BigCookingData

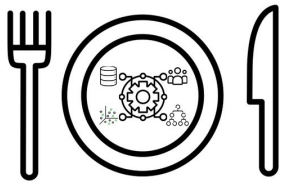
Arthur MIMOUNI
Mamadou Bella DIALLO
Mathieu SAUVAGEOT
Marouane RACHIDY
Imane CHBIRA

Rapportrice : Mme. TZOMPANAKI Katerina
Tuteur technique : M. VODISLAV Dan
Encadrant de gestion de projet : M. LIU Tianxiao

Master IISC - UE Projet de synthèse

CY Cergy Paris Université

16 Juin 2022



Sommaire

Mise en scénario

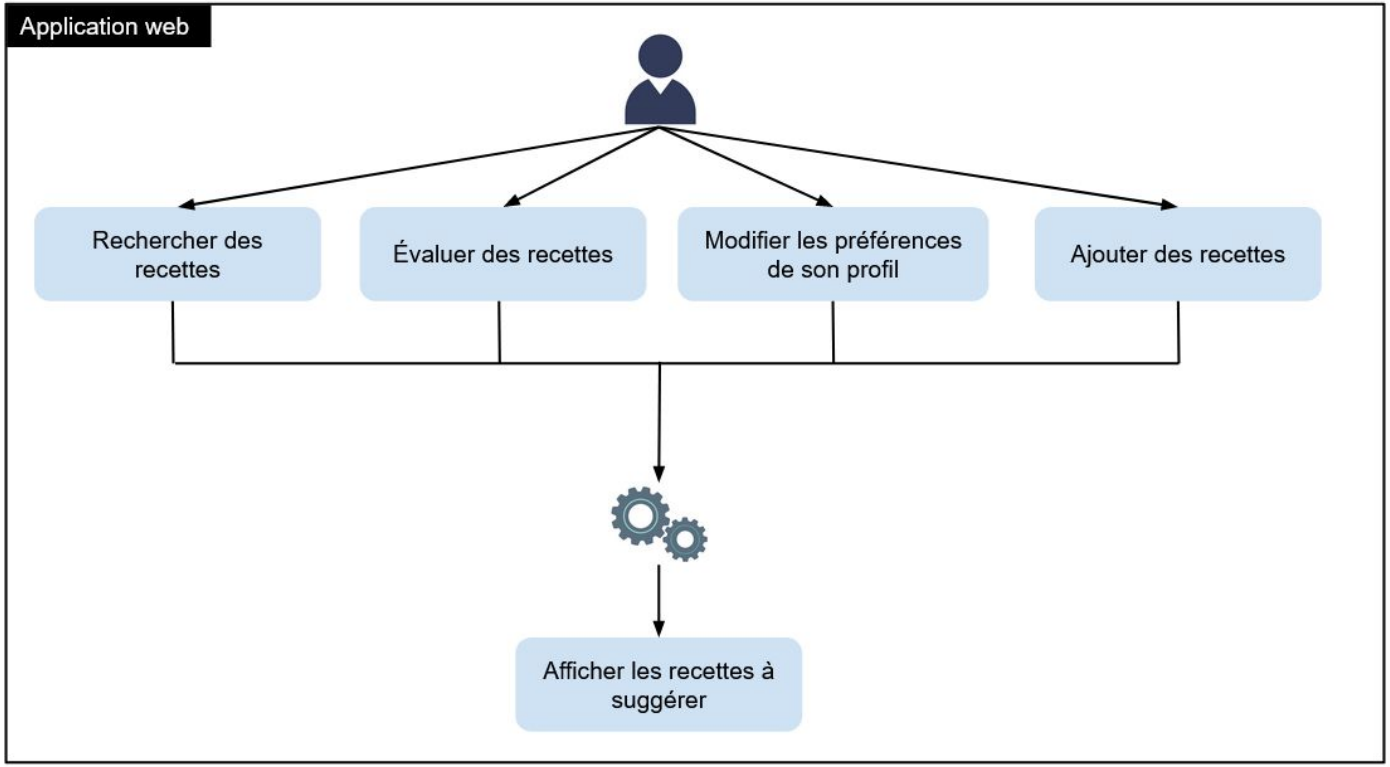


Schéma: Mise en scénario de l'utilisation du système

Objectif du projet

Traitement des recettes	Regroupement des recettes similaires	Classification des recettes	Algorithme de suggestion
Collecte des recettes sur le site web “Marmiton”	Construire un Kmeans clustering	Construire un arbre de décision	Récupération des données de l’utilisateur
Nettoyer et structurer les recettes	Évaluer le partitionnement	Évaluer la précision de l’arbre	Trouver les recettes similaires aux données de l’utilisateur
Insérer les recettes dans un fichier JSON			

