*KHENNICHE Léon*

*ENEDIS |* [48 Bd de l'Artillerie, 69007 Lyon](https://www.google.com/maps/place/data=!4m2!3m1!1s0x47f4eb3e71f15e21:0x238b50fd34b94c9a?sa=X&ved=1t:8290&ictx=111)



*Prediction des interventions vaines*

*Mémoire de fin de cycle Ingénieurs*

École : 3iL Ingénieurs

*Parcours : I3 Cycle Ingénieurs*

*Année : 2023 - 2024*

Date de soutenance : 6 septembre 2024

Mission effectuée du 1er janvier au 6 septembre.

REMERCIEMENT

Je tiens à exprimer ma plus profonde gratitude à toutes les personnes qui ont contribué à l’élaboration de ce mémoire.

Tout d'abord, je remercie sincèrement Monsieur Adil Bajjou, mon tuteur d'entreprise, pour son accompagnement, ses précieux conseils et sa disponibilité tout au long de ce travail. Son soutien a été indispensable à l'aboutissement de ce projet.

Je souhaite également remercier les professeurs que j'ai rencontrés au cours de ma scolarité pour leur enseignement et leurs conseils qui ont enrichi ma réflexion.

Un grand merci à mes camarades de promotion, avec qui j’ai partagé des moments de réflexion, d’entraide et de convivialité. Leurs échanges et leur soutien ont été précieux tout au long de cette aventure académique.

Je tiens à exprimer toute ma reconnaissance à ma famille et à mes amis pour leur patience, leur compréhension et leurs encouragements tout au long de ce parcours. Leur soutien moral a été une source de motivation inestimable.

À tous, je vous dis un grand merci.

**TABLE DES MATIÈRES**

**PARTIE 1 : MISSION D’ENTREPRISE**

[I. Contexte de la Mission 5](#_Toc175152568)

[1.1 L'entreprise 5](#_Toc175152569)

[1.1.1 Présentation de l’entreprise 5](#_Toc175152570)

[1.1.2 Objectif de la mission 5](#_Toc175152571)

[1.2 Le Service et les Interlocuteurs 6](#_Toc175152572)

[1.2.1 Service 6](#_Toc175152573)

[1.2.2 Interlocuteurs 6](#_Toc175152574)

[1.3 L'environnement Matériel et Logiciel 7](#_Toc175152575)

[1.3.1 Matériel 7](#_Toc175152576)

[1.3.2 Logiciel 7](#_Toc175152577)

[II. Déroulement Chronologique de la Mission 8](#_Toc175152578)

[2.1 Planification Initiale et Réalisation 8](#_Toc175152579)

[2.1.1 Préparation des données 8](#_Toc175152580)

[2.1.2 Modélisation 11](#_Toc175152581)

[2.1.3 Évaluation des performances 14](#_Toc175152582)

[2.1.4 Interprétabilité des Données 16](#_Toc175152583)

[2.1.5 Industrialisation et mise à disposition des résultats via Power BI 17](#_Toc175152584)

[2.2 Recherche de nouveaux segments 18](#_Toc175152585)

[2.2.1 Méthodologie 18](#_Toc175152586)

[2.2.2 Réduction de la Dimensionnalité 19](#_Toc175152587)

[2.2.3 Résultats et Analyse 19](#_Toc175152588)

[2.3 Comparaison avec les Prévisions 19](#_Toc175152589)

[2.3.1 Problèmes Rencontrés et Solutions Apportées 19](#_Toc175152590)

[III. Résultats Obtenus et Conclusions 20](#_Toc175152591)

[3.1 Résultats Obtenus 20](#_Toc175152592)

[3.1.1 Performances du Modèle (Accuracy, Recall, Precision) 20](#_Toc175152593)

[3.1.2 Segmentation des Données et Impact sur les Résultats 21](#_Toc175152594)

[3.1.3 Présentation des Résultats par Segments 21](#_Toc175152595)

[3.2 Conclusions 21](#_Toc175152596)

[3.2.1 Efficacité du Modèle dans la Prédiction des Interventions Vaines 21](#_Toc175152597)

[3.2.2 Améliorations Possibles et Prochaines Étapes 21](#_Toc175152598)

[3.2.3 Bénéfices pour l'Entreprise : Réduction des Interventions Vaines, Amélioration de la Planification 21](#_Toc175152599)

[3.2.4 Lien avec le Sujet : Impact de la Qualité des Données sur la Performance des Modèles de Machine Learning 21](#_Toc175152600)

**PARTIE 2 : REFLEXION**

[I. Introduction 23](#_Toc175152601)

[1.1 Contexte et Problématique 23](#_Toc175152602)

[1.1.1 Contexte 23](#_Toc175152603)

[1.1.2 Problématique 23](#_Toc175152604)

[1.1.3 Importance de l'Étude 23](#_Toc175152605)

[1.2 Objectifs du Mémoire 24](#_Toc175152606)

[II. Revue de Littérature 25](#_Toc175152607)

[2.1 Introduction 25](#_Toc175152608)

[2.1.1 Objectif de la revue de littérature 25](#_Toc175152609)

[2.1.2 Importance de la qualité des données dans le contexte des modèles de machine learning 25](#_Toc175152610)

[2.2 Présentation des travaux antérieurs 25](#_Toc175152611)

[2.2.1 Synthèse des études les plus pertinentes sur la qualité des données 25](#_Toc175152612)

[2.2.2 Évolution des recherches dans ce domaine 26](#_Toc175152613)

[2.2.3 Présentation des théories et modèles utilisés pour évaluer et améliorer la qualité des données 27](#_Toc175152614)

[2.3 Problématique et lacunes identifiées 28](#_Toc175152615)

[2.3.1 Lacunes dans les Données et les Études Existantes 28](#_Toc175152616)

[2.3.2 Questions de Recherche Non Résolues 29](#_Toc175152617)

[III. Cadre Théorique et Méthodologie 31](#_Toc175152618)

[3.1 Cadre Théorique pour l’Étude Proposée 31](#_Toc175152619)

[3.1.1 Dimensions de qualité de données 31](#_Toc175152620)

[3.1.2 Description des Données 31](#_Toc175152621)

[3.2 Processus de Préparation des Données 33](#_Toc175152622)

[3.2.1 Nettoyage des données : 33](#_Toc175152623)

[3.2.2 Techniques de réduction de dimensionalité 34](#_Toc175152624)

[3.2.3 Modèles et Algorithmes 34](#_Toc175152625)

[3.2.4 Stratégies d'Optimisation et d'Évaluation 35](#_Toc175152626)

[IV. Résultats et Analyse 36](#_Toc175152627)

[4.1 Impact de la qualité des données par Dimension 36](#_Toc175152628)

[4.1.1 Completeness 36](#_Toc175152629)

[4.1.2 Consistent Representation 37](#_Toc175152630)

[4.1.3 Feature Accuracy 38](#_Toc175152631)

[4.1.4 Unicity 39](#_Toc175152632)

[4.1.5 Target Accuracy 40](#_Toc175152633)

[4.1.6 Target Class Balance 41](#_Toc175152634)

[4.2 Comparaison des Modèles 41](#_Toc175152635)

[4.2.1 Performance des Modèles 42](#_Toc175152636)

[4.2.2 Comparaison des Conclusions 42](#_Toc175152637)

[V. Discussion 43](#_Toc175152638)

[5.1 Limites de l'Étude 43](#_Toc175152639)

[VI. Conclusion 43](#_Toc175152640)

[6.1 Résumé des Constatations Clés 43](#_Toc175152641)

[6.2 Contributions du Mémoire 44](#_Toc175152642)

[6.3 Perspectives Futures 44](#_Toc175152643)

[VII. Bibliographie 44](#_Toc175152644)

[VIII. ANNEXES 46](#_Toc175152645)

[8.1 Détails visuels pour la Complétude 46](#_Toc175152646)

[8.1.1 Regression 46](#_Toc175152647)

[8.1.2 Classification 47](#_Toc175152648)

PARTIE 1 : MISSION EN ENTREPRISE

# Contexte de la Mission

## L'entreprise

### Présentation de l’entreprise

Enedis est une entreprise de service public en charge de la gestion et de la distribution de l'électricité en France. Filiale à 100% d'EDF, Enedis est chargée de gérer le réseau public de distribution d'électricité pour environ 95% du territoire français, ce qui représente plus de 1,3 million de kilomètres de lignes électriques et près de 36 millions de compteurs électriques.

Enedis a pour mission de garantir la qualité et la continuité de la distribution d'électricité, de raccorder les clients au réseau, de réaliser les travaux d'extension et de renforcement du réseau, ainsi que de relever les compteurs et de réaliser les opérations de maintenance et de dépannage.

Enedis s'engage également dans la transition énergétique en favorisant le développement des énergies renouvelables et en accompagnant les clients dans la réduction de leur consommation d'énergie.

Enedis emploie plus de 39 000 collaborateurs et est présente sur l'ensemble du territoire français. L'entreprise est engagée dans une démarche de responsabilité sociétale et environnementale en favorisant notamment la prévention des risques professionnels et la réduction de son impact environnemental.

Le domaine GDI (Gestion des interventions) fait partie du département CESE (Conduite et exploitation du système électrique) au sein du pôle PRISME de la DSI d'ENEDIS. Son but est de fournir les applicatifs bureaux et mobilités pour les équipes de planification et de programmation, ainsi que de la réalisation des interventions terrains (Réseaux et clientèle).

CEVOL pilotage, est un projet de refonte développé par la GDI regroupant à la fois, la consultation de rapport Power BI mais également, l’accès à une base de données en libre-service, la Self-BI permettant au DR de créer leur propre rapport Power BI. Ce projet BI vise à corriger les defaults que pouvait avoir les précédente solution BI (GDI historique).

### Objectif de la mission

L’objectif principal de cette étude est de participer au développement d’un modèle prédictif capable d'anticiper les interventions vaines réalisées par les techniciens de l’entreprise ENEDIS. Une intervention vaine est définie comme une intervention où le technicien n’a pas pu réaliser son travail en raison de facteurs externes. La réduction du nombre d’interventions vaines présente plusieurs avantages opérationnels et économiques.

Les objectifs spécifiques de l’étude sont les suivants :

1. Diminution du nombre d’interventions vaines :

* Réduire les déplacements inutiles des techniciens, permettant ainsi d’économiser du temps et des ressources.
* Minimiser les interruptions et les frustrations pour les clients d’ENEDIS en améliorant la précision des rendez-vous.

2. Anticipation et/ou évitement des reprogrammations :

* Développer des prévisions précises afin d’éviter la nécessité de reprogrammer les interventions, améliorant ainsi la planification des techniciens et la satisfaction client.
* Optimiser l’agenda des techniciens en limitant les interventions annulées ou reprogrammées à la dernière minute.

3. Réduction du nombre de kilomètres parcourus :

* Limiter les trajets inutiles pour les interventions vaines, contribuant ainsi à une réduction des émissions de CO2 et à des économies de carburant.
* Améliorer l’efficacité logistique et réduire l’usure des véhicules de service.

4. Amélioration du contexte de travail des techniciens et des programmateurs :

* Fournir aux techniciens des informations précises et fiables sur les interventions planifiées, réduisant ainsi le stress et l’incertitude associés aux missions annulées.
* Faciliter le travail des programmateurs en leur offrant des outils prédictifs pour mieux gérer les plannings et anticiper les besoins de ressources.

La mission vise donc à participer au développement d’un modèle de machine learning robuste, en utilisant des données historiques et des techniques avancées de modélisation, afin de fournir des prédictions fiables et d’optimiser les opérations d’ENEDIS.

## Le Service et les Interlocuteurs

### Service

La mission s'est déroulée au sein du service DSI Enedis, spécifiquement dans l'équipe FT-BI (équipe Business Intelligence de Enedis). Cette équipe a pour responsabilité de développer des rapports d’analyse avancés en utilisant PowerBI. Elle est également chargée de gérer et de maintenir l’alimentation de la base de données Self-BI, un Data Warehouse en libre-service. Ce système permet aux utilisateurs internes de créer et de personnaliser leurs propres rapports sous PowerBI, facilitant ainsi l'accès aux données et l'analyse autonome.

