Rapport de projet :

Réseaux de Neurones

Membres du groupe :

Cheddadi Radja
Abichou Nour Elhouda

M2 IMSD 2022/2023

Table de matière :

Introduction

L'analyse des séries financières est devenue de plus en plus importante au fil du temps, car les fluctuations des marchés financiers ont un impact significatif sur l'économie mondiale et sur les investisseurs individuels. Les avancées en matière de techniques de machine learning et de deep learning sont un outil qui permet d'exploiter de grandes quantités de données financières historiques, afin d'améliorer considérablement les prévisions des séries financières.

Dans le cadre de notre étude, nous allons nous concentrer sur le SP500, qui est l'un des indices boursiers les plus largement suivis et les plus importants du marché. Nous allons utiliser des techniques de machine learning et de deep learning avancées pour extraire des informations à partir de données historiques du SP500, ce qui nous permettra de fournir des prévisions fiables pour le prix du SP500.

Nous allons commencer par une analyse exploratoire des données historiques du SP500 sur une période donnée, qui nous permettra de comprendre les caractéristiques et les tendances du marché. Nous allons ensuite tester plusieurs modèles de machine learning pour prédire les prix futurs du SP500. Nous allons ensuite chercher à identifier le modèle qui fournit les résultats les plus précis en utilisant le Root Mean Squared Error qui est une mesure couramment utilisée pour évaluer la précision des prévisions de séries temporelles.

Notre objectif ultime est d'aider les investisseurs à prendre des décisions plus éclairées en matière d'investissement en leur fournissant des prévisions pour le prix du SP500. En fournissant des outils de prévisions plus précis, ce qui peut avoir un impact significatif sur leur portefeuille et sur leur avenir financier.

1) Analyse exploratoire des données

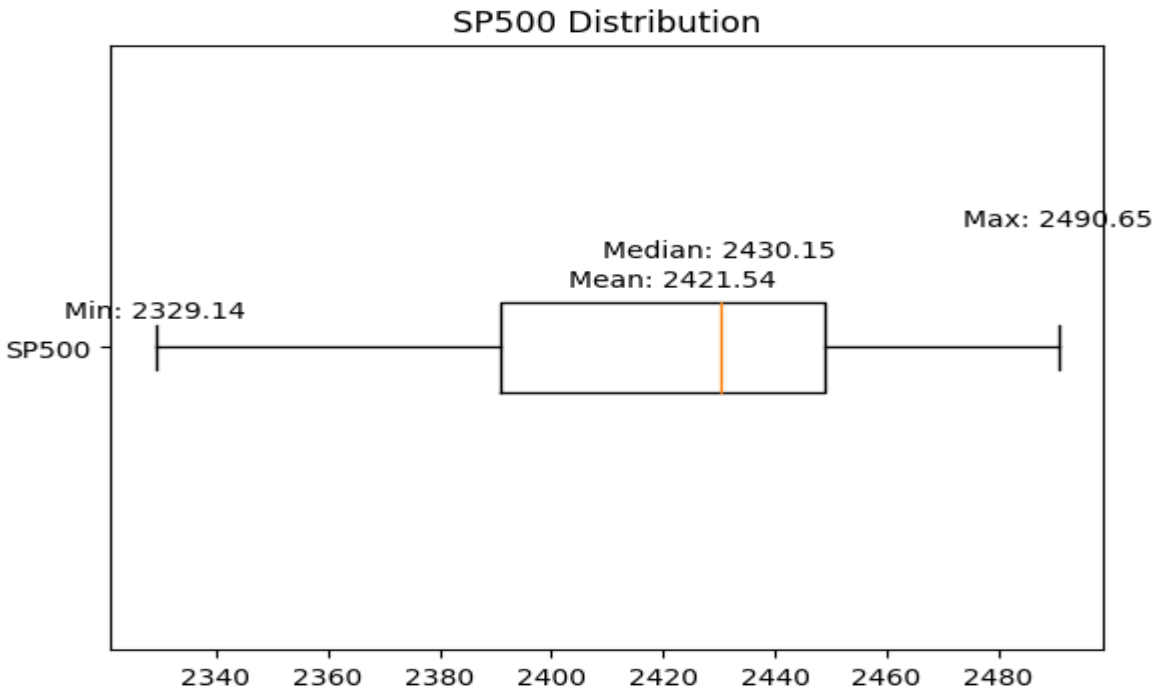
Le SP500 est un indice boursier américain pondéré en fonction de la capitalisation qui suit la performance de 500 grandes sociétés cotées en bourse. Il couvre environ 80% du marché boursier américain par capitalisation boursière. L'analyse statistique de données du SP500 peut fournir des informations précieuses pour les investisseurs, les analystes et les décideurs.

