Projet BIG DATA (Lubabah HAMOUCH & Mohammed FAKIR)

Prédiction de l'Impact des Événements Mondiaux sur les Prix des Actions via l'Analyse des Sentiments et les Modèles de Machine Learning

Résumé du Projet

L'objectif de ce projet est de prédire les fluctuations des prix des actions (par exemple, l'indice S&P 500, actions individuelles) en utilisant les données de sentiment extraites d'articles de presse et de publications sur les réseaux sociaux concernant des événements mondiaux. Ces prédictions sont ensuite réalisées en appliquant des modèles de machine learning tels que **Régression Linéaire**, **XGBoost**, **Random Forest**, et **Gradient Boosting**. Les résultats de ces prédictions sont ensuite visualisés à l'aide de Power BI pour fournir des analyses visuelles et interactives.

1. Collecte et Prétraitement des Données (Big Data et Web Scraping)

A. Collecte de Données via Web Scraping et APIs

Le projet repose sur la collecte de données provenant de plusieurs sources afin de générer des prédictions fiables des fluctuations des prix des actions.

1. Web Scraping :

- Des outils comme **BeautifulSoup** et **Scrapy** sont utilisés pour extraire des données à partir de sites de nouvelles comme **Reuters**, **Bloomberg** et **BBC**. Ce scraping permet de collecter des articles relatifs à des événements mondiaux majeurs (politiques, économiques, catastrophes naturelles, etc.) qui pourraient influencer les marchés financiers.

2. APIs :

- Les **APIs** comme **Twitter** et **Reddit** permettent de récupérer des publications sociales, tweets et commentaires sur des événements mondiaux. L'analyse des sentiments à partir de ces plateformes aide à déterminer la perception publique sur des événements clés et leur impact sur les marchés boursiers.

3. Données Économiques et Boursières :

- Les données historiques sur les prix des actions et les indicateurs économiques (comme le PIB, l'inflation) sont récupérées via des **APIs** comme **Yahoo Finance**, **Alpha Vantage**, et **Quandl**.

B. Base de Données MongoDB

Toutes les données collectées, incluant les articles de presse, les posts sur les réseaux sociaux et les données boursières, sont stockées dans une base de données **MongoDB**. Ce choix de base de données NoSQL permet de gérer efficacement des données non structurées, comme des textes provenant d'articles et de tweets, ainsi que des données financières.

- **Structure des Données** : Les documents dans MongoDB contiennent des informations telles que :
    - **Articles de presse** : Titre, URL, texte, sentiment calculé, date.
    - **Tweets/Réseaux Sociaux** : Texte des posts, date, sentiment, mentions d'événements.
    - **Données boursières** : Prix des actions, volumes d'échanges, dates, indices.

2. Analyse des Sentiments

A. Prétraitement des Textes

Les textes récupérés (articles de presse et posts sur les réseaux sociaux) sont nettoyés pour supprimer les éléments non pertinents comme les stop words, la ponctuation et les liens. Les étapes suivantes incluent la **tokenisation** et l'application de **modèles d'analyse des sentiments** pour évaluer l'émotion associée à chaque événement.

B. Modèles d'Analyse des Sentiments

L'analyse des sentiments permet d'assigner un score de sentiment à chaque article ou tweet. Voici les modèles utilisés :

- **VADER (Valence Aware Dictionary and sEntiment Reasoner)** : Idéal pour les textes courts comme les tweets et les posts, VADER attribue des scores indiquant si le sentiment est **positif**, **négatif** ou **neutre**.
- **TextBlob** : Une méthode plus simple pour extraire les sentiments des textes, qui donne aussi un score pour le sentiment positif ou négatif.

Les scores de sentiment sont ensuite agrégés pour chaque événement afin de déterminer l'impact global de l'événement sur le marché.

3. Modèles de Machine Learning pour la Prédiction des Prix des Actions

A. Objectif de la Modélisation

L'objectif principal de l'utilisation du machine learning est de prédire les **fluctuations des prix des actions** en fonction des sentiments collectés des événements mondiaux et des données économiques.

B. Modèles Utilisés

Voici les modèles de machine learning appliqués pour prédire les mouvements des prix des actions :

1. **Régression Linéaire** :
    - La régression linéaire est utilisée pour prédire les **changements** de prix des actions en fonction des scores de sentiment. Ce modèle est simple, mais efficace pour établir une première relation entre sentiment et prix.
2. **Random Forest** :
    - Le modèle **Random Forest** est un ensemble d'arbres de décision qui permet de capturer des interactions complexes entre les variables, telles que les changements de prix des actions en fonction des différents types de sentiments et des événements.
3. **XGBoost (Extreme Gradient Boosting)** :
    - **XGBoost** est un modèle de boosting qui offre une meilleure performance en raison de sa capacité à gérer des données déséquilibrées et à réduire l'overfitting. Il est utilisé pour prédire les fluctuations boursières avec un niveau de précision plus élevé.
4. **Gradient Boosting** :
    - Un autre modèle de boosting est appliqué pour améliorer les prédictions, en ajustant les erreurs du modèle de base et en obtenant ainsi une meilleure précision pour les prédictions boursières.

C. Évaluation des Modèles

Les modèles sont évalués en utilisant plusieurs métriques d’évaluation pour s’assurer de leur performance :

- **RMSE (Root Mean Squared Error)** : Mesure l'erreur moyenne entre les valeurs prédites et réelles.
- **MAE (Mean Absolute Error)** : Mesure de l’erreur absolue moyenne des prédictions.
- **R^2 (Coefficient de Détermination)** : Indicateur de la capacité du modèle à expliquer la variance des données.