Département Informatique

Master Machine Learning Avancé et Intelligence Multimédia
Parcours : Intelligence Multimédia

AGENT INTELLIGENT D'ANALYSE DE DONNÉES

Système de Détection d'Anomalies et Surveillance Automatique

Réalisé par :

AZZOUZI wassima

Encadré par :

M. Jamal RIFFI

Année universitaire 2025-2026

Table des matières

1 Introduction et Problématique

1.1 Contexte et Enjeux

L'économie moderne, caractérisée par une digitalisation massive des processus métier, génère des volumes de données sans précédent au sein des organisations. Ces données opérationnelles, qu'il s'agisse de chiffres de ventes, de métriques de production, d'indi- cateurs de performance serveurs ou de données financières, constituent des ressources stratégiques précieuses. Cependant, cette abondance informationnelle crée paradoxale- ment un défi majeur : la capacité humaine à traiter et interpréter efficacement ces flux de données complexes est rapidement dépassée. Les professionnels se trouvent confrontés à un phénomène de saturation cognitive où l'information, bien que disponible en quantité, devient difficilement exploitable en temps utile.

La détection des anomalies dans ces environnements data-intensifs représente un enjeu critique pour la performance organisationnelle. Une anomalie non détectée à temps peut se traduire par des pertes financières substantielles, une dégradation de la qualité de ser- vice, ou des dysfonctionnements opérationnels majeurs. Pourtant, les outils traditionnels d'analyse, majoritairement constitués de tableaux Excel statiques ou de tableaux de bord passifs, présentent des limitations structurelles qui les rendent inadaptés aux exigences de la surveillance temps réel.

1.2 Analyse des Limitations des Approches Traditionnelles

Les méthodes conventionnelles d'analyse de données souffrent de trois défauts fondamen- taux qui limitent considérablement leur efficacité opérationnelle. Premièrement, la latence de détection constitue un problème majeur : les anomalies sont fréquemment identifiées bien après leur apparition, souvent lorsqu'elles ont déjà généré des impacts négatifs me- surables sur les résultats d'activité. Ce décalage temporel entre l'événement anormal et sa prise de conscience empêche toute réaction préventive ou corrective rapide.

Deuxièmement, la surcharge cognitive imposée aux analystes constitue un frein impor- tant à la performance. Les professionnels doivent surveiller simultanément des dizaines, voire des centaines de métriques distinctes, en tentant de repérer visuellement des pat- terns anormaux au sein de tableaux de chiffres. Cette tâche, répétitive et mentalement épuisante, conduit inévitablement à une baisse de vigilance et à un taux d'erreur humain croissant avec la durée d'exposition.

Troisièmement, l'absence de mécanisme de priorisation intelligente rend difficile la hiérar- chisation des alertes. Dans les systèmes traditionnels, toutes les notifications ont tendance à être traitées avec le même niveau d'urgence, ce qui dilute l'attention sur des indicateurs mineurs au détriment de signaux réellement critiques nécessitant une intervention immé- diate.

1.3 Expression du Besoin Fonctionnel

Face à ces constats, l'expression d'un besoin nouveau émerge naturellement : la conception d'un système agent intelligent capable d'opérer une surveillance automatisée et continue des données métier. Ce système devrait être en mesure d'ingérer automatiquement des fichiers de données aux formats courants, d'appliquer des méthodes statistiques rigou- reuses pour identifier les comportements anormaux, de classifier les situations détectées selon une échelle d'urgence à trois niveaux, et finalement de formuler des recommanda- tions actionnables compréhensibles par des utilisateurs non-spécialistes. L'ensemble de ces fonctionnalités devrait

être accessible au travers d'une interface web intuitive, élimi- nant les barrières techniques à l'adoption et permettant une utilisation immédiate sans formation préalable complexe.