Table : Objectifs à réaliser pour les quatres parties essentielles

Architecture technique globale

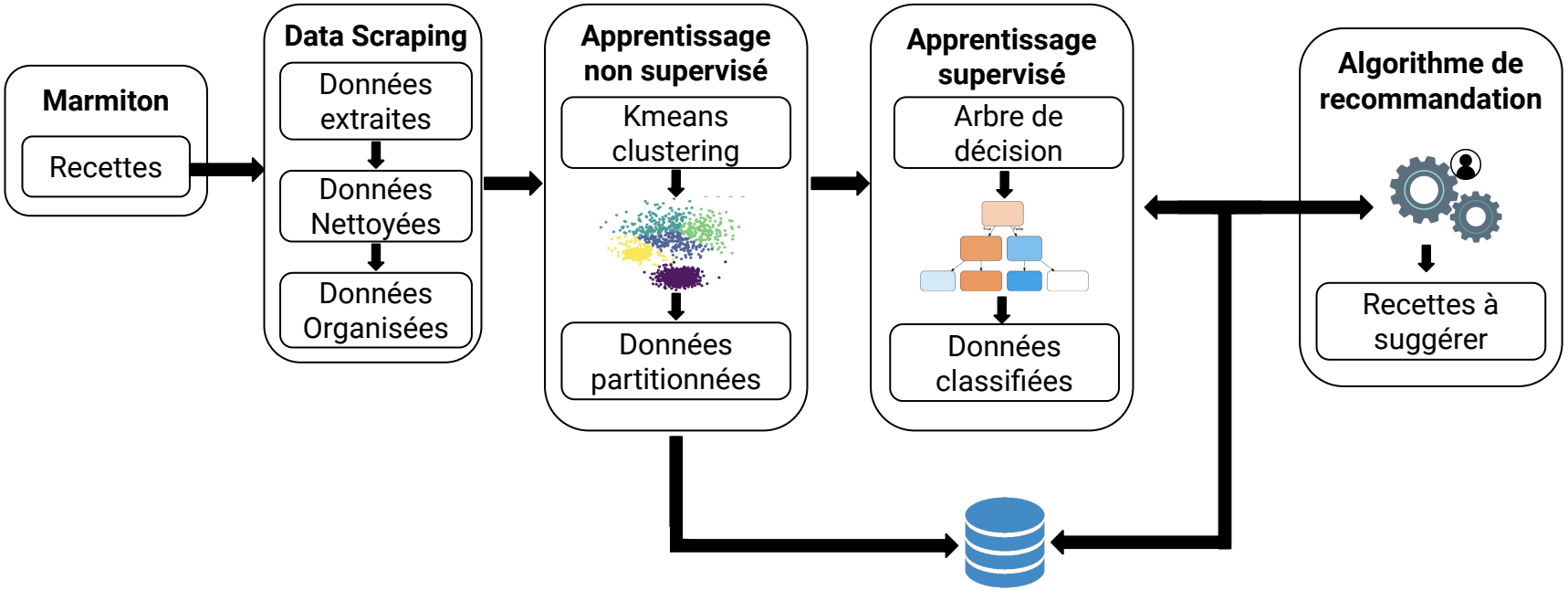
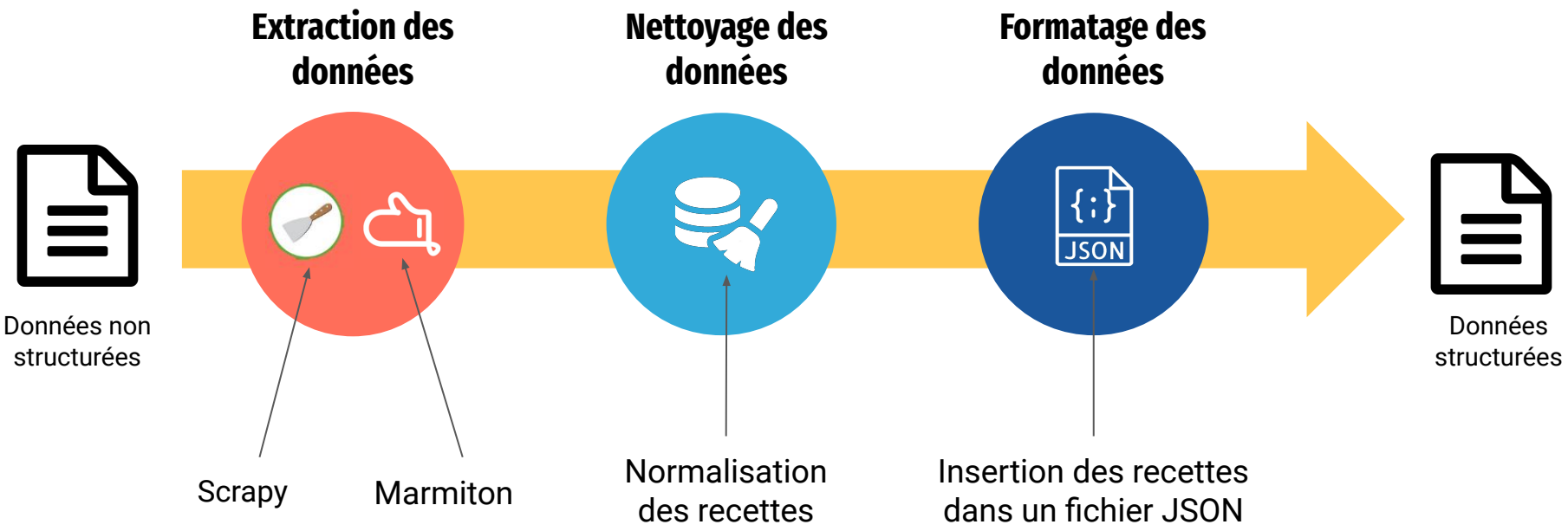


Schéma: Architecture des différentes parties techniques du projet

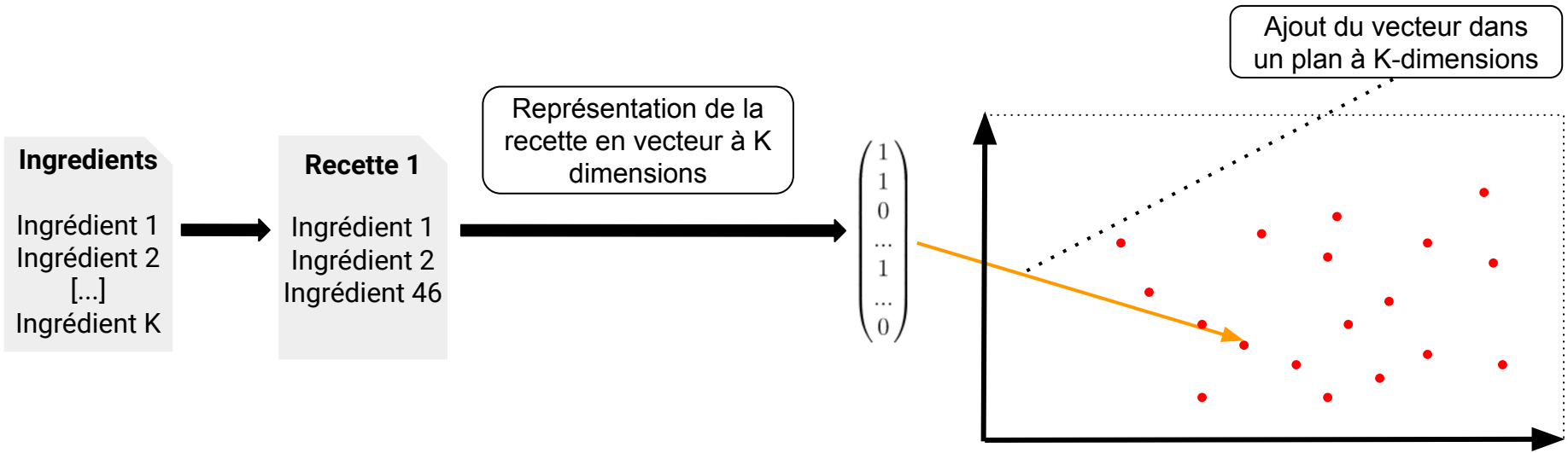
Collecte et nettoyage des données



Kmeans Clustering

- **Objectif du Kmeans ?** → Partitionner les recettes dans des clusters pour les catégoriser

- La première étape est de définir chaque recette comme un point en K-dimension
  - ➢ Chaque dimension représente un ingrédient



Estimation du partitionnement

- Réduction de dimensionnalité via **PCA** pour nos vecteurs de recettes
- Récupération du nombre optimal de cluster par la méthode **Elbow**

