Taille previligier pour les DataSets : 20 et 100 features

Régression : 100 000 – 1 000 000 lignes

<https://archive.ics.uci.edu/dataset/203/yearpredictionmsd> 500k x 90

# fps-in-video-games 400k x 45

Classification : 500 000 – 5 000 000 lignes

<https://archive.ics.uci.edu/dataset/31/covertype> 500k x 54

[https://www.kaggle.com/datasets/ivansher/nasa-nearest-earth-objects-1910-2024 300k x 9](https://www.kaggle.com/datasets/ivansher/nasa-nearest-earth-objects-1910-2024%20300k%20x%209)

<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease> 300k x 18

Clustering : 100 000 – 1 000 000 lignes

[https://www.kaggle.com/datasets/bhaccavikjikadara/retail-transactional-dataset](https://www.kaggle.com/datasets/bhavikjikadara/retail-transactional-dataset) 300k x 30

* <https://archive.ics.uci.edu/dataset/352/online+retail> 500k x 8

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

DataSet Choisie:

* <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

DataSert non choisie:

* <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

Plan pour la revue de littérature

Sujet : Impact de la qualité des données sur la performance des modèles de machine learning

1. Introduction

Objectif de la revue de littérature

* Définir clairement les objectifs de cette revue.
* Expliquer l'importance de la qualité des données dans le contexte des modèles de machine learning.
* Mettre en perspective l'importance de cette revue pour le mémoire de fin d'étude.

2. Présentation des travaux antérieurs

Synthèse des études récentes et classiques sur la qualité des données

* Révision des principales études ayant abordé la qualité des données.
* Comparaison des approches classiques et contemporaines.

Évolution des recherches dans ce domaine

* Traçage de l'évolution des concepts et méthodes depuis les premières études jusqu'aux travaux les plus récents.

Présentation des théories et modèles utilisés pour évaluer et améliorer la qualité des données

* Discussion des principaux modèles théoriques.
* Description des frameworks et standards pour l'évaluation de la qualité des données.

3. Problématique et lacunes identifiées

Lacunes dans les données et les études existantes

* Analyse critique des limites des recherches antérieures.
* Identification des défis persistants liés à la qualité des données.

Problèmes non résolus concernant la qualité des données et leur impact sur les performances des modèles de Machine Learning

* Discussion des questions de recherche encore en suspens.
* Mise en évidence des zones grises dans les connaissances actuelles.

Questions de recherche non résolues

* Identification des questions de recherche pertinentes non abordées.
* Justification de la nécessité de votre recherche.

4. Impact de la Qualité des Données sur les Modèles de Machine Learning

Études empiriques et résultats

* Résumé des études empiriques (références aux documents PDF fournis).
* Synthèse des résultats et des conclusions principales.

Études de cas spécifiques

* Présentation d'études de cas illustrant l'impact de la qualité des données sur les modèles ML.
* Discussion des implications pratiques et théoriques des études de cas.

5. Cadre de référence et méthodologie

Méthodologie utilisée dans les travaux antérieurs

* Description détaillée des méthodes et approches pour évaluer et améliorer la qualité des données.
* Comparaison critique des différentes approches méthodologiques.

Cadre théorique pour l’étude proposée

* Élaboration d'un cadre théorique basé sur les travaux antérieurs.
* Justification des méthodes et approches choisies pour votre recherche.
* Explication de l'adaptation ou l'extension des cadres théoriques existants pour répondre aux nouvelles questions de recherche.

**Sessions, V. et Valtorta, M. (2006). The Effects of Data Quality on Machine Learning Algorithms**

L'étude menée par Sessions et Valtorta (2006) vise à démontrer l'importance de la qualité des données pour les algorithmes de machine learning, en particulier les réseaux bayésiens (BNs), et à proposer des méthodes pour intégrer l'évaluation de la qualité des données dans ces algorithmes.

##### Objectifs principaux

1. **Démontrer l'importance de la qualité des données** : L'étude montre que la qualité des données est cruciale pour la performance des algorithmes de machine learning, une idée largement reconnue dans le milieu des affaires mais encore négligée dans la communauté du machine learning.
2. **Développer et tester des méthodes pour intégrer la qualité des données** : Sessions et Valtorta ont conçu et testé des méthodes pour intégrer des évaluations de la qualité des données dans les algorithmes, créant ainsi des modèles plus robustes et utiles.

##### Méthodologie

* **Focus sur l'exactitude des données** : L'étude se concentre sur l'impact de l'exactitude des données sur l'algorithme PC pour les BNs, en raison de la facilité de manipulation de cette dimension de la qualité dans les ensembles de données de test.
* **Expérimentations** : Des ensembles de données inexactes ont été créés en combinant des données "vraies" avec des données "fausses" à divers pourcentages (10%, 25%, 50%, 75%, et 90% d'inexactitude).
* **Algorithme PC** : L'algorithme PC a été utilisé pour apprendre la structure des BNs à partir des ensembles de données de test, et les résultats ont été évalués en fonction du nombre d'arêtes correctes, incorrectes, mal orientées, positives fausses et négatives fausses.

##### Résultats

* **Impact des données inexactes** : L'algorithme PC s'est avéré très sensible aux données inexactes, avec une dégradation rapide de la performance au-delà de 10% d'inexactitude des données.
* **Méthodes proposées** : Quatre méthodes ont été développées pour améliorer l'algorithme en intégrant des évaluations de la qualité des données :
  1. **Do Nothing** : Utilise les données telles quelles.
  2. **Threshold** : Filtre les ensembles de données en dessous d'un certain seuil de qualité.
  3. **DQ Method Part One et Two** : Attribue des poids de qualité aux ensembles de données.
  4. **DQ Method Part Three** : Utilise différents niveaux de signification pour déterminer l'indépendance des nœuds, ajustant ainsi la tolérance de l'algorithme aux données de faible qualité.

##### Contributions et limitations

* **Contributions** : L'étude a démontré empiriquement l'impact significatif de la qualité des données sur les performances des algorithmes de machine learning et proposé des méthodes pour améliorer ces performances en intégrant des évaluations de la qualité des données.
* **Limitations** : La recherche est limitée à l'exactitude des données et à l'algorithme PC pour les BNs. Les futures recherches devraient explorer d'autres dimensions de la qualité des données et d'autres algorithmes de machine learning.

En résumé, l'étude de Sessions et Valtorta (2006) souligne l'importance cruciale de la qualité des données pour les algorithmes de machine learning et propose des méthodes pratiques pour intégrer cette qualité dans les processus de modélisation, ouvrant ainsi la voie à des améliorations significatives dans la robustesse et l'utilité des modèles de machine learning.

**Gudivada, V. N., Apon, A., et Ding, J. (2017). Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations**

L'étude de Gudivada et al. (2017) examine les problèmes de qualité des données dans le contexte des big data et du machine learning, dépassant les simples opérations de nettoyage et de transformation des données. Les auteurs discutent des différents aspects de la qualité des données et proposent un cadre de gouvernance des données pour gérer le cycle de vie de la qualité des données dans ce nouveau contexte.

