Documentation de l'Evidently AI Test Suite

Vue d'ensemble

L'Evidently AI Test Suite est un framework de tests automatisés pour la validation et le monitoring de la qualité des données et des modèles de machine learning. Elle permet de détecter les dérives de données (data drift), les problèmes de qualité et les anomalies dans les pipelines ML.

Table des matières

Installation

L'installation se fait via pip avec la commande pip install evidently. Pour des versions spécifiques, il est possible de préciser le numéro de version.

Concepts clés

Test Suite vs Report

Test Suite effectue des vérifications binaires (pass/fail), idéale pour les alertes et l'intégration CI/CD. Elle retourne des résultats structurés en JSON et permet l'automatisation des décisions.

Report génère des analyses visuelles détaillées, idéale pour l'exploration et le debugging. Elle produit des rapports HTML interactifs orientés vers l'analyse humaine.

Données de référence vs Données courantes

Les **données de référence** constituent la base de comparaison (généralement les données d'entraînement), tandis que les **données courantes** sont les nouvelles données à comparer avec la référence.

Architecture

La Test Suite est organisée en deux niveaux principaux. Le premier niveau comprend les **Test Presets** qui regroupent plusieurs tests thématiques comme DataDriftTestPreset, DataQualityTestPreset, DataStabilityTestPreset, RegressionTestPreset et ClassificationTestPreset. Le second niveau contient les **Individual Tests** permettant de créer des tests personnalisés spécifiques à chaque besoin.

Types de tests disponibles

1. Data Drift Test Preset

Ce preset détecte les changements dans la distribution des données entre le dataset de référence et le dataset courant.

Tests inclus :

- Dérive au niveau du dataset (proportion de colonnes déviant)
- Dérive par colonne (changements de distribution)
- Tests statistiques automatiques

Cas d'usage :

- Monitoring de production
- Détection de concept drift
- Validation de nouvelles données

---

2. Data Quality Test Preset

Ce preset vérifie la qualité intrinsèque des données.

Tests inclus :

- Détection de valeurs manquantes
- Identification de valeurs dupliquées
- Détection de colonnes constantes
- Validation des valeurs hors limites
- Vérification des types de données

Cas d'usage :

- Validation de pipelines ETL
- Détection d'anomalies de qualité
- Monitoring de l'intégrité des données

---

3. Data Stability Test Preset

Ce preset vérifie la stabilité structurelle des données.

Tests inclus :

- Vérification du nombre de colonnes
- Validation des types de colonnes
- Contrôle des noms de colonnes
- Vérification du nombre de lignes

Cas d'usage :

- Validation de schéma
- Détection de changements structurels
- Tests de régression de données

---

4. Regression Test Preset

Ce preset évalue les performances des modèles de régression en comparant les prédictions avec les valeurs réelles.

Tests inclus :

- Métriques d'erreur (MAE, RMSE, MAPE)
- Analyse des erreurs par segments
- Évaluation de la qualité des prédictions
- Distribution des erreurs résiduelles

Cas d'usage :

- Monitoring de modèles de régression en production
- Validation de performances
- Détection de dégradation du modèle

---

5. Classification Test Preset

Ce preset évalue les performances des modèles de classification.

Tests inclus :

- Métriques de classification (Accuracy, Precision, Recall, F1)
- Aire sous la courbe ROC (ROC AUC)
- Analyse de la matrice de confusion
- Distribution des classes prédites vs réelles

Cas d'usage :

- Monitoring de modèles de classification
- Validation de métriques métier
- Analyse de biais de classe

Tests statistiques et formules

Tests de dérive pour variables numériques

Test de Kolmogorov-Smirnov (KS)

Le test KS mesure la distance maximale entre les fonctions de répartition empiriques de deux échantillons.

Statistique KS :

$$D_{KS} = \sup_x |F_{\text{ref}}(x) - F_{\text{curr}}(x)|$$

où :

- $F_{\text{ref}}(x)$ est la fonction de répartition empirique des données de référence
- $F_{\text{curr}}(x)$ est la fonction de répartition empirique des données courantes
- $\sup_x$ représente le supremum sur toutes les valeurs de x

Interprétation : Un D_{KS} élevé (proche de 1) indique une forte dérive. Le seuil typique est 0.05-0.1.

Test de Population Stability Index (PSI)

Le PSI mesure le changement dans la distribution d'une variable en comparant les proportions dans différents bins.