### Interlocuteurs

Les principaux interlocuteurs pour cette mission ont été :

* **Kevin BURDIN et Rached THAMRI :** Tous deux sont des consultants en Business Intelligence au sein de l’équipe FT-BI. Ils ont repris en charge le projet après le départ du précédent Data Scientist ayant commencé à travailler sur le sujet. Leur rôle a été crucial pour assurer la continuité du projet, en apportant leur expertise et en s’assurant que les objectifs et les méthodes restent alignés avec les besoins de l'entreprise.

Kevin et Rached ont joué un rôle clé en fournissant les orientations nécessaires, en validant les approches méthodologiques, et en facilitant les échanges avec les autres parties prenantes.

## L'environnement Matériel et Logiciel

### Matériel

Pour la réalisation de cette mission, nous avons du configuré et utilisé plusieurs environnements distincts pour assurer un développement sécurisé et efficace. Ces environnements incluent :

* **Environnement de Développement** **:** Utilisé pour le développement initial et les tests des scripts et modèles.
* **Environnement de Recette :** Utilisé pour les tests plus avancés et les validations des fonctionnalités avant déploiement.
* **Environnement de Pré-Production :** Permet de simuler les conditions de production et de valider les performances et l’intégrité des solutions développées.
* **Environnement de Production :** Lieu final de déploiement des solutions validées pour une utilisation réelle et opérationnelle.

Ces environnements sont hébergés sur des machines virtuelles (VM) sous Windows Server, accessibles via une connexion Bureau à Distance, garantissant une flexibilité et une sécurité optimales pour le travail sur les différentes étapes du projet.

### Logiciel

L'ensemble du projet a été développé en utilisant un stack technologique moderne et efficace, principalement basé sur le langage Python. Voici les principaux outils et technologies utilisés :

* **Langage Python :** Langage principal utilisé pour le développement des scripts et des modèles de machine learning.
* **Anaconda :** Distribution de Python permettant de gérer les environnements virtuels et d'installer facilement les dépendances nécessaires.
* **Jupyter Notebook :** Outil utilisé pour le développement interactif et l'expérimentation des modèles, facilitant l'écriture et l'exécution de code en cellules.
* **Librairies Python** :
  + Pandas : Utilisée pour la manipulation et l'analyse des données.
  + Scikit-Learn : Utilisée pour le développement des modèles de machine learning.
  + XGBoost : Utilisée pour les algorithmes de gradient boosting afin d’améliorer les performances des modèles prédictifs.
* **Visual Studio Code :** Environnement de développement intégré (IDE) utilisé pour l’écriture du code avec les extensions Python et Jupyter, permettant une intégration fluide entre le développement de scripts et les notebooks interactifs.
* **GitLab :** Utilisé pour la gestion du code source et le contrôle de version. Grâce à GitLab, nous avons pu assurer un suivi rigoureux des modifications apportées au code, collaborer efficacement avec l'équipe, et gérer les branches et les merges de manière structurée.

# Déroulement Chronologique de la Mission

## Planification Initiale et Réalisation

### Préparation des données

La préparation des données est une étape cruciale dans tout projet de Data Science. Pour le projet de prédiction des interventions vaines chez ENEDIS, cette étape comprend l'extraction des données, leur filtrage, et leur normalisation. Cette étape a été effectuée principalement par le précédent Data Scientist..

#### Extraction des données

Pour cette étude, les données disponibles incluent un historique des interventions depuis 2018, totalisant environ 7 millions d'interventions (85 % réalisées et 15 % vaines). Cependant, l'analyse se concentre sur les interventions effectuées entre octobre 2021 et janvier 2024, représentant environ plus d'un million d'interventions, dont 75 % réalisées et 25 % vaines. Ce sous-ensemble de données permet une analyse ciblée et pertinente, alignée avec les objectifs de l'étude, tout en facilitant la gestion et l'optimisation des ressources pour le développement du modèle prédictif.

Les données sont extraites de différentes sources internes à ENEDIS, notamment le Data Warehouse (DWH) et l'Operational Data Store (ODS). Voici les principales étapes d'extraction :

**1. Requêtes sur la DWH et l'ODS :**

* **Une requête sur la DWH** pour collecter la majorité des informations nécessaires.
* **Trois requêtes sur l'ODS** pour récupérer des informations complémentaires sur les interventions non disponibles dans la DWH.
* **Une requête supplémentaire sur l'ODS** pour récupérer les informations des techniciens.
* **Récupération des informations INSEE** des communes pour enrichir les données avec des informations socio-démographiques.

**2. Fusion des données :**

* Les données extraites sont converties en DataFrames.
* Un processus de fusion (merge) est réalisé pour assembler toutes les informations en un unique DataFrame.

#### Filtrage des données

Après l'extraction, un filtrage rigoureux est appliqué pour s'assurer que seules les données pertinentes sont retenues pour l'analyse :

**1. Restriction temporelle :**

* Le jeu de données est restreint aux informations des deux premières années pour cibler l'analyse.

**2. Intégration des données externes :**

* Les données INSEE, provenant d'un fichier Excel, sont intégrées au DataFrame principal via un merge pour enrichir les informations disponibles.

**3. Critères spécifiques :**

* Les interventions sont filtrées pour exclure les cas où la présence du client n'est pas obligatoire et le compteur n'est pas accessible.
* Une nouvelle colonne ‘nb\_vain\_prev’ est ajoutée pour indiquer le nombre d'interventions infructueuses précédentes.

**4. Épuration des données :**

* Certaines colonnes jugées inutiles sont supprimées pour alléger le jeu de données.
* Les sous-activités et catégories moins représentées sont regroupées sous la catégorie « Autres » pour simplifier l'analyse.

#### Normalisation des données

La normalisation est l'étape où les données sont transformées pour s'assurer qu'elles sont dans un format approprié pour la modélisation :

**1. Remplacement de valeurs :**

* Certaines valeurs du jeu de données sont remplacées en utilisant un dictionnaire spécifique pour améliorer la cohérence des données.

**2. Typage des données :**

* Les types de données sont ajustés pour s'assurer qu'ils correspondent aux besoins du modèle.

**3. Traitement des valeurs manquantes :**

* Les valeurs indiquées comme « à définir » sont remplacées par -1 pour maintenir l'intégrité numérique des données.
* Les interventions avec des valeurs manquantes significatives sont supprimées.

**4. Binarisation :**

* Les modalités des variables catégorielles sont remplacées par des valeurs numériques (0 ou 1) pour faciliter leur interprétation par les modèles de machine learning.

**5. Organisation et tri :**

* Les données sont triées par date pour faciliter les analyses temporelles.

Ces étapes de préparation des données permettent de transformer les données brutes en un jeu de données propre et structuré, prêt pour la modélisation et l'analyse prédictive.

#### Sélection des Variables Importantes

Pour améliorer les performances du modèle de prédiction des interventions vaines, nous avons utilisé plusieurs techniques de sélection des variables. Ces techniques nous ont permis d'identifier et de conserver uniquement les variables les plus pertinentes, réduisant ainsi la complexité du modèle et améliorant sa robustesse. Voici une explication claire et technique des méthodes utilisées :

##### Information Mutuelle

L'information mutuelle mesure la dépendance entre chaque variable explicative et la variable cible (intervention vaine ou réalisée). Elle nous aide à identifier quelles variables contiennent le plus d'information pertinente pour prédire le résultat de l'intervention. En utilisant cette méthode, nous avons pu classer les variables en fonction de leur contribution informative et sélectionner les plus importantes.

##### Test du Chi-Carré

Le test du chi-carré est utilisé pour évaluer l'importance des variables catégorielles binaires. Cette méthode compare les fréquences observées et attendues des résultats pour déterminer si une variable est significativement liée à la variable cible. Les variables avec des p-values faibles sont considérées comme importantes.

##### Variance Threshold

La méthode de la variance threshold consiste à éliminer les variables dont la variance est trop faible, car elles apportent peu d'information discriminante. En évaluant la variance de chaque variable, nous avons pu supprimer celles qui varient trop peu pour être utiles dans la prédiction.

##### Analyse de la Variance (ANOVA)

L'ANOVA est utilisée pour les variables continues. Elle permet de comparer les moyennes de différentes groupes (par exemple, interventions vaines et interventions réalisées) pour voir si elles diffèrent significativement. Les variables avec des p-values faibles sont considérées comme ayant un impact significatif sur la variable cible.

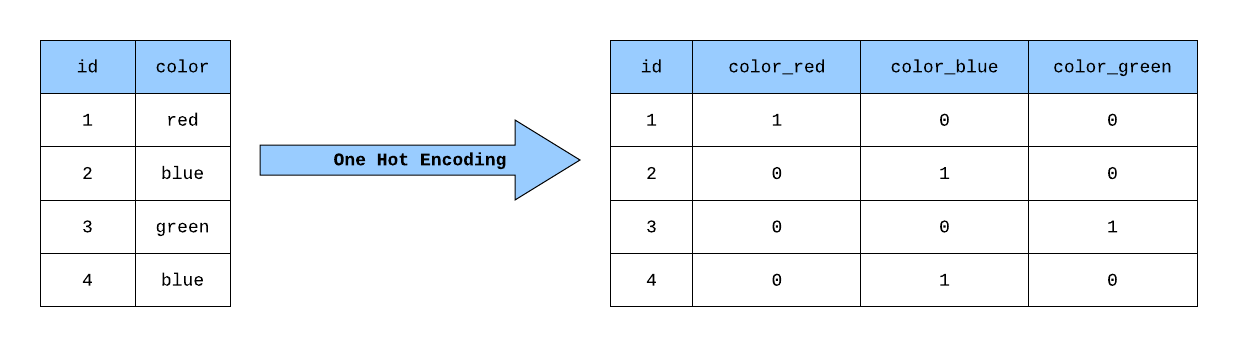
##### Importance des caractéristiques

Enfin, l'importance des caractéristiques est évaluée après l’apprentissage du modèle pour comprendre quelles variables contribuent le plus à la prédiction du modèle. Cela permet de mettre en lumière les facteurs les plus influents et peut guider des améliorations ultérieures du modèle ou de la collecte de données.

##### Fusion des Résultats et Score Global

Après avoir appliqué ces différentes méthodes, nous avons fusionné les résultats pour chaque variable. Chaque variable a été évaluée selon plusieurs critères : information mutuelle, p-value du chi-carré, variance, F-statistic de l'ANOVA, etc. Nous avons ensuite attribué un score à chaque variable basé sur le nombre de critères satisfaits. Seules les variables avec un score élevé, c'est-à-dire celles qui satisfont plusieurs critères d'importance, ont été retenues.

##### Réduction du Nombre de Variables

Grâce à ce processus rigoureux, nous avons pu réduire le nombre de variables explicatives de 2500 (en raison du one-hot encoding des variables catégorielles) à environ 700. Toutefois, cette réduction n'a pas eu l'effet escompté sur la performance du modèle. Bien qu'elle n'ait pas amélioré la précision des prédictions, elle n'a pas non plus diminué ses performances. En somme, la sélection des variables a rendu le modèle plus simple à interpréter et à maintenir sans impacter négativement ses capacités prédictives. 

