

Royaume de Belgique

Projet de spécialisation IA

**Rapport de choix de modèle**

Par Harold Rajaonarison

Année Académique 2024 - 2025

**Table des matières**

[**Liste de figures** i](#_Toc199693671)

[**Introduction** 1](#_Toc199693672)

[**1.** **Filtrage des utilisateurs à segmenter** 2](#_Toc199693673)

[**1.1.** **Justification du seuil minimal** 2](#_Toc199693674)

[**1.2.** **Application du seuil et réduction de la base** 3](#_Toc199693675)

[**2.** **Préparation des variables de segmentation** 3](#_Toc199693676)

[**2.1.** **Sélection des variables pertinentes** 3](#_Toc199693677)

[**2.2.** **Normalisation et transformation** 3](#_Toc199693678)

[**3.** **Modélisation : comparaison de 3 modèles** 4](#_Toc199693679)

[**3.1.** **Modèle KMeans** 4](#_Toc199693680)

[**3.2.** **Modèle Agglomerative Clustering** 5](#_Toc199693681)

[**3.3.** **Modèle DBSCAN** 6](#_Toc199693682)

[**4.** **Analyse des clusters retenus** 7](#_Toc199693683)

[**4.1.** **Choix du modèle optimal** 7](#_Toc199693684)

[**4.2.** **Profilage et segments** 8](#_Toc199693685)

[**4.3.** **Nomenclature des clusters** 9](#_Toc199693686)

[**Conclusion** 10](#_Toc199693687)

**Liste de figures**

[Figure 1 - Distribution du total d'événements par utilisateur 2](#_Toc199687125)

[Figure 2 - Total de données après filtrage 3](#_Toc199687126)

[Figure 3 - Méthode du coude - Modèle KMeans 5](#_Toc199687127)

[Figure 4 - Score de silhouette moyen selon k - Modèle KMeans 5](#_Toc199687128)

[Figure 5 - Score de silhouette du modèle Agglomerative Clustering 6](#_Toc199687129)

[Figure 6 - Score de silhouette du modèle DBSCAN 7](#_Toc199687130)

[Figure 7 - Profil moyen des utilisateurs par cluster 8](#_Toc199687131)

**Introduction**

Face à une baisse du chiffre d’affaires sur certains produits de sa marque propre, la plateforme de commerce en ligne Amazing cherche à mieux comprendre les comportements de ses utilisateurs. Pour cela, une stratégie de segmentation basée sur les données d’interaction est envisagée afin d’identifier des groupes d’utilisateurs homogènes et actionnables. Ces segments permettront d’optimiser les recommandations de produits, les actions marketing ciblées, ainsi que les analyses stratégiques internes.

Ce rapport s’inscrit dans la continuité d’un premier travail ayant permis d’explorer et de structurer les données issues de plus de 15 millions d’utilisateurs. À partir de ce jeu agrégé, le présent document vise à concevoir et comparer plusieurs modèles de segmentation non supervisée afin de catégoriser au mieux les habitudes d’achat.

L’approche adoptée repose sur plusieurs étapes : filtrer les utilisateurs faiblement actifs, sélectionner et normaliser les variables pertinentes, puis appliquer trois algorithmes de clustering différents (KMeans, Agglomerative Clustering, DBSCAN). Le modèle le plus adapté sera ensuite sélectionné sur la base de critères de qualité statistique et de lisibilité métier. Une interprétation des segments permettra enfin de dégager des profils types exploitables par l’entreprise.

1. **Filtrage des utilisateurs à segmenter**
   1. **Justification du seuil minimal**

Avant d’engager un processus de modélisation, il est indispensable d’identifier les utilisateurs ayant une activité suffisamment significative pour être analysés de manière fiable. Pour ce faire, une nouvelle variable appelée total\_events a été construite, regroupant l’ensemble des interactions d’un utilisateur via la somme du nombre de vues (count\_view), d’ajouts au panier (count\_cart) et d’achats (count\_purchase). La figure suivante illustre la distribution de cette variable sur l’ensemble des utilisateurs.

