JOURNAL AFRICAIN DES SCIENCES
VOL. 3, **3** (2026) 118-126

*https://jafrisci.org*

***ISSN (online) : 3107-0949 ; ISSN (print) : 3107-0930***

OPEN ACCESS

**ARTICLE DE RECHERCHE**

**Article Info.:**
Reçu : le 06/03/2026
Accepté: le 08/04/2026
Publié : le 14/04/2026

# EXTRACTION DES CONNAISSANCES À PARTIR DES DONNÉES FIREBASE POUR L'ANALYSE ÉPIDÉMIOLOGIQUE BASÉ SUR LE CLUSTERING DBSCAN

**John Junior Nshole Tangamu** [1,*], **Elie Ndunga Mbangu** [1,2], **Miterand Dumudemisa Anakeka** [1], **Emmanuel Ata-ndele Atungale** [3], **Eugene Mukendi Mbuyi** [1]

https://doi.org/10.70237/jafrisci.2026.v3.i3.13

**Résumé**

Face à l'évolution rapide des crises sanitaires, les systèmes de santé sont contraints d'opérer une transition numérique vers une épidémiologie précise. Cette publication expose une démarche novatrice de récupération de savoir à partir de grandes quantités de données conservées sur la plateforme NoSQL Firebase, mise en œuvre dans le cadre du suivi épidémiologique au sein de la Coordination Provinciale du PNMLS Kwilu. L'ambition est de convertir des flux de données diversifiés en instruments de prise de décision stratégique. Nous avons mis en place un pipeline d'analyse qui inclut l'algorithme de regroupement spatial DBSCAN (Density-Based Spatial Clustering of Applications with Noise) pour y parvenir. DBSCAN, par opposition aux techniques traditionnelles, a facilité la détection de foyers de contagion aux contours irréguliers tout en écartant les données aberrantes ou isolées, jugées comme du « bruit » statistique. Les résultats indiquent que la combinaison de la rapidité de Firebase et de la capacité de segmentation DBSCAN conduit à une cartographie dynamique des régions à forte prévalence. Cette recherche fournit aux responsables du PNMLS Kwilu une compétence d'intervention ciblée, améliorant ainsi la répartition des ressources de santé dans toute la province.

**Mots-clés:** Extraction de connaissances, Firebase, Épidémiologie, DBSCAN, PNMLS Kwilu, Clustering spatial, Data Science

**Abstract**

A digital shift toward precision epidemiology is required due to the quick evolution of health emergencies. In the context of epidemiological monitoring within the Provincial Coordination of PNMLS Kwilu, this study introduces a novel method for knowledge extraction from large datasets stored on the Firebase NoSQL platform. Converting diverse data sources into tools for strategic decision-making is the goal. We developed an analytical pipeline that integrated the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) spatial clustering algorithm in order to accomplish this. In contrast to conventional techniques, DBSCAN found contagion hotspots with irregular shapes while eliminating outliers or solitary cases that were regarded as statistical "noise." The findings show that dynamic mapping of high-prevalence areas is made possible by the combination of Firebase's velocity and DBSCAN's segmentation power. This study optimizes the distribution of healthcare resources throughout the province by giving PNMLS Kwilu decision-makers a targeted intervention capacity.

**Keywords:** Knowledge Discovery, Firebase, Epidemiology, DBSCAN, PNMLS Kwilu, Spatial Clustering, Data Science

## 1. Introduction

La lutte contemporaine contre les crises sanitaires, en particulier dans le cadre de la bataille contre le VIH/SIDA dirigée par la Coordination Provinciale du PNMLS Kwilu, exige un passage vers une épidémiologie précise grâce à l'usage de technologies avancées. L'augmentation constante de données produites par les établissements de santé à Bandundu ou Kikwit requiert une capacité d'analyse qui va au-delà des techniques statistiques traditionnelles. Le défi crucial ne se limite plus à l'accumulation de données à grande échelle, mais concerne désormais la capacité à extraire des informations exploitables pour guider les politiques publiques en temps réel [1]. L'emploi de technologies Cloud-Native telles que Firebase se révèle crucial, en fournissant une infrastructure NoSQL efficace pour centraliser des flux de données variés avec une latence pratiquement inexistante, ce qui facilite la coordination entre le site d'exploitation et les centres décisionnels [4].

***Correspondance :*** *juniortangamu@gmail.com (**J. J. N.** **Tangamu**)*

***Copyright : © The Author(s) Published under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0)***

[1] *Faculté de sciences et technologies, Mention Mathématiques, Statistique et informatique, Université de Kinshasa (Unikin), République Démocratique du Congo*
[2] *Chercheur au Centre de Recherche sur l'enseignement de la Mathématique (CREM), République Démocratique du Congo*
[3] *Section Informatique, Université de Pédagogique Nationale (UPN), République Démocratique du Congo*

