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 Cadre Théorique

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.

### 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.

# 1. Revue de Littérature

## 1.1 Qualité des Données

La qualité des données est un aspect crucial dans le domaine du machine learning (ML). Elle influence directement les performances des modèles en termes de précision, de rappel et de valeur prédictive. Les dimensions de la qualité des données, telles que la complétude, l'exactitude, la cohérence et l'accessibilité, sont essentielles pour garantir des résultats fiables et robustes.

### Dimensions de la qualité des données

1. Complétude : La complétude des données se réfère à la mesure dans laquelle toutes les données nécessaires sont présentes. Par exemple, dans les systèmes IoT, la complétude des données est essentielle pour des prévisions précises, et les données manquantes peuvent réduire significativement la précision, le rappel et la valeur prédictive positive (PPV) des modèles [oai\_citation:1,JDI\_2022\_Explainable-ML-Data-Quality\_CR.pdf] [oai\_citation:2,soft\_v10\_n12\_2017\_1.pdf] Une étude a montré que l'absence de données d'un quartier peut entraîner des résultats biaisés et incorrects dans des modèles ML utilisés pour la prévision [oai\_citation:3,2406.19614v1.pdf]

2. Exactitude : L'exactitude des données concerne la mesure dans laquelle les données reflètent correctement la réalité. Des erreurs dans les données peuvent gravement compromettre la qualité des modèles ML. Par exemple, une étude a montré que l'ajout d'erreurs à un ensemble de données, telles que des valeurs manquantes ou incorrectes, peut réduire la performance des modèles de manière significative [oai\_citation:4,2406.19614v1.pdf] [oai\_citation:5,soft\_v10\_n12\_2017\_1.pdf]

3. Cohérence : La cohérence des données implique que les données doivent être uniformes et sans contradiction. Les données incohérentes peuvent entraîner des modèles instables et des prédictions peu fiables. La cohérence est particulièrement critique dans les systèmes où les données sont collectées à partir de plusieurs sources hétérogènes [oai\_citation:6,soft\_v10\_n12\_2017\_1.pdf]

4. Accessibilité : L'accessibilité des données se réfère à la facilité avec laquelle les données peuvent être obtenues et utilisées. Les restrictions d'accès aux données peuvent limiter la capacité des modèles à apprendre de manière adéquate et à produire des prédictions précises [oai\_citation:7,2406.19614v1.pdf]

### Impact de la qualité des données sur les performances des modèles

Les données de mauvaise qualité peuvent avoir plusieurs impacts négatifs sur les performances des modèles ML, notamment :

- Réduction de la précision : La précision des modèles est fortement influencée par la qualité des données d'entrée. Des données incorrectes ou incomplètes peuvent conduire à des modèles moins précis. Par exemple, une diminution de la complétude des données d'entraînement peut entraîner une diminution significative de la précision des modèles de classification et de régression [oai\_citation:8,2207.14529v4.pdf](file-service://file-6a1OcfPqNtxGcKXbF3Ssz7Ta) [oai\_citation:9,JDI\_2022\_Explainable-ML-Data-Quality\_CR.pdf]

- Diminution de la robustesse : Les modèles de machine learning doivent être robustes face aux variations des données d'entrée. Cependant, des données de mauvaise qualité peuvent rendre les modèles sensibles aux erreurs et aux variations des données, compromettant ainsi leur robustesse et leur fiabilité dans des conditions réelles [oai\_citation:10,2406.19614v1.pdf]

- Augmentation des erreurs de prédiction : Les erreurs de prédiction peuvent augmenter lorsque les données sont de mauvaise qualité. Par exemple, les modèles formés sur des données étiquetées incorrectement ou incomplètes peuvent produire des prédictions erronées, ce qui peut être critique dans des applications sensibles comme la conduite autonome ou le diagnostic médical [oai\_citation:11,2207.14529v4.pdf]

### Exemples concrets

1. Classification de maladies oculaires : Des modèles de détection de maladies oculaires entraînés sur des données sans bruit (bruit-free) peuvent échouer à détecter des images avec de petites variations si la qualité des données d'entraînement est compromise. Les erreurs de classification peuvent être aggravées par des étiquettes incorrectes, ce qui montre l'importance d'une étiquetage précis [oai\_citation:12,2406.19614v1.pdf]

