Calcul de l'impact de la qualité de données sur la performance des modèles de Machine Learning

**TABLE DES MATIERES**

[I. Introduction 4](#_Toc173853971)

[1.1 Contexte et Problématique 4](#_Toc173853972)

[1.1.1 Contexte 4](#_Toc173853973)

[1.1.2 Problématique 4](#_Toc173853974)

[1.1.3 Importance de l'Étude 4](#_Toc173853975)

[1.2 Objectifs du Mémoire 4](#_Toc173853976)

[1.2.1 Identifier les Dimensions Clés de la Qualité des Données 4](#_Toc173853977)

[1.2.2 Évaluer l'Impact de Différentes Méthodes de Traitement des Données Manquantes 5](#_Toc173853978)

[1.2.3 Comparer les Performances de Divers Algorithmes de ML et de Deep Learning 5](#_Toc173853979)

[1.2.4 Proposer des Recommandations pour Améliorer la Qualité des Données 5](#_Toc173853980)

[1.2.5 Synthèse des Résultats et Analyse Critique 5](#_Toc173853981)

[II. Revue de Littérature 6](#_Toc173853982)

[2.1 Introduction 6](#_Toc173853983)

[2.1.1 Objectif de la revue de littérature 6](#_Toc173853984)

[2.1.2 Importance de la qualité des données dans le contexte des modèles de machine learning 6](#_Toc173853985)

[2.1.3 Importance de cette revue pour le mémoire de fin d'étude 6](#_Toc173853986)

[2.2 Présentation des travaux antérieurs 6](#_Toc173853987)

[2.2.1 Synthèse des études récentes et classiques sur la qualité des données 6](#_Toc173853990)

[2.2.2 Évolution des recherches dans ce domaine 9](#_Toc173853991)

[2.2.3 Présentation des théories et modèles utilisés pour évaluer et améliorer la qualité des données 11](#_Toc173853992)

[2.3 Problématique et lacunes identifiées 13](#_Toc173853993)

[2.3.1 Lacunes dans les Données et les Études Existantes 13](#_Toc173853994)

[2.3.2 Questions de Recherche Non Résolues 14](#_Toc173853995)

[III. Cadre Théorique et Méthodologie 15](#_Toc173853996)

[3.1 Cadre Théorique pour l’Étude Proposée 15](#_Toc173853997)

[3.1.1 Définition de la qualité des données 15](#_Toc173853998)

[3.1.2 Dimensions de qualité de données 15](#_Toc173853999)

[3.1.3 Théorie et modèles existant 15](#_Toc173854000)

[3.2 Description des Données 15](#_Toc173854001)

[3.3 Processus de Préparation des Données 15](#_Toc173854002)

[3.4 Modèles et Algorithmes 15](#_Toc173854003)

[3.5 Stratégies d'Optimisation et d'Évaluation 15](#_Toc173854004)

[IV. Résultats et Analyse 15](#_Toc173854005)

[4.1 Impact de la Complétude des Données 15](#_Toc173854006)

[4.2 Impact de la Variabilité Temporelle des Données 16](#_Toc173854007)

[4.3 Étude Simultanée de Plusieurs Dimensions de Qualité 16](#_Toc173854008)

[4.4 Comparaison des Modèles 16](#_Toc173854009)

[V. Discussion 17](#_Toc173854010)

[5.1 Synthèse des Résultats 17](#_Toc173854011)

[5.2 Limites de l'Étude 17](#_Toc173854012)

[5.3 Recommandations 17](#_Toc173854013)

[VI. Conclusion 17](#_Toc173854014)

[6.1 Résumé des Constatations Clés 17](#_Toc173854015)

[6.2 Contributions du Mémoire 17](#_Toc173854016)

[6.3 Perspectives Futures 17](#_Toc173854017)

[VII. Bibliographie 17](#_Toc173854018)

[VIII. Annexes 17](#_Toc173854019)

# Introduction

## Contexte et Problématique

### Contexte

La qualité des données est un aspect fondamental du Machine Learning. En effet, l’exactitude, la complétude, la structure des données sur lesquelles les modèles de Machine Learning optimisent leurs fonctions en dépendent largement. Les entreprises et les chercheurs reconnaissent de plus en plus que la qualité des données affecte inévitablement le fonctionnement et la fiabilité des modèles prédictifs. C’est un élément majeur dans la mesure où la ML prend de plus en plus une importance décisive dans les décisions liées au domaine de la santé, de la finance et de l’énergie.

