# MIDTERM S&P 500

*Prédiction des Prix et Actifs du S&P 500 avec le Machine Learning*

*Timothée Henriot - Eliott Ollivier - Marie Simon - Mathias Buée - Paul Consalès*

# Collecte des données

*Sur une période de 5 ans*

## S&P 500

- Téléchargées via l'API de Yahoo Finance (données de clôture journalières, volumes de transaction, et autres indicateurs financiers.

## Variables macroéconomiques

- **CPIAUCSL** : Indice des  prix à la consommation (CPI)
- **GDPC** : Produit intérieur brut (PIB) , extrait depuis FRED (federal reserve Economic Data).

*Données téléchargées sous forme de fichiers CSV (synchronisation des mesures économiques avec les données de marché)*

# Préparation des données

## Nettoyage et préparation

- Traitement des valeurs manquantes
- Elimination des doublons
- Conversion des prix en rendements logarithmiques

## Alignement des séries temporelles

Fusion des ensembles de données avec la prise en compte des **différences dans les fréquences de publications** : quotidiennes (S&P),  mensuelle ou trimestrielle (indicateurs économiques).


# Feature Engineering

## Indicateurs techniques  calculés à partir des données du S&P :

- Moyennes mobiles
- Indice de force relative (RSI)
- Convergence/ divergence des moyennes mobiles (MACD)

Objectif : enrichir le modèle et améliorer la qualité des prédictions

## Sélection des variables les plus significatives

- Analyses de corrélation
- Méthodes de sélection de features.

Objectif : réduire la dimensionnalité des données tout en conservant les informations les plus critiques pour la prédiction des rendements.

# Développement initial du Modèle

## Mise en place

- Application de techniques de régression linéaire (base de comparaison)
- Données divisées en ensembles d'entraînement et de test selon un ratio 80/20
- Mise en place d'une validation croisée (généralisabilité des résultats)

Les premiers résultats servent de benchmark pour les étapes suivantes du développement, permettant d'aborder les modèles plus sophistiqués comme les Random Forests.

## Analyse de performance des modèles

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Accuracy

# Résultats et analyse

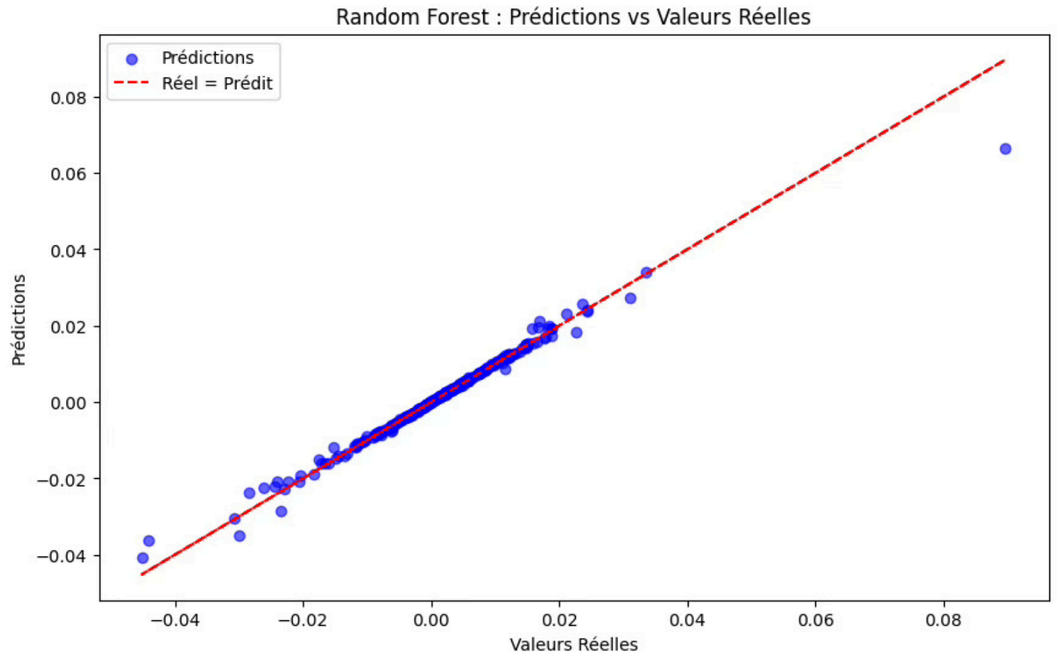
## Evaluation des modèles : Régression linéaire VS Random Forest