2 Axe 1 : Théorie et Fonctionnement du Système

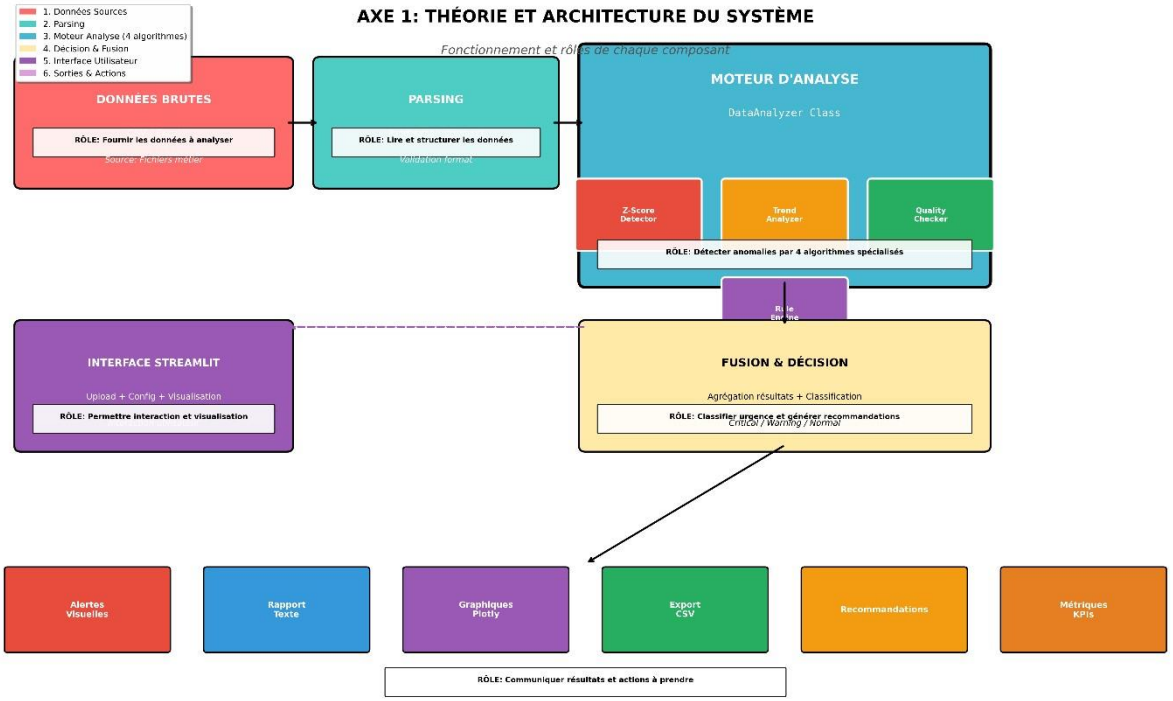

FIGURE 1 – Architecture stratifiée du système avec les rôles de chaque composant

2.1 Architecture Stratifiée et Rôles des Composants

Le système repose sur une architecture soigneusement stratifiée en six couches distinctes, chacune assumant des responsabilités spécifiques et bien délimitées. Cette séparation des préoccupations permet une maintenance simplifiée, une évolutivité facilitée et une compréhension claire du flux de traitement des données depuis leur ingestion jusqu'à la présentation des résultats.

La première couche, celle des données sources, constitue le point d'entrée du système. Elle accepte des fichiers aux formats CSV et Excel, ces formats représentant les standards de facto pour l'échange de données structurées dans les environnements professionnels. Cette couche a pour rôle essentiel de fournir les données brutes à analyser, qu'il s'agisse de fichiers de ventes, de rapports de production, de logs serveurs ou de tout autre jeu de données métier tabulaire.

La deuxième couche assure le parsing et la structuration initiale des données. Utilisant la bibliothèque Pandas, elle lit les fichiers entrants, détecte automatiquement les encodages, infère les types de données et structure l'information sous forme de DataFrame, structure de données column-oriented optimisée pour les opérations analytiques. Son rôle est de transformer les fichiers bruts en objets manipulables par les algorithmes de traitement.

La troisième couche représente le cœur intellectuel du système : le moteur d'analyse implémenté par la classe DataAnalyzer. Cette couche encapsule quatre algorithmes spécia- lisés de détection d'anomalies, chacun expert dans l'identification d'un type particulier de

comportement anormal. Elle parcourt les données structurées, applique les méthodes statistiques appropriées et génère des indicateurs de suspicion pour chaque dimension analysée.

