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.

* 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.
* Completness , choisir différentes méthode de remplacement des valeurs manquantes
  + Evaluation de différentes valeurs de remplacement pour identifier celles qui conduisent aux meilleures performances du modèle
  + Comparaison des valeurs de remplacement
* 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
* Dimensions supplémentaire de qualité des données
  + Elargir l’évaluation pour inclure davantage de dimensions de qualité des données au-delà des six étudiées
* 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
* Modèle de ML plus complexe
* Différent type de pollution/erreur
* Etude simultané de plusieurs dimensions de qualité des données
* Réseau Neuronal plus complexe
* Transfer Learning