Déroulement du stage
Sujet 5 : Machine Learning appliqué en Finance de marché

Méthodologie à suivre : CRISP DM or IBM Master Plan

** Semaine 1: Compréhension du métier

- Durée (1 jour): On demande à l'étudiant de comprendre la méthodologie choisie et effectuer une petite étude bibliographique sur les méthodologies SEMMA, KDP, CRISP DM et IBM Master Plan ⇒ Il doit rendre un document de une ou deux pages à la fin il réalise une comparaison entre les différentes méthodologies afin de conclure que la meilleur est CRISP ou IBM.

- Durée (3 jours): Compréhension propre du métier:
  On demande à l'étudiant de lire et regarder de supports sur la théorie CAPM (il existe plusieur ressources simples et facile sur YouTube). Le but est de lui sensibiliser sur la finance de marché, les risques rencontrés par les investisseurs, les notions de prix d'ouverture, de fermeture, de fermeture ajusté, le volume échangé, les séries temporelles, la granularité, la rentabilité, le risque, la rentabilité annualisée, le risque annualisé, le ratio de Sharpe, le skeweness et le kurtosis.

  L'étudiant doit rendre un document simple selon ce plan:

  ** Prix de fermeture ajusté
  Une description sur un seul paragraphe + formule

  ** Volume échangé:
  Une description sur un seul paragraphe + formule

  ** Rentabilité:
  Un paragraphe + une formule

  Et ainsi de suite ...

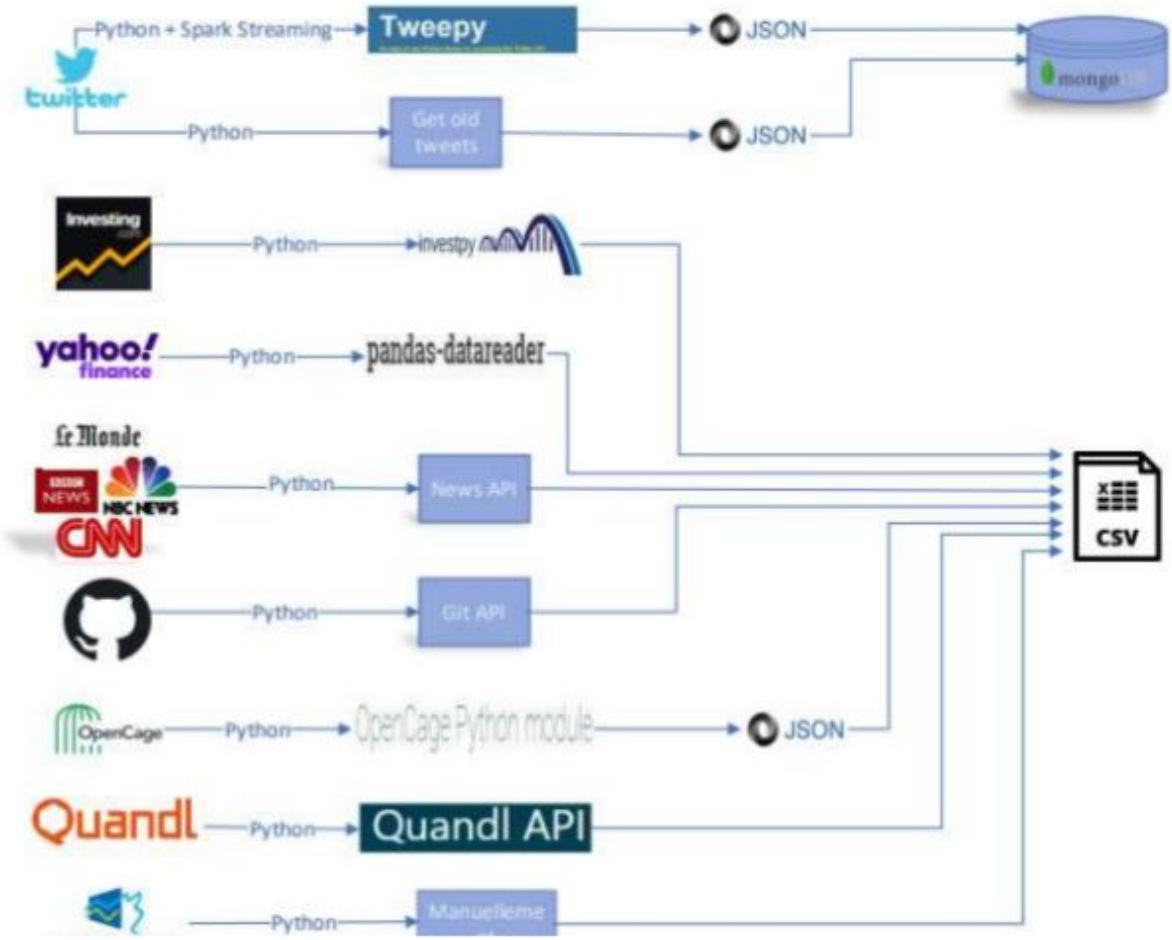
** Semaine 2: Data requirements, Data Collection

- Durée (entre 4 et 6 jours): L'étudiant doit parcourir les sources des données financières: QUANDL, Yahoo Finance, ILBOURSA, INVESTING.COM, XE, BLOOMBERG, REUTERS, OANDA, La banque centrale
- Si le site propose la création d'un compte alors il doit avoir un compte, exemple: quandl demande un compte pour avoir accès aux données, XE aussi, OANDA aussi. investing.com aussi (on doit vérifier qu'il a eu son propre compte sur chaque site financier)
- Collecte des données alternatives à partir des réseaux sociaux et sites d'informations: Twitter ⇒ l'étudiant doit candidater pour avoir un compte sur linkedin

Maintenant, la divergence entre les étudiants peut être dictée par les marchés financiers traités.
Par exemple: On peut imaginer qu'un étudiant peut bosser sur le marché tunisien, un autre sur le marché américain, un autre sur le marché français, un autre sur le marché anglais ou les marchés asiatiques. En effet, chaque marché possède sa propre spécificité et contraintes et connaissances métiers.