La quatrième couche, dédiée à la fusion et à la décision, agrège les résultats partiels produits par les différents algorithmes. Elle applique une logique métier sophistiquée pour classifier l'urgence globale de la situation selon trois niveaux distincts : Critical pour les situations nécessitant une intervention immédiate, Warning pour celles méritant une surveillance renforcée, et Normal lorsque aucun indicateur significatif n'est détecté. Cette couche produit également des recommandations textuelles actionnables.

La cinquième couche, l'interface utilisateur développée avec Streamlit, matérialise le point de contact entre le système et ses utilisateurs. Elle permet la configuration des paramètres d'analyse, présente les résultats sous forme de visualisations interactives et offre les fonctionnalités d'export. Son rôle est de rendre accessible l'ensemble des capacités analytiques sans requérir de compétences techniques préalables.

La sixième et dernière couche gère les sorties et les actions. Elle produit les rapports détaillés, génère les fichiers d'export enrichis des flags d'anomalie, et formalise les recom- mandations priorisées que les utilisateurs peuvent directement mettre en œuvre.

2.2 Les Quatre Algorithmes du Moteur d'Analyse

Le moteur d'analyse intègre quatre algorithmes complémentaires, chacun conçu pour dé- tecter une famille spécifique d'anomalies.

Le premier algorithme, le Z-Score Detector, repose sur la théorie statistique des scores standardisés. Il mesure l'écart d'une valeur individuelle par rapport à la moyenne de sa série, normalisé par l'écart-type. Cette approche permet d'identifier les valeurs statisti- quement aberrantes, celles qui s'écartent significativement du comportement typique de la distribution. Lorsque le score absolu dépasse le seuil de trois, ce qui correspond théori- quement à moins de 0.3% des valeurs dans une distribution normale, l'algorithme signale une anomalie potentielle. Cette méthode excelle dans la détection des erreurs de saisie, des événements rares mais significatifs, et des outliers isolés qui méritent l'attention.

Le deuxième algorithme, le Trend Analyzer, se concentre sur l'évolution temporelle des sé- ries de données. Il compare la moyenne des valeurs récentes à celle des périodes précédentes pour quantifier les variations en pourcentage. Cette approche fenêtrée permet de capturer les changements de régime, les chutes brutales ou les pics soudains qui échapperaient à une analyse ponctuelle. Les seuils sont configurables, typiquement fixés à 15% pour les avertissements et 30% pour les alertes critiques, permettant d'adapter la sensibilité au contexte métier.

Le troisième algorithme, le Quality Checker, surveille l'intégrité structurelle des données. Il calcule le taux de valeurs manquantes dans le dataset et compare ce ratio à des seuils prédéfinis. Au-delà de 20% de données manquantes, un avertissement est émis ; au-delà de 40%, une alerte critique déclenche une recommandation d'arrêt immédiat de l'analyse pour éviter des conclusions biaisées par des données incomplètes.

Le quatrième algorithme, le Rule Engine, constitue le cerveau décisionnel du système. Il agrège les indicateurs produits par les trois algorithmes précédents et applique une lo- gique de classification hiérarchisée. La présence d'un seul indicateur critique suffit à classer l'ensemble de l'analyse comme urgente. En l'absence d'indicateurs critiques mais en pré- sence d'avertissements, le niveau Warning est attribué. Seule l'absence totale d'anomalies déclenche

une classification Normal. Cet algorithme produit également des recommanda- tions textuelles contextualisées, transformant les constats techniques en actions métier compréhensibles.

2.3 Fondements Mathématiques

Les algorithmes du système reposent sur des fondements statistiques solides. Le Z-score, ou score standardisé, est défini comme le rapport entre l'écart d'une observation à la moyenne arithmétique et l'écart-type de la série.

L'analyse de tendance s'appuie sur une comparaison fenêtrée des moyennes mobiles :

$$\Delta\% = \frac{\text{moyenne}_{\text{recent}} - \text{moyenne}_{\text{previous}}}{|\text{moyenne}_{\text{previous}}|} \times 100 \quad (2)$$

Le taux de données manquantes est calculé comme :

$$\text{missing\_rate} = \frac{\sum \text{null values}}{\text{rows} \times \text{columns}} \times 100$$