Les données fournies contiennent les valeurs de l'indice SP500 chaque minute entre le 2017-04-03 et 2017-08-31. La série chronologique contient un total de 41266 observations.

La moyenne des prix de SP500 pour cette période est de 2421,54 USD, avec un écart-type de 39,56. Le prix minimum enregistré était de 2329,14 USD et le prix maximum enregistré était de 2490,65 USD. Les 25% les plus bas des prix étaient inférieurs à 2390,86USD, tandis que les 25% les plus élevés étaient supérieurs à 2448,82 USD.

Ces statistiques révèlent que les prix de SP500 ont une distribution relativement étroite, avec un écart-type relativement faible de 39,56. Cependant, il y a une certaine variabilité, avec une différence de près de 200 points entre le prix minimum et le prix maximum enregistré.

Figure 01: Distribution de l'indice SP500



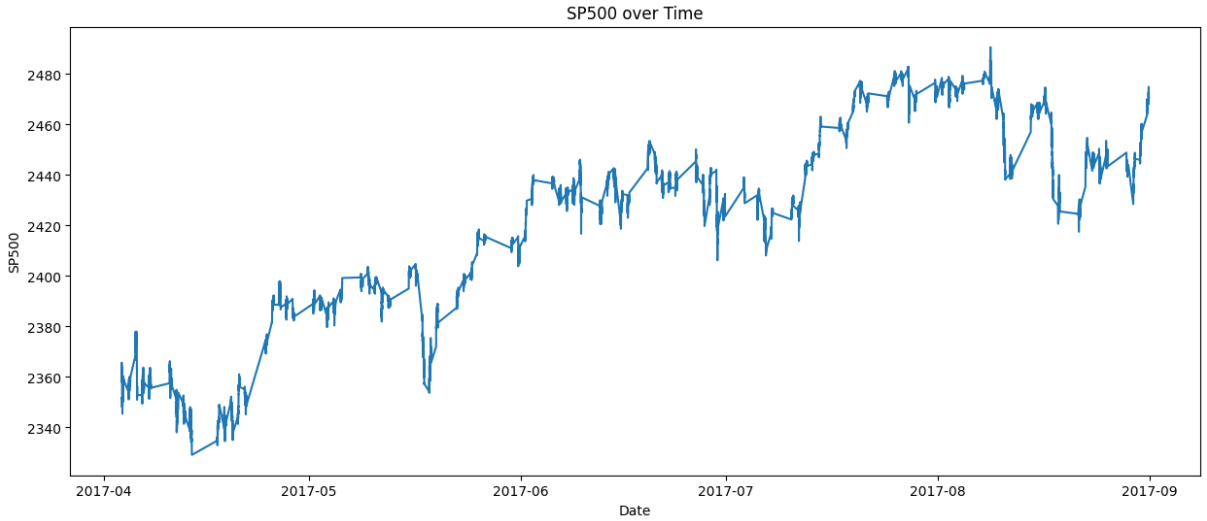
Du 3 avril 2017 au 31 août 2017, l'indice a connu des fluctuations importantes, passant d'un prix de 2358,96 USD le 3 avril 2017 à un sommet de 2446,30 USD le 8 août 2017, avant de redescendre à 2440,69 USD le 31 août 2017. Malgré ces fluctuations, l'indice a globalement augmenté de 3,2 % sur la période.

En analysant les données de manière plus détaillée, l'on observe que l'indice a connu une légère baisse initiale de 0,20% le 3 avril 2017, passant de 2363,61 à 2358,96 USD. Cependant, il s'est ensuite redressé pour atteindre un sommet de 2388,62 USD le 25 avril

2017. Tout au long de la période considérée, la tendance globale a été à la hausse, mais avec des fluctuations à court terme.