Formule PSI :

$$PSI = \sum_{i=1}^{n} (P_{\text{curr},i} - P_{\text{ref},i}) \times \ln\left(\frac{P_{\text{curr},i}}{P_{\text{ref},i}}\right)$$

où :

- $P_{\text{ref},i}$ est la proportion dans le bin i pour les données de référence
- $P_{\text{curr},i}$ est la proportion dans le bin i pour les données courantes
- n est le nombre de bins

Interprétation :

- $PSI < 0.1$: Pas de changement significatif

- $0.1 \leq PSI < 0.25$: Changement modéré
- $PSI \geq 0.25$: Changement significatif nécessitant attention

---

Test de Wasserstein (Earth Mover's Distance)

La distance de Wasserstein mesure le coût minimal pour transformer une distribution en une autre.

Formule :

$$W_p(P, Q) = \left(\inf_{\gamma \in \Gamma(P,Q)} \int |x - y|^p d\gamma(x, y) \right)^{1/p}$$

où :

- P et Q sont les distributions de référence et courante
- $\Gamma(P, Q)$ représente l'ensemble des couplages entre P et Q
- p est généralement égal à 1 (distance de Wasserstein-1)

Pour le cas $p = 1$, la formule se simplifie en fonction des quantiles.

Tests de dérive pour variables catégorielles

Test du Chi-carré (χ^2)

Le test du Chi-carré évalue si les fréquences observées diffèrent significativement des fréquences attendues.

Statistique du Chi-carré :

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

où :

- O_i est la fréquence observée dans la catégorie i (données courantes)
- E_i est la fréquence attendue dans la catégorie i (données de référence)
- k est le nombre de catégories

Degrés de liberté : $df = k - 1$

Interprétation : Une valeur χ^2 élevée avec une p-value < 0.05 indique une dérive significative.

Test Z pour proportions

Le test Z compare les proportions entre deux échantillons.

Statistique Z :

$$Z = \frac{p_{\text{curr}} - p_{\text{ref}}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_{\text{ref}}} + \frac{1}{n_{\text{curr}}}\right)}}$$

où :

- $p_{\text{ref}} = \frac{x_{\text{ref}}}{n_{\text{ref}}}$ est la proportion dans les données de référence
- $p_{\text{curr}} = \frac{x_{\text{curr}}}{n_{\text{curr}}}$ est la proportion dans les données courantes
- $\bar{p} = \frac{x_{\text{ref}} + x_{\text{curr}}}{n_{\text{ref}} + n_{\text{curr}}}$ est la proportion combinée

Interprétation : $|Z| > 1.96$ indique une différence significative au niveau de confiance de 95%.

Tests de qualité des données

Taux de valeurs manquantes

Formule :

$$\text{Missing Rate} = \frac{\text{Nombre de valeurs manquantes}}{\text{Nombre total de valeurs}} \times 100\%$$

Seuils recommandés :

- Acceptable : < 5%
- Modéré : 5-20%
- Critique : > 20%

Taux de duplication

Formule :

$$\text{Duplication Rate} = \frac{\text{Nombre de lignes dupliquées}}{\text{Nombre total de lignes}} \times 100\%$$

Détection de valeurs aberrantes (Outliers)

Méthode IQR (Interquartile Range) :

$$\text{Outlier si } x < Q_1 - 1.5 \times IQR \text{ ou } x > Q_3 + 1.5 \times IQR$$

où :

- Q_1 est le premier quartile (25e percentile)
- Q_3 est le troisième quartile (75e percentile)
- $IQR = Q_3 - Q_1$ est l'écart interquartile

Méthode des N-sigma :

$$\text{Outlier si } |x - \mu| > n \times \sigma$$

où :

- μ est la moyenne
- σ est l'écart-type
- n est généralement égal à 3 (99.7% des données dans cet intervalle)

Métriques de performance

Métriques de régression

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

où :

- y_i est la valeur réelle
- $\hat{y}_i$ est la valeur prédite
- n est le nombre d'observations

Interprétation : Plus le MAE est faible, meilleure est la performance. Il est exprimé dans la même unité que la variable cible.

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Interprétation : Le RMSE pénalise davantage les grandes erreurs que le MAE. Plus sensible aux outliers.

Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Interprétation : Exprime l'erreur en pourcentage. Attention : non défini si $y_i = 0$.

Coefficient de détermination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

où $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ est la moyenne des valeurs réelles.