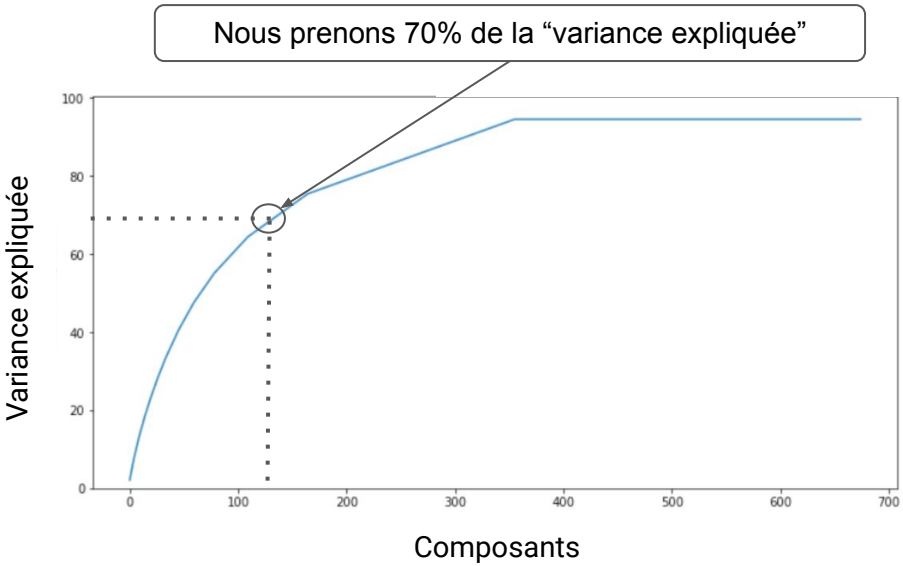


Schéma : Réduction de dimensions via PCA

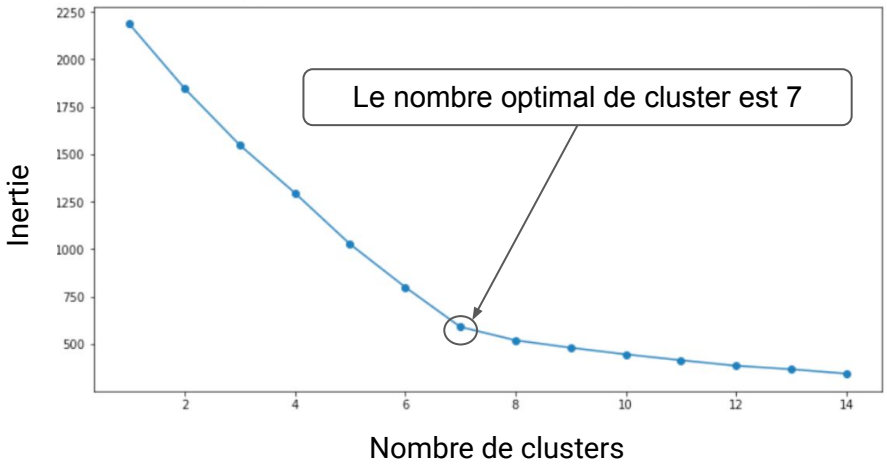


Schéma : Méthode Elbow par l’inertie

Évaluation du partitionnement

- Pour mesurer la qualité de notre partitionnement : coefficient de la **silhouette**

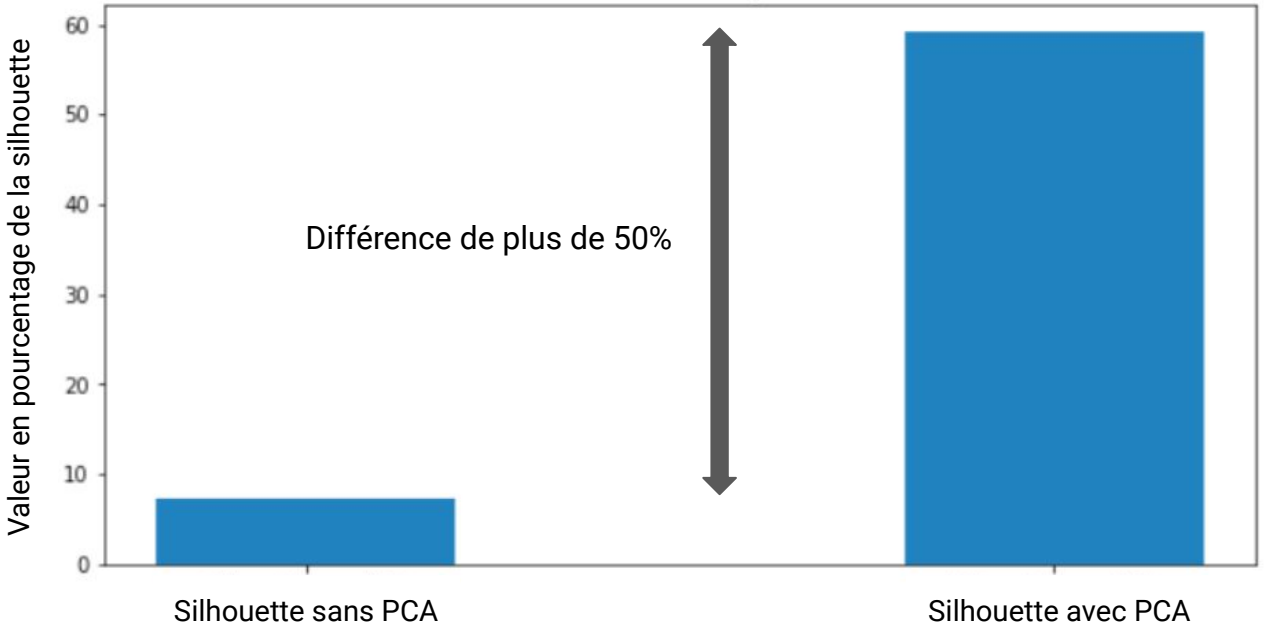
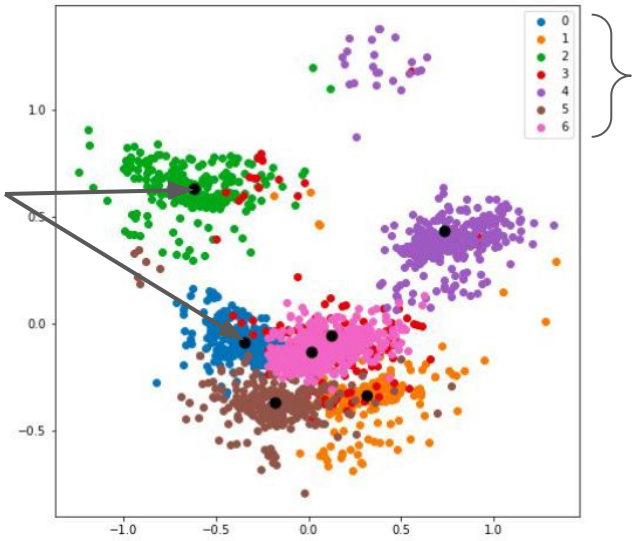


Schéma : Analyse du pourcentage de la silhouette avant et après redimensionnement de nos vecteurs via PCA

Résultat du clustering

- Visualisation des données partitionnées par notre Kmeans Clustering dans un plan 2D et 3D