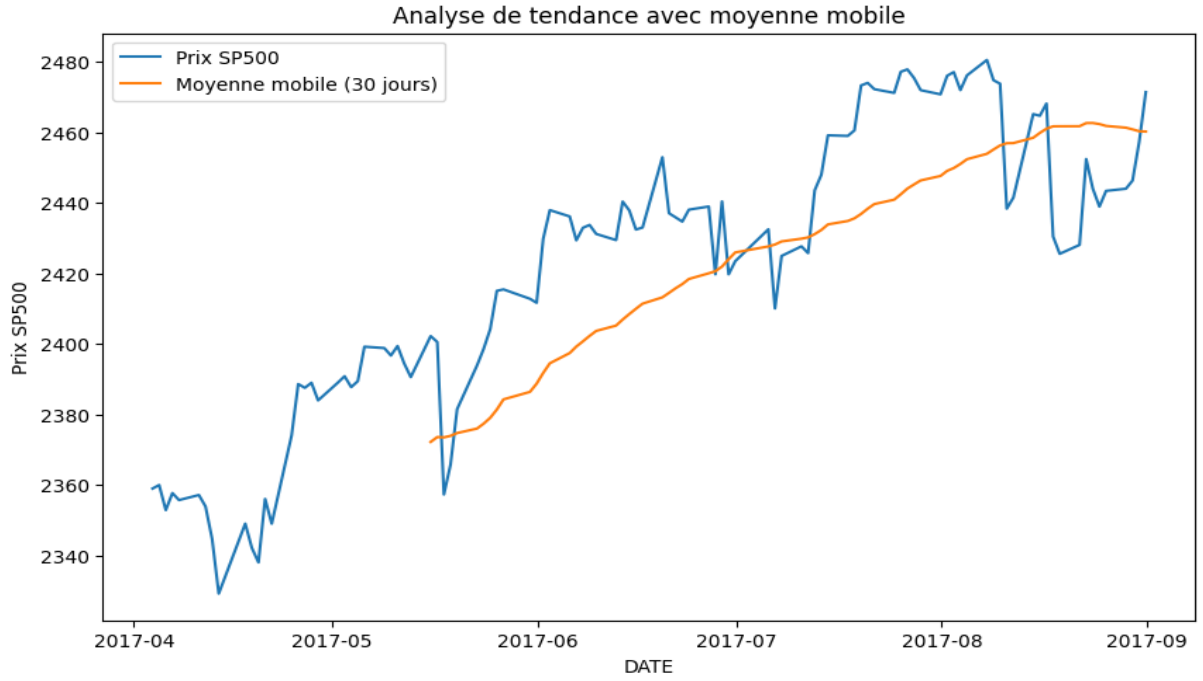
En conclusion, bien que l'indice SP500 ait connu des fluctuations importantes sur la période considérée, il a globalement augmenté de 3,2 %. La tendance globale a donc été à la hausse, mais avec des fluctuations à court terme, indiquant une certaine volatilité sur le marché.

Figure 02: Evolution des prix de l'indice SP500



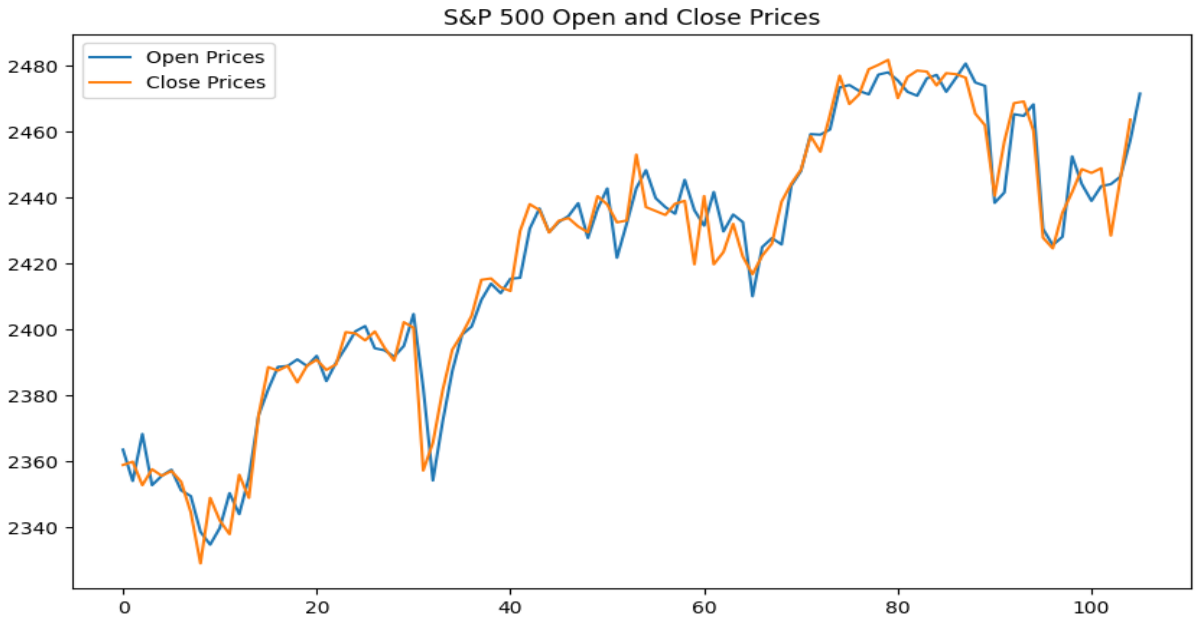
Le calcul et la visualisation de moyenne mobile sur 30 jours (figure 03) confirme notre constat de la tendance générale des valeurs de l'indice SP500. On peut voir que la moyenne augmente progressivement au fil du temps, ce qui suggère une croissance globale de l'indice SP500. Cependant, on remarque aussi les fluctuations constatées plus tôt autour de cette tendance, ce qui peut indiquer des variations saisonnières ou des effets ponctuels.