Objectifs principaux :

1. **Nature des problèmes de qualité des données** : Identifier les problèmes spécifiques de qualité des données dans les environnements de big data et de machine learning.
2. **Cadre de gouvernance des données** : Présenter un cadre de gouvernance des données pour gérer la qualité des données tout au long de leur cycle de vie, incluant l'acquisition, le nettoyage, la transformation et l'intégration des données.

Méthodologie :

* **Examen des dimensions de la qualité des données** : Les auteurs définissent plusieurs dimensions de la qualité des données, telles que l'exactitude, la complétude, la cohérence et la provenance.
* **Framework et outils** : Développement d'un cadre de gouvernance des données pour évaluer et améliorer la qualité des données, et examen des outils disponibles pour la gestion de la qualité des données.

Résultats :

* **Problèmes de qualité des données en big data** : La qualité des données dans les environnements de big data est souvent compromise par des problèmes de données manquantes, de duplication, d'hétérogénéité et de streaming.
* **Méthodes de gestion de la qualité** : Les auteurs proposent des méthodes pour évaluer et améliorer la qualité des données, incluant des techniques de validation des données, d'imputation des données manquantes et de détection des duplications.

Contributions et limitations :

* **Contributions :** L'étude fournit un cadre structuré pour aborder les problèmes de qualité des données dans le contexte des big data et du machine learning, et propose des solutions pratiques pour améliorer la qualité des données.
* **Limitations :** Bien que l'étude couvre de nombreux aspects de la qualité des données, elle est principalement théorique et nécessite des validations empiriques supplémentaires pour confirmer l'efficacité des méthodes proposées.

En résumé, cette étude souligne les défis spécifiques de la qualité des données dans les environnements de big data et de machine learning, et propose un cadre de gouvernance des données pour améliorer la gestion de la qualité des données.

**Qi, Z., Wang, H., et Wang, A. (2021). Impacts of Dirty Data on Classification and Clustering Models: An Experimental Evaluation**

L'étude de Qi et al. (2021) examine les impacts des données de mauvaise qualité sur les modèles de classification et de clustering. Les auteurs ont réalisé une évaluation expérimentale pour analyser comment les données manquantes, incohérentes et conflictuelles affectent les performances des modèles.

Objectifs principaux :

1. **Impact des données de mauvaise qualité :** Étudier comment différents types d'erreurs (manquantes, incohérentes, conflictuelles) affectent les performances des modèles de classification et de clustering.
2. **Proposer des métriques :** Introduire deux nouvelles métriques, la sensibilité et le point d'inflexion de la qualité des données (DQIP), pour évaluer la tolérance des modèles aux données de mauvaise qualité.

Méthodologie :

* **Données générées :** Utilisation de neuf ensembles de données typiques avec des taux d'erreurs variés.
* **Évaluation des modèles :** Douze modèles classiques de machine learning ont été évalués en termes de sensibilité et de tolérance aux données de mauvaise qualité.

Résultats :

* **Impacts variables :** Les impacts des données de mauvaise qualité varient selon le type d'erreur et le modèle utilisé. Les données manquantes ont tendance à avoir des effets plus prononcés que les données incohérentes ou conflictuelles.
* **Modèles tolérants :** Certains modèles, comme KNN et DBSCAN, montrent une meilleure tolérance aux erreurs comparés à d'autres comme les réseaux bayésiens et la régression logistique.
* **Métriques proposées :** Les métriques de sensibilité et de DQIP se sont avérées efficaces pour mesurer et comparer la robustesse des modèles face aux données de mauvaise qualité.

Limitations :

**Portée Limitée des Types d'Erreurs** : L'étude se concentre principalement sur trois types de données sales : les données manquantes, incohérentes et conflictuelles. D'autres types de problèmes de qualité des données, tels que les données dupliquées, les valeurs aberrantes et les données incorrectes, ne sont pas explorés.

**Dimensions de Qualité des Données Restreintes** : L'analyse est limitée à des dimensions spécifiques de la qualité des données comme l'intégralité, la cohérence et l'identité des entités. Des dimensions plus larges telles que l'exactitude, la ponctualité et la pertinence ne sont pas prises en compte.

**Diversité des Jeux de Données** : L'évaluation utilise neuf jeux de données provenant de la UCI Machine Learning Repository, qui, bien que diversifiés, pourraient ne pas représenter de manière exhaustive la variabilité des données du monde réel. Utiliser plus de jeux de données provenant de différents domaines pourrait améliorer la généralisation des résultats.

**Variété des Modèles** : L'étude évalue 12 modèles classiques. Bien que ces modèles soient largement utilisés, l'inclusion de modèles plus contemporains, tels que les algorithmes d'apprentissage profond et les méthodes d'ensemble, pourrait fournir une compréhension plus complète des impacts des données sales.

**Limitations de la Configuration Expérimentale** : Les expériences sont menées avec des paramètres spécifiques et pourraient ne pas refléter l'ensemble du comportement des modèles sous différentes configurations. De plus, seuls les résultats moyens sont rapportés, ce qui pourrait masquer la variabilité des performances.

**Metrics d'Évaluation** : L'étude introduit des nouvelles metrics telles que la sensibilité et le point d'inflexion de la qualité des données (DQIP). Cependant, ces metrics pourraient ne pas être universellement acceptées ou comprises dans la communauté de l'apprentissage machine. Une validation contre des metrics plus établies pourrait renforcer les conclusions.

**Hypothèses dans l'Imputation des Données** : Pour les données manquantes, l'étude utilise les valeurs moyennes pour les attributs numériques et les valeurs maximales pour les attributs catégoriels. Cela pourrait ne pas toujours être la meilleure approche et pourrait biaiser les résultats. Des méthodes d'imputation plus sophistiquées pourraient être explorées.

**Concentration sur les Jeux de Données Statistiques** : L'étude ne considère pas l'impact de la qualité des données dans des environnements dynamiques où les données sont continuellement mises à jour. Les effets des données sales dans les contextes de données en streaming ne sont pas abordés.

Axes d'Amélioration :

**Inclusion de Plus de Problèmes de Qualité des Données** : Les futures études devraient explorer les impacts d'autres types de données sales, comme les doublons, les valeurs aberrantes et les données incorrectes, pour fournir une analyse plus complète.

**Metrics d'Évaluation Plus Larges** : Incorporer des metrics d'évaluation supplémentaires, telles que les mesures de robustesse et l'efficacité computationnelle, pour fournir une vue plus holistique des performances des modèles en présence de données sales.

**Augmentation de la Diversité des Jeux de Données** : Utiliser un plus large éventail de jeux de données, y compris ceux provenant de différents domaines et avec des niveaux variés de bruit et de problèmes de qualité intrinsèques, pour améliorer la généralisation des résultats.

