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

# Introduction

## 1. Contexte et Problématique

### 1.1 Contexte

La qualité des données est un pilier essentiel dans le domaine du Machine Learning (ML). En effet, les modèles de ML dépendent fortement de la précision, de la complétude et de la structure des données sur lesquelles ils sont formés. Les entreprises et les chercheurs reconnaissent de plus en plus que des données de mauvaise qualité peuvent compromettre l'efficacité et la fiabilité des modèles prédictifs. Ce constat est particulièrement pertinent dans un contexte où les décisions basées sur des modèles de ML influencent des domaines cruciaux tels que la santé, la finance, et l'énergie.

Les avancées rapides dans les techniques de ML et d'intelligence artificielle (IA) ont conduit à une adoption croissante de ces technologies. Toutefois, l'impact de la qualité des données sur la performance des modèles reste un domaine de recherche très actif. Cette étude se situe dans ce contexte en cherchant à comprendre comment différentes dimensions de la qualité des données influencent la performance et l'interprétabilité des modèles de ML.

### 1.2 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 les valeurs manquantes affectent-elles la performance des modèles de ML ?**
* **Quel est l'impact du bruit ajouté aux données sur les performances des modèles ?**
* **Les doublons et le déséquilibre des classes dans les jeux de données influencent-ils la précision des modèles ?**
* **La variabilité temporelle des données impacte-t-elle la robustesse des modèles de ML ?**
* **Quels sont les effets combinés de plusieurs dimensions de la qualité des données sur les performances des modèles ?**

### 1.3 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.

### 1.4 Contexte Théorique et Scientifique

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

Le cadre théorique de cette étude repose sur plusieurs dimensions clés de la qualité des données.

* **Complétude** : La présence ou l'absence de valeurs dans les jeux de données.
* **Précision** : Le degré de fidélité des données aux valeurs réelles.
* **Unicité** : L'absence de doublons dans les jeux de données.
* **Balance des Classes** : La distribution égale ou inégale des classes dans les données de classification.
* **Uniformité** : La consistance des données en termes de format et de type.
* **Actualité** : La pertinence des données par rapport à leur temps de collecte.

Chaque dimension sera analysée individuellement et en combinaison avec les autres pour évaluer leur impact sur différents types de modèles de ML, y compris les modèles de régression, de classification et de clustering.

#### Modèles de Machine Learning

Types de modèles (supervisé, non supervisé, semi-supervisé).

Sensibilité des modèles à la qualité des données.

#### Relation entre qualité des données et performance des modèles

Théories et modèles existants

### 1.5 Hypothèses de Recherche

Les hypothèses principales de cette recherche sont les suivantes :

* Les valeurs manquantes et les erreurs dans les données diminuent significativement la performance des modèles de ML.
* La présence de doublons et de déséquilibres dans les classes affecte négativement la précision et la robustesse des modèles.
* Les variations temporelles des données influencent la stabilité des modèles de ML, particulièrement dans les séries temporelles.
* L’analyse simultanée de plusieurs dimensions de qualité des données peut révéler des interactions complexes qui ne sont pas apparentes lorsqu’elles sont examinées séparément.

## 2. Objectifs du Mémoire

### 2.1 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.

### 2.2 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.

### 2.3 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)

### 2.4 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.

### 2.5 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.

# Chapitre 1 : 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.

## 2. 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

* 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.

# Chapitre 2 : Méthodologie

## 1. 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.).

## 2. Processus de Préparation des Données

- Nettoyage des données et traitement des valeurs manquantes.

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

## 3. 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.).

## 4. Stratégies d'Optimisation et d'Évaluation

- Hyperparamètres et configurations spécifiques.

- Méthodes d'évaluation des performances des modèles.

# Chapitre 3 : Résultats et Analyse

## 1. Impact de la Complétude des Données

- Évaluation de différentes valeurs de remplacement.

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

## 2. 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.

## 3. Étude Simultanée de Plusieurs Dimensions de Qualité

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

## 4. Comparaison des Modèles

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

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

# Chapitre 4 : Discussion

## 1. Synthèse des Résultats

- Comparaison des résultats obtenus avec les attentes théoriques.

- Analyse critique des méthodes et des résultats.

## 2. Limites de l'Étude

- Problèmes rencontrés et limitations des méthodes utilisées.

- Suggestions pour des recherches futures.

## 3. 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

## 1. Résumé des Constatations Clés

- Points saillants de l'étude et implications pratiques.

## 2. Contributions du Mémoire

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

## 3. Perspectives Futures

- Directions possibles pour des recherches ultérieures.

# Bibliographie

# Annexes

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