L’adoption accrue des technologies de ML et d’intelligence artificielle a été précédé de développements rapides dans ces domaines. Néanmoins, l’impact du déclin de la qualité des données sur les modèles de performances semble être un sujet relativement actif. L’étude actuelle aborde ce problème à travers l’objectif de recherche suivant: comment diverses dimensions de la qualité des données modélisent-elles la performance et l’interprétabilité des modèles ML.

### Problématique

La problématique centrale de ce mémoire est d’analyser l’impact de la qualité des données sur la performance des modèles de ML. Plus spécifiquement, cette étude vise à répondre aux questions suivantes :

**Comment la performance des modèles de machine learning est-elle influencée par la qualité des données ?**

### Importance de l'Étude

Cette étude est cruciale car elle fournit des perspectives pratiques pour améliorer les processus de ML. En identifiant les principales dimensions de la qualité des données qui affectent les performances des modèles, les entreprises peuvent développer des stratégies plus efficaces pour nettoyer et préparer leurs jeux de données. De plus, cette recherche aide à comprendre les limitations des modèles de ML lorsqu'ils sont confrontés à des données de qualité variable, fournissant ainsi des recommandations pour optimiser l'utilisation de ces modèles dans des contextes réels.

## Objectifs du Mémoire

### Identifier les Dimensions Clés de la Qualité des Données

L’un des principaux objectifs de ce mémoire est d’identifier les dimensions clés de la qualité des données qui affectent la performance des modèles de ML. Cela inclut la complétude, la précision, l’unicité, la balance des classes, l’uniformité et l’actualité des données. Comprendre ces dimensions permettra de mieux évaluer l’impact des différentes techniques de nettoyage et de préparation des données.

### Évaluer l'Impact de Différentes Méthodes de Traitement des Données Manquantes

Les valeurs manquantes sont un problème courant dans les jeux de données utilisés pour le ML. Ce mémoire vise à évaluer l’impact de différentes méthodes de traitement des données manquantes, telles que l’imputation par la moyenne, la médiane, et des techniques plus avancées comme le k-NN. L’objectif est de déterminer quelles méthodes conduisent aux meilleures performances des modèles.

### Comparer les Performances de Divers Algorithmes de ML et de Deep Learning

Un autre objectif est de comparer les performances de divers algorithmes de ML et de Deep Learning en fonction de la qualité des données. Cette comparaison inclura des modèles de régression (Elastic Net, XGBoost), de classification (Naive Bayes, AdaBoost) et de clustering (DBSCAN, Birch), ainsi que des modèles de Deep Learning (DNNs, RNNs)

### Proposer des Recommandations pour Améliorer la Qualité des Données

Enfin, ce mémoire proposera des recommandations pratiques pour améliorer la qualité des données dans les projets de ML. Ces recommandations seront basées sur les résultats obtenus et viseront à optimiser les processus de collecte, de nettoyage et de préparation des données pour maximiser la performance des modèles.

### Synthèse des Résultats et Analyse Critique

Une analyse critique des résultats obtenus permettra de comparer ces derniers avec les attentes théoriques et d’identifier les limites des méthodes utilisées. Cette synthèse offrira également des pistes pour des recherches futures afin de continuer à améliorer la qualité des données et la performance des modèles de ML.