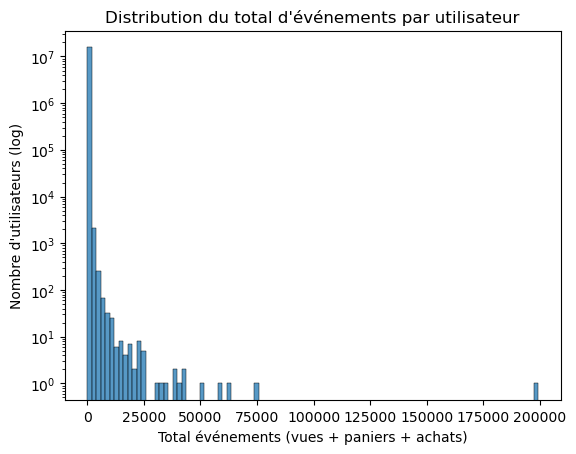


Figure 1 - Distribution du total d'événements par utilisateur

La très forte concentration autour des faibles valeurs montre que la majorité des utilisateurs ont eu une activité limitée, voire marginale. Afin de ne pas introduire de bruit dans la modélisation, un seuil minimal de 5 événements a été fixé. Ce seuil permet de conserver uniquement les utilisateurs ayant exprimé un comportement d’achat suffisant, tout en maintenant un volume de données exploitable pour l’analyse.

* 1. **Application du seuil et réduction de la base**

L’application de ce filtrage a permis d’exclure les comptes passifs et les visites ponctuelles sans réel engagement. Concrètement, seuls les utilisateurs dont le total d’événements est supérieur ou égal à cinq ont été conservés pour la suite du traitement. Cette opération a permis de réduire le volume de données à une portion plus ciblée, correspondant aux profils pertinents pour une segmentation comportementale.

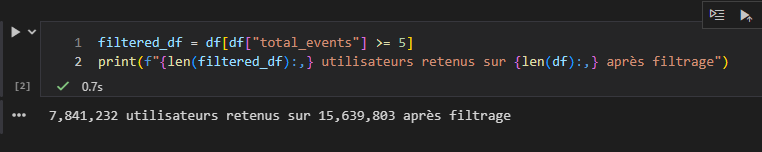


Figure 2 - Total de données après filtrage

1. **Préparation des variables de segmentation**
   1. **Sélection des variables pertinentes**

L’objectif de cette étape est de sélectionner les variables les plus représentatives du comportement utilisateur pour permettre une segmentation pertinente. Le fichier user\_features.parquet, filtré pour ne conserver que les utilisateurs actifs, contient des informations issues de l’agrégation d’événements tels que les vues, les ajouts au panier et les achats.

Les variables sélectionnées pour l’analyse sont exclusivement numériques, continues ou discrètes, et décrivent des aspects variés du comportement d’achat :

* count\_view : nombre de produits consultés
* count\_cart : nombre d’articles ajoutés au panier
* count\_purchase : nombre d’achats réalisés
* active\_days : nombre de jours distincts d’activité
* unique\_sessions : nombre de sessions différentes
* recency\_days : nombre de jours depuis la dernière activité
* total\_spent : montant total dépensé
* avg\_purchase\_price : panier moyen

Ces variables couvrent trois dimensions clés : la fréquence et l’intensité d’interaction, la valeur client, la récence comportementale.

* 1. **Normalisation et transformation**

La qualité d’un algorithme de segmentation repose sur la manière dont il évalue la proximité entre les individus. Mais dans notre cas, les variables choisies pour représenter les comportements utilisateurs présentent à la fois des unités différentes et des distributions très déséquilibrées, souvent marquées par une forte asymétrie à droite. Ces caractéristiques compromettent la validité des mesures de distance, notamment euclidiennes, qui supposent des échelles comparables et une distribution relativement homogène.

Certaines variables telles que count\_view, total\_spent ou avg\_purchase\_price comportent des valeurs extrêmes qui sont rares mais très élevées. Ces extrêmes influencent fortement la distance entre individus et créent des effets de levier injustifiés. L’usage de la fonction log1p(x) permet de réduire l’échelle des grandes valeurs tout en conservant l’ordre relatif des individus. Cela rend les distributions plus proches de la normalité, ce qui est particulièrement adapté pour des méthodes basées sur la distance comme KMeans ou l’agglomératif.

Les variables suivantes ont été transformées de cette manière : count\_view, count\_cart, count\_purchase, total\_spent, avg\_purchase\_price.