**Évaluation de Plus de Modèles** : Inclure des modèles d'apprentissage machine plus modernes, tels que les cadres d'apprentissage profond, les machines à gradient boosting et des méthodes d'ensemble plus complexes, pour comprendre leur résilience aux données sales.

**Analyse de Sensibilité des Paramètres** : Mener une analyse de sensibilité détaillée des paramètres des modèles pour comprendre comment différentes configurations affectent les performances des modèles en présence de données sales.

**Techniques d'Imputation de Données Avancées** : Expérimenter avec des méthodes d'imputation plus avancées, telles que l'imputation par les k-plus proches voisins (KNN), l'imputation multiple et l'imputation basée sur des modèles, pour évaluer leur efficacité à atténuer les impacts des données manquantes.

**Contextes de Données Dynamiques** : Étendre l'analyse aux environnements de données dynamiques, tels que les flux de données, pour comprendre comment les données sales impactent les modèles en applications temps réel.

**Validation des Nouvelles Metrics** : Valider les metrics proposées (sensibilité et DQIP) par rapport aux metrics établies et dans différents contextes pour assurer leur robustesse et leur applicabilité.

**Srivastava, S., Shah, R. N., et Teodoriu, C. (2022). Impact of Data Quality on Supervised Machine Learning: Case Study on Drilling**

L'étude de Srivastava et al. (2022) explore l'impact de la qualité des données sur les modèles de machine learning supervisés dans le contexte des vibrations de forage. Elle se concentre sur des aspects tels que la fréquence d'échantillonnage, l'étiquetage des données, l'extraction des caractéristiques et le déséquilibre des classes.

Objectifs principaux :

1. **Examiner l'impact de la qualité des données :** Analyser comment la fréquence d'échantillonnage, l'étiquetage et l'extraction des caractéristiques affectent les performances des modèles de classification.
2. **Proposer des solutions :** Développer un workflow pour choisir les bons paramètres et techniques à chaque étape, de la collecte des données à l'entraînement des modèles.

Méthodologie :

* **Données expérimentales :** Utilisation de données de vibrations de forage collectées à différentes fréquences (1 Hz, 10 Hz, 100 Hz) pour évaluer l'impact de la fréquence d'échantillonnage.
* **Modèles de classification :** Évaluation de six modèles de machine learning (SVM, LDA, Logistic Regression, Naïve Bayes, CART, KNN) sur des données étiquetées manuellement et automatiquement.

Résultats :

* **Fréquence d'échantillonnage :** Des fréquences d'échantillonnage plus élevées (100 Hz) améliorent la précision des modèles, avec une variabilité moindre des résultats.
* **Étiquetage des données :** L'étiquetage manuel, bien que coûteux, est plus précis que l'étiquetage automatique basé sur des métriques empiriques.
* **Extraction des caractéristiques :** L'utilisation de caractéristiques statistiques et spectrales améliore significativement la performance des modèles comparé à l'utilisation de données brutes.
* **Déséquilibre des classes :** Les données déséquilibrées affectent les performances, mais des techniques comme l'échantillonnage et la pondération des classes peuvent atténuer ces effets.

Contributions et limitations :

* **Contributions :** L'étude fournit des recommandations pratiques pour améliorer la qualité des données et les performances des modèles de machine learning dans le domaine du forage.
* **Limitations :** Les conclusions sont spécifiques aux données de forage et peuvent nécessiter des validations supplémentaires dans d'autres contextes.

En résumé, cette étude souligne l'importance de divers aspects de la qualité des données sur les performances des modèles de machine learning et propose des solutions pour optimiser ces performances dans le contexte des vibrations de forage.

**Luca, A. R., Ursuleanu, T. F., et al. (2022). Impact of Quality, Type, and Volume of Data Used by Deep Learning Models in the Analysis of Medical Images**

L'étude de Luca et al. (2022) explore l'impact de la qualité, du type et du volume des données sur les performances des modèles de deep learning dans l'analyse des images médicales. Les auteurs mettent en évidence que l'accès à des bases de données médicales volumineuses et bien annotées peut améliorer les performances des modèles de deep learning, mais des résultats similaires ou supérieurs peuvent être obtenus avec de petites bases de données bien catégorisées et annotées.

Objectifs principaux :

1. **Analyser l'impact de la qualité, du type et du volume des données** : Évaluer comment ces facteurs influencent les performances des modèles de deep learning en imagerie médicale.
2. **Décrire les composantes du processus de deep learning** : Caractériser les différentes étapes, de la collecte des données aux applications médicales, et leurs corrélations spécifiques.

Méthodologie :

* **Types de données** : Les auteurs examinent différents types de données médicales, notamment les images CT, MRI, et PET, et soulignent l'importance de la labellisation et de l'annotation des données.
* **Cadre de gouvernance des données** : Proposition d'un cadre pour gérer la qualité des données tout au long de leur cycle de vie.

Résultats :

* **Qualité des données** : Les données bien annotées et organisées améliorent significativement les performances des modèles.
* **Volume des données** : Les grandes bases de données bien annotées permettent des performances optimales, mais de petites bases de données bien gérées peuvent également donner d'excellents résultats.
* **Types de données** : L'utilisation de données multimodales et bien structurées augmente la robustesse et l'efficacité des modèles.

Contributions et limitations :

* **Contributions :** L'étude offre une vue d'ensemble des facteurs influençant les performances des modèles de deep learning en imagerie médicale et propose des solutions pratiques pour améliorer la qualité des données.
* **Limitations :** Les conclusions sont principalement théoriques et nécessitent des validations empiriques supplémentaires.

En résumé, cette étude souligne l'importance de la qualité, du type et du volume des données sur les performances des modèles de deep learning dans l'analyse des images médicales et propose un cadre pour gérer et optimiser ces facteurs.

**Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., et Harmouch, H. (2022). The Effects of Data Quality on Machine Learning Performance**

L'étude de Budach et al. (2022) examine l'impact de six dimensions de la qualité des données sur les performances de quinze algorithmes de machine learning couvrant les tâches de classification, régression et clustering. Les dimensions étudiées sont la représentation cohérente, la complétude, la précision des caractéristiques, la précision des cibles, l'unicité et l'équilibre des classes.

Objectifs principaux :

1. **Relation entre la qualité des données et les performances des modèles** : Analyser empiriquement comment différentes dimensions de la qualité des données affectent les performances des algorithmes de machine learning.
2. **Scénarios de pollution des données** : Étudier les impacts de la pollution des données dans différents scénarios (données de formation polluées, données de test polluées, ou les deux).

Méthodologie :

* **Expérimentation systématique** : Pollution des données pour simuler des erreurs dans les six dimensions de qualité et évaluation de la performance de modèles de machine learning sur ces données.
* **Algorithmes testés** : Quinze algorithmes, incluant des modèles simples (arbres de décision) et complexes (réseaux de neurones), appliqués à des tâches de classification, régression et clustering.