Figure 03: tendance avec la moyenne mobile des prix de l'indice SP500



En étudiant les prix de l'indice à la clôture et à l'ouverture de chaque jour (figure 04), on peut observer que la valeur de l'indice à la clôture est souvent inférieure à celle de l'ouverture pour le même jour. Cependant, il y a certaines exceptions à cette tendance, notamment le 10 avril 2017, où la valeur de l'indice à 20h00 est légèrement supérieure à celle de 13h30.

Figure 04: Comparaison des prix d'ouverture et de clôture de l'indice SP500



2) Modélisation

Les modèles de réseaux de neurones sont des algorithmes d'apprentissage automatique puissants qui sont utilisés pour résoudre une variété de problèmes de prédiction, tels que la prédiction de prix d'actifs financiers ou la prévision de tendances de marché. Dans cette partie, nous examinons quatre modèles de réseaux de neurones différents: le modèle classique, le réseau de neurones convolutionnel (CNN), la couche personnalisée CLTN et le modèle Long Short-Term Memory (LSTM).

Nous utilisons également d'autres algorithmes de régression tels que le Support Vector Regression (SVR) et le Perceptron Multicouche de Régression. Ces modèles de régression sont également couramment utilisés pour la prédiction de valeurs numériques et peuvent être très efficaces en fonction des caractéristiques des données.

Pour chaque modèle, nous présentons une brève description de sa structure et de son fonctionnement, ainsi que des résultats d'apprentissage obtenus sur des données de test et de validation. Nous comparons également les performances de ces différents modèles pour voir comment ils se comparent les uns aux autres.

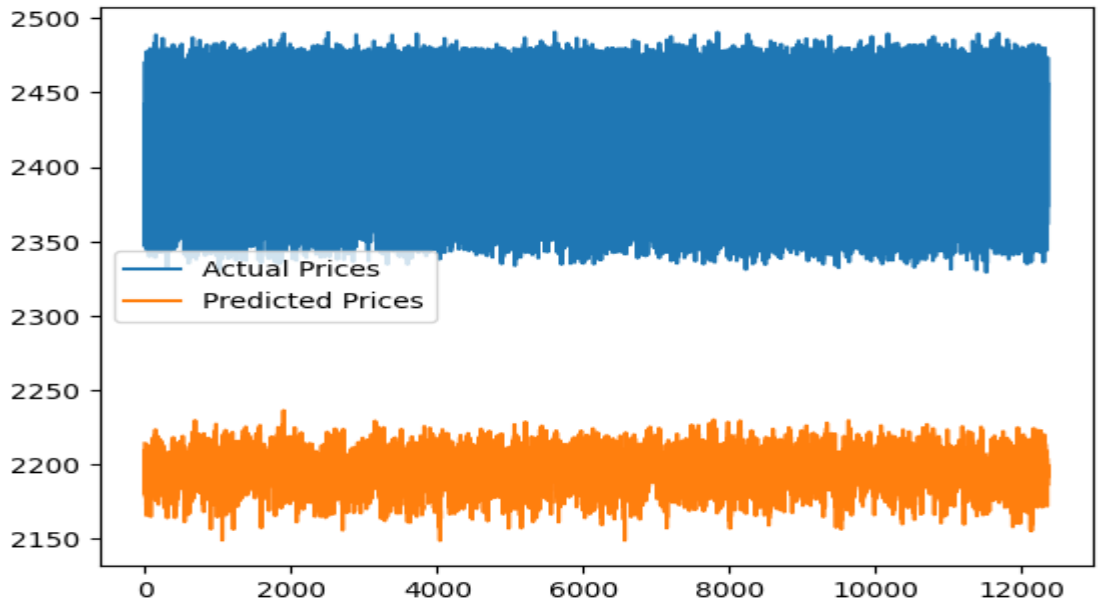
a) Model classique :

Nous avons d'abord construit un modèle de réseau de neurones à l'aide de l'API Keras de TensorFlow. Le modèle est composé de trois couches entièrement connectées, chacune avec une fonction d'activation ReLU, et a été compilé avec l'optimiseur Adam et la fonction de perte de l'erreur quadratique moyenne (MSE) pour résoudre un problème de régression.

Pour éviter l'ajustement excessif, nous avons ajouté des couches d'abandon avec un taux d'abandon de 0,2 et avons utilisé l'arrêt précoce. Nous avons entraîné le modèle sur l'ensemble d'apprentissage pendant 100 époques avec une taille de lot de 32, en utilisant l'ensemble de validation pour contrôler les performances pendant la formation.