2. Prévisions météorologiques : Dans une étude utilisant des données météorologiques, il a été constaté que les modèles prédictifs souffraient significativement lorsque des valeurs extrêmes ou incorrectes étaient introduites. Par exemple, l'introduction de valeurs de précipitations irréalistes a conduit à une diminution notable de la précision, de la précision et du rappel des modèles [oai\_citation:13,soft\_v10\_n12\_2017\_1.pdf]

3. Données de trafic : Une autre étude a montré que dans un ensemble de données de trafic, l'incomplétude des données (p. ex., des rangées ou des caractéristiques manquantes) affectait différemment la précision, la précision et le rappel. L'absence de rangées entières avait un impact plus important que l'absence de certaines caractéristiques, ce qui souligne l'importance de la complétude des données pour des prédictions précises [oai\_citation:14,JDI\_2022\_Explainable-ML-Data-Quality\_CR.pdf]

### Méthodes pour améliorer la qualité des données

Pour atténuer les impacts négatifs des données de mauvaise qualité, plusieurs techniques et méthodes peuvent être utilisées, notamment :

- Nettoyage des données : Le nettoyage des données consiste à identifier et à corriger les erreurs dans les ensembles de données. Cela inclut la gestion des valeurs manquantes, la correction des valeurs incorrectes et l'élimination des doublons. Par exemple, des techniques de nettoyage ont été utilisées pour améliorer les performances des modèles en réduisant les incohérences et en éliminant les erreurs d'étiquetage [oai\_citation:15,2207.14529v4.pdf]

- Imputation des valeurs manquantes : L'imputation des valeurs manquantes est une technique couramment utilisée pour gérer les données incomplètes. Elle consiste à remplacer les valeurs manquantes par des estimations basées sur les autres données disponibles. Cette méthode peut aider à maintenir la complétude des données et à améliorer les performances des modèles ML [oai\_citation:16,1-s2.0-S0950584923001222-main.pdf]

- Utilisation de modèles robustes : Certains modèles de machine learning, comme les forêts aléatoires et les modèles de régression régularisés, sont plus robustes aux erreurs dans les données d'entrée. Ces modèles peuvent mieux gérer les données de mauvaise qualité et produire des prédictions plus fiables [oai\_citation:17,2207.14529v4.pdf]

En conclusion, la qualité des données joue un rôle fondamental dans les performances des modèles de machine learning. Il est crucial de mettre en œuvre des techniques appropriées pour évaluer et améliorer la qualité des données afin de garantir des modèles fiables et efficaces. Les chercheurs et les praticiens doivent accorder une attention particulière à la qualité des données tout au long du cycle de vie du projet de machine learning pour maximiser les bénéfices des modèles développés [oai\_citation:18,soft\_v10\_n12\_2017\_1.pdf] [oai\_citation:19,The\_Effects\_of\_Data\_Quality\_on\_Machine\_Learning\_Al.pdf]

## 2. Techniques de Traitement des Données Manquantes

Les données manquantes représentent un défi important dans le traitement des DataSets pour le Machine Learning. Une gestion inappropriée des valeurs manquantes peut entraîner une dégradation significative des performances des modèles. Pour pallier ce problème, plusieurs techniques d'imputation peuvent être utilisées pour remplacer les valeurs manquantes par des estimations appropriées. Voici une description des méthodes d'imputation couramment utilisées :

### 2.1 Imputation par la Moyenne

L'imputation par la moyenne consiste à remplacer les valeurs manquantes par la moyenne des valeurs observées de la même variable. Cette méthode est simple à implémenter et permet de conserver le nombre total d'échantillons dans le dataset. Cependant, elle présente des limitations car elle ne prend pas en compte la variabilité des données et peut biaiser les résultats si les données manquantes ne sont pas aléatoires.

**Avantages** :

* Facile à implémenter.
* Permet de conserver tous les échantillons.

**Inconvénients** :

* Ne prend pas en compte la variabilité des données.
* Peut introduire un biais si les données manquantes ne sont pas aléatoires.

### 2.2 Imputation par la Médiane

L'imputation par la médiane est similaire à celle par la moyenne, mais elle utilise la médiane des valeurs observées pour remplacer les données manquantes. Cette méthode est particulièrement utile lorsque les données contiennent des valeurs aberrantes, car la médiane est moins sensible aux extrêmes que la moyenne.