Résultats :

* **Impact de la qualité des données** : La qualité des données affecte significativement les performances des modèles, avec des impacts variés selon le type d'erreur et l'algorithme utilisé.
* **Robustesse des modèles** : Certains modèles, comme les réseaux de neurones, montrent une meilleure tolérance aux erreurs que d'autres, comme les arbres de décision.
* **Scénarios de pollution** : La performance des modèles diminue le plus lorsque les données de test sont polluées, bien que la pollution des données de formation ait également un impact notable.

Limitations :

**Absence d'optimisation des hyperparamètres** : L'étude n'a pas inclus l'optimisation des hyperparamètres pour les algorithmes testés. Cela signifie que les performances rapportées peuvent ne pas être les meilleures possibles pour chaque algorithme testé​((2022) The Effects of D…)​.

**Valeurs de remplacement fixes pour la pollution de complétude** : Pour simuler l'absence de données, des valeurs de remplacement spécifiques ont été manuellement définies pour chaque caractéristique, ce qui pourrait influencer négativement les performances des modèles. Explorer différentes valeurs de remplacement pourrait être une piste pour des travaux futurs​((2022) The Effects of D…)​.

**Autoencodeur non optimisé** : L'implémentation de l'autoencodeur utilise un réseau de neurones basique, sans optimisation spécifique pour sa tâche. Des améliorations comme l'utilisation de composants de réseau autres que des couches linéaires ou l'incorporation de la performance de clustering dans la fonction de perte pourraient être envisagées​((2022) The Effects of D…)​.

**Perte d'information lors de l'encodage des données à haute dimension** : Lors de l'évaluation des algorithmes de clustering, la perte d'information due à l'encodage des données à haute dimension en un espace à deux dimensions n'a pas été prise en compte​((2022) The Effects of D…)​.

**Évaluation limitée à l'AMI (Adjusted Mutual Information)** : Seule la mesure AMI a été utilisée pour évaluer les algorithmes de clustering. D'autres métriques comme la taille absolue du chevauchement des clusters d'origine et générés, ainsi que la moyenne et la variance des tailles des clusters dans les résultats des algorithmes, pourraient également être considérées pour une évaluation plus complète​((2022) The Effects of D…)​.

Axes d'amélioration :

**Optimisation des hyperparamètres** : Intégrer une optimisation systématique des hyperparamètres pour chaque algorithme afin de maximiser leurs performances et fournir des résultats plus précis​((2022) The Effects of D…)​.

**Exploration de différentes valeurs de remplacement** : Tester différentes valeurs de remplacement pour les données manquantes afin de déterminer lesquelles minimisent le plus l'impact négatif sur les performances des modèles. Cela pourrait également inclure l'étude de diverses méthodes d'imputation​((2022) The Effects of D…)​.

**Amélioration des architectures d'autoencodeurs** : Utiliser des architectures de réseaux de neurones plus sophistiquées pour les autoencodeurs, comme des couches convolutives ou récurrentes, et ajuster la fonction de perte pour inclure la performance de clustering​((2022) The Effects of D…)​.

**Évaluation multidimensionnelle des algorithmes de clustering** : Élargir l'évaluation des algorithmes de clustering en utilisant une variété de métriques, telles que la taille du chevauchement des clusters et la distribution des tailles des clusters, pour obtenir une vue plus complète de leur performance​((2022) The Effects of D…)​.

**Analyse approfondie par modèle de ML** : Conduire une évaluation plus approfondie et détaillée de chaque modèle de machine learning pour comprendre comment chacun réagit aux différentes dimensions de qualité des données​((2022) The Effects of D…)​.

**Étude de nouvelles dimensions de qualité des données** : Ajouter de nouvelles dimensions de qualité des données dans les analyses futures pour couvrir un éventail plus large de scénarios et de défis liés à la qualité des données​((2022) The Effects of D…)​.

Théorie et modèles utilisés :

1. Dimensions de qualité des données

- Représentation cohérente : Assure que chaque entité ou concept du monde réel est représenté de manière uniforme dans le jeu de données. Par exemple, une ville ne devrait pas être représentée par plusieurs valeurs différentes comme "New York", "NYC", et "NY".

- Complétude : Se réfère à l'absence de valeurs manquantes dans le jeu de données.

- Exactitude des caractéristiques : Mesure la précision des valeurs des caractéristiques par rapport à leur valeur réelle.

- Exactitude des cibles : Concernant la précision des étiquettes de classe dans un ensemble de données de classification.

- Unicité : Garantit qu'il n'y a pas de duplicata exact dans les données.

- Équilibre des classes cibles : Assure une distribution équilibrée des classes dans les jeux de données de classification .

2. Méthodes de pollution des données

- Pollution par représentation incohérente : Ajouter des représentations alternatives pour les valeurs de caractéristiques catégorielles.

- Pollution par incompletude : Introduire des valeurs manquantes de manière systématique dans les données.

- Pollution par exactitude des caractéristiques : Modifier les valeurs des caractéristiques en ajoutant du bruit.

- Pollution par exactitude des cibles : Changer aléatoirement les étiquettes des classes .

4. Approches spécifiques à certaines tâches

- Classification : Utilisation du score F1 pour évaluer la performance des modèles de classification, en tenant compte de l'impact des données déséquilibrées.

- Régression : Utilisation du coefficient de détermination (R²) pour mesurer la fraction de variance expliquée par le modèle.

- Clustering : Utilisation de l'information mutuelle ajustée (AMI) pour évaluer la qualité des clusters produits par les algorithmes de clustering .

3. Évaluation Empirique de la Qualité des Données

• Analyse des Scénarios Réalistes : Les auteurs ont analysé trois scénarios différents basés sur les étapes du pipeline AI alimentées par des données polluées : données d’entraînement polluées, données de test polluées, ou les deux.

• Expérimentation avec des Algorithmes Variés : 15 algorithmes de machine learning, couvrant les tâches de classification, de régression et de clustering, ont été testés pour observer comment les différentes dimensions de qualité des données affectent leurs performances .

**Azimi, S. et Pahl, C. (2022). The Impact of Data Completeness and Correctness on Explainable Machine Learning Models**

L'étude de Azimi et Pahl (2022) explore l'impact de la complétude et de la correction des données sur les modèles de machine learning explicables, en utilisant des données de l'Internet des Objets (IoT) et des arbres de décision.

Objectifs principaux :

1. **Impact des erreurs de données :** Analyser comment la complétude et la correction des données influencent l'exactitude, la précision et le rappel des modèles.
2. **Expérimentations avec des données défectueuses :** Simuler différents niveaux de données manquantes et incorrectes pour observer leurs effets sur les modèles.

Méthodologie :

* **Données expérimentales :** Utilisation de données de trafic et météorologiques avec différentes proportions de lignes et de caractéristiques manquantes ou incorrectes.
* **Évaluation des modèles :** Mesure des impacts sur l'exactitude, la précision et le rappel des modèles de décision.