# Revue de Littérature

## Introduction

### Objectif de la revue de littérature

Cette revue de littérature a pour objectif d'examiner l'impact de la qualité des données sur la performance des modèles de machine learning. Nous chercherons à définir les principales dimensions de la qualité des données, à synthétiser les recherches antérieures, à identifier les lacunes existantes, et à formuler des questions de recherche pertinentes pour le futur. La qualité des données est un aspect crucial dans le développement de modèles de machine learning efficaces, influençant directement leur précision, leur robustesse et leur capacité à généraliser. Comprendre ces impacts est essentiel pour améliorer les pratiques actuelles et orienter les futures recherches dans ce domaine.

### Importance de la qualité des données dans le contexte des modèles de machine learning

La qualité des données est souvent citée comme un facteur déterminant dans la performance des modèles de machine learning. Les données de haute qualité permettent d'entraîner des modèles plus précis, robustes et capables de généraliser correctement à de nouvelles données. À l'inverse, des données de mauvaise qualité peuvent conduire à des modèles biaisés, à des performances médiocres et à des interprétations erronées des résultats. Dans ce contexte, l'évaluation et l'amélioration de la qualité des données sont devenues des priorités pour les chercheurs et les praticiens du machine learning.

### Importance de cette revue pour le mémoire de fin d'étude

Cette revue est cruciale pour le mémoire de fin d'étude car elle établit une base théorique solide et fournit une vue d'ensemble des recherches actuelles et passées sur la qualité des données et leur impact sur les modèles de machine learning. Elle permet également de situer la recherche proposée dans le contexte plus large des travaux existants, d'identifier les lacunes et de justifier la pertinence des nouvelles questions de recherche abordées dans ce mémoire.

## Présentation des travaux antérieurs



### Synthèse des études récentes et classiques sur la qualité des données

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

L'étude de Sessions et Valtorta (2006) démontre l'importance cruciale de la qualité des données pour les algorithmes de machine learning, en particulier les réseaux bayésiens (BNs). Se concentrant sur l'exactitude des données, ils montrent que même de petites inexactitudes (10%) peuvent fortement dégrader les performances de l'algorithme PC.

Les auteurs développent et testent plusieurs méthodes pour intégrer les évaluations de la qualité des données : laisser les données telles quelles (Do Nothing), filtrer les données de faible qualité (Threshold), attribuer des poids de qualité (DQ Method Part One et Two), et ajuster les niveaux de signification (DQ Method Part Three). Ces méthodes visent à améliorer la robustesse des modèles en tenant compte 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. Ils introduisent 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. Les résultats montrent que les impacts varient selon le type d'erreur et le modèle utilisé, les données manquantes ayant généralement des effets plus prononcés. Certains modèles, comme KNN et DBSCAN, montrent une meilleure tolérance aux erreurs comparés à d'autres comme les réseaux bayésiens et la régression logistique. L'étude fournit une analyse détaillée et propose des outils pratiques pour évaluer et améliorer la qualité des données, bien que les conclusions puissent nécessiter des validations supplémentaires dans d'autres contextes.

#### Budach, L., et al. (2022). The Effects of Data Quality on Machine Learning Performance

L'étude de Budach et al. (2022) examine comment six dimensions de la qualité des données (représentation cohérente, complétude, précision des caractéristiques, précision des cibles, unicité et équilibre des classes) influencent les performances de quinze algorithmes de machine learning. Les auteurs testent ces effets sur des tâches de classification, régression et clustering, en simulant des scénarios de pollution des données. Ils constatent que la qualité des données impacte fortement les performances des modèles, avec des effets variés selon le type d'erreur et l'algorithme. Certains modèles, comme les réseaux de neurones, montrent une meilleure tolérance aux erreurs. La pollution des données de test diminue particulièrement les performances. Cette étude souligne l'importance de la qualité des données pour les modèles de machine learning et propose des approches pour améliorer la gestion de cette qualité. Les résultats, bien que basés sur des ensembles de données spécifiques, offrent des insights précieux pour la pratique et la recherche futures.