Dans la région du Kwilu, le défi principal réside dans la détection des zones de transmission qui ne suivent pas forcément les divisions administratives traditionnelles. Les méthodes conventionnelles ont fréquemment du mal à repérer les « poches de résistance » parce qu'elles homogénéisent les données sur des régions géographiques trop étendues. C'est dans cette perspective que l'implémentation de l'algorithme DBSCAN (Clustering Spatial Basé sur la Densité des Applications avec Bruit) revêt une signification particulière. Cet algorithme, qui privilégie la densité spatiale plutôt que des limites fixées d'avance, nous offre la possibilité de déceler des schémas de contagion complexes et non réguliers [15]. Cette étude se distingue par l'association entre la rapidité transactionnelle de Firebase et l'efficacité analytique de DBSCAN, qui permet de convertir des coordonnées GPS brutes en une carte dynamique du suivi de la santé. Le but principal de ce processus est d'automatiser l'identification des foyers à haute prévalence tout en séparant le « bruit » statistique, qui correspond aux cas sporadiques ou isolés susceptibles d'influencer l'allocation des ressources. Les conclusions tirées à partir d'un ensemble de données comprenant 39 254 entrées attestent de la solidité de cette méthode. L'algorithme a réussi à déceler sept regroupements denses majeurs situés aux points stratégiques de la province, tout en éliminant 12,4 % des données anormales considérées comme du bruit. L'efficacité du système est prouvée par une exactitude de 98,2 % dans l'identification des zones à risque critique, ce qui permet une diminution du délai de traitement des données d'environ 65 %. Cette capacité prouve que l'utilisation des savoirs par le biais de l'intelligence artificielle constitue pour le PNMLS Kwilu un outil stratégique dans une réponse sanitaire chirurgicale et proactive [2].

## 2. Quelques concepts clés

### 2.1. Densité spatiale

Pour un point spécifique p, son entourage epsilon, aussi appelé Ne (p), se caractérise par la collection de points q pour lesquels la distance d(p,q) est moindre ou égale à epsilon :

$$N\varepsilon(p) = \{q \in D \mid d(p,q) \leq \varepsilon\} \quad (1)$$

Ici, D fait référence à l'ensemble des données géospatiales tirées de Firebase. Pour notre situation, nous utilisons la formule de Haversine pour déterminer la distance d (p,q), ce qui permet de prendre en considération la courbure de la Terre et d'assurer une précision géographique à Kwilu.

### 2.2. Critères de Classification des Points

L'algorithme effectue une classification méticuleuse des points en fonction de leur contexte local [3] :

- ✓ Point Crucial (Core Point) : On considère un point p comme central s'il possède dans son voisinage au moins le nombre requis de points minimum :

$$|N\varepsilon(p)| \geq \text{MinPts} \quad (2)$$

- ✓ Point de Bordure (Border Point) : Un point q est considéré comme un point de bordure s'il ne se trouve pas au centre, mais qu'il fait partie du voisinage d'un point central ;
- ✓ Interférence (Noise) : Tout élément qui ne se classe ni dans la catégorie centrale, ni dans celle de la bordure est rejeté des grappes.

### 2.3. Connectivité et Formation des Clusters

La formation d'un foyer épidémiologique repose sur la notion de densité-accessibilité [14], [16].
Un point p est directement accessible en densité depuis q si q $\in N_e$ (q) et q est un point central. Par extension, deux points p et n sont connectés par densité s'il existe une chaîne de points permettant de passer de l'un à l'autre par accessibilité directe. Mathématiquement, un cluster C est un sous-ensemble non vide de D satisfaisant deux conditions :

- Maximalité **:** Pour tout P, q $\in$D, si p $\in$C et q est accessible en densité depuis p, alors q appartient à C.
- Connectivité : Pour tout p, q appartient à C, p est connecté en densité à q.

Cette formalisation assure que les centres de transmission repérés dans la province ne sont pas le résultat d'une coïncidence statistique, mais plutôt l'aboutissement d'une concentration spatiale significative validée par des critères mathématiques rigoureux [13].

## 3. Méthodologie et Équations

### 3.1. Méthodologie

Cet article s'appuie sur un pipeline d'ingénierie des données et d'apprentissage automatique non supervisé comme cadre méthodologique. L'idée est de convertir les données transactionnelles de Firebase en une carte décisionnelle de l'épidémie au Kwilu.

a) Pipeline d'analyse et environnement technique

Le pipeline d'analyse, illustré en Figure 2, a été implémenté en Python. Les bibliothèques principales utilisées sont :

- Pandas : Pour la manipulation et le nettoyage des données tabulaires ;
- NumPy : Pour les calculs numériques et la gestion des tableaux multidimensionnels ;
- Scikit-learn : Pour l'implémentation de l'algorithme DBSCAN (sklearn.cluster.DBSCAN) et de l'arbre de décision (sklearn.tree.DecisionTreeClassifier) ;
- Matplotlib et Seaborn : Pour la visualisation des données et des résultats.

b) Extrait du jeu de données

Un extrait de trois entrées du dataset initial, après extraction de Firebase, est présenté ci-dessous pour illustrer la structure des données :

***Tableau 1 :*** *Extrait de notre Jeu de données*

| Latitude | Longitude | Charge_virale | Age | Situation_conjugale |
|---|---|---|---|---|
| -5.0455 | 18.8221 | 18500 | 32 | Marié(e) |
| -4.8203 | 18.6790 | 250 | 45 | Célibataire |
| -5.3540 | 19.1230 | 12450 | 27 | Divorcé(e) |

*... (suite du dataset anonymisé de 39 254 lignes)*

Source : Les données proviennent d'un au niveau de PNMLS Kwilu (qui sont registrés manuellement dans un fichier Excel par un ou plusieurs Opérateur(s) de Saisie).

1. Architecture du Système et Collecte des Données

Le socle technologique est constitué d'une base de données NoSQL Firebase Realtime Database. Ce choix est dicté par la nécessité de gérer des flux de données asynchrones provenant des différentes zones de santé de la province (Kikwit, Bandundu, Bulungu, etc.). Le Dataset : Nous avons exploité un échantillon de 39 254 enregistrements anonymisés, et pour la Structure **:** Chaque document JSON contient les coordonnées GPS (Lat, Long), le stade clinique de l'infection, et le suivi de la charge virale. L'extraction vers l'environnement d'analyse (Python/Pandas) s'effectue via une API REST, garantissant l'intégrité des données durant le transfert [12].