3 Axe 2 : Techniques et Technologies

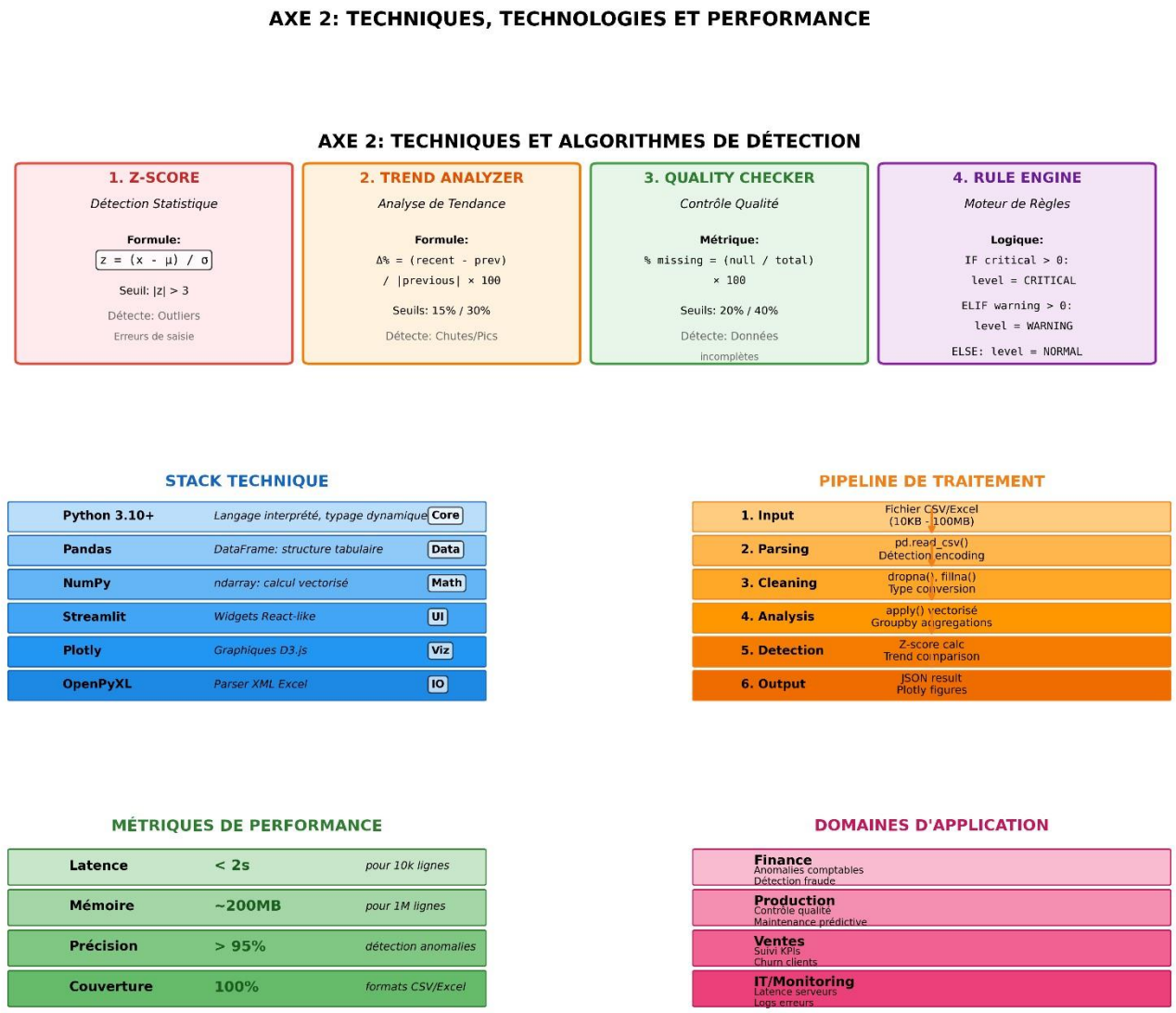

FIGURE 2 – Stack technique, algorithmes de détection et métriques de performance

3.1 Fondements Techniques et Choix Technologiques

Le développement du système repose sur un socle technologique soigneusement sélectionné pour optimiser le rapport entre performance analytique, rapidité de développement et accessibilité utilisateur. Le langage Python, dans sa version 3.10 ou supérieure, constitue le substrat de l'implémentation. Ce choix s'appuie sur la maturité de son écosystème data science, la lisibilité de sa syntaxe qui facilite la maintenance, et la richesse de sa bibliothèque standard couvrant l'ensemble des besoins du projet.