Nous avons constaté que la perte d'apprentissage diminue progressivement au fil des époques, mais la perte de validation diminue initialement puis recommence à augmenter après quelques époques, ce qui peut indiquer un surajustement du modèle aux données d'apprentissage.

Figure 05: Résultats de prédiction de l'indice SP500

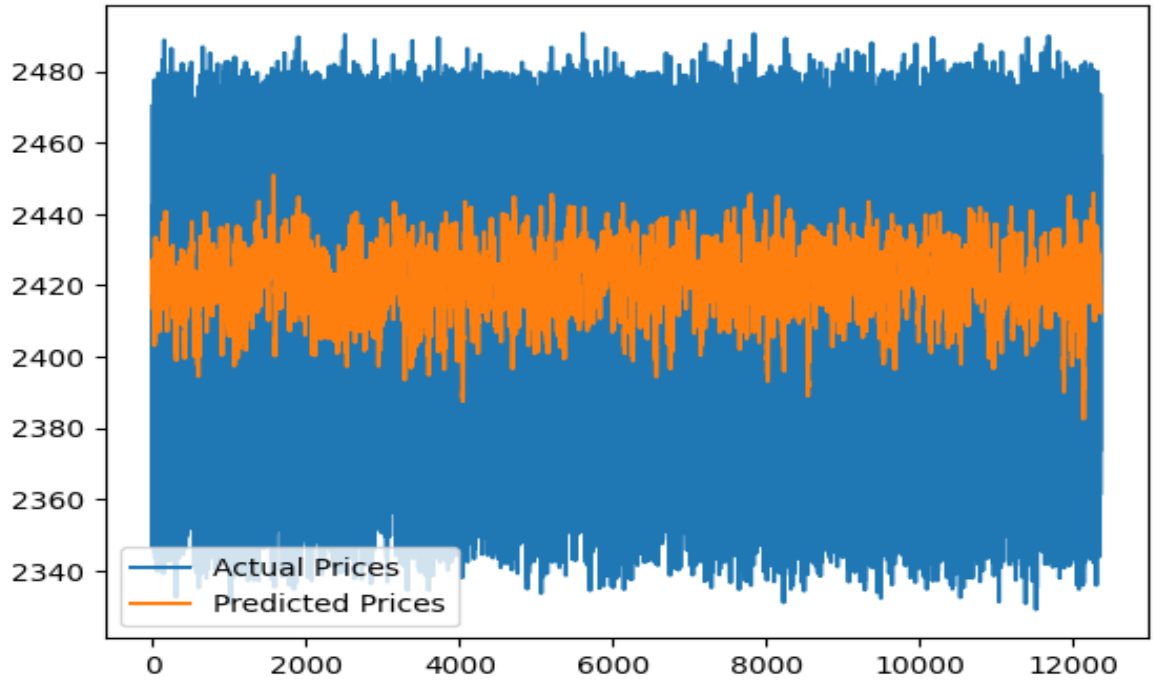


b) Réseau de neurones convolutionnel (CNN) :

Nous avons créé un modèle de réseau de neurones convolutionnel (CNN) afin de prédire les prix de l'indice S&P 500. Nous avons utilisé une couche de convolution 1D, suivie d'une couche de mise en commun, et une couche Flatten pour transformer les sorties de la couche de mise en commun en un vecteur unidimensionnel. Les sorties sont ensuite alimentées dans une couche Dense avec 50 neurones et une activation, suivie d'une dernière couche Dense avec une seule sortie sans activation. Nous avons compilé le modèle avec l'optimiseur Adam et la fonction de perte mean_squared_error.

Nous avons observé que la valeur de perte diminue au fur et à mesure que le nombre d'époques augmente, indiquant que le modèle apprend et devient plus performant dans la prédiction de la variable cible. Cependant, il est important de noter que même une valeur de perte très faible ne garantit pas que le modèle est le meilleur, car il est possible que le modèle soit surajusté aux données d'entraînement et ne puisse pas bien se généraliser à de nouvelles données.