Résultats :

* **Complétude des données :** La perte de lignes a un impact plus important sur les performances des modèles que la perte de caractéristiques.
* **Correction des données :** Les données incorrectes dégradent davantage la performance des modèles que les données manquantes.
* **Patterns observés :** Les modèles montrent des variations spécifiques en termes d'exactitude, de précision et de rappel selon les types et niveaux de défauts des données.

Limitations :

**Nombre Limité de Jeux de Données** : L'étude utilise deux jeux de données spécifiques (trafic et météo) pour évaluer l'impact de la complétude et de la correction des données sur les modèles de machine learning explicables. Cette approche limite la généralisation des résultats à d'autres types de données ou domaines d'application​((2022) The Impact of Da…)​.

**Modèle de Machine Learning Unique** : L'étude se concentre exclusivement sur les arbres de décision comme méthode de machine learning. Bien que les arbres de décision soient utiles pour l'interprétabilité, d'autres modèles comme les réseaux neuronaux ou les modèles basés sur des forêts d'arbres pourraient présenter des comportements différents face à des problèmes de qualité des données​((2022) The Impact of Da…)​.

**Quantité de Données** : Le volume de données utilisé pour les expériences est relativement faible, ce qui pourrait affecter la robustesse des conclusions tirées. Les petits ensembles de données peuvent également amplifier les effets du surapprentissage ou des erreurs spécifiques aux données​((2022) The Impact of Da…)​.

**Nature des Données** : L'étude se limite à des données numériques et ne prend pas en compte les données textuelles ou d'image, ce qui restreint la portée des conclusions aux seules données numériques. Les erreurs et les techniques de gestion des erreurs peuvent varier considérablement entre les différents types de données​((2022) The Impact of Da…)​.

**Analyse Post-Hoc** : La méthode utilisée pour expliquer les modèles est basée sur une analyse post-hoc, ce qui signifie qu'elle examine les résultats après la construction du modèle. Cela peut limiter la capacité à prévenir les erreurs avant qu'elles n'affectent les performances du modèle​((2022) The Impact of Da…)​.

Axes d'Amélioration :

**Augmentation de la Diversité des Jeux de Données** : Pour améliorer la généralisation des résultats, il serait bénéfique d'inclure une variété plus large de jeux de données provenant de différents domaines et types de données (numériques, textuels, images). Cela permettrait d'évaluer si les conclusions tirées sont applicables à un éventail plus large de situations​((2022) The Impact of Da…)​.

**Exploration de Différents Modèles de Machine Learning** : Tester d'autres modèles de machine learning comme les réseaux de neurones, les modèles de forêts d'arbres, et les SVM (Support Vector Machines) pourrait fournir des insights complémentaires sur l'impact de la qualité des données. Certains modèles peuvent être plus résilients aux erreurs de données que d'autres​((2022) The Impact of Da…)​.

**Augmentation du Volume de Données** : Utiliser des ensembles de données plus volumineux pourrait aider à valider les résultats et rendre les conclusions plus robustes. Des volumes de données plus importants peuvent également mieux représenter les scénarios du monde réel où les données sont souvent massives et variées​((2022) The Impact of Da…)​.

**Inclusion d'Analyses Pré-Hoc** : Développer des techniques pour identifier et corriger les erreurs de données avant la phase de construction du modèle pourrait améliorer la qualité des modèles de machine learning dès le départ. Des outils de pré-traitement et de validation des données automatisés pourraient être explorés​((2022) The Impact of Da…)​.

**Étude des Données Textuelles et d'Images** : Intégrer des analyses sur des jeux de données textuelles et d'images permettrait de couvrir un spectre plus large de types de données, augmentant ainsi la portée et l'applicabilité des résultats de l'étude​((2022) The Impact of Da…)​.

**Évaluation de la Robustesse des Modèles** : Tester la robustesse des modèles face à différentes formes d'erreurs et de corruptions de données pourrait fournir des insights plus approfondis sur la résilience des modèles de machine learning aux problèmes de qualité des données. Cela inclut l'évaluation de l'impact des erreurs systématiques versus aléatoires​((2022) The Impact of Da…)​.

**Gong, Y., Liu, G., Xue, Y., Li, R., et Meng, L. (2023). A Survey on Dataset Quality in Machine Learning**

L'étude de Gong et al. (2023) explore l'importance de la qualité des ensembles de données pour les performances des modèles de machine learning. Les auteurs passent en revue les concepts clés, les problèmes de qualité et les risques associés, ainsi que les dimensions et les métriques de la qualité des données tout au long de leur cycle de vie.

Objectifs principaux :

1. **Définir la qualité des ensembles de données :** Analyser les différentes dimensions et métriques pour évaluer la qualité des données.
2. **Cadre d'évaluation :** Proposer un cadre complet d'évaluation de la qualité des données avec huit dimensions et 32 métriques.

Méthodologie :

* **Revue de la littérature :** Synthèse des travaux existants sur la qualité des ensembles de données et identification des dimensions critiques telles que la complétude, l'exactitude, la cohérence et la ponctualité.
* **Cadre d'évaluation :** Développement d'un cadre intégrant des métriques telles que la conformité du format des données, l'unicité des données et la réduction des biais.

Résultats :

* **Dimensions de qualité :** Identification de huit dimensions de qualité, notamment la complétude, l'exactitude, la standardisation et l'absence de biais.
* **Métriques spécifiques :** Introduction de 32 métriques pour évaluer chaque dimension, comme la complétude des éléments de données et la conformité au format des données.

Limitations de l'Étude

**Manque d'uniformité dans les métriques d'évaluation** : La construction des ensembles de données repose sur les normes respectives de chaque agence, rendant difficile l'obtention d'une description unifiée de l'assurance qualité de la construction des ensembles de données​((2023) A survey on data…)​.

**Dimensions de qualité des ensembles de données non exhaustives** : Actuellement, certaines dimensions de qualité des ensembles de données sont encore analysées sur la base de la qualité des bases de données traditionnelles. De nouvelles dimensions de qualité doivent être évaluées en fonction des caractéristiques spécifiques au domaine de l'apprentissage automatique​((2023) A survey on data…)​.

**Besoin de recherches supplémentaires sur les métriques d'évaluation de la qualité des ensembles de données** : Il est nécessaire de poursuivre les recherches pour approfondir la mesure des métriques d'évaluation de la qualité des ensembles de données​((2023) A survey on data…)​.

Axes d'Amélioration

**Clarification des normes de qualité des ensembles de données** : Les recherches futures peuvent se concentrer sur la clarification des normes de qualité des ensembles de données pendant leur construction afin d'atteindre une plus grande uniformité dans la construction et l'utilisation des ensembles de données​((2023) A survey on data…)​.

**Ajout de nouvelles dimensions de qualité** : En se concentrant sur les caractéristiques des tâches dans le domaine de l'apprentissage automatique, de nouvelles dimensions de qualité peuvent être ajoutées. Par exemple, en tenant compte des biais de sélection, des biais d'information et des biais de négativité dans les données d'entraînement​((2023) A survey on data…)​.