#### 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) examine l'impact de la complétude et de la correction des données sur les modèles de machine learning explicables, en utilisant des données IoT et des arbres de décision. Les auteurs montrent que la perte de lignes de données a un impact plus important sur les performances des modèles que la perte de caractéristiques. De plus, les données incorrectes dégradent davantage la performance des modèles que les données manquantes. Les résultats révèlent des variations spécifiques en termes d'exactitude, de précision et de rappel selon les types et niveaux de défauts des données. Cette étude souligne l'importance de la qualité des données pour les modèles de machine learning explicables et propose des méthodes pour améliorer la gestion de cette qualité dans les systèmes IoT.

#### 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) examine l'importance de la qualité des ensembles de données pour les performances des modèles de machine learning, en définissant des concepts clés et en identifiant les problèmes de qualité. Les auteurs proposent un cadre d'évaluation complet comprenant huit dimensions de qualité (comme la complétude, l'exactitude et l'absence de biais) et 32 métriques pour évaluer chaque dimension. Ce cadre aide à améliorer la précision et l'efficacité des modèles de machine learning. Bien que l'étude fournisse une structure utile pour évaluer la qualité des données, elle nécessite des validations supplémentaires dans divers contextes pour confirmer l'efficacité des méthodes proposées. En résumé, cette étude offre des outils précieux pour améliorer la qualité des données et les performances des modèles de machine learning.

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

L'étude de Mazurek et Wielgosz (2023) évalue 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. En utilisant neuf ensembles de données variés, les auteurs montrent que les ensembles de haute qualité produisent des arbres de décision moins complexes après traitement par autoencodeurs, tandis que les ensembles de faible qualité montrent peu de différence entre les arbres formés sur les données brutes et traitées. Cette étude propose un cadre d'évaluation pratique pour améliorer la qualité des données en capturant les caractéristiques essentielles et en éliminant le bruit, démontrant ainsi l'utilité des autoencodeurs pour la gestion de la qualité des données.

#### 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 identifient huit dimensions critiques de la qualité des données (complétude, précision, cohérence, etc.) et passent en revue 17 outils d'évaluation et d'amélioration de la qualité des données, en analysant leurs fonctionnalités et limitations. Ils proposent également un cadre pour le développement futur d'outils open-source, intégrant des tendances émergentes comme les grands modèles de langage et l'IA générative. Cette étude offre une vue d'ensemble des pratiques actuelles et un guide pour l'amélioration future des outils de gestion de la qualité des données.

### Évolution des recherches dans ce domaine

L'évolution des recherches sur la qualité des données dans le domaine du machine learning a suivi un parcours riche et diversifié, passant de simples préoccupations de nettoyage des données à des approches sophistiquées intégrant des techniques avancées de deep learning et des cadres d'évaluation complexes.

#### Débuts et Approches Classiques

Les premières études sur la qualité des données se concentraient principalement sur des problèmes de base tels que la complétude, l'exactitude et la cohérence des données. Les travaux pionniers, comme ceux de Little et Rubin (1987), mettaient l'accent sur le traitement des données manquantes et les techniques de nettoyage des données, en se concentrant sur la correction des erreurs évidentes et la normalisation des jeux de données. Ces études ont établi les bases de la compréhension des impacts directs des données de mauvaise qualité sur les analyses statistiques et les modèles de machine learning simples.

#### Avancées avec les Réseaux Bayésiens et les Algorithmes de Correction

Avec l'avènement des réseaux bayésiens et des méthodes d'inférence plus sophistiquées, les chercheurs ont commencé à examiner plus en profondeur comment la qualité des données affecte la performance des modèles. L'étude de Sessions et Valtorta (2006) a démontré que même des algorithmes avancés comme les réseaux bayésiens peuvent être sévèrement affectés par des données inexactes. Ils ont proposé des méthodes pour intégrer des évaluations de la qualité des données dans les algorithmes, illustrant une approche plus intégrée et proactive de la gestion de la qualité des données.

#### Impact des Big Data et du Deep Learning

Avec l'essor des big data et des techniques de deep learning, les enjeux de la qualité des données ont pris une nouvelle dimension. Gudivada et al. (2017) ont exploré les problèmes spécifiques de qualité des données dans les environnements de big data, soulignant que des simples opérations de nettoyage ne suffisent plus. Ils ont proposé des cadres de gouvernance des données pour gérer la qualité tout au long du cycle de vie des données.