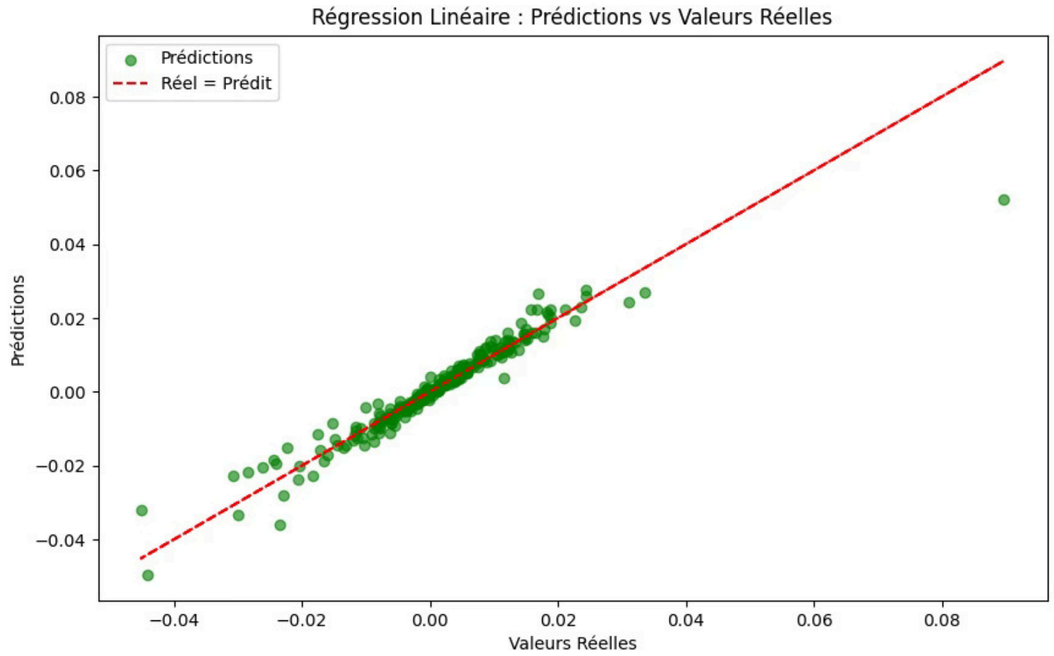
| | Modèle | MAE | RMSE | $R^2$ |
|---|---|---|---|---|
| 0 | Régression Linéaire | 0.001830 | 0.001873 | 0.922304 |
| 1 | Random Forest | 0.000582 | 0.001873 | 0.977799 |

- **MAE** (Mean Absolute Error) : nettement inférieur pour le Random Forest => meilleure précision des prédictions.
- **$R^2$** (coefficient de détermination) : montre que les deux modèles expliquent une grande part de la variance des données, la Random Forest ayant une performance légèrement supérieure.
- **RMSE** (Root Mean Squared Error) : équivalent pour les deux modèles, sûrement en raison d'erreurs systématiques non corrigées.

# Résultats et analyse

## Visualisation des résultats

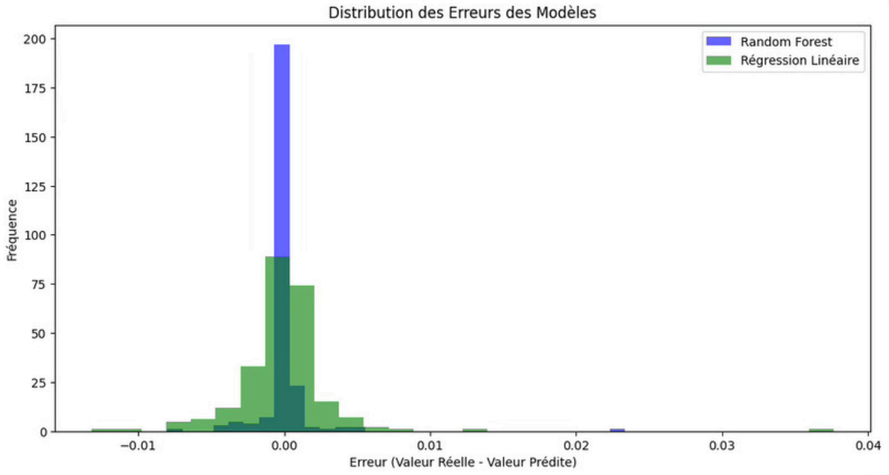



Les prédictions de **Random Forest** suivent de manière plus étroite les valeurs réelles, avec moins de dispersion autour de la ligne idéale (y = x).

# Résultats et analyse

## Distribution des erreurs




**Random Forest** produit une distribution d'erreurs bien centrée autour de zéro avec une variance plus faible par rapport à la régression linéaire.

## Sources d'erreurs des modèles

- Bruitage éventuel des données financières (S&P 500, PIB, inflation).
- Absence de prise en compte de certains facteurs exogènes  (annonces économiques ou chocs externes)


# Résultats et analyse

## Limites

**Régression linéaire** :

- Incapacité à capturer les relations non linéaires.
- Susceptibilité à la multicolinéarité, rendant les coefficients instables.
- Pics d'erreurs résiduelles significatifs autour des périodes de forte volatilité.

**Random Forest** :

- Difficulté à identifier les relations spécifiques entre les variables.
- Sur-ajustement avec une trop grande adaptation aux données d'entraînement.
- Erreurs résiduelles persistantes lors des retournements de tendance rapide.

# Résultats et analyse

## Améliorations envisagées

- Optimiser les hyperparamètres de la Random Forest (nombre d'arbres, profondeur maximale, critère d’évaluation) via une recherche en grille.
- Tester des variantes comme Gradient Boosting ou XGBoost pour des performances accrues.
- Ajouter des variables exogènes et intégrer des indicateurs avancés (flux de capitaux, données sectorielles, annonces macroéconomiques).
- Réduire le bruit des données avec des techniques de débruitage (filtres ou transformations spécifiques).
- Explorer des modèles alternatifs comme LSTM pour les dépendances temporelles complexes ou Ridge  (pour la régression linéaire) pour limiter la multicolinéarité.