|  |  |
| --- | --- |
| Document technique  Modèle OLTP, OLAP en étoile et IA | Résumé  Dans le document qui suit, vous retrouverez la base de données OLTP et OLAP avec leurs explications. Puis les modèles d’IA réalisés l’aide de KNIME. Vous aurez la possibilité d’étudier les scripts et les workflows en pièces jointes.  AUTEURS   * BETEMPS Quentin * EL ABLAK DIDOUCHE Sammy * PERROUX Valentin * TASDEMIR Sefer * TINASTEPE Feyza   SAE S4.C.01 : développement avec une base de données et visualisation |

**SOMMAIRE**

[1. Modèle OLTP 3](#_Toc195259312)

[a) Base 3](#_Toc195259313)

[b) Vues 4](#_Toc195259314)

[c) Optimisation OLTP 4](#_Toc195259315)

[d) Insertion 7](#_Toc195259316)

[2. Modèle OLAP 9](#_Toc195259317)

[a) Base 9](#_Toc195259318)

[b) Explication 10](#_Toc195259319)

[c) Insertion 10](#_Toc195259320)

[3. Modèle KNIME 11](#_Toc195259321)

[a) Modèle de visualisation + clustering 11](#_Toc195259322)

[b) Modèle IA de calcul de Nutri-Score 19](#_Toc195259323)

[c) Modèle IA de l’Open Fact Food Score (OFFS) 26](#_Toc195259324)

[d) Modèle IA d’attribution de l’OFFS 29](#_Toc195259325)

[4. Power BI 30](#_Toc195259326)

[e) Préparation et importation des données 30](#_Toc195259327)

[f) Modèle OLTP 31](#_Toc195259328)

[g) Modèle OLAP 33](#_Toc195259329)

[5. Conclusion : 39](#_Toc195259330)



ABSTRACT

This report is based on Open Food Facts (OFF), an open-source platform that collects food data entered by volunteers. The project involves optimizing an existing database by improving its quality and performance, while also incorporating newly simulated data on the sales volumes of food products. To meet these requirements, we developed an optimized OLTP relational model to reduce redundancy and the overall volume of processed data.  
  
We then carried out data analysis and visualization through a workflow built on Knime. In parallel, we created analytical reports in English using Power BI. We subsequently implemented several artificial intelligence models in Knime, the first to automate the assignment of the Nutri-Score despite the absence of certain essential data (fiber and sodium), thus enabling the detection of potential input errors. A second model was developed to calculate a new nutritional score, based solely on the available data. Finally, a last model was built to automatically assign this score during product entry.

All these AI models were developed according to a rigorous approach aimed at minimizing their environmental impact, using only a standard PC for computation. The outcomes of this project include an optimized database model, powerful data visualization tools, and robust, efficient AI models that bring real added value to the Open Food Facts platform.

# Modèle OLTP

## Base

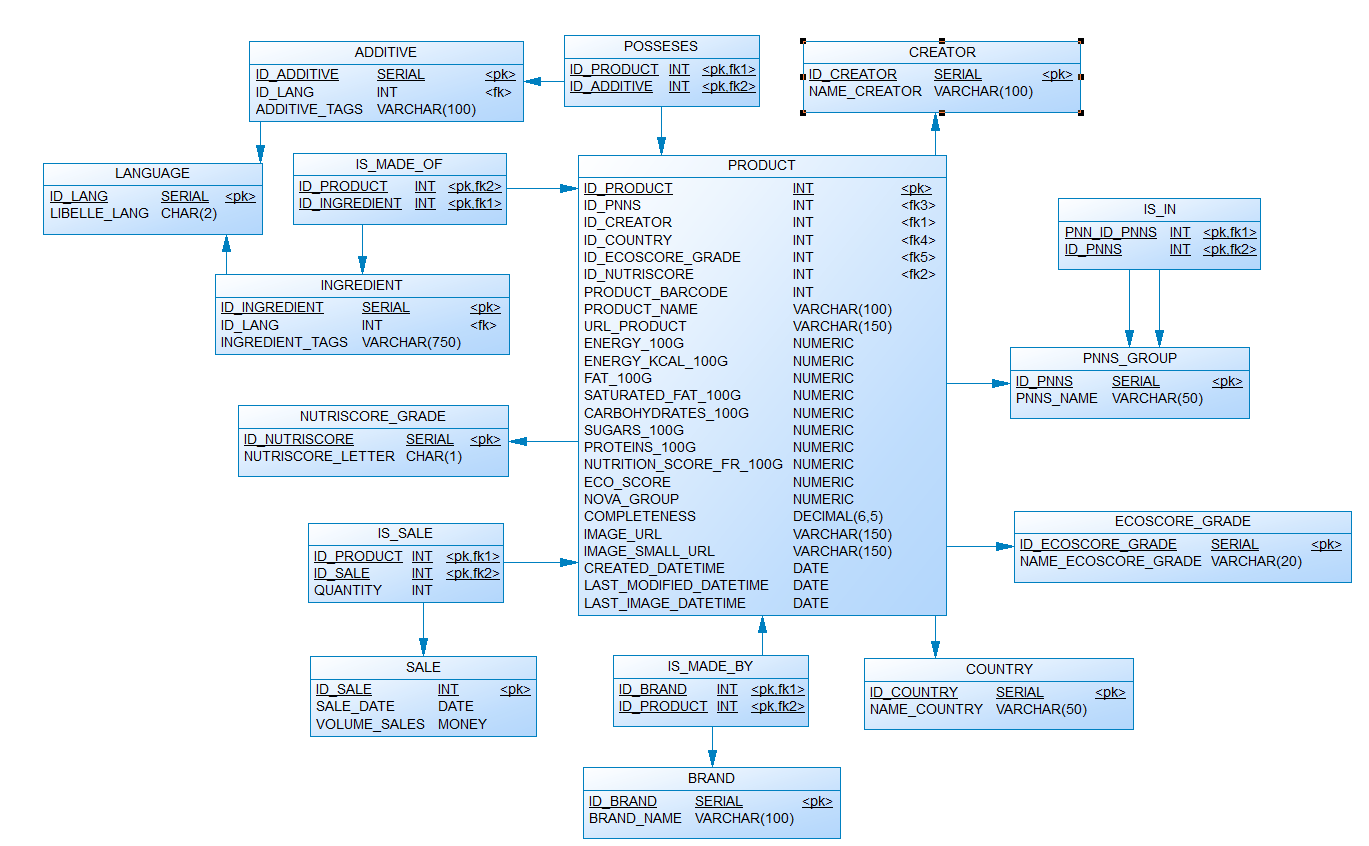


Figure 1 - MLD

## Vues

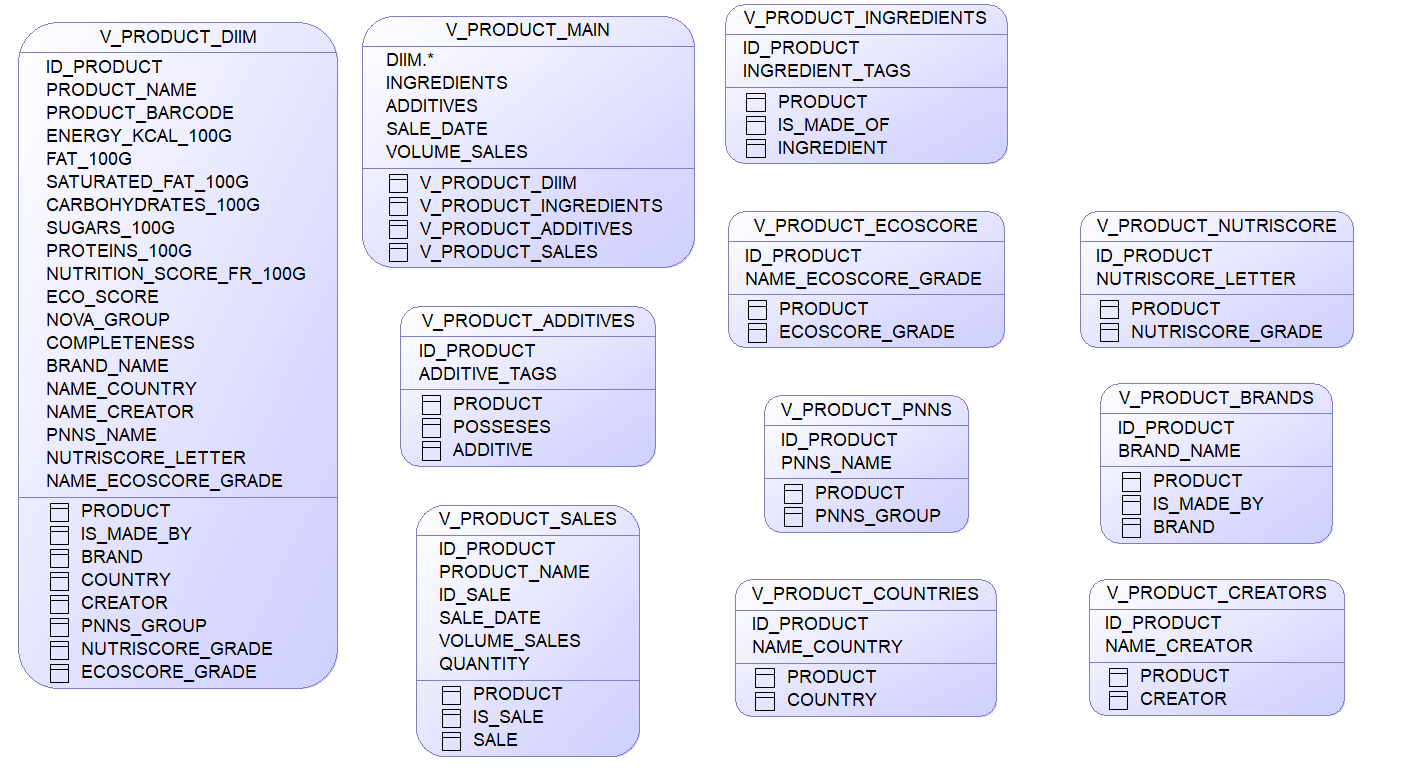


Figure 2 - Vues

## Optimisation OLTP

Tout d’abord, il est important de préciser que la base de données PostgreSQL est hébergée sur Microsoft Azure. La base est située en France, elle peut stocker jusqu’à 32 Gio de données et son niveau de performance a été amélioré pour atteindre 500 iops. À l’heure actuelle, la base de données OLTP coûte exactement $30.89 par mois. Plusieurs raisons se trouve derrière cette décision. Premièrement la disponibilité de la base, en externalisant le stockage de la base de données à un service tier, nous déléguons la maintenance à Azure et donc à Microsoft. Cela nous permet d’avoir un service efficace toujours en ligne sur le cloud prêt à se connecter à notre demande, mais il est nécessaire de configurer la base pour respecter le RGPD et surtout la sécurité de nos données (pare-feu, géo-redondance, rôle, etc…).