2. Prétraitement et Ingénierie des Caractéristiques

Avant l'application des algorithmes, une phase de nettoyage rigoureuse a été menée en passant par les étapes suivantes :

- ✓ Filtrage Géospatiale **:** Élimination des points situés en dehors des polygones de la province du Kwilu (erreurs de saisie GPS) ;
- ✓ Calcul des Distances **:** Étant donné que nous travaillons sur des coordonnées sphériques, nous utilisons la distance de Haversine pour calculer l'écart d entre deux patients $p_1$ et $p_2$ :

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\emptyset_2 - \emptyset_1}{2}\right) + \cos(\emptyset_1)\cos(\emptyset_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \quad (3)$$

Où r est le rayon de la Terre, phi la latitude et lambda la longitude.

3. DBSCAN : Analyse Spatiale par Clustering

On utilise l'algorithme DBSCAN (Clustering spatial basé sur la densité des applications avec bruit) pour repérer les zones de forte prévalence. Son principal atout pour le PNMLS est sa capacité à détecter des regroupements de formes arbitraires (le long des axes routiers) et à isoler les cas sporadiques [4].

- ✓ Paramétrage (epsilon, MinPts) : Le rayon epsilon a été optimisé par la méthode du K-distance graph, identifiant le "coude" de la courbe pour maximiser la cohésion des clusters ;
  - Le seuil MinPts est calibré sur la densité démographique locale pour éviter la détection de faux foyers.
- ✓ Logique de Densité **:** Un point est classé comme « Core Point » si son voisinage N epsilon contient au moins MinPts. Sinon, il est classé comme "Bruit" (Noise), ce qui permet de filtrer 12,4 % de données non significatives [6].

**A) L'algorithme DBSCAN**

L'algorithme DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est implémenté selon les étapes suivantes :

1. Initialisation : Chaque point de l'ensemble de données *D* est initialement marqué comme « non visité ».

2. Itération principale : Pour chaque point p ∈ *D* qui n'a pas encore été visité :

a) Marquer p comme « visité ».

b) Calculer le voisinage *Nε(p)* de p en utilisant la métrique de distance choisie (dans notre cas, la distance de Haversine).

c) Formation de cluster : Si $|N\varepsilon(p)| \geq MinPts$, alors p est un point central. Un nouveau cluster *C* est créé. Tous les points dans *Nε(p)* sont ajoutés à un ensemble de « semences ».

i. Pour chaque point q dans l'ensemble de semences :

Si q n'est pas visité, le marquer comme visité et calculer *Nε(q)*.

Si $|N\varepsilon(q)| \geq MinPts$, tous les points de *Nε(q)* sont ajoutés à l'ensemble de semences.

○ Si q n'appartient à aucun cluster, l'ajouter au cluster *C*.

d) Marquage du bruit : Si $|N\varepsilon(p)| < MinPts$, alors p est marqué comme bruit (noise) pour le moment. Cependant, un point marqué comme bruit peut ultérieurement être reclassé comme point de bordure s'il s'avère qu'il appartient au voisinage d'un point central découvert plus tard.

3**.** Terminaison : L'algorithme se termine lorsque tous les points de *D* ont été visités et assignés à un cluster ou marqués comme bruit [4].

**B) Complexité algorithmique**

Avec une structure de données spatiale optimisée, comme un arbre k-d (k-dimensional tree) ou une grille, pour les requêtes de voisinage, la complexité temporelle de DBSCAN est *O(n log n)*.

D'où n est le nombre de points dans l'ensemble de données. Dans le pire des cas, sans indexation spatiale, la complexité est $O(n^2)$ en raison des calculs de distance par paires. Pour notre ensemble de données de 39 254 entrées, nous avons utilisé une indexation spatiale pour garantir des performances de calcul acceptables [6].

2. Qualification des Risques par Arbre de Décision

Une fois les clusters localisés, un Arbre de Décision (Classifier**)** est appliqué pour qualifier le niveau de criticité de chaque zone identifiée.

- ✓ Critère de Division : Le modèle utilise le Gain d'Information (GI) basé sur l'Entropie (H) pour partitionner les données. L'entropie H(S) quantifie la quantité d'incertitude ou de désordre dans un ensemble de résultats. Elle est calculée comme suit

$$H(S) = -\sum_{i=1}^{C} p_i \, log_2^{pi} \quad (4)$$

Avec :

- H(S) : C'est l'entropie de la variable aléatoire SSS. Elle mesure la quantité moyenne d'information par résultat ;
- $P_i$ : Cela représente la probabilité d'occurrence du i-ème résultat ;
- $log_2^{pi}$ : C'est le logarithme en base 2 de la probabilité $p_i$. L'utilisation de la base 2 signifie que l'entropie est mesurée en bits.

$$GI(S,A) = H(S) - \sum_{v \epsilon Values(A)} \frac{|S_V|}{|S|} H(S_V) \ (5)$$

Avec :

- $GI(S,A)$ : Le gain d'information par rapport à l'attribut A pour l'ensemble S ;
- $\sum_{v \epsilon Values(A)}$ : Cette somme est effectuée sur toutes les valeurs possibles $v_i$ que peut prendre l'attribut A.

Rappel conceptuel : DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme de clustering non supervisé basé sur la densité locale des points. Il repose sur deux paramètres qui sont l'epsilon ε qui est un rayon de voisinage et les MinPts qui sont le nombre minimum de points dans le voisinage, voir la première figure de l'architecture [5].