L'impact de la qualité des données sur les modèles de machine learning est devenu encore plus critique avec les avancées en deep learning. Srivastava et al. (2022) ont montré comment la fréquence d'échantillonnage, l'étiquetage et l'extraction des caractéristiques influencent les performances des modèles de classification dans le contexte des vibrations de forage, mettant en évidence la complexité accrue de la gestion de la qualité des données dans des applications spécifiques.

#### Méthodes Empiriques et Évaluations Systématiques

Des études plus récentes, comme celle de Budach et al. (2022), ont mené des évaluations systématiques des impacts de différentes dimensions de la qualité des données sur les performances des modèles de machine learning. Ces recherches ont utilisé des méthodes empiriques rigoureuses pour démontrer comment des dimensions comme la complétude, la précision et l'équilibre des classes affectent la performance des algorithmes de classification, de régression et de clustering. Elles ont également introduit de nouvelles métriques pour évaluer la tolérance des modèles aux erreurs de données.

#### Intégration des Autoencodeurs et des Modèles Explicables

Les travaux de Mazurek et Wielgosz (2023) ont poussé plus loin en utilisant des autoencodeurs pour traiter les données avant d'évaluer leurs impacts sur les arbres de décision, démontrant que les modèles peuvent devenir plus robustes et moins complexes lorsque les caractéristiques essentielles sont bien capturées. Parallèlement, Azimi et Pahl (2022) ont étudié l'impact de la complétude et de la correction des données sur les modèles explicables, soulignant l'importance de la clarté et de la précision des données pour obtenir des résultats fiables.

#### Cadres d'Évaluation et Outils Open-Source

Plus récemment, des efforts ont été faits pour systématiser l'évaluation de la qualité des données et développer des outils open-source. Gong et al. (2023) ont proposé un cadre complet intégrant huit dimensions de qualité des données et 32 métriques, tandis que Zhou et al. (2024) ont examiné les outils existants et proposé un cadre pour le développement futur d'outils, intégrant des tendances comme les grands modèles de langage et l'IA générative.

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

#### Dimensions de Qualité des Données

Les dimensions de qualité des données sont essentielles pour évaluer et améliorer les performances des modèles de machine learning. Budach et al. (2022) identifient plusieurs dimensions clés : la représentation cohérente, la complétude, l'exactitude des caractéristiques, l'exactitude des cibles, l'unicité, et l'équilibre des classes cibles. La représentation cohérente assure que chaque entité est représentée uniformément, évitant les incohérences comme des dénominations variées pour une même entité. La complétude se réfère à l'absence de valeurs manquantes, cruciales pour éviter des biais dans les modèles. L'exactitude des caractéristiques et des cibles mesure la précision des données par rapport à leur valeur réelle, garantissant la fiabilité des résultats des modèles. L'unicité évite les duplicatas qui pourraient fausser les analyses, tandis que l'équilibre des classes cibles assure une distribution homogène des classes, essentiel pour les modèles de classification.

Gong et al. (2023) proposent des dimensions complémentaires telles que l'auto-cohérence, qui garantit l'absence de contradictions internes dans les données, et l'actualité, qui vérifie que les données sont pertinentes et à jour. La confidentialité, l'exactitude, la standardisation, l'impartialité et la facilité d'utilisation sont également des dimensions critiques. La standardisation assure que les données respectent des formats et normes requis, et l'impartialité évalue l'absence de biais, améliorant ainsi la fiabilité des modèles.

Zhou et al. (2024), en se basant sur le cadre de Wang et Strong (1996), catégorisent ces dimensions en quatre grandes catégories : intrinsèque, contextuelle, représentationnelle, et accessibilité. Les dimensions intrinsèques, comme la précision et la crédibilité, assurent la fiabilité interne des données. Les dimensions contextuelles garantissent que les données répondent aux besoins spécifiques des projets. Les dimensions représentationnelles concernent la cohérence et l’interprétabilité des formats de données, tandis que l'accessibilité évalue la facilité d’accès et de partage des données avec des contrôles de sécurité adéquats.