Dans le but d’optimiser la performance de notre base de données, nous avons décidé d’insérer des index sur toutes les clés étrangères où l’on effectue des recherches, ces index s’ajoutent à celle déjà présente sur les clés primaires générés automatiquement par PostgreSQL.

Pour la partie des traitements des données, nous avons remarqué dans le jeu de données fourni, un préfixe au niveau des ingrédients et des additives. À l’aide d’un script, nous avons séparé ces préfixes, et on l’insère dans une table LANGUAGE s’il n’existe pas, s’il existe nous l’associant avec une clé étrangère. En plus de rendre les données sur les ingrédients et les additives plus lisibles, nous optimisons le stockage alloué à ces tables Outre les tables « classiques », nous avons décidé que la table PRODUCT deviendrait le point central de la base, de ce fait, un produit sera inséré plus facilement à l’aide des clés étrangères. Les tables intermédiaires IS\_MADE\_OF, IS\_MADE\_BY, POSSESES et IS\_SALE permette d’insérer plusieurs clés à un produit.

Pour toutes les données qui peuvent être référées à une liste, nous avons décidé de créer une table pour chacun d’entre eux. Voici une liste de ces tables NUTRISCORE\_GRADE, COUNTRY, PNNS\_GROUP, ECOSCORE\_GRADE et CREATOR. Ainsi, pour l’insertion d’un produit, nous allons simplement remplacer la donnée par sa clé étrangère

Une dernière table qui mérite de l’attention est la table PNNS\_GROUP, on insère dans un premier temps les grandes catégories des PNNS puis les sous-catégories, on relie ensuite les petites catégories aux plus grandes à l’aide de la table IS\_IN. Cette manipulation nous prévient la création de deux tables pour les groupes de PNNS, une pour les grands groupes de PNNS, et l’autre pour les groupes mineurs du PNNS.

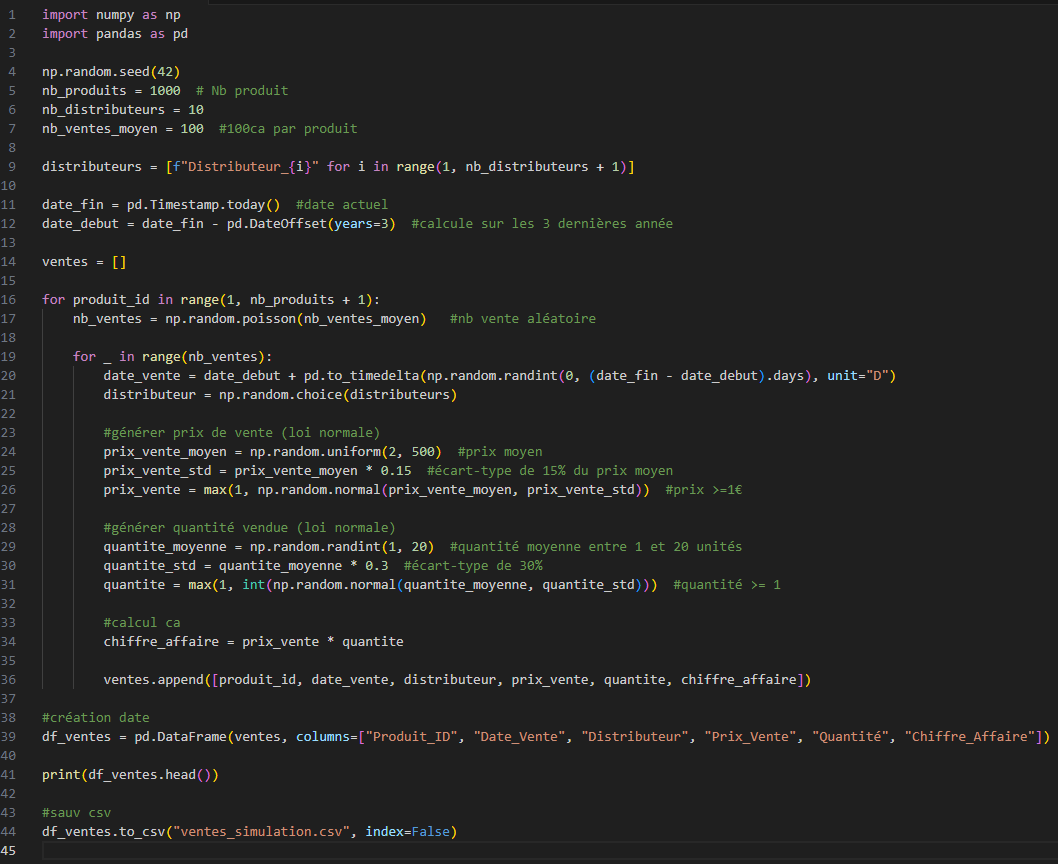
Concernant les attributs des tables, exceptés deux tables, toutes les clés primaires sont en SERIAL (auto-incrémentation) dans le but de simplifier les insertions de données. Les deux tables qui ne sont pas concernés par cette décision sont « PRODUCT ». En effet, la table de produit, cela s’explique par le fait que la sa clé primaire est la référence du produit (qui est unique), ainsi nous perdons par ces données mais en plus elle possède un rôle clé dans notre base de données.

Additionnement, les champs qui attendent comme valeur une monnaie sont déclarés en tant que MONEY. Bien que moins pratique pour les valeurs avec beaucoup de décimal (MONEY arrondit automatiquement les valeurs à 4 décimales), ce type est simple à utiliser pour des petites applis pour des affichages rapides.

Afin de construire une *data visualisation,* nous avons créé des vues pour garder une indépendance et plus de flexibilité vis-à-vis de la base de données. En possédant les mêmes clés, la construction des relations devient plus simple et permet ainsi de construire un modèle plus efficace, à l’instar d’importer les tables.

￼

## Insertion



Afin de simuler un historique réaliste de ventes pour les besoins du modèle OLAP, un script Python a été mis en place. Ce script génère des données de vente pour **1 000 produits** sur une période de **3 années**, en tenant compte d'une distribution aléatoire des dates, des prix et des quantités.

#### Méthodologie :

* **Période couverte** : 3 dernières années jusqu'à la date actuelle.
* **Nombre de ventes** : Généré pour chaque produit selon une **loi de Poisson** avec une moyenne de 100 ventes par produit.
* **Prix de vente** : Tiré d'une **distribution normale**, avec une moyenne aléatoire entre 2€ et 500€, et un écart-type de 15%.
* **Quantité vendue** : Également générée via une **distribution normale**, avec une moyenne aléatoire entre 1 et 20 unités, et un écart-type de 30%.
* **Chiffre d'affaires** : Calculé automatiquement comme le produit du prix de vente par la quantité.

#### Traitement :

Le fichier final contenait plus de **600 000 lignes**, ce qui le rendait trop volumineux pour un traitement direct dans Power BI. Afin d’optimiser la performance, les données ont été **divisées en 5 fichiers CSV plus légers**, puis regroupées dans Power BI après importation.

# Modèle OLAP

## Base

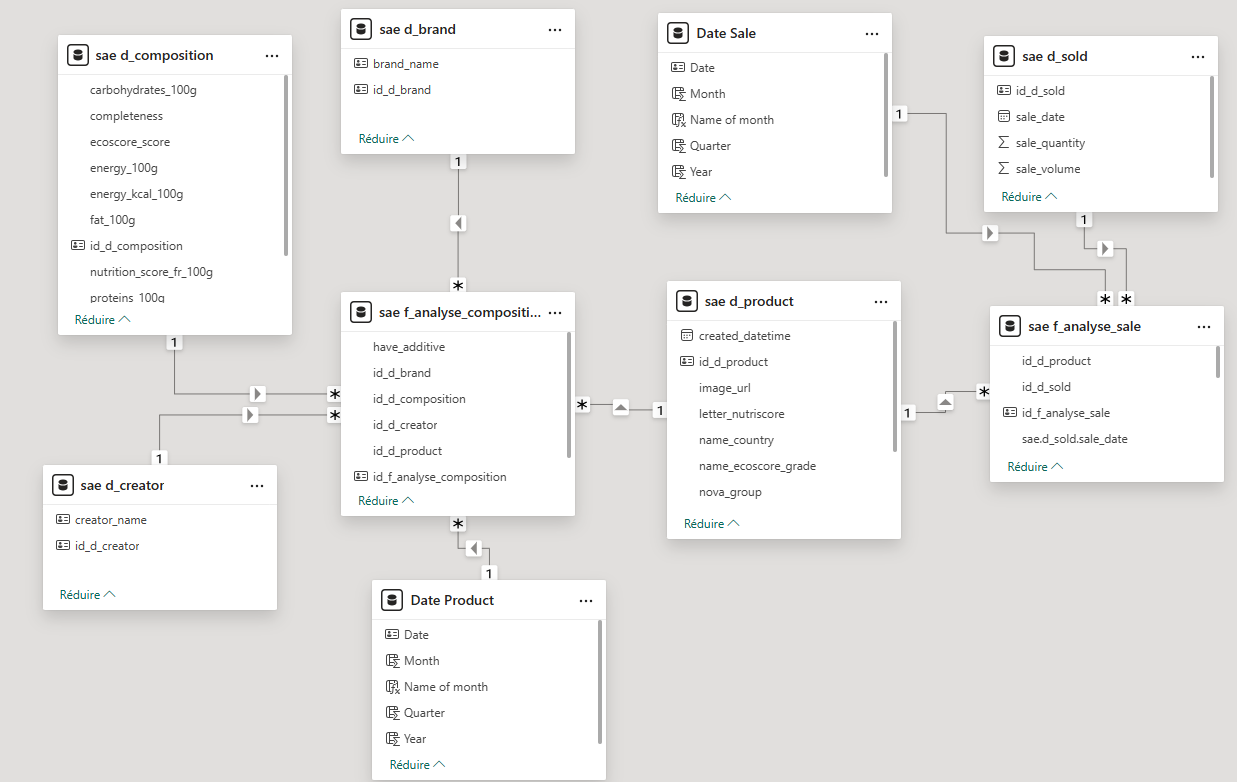


Figure 3 - Base OLAP

## Explication

Le modèle OLAP est une version améliorée du modèle OLTP, nous l’avons construit dans le but d'obtenir des rapports de synthèse tels que ceux utilisés dans la finance. En triant les données et en explicitant les objectifs, nous avons déduit deux tables de faits qui vont être analysées dans les *Dashboard*. Nous avons ajouté pour chacun de ces tables une table de temps, afin de permettre des visualisations pour une période voulues. Autour d’eux gravitent plusieurs tables de dimensions, telles que les produits, les marques ou encore les ventes, avec bien sûr des colonnes simplifiées.