Centroïdes



Clusters

Schéma: Plan 2D - Kmeans Clustering

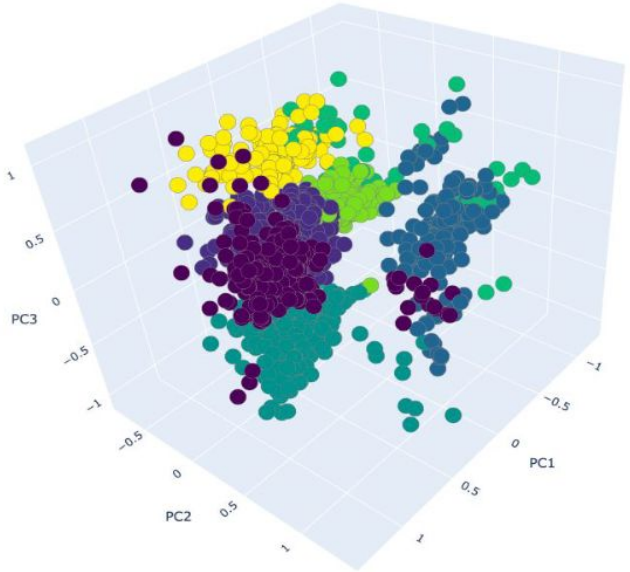
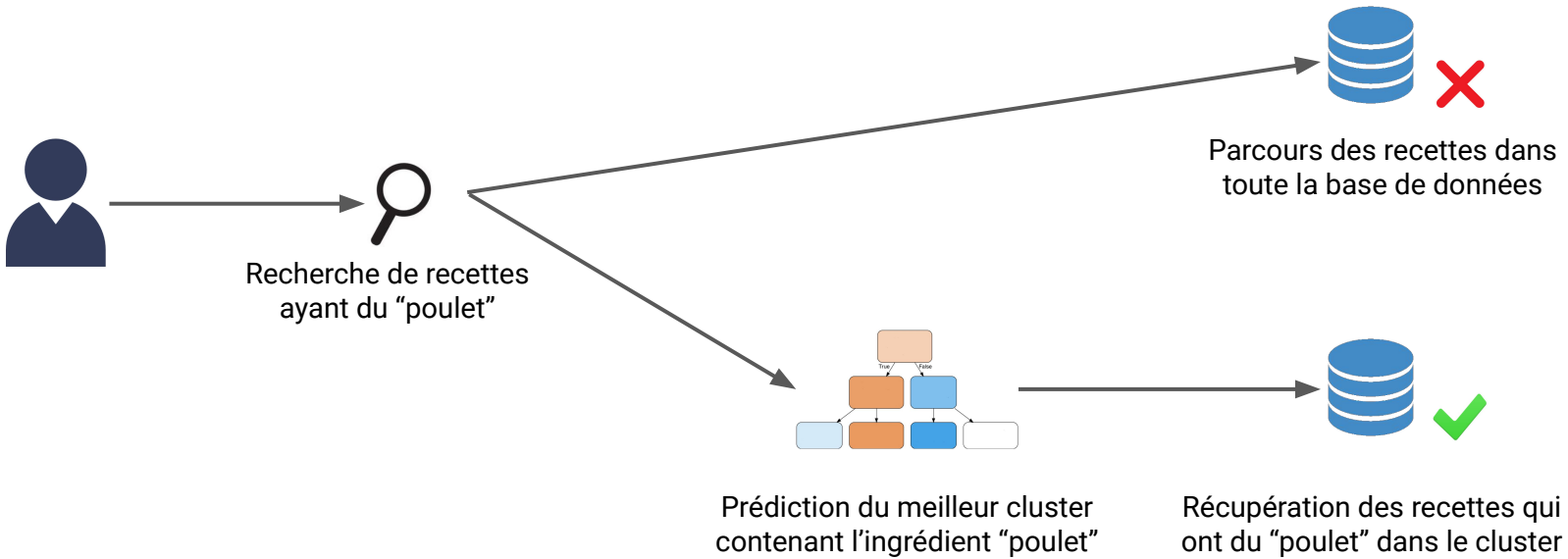


Schéma: Plan 3D - Kmeans Clustering

Arbre de décision

- **Objectif de l'arbre de décision ?** → Trouver le meilleur cluster contenant les ingrédients souhaités
- **Intérêt d'un arbre de décision ?** → Permet de réduire la zone de recherche dans la base de données.



Création des règles de décision

- La première étape est de construire nos règles de décision permettant d'entraîner notre arbre de décision.

Règles de décision

	Ingrédient 1	Ingrédient 2	Ingrédient 3	[...]	Ingrédient N	Numéro de cluster
Recette 1	1	1	0	[...]	1	1
Recette 2	0	0	1	[...]	1	6
[...]	1	1	0	[...]	0	2
Recette M	0	0	1	[...]	1	6

Vecteurs binaires des recettes

Label de la règle

Évaluation de l'arbre de décision

- Utilisation d'une **validation croisée K-Fold** pour trouver la meilleure profondeur de l'arbre de décision.