Même après transformation logarithmique, les variables gardent des échelles différentes. Par exemple, une variable peut varier entre 0 et 20 tandis qu’une autre varie entre 0 et 5. Pour éviter que les plus grandes plages ne dominent le calcul des distances, chaque variable est ensuite centrée réduite à l’aide d’un StandardScaler, ce qui revient à transformer chaque variable pour qu’elle ait une moyenne de 0 et un écart-type de 1. Cette étape est indispensable pour que chaque dimension contribue équitablement à la segmentation.

Nous obtenons ainsi un tableau X\_scaled dans lequel chaque variable est sur une base comparable et aucune ne domine artificiellement la structure des distances. Ce tableau peut désormais être utilisé comme entrée directe pour les algorithmes de clustering, garantissant une évaluation plus juste et fiable des regroupements.

1. **Modélisation : comparaison de 3 modèles**
   1. **Modèle KMeans**

Le KMeans est l’un des algorithmes de clustering les plus populaires pour la segmentation client. Il partitionne les individus en *k* groupes en minimisant la variance intra-groupe, chaque client étant affecté au cluster dont le centroïde est le plus proche selon la distance euclidienne. Sa rapidité d'exécution et sa capacité à générer des groupes bien séparés en font un choix privilégié, bien qu’il repose sur l’hypothèse que les clusters soient sphériques et de taille comparable.

Pour déterminer le nombre optimal de clusters, deux méthodes complémentaires ont été mobilisées. La méthode du coude, représentée en figure 3, permet de visualiser l’évolution de l’inertie intra-cluster selon différentes valeurs de *k*. Le point d’inflexion observé suggère une segmentation optimale au-delà de laquelle l’ajout de nouveaux clusters n’apporte plus de gain significatif.

En complément, le score de silhouette (figure 4) évalue la cohésion et la séparation des clusters formés. Ce score, compris entre -1 et 1, atteint un maximum lorsque les individus sont proches de leur propre groupe tout en étant éloignés des autres. L’analyse conjointe de ces deux indicateurs confirme le choix d’un *k* optimal équilibré, garantissant une segmentation à la fois fiable et interprétable.