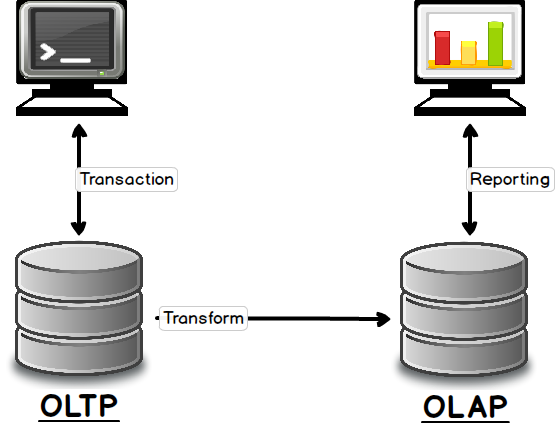


Figure 4 - Schéma explication

En effet, pour obtenir des calculs plus rapides nous avons diviser la table « PRODUCT » de la base OLTP, d’un côté nous avons la table « D\_COMPOSITION » qui contient toutes les valeurs nutritionnelles des produits insérés. Et de l’autre, nous avons « D\_PRODUCT » qui possède les données caractéristiques d’un produit, tels que son nom, son NOVA, sa date de création, etc…

Au vu du grand nombre de marque et de créateur insérés (plus de 1000 pour chacun d’entre eux), nous avons décidé de les séparer dans différentes dimensions pour respecter le modèle OLAP. Grâce à ces tables nous avons pu construire la table de fait « F\_ANALYSE\_COMPOSITION », qui s’occupe des calculs concernant la composition des produits. La deuxième table de fait « F\_ANALYSE\_SALE » combine les données des ventes et des produits, avec bien sûr une table de temps. Avec la présence de deux tables de faits, notre modèle OLAP représente un **modèle en constellation**.

## Insertion

L’insertion des données de ventes dans le modèle OLAP a été réalisée selon la **même logique de génération** que pour le modèle OLTP, assurant ainsi **cohérence et continuité** entre les deux modèles.

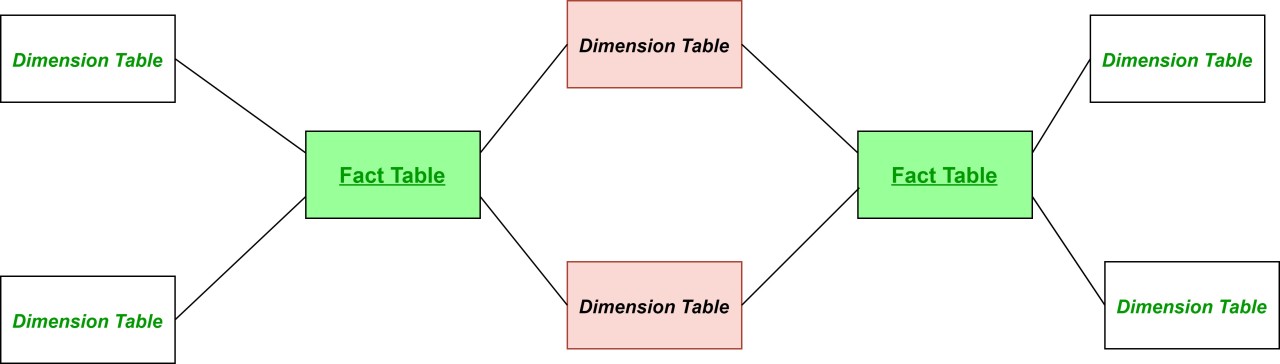


Figure 5 - Exemple de modèle constellation

# Modèle KNIME

## Modèle de visualisation + clustering

Une image contenant texte, diagramme, capture d’écran, ligne

Le contenu généré par l’IA peut être incorrect.

Figure 6 – Vue générale - Visualisation

À l’image des *data visualisation* réalisée sur Power BI, on a produit une représentation graphique des données sur le logiciel KNIME. Bien qu’il possède une interface moins intuitive, il reste néanmoins tout autant efficace dans le monde professionnel. Pour pouvoir observer les données avec des graphiques, vous allez devoir interagir avec le *workflow* dans KNIME, approcher votre souris d’un nœud avec une couleur de fond bleu foncé. Appuyer ensuite sur la loupe située en haut à gauche des nœuds, vous allez ainsi faire apparaître le graphique.

Dans la Figure 6, vous pouvez observer une vue d’ensemble de la visualisation avec l’aide de plusieurs nœuds, tels que :

Une image contenant texte, capture d’écran, affichage, logiciel

Le contenu généré par l’IA peut être incorrect.« Missing Value » pour retirer les lignes dans lesquelles les données manquent (cf. Figure 7).

Figure 7 - Lignes vides retirées

« Rule Engine » pour définir différentes conditions, fonctions ou bien simplement attribuer des lettres à des valeurs numériques (cf. Figure 8).

Une image contenant texte, capture d’écran, affichage, logiciel

Le contenu généré par l’IA peut être incorrect.

Figure 8 - Attribution lettres NutriScore

Une image contenant texte, capture d’écran, affichage, logiciel

Le contenu généré par l’IA peut être incorrect.« Numeric Outliers » pour supprimer les données dites aberrantes (cf. Figure 9).

Figure 9 - Outliers supprimés

« Linear Correlation » pour visualiser les attributs avec une forte corrélation avant de passer à la visualisation (cf. Figure 10).

Une image contenant texte, capture d’écran, nombre, Parallèle

Le contenu généré par l’IA peut être incorrect.

Figure 10 - Corrélation entre attributs (avant optimisation)

Maintenant, des exemples de visualisation des données réalisés à l’aide des nœuds « Scatter Plot » (cf. Figure 11), « Box Plot » (cf. Figure 12) et « Bar Chart » (cf. Figure 13).

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Figure 11 - Bar Chart

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, diagramme, Plan, capture d’écran

Le contenu généré par l’IA peut être incorrect.

Figure 12 - Scatter Plot

Figure 13 - Box Plot

Pour la partie Clustering, 2 méthodes de Clustering ont été retenues, la méthode k-Means, et la méthode DBSCAN.

Tout d’abord, voici une visualisation générale du modèle réalisé sur KNIME.

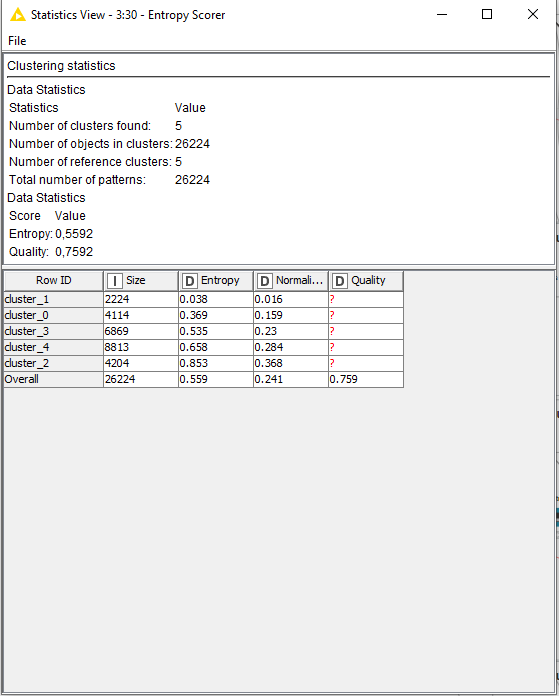
Une image contenant texte, capture d’écran, diagramme, ligne

Le contenu généré par l’IA peut être incorrect.

Figure 14 - Vue générale - Clustering

Comme on peut le constater ci-dessus (cf. Figure 14), le clustering a été appliqué à la fois sur la classe NutriScore et sur les différents attributs qui permettent de le calculer.

Malheureusement, aucune étude spécifique sur une métrique corrélée au NutriScore n’a été trouvé. Donc, malgré le manque d’attributs importants tels que le sodium et les fibres, nous avons effectué le clustering sur les attributs que l’on avait à disposition.

Ci-dessous, vous retrouvez les différentes qualités relevées sur chaque cluster.

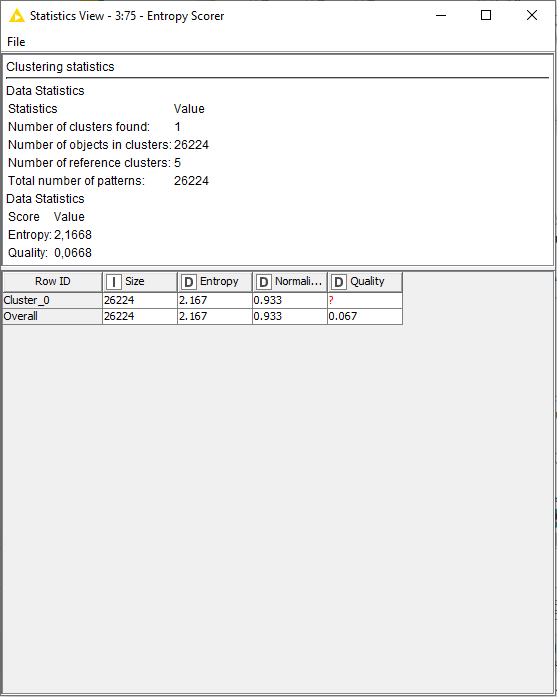


Figure 16 - Entropy Scorer - DBSCAN (NS score 100g)

Figure 15 - Entropy Scorer - k-Means (NS score 100g)

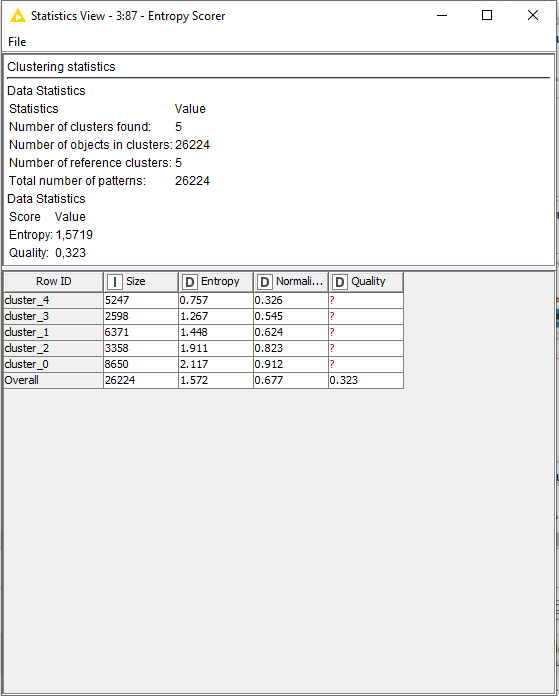
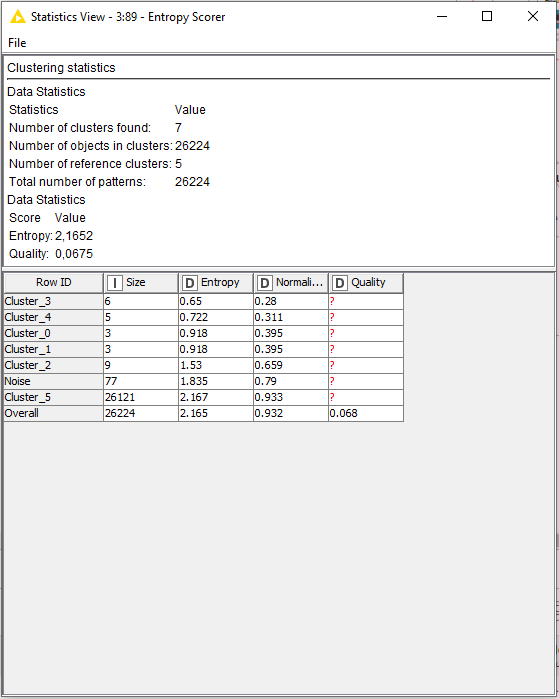


Figure 18 - Entropy Scorer - DBSCAN (attributs)

Figure 17 - Entropy Scorer - k-Means (attributs)

Comme on peut le voir, la méthode de clustering la plus optimisée est bien k-Means.