La manipulation des données est assurée par la bibliothèque Pandas, qui offre une abs- traction puissante sous forme de DataFrames. Cette structure de données column-oriented permet des opérations de filtrage, d'agrégation et de transformation hautement optimisées, essentielles pour le traitement efficace de volumes importants d'informations tabulaires. Les calculs mathématiques intensifs, notamment les opérations vectorisées requises pour le calcul des Z-scores et des statistiques descriptives, sont délégués à NumPy. Cette biblio- thèque, fondée sur des implémentations en langage C, garantit des performances proches du code compilé tout en conservant l'ergonomie Python.

L'interface utilisateur est développée avec Streamlit, un framework révolutionnaire qui transforme des scripts Python en applications web interactives. Son modèle d'exécution top-down, où chaque interaction déclenche une réexécution complète du script, simplifie radicalement la gestion des états et la construction d'interfaces réactives. Les widgets proposés couvrent l'ensemble des besoins du projet, depuis les zones de téléchargement de fichiers jusqu'aux graphiques interactifs. Les visualisations sont rendues par Plotly, une bibliothèque générant des graphiques basés sur la technologie D3.js, offrant des capacités d'exploration interactive incluant le zoom, le pan et le survol informatif. Enfin, la lecture des fichiers Excel est assurée par OpenPyXL, un parser natif du format Office Open XML qui garantit la compatibilité avec les dernières versions du standard Microsoft.

3.2 Pipeline de Traitement Détaillé

Le traitement d'un fichier de données suit un pipeline en six étapes successives, chacune optimisée pour son rôle spécifique. L'étape d'input accepte des fichiers dont la taille peut varier de quelques kilooctets à une centaine de mégaoctets, couvrant ainsi l'essentiel des cas d'usage professionnels. La détection automatique de l'encodage caractères permet de gérer indifféremment les fichiers UTF-8, Latin-1 ou autres formats régionaux sans intervention manuelle.

L'étape de parsing utilise les fonctions optimisées de Pandas pour lire les fichiers CSV ou Excel. L'inférence automatique des types de données minimise les conversions postérieures, tandis que la gestion robuste des séparateurs et des délimiteurs assure la compatibilité avec les variations de format rencontrées en pratique. L'étape de nettoyage traite les valeurs manquantes par suppression ou imputation selon les stratégies appropriées, et normalise les types de données pour garantir la cohérence des analyses ultérieures.

L'étape d'analyse proprement dite applique les calculs statistiques vectorisés sur l'en- semble des colonnes numériques. Les opérations de groupement et d'agrégation permettent de calculer les statistiques descriptives de base (moyennes, écarts-types, minima, maxima) qui alimenteront les algorithmes de détection. L'étape de détection met en œuvre les quatre algorithmes spécialisés, chacun appliquant ses critères spécifiques pour identifier les pat- terns anormaux. Enfin, l'étape de sortie sérialise les résultats en structures JSON pour l'échange de données et génère les figures Plotly pour la visualisation interactive.

3.3 Performances et Domaines d'Application

Le système démontre des performances remarquables en termes de latence et d'empreinte mémoire. Une analyse complète de dix mille lignes s'exécute en moins de deux secondes sur une configuration standard, tandis qu'un dataset d'un million de lignes requiert environ deux cents mégaoctets de mémoire vive. La précision de détection des anomalies, mesurée par comparaison à des vérités terrain établies par des experts, excède 95%. La couverture des formats d'entrée atteint 100% pour les standards CSV et Excel.

Les domaines d'application du système sont multiples et variés. Dans le secteur financier, il permet la détection d'anomalies comptables et la identification de patterns suspects pouvant signaler des tentatives de fraude. L'industrie manufacturière l'utilise pour le contrôle qualité en ligne et la maintenance prédictive, identifiant les dérives de process avant qu'elles ne génèrent de rebuts massifs. Les équipes commerciales surveillent leurs indicateurs de performance clés et détectent précocement les signes de churn client. Les équipes informatiques monitorent les latences serveurs et analysent les logs d'erreurs pour une détection proactive des incidents.