**Utilisation de technologies avancées pour la vectorisation des métriques d'évaluation** : Des technologies telles que l'apprentissage profond et le traitement du langage naturel peuvent être utilisées pour vectoriser les métriques d'évaluation des ensembles de données, améliorant ainsi la mesure des métriques d'évaluation de la qualité des ensembles de données​((2023) A survey on data…)​.

**Amélioration de la gestion et du traitement des données** : Il est important de prendre en compte des facteurs tels que la gestion des données, leur traitement et les besoins des utilisateurs pour mesurer pleinement la qualité des données volumineuses dans des domaines comme l'intelligence artificielle, la médecine et les affaires​((2023) A survey on data…)​.

Théorie et modèles utilisés :

1. Dimensions de Qualité des Données=

- Complétude : Évalue si toutes les données nécessaires sont présentes.

- Auto-cohérence : Assure que les données sont logiquement cohérentes et exemptes de contradictions.

- Actualité : Vérifie que les données sont à jour et pertinentes dans le contexte actuel.

- Confidentialité : Garantit que l'accès aux données est limité aux utilisateurs autorisés.

- Exactitude : Mesure la précision des données par rapport à leur valeur réelle.

- Standardisation : Assure que les données respectent les normes et formats requis.

- Impartialité : Évalue l'absence de biais dans la distribution des données.

- Facilité d'utilisation : Mesure la simplicité d'utilisation et d'interprétation des données par les utilisateurs finaux [oai\_citation:1,(2023) A survey on dataset quality in machine learning.pdf](file-service://file-2gIuHgdd83S370OtM5UOlx4H).

Modèles Pratiques et Méthodes

1. Méthodes de Mesure Basées sur des Modèles de Machine Learning

Gong et al. (2023) ont détaillé plusieurs méthodes pratiques pour évaluer la qualité des données :

- Validité des Données : Utilisation de modèles de machine learning pour identifier et éliminer les données invalides ou bruitées. Par exemple, un modèle Autoencoder peut être utilisé pour encoder et décoder toutes les données, et les distances des vecteurs par rapport au centre des caractéristiques peuvent être utilisées pour identifier les données invalides.

- Exactitude des Étiquettes : Algorithmes basés sur l'apprentissage automatique pour filtrer les erreurs d'étiquetage, en utilisant des méthodes comme la validation croisée pour estimer la distribution conjointe des étiquettes de bruit et des étiquettes vraies [oai\_citation:2,(2023) A survey on dataset quality in machine learning.pdf](file-service://file-2gIuHgdd83S370OtM5UOlx4H).

2. Outils de Surveillance et de Validation des Données

Gong et al. (2023) ont également mentionné plusieurs outils pour surveiller et valider la qualité des données :

- Great Expectations : Permet de tester, documenter et profiler la qualité des données, en fournissant des résultats rapides avec des données volumineuses.

- Deequ : Une bibliothèque pour la validation continue de la qualité des données, développée par Amazon, qui utilise des tests de qualité des données basés sur des règles [oai\_citation:3,(2023) A survey on dataset quality in machine learning.pdf](file-service://file-2gIuHgdd83S370OtM5UOlx4H).

3. Évaluation de la Qualité Basée sur des Métriques Statistiques

- Complétude : Mesure la correspondance entre les éléments de données et les exigences, ainsi que la présence d'annotations correspondantes.

- Actualité : Évalue la pertinence des données au fil du temps, en comparant les listes d'attributs uniques avec des règles de temporisation.

- Imbalance : Mesure la différence de distribution entre les catégories de données pour identifier les déséquilibres potentiels [oai\_citation:4,(2023) A survey on dataset quality in machine learning.pdf](file-service://file-2gIuHgdd83S370OtM5UOlx4H).

**Mazurek, S. et Wielgosz, M. (2023). Assessing Dataset Quality Through Decision Tree Characteristics in Autoencoder-Processed Spaces**

L'étude de Mazurek et Wielgosz (2023) analyse l'impact de la qualité des ensembles de données sur les performances des modèles de machine learning, en utilisant des caractéristiques des arbres de décision dans des espaces traités par autoencodeurs. Les auteurs utilisent neuf ensembles de données variés, y compris des données à haute entropie et à haute redondance, pour évaluer les effets de la qualité des données sur la formation et la performance des modèles.

Objectifs principaux :

1. **Impact de la qualité des données :** Évaluer comment la qualité des ensembles de données affecte la profondeur et le nombre de feuilles des arbres de décision.
2. **Proposer un cadre d'évaluation :** Introduire un cadre pour évaluer la qualité des ensembles de données à l'aide d'autoencodeurs et d'arbres de décision.

Méthodologie :

* **Données expérimentales :** Utilisation de neuf ensembles de données, standardisés et traités par autoencodeurs.
* **Évaluation des modèles :** Formation d'arbres de décision sur les ensembles de données bruts et traités, en mesurant la profondeur et le nombre de feuilles.

Résultats :

* **Données de haute qualité :** Les ensembles de données de haute qualité produisent des arbres de décision moins complexes après traitement par autoencodeurs.
* **Données de faible qualité :** Les ensembles de données de faible qualité montrent peu de différence de complexité entre les arbres de décision formés sur les données brutes et traitées.
* **Utilité des autoencodeurs :** Les autoencodeurs aident à réduire la complexité des modèles en capturant les caractéristiques essentielles et en éliminant le bruit.

Contributions et limitations :

* **Contributions :** Fournit un cadre pour évaluer la qualité des ensembles de données et démontre l'utilité des autoencodeurs pour améliorer la qualité des données.
* **Limitations :** Nécessite des validations supplémentaires dans d'autres contextes et types de données.

En résumé, cette étude met en évidence l'importance de la qualité des ensembles de données pour les performances des modèles de machine learning et propose des outils pratiques pour améliorer cette qualité en utilisant des autoencodeurs et des arbres de décision.

**Danilov, G., Kotik, K., et al. (2023). Data Quality Estimation Via Model Performance: Machine Learning as a Validation Tool**

L'étude de Danilov et al. (2023) évalue l'impact de la qualité des données sur la performance des modèles de machine learning en utilisant des rapports opératoires en neurochirurgie. En redéfinissant les variables cibles selon des principes stricts, les auteurs montrent une amélioration significative de la classification des textes courts.

Objectifs principaux :

* **Impact de la qualité des données :** Évaluer comment la qualité des données influence la performance des modèles de classification de textes courts.
* **Validation des cibles :** Utiliser la performance des modèles comme outil de validation pour la qualité des données textuelles.

Méthodologie :

* **Données expérimentales :** Utilisation de rapports opératoires de neurochirurgie, nettoyés et tokenisés.
* **Reclassification :** Redéfinir les variables cibles basées sur le type de pathologie, la localisation et le type de manipulation.
* **Modèles de deep learning :** Utilisation de BiRNN-GRU pour la classification, avec des ensembles de données divisés en formations, validations et tests.