Cependant, malgré le fait que k-Means soit plus optimisé, on obtient seulement 75,9% de qualité en utilisant comme seul attribut « nutriscore\_fr\_100g ».

Avec les autres attributs que l’on a, c’est à dire « sugar\_100g », « fats\_100g », « saturated\_fat\_100g », « proteins\_100g », « carbohydrates\_100g » et « energy\_kcal\_100g » (« energy\_100g a été omis pour sa version en kcal), la qualité est encore plus basse, 32,9%.

On remarque clairement que pour un clustering optimisé, il faut au préalable tous les attributs nécessaires pour calculer la métrique (dans notre cas le NutriScore).

## Modèle IA de calcul de Nutri-Score

IMAGE

Dans le cas de notre jeu de données, l’utilisation de la formule de référence n’est pas possible, car les données sur les quantités de fibres et de sodium contenues dans les produits référencés ne sont pas souvent disponibles (données non fournies par le fabricant/vendeur ou non saisies par les volontaires). Pour information, la formule du Nutri-Score a été proposée par l'EREN, une équipe de recherche publique française sur la nutrition, dirigée par le professeur Serge Hercberg. Il est basé sur le score nutritionnel de la FSA créé par la Food Standards Agency du Royaume-Uni. Ce modèle KNIME va imiter le calcul du Nutri-Score pour produire un résultat similaire, et ensuite, il va insérer localement ce nouveau score pour les produits qui n’ont pas de Nutri-Score inséré.

Voici les différents modèles que nous avons construits pour déterminer le nutri-score des produits.

Tout d’abord, nous nous sommes basés sur les modèles d’arbre de décision vus en TD :Une image contenant texte, capture d’écran, diagramme, ligne

Le contenu généré par l’IA peut être incorrect.

Figure 19 – Test modèles Arbres de décision

Dans ces premiers modèles, nous avons effectué un pré-traitement des données pour que seulement les données significatives soient prises en compte. Nous avons aussi pris en compte les possibles erreurs d’insertion pour le nutri-score. Pour que le modèle puisse apprendre correctement, nous avons retiré les produits ayant des valeurs manquantes. Nous séparerons les modèles en 4 parties, la partie train, test, validation et test validation.

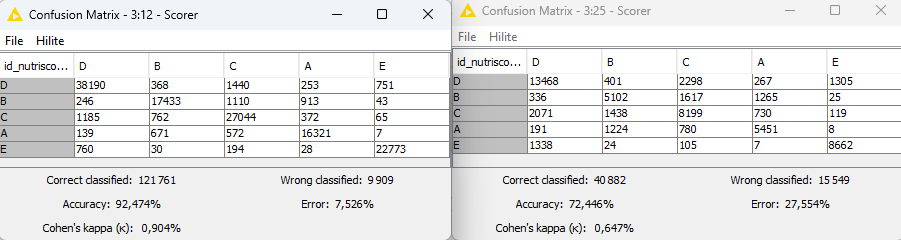


Figure 20 - Scores modèle NutriScore

Pour la partie train et test, nous pouvons observer que le modèle est très performant sur les données qu’il connaît.

On obtient une précision à hauteur de 92% alors que sur la partie test le modèle perd beaucoup en précision, on obtient 72% de précision. Cela démontre que lorsque le modèle ne connaît pas les données, il n’est pas très performant.

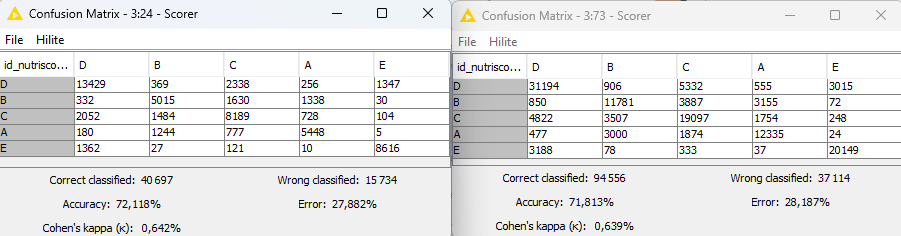


Figure 21 - Scores modèle NutriScore

Sur le modèle validation, on perd encore très légèrement en précision. Cela est dû à la validation croisée qui permet la création de sous-ensembles des données et de créer plusieurs modèles.

Lorsque nous avons testé ce modèle de validation croisée, nous avons encore perdu en précision.

Nous nous sommes par la suite de ces observations, penchés sur l’utilisation de modèles basés sur le Random Forest, voici notre premier modèle :

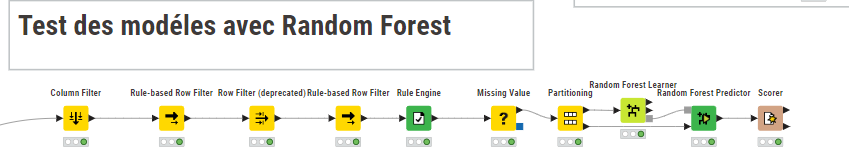


Figure 22 - Test modèles Random Forest

Ce modèle permet l’observation du comportement du modèle lorsqu’il est utilisé sur des données qu’il ne connaît pas, nous avons effectué le même tri de données que sur le modèle d’arbre de décisions.

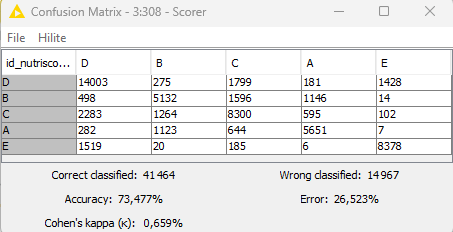


Figure 23 - Score modèle NutriScore

Les résultats de ce premier modèle dépassent ceux observés précédemment, nous avons donc amélioré ce dernier pour l’optimiser et augmenter la précision.

Voici le troisième modèle que nous avons créé :

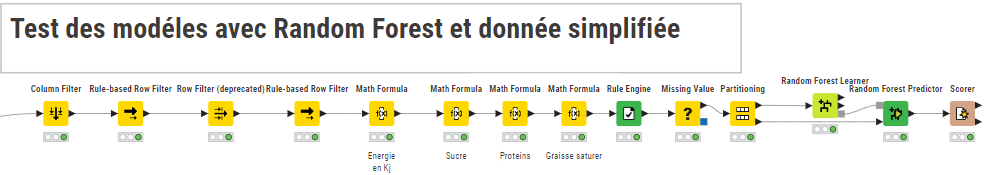


Figure 24 - Test modèles Random Forest et données simplifiées

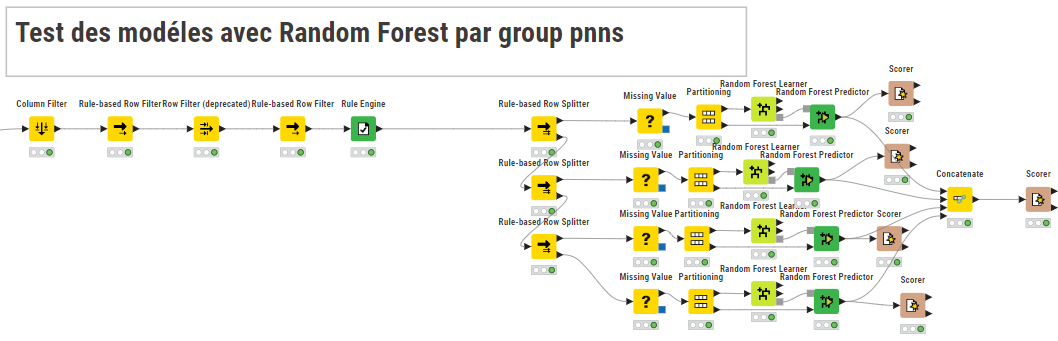
Dans ce modèle, nous avons simplifié les données, en créant des points allant de 1 à 10 pour les valeurs négatives du nutri-score (énergie en Kj, Sucre et graisse saturée) et de 1 à 5 pour les valeurs positives (protéines). Ces valeurs sont basées sur le calcul du nutri-score et permettent de catégoriser les plages de données.

Une image contenant texte, capture d’écran, nombre, Police

Le contenu généré par l’IA peut être incorrect.

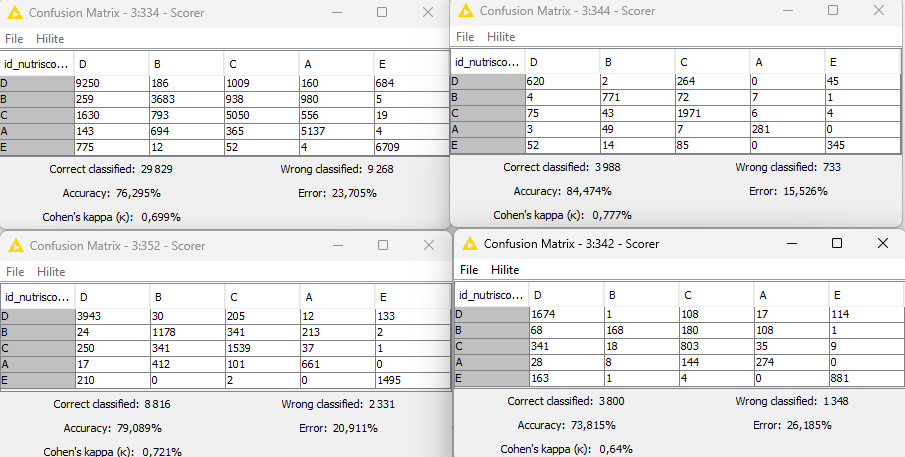
Suite à cette simplification, la précision du modèle s'est améliorée de 2%, nous avons donc essayé une autre méthode pour augmenter la précision

Voici notre quatrième modèle :

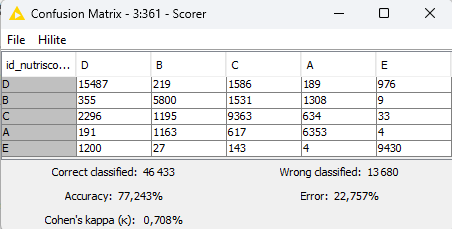


Pour ce modèle, nous sommes reparties sur le premier modèle de Random Forest. Nous avons ensuite établi que le Nutriscore a 4 manières d'être calculé, la première étant les matières grasses catégorisées dans le groupe\_PNNS numéro 8 dans notre base de données, le deuxième étant les fromages catégorisés dans le groupe\_pnns numéro 6 , le troisième étant les boissons catégorisées dans les groupes\_pnns numéro 7 et 11 et pour finir le cas général.

