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

La qualité des données est cruciale pour la performance des modèles de machine learning. Des données de mauvaise qualité peuvent entraîner des prédictions erronées et des décisions basées sur des informations incorrectes. Cette revue de littérature vise à explorer les travaux existants sur la qualité des données, identifier les lacunes et les questions non résolues, et établir un cadre théorique pour la recherche.

## Présentation des travaux antérieurs

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

La qualité des données est un aspect fondamental dans la performance des modèles de machine learning. Diverses études ont exploré les différentes dimensions de la qualité des données, y compris la complétude, la correction, la diversité, et la représentativité des données.

#### 1.1.1 Étude sur l'impact de la qualité des données sur les modèles de machine learning

Danilov et al. (2023) ont exploré l'estimation de la qualité des données à travers la performance des modèles de machine learning. Leur étude a montré que l'amélioration des variables cibles en suivant des principes stricts améliore considérablement la performance de la classification des rapports opératoires en neurochirurgie. Les résultats de cette étude montrent que des textes bien séparés en termes de signification permettent une classification plus précise par apprentissage automatique [oai\_citation:1,data-quality-estimation-via-model-performance-machine-18f5ucey.pdf](file-service://file-lyM3k4Kxm3n7rXow2BpwdHge).

#### 1.1.2 Évaluation de la qualité des ensembles de données via les caractéristiques des arbres de décision

Mazurek et Wielgosz (2023) ont proposé un cadre d'évaluation de la qualité des ensembles de données en utilisant les caractéristiques des arbres de décision dans des espaces traités par autoencodeur. Ils ont illustré que des ensembles de données de haute qualité permettent aux modèles d'atteindre de meilleures performances, tandis que des ensembles de données de faible qualité augmentent la complexité des arbres de décision, soulignant l'importance de la sélection appropriée des caractéristiques et du volume des données [oai\_citation:2,assessing-dataset-quality-through-decision-tree-25ue5ud9.pdf](file-service://file-9k8V5GFEGMDhOsxDBQENrX3C).

#### 1.1.3 Impact de la complétude et de la correction des données sur les modèles explicables de machine learning

Azimi et Pahl (2022) ont étudié l'impact de la complétude et de la correction des données sur les modèles explicables de machine learning en utilisant des arbres de décision. Leur recherche a montré que des données incomplètes ou incorrectes peuvent diminuer significativement la précision, la sensibilité et le rappel des modèles, démontrant l'importance de la qualité des données dans les systèmes IoT [oai\_citation:3,JDI\_2022\_Explainable-ML-Data-Quality\_CR.pdf](file-service://file-vHiSZUFV8APQFNNByKkiKSub).

#### 4. Étude sur les effets de la qualité des données sur les algorithmes de machine learning

Valtorta (2006) a investigué les effets de la qualité des données sur les algorithmes de machine learning. Cette étude historique a mis en lumière comment les erreurs dans les données peuvent entraîner des performances sous-optimales des algorithmes de machine learning, en soulignant l'importance d'une bonne préparation et validation des données avant leur utilisation dans les modèles [oai\_citation:4,data-quality-estimation-via-model-performance-machine-18f5ucey.pdf](file-service://file-lyM3k4Kxm3n7rXow2BpwdHge).

#### 5. Estimation de la qualité des données à l'aide de la performance des modèles

Une étude menée par un groupe de chercheurs, détaillée dans le document "Data Quality Estimation Via Model Performance", a exploré comment la performance des modèles de machine learning peut être utilisée comme un outil de validation pour estimer la qualité des données. Ils ont démontré que l'amélioration de la variable cible basée sur des principes stricts peut considérablement améliorer les résultats de classification des rapports opératoires, soulignant ainsi l'importance d'une représentation textuelle non ambiguë pour des performances optimales des modèles [oai\_citation:5,data-quality-estimation-via-model-performance-machine-18f5ucey.pdf](file-service://file-lyM3k4Kxm3n7rXow2BpwdHge).

#### 6. Qualité des ensembles de données et leurs impacts sur les tâches de classification

Dans l'étude intitulée "Assessing Dataset Quality Through Decision Tree Characteristics in Autoencoder-Processed Spaces", Mazurek et Wielgosz ont examiné l'impact de la qualité des ensembles de données sur les tâches de classification. Ils ont montré que des ensembles de données bien structurés et diversifiés permettent d'obtenir des modèles de machine learning plus robustes et performants. Leur recherche propose un cadre complet pour l'évaluation de la qualité des ensembles de données, offrant des insights précieux pour les chercheurs et les praticiens dans le domaine du machine learning [oai\_citation:6,assessing-dataset-quality-through-decision-tree-25ue5ud9.pdf](file-service://file-9k8V5GFEGMDhOsxDBQENrX3C).