Résultats :

* **Amélioration de la classification :** Les nouvelles variables cibles ont permis d'atteindre une précision de 99,5% et un score F1 de 0,99.
* **Efficacité des modèles :** Les modèles ont montré une meilleure performance avec des données reclassifiées comparées aux anciennes classifications.

Contributions et limitations :

* **Contributions :** Démontre que des cibles bien définies et une représentation textuelle claire améliorent la performance des modèles de machine learning.
* **Limitations :** Nécessite des validations supplémentaires avec d'autres types de données textuelles.

En résumé, cette étude souligne l'importance d'une classification précise et bien définie pour obtenir des résultats optimaux en machine learning et propose d'utiliser la performance des modèles comme outil de validation pour la qualité des données textuelles.

**Zhou, Y., Tu, F., Sha, K., Ding, J., et Chen, H. (2024). A Survey on Data Quality Dimensions and Tools**

L'étude de Zhou et al. (2024) examine les dimensions et outils de la qualité des données pour le machine learning. Les auteurs passent en revue 17 outils d'évaluation et d'amélioration de la qualité des données, en analysant leurs forces et limitations, et proposent un cadre pour le développement futur d'outils open-source.

**Objectifs principaux :**

* **Analyser les dimensions de la qualité des données :** Identifier et définir les dimensions critiques pour évaluer la qualité des données dans les projets de machine learning.
* **Revoir les outils existants :** Passer en revue les outils d'évaluation et d'amélioration de la qualité des données développés au cours des cinq dernières années.
* **Proposer un cadre de développement :** Présenter un roadmap pour la création de nouveaux outils de qualité des données adaptés au machine learning.

**Méthodologie :**

* **Revue de la littérature :** Synthèse de 20 articles et identification de 8 dimensions de qualité des données, telles que la complétude, la précision, la cohérence et la ponctualité.
* **Évaluation des outils :** Analyse comparative de 17 outils open-source sur leurs fonctionnalités, métriques de qualité des données et utilisabilité.

**Résultats :**

* **Dimensions de qualité des données :** Les huit dimensions identifiées comprennent la complétude, la précision, la cohérence, et la contextualité, entre autres.
* **Outils d'évaluation :** Les outils analysés varient en termes de capacités de profilage des données, détection des anomalies, et suivi continu de la qualité des données.
* **Cadre de développement :** Proposition d'un cadre intégrant les tendances émergentes, telles que l'utilisation des grands modèles de langage et l'intelligence artificielle générative pour l'évaluation et l'amélioration de la qualité des données.

**Contributions et limitations :**

* **Contributions :** Offre une vue d'ensemble des dimensions de qualité des données et des outils existants, tout en proposant un cadre pour le développement futur d'outils adaptés au machine learning.
* **Limitations :** Les outils évalués peuvent nécessiter des validations supplémentaires dans des contextes différents pour confirmer leur efficacité.

En résumé, cette étude fournit une analyse détaillée des dimensions et outils de la qualité des données pour le machine learning et propose un cadre pour le développement de futurs outils, renforçant ainsi les pratiques de gestion de la qualité des données dans ce domaine.

Limitations

**Complexité des termes techniques et de l'installation** : L'étude souligne que de nombreux outils de qualité des données (DQ) nécessitent une configuration par code et ont des restrictions environnementales spécifiques, ce qui peut décourager les utilisateurs qui ne sont pas techniquement compétents​((2024) A Survey on Data…)​.

**Manque de personnalisation** : Il y a un manque d'options de personnalisation pour les règles de vérification de la qualité des données dans de nombreux outils. Cette limitation empêche les utilisateurs d'adapter les outils à des besoins et des scénarios spécifiques​((2024) A Survey on Data…)​.

**Méthodes de mesure de la qualité des données peu claires** : Les méthodes de mesure pour évaluer la qualité des données ne sont pas clairement définies ou standardisées à travers les outils. Cela rend difficile la comparaison de l'efficacité des différents outils et la compréhension de leurs résultats​((2024) A Survey on Data…)​.

**Support limité pour les tâches de Machine Learning** : Peu d'outils sont conçus spécifiquement pour évaluer les problèmes de qualité des données liés aux tâches de machine learning (ML). La plupart des outils se concentrent sur l'analyse des données générales plutôt que sur les besoins spécifiques du ML​((2024) A Survey on Data…)​.

**Gestion des grands volumes de données** : Les outils actuels peuvent avoir des difficultés à traiter et surveiller en continu les grands ensembles de données. Il existe des défis dans la gestion de la volatilité des données et pour s'assurer que les outils peuvent évoluer efficacement​((2024) A Survey on Data…)​.

**Interfaces utilisateur obsolètes** : Certains outils n'ont pas été mis à jour depuis longtemps, ce qui entraîne des interfaces utilisateur obsolètes qui manquent d'interactivité et d'informations suffisantes pour les utilisateurs​((2024) A Survey on Data…)​.

Axes d'amélioration

**Amélioration de l'interface utilisateur et de l'expérience utilisateur** : Développer des interfaces plus conviviales avec des guides clairs, des cas d'exemple et une navigation facile peut aider les utilisateurs non techniques à mieux utiliser les outils de qualité des données. Cela inclut la prise en charge des commandes low-code et la fourniture de visualisations et de tableaux de bord intuitifs​((2024) A Survey on Data…)​.

**Personnalisation et flexibilité** : Les outils doivent permettre aux utilisateurs de personnaliser les règles de vérification de la qualité des données et de réviser les règles actuelles pour mieux répondre à des demandes spécifiques. Cela peut être réalisé en permettant des métriques et des règles plus flexibles et définies par l'utilisateur​((2024) A Survey on Data…)​.

**Méthodes de mesure claires et standardisées** : Établir des définitions claires et standardisées pour les méthodes de mesure de la qualité des données à travers les outils améliorera la comparabilité et la fiabilité des évaluations de la qualité des données. Cela implique d'adopter des métriques communes et de s'assurer qu'elles sont bien définies et appliquées de manière cohérente​((2024) A Survey on Data…)​.

**Support spécifique pour les tâches de ML** : Concevoir des outils spécifiquement pour le machine learning peut aider à résoudre les problèmes uniques de qualité des données dans les tâches de ML. Ces outils doivent inclure des métriques pertinentes pour le ML et surveiller comment les améliorations de la qualité des données affectent la performance des modèles​((2024) A Survey on Data…)​.

**Évolutivité pour les grands volumes de données** : Améliorer la capacité des outils à gérer et surveiller en continu les grands ensembles de données sera crucial. Cela inclut l'optimisation des performances pour les scénarios de grands volumes de données et s'assurer que les outils peuvent évoluer sans dégradation significative des performances​((2024) A Survey on Data…)​.