Ce modèle est donc composé de 4 modèles d’apprentissage basés sur les groupes PNNS.

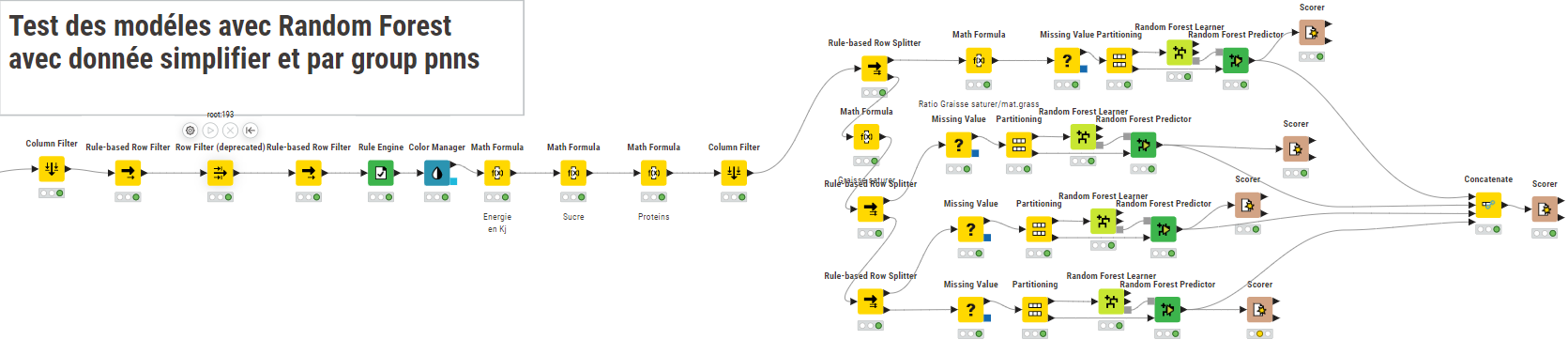


On observe une grande amélioration de la précision pour les modèles indépendants atteignant jusqu’à 84% de précision,

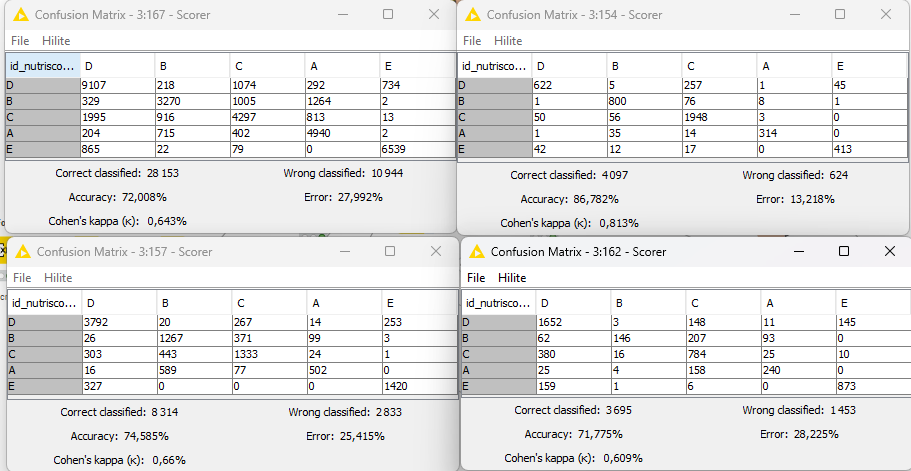


Nous avons ensuite concaténé ses résultats pour avoir une précision globale du modèle, le modèle est donc précis à 77%. C’est le meilleur résultat obtenu jusqu’à présent, nous avons donc décidé de créer un dernier modèle.

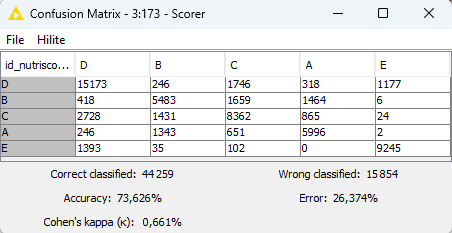
Voici notre cinquième modèle :



Ce modèle est une fusion des 2 derniers modèles, nous avons décidé de simplifier les données et de les séparer par groupes pnns.

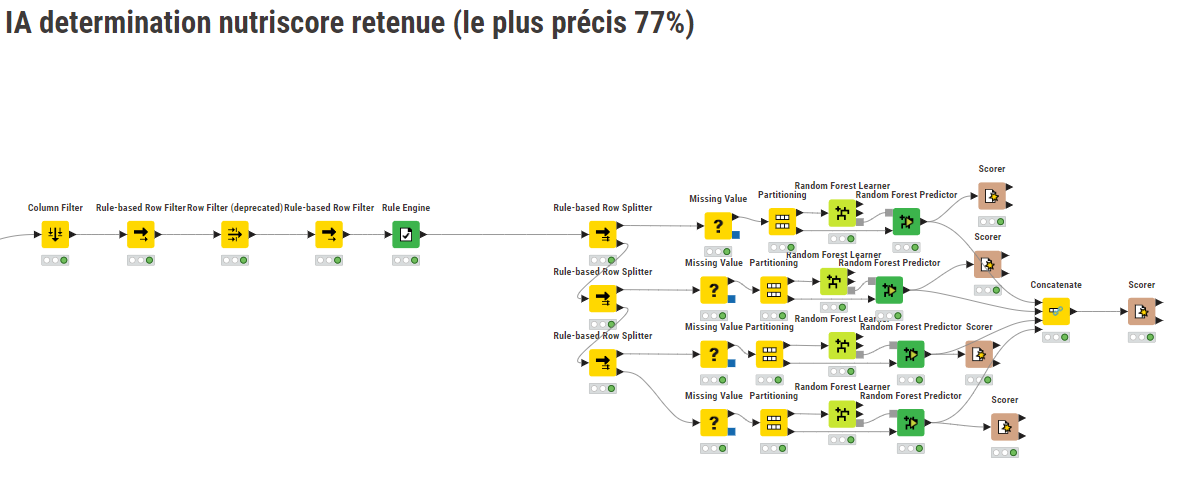


La simplification des données fait baisser la précision des modèles.



La précision globale a aussi été impactée.

Parmi ces modèles, un seul doit être retenu, le modèle de prédiction du nutri-score en fonction des groupesPNNS a été le plus performant en termes de précision, c’est pour cela que nous l’utiliserons par la suite.



## Modèle IA de l’Open Fact Food Score (OFFS)

IMAGE

Avec notre équipe, nous avons créé une nouvelle métrique pour venir compléter les autres scores (Nutri-Score, NOVA et Éco-score), baptisée « Open Fact Food Score » ou OFFS. Elle a été conçue sur KNIME à l’aide de procédé complexe mais simple d’exécution, vous avez la liberté d’examiner chacun des nœuds pour retrouver les formules ou encore les colonnes choisis pour les colonnes. Comme pour le Nutri-Score, l’OOFS va insérer localement ce nouveau score, mais cette fois-ci pour tous les produits, sous forme de grade dans cette ordre S, A, B, C et D, avec « S » la meilleure note et « D » la pire note possible.

Pour la création de score, nous nous sommes basés sur les données que nous possédons pour chaque produit, c’est-à-dire leurs nova-groupe, leurs éco-score et leur nutri-score.

Le nova- group se base sur la transformation d’un produit, plus il obtient un nova- group élevé, plus un produit est transformé.

L’éco-score représente l’impact environnemental d’un produit, plus il est élevé, moins un produit a un impact sur l’environnement.

Le Nutri-Score se base sur la qualité nutritionnelle du produit, plus il est élevé, plus le produit est de qualité.

L’Open Food Fact Score se basera donc sur l’addition de ces dernières, plus un produit aura un score élevé, plus la note sera basse.

Pour prédire ce score, nous allons avoir besoin de prédire la nova score, l’eco score et le nutri-score.

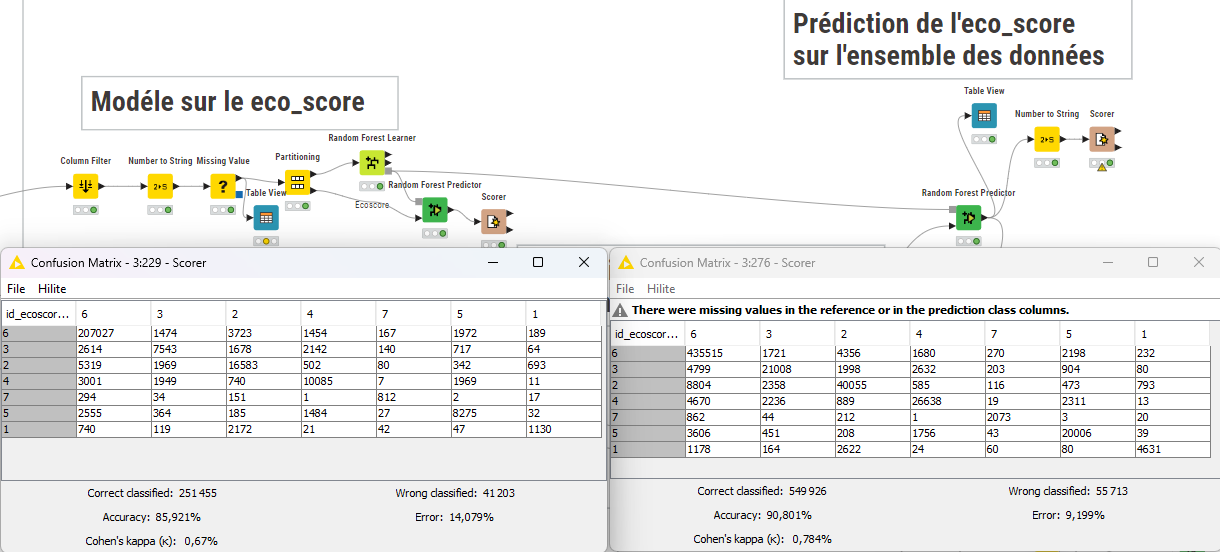
Voici le modèle complet que nous proposons :