#### Évaluation Empirique de la Qualité des Données

Les approches empiriques jouent un rôle essentiel dans la validation des théories et modèles proposés. Budach et al. (2022) ont mené une évaluation empirique rigoureuse en analysant des scénarios réalistes et en expérimentant avec divers algorithmes de machine learning. Ils ont étudié 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. En testant 15 algorithmes de machine learning couvrant les tâches de classification, de régression et de clustering, ils ont observé comment les différentes dimensions de qualité des données affectaient leurs performances.

#### Approches Spécifiques à Certaines Tâches

Les études ont également proposé des approches spécifiques pour évaluer la qualité des données en fonction des tâches de machine learning. Pour la classification, le score F1 est utilisé pour évaluer la performance des modèles, tenant compte de l'impact des données déséquilibrées. Pour la régression, le coefficient de détermination (R²) mesure la fraction de variance expliquée par le modèle. Pour le clustering, l'information mutuelle ajustée (AMI) évalue la qualité des clusters produits par les algorithmes de clustering. Ces théories et modèles offrent un cadre robuste pour comprendre et améliorer la qualité des données, essentielle pour optimiser la performance des modèles de machine learning.

## Problématique et lacunes identifiées

### Lacunes dans les Données et les Études Existantes

La qualité des données est un facteur crucial influençant la performance des modèles de machine learning. Toutefois, plusieurs études récentes ont mis en évidence des limitations significatives dans ce domaine, pointant vers des lacunes spécifiques qui nécessitent une attention et une exploration plus approfondies.

#### Optimisation des hyperparamètres et architectures des modèles

**Budach et al. (2022)** soulignent que l'absence d'optimisation des hyperparamètres constitue une limitation majeure. L'étude n'a pas inclus cette optimisation pour les algorithmes testés, ce qui implique que les performances rapportées peuvent ne pas refléter les meilleures performances possibles pour chaque algorithme. De plus, l'autoencodeur utilisé dans l'étude est basé sur un réseau de neurones basique, sans optimisation spécifique pour sa tâche. Des architectures plus sophistiquées et une optimisation rigoureuse des hyperparamètres pourraient considérablement améliorer les résultats obtenus.

#### Traitement des données manquantes

La gestion des données manquantes est également un point critique. Dans l'étude de **Budach et al. (2022)**, des valeurs de remplacement fixes ont été utilisées pour simuler l'absence de données, ce qui pourrait influencer négativement les performances des modèles. L'exploration de différentes valeurs de remplacement et méthodes d'imputation pourrait offrir des solutions plus robustes et précises.

#### Évaluation limitée des algorithmes de clustering

L'évaluation des algorithmes de clustering dans l'étude de **Budach et al. (2022)** est limitée à la mesure AMI (Adjusted Mutual Information). Une évaluation plus complète, incluant des métriques telles que la taille du chevauchement des clusters et la distribution des tailles des clusters, fournirait une vue plus globale de leur performance.

#### Généralisation des résultats

**Azimi et Pahl (2022)** ont utilisé un nombre limité de jeux de données (trafic et météo), ce qui restreint la généralisation des résultats à d'autres types de données ou domaines d'application. De plus, cette étude se concentre exclusivement sur les arbres de décision, bien que d'autres modèles, comme les réseaux neuronaux ou les SVM (Support Vector Machines), pourraient présenter des comportements différents face à des problèmes de qualité des données.

#### Volume et nature des données

Le volume de données utilisé dans l'étude de **Azimi et Pahl (2022)** est relativement faible, ce qui peut affecter la robustesse des conclusions tirées. De plus, l'étude se limite à des données numériques, négligeant les données textuelles ou d'image. Une augmentation du volume de données et l'inclusion de différents types de données permettraient de valider les résultats et de rendre les conclusions plus robustes.