Fonctionnement du One Hot Encoding

### Modélisation

La phase de modélisation dans un projet de Data Science consiste à sélectionner les algorithmes appropriés, à entraîner les modèles sur les données préparées, et à évaluer leurs performances. Voici les étapes suivies pour la modélisation des interventions vaines chez ENEDIS :

A diagram of a test

Description automatically generated

Shéma de fonctionnement d'un projet Data Science

#### Sélection d'algorithmes

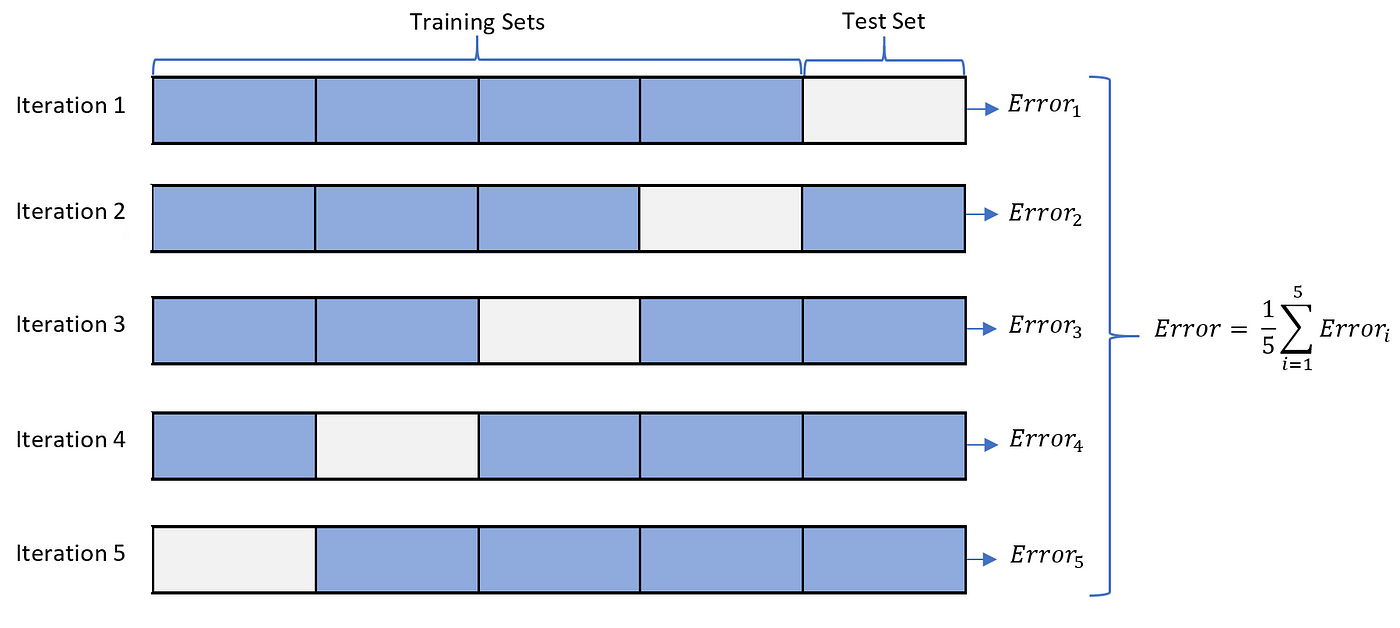
Pour prédire les interventions vaines, plusieurs algorithmes de classification ont été envisagés et testés. En plus de XGBoost (eXtreme Gradient Boosting), d'autres modèles de Gradient Boosting tels que CatBoost et LightGBM ont également été expérimentés. Cependant, bien que ces derniers aient montré des performances satisfaisantes, XGBoost s'est révélé être le plus performant dans notre contexte. En raison de sa robustesse et de ses résultats supérieurs, le choix final s'est porté sur cet algorithme. XGBoost, en tant qu'algorithme de Gradient Boosting, utilise une série d'arbres de décision pour corriger les erreurs des arbres précédents, améliorant ainsi les performances globales du modèle. Il inclut également des mécanismes de régularisation pour prévenir le surapprentissage, ce qui est crucial pour obtenir un modèle généralisable.



Fonctionnement du XGBoost

#### Entraînement des modèles

Une fois l'algorithme sélectionné, le modèle a été entraîné sur les données préparées. Plutôt que de simplement diviser les données en deux ensembles (ensemble d'apprentissage et ensemble de test), une validation croisée a été utilisée pour évaluer de manière plus rigoureuse les performances du modèle. Les données ont été divisées en 5 sous-ensembles et le modèle a été entraîné et évalué sur ces différents sous-ensembles, ce qui permet de mieux estimer sa capacité à généraliser sur des données non vues auparavant. À chaque itération, le modèle a été entraîné sur 80% des données et testé sur le reste, avec une rotation des sous-ensembles pour que chaque portion des données serve à la fois pour l'entraînement et pour le test.



Fonctionnement de la Cross-Validation

Ensuite, le modèle XGBoost a été initialisé avec des paramètres par défaut adaptés au problème de classification et a été entraîné sur l'ensemble des données d'entraînement dans chaque itération de la validation croisée, ajustant ses paramètres internes pour minimiser les erreurs de prédiction. Les résultats de chaque itération ont été agrégés pour obtenir une estimation plus robuste des performances du modèle.

#### Équilibrage des Classes

|  |  |
| --- | --- |
| Nombre d'intervention vaines | Nombre d'intervention réalisée |
| 112 942 | 272 788 |

Étant donné que les données des interventions vaines sont déséquilibrées par rapport aux interventions réalisées, nous avons mis en place des techniques d'équilibrage pour améliorer la performance du modèle :

Pour évaluer de manière robuste la performance du modèle, nous avons utilisé la validation croisée stratifiée. Comme citer précédemment, Cette technique divise les données en plusieurs sous-ensembles tout en conservant la proportion de chaque classe dans chaque pli de validation. Cela permet de s'assurer que chaque pli est représentatif de la distribution des classes et que le modèle est évalué de manière équitable sur des données équilibrées.

#### Optimisation des hyperparamètres

Pour améliorer encore les performances du modèle, une recherche d'hyperparamètres a été effectuée. Cela inclut l'ajustement des paramètres de l'algorithme pour trouver la combinaison optimale qui maximise les performances. Une fois le modèle initial entraîné et évalué, plusieurs cycles d'amélioration ont été entrepris pour affiner les résultats.

#### Segmentation des données

La décision de segmenter les données pour améliorer la précision des prédictions a été prise par le précédent Data Scientist. Ce dernier a opté pour une segmentation des interventions en groupes homogènes partageant des caractéristiques similaires, permettant ainsi de créer des modèles de prédiction plus précis. Cette approche a l'avantage d'ajuster les modèles pour refléter les spécificités de chaque segment, plutôt que d'essayer d'appliquer une règle générale à un ensemble de données très varié. À ce jour, un segment a été identifié : les interventions où la présence du client n'est pas obligatoire et où le compteur est inaccessible. Bien que cette segmentation ait permis de créer un modèle de prédiction plus précis, elle s'applique à un périmètre plus restreint, représentant 25% des interventions totales.

### Évaluation des performances

L'évaluation des performances est une étape cruciale dans le processus de modélisation, car elle permet de mesurer l'efficacité du modèle à répondre au problème posé. Pour le projet de prédiction des interventions vaines chez ENEDIS, plusieurs métriques et techniques ont été utilisées pour évaluer les performances du modèle.

#### Principales métriques utilisées

Les principales métriques utilisées pour évaluer les performances du modèle incluent l'accuracy (précision), le recall (rappel), la précision et le F1-Score. L'accuracy mesure le nombre de prédictions correctes faites par le modèle divisé par le nombre total de prédictions. Le recall indique la capacité du modèle à identifier correctement tous les cas positifs réels, tandis que la précision mesure la proportion de vrais positifs parmi les prédictions positives faites par le modèle.

* **Accuracy (Précision globale) :**

L'accuracy est la mesure la plus intuitive pour évaluer la performance d'un modèle. Elle est définie comme le nombre de prédictions correctes faites par le modèle divisé par le nombre total de prédictions. Cette métrique donne une vue d'ensemble de la performance du modèle, mais peut être trompeuse en cas de classes déséquilibrées.

* **Précision :**

La précision mesure la proportion de vrais positifs parmi les prédictions positives faites par le modèle. Elle est cruciale lorsque le coût des faux positifs est élevé. Une haute précision signifie que lorsqu'un modèle prédit un cas positif, il est très probablement correct.

* **Recall (Rappel) :**

Le recall est une mesure qui indique la capacité du modèle à identifier correctement tous les cas positifs réels. Autrement dit, parmi tous les cas réellement positifs, combien le modèle a-t-il réussi à identifier ? Le recall est particulièrement important dans les situations où il est crucial de détecter tous les cas positifs, même au risque de faire quelques erreurs positives (faux positifs).

* **F1-Score :**

Le F1-Score est la mesure harmonique entre la précision et le rappel. Il est utilisé pour évaluer l'équilibre entre ces deux métriques, en particulier dans les situations où il y a un compromis entre la précision et le rappel. Le F1-Score est particulièrement utile lorsque les données sont déséquilibrées, car il prend en compte à la fois les faux positifs et les faux négatifs. Une valeur élevée du F1-Score indique que le modèle a une bonne balance entre la précision et le rappel, ce qui en fait une mesure complète pour évaluer la performance globale du modèle.

A screenshot of a computer screen

Description automatically generated

Performance du modèle

#### Matrice de confusion

La matrice de confusion est un outil utile pour visualiser les performances du modèle. Elle compare les valeurs prédites par le modèle aux valeurs réelles, en décomposant les résultats en quatre catégories : vrais positifs, vrais négatifs, faux positifs et faux négatifs. Cela permet de comprendre non seulement la performance globale du modèle, mais aussi où il fait des erreurs.

Matrice de confusion

#### Courbe ROC et AUC

La courbe ROC (Receiver Operating Characteristic) et la valeur AUC (Area Under the Curve) sont également utilisées pour évaluer les performances du modèle. La courbe ROC trace le taux de vrais positifs (recall) en fonction du taux de faux positifs, tandis que l'AUC quantifie cette courbe. Une AUC proche de 1 indique un modèle performant, tandis qu'une AUC proche de 0.5 indique une performance proche du hasard.

#### A graph with a line Description automatically generated

Courbe ROC et AUC

#### Précision moyenne et courbe de rappel-précision

La précision moyenne est calculée à partir des prédictions et permet d'évaluer la balance entre la précision et le rappel. La courbe de rappel-précision est utile pour comprendre comment ces deux métriques évoluent en fonction du seuil de classification choisi. Elle aide à trouver le bon équilibre entre détecter le plus de vrais positifs possibles tout en minimisant les faux positifs.

A graph on a screen

Description automatically generated

Courbe Rappel-Précision

### Interprétabilité des Données

L'une des problématiques majeures rencontrées lors de l'utilisation des rapports générés était l'absence de visualisation claire des variables ayant un impact significatif sur les prédictions, notamment pour les classes d'intervention qualifiées de "vaines". Les utilisateurs bénéficiaient d'une représentation visuelle de l'importance globale des variables pendant l'apprentissage, mais il leur manquait des outils pour comprendre quelles variables spécifiques influençaient la prédiction d'une intervention vaine. Cette lacune empêchait les utilisateurs de déterminer les actions à entreprendre pour éviter qu'une intervention ne soit classée comme vaine et qu'elle soit plutôt réalisée avec succès.

Face à cette limitation, il était essentiel de fournir une méthode permettant de rendre ces prédictions plus transparentes et interprétables. Après des recherches approfondies, l'utilisation de l'outil SHAP (SHapley Additive exPlanations) a été adoptée. SHAP permet de décomposer les prédictions des modèles de machine learning en attribuant à chaque variable une valeur indicative de sa contribution positive ou négative à la décision finale. Cette approche fournit non seulement une explication globale des facteurs influençant les prédictions, mais permet également d'identifier les variables spécifiques qui ont le plus impacté les résultats pour chaque prédiction individuelle. Cela offre aux utilisateurs une meilleure compréhension des modèles et leur permet d'ajuster les interventions de manière plus informée afin de réduire le nombre d'interventions vaines.

IMAGE SHAP + RESULTAT IMAGE RAPPORT

### Industrialisation et mise à disposition des résultats via Power BI

L’industrialisation et la mise à disposition des résultats sont des étapes cruciales pour transformer les insights obtenus à partir des modèles de Data Science en actions concrètes. Pour le projet de prédiction des interventions vaines chez ENEDIS, ces étapes ont été soigneusement planifiées et exécutées pour garantir que les résultats soient accessibles et utiles aux parties prenantes.

#### Industrialisation

Le processus d’industrialisation implique deux scripts principaux, chacun ayant un rôle spécifique dans l'entraînement et l'exploitation du modèle.