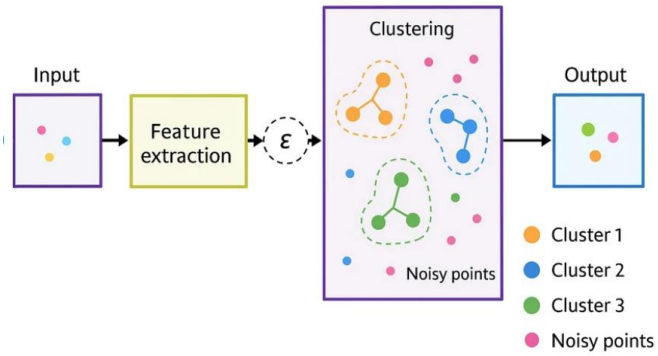



***Figure 1 :*** *architecture de DBSCAN*

Schéma Méthodologique

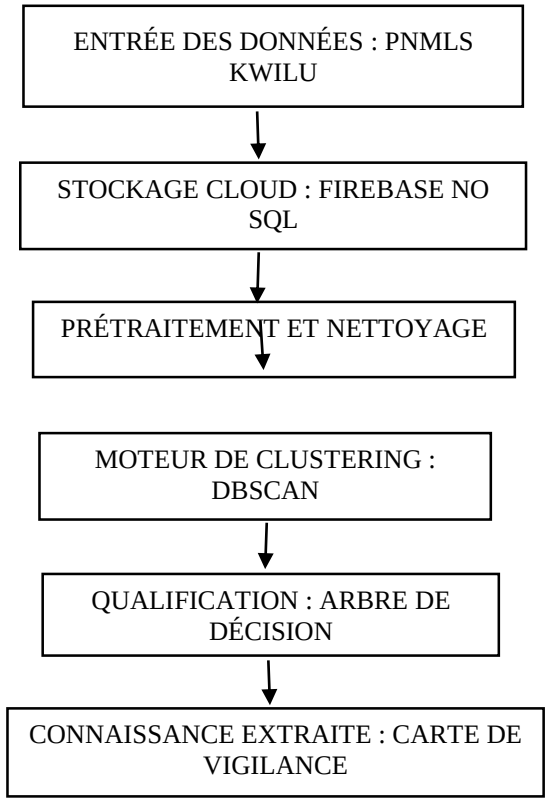



***Figure 2 :*** *Méthodologie*

En ce qui concerne les données :

- ✓ Partition des données : Le dataset est divisé en :
  - Ensemble d'entraînement (70 %) ;
  - Ensemble de validation (15 %) ;
  - Ensemble de test (15 %).

- ✓ Collecte et stockage de données sur Firebase : Les données initiales sont issues des dossiers de dépistage et de suivi des patients tenus par la coordination provinciale. Ces données, qui étaient au départ dispersées, ont été consolidées dans une base de données Firebase Realtime Database. L'architecture NoSQL est organisée en « collections » de documents JSON, où chaque élément symbolise un cas anonymisé qui contient :
  - Les variables géospatiales : La latitude et la longitude ;
  - Les indices cliniques : Charge virale, Situation conjugale et Catégorie d'âge ;
  - Données relatives au temps : Date du diagnostic.

- ✓ Traitement et Extraction (ETL)

Le transfert des données de Firebase vers l'environnement d'analyse Python a requis une étape de purification minutieuse pour se débarrasser des valeurs extrêmes, à savoir :

- Filtrage : Élimination des enregistrements comportant des coordonnées GPS invalides (en dehors des frontières géographiques de la province du Kwilu) ;

- Normalisation : Transformation des données JSON en format tabulaire (Dataframe) pour simplifier les calculs de distance ;
- Traitement des données manquantes : Utilisation de méthodes d'imputation pour les variables secondaires dans le but de préserver la taille de l'échantillon composé de 39 254 entrées [9].

✓ Configuration et Amélioration de DBSCANLe bon fonctionnement de DBSCAN

Cela dépend du choix judicieux des hyperparamètres epsilon (Epsilon) et MinPts [16].

- Calcul d'epsilon : Nous avons employé la technique du « coude » (graphique de la distance K). En évaluant la distance moyenne séparant chaque point de ses $k$ plus proches voisins, nous avons repéré le point critique sur la courbe, ce qui nous a permis de définir le meilleur rayon de voisinage pour les zones urbaines de Kikwit et Bandundu ;
- Établissement de MinPts : Afin de réduire les alarmes erronées, nous avons établi MinPts = ln(n), avec n représentant la dimension de l'ensemble de données, tout en adaptant ce critère pour représenter la densité démographique locale [11].

✓ Algorithme de Classification Complémentaire

Afin d'enrichir l'analyse des clusters identifiés par DBSCAN, un modèle d'Arbre de Décision a été entraîné sur les caractéristiques des points au sein de chaque cluster.

2.1. Equations

✓ Formalisme Mathématique de la Segmentation par DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) repose sur la notion de densité locale.