Interprétation :

- $R^2 = 1$: prédiction parfaite
- $R^2 = 0$: le modèle n'est pas meilleur qu'une prédiction constante (moyenne)
- $R^2 < 0$: le modèle est pire qu'une prédiction constante

Métriques de classification

Accuracy (Exactitude)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

où :

- TP (True Positives) : vrais positifs
- TN (True Negatives) : vrais négatifs
- FP (False Positives) : faux positifs
- FN (False Negatives) : faux négatifs

Precision (Précision)

$$\text{Precision} = \frac{TP}{TP + FP}$$

Interprétation : Proportion de prédictions positives qui sont correctes. Important quand le coût des faux positifs est élevé.

Recall (Rappel / Sensibilité)

$$\text{Recall} = \frac{TP}{TP + FN}$$

Interprétation : Proportion de vrais positifs détectés parmi tous les positifs réels. Important quand le coût des faux négatifs est élevé.

F1-Score

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

Interprétation : Moyenne harmonique de la précision et du rappel. Utile pour équilibrer les deux métriques.

F-beta Score (généralisation)

$$F_\beta = (1 + \beta^2) \times \frac{\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$

où :

- $\beta < 1$: favorise la précision
- $\beta > 1$: favorise le rappel
- $\beta = 1$: équilibre (F1-Score)

ROC AUC (Area Under the ROC Curve)

La courbe ROC trace le **True Positive Rate** (TPR) en fonction du **False Positive Rate** (FPR) pour différents seuils de classification.

$$TPR = \frac{TP}{TP + FN} = \text{Recall}$$

$$FPR = \frac{FP}{FP + TN}$$

L'aire sous la courbe (AUC) est calculée par intégration :

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

Interprétation :

- $AUC = 1.0$: classificateur parfait
- $AUC = 0.5$: classificateur aléatoire
- $AUC < 0.5$: classificateur pire que le hasard

Matthews Correlation Coefficient (MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Interprétation :

- $MCC = 1$: prédiction parfaite
- $MCC = 0$: prédiction aléatoire
- $MCC = -1$: désaccord total

Le MCC est particulièrement utile pour les datasets déséquilibrés car il prend en compte les quatre catégories de la matrice de confusion.

Métriques pour classes déséquilibrées

Weighted Average

Pour les métriques multi-classes, la moyenne pondérée prend en compte la fréquence de chaque classe :

$$\text{Metric}_{\text{weighted}} = \sum_{i=1}^{k} w_i \times \text{Metric}_i$$

où :

- $w_i = \frac{n_i}{n}$ est le poids de la classe i
- n_i est le nombre d'échantillons de la classe i
- n est le nombre total d'échantillons
- k est le nombre de classes

Balanced Accuracy

$$\text{Balanced Accuracy} = \frac{1}{k} \sum_{i=1}^{k} \frac{TP_i}{TP_i + FN_i}$$

Interprétation : Moyenne des recalls par classe, donnant un poids égal à chaque classe indépendamment de sa fréquence.

Interprétation des résultats

Structure des résultats

Les résultats de la Test Suite sont organisés en deux parties principales. Le **summary** contient le nombre total de tests, le nombre de tests réussis et échoués, ainsi qu'une répartition par statut (SUCCESS, FAIL, WARNING, ERROR).

La section **tests** détaille chaque test individuel avec son nom, sa description, son statut, ses paramètres de configuration et ses résultats numériques.

Statuts des tests

SUCCESS indique que le test a été passé avec succès et qu'aucune action n'est requise.

FAIL signale un test échoué nécessitant une investigation et potentiellement une alerte.

WARNING indique qu'un seuil d'avertissement a été atteint, suggérant une surveillance accrue sans urgence immédiate.

ERROR révèle une erreur d'exécution du test, souvent due à des problèmes de données ou de configuration.

Analyse des dérives

Lorsqu'une dérive est détectée, plusieurs facteurs doivent être analysés. La **magnitude** de la dérive indique son importance via le drift score. Le **nombre de features** affectés révèle si la dérive est localisée ou généralisée. Les **patterns temporels** montrent si la dérive est progressive, soudaine ou cyclique. Enfin, l'**impact métier** évalue les conséquences réelles sur les prédictions et les décisions.

Bonnes pratiques

Choix des données de référence

Les données de référence doivent être représentatives de la distribution "normale" attendue. Il est recommandé d'utiliser les données d'entraînement du modèle ou une période stable récente en production. Les données de référence doivent être suffisamment volumineuses (minimum 1000 échantillons pour des tests statistiques robustes) et nettoyées des anomalies évidentes.

