MID TERM - PRÉDICTION DES PRIX ET ACTIFS DU S&P 500 GRÂCE À DU MACHINE LEARNING SIMPLE

PROJET DATA

DONNÉES ET DÉCISION FINANCIÈRE ET ÉCONOMIQUE – 2025

COMASSETO Vitória
ITURRALDE Martín
URTUBIA Carolina
VIOLA CARVALHO Henrique

OBJECTIFS

Objectif Principal :

- Développer un modèle prédictif basé sur le machine learning pour estimer les prix et rendements du S&P 500.

Objectif Spécifiques :

- Collecter et traiter des données historiques et macroéconomiques.
- Comparer différents modèles prédictifs (ARIMA, Random Forest, LSTM, etc.).
- Optimiser les performances via une analyse approfondie.
- Fournir des recommandations d'investissement précises, axées sur la gestion des risques et les opportunités.

MÉTHODOLOGIE ET ANALYSE

Partie 1 : Collecte et Préparation des Données

Partie 2 : Développement du Modèle Prédictif

- **Partie 2.1 :** Régression Linéaire
- **Partie 2.2 :** ARIMA (AutoRegressive Integrated Moving Average)
- **Partie 2.3 :** Random Forest
- **Partie 2.4 :** LSTM

PARTIE 1 : COLLECTE ET PRÉPARATION DES DONNÉES

Étape Cruciale :

- Collecter et préparer des données pour garantir la précision et l'efficacité des modèles de Machine Learning en intégrant des facteurs historiques et macroéconomiques (Yahoo Finance et FRED).

Variables du Modèle :

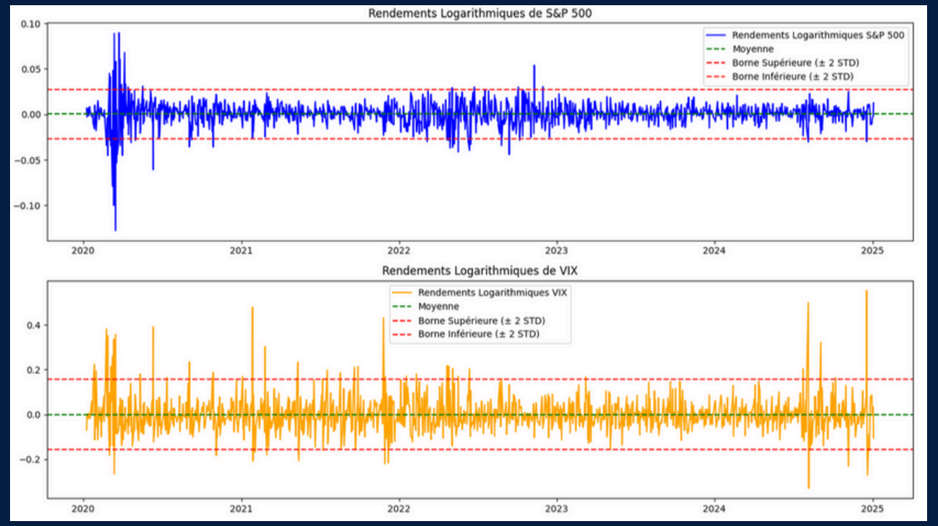
- **Variable dépendante** : Rendement logarithmique du S&P 500.
- **Variables indépendantes** :
  - **Indicateurs techniques** : SMA, RSI, MACD.
  - **Variables exogènes** : Taux d'intérêt, indice de volatilité (VIX), inflation, taux de chômage.

Traitement des Données :

- **Resampling** : Fréquence journalière à mensuelle.
- Suppression des doublons et des données manquantes.
- Fusion et alignement des séries temporelles.

PARTIE 2 : DÉVELOPPEMENT DU MODÈLE PRÉDICTIF

- **Transformation** : Conversion en rendements logarithmiques (log-returns) pour stationnariser les séries temporelles.
- **Enrichissement** : Ajout d'indicateurs techniques (SMA, RSI, MACD) pour analyser les tendances du marché.
- **Division des données** :
  - **Entraînement** : 80 %.
  - **Test** : 20 %.



Graphique 1: Évolution des Rendements Logarithmiques du S&P 500 et du VIX avec Bornes de Confiance

PARTIE 2.1 : RÉGRESSION LINÉAIRE

- **Description :** Modèle statistique reliant une variable cible (rendements financiers) à des variables explicatives pour prédire des valeurs continues.

- **Résultats :**
  - **MAE :** Faible (0.0045), montrant des prédictions proches des valeurs réelles.
  - **RMSE :** Modéré (0.0057), soulignant les grandes erreurs.
  - **Précision :** 76.89 %, capable de prédire correctement la direction des rendements.

- **Limites :** Simple et efficace pour les tendances générales, mais limité dans la capture des relations complexes ou non linéaires.