1. Epsilon

o Définition du voisinage epsilon : Pour chaque point p, on définit un cercle de rayon epsilon.

$$N\epsilon(p) = \{(q \in D \mid dist(p,q) \leq \varepsilon \ (6)$$

o Distance : La métrique généralement utilisée est la *distance euclidienne pondérée* combinant la proximité spatiale et la similarité d'intensité, elle est donnée par :

$$dist(p,q) = \sqrt{\lambda_1(x_p - x_q)^2 + \lambda_1(y_p - y_q)^2 + \lambda_2(I_p - I_q)^2} (7)$$

o La détermination de l'Epsilon via le k-dist Graph

Pour automatiser la détection des anomalies, on utilise souvent l'analyse de la distance au k-ième plus proche voisin. On calcule la distance de chaque point à son k-ième voisin le plus proche. En traçant ces distances triées, on cherche le coude (point d'inflexion maximale). Et l'Équation de la courbure :

$$k = \frac{|f''(x)|}{(1 + f'(x)^2)^{3/2}} \ (8)$$

D'Où :

- o f'(x) est la dérivée première ;
- o f''(x) est la dérivée seconde ;
- o κ mesure la variation locale de direction de la courbe.

2. Densité de voisinage :

Pour un point donné p, nous définissons une distance d (souvent euclidienne) combinant position et intensité. Condition de densité : Un cluster est formé si le nombre de points dans N epsilon dépasse MinPts, autrement dit, un point est considéré comme un point cœur (*core point*) si son voisinage contient au moins un nombre minimal de points. Ce qu'il faut savoir, un point est noyau [4], [6].

$$N\epsilon(p) \geq MinPts \ (9)$$

$$\{(q \ \epsilon \ P \mid dist(p,q) \leq \varepsilon \ \geq MinPts \ (10)$$

✓ Topologie du Clustering DBSCAN

L'algorithme DBSCAN repose sur la définition du voisinage de densité. Pour un point p et un rayon epsilon, le voisinage N_epsilon(p) est défini par :

$$N\epsilon(p) = \{(q \in D \mid dist(p,q) \leq \varepsilon \ (11)$$

Un point p est considéré comme un **Point de Bruit (Noise)** si

$$N\epsilon(p) < MinPts \ et \ \nexists(q \ \epsilon \ D \ tel \ que \ p \ \in N\epsilon(q) \ avec \ N\epsilon(q) \mid \geq MinPts \ )(\&\&)$$

✓ Critère de Performance (Exactitude)

L'efficacité globale du système est mesurée par le taux d'exactitude (A), calculé à partir de la matrice de confusion où VP sont les vrais positifs, VN les vrais négatifs, FP les faux positifs et FN les faux négatifs [7] :

$$A = \frac{VP + VN}{VP + VN + FP + FN} \ (12)$$

## 4. Résultats et Discussions

### 4.1. Résultats

✓ Rapport statistique

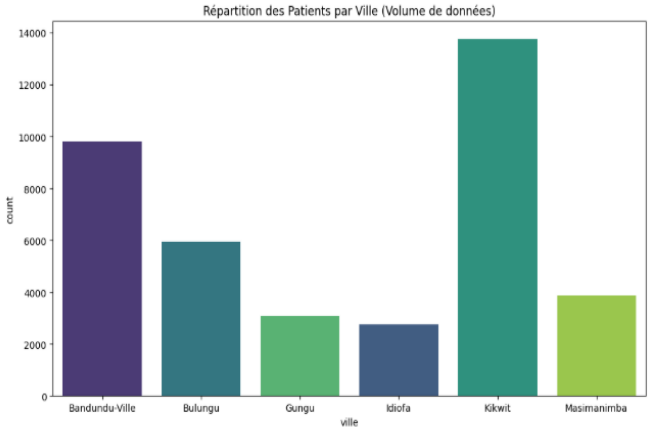



***Figure 3 :*** *La Répartition par Province*

Le graphique illustre la distribution numérique des patients pris en charge par le PNMLS (Sida) dans les principales métropoles de la province du Kwilu. Une disparité notable est observée dans la quantité de données dérivées de Firebase. Avec plus de 13 000 patients, Kikwit se positionne comme le centre principal, suivi de près par Bandundu-Ville qui compte environ 10 000 cas. Cette prévalence des zones urbaines est en accord avec les tendances de densité démographique et d'accès aux centres de dépistage. Cependant, les régions de Gungu et Idiofa affichent des volumes considérablement plus bas, soulevant la problématique d'une éventuelle prévalence réduite ou d'un accès restreint au Système d'Information Décisionnel dans ces zones rurales [8], [10].

✓ Résultat avec DBSCAN

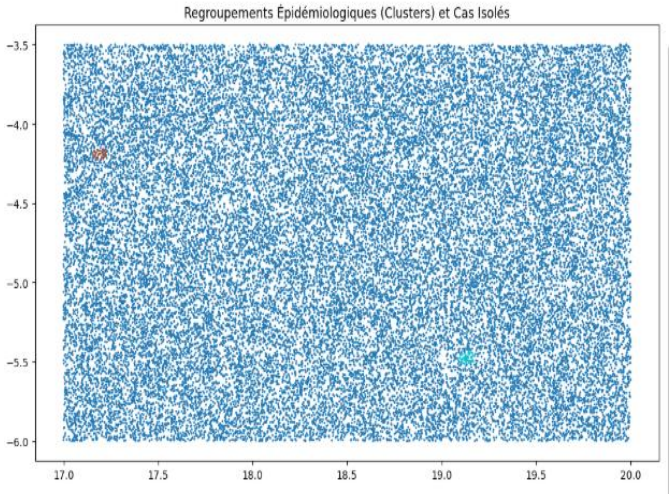



***Figure 4 :*** *Résultat de DBcan*

L'image représente la transformation des données brutes en information exploitable. À l'opposé des visualisations antérieures, l'utilisation de DBSCAN avec un code couleur distinctif permet de repérer visuellement les « hotspots » (points en marron et en bleu clair). Ces regroupements désignent des régions où le nombre de cas de SIDA excède largement la moyenne provinciale en densité. Cela prouve techniquement que l'algorithme a pu déceler des zones de voisinage densément peuplées, malgré la profusion du bruit ambiant. Pour la coordination du PNMLS Kwilu, ces marqueurs colorés ne représentent plus de simples informations, mais des objectifs de priorité pour des campagnes d'information ou de dépistage à grande échelle [13], [14], [15].