#### Normes de qualité des ensembles de données

**Gong et al. (2023)** notent un manque d'uniformité dans les métriques d'évaluation des ensembles de données, ce qui rend difficile l'obtention d'une description unifiée de l'assurance qualité. De nouvelles dimensions de qualité doivent être évaluées en fonction des caractéristiques spécifiques au domaine de l'apprentissage automatique, et des recherches supplémentaires sont nécessaires pour approfondir la mesure des métriques d'évaluation de la qualité des ensembles de données.

#### Complexité des outils de qualité des données

**Zhou et al. (2024)** mettent en évidence la complexité des termes techniques et de l'installation des outils de qualité des données (DQ), ainsi que le manque de personnalisation des règles de vérification de la qualité des données. De plus, les méthodes de mesure de la qualité des données ne sont pas clairement définies ou standardisées à travers les outils, rendant difficile la comparaison de leur efficacité. Il est également noté que peu d'outils sont spécifiquement conçus pour évaluer les problèmes de qualité des données liés aux tâches de machine learning.

#### Surveillance continue et automatisation

Une autre lacune importante identifiée par **Zhou et al. (2024)** concerne la gestion des grands volumes de données et la surveillance continue. Les outils actuels peuvent avoir des difficultés à traiter et surveiller en continu les grands ensembles de données, ce qui est crucial pour maintenir une haute qualité des données au fil du temps.

### Questions de Recherche Non Résolues

#### Optimisation des Hyperparamètres et Architectures des Modèles

Question de Recherche : Comment l'optimisation des hyperparamètres influence-t-elle la robustesse des modèles de machine learning face à des données de qualité variable ?

Justification : Les études mentionnées indiquent un manque d'optimisation des hyperparamètres. Explorer cette dimension pourrait révéler des stratégies pour améliorer la performance des modèles même avec des données de qualité imparfaite.

#### Méthodes d'Imputation des Données Manquantes

Question de Recherche : Quelle méthode d'imputation des données manquantes (moyenne, médiane, k-NN, MICE, etc.) offre les meilleures performances pour différents types de modèles de machine learning ?

Justification : Les méthodes d'imputation des données manquantes peuvent avoir des impacts variés sur les performances des modèles. Comparer ces méthodes peut aider à identifier celle(s) qui maximisent les performances des modèles dans divers contextes.

#### Évaluation de la Qualité des Données pour le Clustering

Question de Recherche : Quelles métriques supplémentaires peuvent être utilisées pour évaluer la qualité des clusters dans des ensembles de données polluées ?

Justification : Les études actuelles utilisent principalement l'AMI pour évaluer les clusters. Explorer d'autres métriques pourrait fournir une évaluation plus complète des algorithmes de clustering en présence de données de qualité variable.

#### Généralisation des Résultats à Divers Types de Données

Question de Recherche : Comment les conclusions tirées de l'impact de la qualité des données sur les modèles de machine learning varient-elles avec différents types de données (numériques, textuelles, images) ?

Justification : La plupart des études se concentrent sur des données numériques. Étendre la recherche à d'autres types de données permettrait de généraliser les résultats et de proposer des solutions adaptées à divers contextes.

#### Impact de la Variabilité Temporelle des Données

Question de Recherche : Comment la variabilité temporelle des données influence-t-elle la performance des modèles de machine learning, et quelles techniques peuvent atténuer cet impact ?

Justification : Étudier l'impact des changements temporels dans les données pourrait aider à développer des modèles plus robustes aux variations dans les données de test et de formation.

#### Cadre d'Évaluation de la Qualité des Données

Question de Recherche : Quel cadre d'évaluation unifié peut être développé pour mesurer de manière systématique la qualité des ensembles de données utilisés dans les projets de machine learning ?

Justification : Actuellement, il existe un manque d'uniformité dans les métriques d'évaluation de la qualité des données. Développer un cadre standardisé aiderait à améliorer la comparabilité et la robustesse des études futures.

#### Complexité des Outils de Qualité des Données:

Question de recherche: Comment simplifier l'utilisation et l'installation des outils de qualité des données pour les non-experts ?