**Script d’entraînement du modèle**

Ce script est conçu pour entraîner le modèle de prédiction en utilisant uniquement les interventions labellisées. L'entraînement du modèle est une opération qui n'est pas effectuée quotidiennement mais est relancée manuellement si le modèle montre des signes de perte de fiabilité au fil du temps. Cette approche permet de maintenir la robustesse et l'exactitude du modèle sans nécessiter un réentraînement constant.

**Script d’Alimentation des tables Self-BI**

Le deuxième script, écrit en Python, est exécuté quotidiennement pour alimenter plusieurs tables utilisées dans le rapport PowerBI. Ce script se lance automatiquement chaque matin à 9h via un scheduler Windows, assurant ainsi que les données sont toujours à jour. Il met à jour trois tables spécifiques :

* **Prédiction** **:** Cette table comprend les interventions labellisées et non labellisées avec les résultats de la prédiction du modèle.
* **Importance\_variable :** Cette table contient l'importance des variables au global pour toutes les interventions, fournissant une vue d'ensemble sur les facteurs influençant les prédictions.
* **Explication\_prediction :** Pour chaque intervention, cette table répertorie les cinq variables qui ont le plus impacté la prédiction du modèle vers la classe "vaine".

#### Mise à disposition des résultats via PowerBI

Pour rendre les résultats accessibles et exploitables, un rapport PowerBI a été créé et est rafraîchi quotidiennement sur la plateforme mutualisée PowerBI. Ce rapport se concentre sur le segment des interventions où la présence du client n'est pas obligatoire et le compteur est inaccessible.

Le rapport PowerBI présente divers onglets pour offrir une vue détaillée et segmentée des données :

* **Suivi de la réalisation sur le segment :** Cet onglet permet de suivre l'état d'avancement des interventions sur le segment spécifique.
* **Taux d'intervention vaines et réalisées par entité :** Ici, les taux d'interventions vaines et réalisées repérées par le modèle sont affichés, répartis par entité.
* **Poids des différentes variables et de leur modalité :** Cet onglet montre l'importance et l'impact des différentes variables utilisées par le modèle, permettant d'identifier les facteurs clés influençant les prédictions.
* **Tableau de détails des interventions :** Ce tableau répertorie les interventions avec les cinq variables ayant le plus impacté la prédiction vers la classe "vaine", offrant une transparence et une compréhension approfondie des raisons derrière chaque prédiction.

## Recherche de nouveaux segments

Etant donnée le perimetre trop limité du segment (25% pourcent du perimetre total). Nous avons exploré la possibilité d'identifier de nouveaux segments au sein des données en utilisant diverses techniques de clustering. L'objectif était de découvrir des groupes homogènes au sein des 14 millions de lignes de données, ce qui pourrait potentiellement révéler des segments non détectés jusque-là et offrir une base pour des analyses et des actions plus ciblées piur ainsi augmenter le perimetre.

### Méthodologie

Pour accomplir cette tâche, plusieurs algorithmes de clustering ont été testés, notamment KMeans, DBSCAN, et HDBSCAN. Voici un résumé des approches utilisées :

* **KMeans** : Nous avons débuté avec l'algorithme KMeans, reconnu pour sa simplicité et son efficacité dans des scénarios où les clusters sont relativement bien définis et de taille similaire. Toutefois, KMeans nécessite de définir le nombre de clusters à l'avance, ce qui peut être un désavantage lorsque ce nombre est inconnu ou variable.
* **DBSCAN** : Ensuite, nous avons exploré l'algorithme DBSCAN, qui n'exige pas de prédéfinir le nombre de clusters. DBSCAN est capable de détecter des clusters de forme irrégulière et de gérer efficacement les outliers. Cependant, dans notre contexte de grande volumétrie de données, cet algorithme s'est avéré peu efficace pour identifier des segments clairs.
* **HDBSCAN** : Enfin, l'algorithme HDBSCAN, une version hiérarchique de DBSCAN, a été mis en œuvre. HDBSCAN permet de détecter des clusters avec une plus grande complexité et est particulièrement adapté à de grands volumes de données. Néanmoins, l'application de HDBSCAN a généré plus de 1600 clusters, ce qui, bien que révélateur de la diversité des données, a posé des défis en termes d'interprétation et d'exploitation des résultats.

### Réduction de la Dimensionnalité

Pour faciliter l'analyse et améliorer la performance des algorithmes de clustering, une réduction de la dimensionnalité a été effectuée avant l'application de ces algorithmes. Nous avons d'abord utilisé la méthode PCA (Principal Component Analysis) pour réduire les dimensions à un niveau gérable (10 dimensions). Ensuite, UMAP (Uniform Manifold Approximation and Projection) a été appliqué pour conserver la structure globale et locale des données tout en réduisant davantage la dimensionnalité.

### Résultats et Analyse

Malgré ces efforts, les résultats du clustering n'ont pas été à la hauteur des attentes. Le score silhouette moyen obtenu était de -0.4221, indiquant une mauvaise définition des clusters et un manque de séparation claire entre les groupes identifiés. Cela a mis en évidence les difficultés à segmenter ces données de manière significative en utilisant les techniques testées.

A screen shot of a computer screen

Description automatically generated

Visualisation des Clusters

## Comparaison avec les Prévisions

### Problèmes Rencontrés et Solutions Apportées

Lors de la mise en œuvre du modèle de prédiction des interventions vaines chez ENEDIS, plusieurs défis majeurs ont été identifiés, nécessitant des ajustements et des solutions adaptées pour améliorer les performances globales du modèle.

#### **Déséquilibre des Classes**

**Problème :** Le jeu de données était fortement déséquilibré, avec une proportion significativement plus élevée d'interventions réalisées par rapport aux interventions vaines. Cela a créé un biais dans l'apprentissage du modèle, le poussant à privilégier la classe majoritaire, au détriment de la précision des prédictions pour la classe minoritaire (interventions vaines).

**Solution :** Pour contrer ce déséquilibre, des techniques de rééchantillonnage ont été mises en place. Cela comprenait l'utilisation de méthodes comme le suréchantillonnage des classes minoritaires et le sous-échantillonnage des classes majoritaires. De plus, la validation croisée stratifiée a été utilisée pour s'assurer que chaque pli de validation représentait fidèlement la distribution des classes, ce qui a permis d'améliorer la robustesse des résultats du modèle.

#### **Difficulté à Trouver de Nouveaux Segments**

**Problème :** Le périmètre initial de la segmentation des données était limité, ce qui a restreint la capacité du modèle à identifier des segments supplémentaires pertinents au sein du jeu de données. Cela a conduit à une couverture insuffisante des cas d'interventions vaines, limitant ainsi l'efficacité du modèle.

**Solution :** Pour pallier cette difficulté, diverses techniques de clustering ont été explorées, telles que KMeans, DBSCAN, et HDBSCAN. Bien que ces méthodes n'aient pas toujours permis de dégager des segments clairs, elles ont offert une base pour ajuster les approches futures. Une réduction de la dimensionnalité a également été effectuée via l'utilisation de techniques telles que l'UMAP pour faciliter l'identification de clusters potentiels.

#### **Clustering pas au Point Malgré plusieurs Essais**

**Problème :** Malgré de nombreux essais, les techniques de clustering n'ont pas produit de résultats satisfaisants. Le score silhouette moyen obtenu était faible, indiquant un manque de définition des clusters et une faible séparation entre les groupes identifiés. Cette situation a complexifié l'interprétation des résultats et la formulation de stratégies basées sur ces clusters.

**Solution :** Afin d'améliorer les résultats du clustering, une approche plus rigoureuse de la réduction de dimensionnalité a été adoptée. En complément, des ajustements dans les paramètres des algorithmes de clustering et l'exploration de nouvelles techniques de segmentation ont été envisagés pour optimiser les performances futures du modèle.

# Résultats Obtenus et Conclusions

## Résultats Obtenus

### **Performances du Modèle (Accuracy, Recall, Precision)**

Le modèle développé a montré des performances satisfaisantes en termes de précision globale (accuracy), de rappel (recall) et de précision (precision), bien que celles-ci aient varié en fonction des segments et des techniques d’équilibrage des classes. Le modèle XGBoost, en particulier, a offert les meilleures performances en exploitant les capacités du gradient boosting pour corriger les biais présents dans les données.

### **Segmentation des Données et Impact sur les Résultats**

La segmentation des données a joué un rôle clé dans l'amélioration des performances du modèle. Toutefois, la création de segments supplémentaires s’est avérée complexe, les techniques de clustering testées n'ayant pas produit des résultats concluants. Le segment initial, qui représentait 25 % du périmètre total, a cependant permis d'améliorer la précision des prédictions dans ce sous-ensemble spécifique.

### **Présentation des Résultats par Segments**

Les résultats ont été présentés par segments, avec un focus particulier sur le segment où la présence du client n’était pas obligatoire et où le compteur était inaccessible. Ce segment a démontré une amélioration significative des prédictions, mais son périmètre limité a contraint l’applicabilité du modèle à l’ensemble des interventions.

## Conclusions

### **Efficacité du Modèle dans la Prédiction des Interventions Vaines**

Le modèle a montré une efficacité certaine dans la prédiction des interventions vaines, bien que des défis subsistent, notamment en ce qui concerne l’équilibrage des classes et la segmentation des données. Les performances globales indiquent que, malgré les limitations, le modèle peut être un outil utile pour réduire le nombre d’interventions vaines, contribuant ainsi à une optimisation des ressources.

### **Améliorations Possibles et Prochaines Étapes**

**Nouvelle Segmentation :** L'identification de nouveaux segments reste une priorité pour élargir le périmètre des prédictions. Des efforts continus dans l'exploration de techniques avancées de clustering et de réduction de dimensionnalité sont recommandés.

**Intégration de Nouvelles Données :** L'intégration de nouvelles sources de données pourrait également enrichir le modèle et améliorer sa robustesse. Des données additionnelles telles que des informations météorologiques ou des données socio-économiques pourraient offrir de nouvelles perspectives pour la segmentation et la prédiction.

### **Bénéfices pour l'Entreprise : Réduction des Interventions Vaines, Amélioration de la Planification**

Les bénéfices anticipés pour l’entreprise incluent une réduction significative des interventions vaines, permettant une meilleure allocation des ressources et une réduction des coûts associés aux déplacements inutiles des techniciens. En outre, une amélioration de la planification des interventions pourrait être réalisée, augmentant la satisfaction des clients et l’efficacité opérationnelle.

### **Lien avec le Sujet : Impact de la Qualité des Données sur la Performance des Modèles de Machine Learning**

Les résultats obtenus confirment l'importance critique de la qualité des données sur les performances des modèles de machine learning. Les dimensions telles que l'équilibre des classes, la précision des caractéristiques et la complétude des données se sont révélées déterminantes. Les défis rencontrés lors de ce projet illustrent parfaitement comment une qualité de données sous-optimale peut limiter l'efficacité des modèles et soulignent la nécessité d'améliorer continuellement les processus de gestion de la qualité des données pour maximiser les bénéfices des modèles de machine learning.

PARTIE 2 : MÉMOIRE

# Introduction

## Contexte et Problématique

### Contexte

La qualité des données est un aspect fondamental du Machine Learning. En effet, l'exactitude, la complétude, la structure des données sur lesquelles les modèles de Machine Learning optimisent leurs fonctions en dépendent largement. Les entreprises et les chercheurs reconnaissent de plus en plus que la qualité des données affecte inévitablement le fonctionnement et la fiabilité des modèles prédictifs. Comme le soulignent Zhou et ses collègues (2024), **« Improving the quality of the dataset used in ML can be more efficient than enlarging the quantity. Inaccurate, biased, and incomplete datasets may cause unfair, bad-performance ML models that cannot be used in decision-making and mislead downstream artificial intelligence (AI) applications. »** **(Zhou et al., 2024, p. 1)**. C’est un élément majeur dans la mesure où le ML prend de plus en plus une importance décisive dans les décisions liées au domaine de la santé, de la finance et de l’énergie.

