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
    - Pollution synthetique si pas de valeurs manquante : valeurs manquantes sont introduites dans un certain pourcentage des échantillons.
    - Évaluation 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
    - Pollution : Ajouté du bruit aux valeurs des caractéristiques numérique selon une distribution normale, avec une variance proportionnelle à la moyenne des valeurs originales et en permutante les valeurs pour les catégorique
  + Uniqueness
    - Pollution : Ajouter des lignes en double de manière aléatoire.
  + Target Class Balance
    - Pollution : Introduire un déséquilibre
  + Uniformité
    - Tester le modele avec un DataSet normaliser et un DataSet non normaliser et comparer la fiabilité du modele avant et apres.
  + 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
    - Tous les dataset comportes des series temporelles. Nous allons créer different version des dataset décaler de une heure, jour, semaine ou mois et comparer. Pour la regression et la classification, l’idee sera d’entrainer le modele sur differentes version et de regarder la fiabiliter que obtient le modele sur les autres versions. Pour le clusteering, l’idée sera d’appliquer le clustrer sur differentes versions de données et d’analyser les cluster ainsi former.
* **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
    - Gaussian Process Regression
  + Classification
    - Gradient Boosting Machine : XGBoost, CatBoost, LightBoost
    - TabNeT
    - DeepFM (Deep Factorization Machines)
    - Node
    - AdaBoost Classifier
    - Naive Bayes Classifier
  + Clustering
    - HDBSCAN
    - Birch (Balanced Iterative Reducing and Clustering using Hierarchies)
    - AutoML-based Clustering Approaches (e.g., AutoCluster)
    - Deep Embedded Clustering (DEC)
  + Deep Learning
    - DNN avec Neural Architecture Search (NAS)
  + 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 Choisie pour les experiences :

* <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