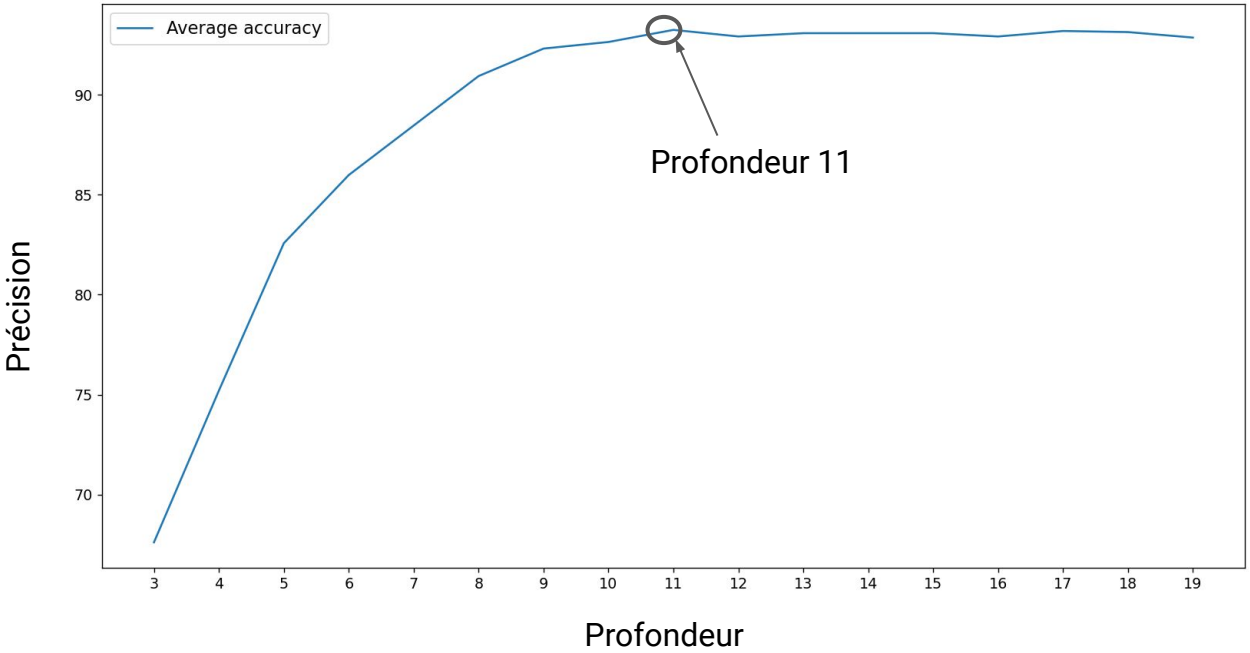


Schéma: Précision d'un arbre avec différentes profondeurs

Extrait de l'arbre de décision

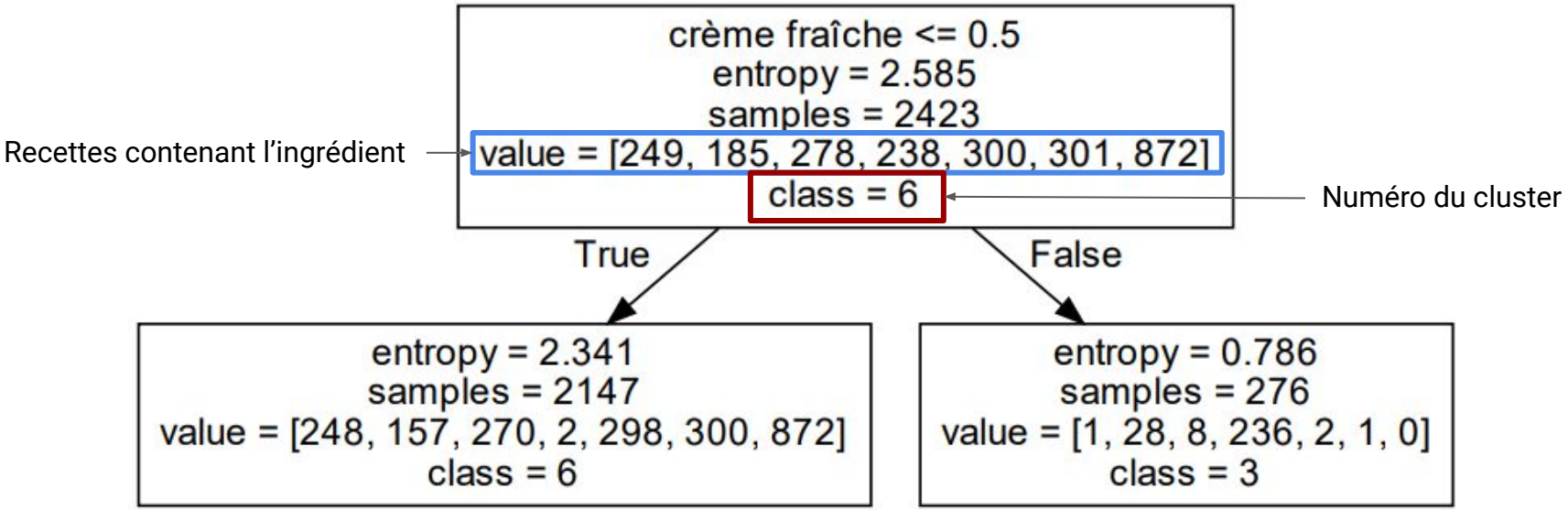
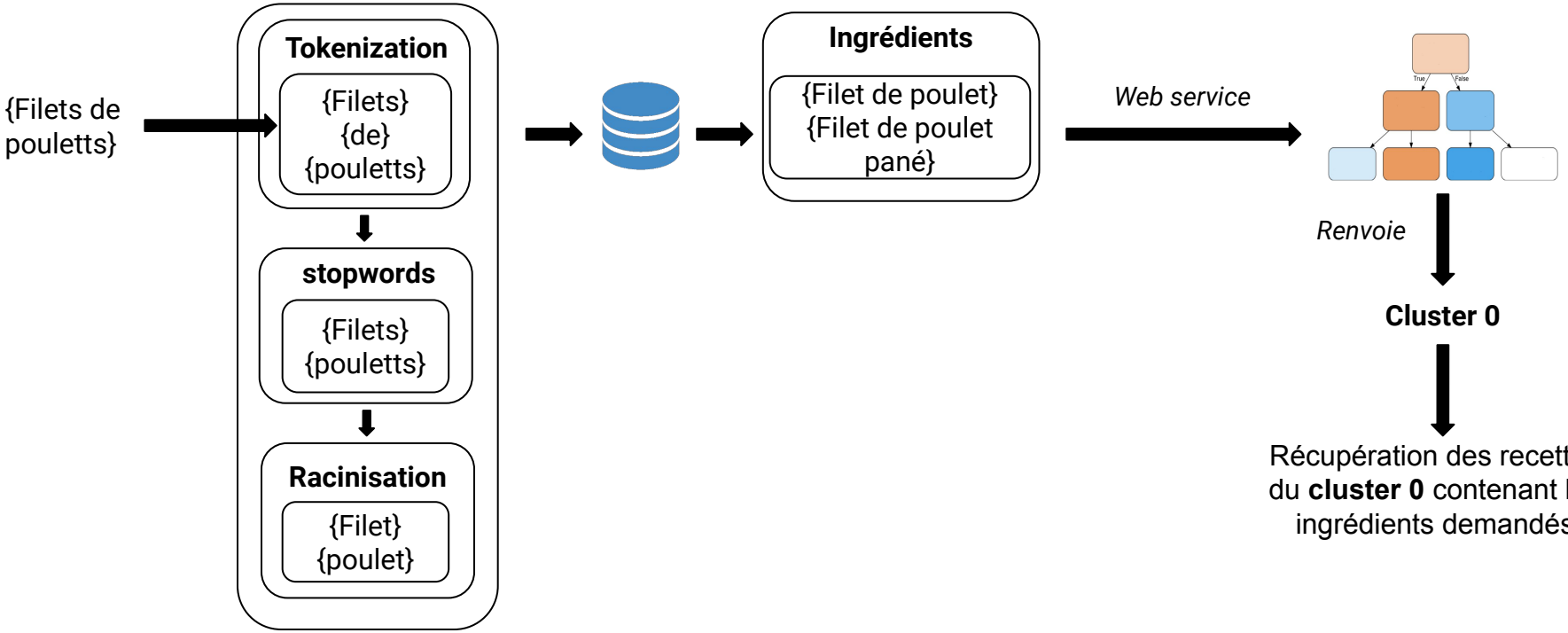


Schéma: Exemple d'arbre de décision avec une profondeur de 1

Cas d'utilisation de l'arbre de décision

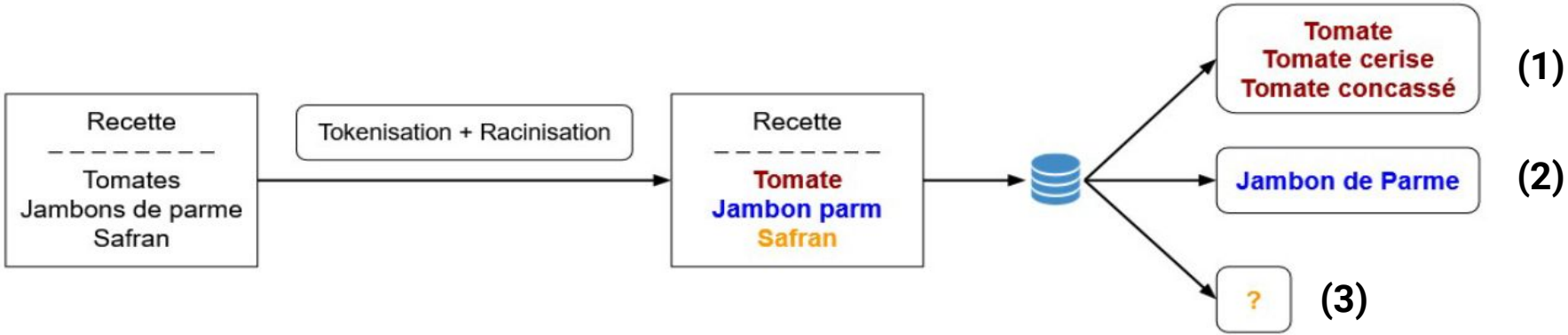
- L'utilisateur à la possibilité d'effectuer des recherches de recettes basées sur des ingrédients.