L’adoption accrue des technologies de Machine Learning et d’intelligence artificielle a été précédé de développements rapides dans ces domaines. Néanmoins, l’impact du déclin de la qualité des données sur les modèles de performances semble être un sujet relativement actif. L’étude actuelle aborde ce problème à travers l’objectif de recherche suivant : comment diverses dimensions de la qualité des données modélisent-elles la performance et l’interprétabilité des modèles ML.

### 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 la performance des modèles de Machine Learning est-elle influencée par la qualité des données ?**

### 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.

## Objectifs du Mémoire

#### 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 des variables explicatives et cibles, la balance des classes, l’unicité et la représentation consistante. Comprendre ces dimensions permettra de mieux évaluer l’impact des différentes techniques de nettoyage et de préparation des données.

#### 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. Étant donné que les modèles n’aiment pas les valeurs manquantes, il faudra résoudre ce problème et donc explorer différentes techniques d’imputation et de leur impact sur la fiabilité des modèles de machine Learning. Ce mémoire visera donc également 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.

#### Comparer les Performances de Divers Algorithmes de Machine Learning

Un autre objectif est de comparer les performances de divers algorithmes de Machine Learning en fonction de la qualité des données. Cette comparaison inclura des modèles de régression, de classification et de clustering.

#### 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 Machine Learning.

# **Revue de Littérature**

## **Introduction**

### **Objectif de la revue de littérature**

Cette revue de littérature a pour objectif d'examiner les études sur l'impact de la qualité des données sur la performance des modèles de machine learning. Elle cherche 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 pour le futur. La qualité des données est un aspect crucial pour 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. Des 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. Par conséquent, l'évaluation et l'amélioration de la qualité des données sont devenues des priorités pour les chercheurs et les praticiens du machine learning. **« The quality, size, and diversity of the dataset largely determine the effectiveness of a machine learning algorithm, particularly in classification tasks. » [9, p. 1]**

## **Présentation des travaux antérieurs**

### **Synthèse des études les plus pertinentes sur la qualité des données**

Les travaux sur la qualité des données ont exploré divers aspects de leur impact sur les modèles de machine learning. Ces études montrent l'importance des dimensions de qualité des données et leur influence sur la performance des modèles.

**Qi, Z., Wang, H., et Wang, A. (2021)** ont étudié les impacts des données de mauvaise qualité, telles que les données manquantes, incohérentes et conflictuelles, sur les modèles de classification et de clustering. Les auteurs ont introduit deux métriques, la sensibilité et le point d'inflexion de la qualité des données, pour évaluer la tolérance des modèles aux données de mauvaise qualité. Comme le soulignent Qi et al. (2021), **« To evaluate dirty-data impacts on models, we propose two novel metrics, sensibility and data quality inflection point. These two metrics are used to measure the fluctuation degree and the dirty-data tolerability of a model. » [3, p. 807]**. Ils ont montré que certains modèles, comme KNN et DBSCAN, étaient plus tolérants aux erreurs que d'autres, comme les réseaux bayésiens et la régression logistique.

**Budach, L., et al. (2022)** ont examiné comment six dimensions de la qualité des données, y compris la représentation cohérente, la complétude et la précision, 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. Comme le soulignent Budach et al. (2022), **« We explore empirically the relationship between six data quality dimensions and the performance of fifteen widely used machine learning (ML) algorithms covering the tasks of classification, regression, and clustering, with the goal of explaining their performance in terms of data quality. »** **[6, p. 2].** Leur étude a révélé que la pollution des données de test diminuait particulièrement les performances, et que certains modèles, comme les réseaux de neurones, montraient une meilleure tolérance aux erreurs.

**Azimi, S., et Pahl, C. (2022)** se sont concentrés sur l'impact de la complétude et de la correction des données sur les modèles de machine learning explicables. Ils ont montré que les données incorrectes avaient un impact plus important sur les performances des modèles que les données manquantes, en se basant sur des études utilisant des données IoT et des arbres de décision. Comme le soulignent **Azimi et Pahl (2022),** **« Incorrectness has a bigger effect on the accuracy than the incompleteness. The most probable reason for it is that in incompleteness the machine learning tool may ignore the missing rows or features and not engage them in the predictions and calculations, but regarding incorrectness the tool is forced to use all the values either correct or incorrect therefore it cannot control or minimize the damage to the accuracy » [7, p. 7]**.

**Gong, Y., et al. (2023)** ont proposé un cadre d'évaluation de la qualité des ensembles de données comprenant huit dimensions de qualité et 32 métriques pour évaluer chaque dimension. Ce cadre aide à améliorer la précision et l'efficacité des modèles de machine learning en fournissant une méthode complète pour évaluer la qualité des données. Comme le précisent **Gong et al. (2023),** **« We present an evaluation framework consisting of eight quality dimensions and 32 evaluation metrics, which provides a comprehensive method for assessing dataset quality and can help improve the accuracy, efficiency, and generalization ability of machine learning models » [8, p. 4]**.

**Zhou, Y., et al. (2024)** ont examiné les dimensions critiques de la qualité des données et passé en revue 17 outils d'évaluation et d'amélioration de la qualité des données, en proposant un cadre pour le développement futur d'outils open-source intégrant des tendances émergentes comme les grands modèles de langage et l'IA générative​.

### **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.

**Les premières études** se concentraient sur des problèmes fondamentaux tels que la complétude, l'exactitude et la cohérence des données. Par exemple, **Sessions et Valtorta (2006)** ont démontré que même des algorithmes avancés comme les réseaux bayésiens peuvent être sévèrement affectés par des données inexactes, mettant en lumière l'importance de la qualité des données dès les premiers stades de la modélisation. Comme le mentionnent les auteurs, **« After demonstrating how severely the PC algorithm is affected by inaccurate data, we propose four enhancements to the algorithm which incorporate data quality assessments » [4, p. 2]**. 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.

**Avec l'avènement des réseaux bayésiens et des méthodes d'inférence sophistiquées**, les chercheurs ont approfondi l'examen de l'impact de la qualité des données, intégrant des évaluations de la qualité dans les algorithmes pour améliorer la robustesse des modèles. Cela marque une transition vers une approche plus intégrée de la gestion de la qualité des données. Comme le mentionnent les auteurs, **« After demonstrating how severely the PC algorithm is affected by inaccurate data, we propose four enhancements to the algorithm which incorporate data quality assessments » [1, p. 2]**.

**L'essor des big data et des techniques de deep learning** a amplifié les enjeux liés à la qualité des données. Des études comme celle de Gudivada et al. (2017) ont souligné que les simples opérations de nettoyage ne suffisaient plus et que des cadres de gouvernance des données étaient nécessaires pour gérer la qualité tout au long du cycle de vie des données. Comme le précisent les auteurs, **« The recent emergence of big data analytics and renaissance in machine learning necessitates evaluating the suitability of relational database-centric approaches to data quality. It is no longer sufficient to simply clean the data; a more comprehensive data governance framework is required » [2, p. 2]**.

**Des recherches récentes**, comme celle de Mazurek et Wielgosz (2023), ont exploré l'utilisation d'autoencodeurs pour traiter les données avant d'évaluer leurs impacts sur les modèles. Ils ont démontré que les modèles peuvent devenir plus robustes et moins complexes lorsque les caractéristiques essentielles sont bien capturées par ces techniques de compression de données. Comme le soulignent les auteurs, **« We found that for high-quality datasets, decision trees trained on autoencoder-compressed data generally exhibited lower complexity compared to trees trained on raw data. This underlines the value of autoencoders in extracting salient features and reducing noise, which assists decision trees in making efficient, meaningful splits » [9, p. 8]**.

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

Les théories et modèles pour évaluer la qualité des données dans le machine learning incluent plusieurs dimensions clés, essentielles pour optimiser les performances des modèles.

**Budach et al. (2022)** identifient plusieurs dimensions cruciales, telles que la représentation cohérente, la complétude, l'exactitude des caractéristiques, l'exactitude des cibles, l'unicité, et l'équilibre des classes cibles. Ces dimensions sont essentielles pour assurer que les données sont suffisamment représentatives et précises pour être utilisées dans les modèles de machine learning.

**Gong et al. (2023)** proposent des dimensions complémentaires telles que l'auto-cohérence, qui garantit l'absence de contradictions internes dans les données, et l'actualité, qui vérifie que les données sont pertinentes et à jour. Ces dimensions sont cruciales pour évaluer la pertinence des données dans des contextes spécifiques.

**Zhou et al. (2024)** catégorisent ces dimensions en quatre grandes catégories : intrinsèque, contextuelle, représentationnelle, et accessibilité. Ces catégories offrent un cadre robuste pour comprendre et améliorer la qualité des données en fonction des besoins spécifiques des projets de machine learning.

Les approches empiriques sont également cruciales pour valider ces théories et modèles. Par exemple, les études ont proposé des méthodes spécifiques pour évaluer la qualité des données en fonction des tâches de machine learning, telles que l'utilisation du score F1 pour la classification ou du coefficient de détermination (R²) pour la régression. Ces méthodes permettent une évaluation plus fine de l'impact de la qualité des données sur la performance des modèles.

## Problématique et lacunes identifiées

### Lacunes dans les Données et les Études Existantes

La qualité des données est un facteur crucial influençant la performance des modèles de machine learning. Toutefois, plusieurs études récentes ont mis en évidence des limitations significatives dans ce domaine, pointant vers des lacunes spécifiques qui nécessitent une attention et une exploration plus approfondies.

#### Optimisation des hyperparamètres et architectures des modèles

**Budach et al. (2022)** soulignent que l'absence d'optimisation des hyperparamètres constitue une limitation majeure. L'étude n'a pas inclus cette optimisation pour les algorithmes testés, ce qui implique que les performances rapportées peuvent ne pas refléter les meilleures performances possibles pour chaque algorithme. De plus, l'autoencodeur utilisé dans l'étude pour les modèles de clustering est basé sur un réseau de neurones basique, sans optimisation spécifique pour sa tâche. Des architectures plus sophistiquées et une optimisation rigoureuse des hyperparamètres pourraient considérablement améliorer les résultats obtenus.

#### Traitement des données manquantes

La gestion des données manquantes est également un point critique. Dans l'étude de **Budach et al. (2022)**, des valeurs de remplacement fixes ont été utilisées pour simuler l'absence de données, ce qui pourrait influencer négativement les performances des modèles. L'exploration de différentes valeurs de remplacement et méthodes d'imputation pourrait offrir des solutions plus robustes et précises.

#### Évaluation limitée des algorithmes de clustering

L'évaluation des algorithmes de clustering dans l'étude de **Budach et al. (2022)** est limitée à la mesure AMI (Adjusted Mutual Information). Une évaluation plus complète, incluant des métriques telles que la taille du chevauchement des clusters et la distribution des tailles des clusters, fournirait une vue plus globale de leur performance.

#### Généralisation des résultats

**Azimi et Pahl (2022)** ont utilisé un nombre limité de jeux de données (trafic et météo), ce qui restreint la généralisation des résultats à d'autres types de données ou domaines d'application. De plus, cette étude se concentre exclusivement sur les arbres de décision, bien que d'autres modèles, comme les réseaux neuronaux ou les SVM (Support Vector Machines), pourraient présenter des comportements différents face à des problèmes de qualité des données.

#### Volume et nature des données

Le volume de données utilisé dans l'étude de **Azimi et Pahl (2022)** est relativement faible, ce qui peut affecter la robustesse des conclusions tirées. De plus, l'étude se limite à des données numériques, négligeant les données textuelles ou d'image. Une augmentation du volume de données et l'inclusion de différents types de données permettraient de valider les résultats et de rendre les conclusions plus robustes.

#### Normes de qualité des ensembles de données

**Gong et al. (2023)** notent un manque d'uniformité dans les métriques d'évaluation des ensembles de données, ce qui rend difficile l'obtention d'une description unifiée de l'assurance qualité. De nouvelles dimensions de qualité doivent être évaluées en fonction des caractéristiques spécifiques au domaine de l'apprentissage automatique, et des recherches supplémentaires sont nécessaires pour approfondir la mesure des métriques d'évaluation de la qualité des ensembles de données.