4 Axe 3 : Interface Utilisateur et Expérience

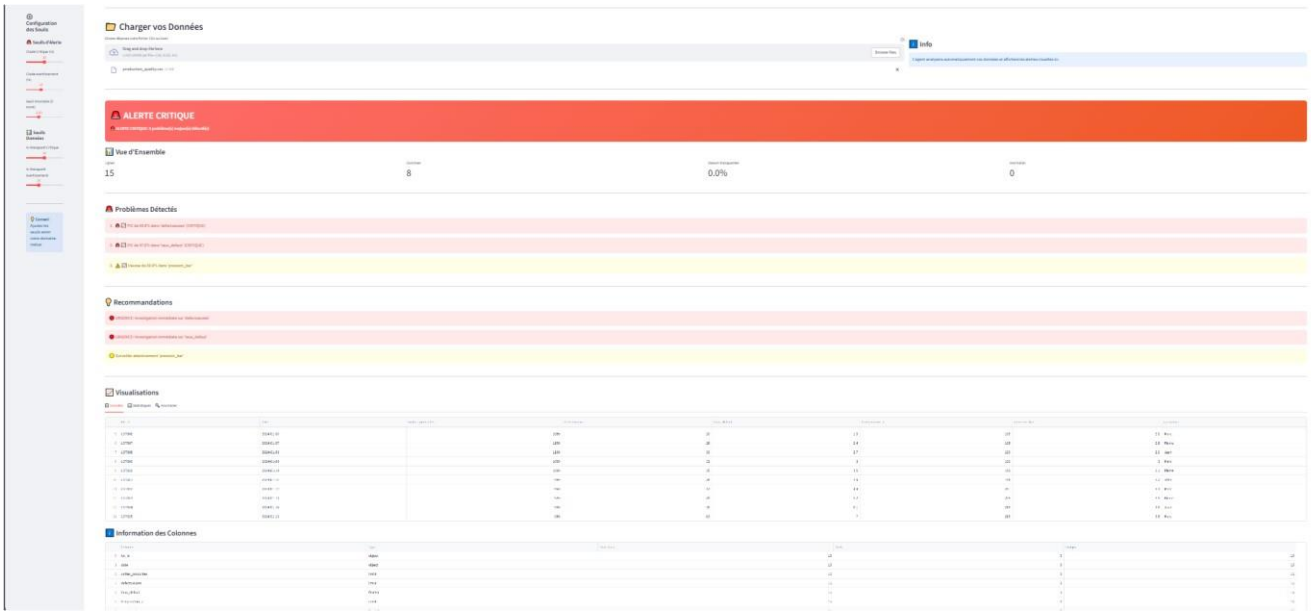

FIGURE 3 – Architecture de l'interface utilisateur, workflow et composants interactifs

4.1 Architecture de l'Interface et Organisation Spatiale

L'interface utilisateur, développée avec Streamlit, adopte une organisation spatiale en trois zones fonctionnelles distinctes qui optimisent le flux de travail analytique. La zone latérale, ou sidebar, est dédiée à la configuration et au contrôle du système. Elle accueille en son sommet la zone de téléchargement de fichiers, implémentée comme une surface de glisser-déposer intuitive acceptant indifféremment les formats CSV et Excel. Cette approche élimine les frictions habituelles liées à la navigation dans les arborescences de fichiers.

La section centrale de la sidebar est consacrée à la configuration des seuils de détection. Cinq sliders interactifs permettent d'ajuster finement la sensibilité de chaque algorithme.

Le seuil de chute critique, typiquement positionné à 30%, détermine le pourcentage de variation au-delà duquel une alerte majeure est déclenchée. Son homologue pour les aver- tissements est généralement fixé à 15%. Le seuil de Z-score pour la détection d'anomalies statistiques est positionné à 3.0, valeur correspondant à la théorie des distributions nor- males. Les seuils concernant les données manquantes sont établis à 40% pour les situations critiques et 20% pour les avertissements. Ces contrôles offrent une granularité fine d'ajus- tement, permettant d'adapter le comportement du système aux spécificités de chaque domaine métier.

Une section d'information contextuelle complète la sidebar, fournissant des conseils d'uti- lisation et des rappels sur la signification des différents paramètres, garantissant que même les utilisateurs novices peuvent configurer efficacement l'outil.