Une image contenant texte, capture d’écran, diagramme, Police

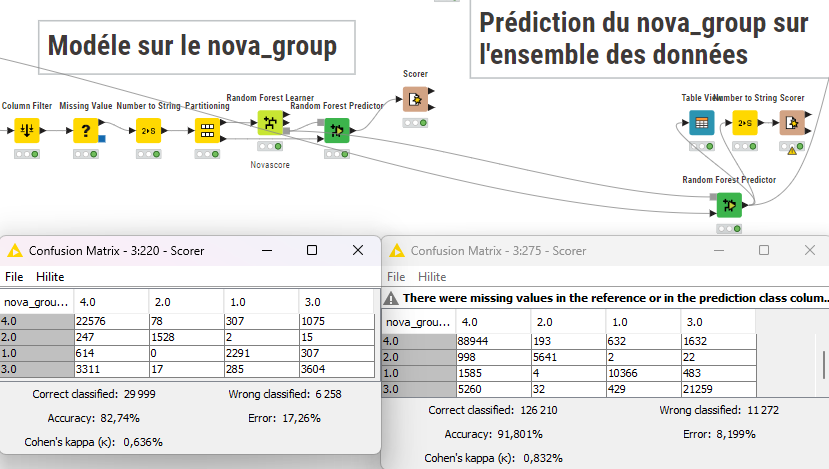
Le contenu généré par l’IA peut être incorrect.

Rentrons plus en détails dans ce modèle.

Voici notre modèle pour la prédiction de l’eco score :

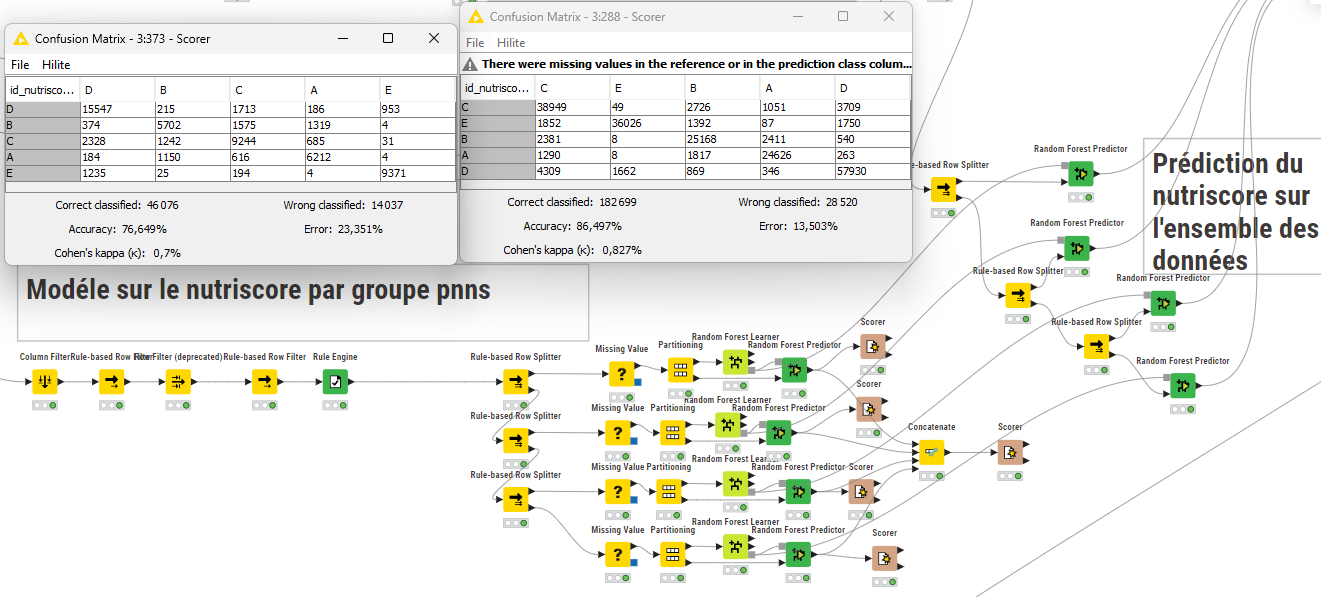


Ce modèle fonctionnant avec un Random Forest learner est très performant, on obtient 85% de précision sur les données qu’il ne connaît pas. Lorsque nous avons prédit cette valeur sur nos données, on obtient une précision de 90%.

Nous avons par la suite créé un modèle pour prédire la nova group d’un produit 

Ce modèle est aussi performant, 82% de précision sur les données qu’il ne connaît pas.

Avec une précision de 91% sur toutes nos données.Pour finir, nous avons repris le modèle de prédiction du Nutri-score :



Pour finir, nous avons une partie qui additionne toutes les prédictions afin de créer notre Open Food Fact Score :

Une image contenant texte, capture d’écran, diagramme, Police

Le contenu généré par l’IA peut être incorrect.

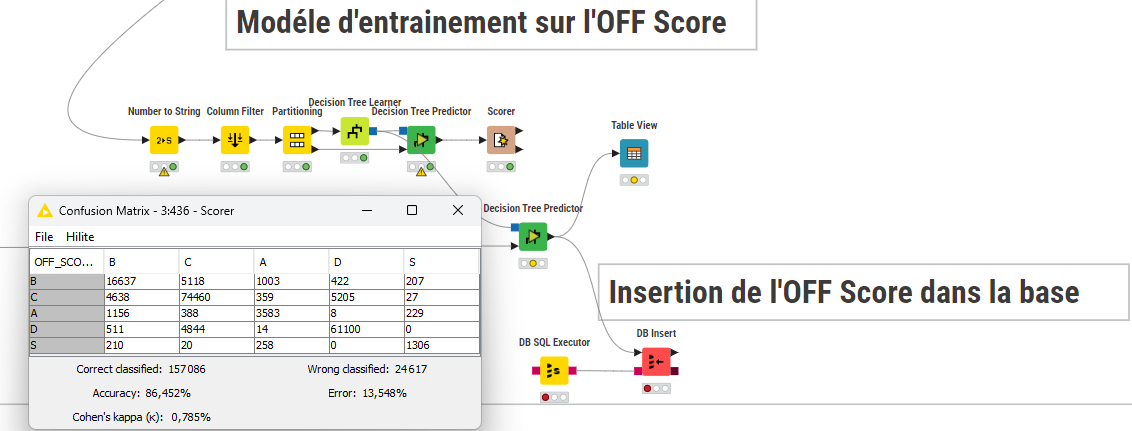
Ce score est compris en 3 et 16 qui sont ensuite répartis entre les différentes lettres S, A, B, C, D.

## Modèle IA d’attribution de l’OFFS

IMAGE

Ce dernier modèle KNIME va prédire le grade de l’OFFS pour n’importe quel produit nouvellement inséré dans la base de données Open Food Fact. En effet, l’OFFS était calculé en fonction des données disponible, qui étaient précédemment insérés. Ainsi, ce modèle va simplifier l’attribution de l’OFFS pour permettre une insertion plus rapide et optimisée.

Pour ce modèle, nous avons entraîné un arbre de décisions sur les données précédemment générées sur les OFFS,

Voici le modèle :

Le modèle possède une précision de 86% sur les données qu’il ne connaît pas.

Nous avons ensuite une liaison directe à la base de données qui crée la colonne OFF score et insère les valeurs prédites à l’intérieur de celle-ci.

# Power BI

Nous avons réalisé deux tableaux de bord Power BI, chacun basé sur un modèle de données différent : un modèle OLTP (relationnel) et un modèle OLAP (analytique). Chaque rapport comporte plusieurs pages organisées de manière à offrir une analyse claire, visuelle et interactive des données issues d’Open Food Facts.

## Préparation et importation des données

Avant la création du tableau de bord, nous avons effectué plusieurs étapes de préparation pour rendre les données exploitables dans Power BI.

Nous avons importé les fichiers CSV dans Power BI via **l'option "Obtenir des données" > "Fichier texte/CSV".**

Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.

Nous avons ensuite utilisé l’éditeur de requêtes Power Query pour :

* + Renommer les colonnes
  + Appliquer les bons types de données (texte, nombre, date…)
  + Ajouter des colonnes calculées (ex : identifier si un produit est sain ou non)
  + Filtrer et structurer les données pour les rendre compatibles avec nos modèles OLTP et OLAP

Une image contenant texte, diagramme, Police, capture d’écran

Le contenu généré par l’IA peut être incorrect.

## Modèle OLTP

Ce premier tableau de bord repose sur un modèle relationnel, structuré autour des produits. Il est composé de quatre pages :

**Product Statistics** : Dans cette première page, nous présentons une vue d’ensemble descriptive de notre base de données produits à travers plusieurs indicateurs clés.

Une image contenant texte, capture d’écran, nombre, Police

Le contenu généré par l’IA peut être incorrect.

Cette page sert à fournir une vue macroscopique et synthétique de notre dataset : nombre de produits, marque, nutrition, et chiffre d’affaires par typologie. Elle pose les bases de l’analyse qui sera approfondie dans les autres pages du tableau de bord.

**Product Analysis :** Visualisations permettant d’explorer plus en détail certaines dimensions.

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Le contenu généré par l’IA peut être incorrect.

Cette section propose des visualisations détaillées sur certaines dimensions spécifiques. Par exemple, nous avons analysé la répartition du nombre de produits par catégorie. Un graphique radar permet également de visualiser la proportion de produits selon les différents groupes Nutri-Score, offrant une lecture rapide de la qualité nutritionnelle globale. Une barre de recherche placée en haut de la page permet à l’utilisateur de filtrer les visualisations en recherchant un produit spécifique, rendant l’exploration plus intuitive et interactive.

**Product Search :** Une interface pour effectuer des recherches ciblées dans les données disponibles.

Une image contenant texte, capture d’écran, nombre, Police

Le contenu généré par l’IA peut être incorrect.

Cette section permet d’effectuer des recherches ciblées sur les produits grâce à une barre de recherche par nom. Des filtres sont disponibles pour affiner les résultats selon la catégorie, la sous-catégorie de produit, ainsi que le type d’Eco-Score, afin de repérer les produits les plus ou les moins vertueux écologiquement. Un tableau interactif affiche ensuite l’ensemble des informations nutritionnelles pour chaque produit, permettant d’accéder rapidement aux détails d’un produit sélectionné.