Ajout de nouvelle recette (1)

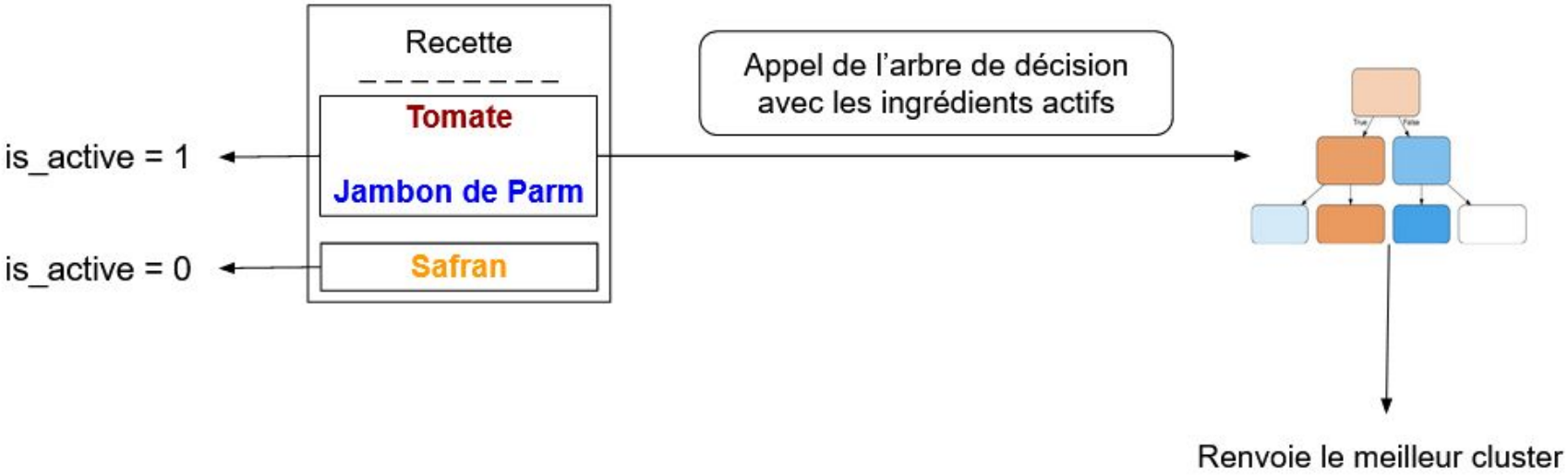
- Prétraitement des ingrédients de la nouvelle recette

- Récupération des ingrédients similaires dans la base de données
  - ➢ Plusieurs ingrédients similaires **(1)**
  - ➢ Un seul ingrédient similaire **(2)**
  - ➢ Aucun ingrédient similaire **(3)**



Ajout de nouvelle recette (2)

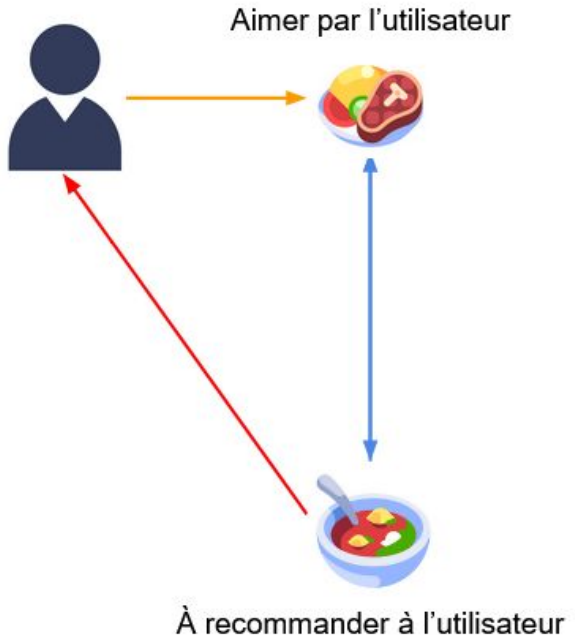
- Recherche du meilleur cluster pour la nouvelle recette
  - ➢ Appel de l'arbre avec les ingrédients actifs de la recette
  - ➢ Prédiction de l'arbre et renvoie du meilleur cluster



Algorithme de filtrage par contenu

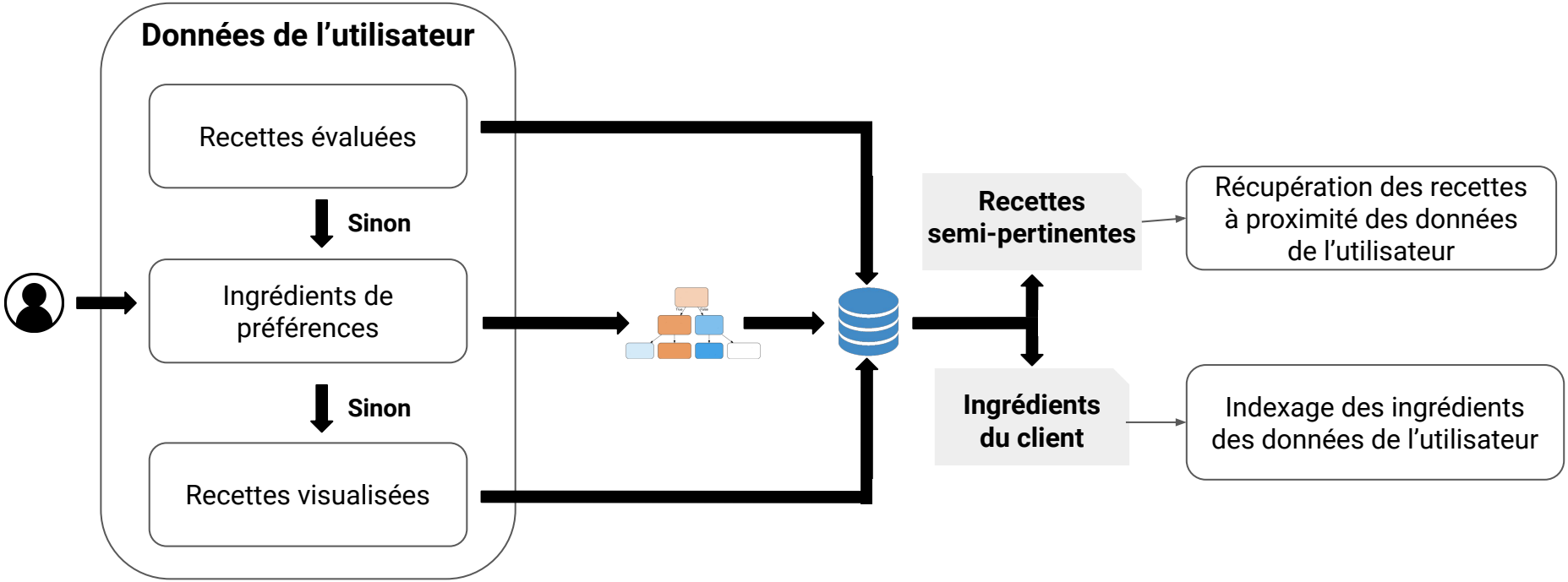
- **Objectif de l'algorithme ?** → Proposer des recettes selon le contenu des recettes aimé par l'utilisateur

- **Pourquoi un filtrage par contenu ?** → Permet de proposer des recettes populaires ET non populaires



Récupération des données

- La première étape est de récupérer les <u>données de l'utilisateur</u> et les <u>recettes semi-pertinentes</u>