Il faut éviter d'utiliser des données trop anciennes qui ne reflètent plus la réalité actuelle, ainsi que des périodes atypiques (fêtes, promotions, incidents).

Définition des seuils

Les seuils doivent être adaptés au contexte métier et aux conséquences des faux positifs/négatifs. Pour les **features critiques** ayant un impact direct sur les décisions, utiliser des seuils stricts (drift threshold = 0.1). Pour les **features normales** à importance modérée, appliquer des seuils standard (drift threshold = 0.3). Les **features non-critiques** peuvent tolérer des seuils permissifs (drift threshold = 0.5).

Les seuils doivent être régulièrement réévalués basés sur l'historique des alertes et ajustés selon les retours d'expérience.

Fréquence de monitoring

La fréquence de monitoring dépend de la criticité du système et de la vitesse de changement des données. Pour les **systèmes critiques** (finance, santé), effectuer un monitoring en temps réel ou horaire. Pour les **systèmes à risque modéré**, un monitoring quotidien ou hebdomadaire suffit. Les **systèmes à faible risque** peuvent se contenter d'un monitoring mensuel ou trimestriel.

Échantillonnage des données

Pour les grands volumes de données, l'échantillonnage est nécessaire pour des performances optimales. Un échantillonnage stratifié préserve la distribution des classes importantes. L'échantillonnage aléatoire avec seed fixe assure la reproductibilité. La taille minimale recommandée est de 10000 échantillons pour les données de référence et 5000 pour les données courantes.

Il est important de vérifier que l'échantillon reste représentatif de la population complète après échantillonnage.

Gestion des alertes

Un système d'alertes efficace nécessite plusieurs niveaux. Les **alertes critiques** (multiple drifts, chute de performance majeure) requièrent une notification immédiate et une intervention dans l'heure. Les **alertes importantes** (drift sur features clés, dégradation modérée) nécessitent une investigation dans les 24 heures. Les **alertes informatives** (drift mineur, avertissements) peuvent être traitées lors de la revue hebdomadaire.

Il faut éviter la fatigue d'alerte en ajustant les seuils pour réduire les faux positifs, en agrégeant les alertes similaires et en fournissant un contexte actionnable pour chaque alerte.

Documentation et traçabilité

Chaque exécution de Test Suite doit être documentée avec horodatage précis, version du modèle testé, source et période des données, environnement d'exécution, paramètres de configuration utilisés et résultats complets.

Conserver un historique des résultats permet d'analyser les tendances à long terme, d'identifier les patterns saisonniers, de calibrer les seuils et de démontrer la conformité réglementaire.

Intégration dans le workflow ML

La Test Suite doit être intégrée à différentes étapes du workflow. Lors du **développement**, valider les nouvelles données avant l'entraînement et tester les performances du modèle. En **staging**, vérifier la stabilité avant le déploiement et valider sur des données représentatives de production. En **production**, monitorer continuellement les dérives et tester les performances régulièrement. Pour la **maintenance**, détecter les besoins de réentraînement et valider les nouveaux modèles.

Analyse de cause racine

Lorsqu'un test échoue, une analyse systématique doit être conduite. Vérifier d'abord les **changements de source de données** (nouveau système, migration, modification de pipeline). Examiner ensuite les **changements métier** (nouvelle offre, changement de marché, modification de processus). Investiguer les **problèmes techniques** (bugs, interruptions, problèmes de qualité). Enfin, considérer les **facteurs externes** (saisonnalité, événements exceptionnels, changements réglementaires).

Tests complémentaires

En plus des tests automatiques d'Evidently, il est recommandé d'effectuer des tests de sanity check manuels, des analyses exploratoires des cas d'échec, des comparaisons avec des métriques métier et des validations avec les experts du domaine.

Ressources supplémentaires

La documentation officielle complète est disponible sur le site d'Evidently AI. Le dépôt GitHub contient des exemples pratiques et le code source. Une communauté active est disponible sur Discord pour échanger avec d'autres utilisateurs et l'équipe de développement.

Conclusion

L'Evidently AI Test Suite fournit un framework robuste et mathématiquement fondé pour le monitoring des données et des modèles de machine learning. En combinant des tests statistiques éprouvés, des métriques de performance complètes et une architecture flexible, elle permet de détecter proactivement les problèmes avant qu'ils n'impactent les résultats métier.