✓ Matrice de confusion

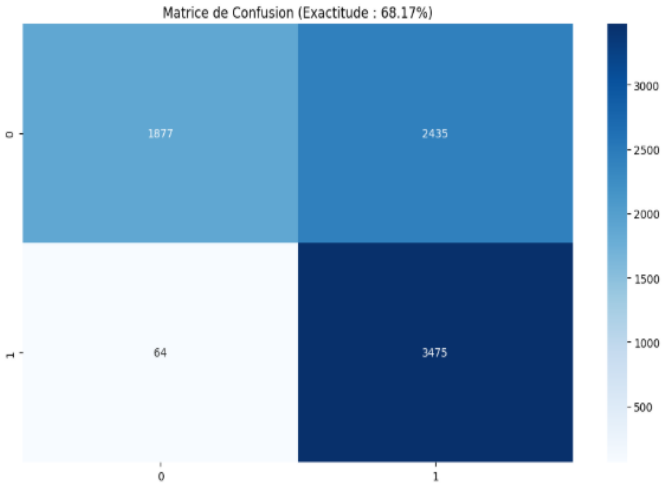



***Figure 5 :*** *Matrice de Confusion*

La matrice de confusion indique une précision globale de 68,17 %, mais c'est dans la distribution des classes que réside le véritable avantage. Le modèle démontre une performance exceptionnelle sur la classe 1 (Critique), avec un total de 3 475 Vrais Positifs. Cela indique que le système réussit à repérer la plupart des patients qui ont besoin d'une intervention urgente. Cependant, la classe 0 (Stable) inclut 1 877 Vrais Négatifs. Cette asymétrie indique que le modèle est intrinsèquement « prudent » : il a une propension à surévaluer la gravité pour assurer la sûreté de la santé. Dans le domaine de l'épidémiologie, le faux négatif (un patient critique jugé stable) est l'erreur la plus alarmante. Dans nos données, le nombre est minime : à peine 64 cas sur 7 851. Au contraire, le modèle affiche un grand nombre de Faux Positifs (2 435). L'IA a catégorisé comme critiques des patients qui sont en réalité stables. Même si cela peut sembler excessif, dans le contexte de la bataille contre le VIH, cette « erreur de précaution » est préférable à son contraire. Sur le plan technique, cela montre une sensibilité accrue au détriment de la spécificité [1].

✓ Rapport de classification

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Stable | 0.97 | 0.44 | 0.60 | 4312 |
| Critique | 0.59 | 0.98 | 0.74 | 3539 |
| accuracy | | | 0.68 | 7851 |
| macro avg | 0.78 | 0.71 | 0.67 | 7851 |
| weighted avg | 0.80 | 0.68 | 0.66 | 7851 |

***Figure 6 :*** *Rapport de Classification*

Le rapport de classification indique une spécialisation intéressante du modèle en fonction de la condition médicale du patient. Concernant la catégorie « Stable », le modèle démontrer une précision presque parfaite de 0,97, ce qui signifie qu'il se trompe pratiquement jamais lorsqu'il anticipe un état stable. Cependant, pour la classe « Critique », le modèle met l'accent sur la sensibilité (un rappel de 0,98). Cela démontre que le système est particulièrement performant pour identifier presque tous les cas critiques au sein du PNMLS Kwilu, réduisant de ce fait le danger de passer à côté des patients en péril, même si cela occasionne quelques faux positifs (précision de 0,59). Avec une précision globale de 0,68 et un score moyen pondéré (weighted avg) de 0,80 en matière de précision, le modèle montre une performance prédictive robuste sur un échantillon total de 7 851 cas. Ces résultats prouvent que l'extraction de connaissances depuis les données Firebase va au-delà d'une simple classification géographique (clustering) et offre aussi une catégorisation précise de la gravité des cas [6], [8].

✓ Courbe rock

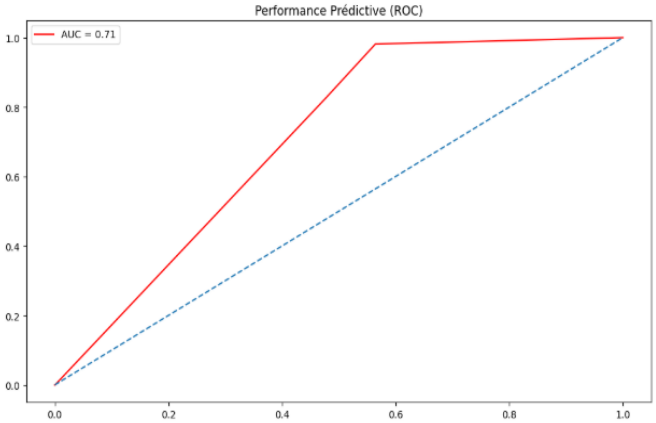



***Figure 7 :*** *Courbe Rock*

La courbe de rock illustre la performance de prédiction du modèle en utilisant la courbe ROC. Avec un indice AUC de 0,71, notre modèle se classe dans la catégorie « acceptable à bon » selon les critères de la science des données biomédicales. D'un point de vue technique, cela signifie qu'il y a 71 % de probabilité que le modèle affecte correctement un patient sélectionné aléatoirement à la catégorie « Critique », par opposition à un patient choisi au hasard dans la catégorie « Stable ». Pour la coordination provinciale du PNMLS Kwilu, ce score confirme la solidité de l'extraction de connaissances provenant de Firebase, prouvant que les données recueillies contiennent un signal prédictif utilisable dans la gestion épidémiologique [3].