**Avantages** :

* Moins sensible aux valeurs aberrantes.
* Facile à implémenter.

**Inconvénients** :

* Ne prend pas en compte les relations entre les variables.
* Peut biaiser les résultats si les données manquantes ne sont pas aléatoires.

### 2.3 Imputation par k-Nearest Neighbors (k-NN)

L'imputation par k-Nearest Neighbors (k-NN) est une méthode plus avancée qui utilise les valeurs des k observations les plus proches pour estimer la valeur manquante. Les voisins les plus proches sont déterminés en fonction de la distance entre les observations, généralement calculée par la distance euclidienne. Cette méthode peut capturer la structure locale des données et est adaptée pour les datasets où les valeurs manquantes sont liées aux valeurs observées des autres variables.

**Avantages** :

* Prend en compte la structure locale des données.
* Adaptée aux données avec des relations complexes.

**Inconvénients** :

* Plus complexe et coûteuse en termes de calcul.
* Sensible au choix du nombre de voisins k.

### 2.4 Imputation Basée sur des Modèles

L'imputation basée sur des modèles implique l'utilisation de modèles de Machine Learning pour prédire les valeurs manquantes en fonction des autres variables du dataset. Par exemple, une régression linéaire, une régression logistique ou des modèles plus complexes comme les forêts aléatoires peuvent être utilisés pour estimer les valeurs manquantes. Cette approche permet de capturer les relations entre les variables et offre des estimations plus précises, surtout lorsque les données manquantes dépendent de plusieurs variables.

**Avantages** :

* Prend en compte les relations entre les variables.
* Peut fournir des estimations précises.

**Inconvénients** :

* Nécessite une modélisation préalable.
* Peut-être coûteuse en termes de calcul et de temps.

## 3. Modèles de Machine Learning

L’objectif principal de cette section est d’explorer l’importance d’utiliser différents types de modèles de machine learning, à savoir les modèles de régression, de classification, de clustering et de deep learning, dans le cadre de l’étude sur l’impact de la qualité des données sur la performance des modèles. Il est essentiel de justifier l’utilisation de ces diverses approches pour obtenir une compréhension globale et approfondie des dynamiques complexes entre la qualité des données et les performances des modèles de machine learning.

### 3.1 Importance de la Diversité des Modèles

**Régression** : Les modèles de régression sont utilisés pour prédire des valeurs continues et sont particulièrement sensibles aux données de mauvaise qualité, telles que les valeurs manquantes ou les valeurs aberrantes. En utilisant différents modèles de régression (e.g., XGBoost, Elastic Net, Support Vector Regression), cette étude peut identifier comment chaque modèle réagit aux variations de qualité des données et déterminer les meilleures pratiques pour le traitement des données dans des contextes de régression.

**Classification** : Les modèles de classification sont utilisés pour prédire des catégories discrètes. La qualité des données, notamment l’équilibre des classes et la précision des étiquettes, joue un rôle crucial dans la performance des modèles de classification. En comparant les performances de divers algorithmes de classification (e.g., AdaBoost, Naive Bayes), cette étude peut évaluer l’impact de différentes dimensions de la qualité des données sur la précision et la robustesse des modèles de classification.

**Clustering** : Les algorithmes de clustering, tels que DBSCAN et HDBSCAN, sont utilisés pour regrouper des données non étiquetées en fonction de leurs similarités. L’intégrité et la consistance des données sont essentielles pour des clusters significatifs. Cette étude explorera comment les algorithmes de clustering réagissent aux données de qualité variable et comment la réduction de la dimensionalité peut influencer les résultats du clustering.

**Deep Learning** : Les modèles de deep learning, y compris les réseaux de neurones profonds (DNNs) et les réseaux de neurones récurrents (RNNs), sont capables de capturer des relations complexes dans les données. Cependant, ils sont également très sensibles à la qualité des données. En utilisant des modèles de deep learning, cette étude peut comparer leur robustesse et leur sensibilité par rapport aux algorithmes de machine learning plus classiques.

## 4. Évaluation des Modèles

- Métriques d'évaluation (précision, rappel, F1-score, AMI, etc.).

- Impact des hyperparamètres et des configurations de modèles.

# 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.