**Sales :** Une interface pour analyser des données de vente

Une image contenant texte, capture d’écran, Police, diagramme

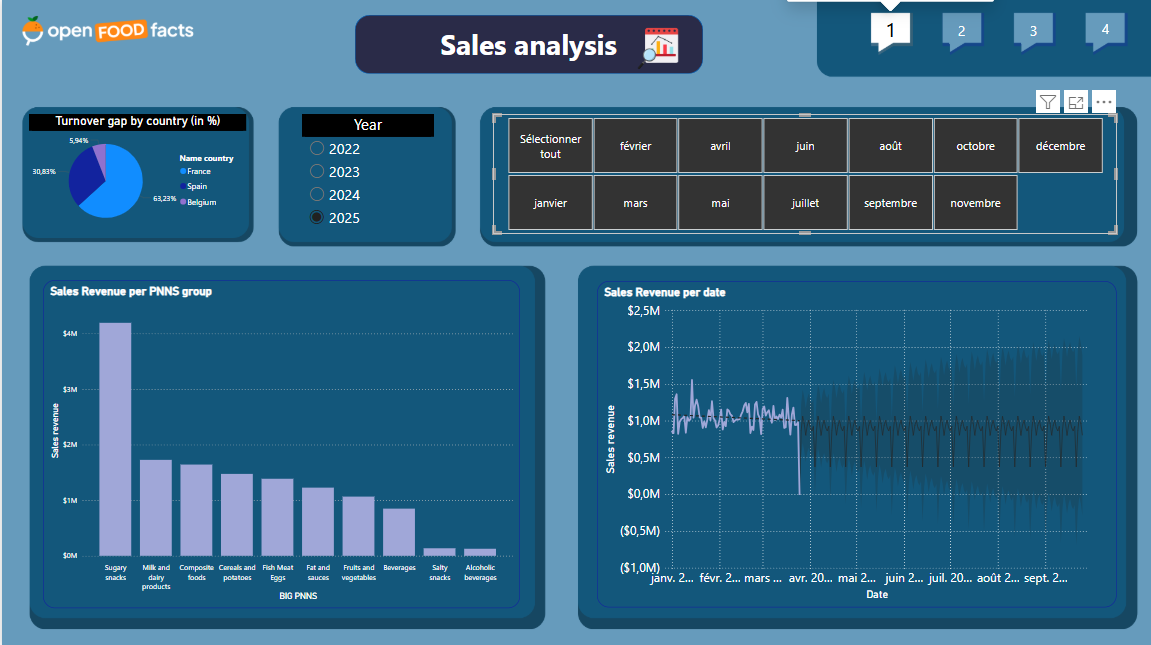
Le contenu généré par l’IA peut être incorrect.

Cette section permet d’analyser les ventes réalisées par pays, avec un filtre permettant de sélectionner l’un des trois pays présents dans la base de données. Les indicateurs affichés incluent le nombre total de produits vendus, la somme du chiffre d’affaires généré, ainsi que le nombre de transactions enregistrées. Un second filtre permet de cibler les sous-catégories de produits pour affiner l’analyse.

Un tableau présente le Top 20 des produits les plus rentables, c’est-à-dire ceux qui génèrent le plus de bénéfices, un bon indicateur des produits les plus vendus et les plus consommés. Enfin, un graphique linéaire illustre l’évolution du chiffre d’affaires au fil du temps, couvrant les années 2022, 2023 et 2024 jusqu’à aujourd’hui, avec une ventilation par catégorie de produit.

## Modèle OLAP

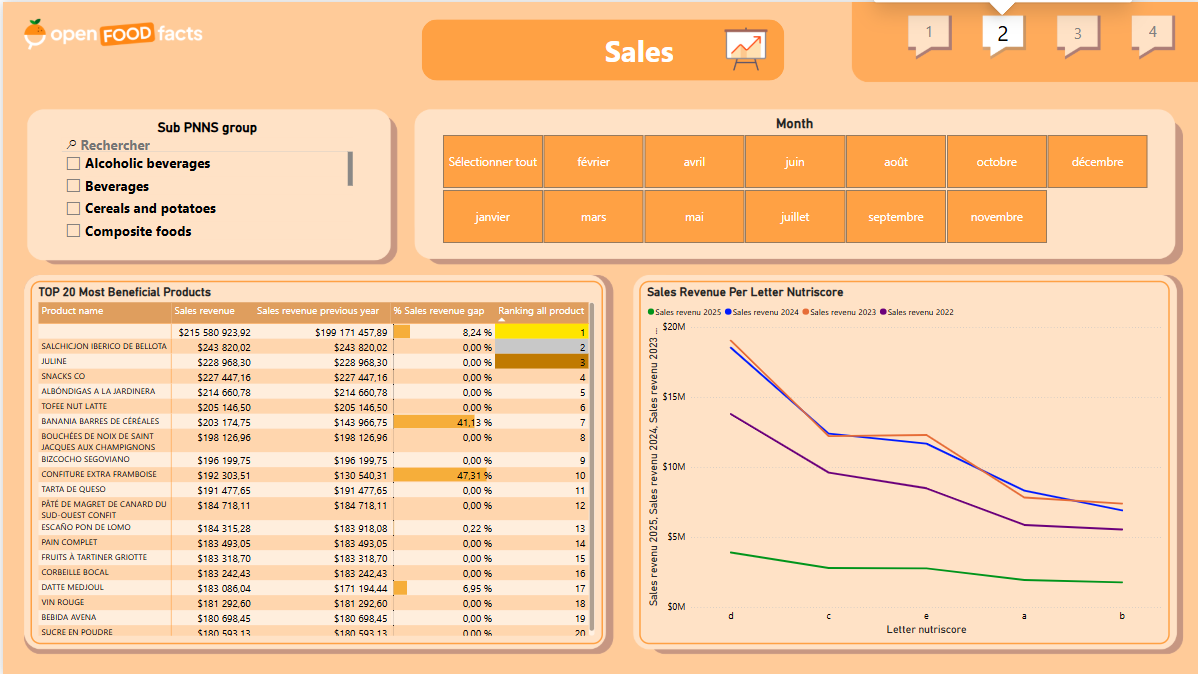
Le second rapport a été conçu sur la base d’un modèle en étoile, propre aux architectures OLAP, afin de permettre une analyse multidimensionnelle des données. Ce type de modélisation repose sur une table de faits centrale connectée à plusieurs tables de dimensions (temps, produits, géographie, etc.), ce qui facilite l’exploration des données selon différents axes d’analyse, ainsi que le calcul efficace de mesures agrégées.

**Sales analysis** : Une interface pour analyser les performances commerciales et les écarts de revenue 

Cette première page du rapport OLAP offre une vue d’ensemble des performances commerciales. Elle met en avant le **pourcentage d’écart de chiffre d’affaires (Sales Revenue Gap %) par pays**, permettant d’identifier rapidement les écarts de performance entre les marchés.  
Des **filtres interactifs** permettent de naviguer facilement dans les données, notamment un filtre par **année** ainsi qu’un filtre de type **vignette** pour sélectionner un ou plusieurs **mois**.

Un **graphique en barre** présente les résultats de vente regroupés par **groupe PNNS (catégorie nutritionnelle**

À droite, un **graphique linéaire** montre l’évolution du chiffre d’affaires au fil des mois et des années. Ce graphique intègre également une **courbe de prévision** projetée sur **une année**, permettant d’anticiper les tendances futures.

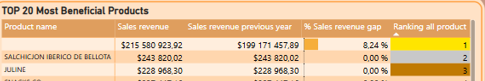
**Sales :** Une interface pour comparer les produits les plus performants sur plusieurs années

Cette seconde page dédiée à l’analyse des ventes offre une vision comparative détaillée des **meilleurs produits en termes de rentabilité**. Plusieurs **filtres** sont disponibles, notamment par **catégorie** de produit et par **mois**, pour affiner l’analyse.

Un **Top 20 des produits les plus bénéfiques** est présenté sous forme de tableau, affichant pour chaque produit :

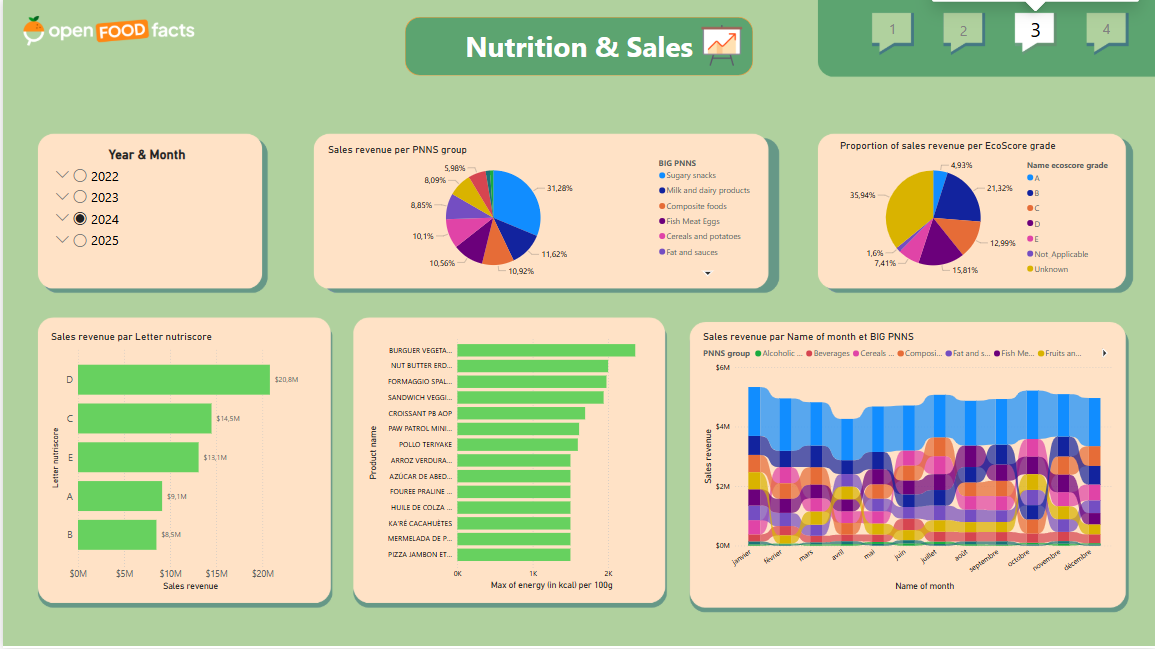
* Le **chiffre d’affaires actuel (Sales Revenue)**,
* Le **chiffre d’affaires de l’année précédente (Sales Revenue Previous Year)**,
* Le **pourcentage d’évolution (% Sales Revenue Gap)**,
* Ainsi que le **classement global** du produit (Ranking de 1 à 20).

Un **graphique linéaire** complète l’analyse en illustrant le chiffre d’affaires par **lettre de Nutri-Score**, avec une **légende temporelle** permettant de comparer les performances sur les 3 dernières années à partir d’aujourd’hui.