### 1.2 Évolution des recherches dans ce domaine

Les recherches sur la qualité des données dans le machine learning ont considérablement évolué au fil du temps. Initialement, les efforts se concentraient sur la gestion des données manquantes et incorrectes grâce à des techniques d'imputation et de nettoyage. Avec le développement des techniques de machine learning, l'accent a été mis sur l'utilisation de modèles pour évaluer et améliorer la qualité des données.

Les études récentes intègrent des techniques avancées telles que les autoencodeurs pour traiter les données avant leur utilisation dans les modèles, comme le montre l'étude de Mazurek et Wielgosz (2023) [oai\_citation:7,assessing-dataset-quality-through-decision-tree-25ue5ud9.pdf](file-service://file-9k8V5GFEGMDhOsxDBQENrX3C). De plus, des approches modernes incluent l'évaluation de la validité des codifications humaines via des modèles de machine learning, une méthode explorée par Danilov et al. (2023) [oai\_citation:8,data-quality-estimation-via-model-performance-machine-18f5ucey.pdf](file-service://file-lyM3k4Kxm3n7rXow2BpwdHge).

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

Les théories et modèles utilisés pour évaluer et améliorer la qualité des données incluent divers cadres et algorithmes. Les autoencodeurs sont couramment utilisés pour réduire la dimensionnalité des données et extraire des caractéristiques pertinentes tout en éliminant le bruit, comme démontré par Mazurek et Wielgosz (2023) [oai\_citation:9,assessing-dataset-quality-through-decision-tree-25ue5ud9.pdf](file-service://file-9k8V5GFEGMDhOsxDBQENrX3C).

Les arbres de décision sont également fréquemment utilisés pour évaluer la qualité des ensembles de données. Ces modèles permettent d'identifier les caractéristiques importantes et de mesurer l'impact des données de faible qualité sur la complexité et la performance des modèles, comme le montrent les études de Azimi et Pahl (2022) [oai\_citation:10,JDI\_2022\_Explainable-ML-Data-Quality\_CR.pdf](file-service://file-vHiSZUFV8APQFNNByKkiKSub). En outre, les techniques de redéfinition des variables cibles et de reclassification, comme celles utilisées par Danilov et al. (2023), sont essentielles pour améliorer la qualité des données avant l'entraînement des modèles de machine learning [oai\_citation:11,data-quality-estimation-via-model-performance-machine-18f5ucey.pdf](file-service://file-lyM3k4Kxm3n7rXow2BpwdHge).

En résumé, les recherches récentes et classiques convergent vers l'importance cruciale de la qualité des données pour la performance des modèles de machine learning. Les approches méthodologiques évoluent continuellement pour intégrer des techniques d'apprentissage automatique et des algorithmes sophistiqués afin de quantifier et améliorer cette qualité.

## 2. Problématiques et Lacunes Identifiées

### 2.1. Lacunes dans les Données et les Études Existantes

* Analyse des lacunes identifiées dans les recherches antérieures.
* Problèmes non résolus concernant la qualité des données et leur impact sur les performances des modèles de Machine Learning.

### 2.2. Questions de Recherche Non Résolues

* Identification des questions de recherche pertinentes qui n'ont pas encore été abordées.
* Justification de la nécessité de la recherche proposée.

## Impact de la Qualité des Données sur les Modèles de Machine Learning

### **3.1. Études Empiriques et Résultats**

* Résumé des études empiriques sur l'impact de la qualité des données (Références aux documents PDF fournis).
* Résultats et conclusions principales des études.

### **3.2. Études de Cas Spécifiques**

* Analyse détaillée de quelques études de cas spécifiques :
  + [Danilov et al., 2023]: "Data Quality Estimation Via Model Performance"​(data-quality-estimation…)​.
  + [Mazurek et Wielgosz, 2023]: "Assessing Dataset Quality Through Decision Tree Characteristics"​(assessing-dataset-quali…)​.
  + [Valtorta, 2006]: "The Effects of Data Quality on Machine Learning Algorithms"​(The\_Effects\_of\_Data\_Qua…)​.

## 4. Cadre de Référence et Méthodologies

* **4.1. Méthodologies Utilisées dans les Travaux Antérieurs**
  + Description des méthodes et approches adoptées pour évaluer et améliorer la qualité des données.
  + Comparaison des différentes approches méthodologiques.
* **4.2. 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.

# 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.