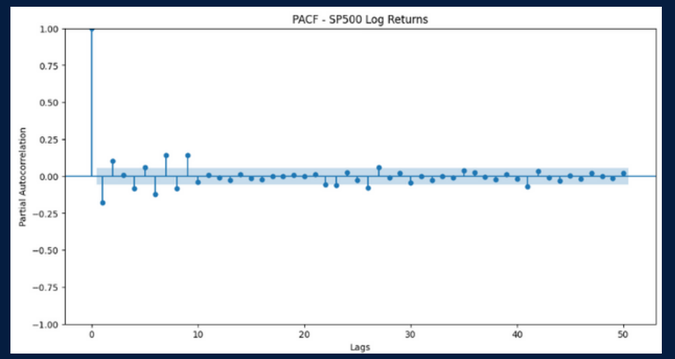
PARTIE 2.2 : ARIMA (AUTOREGRESSIVE INTEGRATED MOVING AVERAGE)

- **Description :**
  - Modèle utilisé pour analyser et prédire les séries temporelles stationnaires, basé sur les termes auto-régressifs (AR), de différenciation (I) et de moyenne mobile (MA).
  - Identifie les lags pertinents via les graphiques ACF et PACF.

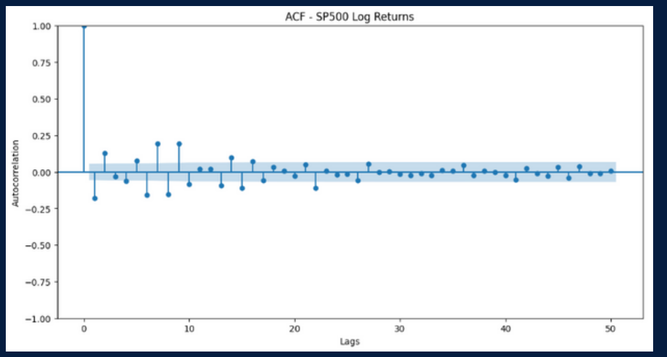
- **Paramètres déterminés :**
  - **p = 1 (AR) :** Autocorrélation significative au premier lag.
  - **q = 1 (MA) :** Moyenne mobile ajustée au premier lag.
  - **Modèle final : ARIMA(1, 0, 1).**

- **Résultats :**
  - **MAE :** 0.0059.
  - **RMSE :** 0.0081 (erreurs supérieures à la régression linéaire).
  - **Précision :** 56.13 %, faible pour prédire les directions (hausse/baisse).

- **Limites :**
  - Performant pour des tendances linéaires à court terme sur des séries stationnaires.
  - Moins adapté aux données volatiles et aux fluctuations complexes.



Graphique 3: PACF - SP500 Log Returns



Graphique 2: ACF - SP500 Log Returns

PARTIE 2.3 : RANDOM FOREST

- **Description :**
  - Modèle d'ensemble supervisé basé sur plusieurs arbres de décision indépendants.
  - Combine leurs prédictions via moyenne ou vote.
  - Adapté aux relations non linéaires et limite les risques de surajustement.

- **Mise en œuvre dans le projet :**
  - Entraîné sur les données d'entraînement (X_train, y_train).
  - Prédictions générées sur l'ensemble de test (X_test).

- **Avantages :**
  - Captures efficaces des relations complexes.
  - Résistant aux surajustements grâce à l'approche en sous-échantillons.

MAE (Mean Absolute Error)	0.0041	Précision globale satisfaisante
RMSE (Root Mean Squared	0.0052	Accent sur les grandes erreurs
Error)		
Accuracy	75.47 %	Le modèle prédit correctement la direction des rendements dans la majorité des cas

Table 1: MAE, RMSE et Accuracy du Random Forest

PARTIE 2.4 : LSTM

- **Description :**
  - Réseaux neuronaux récurrents conçus pour apprendre les dépendances temporelles dans les séries chronologiques.
  - Idéal pour détecter des modèles complexes dans des séries temporelles volatiles.

- **Mise en œuvre :**
  - Modèle configuré avec une couche LSTM (50 neurones) pour capturer les dépendances temporelles.
  - Suivi d'une couche dense pour produire une sortie unique.
  - Optimiseur : Adam.
  - Fonction de perte : MSE (Minimisation des erreurs).

- **Résultats :**
  - **MAE** : 0,0055.
  - **RMSE** : 0,0070 (erreurs modérées, légèrement supérieures à Random Forest).
  - **Précision** : 77,35 %, bonne performance pour prédire la direction des rendements.

- **Limites :**
  - Captures efficaces des dépendances temporelles complexes, mais la sensibilité aux données peut augmenter légèrement les erreurs.

ANALYSE DES RÉSULTATS

- **Régression Linéaire :**
  - Simple et fiable, avec une précision de 76,88 %.
  - MAE et RMSE modérées, adaptées pour des tendances simples.

- **ARIMA :**
  - Performances les plus faibles : précision de 56,13 %, MAE et RMSE les plus élevées.
  - Non adapté aux séries temporelles volatiles.

- **Random Forest :**
  - Modèle le plus performant avec des MAE et RMSE très faibles.
  - Précision élevée de 77,83 %, capturant efficacement les relations non linéaires.