Ici on constate que les top 3 des produits le plus vendue sont **un produit ayant pas de nom, le salchicjon iberico de bellota et juline.**

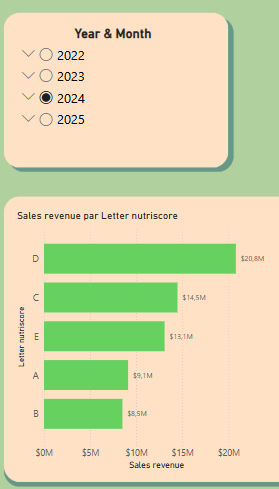
**Nutrition & Sales :** Une interface croisant données nutritionnelles et performances commerciales



Cette page combine l’analyse des **ventes** avec les **indicateurs nutritionnels et environnementaux**, afin de mieux comprendre l’impact de la qualité produit sur les performances commerciales.  
Des **filtres par année et par mois** permettent de cibler la période d’analyse.

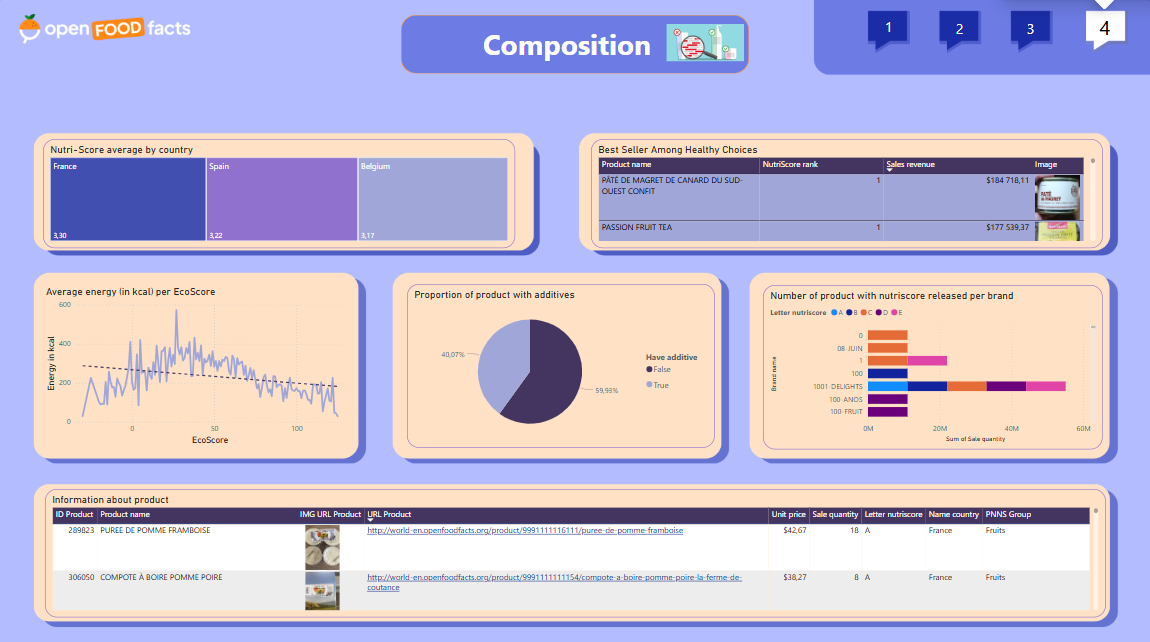
Elle se compose de plusieurs visualisations clés :

* Un **graphique en secteurs** affichant la répartition du **chiffre d’affaires par groupe de catégories**.
* Un second **graphique en secteurs** représentant les **ventes par score environnemental (Eco-Score)**.
* Un **graphique en barres** illustrant la relation entre les **revenus de vente et le Nutri-Score**, permettant d’identifier les groupes les plus consommés selon leur qualité nutritionnelle.
* Un **Top 10 des produits les plus caloriques**, pour mettre en évidence les produits les plus énergétiques présents dans les ventes.
* Un **graphique en ruban** montrant l’évolution du **chiffre d’affaires par mois**, ventilé par **groupe de catégories** (affichés en légende), permettant une analyse temporelle croisée.



On observe qu’en **2024**, les produits affichant un **Eco-Score D** ont généré le **chiffre d’affaires le plus élevé**, atteignant **20,6 millions de dollars**.

**Composition :** Une interface pour explorer les produits sains et leur performance commerciale

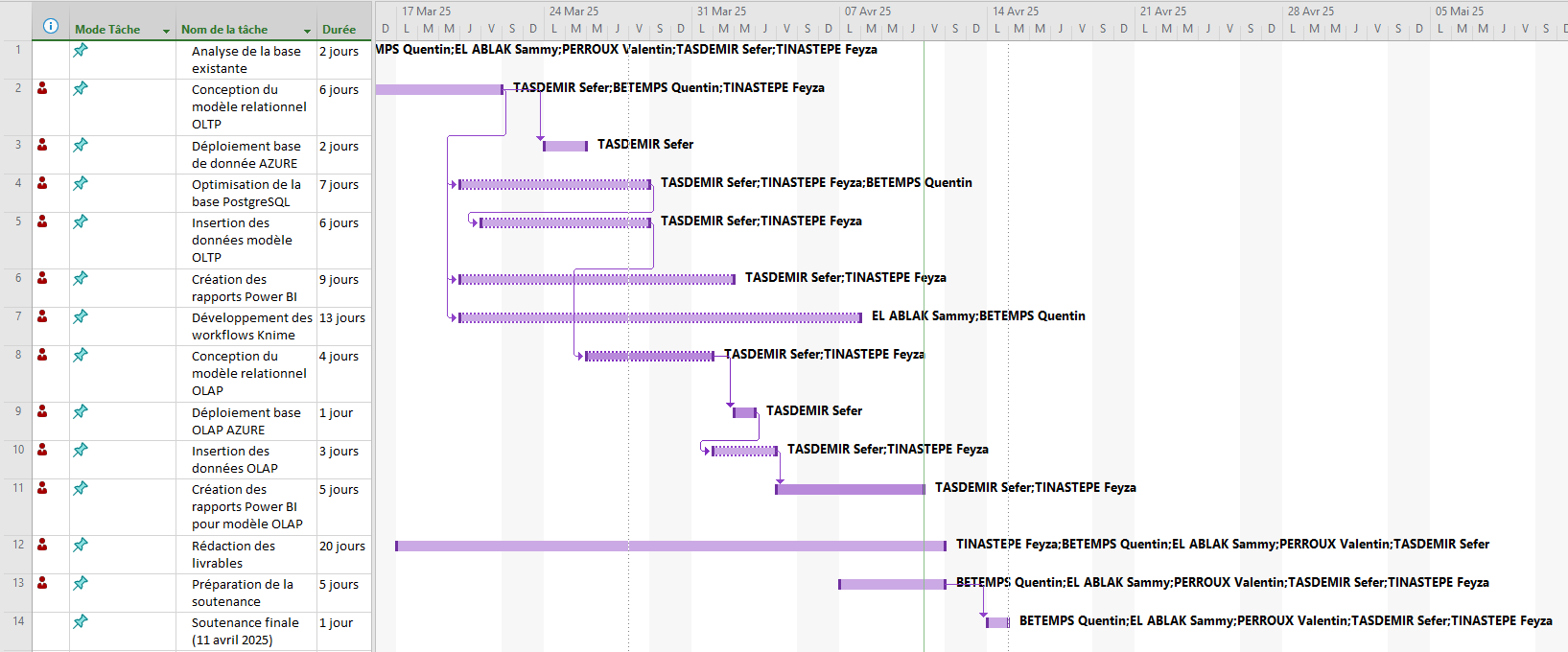


Cette page est centrée sur l’analyse des **produits considérés comme "healthy"**, en croisant des données nutritionnelles, commerciales et descriptives. Un **filtre par pays** (parmi les trois disponibles dans la base) permet d’adapter l’analyse selon le marché.

Elle inclut plusieurs éléments clés :

* Un **tableau présentant le Top des meilleures ventes parmi les produits sains**, avec le **nom du produit**, son **classement Nutri-Score**, le **chiffre d’affaires généré** et une **image du produit**.
* Un **graphique en secteurs** illustrant la proportion de produits **avec ou sans additifs**, pour évaluer la qualité des produits consommés.
* Un **graphique linéaire montrant** l’énergy moyenne en kcal par ecoscore**.**
* Un **graphique en barres** montrant la répartition des **Nutri-Scores par marque**, permettant de comparer la qualité nutritionnelle des différentes enseignes.
* Un **tableau détaillé** listant l’ensemble des produits avec les informations suivantes : **ID**, **nom**, **image**, **URL web**, **prix unitaire**, **quantité vendue**, **lettre Nutri-Score**, et **pays de vente**.

# GANTT



Nous avons réalisé un **diagramme de Gantt** afin de visualiser les tâches effectuées au fil du temps et d’identifier **quel membre du groupe** était responsable de chaque étape. Cela nous a permis d’évaluer la **répartition du travail** et le **taux d’implication de chacun** tout au long du projet.

# Répartition

|  |  |
| --- | --- |
| BETEMPS Quentin | 24 |
| EL ABLAK DIDOUCHE Sammy | 24 |
| PERROUX Valentin | 4 |
| TASDEMIR Sefer | 24 |
| TINASTEPE Feyza | 24 |

# Conclusion :

ce projet nous a permis le développer une solution complète pour optimiser et exploiter les données d’Open Food Facts, en passant d’un fichier plat volumineux à un modèle OLTP robuste et normalisé, puis à un modèle OLAP mieux adapté aux analyses décisionnelles.  
  
Nous avons soigneusement nettoyé et structuré la base, hébergée sur Azure pour des raisons de disponibilité et de maintenance, en ajoutant des index et en centralisant les informations principales sur la table PRODUCT.

Ensuite, nous avons construit un modèle OLAP en constellation, avec deux tables de faits (F\_ANALYSE\_COMPOSITION et F\_ANALYSE\_SALE) permettant d’établir des rapports multidimensionnels.

Nous avons aussi intégré des méthodes de Random Forest, d’arbres de décision et de validation croisée pour affiner la précision des modèles.

Enfin, Power BI a servi de support pour la création de tableaux de bord interactifs, chacun reposant respectivement sur le modèle OLTP et OLAP afin de couvrir des usages variés comme la recherche de produits, l’analyse des ventes, et la mesure d’indicateurs nutritionnels et environnementaux.

Ainsi cette démarche de la conception d’un modèle relationnel optimisé jusqu’à l’élaboration de modèles IA et de rapports visuels, démontre une approche cohérente et polyvalente pour la gestion et l’exploitation des données Open Food Facts, tout en proposant de nouvelles perspectives d’analyse et de prévision.