Figure 06: Résultats de prédiction de l'indice SP500

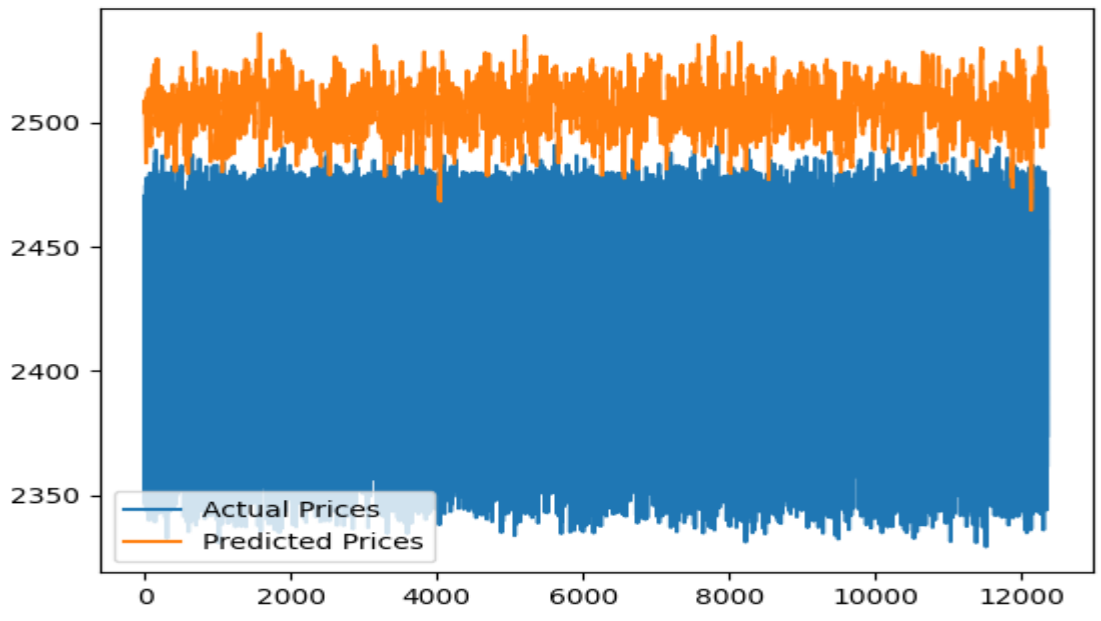


c) Complex-valued Linear Transformation Network (CLTN) :

Ce modèle est basé sur une couche personnalisée appelée CLTNLayer dans Keras, qui est basée sur l'architecture CLTN (Complex-valued Linear Transformation Network) qui applique une transformation linéaire à valeurs complexes à l'entrée. La couche se compose de poids entraînables, qui sont initialisés à l'aide de l'initialisateur uniforme Glorot et optimisés pendant l'entraînement à l'aide de la rétropropagation.

Les résultats suggèrent que le modèle apprend des données et améliore ses prédictions au fil du temps, mais que l'amélioration a peut-être atteint un plateau. Cependant, les valeurs de perte sont assez élevées, ce qui suggère que le modèle n'est peut-être pas très performant.

Figure 07: Résultats de prédiction de l'indice SP500

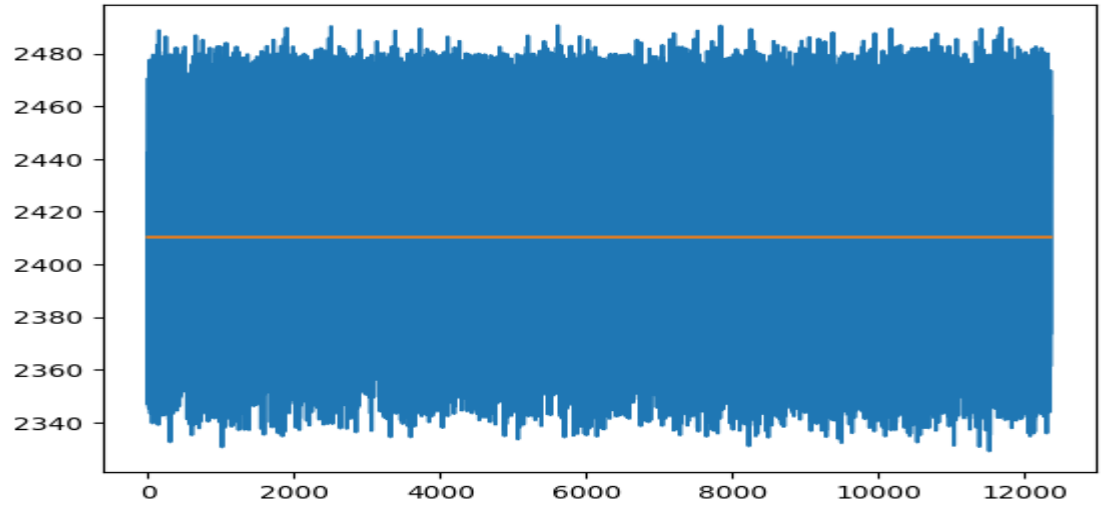


d) Long Short-Term Memory (LSTM):

Nous avons utilisé un modèle LSTM pour prédire les prix historiques du S&P500. Le modèle est composé d'une couche LSTM avec 32 neurones, d'une couche Dropout et d'une couche Dense. La classe TimeseriesGenerator est utilisée pour créer des séries temporelles pour l'apprentissage du modèle.