Sinon, on peut concentrer le travail sur un seul marché, par exemple le marché tunisien. Autrement dit, les 4 ou 3 vont bosser sur le marché tunisien.

- **Une fois on fixe le marché,**
  L'étudiant commence à collecter des données financières. On peut commencer par les séries temporelles (high price, low price, close price, open price, adjusted close price, volume)



L'idéal peut être sera de suivre ce diagramme pour la collecte des données.

La semaine 2 finira par avoir deux bases de données:

- Une base SQL ou un répertoire contenant des fichiers .csv des données financières numériques (les séries temporelles)
- Une base No SQL pour les données alternatives tweets, articles des journaux ...

** Semaine 3: Data Preparation & Data Understanding

- L'étudiant applique ses connaissances sur la préparation des données dans le cas des séries temporelles.
- Missing values imputation

- Elimination des doublons
- Encodage des variables qualitatives
- Vérification de la granularité: jour, semaine, mois, année
- L'étudiant applique aussi ses connaissances sur la compréhension des données:
  - Matrice de corrélation
  - Chercher les corrélations linéaires ⇒ Pearson
  - Chercher les corrélations non linéaires ⇒ Spearman

  - Bien évidemment pendant le développement informatique: on suit le cycle suivant: Fichier Jupyter notebook ⇒ Transformation des codes vers des fonctions ⇒ Eventuellement transformation des fonctions vers des classes

  Jupyter ⇒ .py ⇒ POO

** Calcul du rendement
A partir de la formule trouvée pendant la première semaine, l'étudiant doit développer sa propre fonction qui calcule le rendement.
Il doit comparer ses résultats avec la fonction implémentée en pandas

** Calcul Volatilité
A partir de la formule trouvée pendant la première semaine, l'étudiant doit développer sa propre fonction qui calcule la volatilité.
Il doit comparer ses résultats avec la fonction implémentée en pandas

** Calcul du rendement annuel
Fonction rendement_annuel

** Calcul de la volatilité annuelle
volatilité_annuelle

** Calcul du Ratio de Sharpe
ratio_sharpe

** Calcul de VaR (Value At Risk): Il existe 4 méthodes différentes permettant le calcul du VaR, l'étudiant doit savoir les 4 méthodes et doit implémenter 4 fonctions pythons

** Calcul de CVaR
Fonction CVaR

** Calcul de Skweness

A partir de la formule trouvée pendant la première semaine, l'étudiant doit développer sa propre fonction qui calcule du skwness.
Il doit comparer ses résultats avec la fonction implémentée en scipy

** Calcul de Kurtosis
A partir de la formule trouvée pendant la première semaine, l'étudiant doit développer sa propre fonction qui calcule le kurtosis.
Il doit comparer ses résultats avec la fonction implémentée en scipy

** Calcul du drawdown
L'étudiant implémente la fonction qui calcule le drawdown

** Calcul du ratio de Sortino
L'étudiant implémente la fonction qui calcule le ratio de Sortino

Il est clair qu'il existe plusieurs autres indicateurs financiers utilisés.
On peut regrouper toutes ces fonctions dans un fichier .py ou dans une classe avec un diagramme de classe bien défini

** Semaine 4&5: Modelling

Modélisation

Apprentissage non supervisé

On suppose que l'étudiant a déjà calculé pour toutes les fonctions présentes dans la semaine 3.

On commence tout simplement avec

	Rendement	Volatilité	Volume	Skewness	Kurtosis	Ratio de Sharpe	VaR
Entreprise 1	XXX	XXXX	XXXX	XX	XXX	XXX	XXX	XXX
Entreprise 2								
Entreprise 3								
.								
.								
.								
Entreprise n								

A partir de ce fichier, on applique les algorithmes connus de segmentation et de réduction des dimensions.

*** Application de l'algorithme ACP avec skleanr
Evaluation
*** Application de l'algorithme ACP avec le module PRINCE en Python
*** CROSS CHECK avec les modules stables de R comme factorMineR et factoMineR Shiny ⇒ l'étudiant sort le rapport à partir du module factoMineRShiny (très performant en analyse factorielle)
Evaluation

*** Application de l'algorithme t-SNE
Evaluation
*** Application de l'algorithme U-MAP
Evaluation
*** Application de l'algorithme Kmeans
Evaluation
*** CROSS CHECK avec les modules stables de R comme factorMineR et factoMineR Shiny ⇒ l'étudiant sort le rapport à partir du module factoMineRShiny (très performant en analyse factorielle)
Evaluation

*** Application de l'algorithme CAH
Evaluation
*** Application de l'algorithme DB SCAN
Evaluation

A la fin de cette partie l'étudiant rédige un petit rapport pour comparer entre les différents algorithmes

** Semaine 6: Retour sur les données alternatives

Il est essentiel de revenir aux données alternatives pour comprendre les résultats donnés par les modèles.

Quand on remarque qu'une entreprise sort un peu de la distribution, on essaye de vérifier la réalité économique, si cette entreprise est citée dans un article ou dans un tweet.

On essaye de détecter les signaux comme Bankrupt, difficultés économiques….