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

# Introduction

## 1. Contexte et Problématique

- Importance de la qualité des données dans les modèles de Machine Learning.

- Problématique de l'étude : impact de la qualité des données sur la performance et l'interprétabilité des modèles.

## 2. Objectifs du Mémoire

- Identifier les dimensions clés de la qualité des données.

- Évaluer l'impact de différentes méthodes de traitement des données manquantes.

- Comparer les performances de divers algorithmes de ML et de Deep Learning en fonction de la qualité des données.

- Proposer des recommandations pour améliorer la qualité des données dans les projets de ML.

# Chapitre 1 : Revue de Littérature

## 1. Qualité des Données

- Définitions et dimensions (completeness, accuracy, uniqueness, etc.).

- Importance de la qualité des données dans les processus d'IA.

## 2. Techniques de Traitement des Données Manquantes

- Méthodes d'imputation (moyenne, médiane, k-NN, etc.).

- Comparaison des méthodes et impact sur les performances des modèles.

## 3. Modèles de Machine Learning et Qualité des Données

- Classification, régression et clustering.

- Algorithmes de Deep Learning et leur sensibilité à la qualité des données.

## 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

- Liste des références bibliographiques utilisées (plus de dix références et au moins dix citations provenant de sources différentes).

# Annexes

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