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 combinés de plusieurs dimensions de qualité des données simultanément pourrait offrir des insights supplémentaires sur les interactions complexes entre différentes dimensions de qualité
* **Effets des données sales sur l’Interprétabilité des modèles**
* **Modèle de ML plus complexe** 
  + Boosting (Regression et Classification)
    - XGBoost
    - CatBoost
  + Ensemble Methods
    - AdaBoost
    - Stacking
  + Deep Learning
    - CNN (ResNet, DenseNet)
    - RNN (LSTM, GRU)
    - Transformer based model (BERT, GPT)
    - GNN (GCN)
  + Comparer l'impact de différentes qualités de données sur des réseaux de neurones profonds versus des algorithmes plus classiques.
* **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

Analyse supplémentaire :

* Compare l’efficacité de techniques avancées comme les méthodes basées sur l’apprentissage automatique pour le nettoyage des données

DataSet :

* Regression
  + Boston Housing DataSet
  + California Housing Prices DataSet
  + Air Quality DataSet
  + Energy Efficiency DataSet
  + Auto MGP DataSet
  + Bike Sharing Demand
  + House Prices : Advanced Regression Techniques
* Classification
  + Iris DataSet
  + Titanic DataSet
  + Breast Cancer Wisconsin DataSet
  + Wine Quality DataSet
  + MNIST Handwritten Digits DataSet
  + Credit Card Fraud Detection DataSet
  + Titanic: Machine Learning from Disaster
  + Heart Disis DataSet (UCI)
  + Adult Income DataSet (UCI)
* Clustering
  + Mall Customers DataSet
  + Wholesale Customers DataSet
  + Handwritten Digits DataSet
  + Customer Personality Analysis Dataset
  + Human Activity Recognition (HAR) Using Smartphones DataSet
  + Customer Segmentation DataSet
* Intervention Vaines DataSet for Classification and Clustering