Consignes :   
Cette réflexion peut porter aussi bien sur un aspect technique que sur son environnement.  
o Pour la mener, il faut se demander quelles sont les leçons à tirer de l'expérience vécue. En informatique, il y a des domaines bien balisés, alors que d'autres sont encore en cours d'expérimentation par les entreprises. La réflexion doit donc vous amener à vous former un point de vue sur un sujet pour lequel tout n'a pas encore été dit. Ensuite, ce point de vue devra être confronté avec différentes sources - recherche bibliographique, comparaison avec des expériences similaires - de façon à pouvoir être argumenté.  
o Cette partie doit contenir plus de dix références bibliographiques, et au moins dix citations provenant d’environ dix sources différentes listées en bibliographie. Ne pas respecter ce critère expose à l’invalidation du mémoire.  
o Pour rappel « Une citation directe, reprend une phrase d’un document sans la modifier. Elle permet d’illustrer son propos. Il est obligatoire de la mettre entre guillemets et d’en indiquer la source » [1, p. 6]. La construction d’une bibliographie est décrite pages 14 à 24 de [1]. De telles informations figurent également dans la charte anti-plagiat de l’école.

o Votre objectif doit être de faire profiter de votre expérience les professionnels qui composeront votre jury. Les questions que vous vous êtes posées sont certainement aussi les leurs. Leur intérêt sera de savoir quelles réponses vous (ou l'entreprise qui vous accueille) y donnez et pourquoi.

o Surtout, ne faites pas une réflexion abstraite, elle doit être personnelle et se rattacher à votre vécu. Prenez contact avec votre référent pour valider avec lui le thème choisi et le plan à adopter Faites-le assez tôt - à la moitié de la mission - pour avoir le temps d'effectuer les recherches nécessaires et de laisser mûrir votre réflexion.

Sujets : Calcul de l’impact de la qualité de données sur la performance des modèles de Machine Learning

Action à mener par ordre d’importance :

* **Dimensions à étudier**
  + Completeness
    - Evaluation de différentes valeurs de remplacement pour identifier celles qui conduisent aux meilleures performances du modèle
    - Comparaison des valeurs de remplacement methode d’imputation
  + Target and Feature Accuracy
  + Uniqueness
  + Target Class Balance
  + Uniformity
  + Timeliness : L’impact de la variabilité temporelle des données sur la performance des modèles de machine learning n’a pas été abordé. Par exemple, comment les changements dans les données au fil du temps affectent la robustesse et la performance des modèles formés sur des ensembles de données de qualité variable
* **Etude simultané de plusieurs dimensions de qualité des données** 
  + Examine les dimensions de qualité des données de manière individuelle. Une analyse des effets combinleés de plusieurs dimensions de qualité des données simultanément pourrait offrir des insights supplémentaires sur les interactions complexes entre différentes dimensions de qualité
* **Modèle de ML plus complexe** 
  + Regresssion
    - Gradient Boosting Machine : XGBoost, CatBoost, LightBoost
    - Elastic Net Regression
    - Support Vector Regression
  + Classification
    - Gradient Boosting Machine : XGBoost, CatBoost, LightBoost
    - AdaBoost Classifier
    - Naive Bayes Classifier
  + Clustering
    - DBSCAN
    - HDBSCAN
    - Spectral Clustering
    - Birch (Balanced Iterative Reducing and Clustering using Hierarchies)
  + Deep Learning
    - Deep Neural Networks (DNNs)
    - Recurrent Neural Networks (RNNs) et Long Short-Term Memory Networks (LSTMs)
  + Comparer l'impact de différentes qualités de données sur des réseaux de neurones profonds versus des algorithmes plus classiques.
* **Effets des données sales sur l’Interprétabilité des modèles**
* **Inclure de l’optimisation des Hyper paramètre**
  + Réaliser une évaluation plus détaillée pour chaque modèle d'apprentissage machine, en considérant divers hyperparamètres et configurations
  + Peut fournir des insights sur la manière dont différents modèles traitent les problèmes de qualité des données.
* **Revoir la méthode pour évaluer les méthodes de Clustering** 
  + L’étude a principalement utilisé l’Adjusted Mutual Information (AMI) pour évaluer les algorithmes de clustering. D’autres métriques comme la taille absolue de l’overlap des clusters ou la variance de la taille des clusters pourraient fournir une vue plus complète des performances des algorithmes
* **Techniques Avancées de Réduction de Dimensionalité**
  + Pour les algorithmes de clustering, l’étude a utilisé des réseaux de neurones autoencodeurs basiques sans optimiser pour la tâche spécifique de clustering. Des techniques plus avancées de réduction de dimensionalité ou d’autres architectures de réseaux pourraient potentiellement améliorer les résultats
* **Données plus spécifiques**
  + Exploré l’impact de la qualité des données sur des domaines très spécifiques ou hautement spécialisés. Des études supplémentaires pourraient se concentrer sur des domaines tels que la santé, la finance, ou d’autres industries avec des caractéristiques de données uniques

DataSet :

* <https://archive.ics.uci.edu/dataset/360/air+quality> (Regression)
  + - 10k lignes
    - 15 features
    - Missing Values : Yes
    - Choisir la Target entre plusieurs
  + <https://archive.ics.uci.edu/dataset/849/power+consumption+of+tetouan+city> (Regression)
    - 50k Lignes
    - 6 Feature
    - Trois Target
  + <https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction> (Regression)
    - 30k lignes
    - 28 Features
    - Missing Values : No
* <https://archive.ics.uci.edu/dataset/357/occupancy+detection(Classification)>
  + - 20k lignes
    - 6 features
    - Missing Values : No
    - Target Occupation 0 ou 1
  + <https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset> (Classification)
    - 100k lignes
    - 16 Features
    - Missing Values : No
    - Target : Diabete 0 ou 1
* <https://archive.ics.uci.edu/dataset/352/online+retail> (Clustering)
  + - 500k lignes
    - 8 features
    - Missing Values : No
  + <https://archive.ics.uci.edu/dataset/963/ur3+cobotops>
    - 7k lignes
    - 20 features
  + Intervention Vaines DataSet (Classification, Clustering) DataSet de l’entreprise
    - 1M lignes
    - 30 Features
    - Missing Values : Yes

Pas bon :

* <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraudClustering> (Classification, Clustering)
  + - 300k lignes
    - 10 features
    - Missing Values : No
    - Pas assez d’uniformité
  + <https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption> (Regression, Clustering)
    - 2M de lignes
    - 9 Features
    - Missing Values : Yes
    - Trop de lignes

Régression : 1,000 - 10,000 lignes

Classification : 1,000 - 50,000 lignes

Clustering : 1,000 - 20,000 lignes

Deep Learning (DNN, CNN, RNN, etc.)