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

Objectifs de cette revue de littérature

Présentation des travaux antérieurs

* + Synthèse des études récentes et classiques sur la qualité des données.
  + Évolution des recherches dans ce domaine.
  + Présentation des théories et modèles utilisés pour évaluer et améliorer la qualité des données.

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.