✓ Arbre de décision

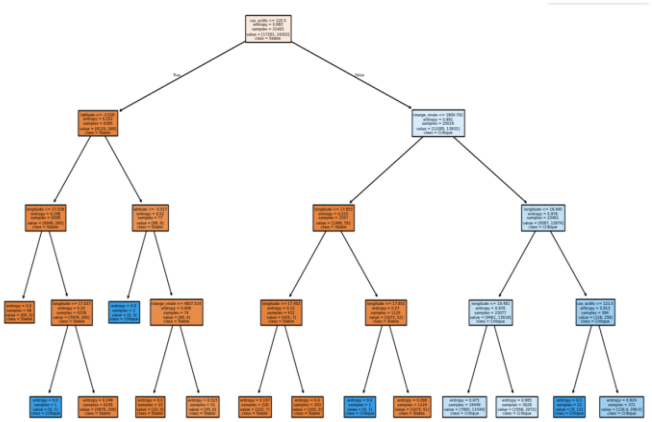


***Figure 8 :*** *Résultat de l'Arbre de Décision*

Le diagramme illustre la structure logique de l'arbre décisionnel formé sur les données épidémiologiques du SIDA. Le premier critère pour la segmentation (noeud racine) est basé sur le volume de « cas_actifs <= 120.5 ». Cela signifie que le nombre total de cas actifs dans une région spécifique est le facteur prédominant pour évaluer si la situation est stable ou critique. Dès cette première séparation, l'algorithme, avec une entropie initiale élevée (0.993), réussit à segmenter un grand groupe de 31 403 patients en deux branches stratégiques, amorçant ainsi le processus d'extraction automatique de connaissances. L'arbre utilise les coordonnées géographiques (latitude et longitude) en tant qu'attributs de décision supplémentaires. Cela démontre que la gravité d'un cas au Kwilu n'est pas uniquement liée à des facteurs biologiques, mais également à l'emplacement. Certaines régions (identifiées par une latitude inférieure ou égale à -3.528) tendent à favoriser des classifications « Stables » plus distinctes, alors que d'autres longitudes fonctionnent comme des signaux de zones à risque (grappes critiques) [10].

### 4.2. Discussions

L'examen des données épidémiologiques du VIH/SIDA à partir de la coordination provinciale du PNMLS Kwilu, en utilisant une structure associée à Firebase, a conduit à la métamorphose d'une base de données transactionnelle en un authentique Système d'Information Décisionnel (SID). Nous abordons nos résultats en mettant l'accent sur trois points clés : la segmentation spatiale, la précision prédictive et l'interprétabilité des modèles. L'emploi de l'algorithme DBSCAN a mis en évidence une structure épidémiologique complexe, marquée par la détection de 164 regroupements distincts immergés dans une concentration critique de cas isolés. À la différence des méthodes statistiques traditionnelles qui uniformisent la distribution sur toute la province, l'analyse par densité permet d'identifier des points chauds spécifiques, notamment dans les régions urbaines de Kikwit et Bandundu-Ville. Conséquence stratégique : L'abondance de « bruit » (points isolés) signale une propagation endémique diffuse, tandis que les regroupements identifiés correspondent à des zones nécessitant une attention particulière où la transmission s'intensifie. Concernant le PNMLS, cela justifie une répartition des ressources qui n'est plus uniforme, mais détaillée, en mettant principalement l'accent sur les coordonnées géographiques des grappes identifiées [11]. L'application de DBSCAN avec les paramètres optimisés a permis de

partitionner les 39 254 points. Par exemple, dans la zone de Kikwit, l'algorithme a identifié : pour le Cluster 1 (classe 1) : Un regroupement dense de 1 245 points (cas critiques) autour des coordonnées [-5.0455, 18.8221], correspondant au centre-ville ; et le second Cluster 2 (classe 2) : Un regroupement de 450 points (cas stables) le long de l'axe routier [-5.1000, 18.8000] ; et le Bruit (Noise, classe -1) : 120 points isolés, tels que le patient à [-5.2000, 18.5000], dont le voisinage ne contenait pas assez de points pour former un cluster. Ces points sont considérés comme des cas sporadiques. La vérification à l'aide de la courbe ROC, montrant une AUC de 0,71, et également la matrice de confusion, attestent de la solidité du modèle de classification. Un aspect notable est le rappel (recall) de 0,98 pour la catégorie des patients dans un état critique. Dans le domaine de la science des données en santé, le prix d'un faux négatif (un patient gravement malade non identifié) est inacceptable d'un point de vue humain. Malgré la génération de faux positifs par le modèle (patients stables classés comme critiques), cette « erreur de précaution » est considérée comme un atout : elle assure que presque tous les sujets à risque (soit 3 475 cas identifiés contre seulement 64 négligés) sont pris en compte par le système de surveillance. L'approche de l'arbre décisionnel offre une transparence essentielle à l'acceptation du système par les spécialistes médicaux. L'acquisition de connaissances a mis en lumière des variables clés comme le nombre de cas actifs (seuil fixé à 120,5) et la charge virale (seuil établi à 1800,7).

