Les sujets identifiés :

**Proposition 1 : Impact de la Qualité des Données sur la Performance des Modèles de Machine Learning**

**Réflexion**

Ce sujet explore l'influence de la qualité des données sur la performance des modèles de machine learning, en se basant sur l'expérience acquise lors du projet de prédiction des interventions vaines. Il s'agit d'analyser comment la gestion des valeurs manquantes, le nettoyage des données et la réduction des modalités ont impacté les résultats des modèles prédictifs.

**Questions à traiter**

* Quelle est l'importance de la qualité des données dans le cycle de vie d'un projet de data science ?
* Quelles leçons tirer des techniques de nettoyage et de préparation des données utilisées dans ce projet ?
* Comment la qualité des données a-t-elle influencé les performances et la robustesse des modèles de machine learning ?

**Sources et comparaison**

* Recherche bibliographique sur l'impact de la qualité des données dans d'autres projets similaires.
* Études de cas et comparaison avec des projets de data science réalisés dans d'autres entreprises.

**Proposition 2 : Segmentation des Données et Personnalisation des Modèles en Machine Learning**

**Réflexion**

Ce sujet se focalise sur la stratégie de segmentation des données et son impact sur la précision et la pertinence des modèles de machine learning. L'étude analysera comment la segmentation des données (par exemple, interventions où la présence du client n'est pas obligatoire et le compteur est inaccessible, segment découvert pour le modèle) a permis d'améliorer les résultats des prédictions dans le contexte du projet.

**Questions à traiter**

* Quels sont les avantages et les défis de la segmentation des données dans un projet de data science ?
* En quoi la segmentation a-t-elle amélioré les performances des modèles dans ce projet spécifique ?
* Quelles sont les meilleures pratiques pour effectuer une segmentation efficace des données ?

**Sources et comparaison**

* Analyse de la littérature sur les techniques de segmentation des données et leur application en machine learning.
* Comparaison avec des méthodes de segmentation utilisées dans d'autres industries ou contextes.

**Proposition 3 : Optimisation et Régularisation des Modèles de Machine Learning : Cas du Gradient Boosting**

**Réflexion**

Ce sujet examine les techniques d'optimisation et de régularisation appliquées aux modèles de machine learning, avec un focus particulier sur l'algorithme XGBoost utilisé dans le projet. Il s'agit d'évaluer comment ces techniques ont contribué à la réduction des erreurs et à l'amélioration des prédictions.

**Questions à traiter**

* Quelles sont les techniques d'optimisation et de régularisation les plus efficaces pour améliorer les modèles de machine learning ?
* Comment XGBoost utilise-t-il ces techniques et quels en sont les avantages dans le contexte de ce projet ?
* Quels sont les défis rencontrés lors de l'optimisation des modèles et comment ont-ils été surmontés ?

**Sources et comparaison**

* Revue des travaux académiques et professionnels sur l'optimisation et la régularisation des modèles de machine learning.
* Étude comparative avec d'autres projets utilisant XGBoost ou des techniques similaires.

**Idées de Points de Vue Innovants sur l'Impact de la Qualité des Données sur la Performance des Modèles de Machine Learning**

1. **Impact des Données Synthétiques sur les Modèles de Machine Learning :**
   * **Problématique :** Dans quelle mesure les données synthétiques peuvent-elles compenser les données manquantes ou de mauvaise qualité ?
   * **Exploration :** Évaluer comment les données générées artificiellement peuvent être utilisées pour améliorer la qualité des jeux de données et, par conséquent, la performance des modèles.
2. **Évaluation de la Qualité des Données en Temps Réel :**
   * **Problématique :** Comment développer des mécanismes pour évaluer et améliorer en temps réel la qualité des données entrant dans les systèmes de machine learning ?
   * **Exploration :** Étudier les techniques et les outils permettant la détection et la correction en temps réel des erreurs dans les flux de données.
3. **Impact des Biais Inconnus dans les Données :**
   * **Problématique :** Quels types de biais restent souvent non détectés dans les jeux de données et comment affectent-ils les modèles de machine learning ?
   * **Exploration :** Analyser comment les biais cachés influencent les résultats et quelles méthodes peuvent être développées pour les identifier et les atténuer.
4. **Interprétabilité des Modèles par Rapport à la Qualité des Données :**
   * **Problématique :** Comment la qualité des données influence-t-elle l'interprétabilité et la transparence des modèles de machine learning ?
   * **Exploration :** Examiner les liens entre des données de haute qualité et la capacité des modèles à fournir des explications compréhensibles sur leurs prédictions.
5. **Adaptabilité des Modèles aux Données Changeantes :**
   * **Problématique :** Dans quelle mesure la qualité des données affecte-t-elle la capacité des modèles de machine learning à s'adapter à des environnements de données changeants ?
   * **Exploration :** Étudier les méthodes permettant aux modèles de rester robustes et performants face à des changements dans les caractéristiques des données d'entrée.
6. **Rôle de la Qualité des Données dans les Applications de l'IA Éthique :**
   * **Problématique :** Comment la qualité des données peut-elle être assurée pour répondre aux préoccupations éthiques dans les applications de machine learning ?
   * **Exploration :** Discuter des implications éthiques de la qualité des données, notamment en termes de discrimination, de confidentialité et d'équité.
7. **Utilisation des Techniques de Prétraitement Automatique des Données :**
   * **Problématique :** Quels sont les impacts de l'automatisation du prétraitement des données sur la qualité des données et la performance des modèles ?
   * **Exploration :** Investiguer les outils d'automatisation qui peuvent aider à améliorer la qualité des données tout en réduisant les efforts humains.
8. **Qualité des Données et Apprentissage Profond :**
   * **Problématique :** Comment la qualité des données affecte-t-elle spécifiquement les modèles d'apprentissage profond par rapport aux modèles traditionnels de machine learning ?
   * **Exploration :** Comparer l'impact de différentes qualités de données sur des réseaux de neurones profonds versus des algorithmes plus classiques.
9. **Qualité des Données dans les Applications Multimodales :**
   * **Problématique :** Comment gérer la qualité des données dans des applications utilisant des sources de données multimodales (texte, image, audio) ?
   * **Exploration :** Étudier les défis uniques posés par la combinaison de plusieurs types de données et leurs impacts sur la performance des modèles.
10. **Techniques de Fusion de Données pour Améliorer la Qualité :**
    * **Problématique :** Comment la fusion de données provenant de différentes sources peut-elle améliorer la qualité globale des données et la performance des modèles ?
    * **Exploration :** Analyser les méthodes de fusion de données et leur efficacité pour augmenter la richesse et la qualité des jeux de données utilisés en machine learning.