Au début de l'entraînement, la perte est relativement élevée, mais elle diminue progressivement au fil des époques. On peut voir que la perte atteint une valeur stable à partir de l'époque 20. Cela suggère que le modèle a atteint un état stable et que l'entraînement supplémentaire s'améliorera probablement pas les performances du modèle.

Figure 08: Résultats de prédiction de l'indice SP500

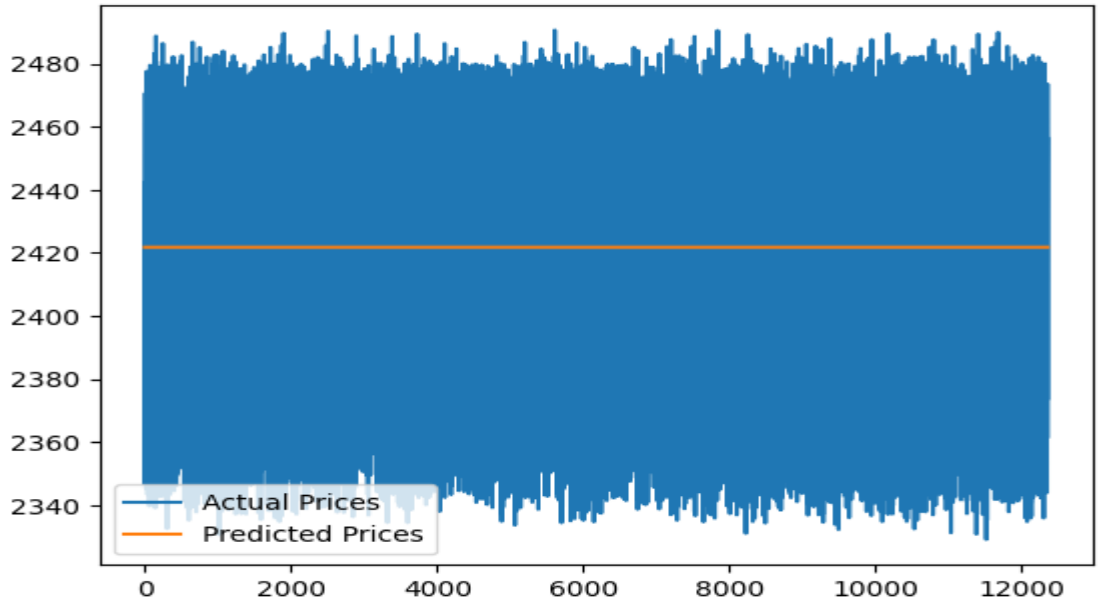


e) Perceptron Multicouche de Régression (MLP) :

Le modèle MLP est constitué de plusieurs couches cachées en plus de la couche d'entrée et de la couche de sortie. Les couches cachées comprennent chacune 500 neurones. La fonction d'activation choisie est la fonction "tanh" qui est responsable d'ajouter la non-linéarité dans la relation.

Les résultats de prédiction de notre modèle sont présentés dans le graphique ci-dessous (Figure 09)**.** Nous analysons le RMSE pour discuter de l'efficacité de notre modèle. La valeur de ce dernier s'avère être de 39.81. Autrement dit, l'erreur moyenne pondérée entre les prédictions et les valeurs réelles des valeurs de l'indice dans cet ensemble de données est de 39.81.

Figure 09: Résultats de prédiction de l'indice SP500

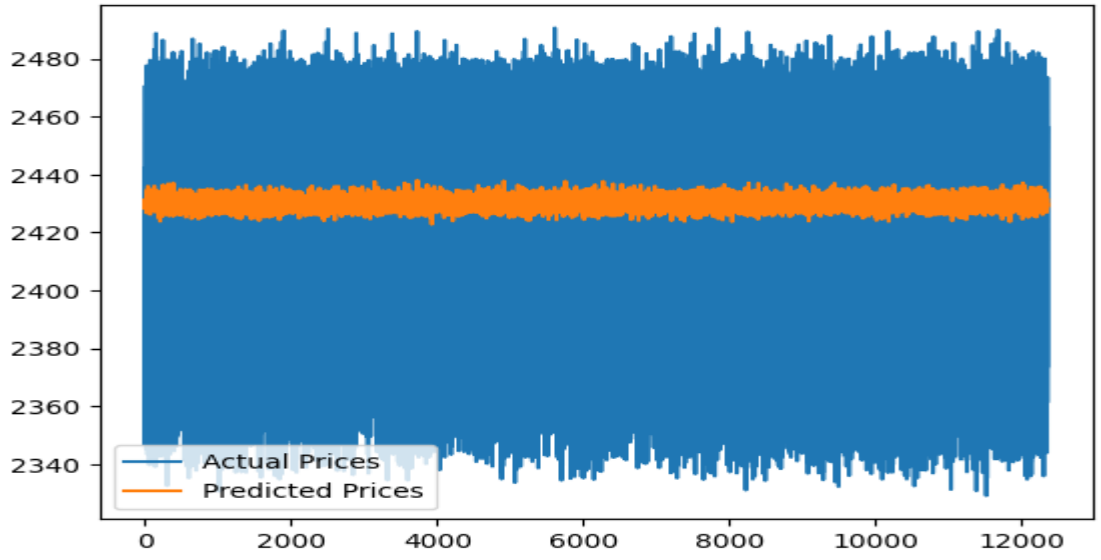


f) Support vector regression (SVR)