**Intégration de l'IA et des techniques avancées** : L'intégration de l'IA et des techniques avancées comme l'augmentation de données générative et le low-code AI peut améliorer les capacités des outils de qualité des données. Ces avancées peuvent aider à automatiser et améliorer l'efficacité des évaluations et des transformations de la qualité des données​((2024) A Survey on Data…)​.

**Surveillance continue et automatisation** : Développer des fonctionnalités pour la surveillance continue des données et les vérifications de qualité automatisées aidera à maintenir une haute qualité des données au fil du temps. Cela implique de mettre en place des workflows qui se réactivent automatiquement lorsque de nouvelles données sont ingérées et de générer des rapports à jour​((2024) A Survey on Data…)​.

**Théorie et modèles utilisés :**

1. Dimensions de Qualité des Données de Wang et Strong

Wang et Strong (1996) ont proposé un cadre de dimensions de qualité des données comprenant quatre catégories principales :

Intrinsèque : Évalue les caractéristiques internes des données, comme la précision, la fiabilité et la crédibilité.

Contextuelle : Assure que les données répondent aux besoins spécifiques des projets de machine learning.

Représentationnelle : Concerne les formats et structures des données, incluant la cohérence et l’interprétabilité.

Accessibilité : Évalue la facilité d’accès et de partage des données avec des contrôles de sécurité .

**Outils et Méthodes Pratiques :**

1. Outils Open Source pour la Qualité des Données

De nombreux outils open source ont été développés pour évaluer et améliorer la qualité des données, parmi lesquels :

Great Expectations : Permet de tester, documenter et profiler la qualité des données, offrant des résultats rapides avec des données volumineuses et un vocabulaire extensible et lisible par l’homme .

Talend Open Studio : Un outil ETL (Extract, Transform, Load) gratuit et open source qui prend en charge le nettoyage des données, la visualisation et la standardisation des données, et l’identification automatique des types de données et des erreurs potentielles .

Soda Core : Une bibliothèque Python qui permet de trouver des données insuffisantes, de tester les données dans les pipelines de développement et de définir des rapports lisibles par l’homme, avec des alertes sur les problèmes de qualité des données .

3. Surveillance et Automatisation de la Qualité des Données

Outils de Surveillance : Des outils comme Great Expectations et Deequ permettent de profiler, valider et surveiller continuellement la qualité des données, en détectant les anomalies et en suggérant des stratégies d’amélioration .

Automatisation : La surveillance automatique des métriques définies pour les ensembles de données mis à jour permet de générer de nouveaux rapports et de faciliter les plans d’amélioration selon les résultats d’évaluation. Cela inclut la possibilité de définir des fréquences de rapports de surveillance et de valider les résultats de transformation de manière continue pour garantir la qualité des données à chaque étape de traitement

Pour explorer efficacement les combinaisons de techniques de complétude sur les données numériques et catégoriques, il est judicieux de sélectionner quelques combinaisons représentatives qui couvrent une gamme de complexité et d’approches différentes. Voici quelques combinaisons intéressantes à explorer :

**1. Combinaison Simple et Rapide**

• **Données Numériques** : Remplacement par la moyenne.

• **Données Catégoriques** : Remplacement par la valeur la plus fréquente (mode).

• **Pourquoi ?** : Cette combinaison est simple à mettre en œuvre et représente une méthode de base largement utilisée dans la pratique. Elle permet de tester la performance minimale de base.

**2. Combinaison Sensible aux Outliers**

• **Données Numériques** : Remplacement par la médiane.

• **Données Catégoriques** : Imputation par une nouvelle catégorie.

• **Pourquoi ?** : Cette combinaison est utile pour les ensembles de données où les outliers (valeurs aberrantes) sont présents. La médiane est moins affectée par les outliers que la moyenne, et l’ajout d’une nouvelle catégorie pour les valeurs manquantes dans les données catégoriques peut être une stratégie intéressante pour observer comment le modèle traite ces catégories supplémentaires.

**3. Combinaison Basée sur des Modèles**

• **Données Numériques** : Imputation par régression.

• **Données Catégoriques** : Imputation par un modèle prédictif (comme un arbre de décision).

• **Pourquoi ?** : Cette combinaison est plus avancée et exploite les relations entre les variables pour prédire les valeurs manquantes. Cela peut être particulièrement efficace si les variables sont fortement corrélées entre elles.

**4. Combinaison Basée sur la Proximité des Données**

• **Données Numériques** : K-Nearest Neighbors (KNN) imputation.

• **Données Catégoriques** : Hot Deck Imputation.

• **Pourquoi ?** : Ces techniques utilisent des méthodes basées sur la proximité ou la similarité des observations. Cette approche peut capturer les patterns locaux des données et être utile lorsque les données présentent des structures complexes.

**5. Combinaison pour Capturer l’Incertitude**

• **Données Numériques** : Multiple Imputation.

• **Données Catégoriques** : Multiple Imputation pour les catégories.

• **Pourquoi ?** : Cette combinaison est la plus complète et permet de capturer l’incertitude inhérente à l’imputation en générant plusieurs jeux de données imputés. Cela permet de mieux comprendre l’impact des valeurs manquantes sur les résultats des modèles de machine learning.

**6. Combinaison de Base avec Ajout de Catégories**

• **Données Numériques** : Remplacement par la moyenne.

• **Données Catégoriques** : Imputation avec une nouvelle catégorie.

• **Pourquoi ?** : Cette combinaison teste une approche de base pour les données numériques tout en explorant l’ajout de nouvelles catégories pour les données catégoriques manquantes, ce qui permet de voir comment un modèle traite ces nouvelles catégories.

**7. Combinaison de Réduction de Données**

• **Données Numériques** : Suppression des lignes avec valeurs manquantes.

• **Données Catégoriques** : Suppression des lignes avec valeurs manquantes.

• **Pourquoi ?** : Bien que cela puisse réduire considérablement la taille de l’ensemble de données, c’est intéressant pour évaluer si la suppression des données manquantes produit un meilleur modèle qu’une imputation simple.

**8. Combinaison Hybride : Simple vs Avancée**

• **Données Numériques** : Remplacement par la médiane.

• **Données Catégoriques** : Imputation par un modèle prédictif.

• **Pourquoi ?** : Cela combine une méthode simple pour les données numériques avec une méthode plus avancée pour les données catégoriques, permettant de comparer les approches directes et complexes dans un cadre hybride.

**9. Combinaison Optimale pour les Données Denses**

• **Données Numériques** : K-Nearest Neighbors (KNN) imputation.

• **Données Catégoriques** : Remplacement par la valeur la plus fréquente (mode).

• **Pourquoi ?** : Cette combinaison est efficace pour les jeux de données denses où les valeurs manquantes sont peu fréquentes mais importantes à combler de manière intelligente pour préserver les relations locales.

En testant ces combinaisons, vous pouvez évaluer une variété de stratégies allant de simples à complexes, et ainsi déterminer lesquelles sont les plus efficaces pour améliorer les performances de vos modèles de machine learning.