Ces idées de points de vue innovants peuvent servir de base pour approfondir vos recherches et développer une réflexion originale sur l'impact de la qualité des données sur la performance des modèles de machine learning.

Dimension déjà étudier :

 **Consistent Representation**: Ensuring uniform representation of data values to avoid semantic discrepancies.

 **Completeness**: Measuring the extent of missing values in the dataset.

 **Feature Accuracy**: Assessing the deviation of feature values from their true values.

 **Target Accuracy**: Evaluating the correctness of target labels.

 **Uniqueness**: Determining the presence of duplicate records.

 **Target Class Balance**: Checking the distribution balance of target classes.

Dimension de données à étudier :

**Ponctualité** (Timeliness):

* **Description** : Évalue la fraîcheur et la pertinence des données pour leur usage prévu.
* **Exemple** : Évaluer la fraîcheur des horodatages ou des données d'événements. Par exemple, analyser si les données transactionnelles sont à jour pour des analyses en temps réel.

**Validité (Validity)** :

* **Description** : Vérifie si les données respectent les règles métier ou contraintes définies.
* **Exemple** : Valider les données catégorielles par rapport à des listes prédéfinies de valeurs acceptables. Par exemple, s'assurer que toutes les entrées dans un champ "pays" sont des noms de pays valides.

**Intégrité (Integrity)** : La qualité des relations et des dépendances entre les différentes parties des données. Des données corrompues peuvent conduire à des prédictions incorrectes.

 **Description** : Garantit l'exactitude et la fiabilité des données tout au long de leur cycle de vie.

 **Exemple** : Détecter et gérer les enregistrements en double dans le jeu de données. Assurer également l'intégrité référentielle entre les ensembles de données connexes lorsque cela est applicable.

**Uniformité (Uniformity)** : La standardisation des formats et des unités de mesure dans les données pour éviter les ambiguïtés.

Action à mener par ordre d’importance :

* Dimensions à étudier :
  + Completeness
    - Evaluation de différentes valeurs de remplacement pour identifier celles qui conduisent aux meilleures performances du modèle
    - Comparaison des valeurs de remplacement
  + Target and Feature Accuracy
  + Uniqueness
  + Target Class Balance
  + Uniformity
  + Timeliness
* Différent type de pollution/erreur
  + Soit prendre des DataSets déjà polluer ou polluer soit meme le dataset
* Etude simultané de plusieurs dimensions de qualité des données
* Modèle de ML plus complexe
* Inclure de l’optimisation des Hyper paramètre
  + Réaliser une évaluation plus détaillée pour chaque modèle d'apprentissage machine, en considérant divers hyperparamètres et configurations
  + Peut fournir des insights sur la manière dont différents modèles traitent les problèmes de qualité des données.
* Revoir la méthode pour évaluer les méthodes de Clustering
  + Réseau neuronal classique avec des couches lineaire
  + Explorer des composants de réseau neuronal au-delà des couches linéaires et intégrer la performance de clustering dans la fonction de perte de l'autoencodeur pourrait améliorer l'évaluation des algorithmes de clustering
  + Perte d’information dans la Réduction de Dimensionnalité
    - Trouver un moyen de prendre en compte la perte d’information causée par l’encodage des données de haute dimensionnalité pour un espace bidimensionnel, pour le clustering
  + Métrique d’algorithmes de Culstering
    - Utiliser non seulement l'Information Mutuelle Ajustée (AMI), mais également considérer des métriques comme la taille absolue du chevauchement des clusters de données originales et générées, ainsi que la moyenne et la variance des tailles de clusters dans les sorties des algorithmes
* Réseau Neuronal plus complexe
* Transfer Learning

DataSet :

* Regression
  + Boston Housing DataSet
  + California Housing Prices DataSet
  + Air Quality DataSet
  + Energy Efficiency DataSet
  + Auto MGP DataSet
  + Bike Sharing Demand
  + House Prices : Advanced Regression Techniques
* Classification
  + Iris DataSet
  + Titanic DataSet
  + Breast Cancer Wisconsin DataSet
  + Wine Quality DataSet
  + MNIST Handwritten Digits DataSet
  + Credit Card Fraud Detection DataSet
  + Titanic: Machine Learning from Disaster
  + Heart Disis DataSet (UCI)
  + Adult Income DataSet (UCI)
* Clustering
  + Mall Customers DataSet
  + Wholesale Customers DataSet
  + Handwritten Digits DataSet
  + Customer Personality Analysis Dataset
  + Human Activity Recognition (HAR) Using Smartphones DataSet
  + Customer Segmentation DataSet
* Intervention Vaines DataSet for Classification and Clustering