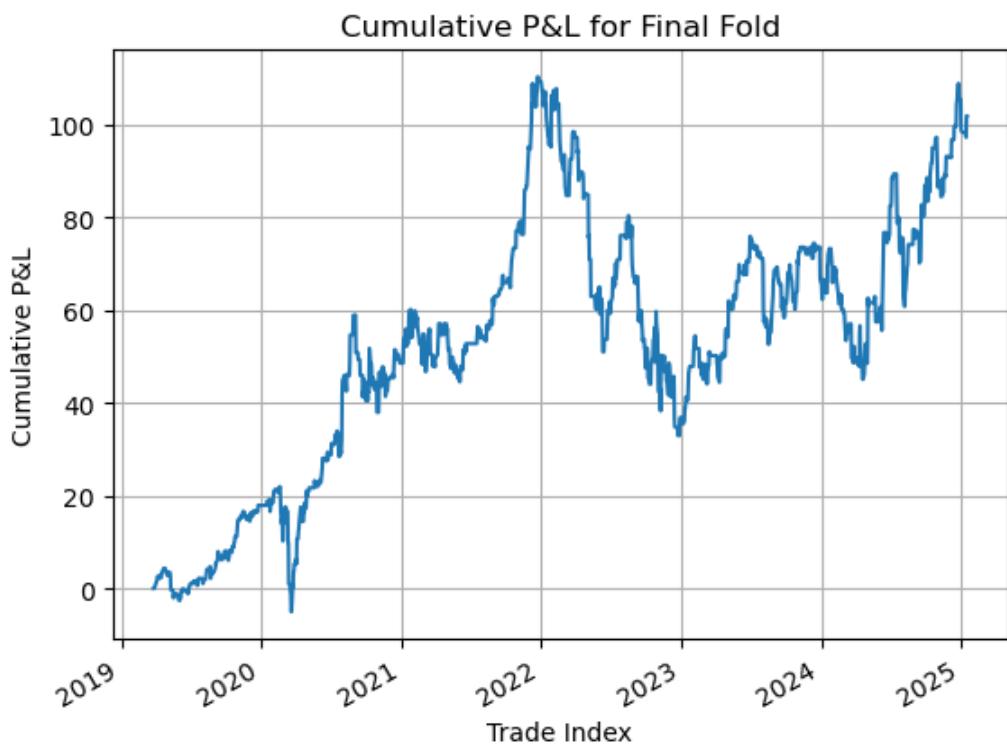
$$P\&L = \sum_{i=1}^N \text{Prediction}_i \cdot (\text{Close}_{i+1} - \text{Close}_i) \quad (36)$$

où Prediction_i est la prédiction binaire du modèle pour le jour i , Close_{i+1} et Close_i sont les prix de clôture respectivement du jour $i + 1$ et du jour i .

Ce calcul permet d'évaluer les performances du modèle en termes financiers, en supposant que chaque prédiction correcte génère un profit équivalent à la variation du prix de l'action.

5.4.4 Résumé des résultats

En résumé, le modèle de Forêt Aléatoire, optimisé via la recherche par grille et évalué avec la F1-score, a permis de prédire efficacement les mouvements du prix de l'action d'Apple. Les profits et pertes calculés montrent la pertinence des variables explicatives sélectionnées et du processus d'optimisation.



6 Répertoire GitHub

Vous pouvez consulter le répertoire GitHub du projet en suivant ce lien : <https://github.com/zikous/pole-projet-data-s7/>.

7 Conclusion

Ce projet a mis en évidence l'efficacité des approches interdisciplinaires pour relever les défis de la gestion de portefeuilles financiers dans un contexte de complexité croissante et de volatilité accrue des marchés. En combinant des outils avancés comme les modèles graphiques parcimonieux, l'analyse des sentiments financiers via FinBERT et des techniques d'apprentissage automatique telles que les Random Forests, nous avons démontré la pertinence de méthodes modernes pour optimiser le compromis entre risque et rendement. Les résultats obtenus, notamment à travers les simulations Monte Carlo, la construction de la frontière efficiente et l'analyse des corrélations, montrent une capacité accrue à modéliser les interactions complexes entre actifs et à améliorer la prise de décision. Par ailleurs, l'application web développée traduit concrètement ces avancées en offrant une interface intuitive et interactive qui démocratise l'accès à ces outils complexes. Cependant, ce projet ouvre également la voie à de nombreuses perspectives prometteuses, telles que l'intégration de modèles dynamiques pour capturer les évolutions temporelles des corrélations, l'exploration des impacts des critères ESG dans les stratégies d'investissement ou encore l'amélioration technologique de l'application avec des fonctionnalités de gestion en temps réel et des visualisations interactives avancées. En définitive, ce travail illustre comment les synergies entre finance, data science et technologie peuvent transformer la gestion d'actifs en une discipline plus robuste, agile et inclusive, répondant mieux aux défis contemporains des marchés financiers.

Références

- [1] L. Boyd, S. Vandenberghe. Convex optimization. 2004.
- [2] L. Breiman. Random forests. 2001.
- [3] Chang M.-W. Lee K. Toutanova K. Devlin, J. Bert : Pre-training of deep bidirectional transformers for language understanding. 2019.
- [4] Dueck D. Frey, B. J. Clustering by passing messages between data points. 2007.
- [5] B. Liu. Sentiment analysis and opinion mining. 2012.
- [6] H Markowitz. Portfolio selection : Efficient diversification of investments. 1959.
- [7] W. F. Sharpe. Capital asset prices : A theory of market equilibrium under conditions of risk. 1964.