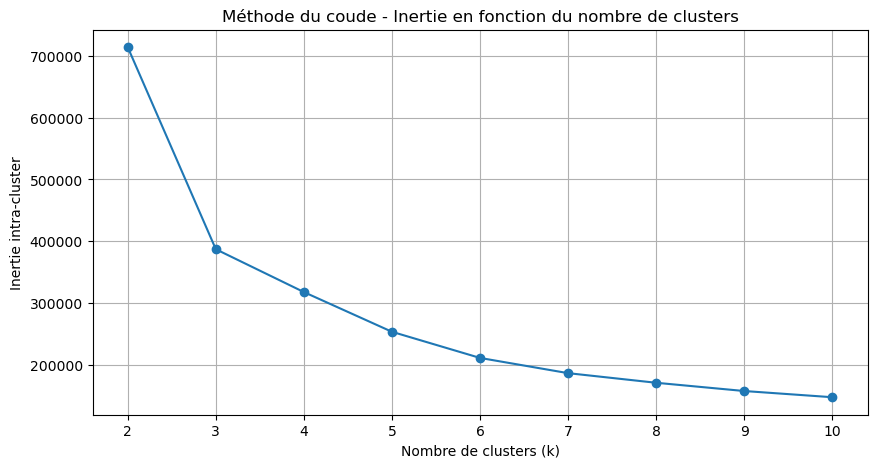


Figure 3 - Méthode du coude - Modèle KMeans

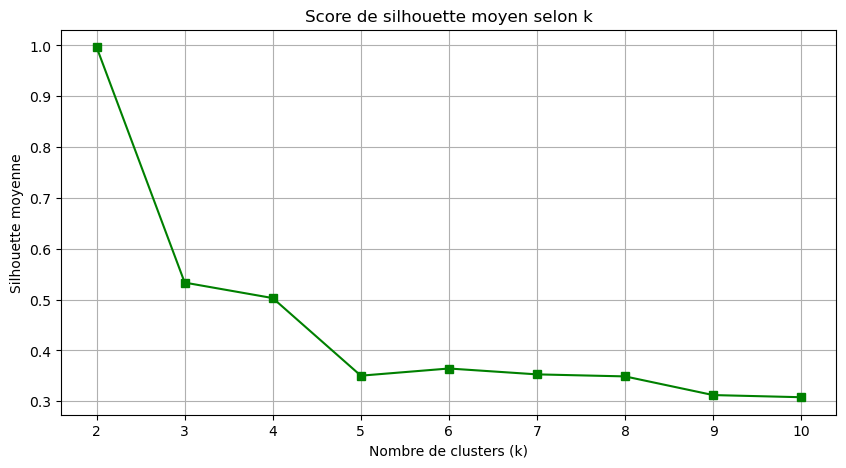


Figure 4 - Score de silhouette moyen selon k - Modèle KMeans

* 1. **Modèle Agglomerative Clustering**

L’agglomérative clustering appartient à la famille des algorithmes hiérarchiques, construisant une arborescence de regroupement à partir des données. Le processus débute avec chaque client comme un cluster individuel, puis fusionne itérativement les paires de clusters les plus similaires jusqu’à atteindre un nombre défini de groupes. Cette approche permet de révéler des structures de données complexes sans imposer de forme prédéfinie aux clusters. Dans le cadre de cette étude, l’algorithme a été appliqué sur un échantillon réduit pour éviter une surcharge mémoire, du fait de sa forte complexité algorithmique. L’évaluation du nombre de clusters a été conduite à l’aide du score de silhouette, mesuré pour différentes valeurs de *k*.

Comme l’illustre la figure 5, la silhouette moyenne suggère une structuration modérée des groupes, avec des performances inférieures à celles observées avec KMeans. Cette observation confirme que, bien que plus souple dans la détection de formes variées, l’approche agglomérative est moins adaptée à une segmentation massive comme celle des utilisateurs de la plateforme étudiée, où la scalabilité et la compacité des clusters sont prioritaires.

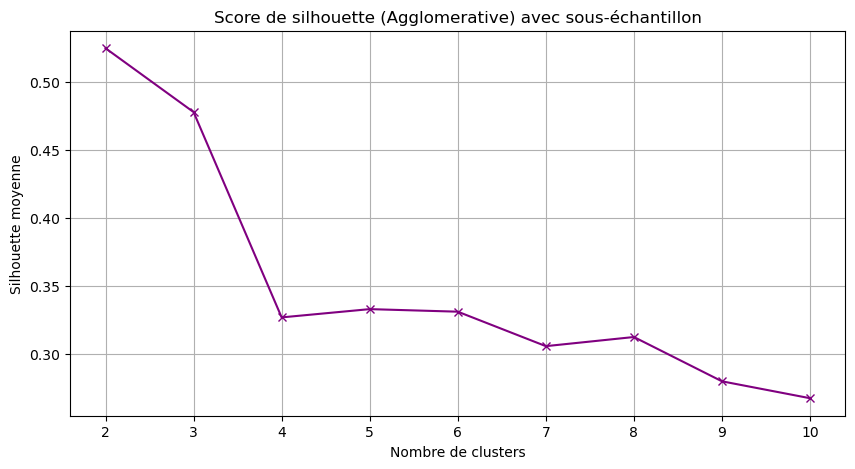


Figure 5 - Score de silhouette du modèle Agglomerative Clustering

* 1. **Modèle DBSCAN**

Le modèle DBSCAN (Density-Based Spatial Clustering of Applications with Noise) repose sur une approche non paramétrique qui regroupe les clients en fonction de la densité locale des points dans l’espace des variables. Contrairement à KMeans ou à l’agglomérative clustering, il ne nécessite pas de spécifier le nombre de clusters à l’avance. Il identifie automatiquement les régions denses comme des clusters et considère les zones isolées comme du bruit, ce qui en fait une méthode particulièrement robuste pour détecter des structures arbitraires et résister aux valeurs aberrantes.

Toutefois, DBSCAN est très sensible au choix des deux paramètres ε rayon de voisinage) et minPts (nombre minimum de points dans un voisinage), et peut échouer à capter des regroupements significatifs si les densités sont trop variables. Dans cette étude, bien que la méthode ait correctement détecté des noyaux compacts, elle a également marqué une part importante des données comme bruit, comme illustré dans la figure 6. Cette limitation réduit la pertinence du modèle pour une segmentation exhaustive des clients, surtout dans un contexte aussi vaste.

En comparaison avec les autres algorithmes testés, DBSCAN présente une performance limitée sur des volumes massifs aux profils hétérogènes, malgré son intérêt théorique pour la détection d'anomalies.