#### Complexité des outils de qualité des données

**Zhou et al. (2024)** mettent en évidence la complexité des termes techniques et de l'installation des outils de qualité des données (DQ), ainsi que le manque de personnalisation des règles de vérification de la qualité des données. De plus, les méthodes de mesure de la qualité des données ne sont pas clairement définies ou standardisées à travers les outils, rendant difficile la comparaison de leur efficacité. Il est également noté que peu d'outils sont spécifiquement conçus pour évaluer les problèmes de qualité des données liés aux tâches de machine learning.

#### Surveillance continue et automatisation

Une autre lacune importante identifiée par **Zhou et al. (2024)** concerne la gestion des grands volumes de données et la surveillance continue. Les outils actuels peuvent avoir des difficultés à traiter et surveiller en continu les grands ensembles de données, ce qui est crucial pour maintenir une haute qualité des données au fil du temps.

#### Modèles récents

Dans l'étude de **Budach et al. (2022)**, une série de modèles classiques ont été utilisés pour évaluer l'impact de la qualité des données sur les performances des algorithmes de machine learning. Les modèles sélectionnés couvrent trois tâches principales : la classification, la régression et le clustering. Parmi ces modèles, on trouve des algorithmes relativement simples comme la régression logistique, les arbres de décision, et des modèles plus complexes comme les réseaux de neurones multicouches (MLP). Cependant, ces modèles, bien que largement utilisés, ne représentent pas nécessairement l'état de l'art actuel dans le domaine du machine learning.

### Questions de Recherche Non Résolues

Lors de notre experience, nous nous concentrerons sur ces questions.

#### Méthodes d'Imputation des Données Manquantes

**Quelle méthode d'imputation des données manquantes (moyenne, médiane, Decision Tree etc.) offre les meilleures performances pour différents types de modèles de Machine Learning ?**

Les méthodes d'imputation des données manquantes peuvent avoir des impacts variés sur les performances des modèles. Comparer ces méthodes peut aider à identifier celle(s) qui maximisent les performances des modèles dans divers contextes.

#### Évaluation de la Qualité des Données pour le Clustering

**Quelles métriques supplémentaires peuvent être utilisées pour évaluer la qualité des clusters dans des ensembles de données polluées ?**

Les études actuelles utilisent principalement l'AMI pour évaluer les clusters. Explorer d'autres métriques pourrait fournir une évaluation plus complète des algorithmes de clustering en présence de données de qualité variable.

#### Généralisation des Résultats à Divers Volumes de Données

**Comment les conclusions tirées de l'impact de la qualité des données sur les modèles de Machine Learning varient-elles avec différents volumes de données ?**

La plupart des études se concentrent sur des DataSet a faible volume de données Étendre la recherche à des volume plus grand des données permettrait de généraliser les résultats et de proposer des solutions adaptées à divers contextes.

# Cadre Théorique et Méthodologie

## Cadre Théorique pour l’Étude Proposée

### **Dimensions de qualité de données**

Dans le cadre de cette étude, il est essentiel de prendre en compte plusieurs dimensions de qualité des données, car elles influencent directement la performance des modèles de Machine Learning. En m'inspirant de l’étude de **Budach et al. (2022)**, j'ai identifié et choisi les dimensions suivantes : **Complétude (Completeness), Représentation cohérente (Consistent Representation), Précision des caractéristiques (Feature Accuracy), Précision de la cible (Target Accuracy), Équilibre des classes cibles (Target Class Balance), et Unicité (Unicity)**.

* **Complétude (Completeness)** : Cette dimension se réfère à l'absence de valeurs manquantes dans les jeux de données. Une donnée incomplète peut introduire des biais dans les modèles et affecter négativement leur capacité de généralisation.
* **Représentation cohérente (Consistent Representation)** : La cohérence dans la représentation des données est cruciale pour éviter les incohérences telles que des variations non uniformes dans la dénomination des entités. Par exemple, une variable catégorielle représentant des lieux ne doit pas inclure des valeurs telles que "Paris", "Paris\_city", et "PARIS" simultanément.
* **Précision des caractéristiques (Feature Accuracy)** : Cette dimension mesure la fidélité des caractéristiques par rapport à leur valeur réelle. Des données imprécises peuvent mener à des prédictions incorrectes et fausser les résultats de l'analyse.
* **Précision de la cible (Target Accuracy)** : Il est essentiel que la variable cible soit précise pour assurer la validité des modèles de prédiction. Une cible erronée conduit à un modèle biaisé qui peut échouer à capturer la réalité des données.
* **Équilibre des classes cibles (Target Class Balance)** : Cette dimension concerne la distribution équilibrée des classes dans les jeux de données, ce qui est particulièrement important pour les tâches de classification. Un déséquilibre significatif peut amener les modèles à privilégier la classe majoritaire, réduisant ainsi la performance globale.
* **Unicité (Unicity)** : L'unicité fait référence à l'absence de duplicatas dans les données, ce qui est crucial pour éviter des biais dans l'apprentissage et assurer que chaque observation est traitée comme une entité distincte.

### Description des Données

#### **Présentation des datasets utilisés**

Pour mener à bien cette étude, nous sélectionnerons trois jeux de données distincts, chacun étant représentatif d'une tâche spécifique en machine learning : la régression, la classification et le clustering. Ces datasets proviennent de différentes sources fiables, et ils ont été choisis en fonction de leur pertinence, leur structure et leur richesse en termes de variables.

**Dataset pour la régression : FPS in Video Games**

* **Origine :** Ce jeu de données provient de la plateforme OpenML. OpenML est une communauté en ligne qui partage des datasets, des expériences et des résultats d’apprentissage machine.
* **Structure :** Ce dataset contient des informations sur les performances (FPS) dans différents jeux vidéo, avec un total de 400 000 lignes et 45 features, dont 31 numérique et 14 catégorique. Les variables incluent des types numériques comme les fréquences des composants et la résolution d'affichage, ainsi que des variables catégoriques comme le nom du jeu.
* **Pertinence :** Ce dataset est pertinent pour la régression car il permet de prédire les FPS (frames per second) en fonction de plusieurs paramètres techniques, ce qui est une tâche typique dans les contextes de modélisation prédictive des performances.

**Dataset pour la classification : Personal Key Indicators of Heart Disease**

* **Origine :** Ce jeu de données est extrait de Kaggle, une plateforme de compétitions de data science bien connue. Le dataset a été créé dans le cadre d'une enquête aux États-Unis pour recueillir des indicateurs de santé personnelle.
* **Structure :** Il comporte environ 250 000 lignes et 40 colonnes, avec des variables aussi bien numériques que catégoriques, telles que l'âge, le sexe, le niveau d'activité physique, etc. La variable cible est binaire, indiquant la présence ou l'absence de maladie cardiaque.
* **Pertinence :** Ce dataset est idéal pour la classification car il permet de développer un modèle prédictif pour identifier les individus à risque de maladie cardiaque, une application courante dans les domaines de la santé et de l’assurance.

**Dataset pour le clustering : Retail Transactional Dataset**

* **Origine :** Ce dataset provient également de Kaggle et comprend des transactions de détail. Il est conçu pour être utilisé dans des tâches d'analyse de la clientèle.
* **Structure :** Le jeu de données contient environ 300 000 transactions, avec des variables telles que l'identifiant du client, la date de transaction, le montant dépensé, etc. Les données sont numériques ainsi que catégoriques.
* **Pertinence :** Ce dataset est pertinent pour le clustering car il permet de segmenter les clients en groupes homogènes, ce qui est une technique fréquemment utilisée en marketing pour personnaliser les offres et optimiser les campagnes.

#### Justification du choix des datasets

Le choix de ces datasets a été guidé par plusieurs critères essentiels, tels que la diversité des types de données, la représentativité des variables, et la gestion des valeurs manquantes. Chaque dataset a été sélectionné pour offrir une richesse suffisante en termes de lignes et de colonnes, avec un nombre de features significatif (environ 30 pour chacun), ce qui reflète les conditions réelles de travail avec de grandes bases de données en entreprise.

**Diversité :** Les datasets couvrent trois types d’analyses courantes en machine learning : régression, classification et clustering. Cette diversité est cruciale pour tester la robustesse des méthodes de traitement de données sur différents types de problèmes.

**Représentativité :** Chacun des jeux de données représente un cas typique d'application en entreprise : prédiction des performances pour la régression, identification des risques de santé pour la classification, et segmentation de la clientèle pour le clustering.

**Volume de données :** Avec environ 300 000 lignes pour chaque dataset, ces jeux de données permettent de simuler des scénarios réalistes, où les défis de la volumétrie, de la diversité des variables, et de l’hétérogénéité des données sont présents.

L'objectif principal de cette sélection est de se rapprocher autant que possible des situations rencontrées en entreprise, où les datasets sont souvent volumineux, hétérogènes et nécessitent des traitements spécifiques pour assurer la qualité des modèles prédictifs.

## Processus de Préparation des Données

### **Nettoyage des données** :

Le nettoyage des données est une étape essentielle dans tout projet de machine learning, car la qualité des données influe directement sur la performance des modèles. Dans cette étude, plusieurs stratégies ont été élaborées pour traiter les valeurs manquantes et les anomalies dans les datasets. Ces stratégies visent à garantir la complétude des données tout en minimisant l'impact négatif des données de mauvaise qualité sur les performances des modèles.

#### Traitement des Valeurs Manquantes

Les datasets utilisés dans cette étude contiennent à la fois des données numériques et des données catégoriques, nécessitant des approches distinctes pour l'imputation des valeurs manquantes. Quatre stratégies d'imputation ont été mises en place, chacune combinant une méthode pour les données numériques avec une autre pour les données catégoriques :

* **Mean and Mode** : Cette stratégie consiste à remplacer les valeurs manquantes des variables numériques par la moyenne de la colonne correspondante et les valeurs manquantes des variables catégoriques par la modalité la plus fréquente.
* **Median and New** : Pour les variables numériques, les valeurs manquantes sont remplacées par la médiane, tandis que pour les variables catégoriques, une nouvelle catégorie est introduite pour indiquer les valeurs manquantes.
* **Decision Tree** : Ici, les valeurs manquantes des variables numériques et catégoriques sont imputées à l'aide de modèles d'arbres de décision, qui prennent en compte les autres variables pour prédire les valeurs manquantes.
* **Mean and New** : Cette stratégie combine l'imputation par la moyenne pour les variables numériques et la création d'une nouvelle catégorie pour les variables catégoriques.

#### Introduction de Pollutions dans les Données

Pour évaluer la robustesse des modèles face à des données de qualité variable, différentes pollutions ont été introduites dans les datasets. Ces pollutions ont été appliquées aux dimensions de qualité des données suivantes :

* **Completeness** : Des valeurs manquantes (NaN) ont été introduites de manière aléatoire dans chaque colonne des features, avec des taux de pollution variant de 0% à 80%.
* **Consistent Representation** : Pour chaque pourcentage de pollution, un sous-ensemble de lignes du DataFrame a été sélectionné aléatoirement. Les valeurs des colonnes catégoriques dans ces lignes ont été remplacées par des variantes de ces valeurs (ex. : ajout de suffixes comme \_var1, \_var2), tout en conservant l'apparence des données originales.
* **Feature Accuracy** : Pour les colonnes catégoriques, un sous-ensemble de lignes a été sélectionné aléatoirement et les valeurs ont été remplacées par d'autres valeurs aléatoires parmi les valeurs uniques de la colonne, à l'exclusion de la valeur actuelle. Pour les colonnes numériques, un bruit gaussien a été ajouté aux valeurs sélectionnées.
* **Target Accuracy** : Pour une colonne cible catégorielle, les valeurs ont été remplacées par des valeurs différentes choisies aléatoirement parmi les autres valeurs uniques disponibles. Pour une colonne cible numérique, un bruit gaussien a été ajouté aux valeurs, avec un arrondi selon la précision originale des données.
* **Target Class Balance** : La classe majoritaire a été réduite en nombre d'occurrences en appliquant un facteur de déséquilibre. Les occurrences des autres classes ont été augmentées en fonction du niveau de déséquilibre.
* **Unicity** : Des duplicatas ont été créés pour un sous-ensemble de lignes, augmentant artificiellement le nombre de lignes similaires ou identiques dans le DataFrame.