Recherche des recettes similaires

- La deuxième étape est d'effectuer les calculs vectoriels entre recettes à l'aide de la **similarité cosinus**

$$u(\boldsymbol{x}, \boldsymbol{i}) = \cos(\boldsymbol{x}, \boldsymbol{i}) = \frac{\boldsymbol{x} \cdot \boldsymbol{i}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{i}||}$$

	I1	I2	I3	I4	I5	I6
R1	1	0	0	0	1	0
R2	0	1	0	0	1	0
R3	0	1	0	0	0	1
R4	0	0	1	0	1	0
Client	0.2	0.004	0.24	0.3	0.15	0.106

$(0.2*1 + 0.15*1) / (\sqrt{0.2^2 + 0.15^2} * \sqrt{1^2 + 1^2}) = 0.99$

0.72

0.73

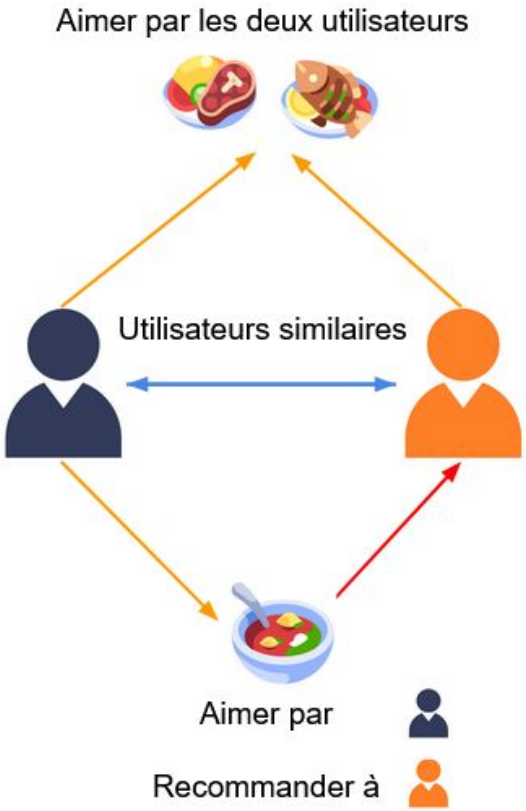
0.97

On récupère les recettes dont le score est plus élevé que la moyenne des scores

Filtrage collaboratif User-User

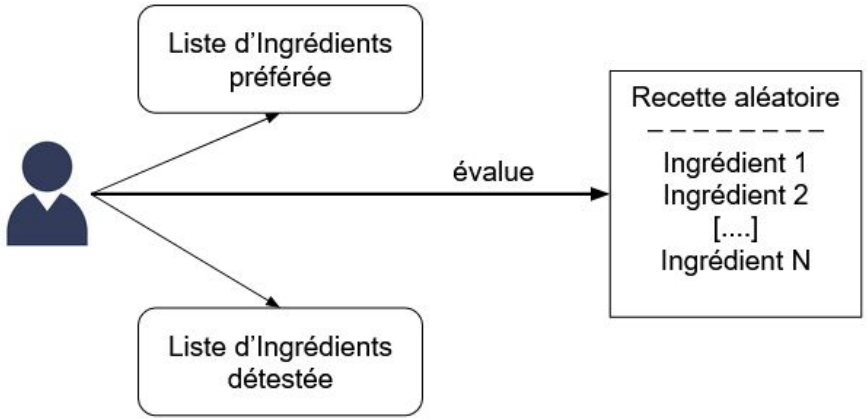
- **Objectif de l'algorithme ?** → Effectuer des comparaisons de performance entre nos deux algorithmes

- **Pourquoi un filtrage collaboratif ?** → Permet de proposer des recettes selon les préférences des utilisateurs similaires



