Data Quality Scorecard

Auditer un dataset en 5 dimensions avant tout projet IA

Pourquoi cette grille

Un modèle IA est aussi fiable que les données qui l'alimentent.
Cette scorecard permet d'évaluer objectivement la qualité d'un jeu de données avant de lancer un projet - et d'identifier où concentrer le travail de préparation.

Les 5 dimensions de la qualité

01 Complétude

Proportion de valeurs renseignées par rapport aux valeurs attendues.

Question clé : *Quel pourcentage de champs obligatoires est rempli sur l'ensemble du dataset ?*

Exemple concret : Un CRM ou 40 % des fiches clients n'ont pas de numéro de téléphone est incomplet pour un projet de relance.

Mesure : % de valeurs non nulles sur champs obligatoires. Cible : > 95 % pour un usage en production.

02 Exactitude

Degré de conformité des valeurs par rapport à la réalité qu'elles sont censées représenter.

Question clé : *Les valeurs correspondent-elles à la réalité ? Peuvent-elles être vérifiées contre une source de référence ?*

Exemple concret : Un prix produit dans le catalogue qui ne correspond pas au prix facturé. Une adresse postale erronée.

Mesure : % de valeurs validées contre une source de référence. Ou : taux d'anomalies détectées par échantillonnage.

03 Cohérence

Absence de contradictions entre plusieurs sources ou au sein du même dataset.

Question clé : *Les mêmes entités (clients, produits…) sont-elles décrites de manière identique dans tous les systèmes ?*

Exemple concret : Le chiffre d'affaires client dans le CRM ne correspond pas à celui dans l'ERP. Le même client enregistré avec deux noms différents.

Mesure : % de concordance entre systèmes sources sur un échantillon de champs clés.

04 Fraîcheur

Âge des données par rapport au moment où elles doivent être utilisées. Aussi appelé 'timeliness'.

Question clé : *Les données sont-elles suffisamment récentes pour le cas d'usage vise ? La fréquence de mise à jour est-elle adéquate ?*

Exemple concret : Utiliser des préférences clients datant de 18 mois pour un moteur de recommandation en temps réel.

Mesure : Âge moyen des données vs. fréquence requise par le cas d'usage. Taux de données périmées par rapport à un seuil défini.

05 Unicité

Absence de doublons : chaque entité est représentée une seule fois dans le dataset.

Question clé : *Y a-t-il des enregistrements en double ? Le même client, produit ou événement apparaît-il plusieurs fois ?*

Exemple concret : 3 fiches pour le même client avec des variantes d'orthographe : 'Dupont Jean', 'Jean DUPONT', 'j.dupont'. Le modèle apprend trois fois la même personne.

Mesure : Taux de doublons détectés. Outil : OpenRefine, deduplication SQL, ou module de qualité dans Dataiku.

Grille de scoring - a remplir

Instructions : évaluer chaque dimension de 1 à 5. Calculer le score global (moyenne). Tout score < 3 sur une dimension est un signal d'alerte à traiter avant de lancer le projet.

Dimension	Observations	Score (1-5)	Priorité
Complétude		/5	
Exactitude		/5	
Cohérence		/5	
Fraîcheur		/5	
Unicité		/5	
SCORE GLOBAL	Moyenne des 5 dimensions	**/5**	

Interprétation des scores

1 - Inutilisable	Problèmes critiques. Ne pas utiliser en production. Nettoyage majeur requis avant tout projet.
2 - Faible	Nombreux problèmes identifiés. Usage possible uniquement pour exploration. Budgéter le nettoyage.
3 - Moyen	Utilisable avec précautions. Documenter les limites connues. Validation manuelle recommandée.
4 - Bon	Qualité satisfaisante pour la majorité des cas d'usage. Surveillance continue recommandée.
5 - Excellent	Données propres, bien documentées, fiables. A maintenir activement.