Ces étapes de pollution visent à simuler les imperfections des données dans des conditions réelles et à évaluer l'impact de ces imperfections sur les modèles de machine learning.

### **Techniques de réduction de dimensionalité**

Pour améliorer l’approche de **Budach et al. (2022)**, qui ont utilisé des autoencodeurs basiques pour la réduction de dimension dans les algorithmes de clustering sans optimisation spécifique, nous avons choisi d'utiliser UMAP (Uniform Manifold Approximation and Projection). UMAP a été sélectionné pour sa capacité à offrir une réduction de dimension plus performante et adaptée aux besoins spécifiques du clustering.

**Avantages de l'UMAP :**

* **Captation des Structures :** UMAP excelle à préserver les structures globale et locale des données. Cette caractéristique est particulièrement utile pour améliorer la qualité des clusters, car elle maintient les relations intrinsèques entre les points de données, ce qui facilite un clustering plus précis et cohérent.
* **Rapidité :** UMAP est connu pour sa rapidité de traitement, surpassant les autoencodeurs en termes de temps de calcul. Cette efficacité est cruciale, surtout lorsqu'il s'agit de travailler avec des ensembles de données volumineux, où la réduction de dimension doit être effectuée rapidement sans sacrifier la qualité des résultats.
* **Robustesse :** UMAP est moins sensible aux bruits et aux variations dans les données, ce qui permet d'obtenir des résultats plus stables et fiables. Contrairement aux autoencodeurs qui peuvent nécessiter des ajustements minutieux pour chaque application, UMAP fonctionne bien sans besoin de fine-tuning complexe, ce qui simplifie son utilisation dans divers contextes.

### Modèles et Algorithmes

Pour évaluer l'impact des données sur les modèles de machine learning, nous avons sélectionné des modèles récents, reconnus pour leurs performances, mais qui n'ont pas encore été largement étudiés dans ce contexte, nottamenent dans l'étude de **Budach et al. (2022).** De plus, nous avons utilisé ses algorithmes pour la prédiction des interventions vaines. Nous analyserons quatre modèles pour la régression, quatre pour la classification, et deux pour le clustering.

#### Modèles de régression : Random Forest, XGBoost, CatBoost, LightGBM

* **Random Forest** : Utilise une combinaison de nombreux arbres de décision pour prédire une valeur continue. Il est robuste contre le surapprentissage et gère bien les jeux de données complexes.
* **XGBoost** : Un algorithme de boosting très performant qui ajuste les prédictions en corrigeant les erreurs des arbres précédents. Efficace sur de grands datasets.
* **CatBoost** : Spécialement optimisé pour les variables catégoriques, CatBoost est performant et limite l'overfitting en ajustant les prédictions via le gradient boosting.
* **LightGBM** : Conçu pour être rapide et efficace, LightGBM traite les données volumineuses avec une complexité réduite grâce à une approche basée sur les histogrammes.

#### Modèles de classification : Random Forest, XGBoost, CatBoost, LightGBM

* **Random Forest** : Classifie les données en agrégeant les prédictions de plusieurs arbres de décision, rendant le modèle robuste et moins sujet aux biais.
* **XGBoost** : Optimise une fonction de perte pour les cibles catégoriques, souvent utilisé pour sa précision dans les tâches complexes de classification.
* **CatBoost** : Gère naturellement les variables catégoriques, ajustant les prédictions pour améliorer la précision tout en évitant l'overfitting.
* **LightGBM** : Idéal pour les données déséquilibrées, il offre une classification rapide et précise, particulièrement utile pour les grands volumes de données.

#### Algorithmes de clustering : DBSCAN, HDBSCAN

* **DBSCAN** : Identifie des clusters basés sur la densité des données, efficace pour détecter des formes complexes et gérer les outliers sans définir le nombre de clusters à l'avance.
* **HDBSCAN** : Une version hiérarchique de DBSCAN, capable de détecter des clusters plus complexes et de mieux gérer les grandes quantités de données, avec une robustesse accrue.\

### Stratégies d'Optimisation et d'Évaluation

Pour évaluer les performances des modèles de machine learning utilisés dans cette étude, nous avons choisi une approche rigoureuse basée sur la validation croisée et plusieurs métriques adaptées à chaque type de tâche (régression, classification, clustering). Contrairement à l’étude de **Budach et al. (2022)**, qui a mis en place trois stratégies distinctes (pollution des données d'entraînement, pollution des données de test, et pollution des deux), nous avons opté pour une stratégie intégrant la pollution dans l'ensemble du dataset. Cette approche reflète plus fidèlement les conditions réelles, où il est rare de disposer de données de test ou de données d'entrainement parfaitement propres.

#### Régression

Pour les modèles de régression, nous utliserons trois métriques clés :

* **Mean Square Error (MSE)** : Mesure la moyenne des carrés des erreurs entre les valeurs prédites et les valeurs réelles. C'est une métrique couramment utilisée qui pénalise fortement les grandes erreurs.
* **Mean Absolute Error (MAE)** : Calcule la moyenne des valeurs absolues des erreurs, offrant une vision claire des écarts moyens entre les prédictions et les résultats réels.
* **R² (Coefficient de détermination)** : Indique la proportion de la variance des données qui est expliquée par le modèle. Une valeur proche de 1 signifie que le modèle s'ajuste bien aux données.

#### Classification

Pour les tâches de classification, nous retiendrons quatre métriques pour évaluer les performances :

* **Accuracy** : Mesure la proportion de prédictions correctes parmi l'ensemble des prédictions effectuées par le modèle.
* **Precision** : Évalue la proportion de vraies prédictions positives parmi toutes les prédictions positives faites par le modèle. Cruciale dans les contextes où les faux positifs ont un coût élevé.
* **Recall** : Indique la capacité du modèle à détecter tous les cas positifs réels parmi les données. C'est une métrique importante lorsque les faux négatifs doivent être minimisés.
* **F1-Score** : Représente la moyenne harmonique entre la précision et le rappel, offrant un bon équilibre entre ces deux métriques, particulièrement utile en cas de classes déséquilibrées.

#### Clustering

Pour le clustering, nous avons évalué les performances à l'aide de trois métriques principales :

* **Adjusted Rand Index (ARI)** : Mesure la similarité entre les clusters obtenus sur des données propres et des données bruitées, en ajustant la similarité pour tenir compte du hasard. Une valeur élevée indique une bonne correspondance entre les clusters.
* **Stability Index (SI)** : Calculé à l'aide de l'Indice de Normalized Mutual Information (NMI), le SI évalue la stabilité des clusters en comparant la similarité entre les clusters des données propres et ceux des données bruitées.
* **Taille de chevauchement** : Cette métrique évalue à quel point les clusters se chevauchent, ce qui est crucial pour comprendre la séparation des clusters dans des données bruitées.

Ces méthodes d’évaluation, sélectionnées pour leur pertinence et leur capacité à capturer les nuances des performances des modèles dans des conditions de données réalistes, permettent une analyse exhaustive de l'impact de la qualité des données sur les résultats des modèles de machine learning.

# Résultats et Analyse

## Impact de la qualité des données par Dimension

L’analyse de l’impact de la qualité des données par dimension sur les performances des modèles de Machine Learning est cruciale pour comprendre comment les différentes caractéristiques des données influencent les résultats prédictifs. Cette section explore l'effet des dimensions spécifiques de la qualité des données, notamment la complétude, la représentation cohérente, la précision des caractéristiques, la précision de la cible, l'équilibre des classes cibles et l'unicité, sur les performances des modèles de régression, de classification et de clustering.

### Completeness

La complétude des données, c'est-à-dire l'absence de valeurs manquantes, est un facteur déterminant dans la performance des modèles. Les résultats montrent que, pour les trois premières stratégies d'imputation (Mean and Mode, Median and New, Mean and New), les modèles perdent progressivement en fiabilité à mesure que le pourcentage de données manquantes augmente. Cependant, la stratégie utilisant des arbres de décision pour prédire les valeurs manquantes semble paradoxalement améliorer les performances du modèle au fur et à mesure que le niveau de pollution augmente, suggérant que cette approche pourrait permettre de compenser partiellement les effets des données manquantes.

**Regression**

A graph of different colored lines

Description automatically generated

**Classification**

A graph of different colored lines

Description automatically generated

### Consistent Representation

L'impact de la pollution de la représentation cohérente est relativement faible pour les modèles XGBoost, CatBoost et LightGBM, tant en régression qu’en classification. Néanmoins, le modèle Random Forest a montré une sensibilité notable à cette dimension, perdant une grande partie de sa fiabilité dès les premiers 10% de pollution. Cela souligne l'importance de la cohérence dans la dénomination des entités, particulièrement pour les modèles moins robustes aux incohérences de représentation.

**Régression**

A graph with different colored lines

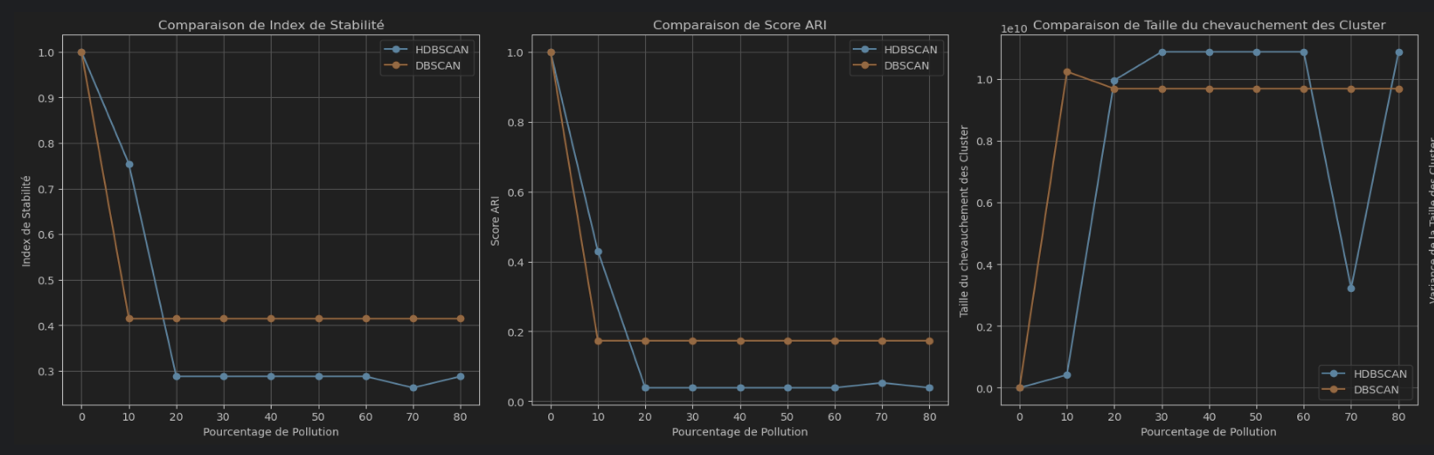
Description automatically generated

**Classification**

A graph of different colored lines

Description automatically generated

**Clustering**

****

### Feature Accuracy

La précision des caractéristiques, ou l'exactitude des variables explicatives, a un impact direct et linéaire sur les performances des modèles de régression. Dans les tâches de classification, cette dimension s'est avérée particulièrement critique, avec une chute drastique de la fiabilité des modèles dès les premiers pourcentages de pollution. Tous les modèles testés ont réagi de manière similaire, indiquant que cette dimension est universellement critique pour la qualité des prédictions. Le clustering, quant à lui, a également montré une forte dégradation de ses performances dès les 10% de pollution.