Création des clients et de la matrice d'utilité

- **Phase 1** : Création des utilisateurs / clients



- **Phase 2** : Création de la matrice d'utilité

	R1	R2	R3	R4	R5
U1	4	?		5	
U2	5	5	4		
U4		3		4	2

Calcul des prédictions de recettes

- **Phase 3** : Calcul des prédictions des recettes pour l'utilisateur ciblé

$$S(x,y) = cos(x,y) = \frac{r_x \cdot r_y}{\|r_x\| \cdot \|r_y\|}$$

Similarité cosinus entre l'utilisateur x et y

$$r_{xi} = \frac{\sum_{y \in N} S_{xy} * r_{yi}}{\sum_{y \in N} S_{xy}}$$

Prédiction de la recette i pour l'utilisateur x

$S(U1, U2) = 0.092$

$S(U1, U4) = 0.82$

	R1	R2	R3	R4	R5
U1	4	?		5	
U2	5	5	4		
U4		3		4	2

$$r_{U1,R2} = \frac{(0.092 * 5) + (0.82 * 3)}{0.092 + 0.96} = 2.77$$

Comparaison des performances

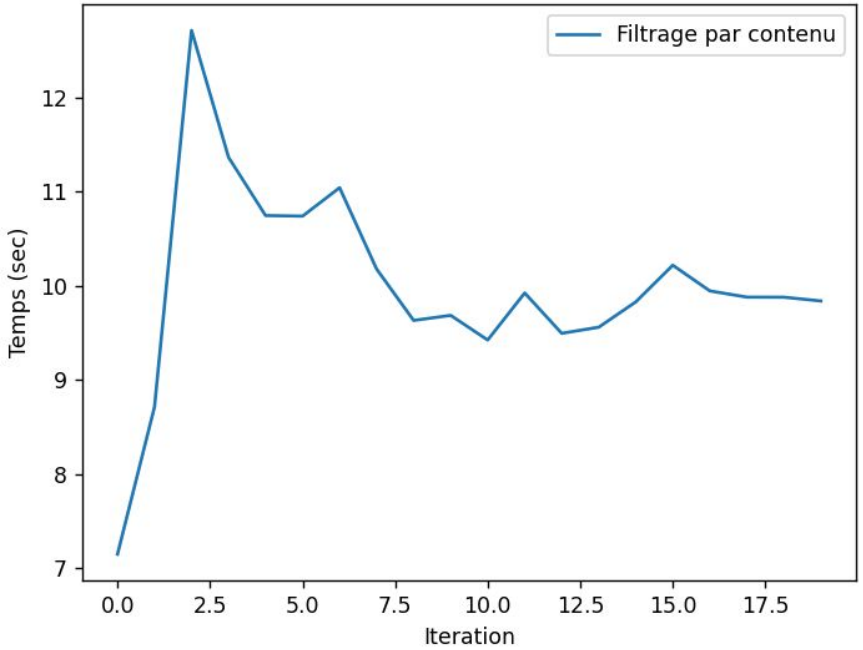


Schéma: Temps d'exécution du filtrage par contenu

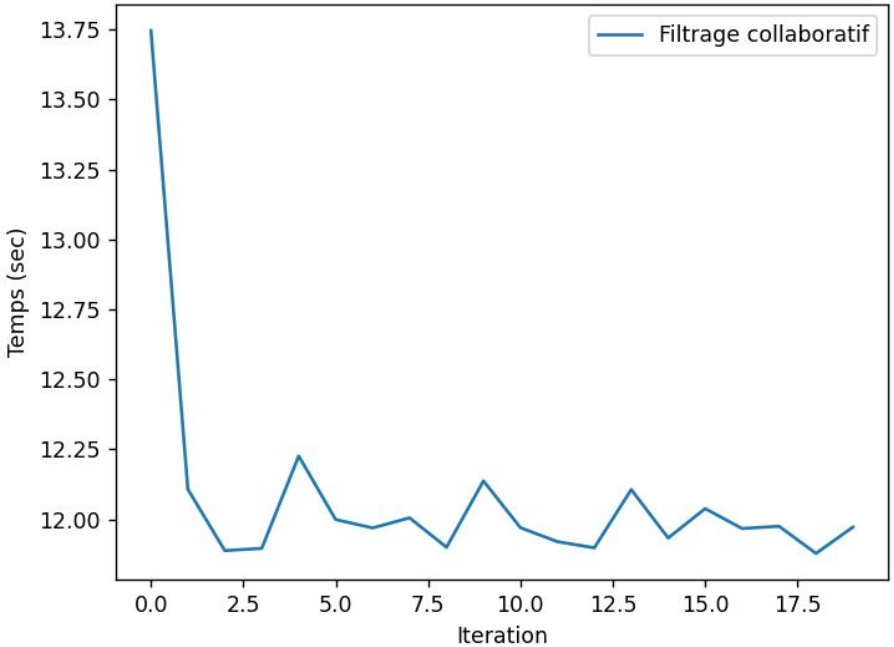


Schéma: Temps d'exécution du filtrage collaboratif

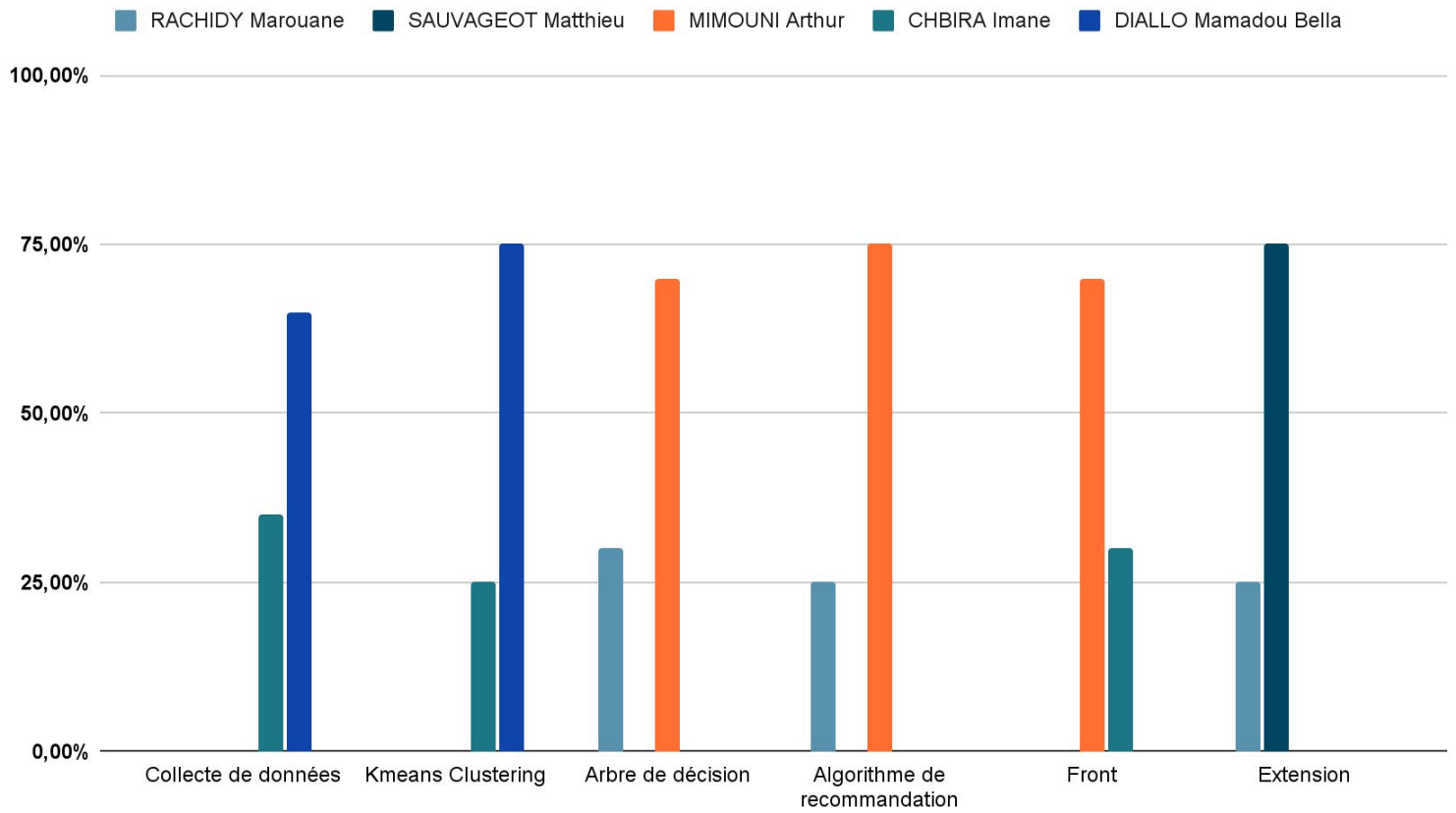
Méthode de gestion

- **Cycle de vie du projet**
  - ➢ Identification des besoins du client
  - ➢ Planification des tâches et constitution des groupes
  - ➢ Exécution du plan d'action et communication entre les membre de l'équipe
  - ➢ Évaluation et analyse des résultats finaux

- **Versions du projet**

Version 1	Version 2	Version 3	Version finale
Extraction et nettoyage des recettes	Création des règles de décision de l'arbre	Création de l'algorithme de filtrage par contenu	Création de l'algorithme de filtrage collaboratif
Transformation des recettes en vecteurs	Évaluation et Implémentation de l'arbre de décision	Implémentation de l'IHM graphique	Implémentation des extensions du projet.
Implémentation du Kmeans Clustering		Fonctionnalité d'ajout de nouvelles recettes	

Répartition de tâches



Conclusion et état d'avancement

Collecte des données

Ce qui a été fait : la collecte et le nettoyage des données ont été conçus

Ce qu'il reste à faire : RAS

Algorithmes de suggestion

Ce qui a été fait : les deux algorithmes ont été correctement implémentés

Ce qu'il reste à faire : RAS

Arbre de décision

Ce qui a été fait : L'arbre de décision a été correctement conçu et permet de faire des prédictions correctes

Ce qu'il reste à faire : Interface en temps réel pour détecter la précision de l'arbre

Kmeans Clustering

Ce qui a été fait : Le partitionnement a été effectué de la manière la plus optimale possible

Ce qu'il reste à faire : RAS

Perspectives et extensions au projet

- Nous avons trois extensions possibles pour notre projet :

Planificateur de repas *Difficulté : 3/5*	**KMeans Clustering avec les quantités des ingrédients** *Difficulté : 2/5*	**Inférence en temps réel de l'arbre de décision** *Difficulté : 4/5*
Planifier des repas sur une semaine selon les préférences de l'utilisateur	Prendre en considération la quantité des ingrédients au lieu de leur appartenance	Détecter en temps réel, les anomalies des prédictions de l'arbre de décision
Créer un calculateur de calories selon les caractéristiques de l'utilisateur	Effectuer une comparaison de performance avec le Kmeans Clustering implémenté	