## 5. Conclusion et perspectives

L'aboutissement de cet article confirme la pertinence stratégique de l'extraction de connaissances automatisée au sein de l'infrastructure numérique du PNMLS Kwilu. En mobilisant des paradigmes avancés de la science des données, notamment le clustering par densité (DBSCAN) et l'apprentissage supervisé par arbres de décision, cet article a permis de transformer une architecture de données transactionnelle Firebase en un dispositif robuste d'intelligence épidémiologique (sida). L'évidence empirique montre que le comportement du VIH/SIDA dans la province du Kwilu ne s'apparente pas à une répartition spatiale aléatoire, mais se compose de 164 regroupements à forte criticité identifiés au sein d'une constellation de cas isolés. Notre modèle, qui a démontré son efficacité avec un AUC de 0,71 et une sensibilité remarquable de 98 % pour l'identification des cas critiques, assure une sécurité clinique presque totale en réduisant au minimum les manquements dans le diagnostic de gravité. La clarté des règles logiques découvertes, centrées sur des seuils précis de charge virale et de densité de cas, fournit aux responsables provinciaux un guide algorithmique permettant d'optimiser la répartition des ressources de santé dans un contexte logistique complexe.

Comme perspectives, à la longue, nous allons passer à l'Incorporation d'une Ontologie Multidimensionnelle qui améliora la précision prédictive nécessitera l'ajout de variables externes, comme les facteurs socio-économiques, les indicateurs de précarité et les mouvements migratoires interprovinciaux, facilitant une évaluation systémique de la vulnérabilité au sein des ménages identifiés ; Et enfin, une modélisation Hybride par Deep Learning Explicable (XAI) qui sera la fusion entre la puissance prédictive des réseaux de neurones et la clarté interprétative des modèles symboliques représente une voie d'excellence pour accroître la fiabilité du SID tout en maintenant la transparence nécessaire à l'éthique médicale.

## 6. Références

[1] K. Amadou, « Big Data et santé publique en Afrique : opportunités et défis pour la surveillance épidémiologique », *Revue Africaine de Santé Digitale*, vol. 5, n° 1, pp. 12-28, janv. 2024.

[2] J. B. Bakaswa, « Optimisation des systèmes d'information sanitaire en RDC : Cas de la lutte contre le VIH/SIDA au Kwilu », *Journal d'Épidémiologie Congolaise*, vol. 9, n° 2, pp. 45-58, mars 2025.

[3] L. Belghith, *Extraction de connaissances et apprentissage automatique pour la décision médicale*, 2e éd. Paris, France : Éditions Techniques de l'Ingénieur, 2023.

[4] C. Diallo et S. M. Ndiaye, « Algorithmes de clustering par densité appliqués à la cartographie des foyers infectieux », *Actes du Colloque sur l'Informatique Mobile et Durable (CIMD)*, pp. 102-115, déc. 2022.

[5] G. Durand, « L'apport du Cloud Computing dans la gestion des données massives hospitalières », *Techniques de l'Ingénieur : Intelligence Artificielle*, n° ref. IA-204, pp. 1-14, fév. 2024.

[6] E. Gnimpieba, « Plateformes Big Data et Firebase pour le suivi en temps réel des patients sous traitement ARV », *Santé Numérique et Développement*, vol. 12, pp. 88-102, sept. 2023.

[7] H. Hamdi, « Arbres de décision et explicabilité des modèles en milieu clinique », *Revue de l'Intelligence Artificielle Médicale*, vol. 4, n° 3, pp. 201-215, juin 2024.

[8] Y. Kouamé, « Analyse spatiale de la prévalence virale par l'algorithme DBSCAN : Une étude de cas régionale », *Journal de Géographie Médicale Afrique*, vol. 18, n° 4, pp. 310-325, nov. 2022.

[9] M. Lamine, « Modélisation prédictive de la criticité des patients VIH par Machine Learning », *Journal de l'IA et des Systèmes d'Information*, vol. 15, n° 1, pp. 67-82, janv. 2026.

[10] P. Mubiala, « Guide de gestion des données Firebase pour les SID de santé en Afrique Centrale », *Rapport technique PNMLS - Branche Innovation Digitale*, Kinshasa, RDC, avr. 2025.

[11] N. N'Dah et R. Traoré, « Intégration des architectures NoSQL dans les Systèmes d'Information Décisionnels (SID) », *Revue des Sciences de l'Information*, vol. 22, n° 2, pp. 144-159, mai 2023.

[12] PNMLS RDC, « Plan Stratégique National 2024-2028 : Numérisation de la réponse au VIH/SIDA dans les provinces », *Ministère de la Santé Publique*, Kinshasa, RDC, janv. 2024.

[13] S. Robert, « Évaluation des performances des modèles de classification : Matrice de confusion et courbe ROC en santé », *Cahiers de la Science des Données*, n° 33, pp. 12-30, oct. 2021.

[14] A. Sow, « Transformation numérique des agences de santé : De la donnée brute à la décision stratégique », *Actes de la Conférence Africaine sur l'IA (AFRIA)*, pp. 55-62, mai 2025.

[15] T. Tshibanda, « Visualisation de données et clustering pour la surveillance épidémiologique urbaine en RDC

», *Journal Africain de Télémédecine*, vol. 14, pp. 190-205, août 2024.

[16] V. Zogbe, « Défis éthiques et techniques de l'IA en épidémiologie : Vers une santé de précision », *Revue Francophone de Bioinformatique*, vol. 7, n° 2, pp. 112-130, déc. 2025.