**Régression**

A graph with colored lines

Description automatically generated

**Classification**

A graph of different colored lines

Description automatically generated

**Clustering**

#### A graph with lines and dots Description automatically generated

### Unicity

L'unicité, qui fait référence à l'absence de duplicatas dans les données, a montré des résultats intéressants. Les modèles de régression ont bien réagi à cette dimension, n'étant que légèrement impactés par les duplicatas. Étonnamment, le modèle Random Forest a vu ses performances en classification s'améliorer avec l'augmentation du niveau de pollution, contrairement aux autres modèles qui ont montré une stagnation. Cela pourrait suggérer que dans certains cas, une certaine redondance dans les données peut renforcer la robustesse du modèle.

**Régression**

A graph with different colored lines

Description automatically generated

**Classification**

A graph of different colored lines

Description automatically generated

**Clustering**

**A graph on a black background

Description automatically generated**

### Target Accuracy

Dans les tâches de régression, la précision de la cible a montré un impact quasi linéaire similaire à celui observé pour la précision des caractéristiques. En classification, une particularité notable a été observée : au-delà de 50% de pollution, les performances des modèles se sont améliorées. Ce phénomène s'explique par le fait que, à 100% de pollution, les classes sont inversées, rétablissant ainsi artificiellement l’équilibre entre les classes.

**Régression**

A graph with colored lines

Description automatically generated

**Classification**

#### A graph of different colored lines Description automatically generated

### Target Class Balance

Sans surprise, les modèles obtiennent de meilleurs résultats lorsque les classes sont équilibrées. Cela est particulièrement vrai pour les modèles de Gradient Boosting, qui ont montré une plus grande sensibilité à cette dimension comparativement au modèle Random Forest. Cette observation souligne l'importance d'assurer un bon équilibre des classes pour optimiser les performances des modèles de classification.

#### **Classification**

A graph of different colored lines

Description automatically generated

## Comparaison des Modèles

Dans cette partie, nous allons comparer les performances des modèles de machine learning évalués dans notre étude à celles obtenues dans les études précédentes, tout en analysant les conclusions de chaque étude par rapport à celles tirées de notre propre expérience.

### Performance des Modèles

Les études examinées, notamment celles de **Azimi et Pahl (2022)** et de **Qi et al. (2021)**, ont évalué l'impact de la qualité des données sur divers modèles de machine learning, tels que les arbres de décision, les réseaux de neurones, et des méthodes de clustering comme DBSCAN et HDBSCAN. Dans notre étude, nous avons principalement travaillé avec des modèles de régression et de classification, tels que **Random Forest**, **XGBoost**, **CatBoost**, et **LightGBM**. Les résultats montrent que les performances de ces modèles sont fortement influencées par la qualité des données, en particulier par la complétude et la précision des caractéristiques.

**Azimi et Pahl (2022)** ont trouvé que les arbres de décision sont particulièrement sensibles à la complétude des données, avec une dégradation notable des performances en présence de données incomplètes. Ce constat est cohérent avec nos résultats, où nous avons observé une chute significative de l'accuracy et du recall lorsque les données étaient polluées par des valeurs manquantes. Cependant, notre étude montre que **LightGBM** et **CatBoost**présentent une certaine résilience par rapport à d'autres modèles comme **Random Forest**, ce qui n'est pas aussi explicitement détaillé dans les études précédentes.

**Qi et al. (2021)** ont introduit les métriques de sensibilité et de point d'inflexion de la qualité des données pour évaluer la tolérance des modèles à la pollution des données. Nos résultats montrent que les modèles de Gradient Boosting, comme **XGBoost** et **CatBoost**, maintiennent des performances relativement élevées même avec des données de faible qualité, ce qui corrobore les conclusions de Qi et al., bien que nos métriques aient été davantage axées sur l'accuracy, la précision, et le rappel.

### Comparaison des Conclusions

Les conclusions tirées de notre expérience sont en grande partie cohérentes avec celles des études antérieures. Par exemple, l'impact de la complétude et de la précision des données sur la performance des modèles est un thème récurrent. Cependant, notre étude se distingue par une analyse plus détaillée des performances spécifiques des modèles de **Gradient Boosting**. Alors que les études précédentes comme celle de **Budach et al. (2022)** mettent en avant la robustesse générale des modèles face à des variations de qualité de données, notre travail démontre que **LightGBM**, bien que performant, reste légèrement en retrait par rapport à **CatBoost** et **XGBoost** en termes de fiabilité lorsque les données sont polluées.

Enfin, bien que nos résultats corroborent en grande partie les conclusions des études antérieures, ils apportent également des précisions sur la robustesse relative des modèles face à la pollution des données, en particulier dans le contexte des modèles de **Gradient Boosting**. Nos observations suggèrent que l'amélioration de la qualité des données, notamment par des méthodes de nettoyage avancées et une meilleure gestion des valeurs manquantes, pourrait permettre de maximiser les performances des modèles de machine learning dans des contextes réels.

# Discussion

## Limites de l'Étude

**Problèmes rencontrés et limitations des méthodes utilisées**: Cette étude a rencontré plusieurs défis majeurs qui ont limité la portée des résultats obtenus. L'un des principaux problèmes a été le déséquilibre des classes dans les données, avec une surreprésentation des interventions réalisées par rapport aux interventions vaines. Cela a créé un biais dans l'apprentissage du modèle, ce qui a nécessité l'utilisation de techniques de rééchantillonnage pour corriger ce déséquilibre. Cependant, même avec ces ajustements, la performance du modèle n'a pas atteint les niveaux escomptés, ce qui souligne la difficulté de traiter des jeux de données fortement déséquilibrés. En outre, les techniques de clustering utilisées pour identifier de nouveaux segments de données n'ont pas produit les résultats attendus, avec un score silhouette faible indiquant un manque de définition claire des clusters. Cette limitation a restreint la capacité du modèle à généraliser ses prédictions à l'ensemble du jeu de données.

**Suggestions pour des recherches futures**: Pour surmonter ces limitations, des recherches futures devraient explorer des techniques plus avancées de gestion des données déséquilibrées, comme l'utilisation de méthodes de génération de données synthétiques ou l'intégration de modèles de coûts différenciés pour les classes minoritaires. De plus, des efforts supplémentaires devraient être consacrés à l'amélioration des techniques de clustering, en explorant des algorithmes plus sophistiqués ou en combinant plusieurs méthodes pour mieux capturer la structure sous-jacente des données. L'intégration de nouvelles sources de données, comme des informations socio-économiques ou météorologiques, pourrait également enrichir le modèle et améliorer sa robustesse.

#### 5.2 Recommandations

**Meilleures pratiques pour améliorer la qualité des données**: Il est crucial d'adopter des pratiques rigoureuses pour assurer la qualité des données utilisées dans les modèles de Machine Learning. Cela inclut un nettoyage minutieux des données, l'imputation des valeurs manquantes avec des techniques appropriées et l'assurance d'une représentation cohérente des variables explicatives et cibles. L'utilisation d'outils d'analyse tels que SHAP pour interpréter les contributions des différentes variables peut également aider à identifier les domaines nécessitant une amélioration de la qualité des données.

**Conseils pour la sélection et l'utilisation des modèles en fonction de la qualité des données**: La sélection du modèle doit être étroitement liée à la qualité des données disponibles. Par exemple, les modèles de Gradient Boosting comme XGBoost et CatBoost ont montré une certaine résilience face à la pollution des données, ce qui les rend particulièrement adaptés dans des contextes où la qualité des données peut être variable. Il est recommandé de tester plusieurs algorithmes et d'ajuster les hyperparamètres en fonction des caractéristiques spécifiques des données pour maximiser les performances.

# Conclusion

## Résumé des Constatations Clés

**Points saillants de l'étude et implications pratiques**: Cette étude a confirmé l'impact critique de la qualité des données sur les performances des modèles de Machine Learning. Les dimensions telles que la complétude, l'équilibre des classes et la précision des caractéristiques ont été identifiées comme étant particulièrement influentes. Les résultats obtenus montrent que l'amélioration de la qualité des données peut significativement augmenter l'efficacité des modèles, réduisant ainsi le nombre d'interventions vaines et optimisant l'allocation des ressources​(Mémoire).

## Contributions du Mémoire

**Apports à la recherche et à la pratique professionnelle en Machine Learning**: Ce mémoire contribue à la compréhension des défis posés par la qualité des données dans le Machine Learning, en offrant des recommandations pratiques pour améliorer les performances des modèles. Les résultats apportent une valeur ajoutée à la recherche en Machine Learning en démontrant l'importance de la gestion des données dans la prédiction des interventions vaines, un domaine critique pour les entreprises comme ENEDIS.

## Perspectives Futures

**Directions possibles pour des recherches ultérieures**: Les recherches futures pourraient se concentrer sur l'intégration de nouvelles dimensions de qualité des données, l'amélioration des algorithmes de clustering, et l'exploration de modèles plus robustes face aux données de faible qualité. De plus, l'application de ces techniques à d'autres domaines d'intervention pourrait fournir des insights supplémentaires sur la généralisation des résultats obtenus​(Mémoire)​(Mémoire).

# Bibliographie

1. SESSIONS, Valerie, VALTORTA, Marco, 2006. **The Effects of Data Quality on Machine Learning Algorithms.**Columbia : University of South Carolina. URL : https://arxiv.org/pdf/2207.14529
2. GUDIVADA, Venkat N., APON, Amy, DING, Junhua, 2017. **Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations.** International Journal on Advances in Software, vol. 10, no. 1 & 2, pp. 1-15. USA: IARIA. URL : https://personales.upv.es/thinkmind/dl/journals/soft/soft\_v10\_n12\_2017/soft\_v10\_n12\_2017\_1.pdf
3. QI, Zhi-Xin, WANG, Hong-Zhi, WANG, An-Jie [et al.], 2021. **Impacts of Dirty Data on Classification and Clustering Models: An Experimental Evaluation.** Harbin : Harbin Institute of Technology. URL : https://arxiv.org/pdf/1803.06071
4. SRIVASTAVA, Saket, SHAH, Rushit N., TEODORIU, Catalin [et al.], 2022. **Impact of data quality on supervised machine learning: Case study on drilling vibrations.** Norman : University of Oklahoma
5. LUCA, Andreea Roxana, URSULEANU, Tudor Florin, GHEORGHE, Liliana [et al.], 2022. **Impact of Quality, Type, and Volume of Data Used by Deep Learning Models in the Analysis of Medical Images.** Iasi : "Grigore T. Popa" University of Medicine and Pharmacy.
6. BUDACH, Lukas, FEUERPFEIL, Moritz, IHDE, Nina [et al.], 2022. **The Effects of Data Quality on Machine Learning Performance.** Potsdam : University of Potsdam. URL : https://arxiv.org/pdf/2207.14529
7. AZIMI, Shelernaz, PAHL, Claus, 2022. **The Impact of Data Completeness and Correctness on Explainable Machine Learning Models.** Bolzano : Free University of Bozen-Bolzano.
8. GONG, Youdi, LIU, Guangzhen, XUE, Yunzhi [et al.], 2023. **A Survey on Dataset Quality in Machine Learning.**Beijing : Institute of Software, Chinese Academy of Sciences.
9. MAZUREK, Szymon, WIELGOSZ, Maciej, 2023. **Assessing Dataset Quality Through Decision Tree Characteristics in Autoencoder-Processed Spaces.** Krakow : AGH University of Science and Technology.
10. DANILOV, Gleb, KOTIK, Konstantin, SHIFRIN, Michael [et al.], 2023. **Data Quality Estimation Via Model Performance: Machine Learning as a Validation Tool.** Moscow : National Medical Research Center for Neurosurgery named after N.N. Burdenko.
11. ZHOU, Yuhan, TU, Fengjiao, SHA, Kewei [et al.], 2024. **A Survey on Data Quality Dimensions and Tools for Machine Learning.** Denton : University of North Texas.

# ANNEXES

## Détails visuels pour la Complétude

### Regression

A screenshot of a graph

Description automatically generatedA graph on a black background

Description automatically generatedA graph on a black background

Description automatically generated

A screenshot of a graph

Description automatically generated

### Classification

A graph of different colored lines

Description automatically generated

A graph of different colored lines

Description automatically generatedA graph of different colored lines

Description automatically generatedA graph of different colored lines

Description automatically generated