En supposant la non-linéarité de nos données, nous utiliserons la capacité du SVM à tenir compte de la non-linéarité des données pour les besoins de la régression, ce qui fait du SVR un outil efficace pour la prévision des séries temporelles. Le SVM a pour objectif de trouver la meilleure ligne d'ajustement afin de prédire les prix de l'indice S&P 500.

De la même façon que le modèle MLP, le RMSE pour ce modèle de régression s'avère être de 40.92. Autrement dit, l'erreur moyenne pondérée entre les prédictions et les valeurs réelles des valeurs de l'indice dans cet ensemble de données est de 40.92.

Figure 10: Résultats de prédiction de l'indice SP500



3) Comparaison des performances des models :

En comparant les performances des 6 modèles en termes de RMSE (Root Mean Squared Error), nous pouvons constater que le modèle de réseau de neurone classique a le plus haut RMSE, ce qui suggère qu'il a la moins bonne performance. Les modèles CNN, LSTM, MLP régression et SVR ont tous des RMSE assez similaires, avec des scores compris entre 39.81 et 41.22. Cela suggère que ces modèles ont des performances similaires.

Le modèle CLTN a un RMSE plus élevé que les modèles CNN, LSTM, MLP régression et SVR, mais il est inférieur au modèle classique. Cela suggère que le modèle CLTN a une performance intermédiaire.

En conclusion, en termes de RMSE, les modèles CNN, LSTM, MLP régression et SVR ont des performances similaires et sont meilleurs que le modèle classique, tandis que le modèle CLTN a une performance intermédiaire.

Tableau 01 : Résultats de prédiction de l'indice SP500

Modèle	**RMSE (Root Mean Squared Error)**
Classique	271.07
CNN	41.21
CLTN	59.07
LSTM	41.22
MLP régression	39.81
SVR	40.92

Conclusion

Notre étude sur la prédiction du prix du SP500 à l'aide de techniques de machine learning et de deep learning avancées a permis de constater que les fluctuations des marchés financiers sont complexes et difficiles à prévoir. Dans le cadre de notre analyse, nous avons exploré plusieurs modèles de prédiction des séries temporelles financières, y compris des modèles classiques et des modèles de réseaux de neurones tels que CNN, CLTN, LSTM, MLP régression et SVR.

En utilisant le Root Mean Squared Error (RMSE) comme mesure de performance, nous avons comparé les performances de ces modèles et avons constaté que le modèle MLP régression était le plus précis, suivis par les modèles CNN, LSTM et SVR. Le modèle classique de réseau de neurones avait la pire performance, tandis que le modèle CLTN avait une performance intermédiaire.

Ces résultats soulignent l'importance des techniques de machine learning et de deep learning pour la prédiction des séries temporelles financières, ainsi que la nécessité de choisir le bon modèle pour chaque tâche de prédiction. En fournissant des prévisions plus précises pour le prix du SP500, notre étude peut aider les investisseurs à prendre des décisions plus éclairées en matière d'investissement, ce qui peut avoir un impact significatif sur leur portefeuille et leur avenir financier.

Une contrainte importante de notre travail était la réalisation d'un grid search efficace pour choisir les nombres de couches et de neurones les plus adaptés. Cette contrainte était liée à des limitations techniques, car la mise en place d’un grid search efficace a été limitée par le manque de puissance de calcul et de mémoire disponible sur nos machines.

En ce qui concerne les perspectives d'amélioration, nous aurions donc voulu utiliser des techniques d'optimisation plus avancées pour chercher plus efficacement les hyper paramètres optimaux de nos modèles. En outre, nous aurions voulu évaluer la performance de nos modèles sur des ensembles de données plus variés pour évaluer leur capacité de généralisation. Enfin, une autre possibilité aurait été d'implémenter des algorithmes de deep learning plus complexes pour améliorer la précision et la rapidité de nos modèles.

Annexes

Annexe 01: statistiques descriptives

	SP500
count	41266.000000
mean	2421.537882
std	39.557135
min	2329.139900
25%	2390.860100
50%	2430.149900
75%	2448.820100
max	2490.649900