PROJET IA : Prédiction des mouvements des marchés financiers (Marché des matières premières / « Commodities ») (Apprentissage supervisé)

1/ Collecte des données : trouver sources de données (ex : API de fournisseurs de données financières) sur une période suffisamment longue (entrainement efficace du modèle)

* Kaggle : fichier csv (période de 10 ans de 2012 à 2022)
* Investing.com : fichier csv (période de 30 ans de 1974 à 2024) + webscrapping prix actuel

2/ Pré-traitement des données : Nettoyer ces données (supp les valeurs aberrantes et compléter les valeurs manquantes + éliminer les doublons), les normaliser + si besoin créer des variables supp pertinentes (ex : indicateurs techniques)

3/ Exploration des données : visualisation, cycles, corrélations entre les données (séries temporelles, histogramme…

Voir exploratory\_data.py

Réalisé sur le fichier csv « commodities » qui contient le prix d’énergies fossiles et de ressources minérales qui sont les suivantes : pétrole brut, pétrole brent, argent, or, cuivre et gaz naturel

Observation du prix en fonction du temps : Les structures ne sont pas stables dans le temps => les séries temporelles ne sont pas stationnaires.

Globalement, des valeurs très faibles vers 2016 et 2020

Les données du fichier csv (prix) vont de 2012 à 2022

* Réalisation d’un graphique linéaire pour chaque matériau affichant l’évolution de son prix au fil du temps sur les 10 dernières années
* Réalisation de 2 matrices de corrélation (1 pour les 5 dernières années et l’autres sur les 10 dernières années)
* Réalisation d’un lot de sous-graphiques qui affiche les relations sous forme de nuage de points entre le matériau et tous les autres matériaux du df

4 / Conception du modèle : sélectionner les caractéristiques qui seront utilisées pour entraîner le modèle, diviser les données (test, validation, etc..)

* **1er modèle = Réseau de neurones (LSTM / RNN)**  :

Prédictions du prix de l’or

retrait des valeurs manquantes, 500 derniers jours du dataset pour les données test, préparer les données de test et d’entrainement pour l’apprentissage supervisé :

* On sélectionne les données de la colonne Gold pour l’ensemble d’entrainement. Ces données sont ensuite converties en un tableau numpy et remodelées en une seule colonne (les prix de l’or)
* Idem pour les données test
* On applique la normalisation (MinMaxScaler) à l’ensemble des données d’entrainement et de test

Cela permet de préparer les données à être utilisées avec un modèle de réseau de neurones récurrents, en particulier pour cette tâche de prédiction de séries temporelles comme les prix de l’or. Les données sont organisées en séquences temporelles où chaque séquence représente une période de 30 jours avec des observations successives des prix de l’or

* **2e modèle : SARIMA (qui prend en compte la saisonnalité d’une série temporelle)**

Série non stationnaire : decroissante lente vers 0 (autocorrélogramme ACF)

Dataset avec prix et date uniquement (capturer les effets saisonniers et tendances a long terme)

* **3e modele :**

Dataset avec comme features le prix, la date, le prix d’ouverture (« open »), le prix le + haut (dans la journée/dans le mois..), le prix + le bas (dans la journée/..), le volume et le pourcentage de changement (« Change)

* Le prix d'ouverture reflète le sentiment du marché au début de la journée de trading. Il peut indiquer comment les événements survenus en dehors des heures de marché (comme les nouvelles économiques, les décisions politiques, ou les catastrophes naturelles) ont affecté les attentes des investisseurs
* Le prix le plus élevé de la journée peut montrer le niveau de résistance, c’est-à-dire un point où les prix ont du mal à augmenter davantage. Cela peut indiquer une pression de vente ou un point où les investisseurs ont décidé de réaliser leurs bénéfices.
* Le volume représente le nombre de contrats échangés et est un indicateur clé de la liquidité et de l'intérêt du marché. Un volume élevé lors de mouvements de prix significatifs peut confirmer la force d'une tendance (hausse ou baisse).
* Une augmentation du volume peut précéder des mouvements de prix importants, car elle indique un intérêt croissant des investisseurs.
* Le pourcentage de changement par rapport à la journée précédente donne une mesure de la volatilité du marché et de la force du mouvement de prix.

Nous avons pris un dataset de données journalières sur 24 ans (depuis 2000) => prediction a court et moyen terme avec une bonne réactivité aux changements recents

5/ Entrainement et validation du modèle + validation croisée et optimisation des hyperparamètres

6/ Evaluation du modèle et comparaisons de plusieurs modèles entre eux (métriques)

Un marché financier est un lieu ou se rencontrent des vendeurs/acheteurs pour échanger des produits financiers

Le marché des matières premières regroupe des ressources brutes et des produits de base et se sépare en 2 catégories :

* Ressources minérales extraites ou exploitées (or, cuivre ..) + énergies fossiles (pétrole, gaz naturel .. ) = « hard commodities »
* Ressources agricoles (sucre, bétail ..) = « soft commodities »

Les matières 1eres sont principalement négociées par le biais de produits dérivées (tels que les contrats à terme ou les contrats futurs, les options ou les produits de bourses). Les transactions peuvent être effectuées de gré à gré ou sur des marché organisés (ex : NY Mercantile Exchange, London Metal Exchange). Plusieurs facteurs peuvent influencer le prix des matières 1eres : contexte géopolitique, économique, offre et la demande, aléas de la nature, etc. Les mouvements sur ce marché sont probablement ceux ressentis par le plus grand monde car ils affectent notre vie quotidienne à tous.

<https://www.kaggle.com/datasets/faseeh001/commoditiesinvestingcom?resource=download>

afficher les prediction a différentes échelles (ex : dans 1 mois, dans 6 mois) (et préciser le degré de fiabilité)

identifier la cible (appli destinée à ..) et tester l’influence de paramètres extérieures (inflation, guerre, etc) sur les predictions