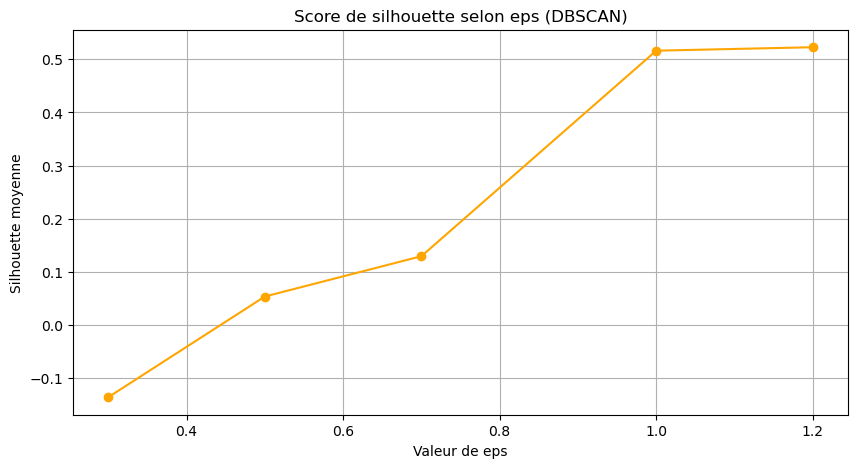


Figure 6 - Score de silhouette du modèle DBSCAN

1. **Analyse des clusters retenus**
   1. **Choix du modèle optimal**

Après avoir appliqué et analysé les trois algorithmes de segmentation, KMeans, Agglomerative Clustering et DBSCAN sur des échantillons représentatifs de notre base utilisateur, le modèle KMeans s’est distingué par sa robustesse, sa rapidité d’exécution et sa capacité à produire des clusters bien séparés et interprétables. L’analyse de la courbe d’inertie a permis d’identifier un point d’inflexion clair, indiquant une configuration optimale à cinq clusters.

Cette conclusion a été renforcée par le score de silhouette moyen, qui a atteint un pic notable pour cette même valeur, traduisant une bonne cohésion intra-cluster et une séparation inter-clusters satisfaisante. Le modèle Agglomerative Clustering a montré une performance relativement comparable, mais avec une latence significative sur des échantillons de grande taille et une instabilité potentielle liée à la complexité quadratique de l’algorithme. Quant à DBSCAN, bien qu’il soit pertinent pour détecter des regroupements denses et des profils atypiques, il a échoué à couvrir la totalité des données, excluant une part importante d’utilisateurs comme bruit.

Au regard de ces éléments, le modèle KMeans est retenu comme modèle optimal pour la catégorisation des comportements d’achat des clients de la plateforme Amazing. Il présente un compromis équilibré entre performance, scalabilité et interprétabilité, ce qui en fait un candidat solide pour une intégration dans une interface de simulation destinée à prédire en temps réel le profil d’un nouvel utilisateur.

* 1. **Profilage et segments**

Une fois le modèle KMeans final entraîné avec **k = 6**, choix validé sur la base du **compromis optimal entre performances (score de silhouette) et interprétabilité métier**, chaque utilisateur a été assigné à l’un des six segments identifiés. Afin de caractériser ces segments, une analyse du **profil moyen** a été menée à partir des principales variables comportementales : nombre de vues, ajouts au panier, achats réalisés, total dépensé, panier moyen, nombre de jours actifs et récence d’activité.

Cette analyse révèle **six profils clients bien distincts**. Le **cluster 5** regroupe les **acheteurs intensifs**, très actifs sur tous les plans et fortement contributeurs au chiffre d'affaires. Le **cluster 0** correspond aux **acheteurs engagés**, réguliers dans leurs interactions et avec une récence d’activité très faible. Le **cluster 2**, avec un panier moyen élevé mais une fréquence d’achat plus faible, est identifié comme un groupe d'**acheteurs premium**. Le **cluster 3** se compose de visiteurs relativement curieux, mais qui ne concrétisent pas leurs intentions d’achat, des **explorateurs non convertis**. Le **cluster 1** rassemble des **visiteurs passifs**, très peu actifs et sans contribution commerciale. Enfin, le **cluster 4** se distingue par une faible activité globale et une récence élevée, typique des **clients inactifs**.

Ces profils sont représentés dans une **carte thermique** facilitant la visualisation des écarts comportementaux entre segments. Cette représentation permet d’orienter des **stratégies marketing différenciées**, en activant des leviers spécifiques pour chaque type de client : fidélisation, réactivation, conversion ou valorisation.