- **LSTM :**
  - Bonne capacité à détecter les dépendances temporelles complexes.
  - MAE et RMSE légèrement plus élevés que Random Forest, précision de 77,35 %.

Comparaison des Modèles:

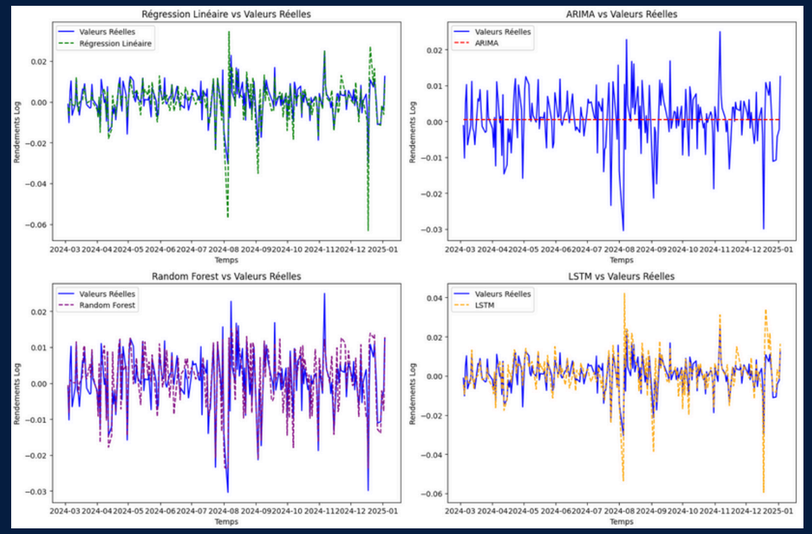
	Modèle	MAE	RMSE	Accuracy
0	Régression Linéaire	0.004050	0.005749	0.768868
1	ARIMA	0.005988	0.008109	0.561321
2	Random Forest	0.004111	0.005183	0.754717
3	LSTM	0.004458	0.006424	0.778302

Table 2: Comparaison entre les modèles

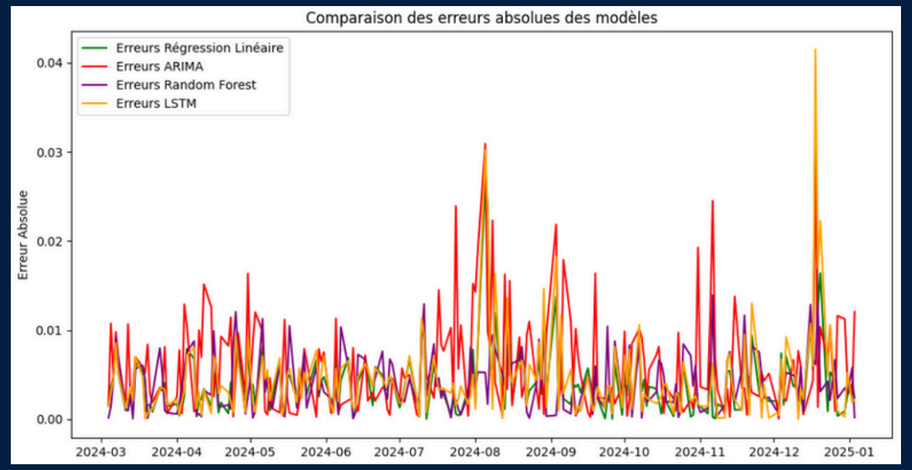
ANALYSE DES RÉSULTATS

- **Régression Linéaire et Random Forest :** erreurs plus faibles et stables.
- **LSTM :** des variations marquées dans des conditions de forte volatilité.
- **ARIMA :** pics fréquents, révélant une faible capacité à capturer les fluctuations du marché.

→ **Random Forest** (idéal pour les relations non linéaires) **et LSTM** (préférable pour les dépendances temporelles) **s'imposent comme les meilleurs choix selon les objectifs.**



Graphique 4: Modèles vs valeurs réelles



Graphique 4: Modèles vs valeurs réelles

CONCLUSION

- **Performances des Modèles :**
  - **Régression Linéaire :** Simple, précision modérée, faible MAE, mais limité pour les relations non linéaires.
  - **ARIMA :** Faibles performances, inadapté aux données volatiles du S&P 500.
  - **Random Forest : Meilleur modèle pour les dépendances non linéaires, avec des MAE et RMSE compétitifs.**
  - **LSTM :** Excellente capacité à capturer les dépendances temporelles complexes.

- **Enseignements :**
  - Le choix du modèle dépend des caractéristiques des données et des objectifs spécifiques.
  - Pour des marchés volatiles comme le S&P 500, des modèles comme Random Forest et LSTM sont essentiels pour des prédictions robustes.

- **Implications :**
  - Importance d'une préparation minutieuse des données et d'une analyse rigoureuse.
  - Ces approches permettent de développer des stratégies d'investissement robustes et informées.

Centrale Méditerranée

MERCI