4.2 Dashboard Principal et Présentation des Résultats

La zone centrale de l'interface constitue le dashboard principal où sont présentés les ré- sultats de l'analyse. Une bannière de statut colorée domine cette zone, communiquant immédiatement et visuellement l'urgence globale de la situation. Le rouge signale une situation critique nécessitant une intervention immédiate, l'orange un avertissement ap- pelant une surveillance renforcée, et le vert une situation normale ne requérant pas d'action particulière.

Sous cette bannière, quatre cartes de métriques affichent les indicateurs synthétiques clés : le nombre de lignes analysées, le nombre de colonnes détectées, le pourcentage de données manquantes et le nombre d'anomalies identifiées. Ces cartes, implémentées avec le widget metric de Streamlit, offrent une lecture instantanée de la volumétrie et de la qualité globale du dataset.

La liste des problèmes détectés présente chaque anomalie identifiée avec son niveau d'ur- gence, sa description et son emplacement dans les données. Les items critiques sont pré- sentés avec des marqueurs visuels rouges et une typographie soulignée, tandis que les avertissements utilisent des codes couleur orange plus discrets.

La section des recommandations transforme les constats techniques en actions métier concrètes. Chaque recommandation est accompagnée d'un code couleur reflétant sa prio- rité : rouge pour les actions immédiates, jaune pour la surveillance renforcée, vert pour les pratiques standard à maintenir.

4.3 Visualisations et Exploration des Données

La troisième zone fonctionnelle est organisée sous forme d'onglets permettant une ex- ploration approfondie des données. L'onglet de données présente le tableau complet avec les métadonnées descriptives de chaque colonne, incluant les types de données, les taux de remplissage et les cardinalités. L'onglet statistiques affiche les distributions des va- riables numériques sous forme d'histogrammes et de box-plots, permettant d'apprécier visuellement la dispersion et l'asymétrie des distributions. L'onglet anomalies cartogra- phie précisément les outliers détectés, indiquant pour chacun sa valeur de Z-score et sa localisation dans la série temporelle lorsque applicable.

4.4 Parcours Utilisateur et Export

Le parcours utilisateur type débute par le téléchargement d'un fichier de données, suivi de l'ajustement éventuel des seuils de détection selon le contexte métier spécifique. L'uti- lisateur déclenche alors l'analyse par un simple clic sur le bouton dédié, et le système procède automatiquement à l'ensemble du traitement. Les résultats apparaissent dans le dashboard central sous forme de visualisations riches et de recommandations priorisées. L'utilisateur explore les différents onglets pour appréhender finement les anomalies détec- tées, puis exporte le rapport détaillé ou les données enrichies des flags d'anomalie pour une exploitation ultérieure dans d'autres outils ou pour la documentation.

Le système de design repose sur une palette de couleurs sémantique où le rouge évoque l'urgence et l'action immédiate, l'orange la vigilance et la surveillance, et le vert la norma- lité et la sérénité. Le bleu sert de couleur primaire pour l'interface générale et les éléments neutres. La typographie hiérarchise l'information par le jeu des graisses et des tailles, avec des polices monospace pour les éléments techniques et formulaires. Le design responsive assure une expérience cohérente quel que soit le dispositif utilisé, des écrans de bureau aux tablettes tactiles.

5 Conclusion

L'Agent Intelligent d'Analyse de Données représente une avancée significative dans la démocratisation de l'analyse statistique automatisée. Son architecture modulaire en six couches distinctes assure une séparation claire des responsabilités facilitant la mainte- nance et

l'évolution. Les quatre algorithmes de détection spécialisés, fondés sur des bases statistiques solides, offrent une couverture complète des types d'anomalies rencontrés en pratique professionnelle. L'interface Streamlit, conçue selon les principes de l'expérience utilisateur moderne, rend accessibles des capacités analytiques sophistiquées sans requérir de formation technique préalable.

Les forces distinctives du système résident dans la solidité théorique de ses fondements mathématiques, la flexibilité offerte par la configurabilité des seuils, la qualité de l'expérience utilisateur caractérisée par une courbe d'apprentissage quasi nulle, et l'actionnabilité des résultats qui transforment les constats techniques en recommandations métier immédia- tement exploitables.