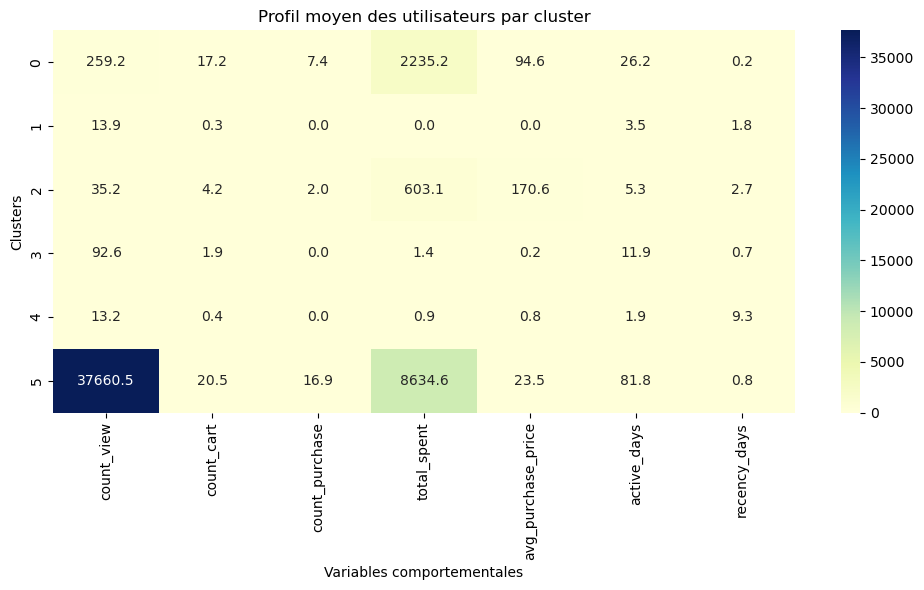


Figure 7 - Profil moyen des utilisateurs par cluster

* 1. **Nomenclature des clusters**

Après l’analyse, des noms sont attribués à chaque cluster pour une meilleure compréhension.

|  |  |  |
| --- | --- | --- |
| **Cluster** | **Description synthétique** | **Nom** |
| 0 | Activité régulière, achats modérés, récence faible | Acheteur engagé |
| 1 | Très faible activité, aucun achat | Visiteur passif |
| 2 | Achats occasionnels, panier moyen élevé | Acheteur premium |
| 3 | Navigation moyenne, peu ou pas d’achats | Explorateur non converti |
| 4 | Faible activité, récence élevée | Client inactif |
| 5 | Très actif, panier élevé, nombreux achats, contributeur clé | Acheteur intensif |

Figure 8 - Nomenclature des clusters

**Conclusion**

Ce projet avait pour objectif de construire un modèle de segmentation des utilisateurs de la plateforme Amazing afin de mieux comprendre et catégoriser leurs comportements d’achat. Après une phase rigoureuse de filtrage des utilisateurs actifs, les variables comportementales ont été sélectionnées, normalisées et utilisées comme base pour l'entraînement de plusieurs algorithmes de clustering. Trois méthodes ont été comparées : KMeans, Agglomerative Clustering et DBSCAN.

L’analyse a mis en évidence que le modèle **KMeans**, appliqué sur un échantillon de 100 000 utilisateurs, offrait un excellent compromis entre performance, lisibilité et applicabilité opérationnelle. Sa capacité à générer des clusters bien séparés, associée à sa rapidité d’exécution et à sa compatibilité avec la prédiction en temps réel, en fait un choix robuste pour une future intégration dans une interface interactive.

Le profilage des segments obtenus a permis d’identifier des comportements variés : utilisateurs passifs, acheteurs réguliers, clients très engagés ou profils atypiques. Cette segmentation ouvre la voie à des actions marketing personnalisées, telles que des campagnes ciblées, des offres adaptées ou des alertes de réengagement.

Enfin, le modèle retenu a été exporté et préparé pour être utilisé dans une interface Streamlit, dans le but de simuler en temps réel le cluster auquel un nouvel utilisateur pourrait appartenir, sur la base de ses premières interactions. Cette étape finale marque le lien entre l’analyse de données et l’expérience utilisateur concrète, illustrant ainsi l’utilité directe de l’intelligence artificielle appliquée à la personnalisation des services.