**Justification** : La complexité technique des outils de qualité des données est un obstacle majeur à leur adoption par les praticiens non-experts. De nombreux outils disponibles nécessitent une compréhension approfondie de concepts avancés en machine learning et en gestion de données, ce qui limite leur utilisation à un cercle restreint de spécialistes. Simplifier ces outils permettrait une adoption plus large et une meilleure gestion de la qualité des données dans divers contextes professionnels.

#### Efficacité des Outils de Qualité des Données :

Question de recherche : Quelle est l'efficacité des outils open-source actuels pour la qualité des données lorsqu'ils sont intégrés dans des pipelines de ML complexes?

**Justification** : L'intégration des outils de qualité des données dans des pipelines de machine learning complexes représente un défi technique important. Les outils open-source offrent de nombreuses possibilités, mais leur efficacité réelle lorsqu'ils sont déployés dans des environnements de production complexes reste à évaluer. Cette question vise à analyser la performance, la fiabilité, et la facilité d'intégration de ces outils dans des workflows de ML.

#### Automatisation et Surveillance Continue de la Qualité des Données

Question de Recherche : Quels outils et techniques peuvent être développés pour l'automatisation et la surveillance continue de la qualité des données dans des environnements de big data ?

Justification : La gestion continue de la qualité des données est cruciale dans les environnements de big data. Des outils automatisés peuvent améliorer la qualité des données de manière proactive.

# Cadre Théorique et Méthodologie

## Cadre Théorique pour l’Étude Proposée

### **Définition de la qualité des données**

### Dimensions de qualité de données

### Théorie et modèles existant

## Description des Données

- Présentation des datasets utilisés : Air Quality, Occupancy Detection, Diabetes Clinical, Power Consumption, Online Retail, DataSet de l’entreprise.

- Justification du choix des datasets en fonction des critères (nombre de lignes, features, valeurs manquantes, etc.).

## Processus de Préparation des Données

- Nettoyage des données et traitement des valeurs manquantes.

- Techniques de réduction de dimensionalité utilisées (autoencodeurs, PCA, etc.).

## Modèles et Algorithmes

- Présentation des modèles de régression (Elastic Net, XGBoost, etc.).

- Modèles de classification (Naive Bayes, AdaBoost, etc.).

- Algorithmes de clustering (DBSCAN, Birch, etc.).

- Modèles de Deep Learning (DNNs, RNNs, etc.).

## Stratégies d'Optimisation et d'Évaluation

- Hyperparamètres et configurations spécifiques.

- Méthodes d'évaluation des performances des modèles.

# Résultats et Analyse

## Impact de la Complétude des Données

- Évaluation de différentes valeurs de remplacement.

- Analyse des performances des modèles en fonction des méthodes d'imputation.

## Impact de la Variabilité Temporelle des Données

- Étude de l'impact des changements dans les données au fil du temps sur la performance des modèles.

## Étude Simultanée de Plusieurs Dimensions de Qualité

- Analyse combinée des effets de différentes dimensions de qualité sur les performances des modèles.

## Comparaison des Modèles

- Performance des modèles de Machine Learning classiques vs Deep Learning.

- Impact des données de qualité variable sur l'interprétabilité des modèles.

# Discussion

## Synthèse des Résultats

- Comparaison des résultats obtenus avec les attentes théoriques.

- Analyse critique des méthodes et des résultats.

## Limites de l'Étude

- Problèmes rencontrés et limitations des méthodes utilisées.

- Suggestions pour des recherches futures.

## Recommandations

- Meilleures pratiques pour améliorer la qualité des données.

- Conseils pour la sélection et l'utilisation des modèles en fonction de la qualité des données.

# Conclusion

## Résumé des Constatations Clés

- Points saillants de l'étude et implications pratiques.

## Contributions du Mémoire

- Apports à la recherche et à la pratique professionnelle en Machine Learning.

## Perspectives Futures

- Directions possibles pour des recherches ultérieures.

# Bibliographie

# Annexes

- Tableaux de données, codes sources, résultats supplémentaires, etc.