

URCA Reims

RAPPORT DE QUINZAINÉ

PRÉVISION SUR LE
NUTRI-SCORE



2023

Brunet Alexandre
Ertas Elif

Jupin Manon
Gabet Léo

Kpadondou Carlos
Otogondoua Ememag
Jordhy Jean Jaurès

SOMMAIRE

| | |
|---|----|
| Introduction | 3 |
| 1. Première quinzaine (du 06/10 au 20/10) | 5 |
| 1.1. Interface utilisateur | 5 |
| 1.2. Préparation des données | 5 |
| 1.3. Algorithme de prévision | 6 |
| 2. Seconde quinzaine (du 21/10 au 27/10) | 7 |
| 2.1. Interface utilisateur | 7 |
| 2.2. Data preparation | 7 |
| 2.3. Algorithme de prédiction | 8 |
| Conclusion..... | 9 |
| Annexes..... | 10 |
| Table des annexes | 11 |
| Annexe 1. Titre de l'annexe | 12 |
| Annexe 2. Titre de l'annexe | 13 |
| Annexe 3. Titre de l'annexe | 14 |
| table des figures..... | 15 |
| Table des tableaux..... | 16 |
| Table des matières..... | 17 |

INTRODUCTION

Depuis quelques années, l'émergence des enjeux environnementaux s'est accompagnée de la question de la santé et du bien-être physique par l'alimentation. Ces deux questions combinées, en plus de la volonté d'informer le consommateur, a poussé l'Etat à instaurer une norme d'étiquetage des produits transformés : le nutri-score.

Cette norme classe les produits selon leur impact positif ou négatif pour la santé des consommateurs. Le classement en question est illustré par des lettres allant de A (très bon pour la santé) à E (très mauvais pour la santé). La base de ce classement repose sur les composants nutritionnels des produits. Notre objectif est de déterminer quels sont les éléments nutritionnels qui influent sur le classement des produits et dans quelle mesure.

Pour répondre à cette question, nous nous sommes appuyés sur une base de données regroupant un peu plus de 52 260 produits ayant chacun leurs caractéristiques nutritionnelles (variables explicatives) et leur nutri-score variable cible (tableau 1).

| Variables | Description |
|---------------------------|--|
| <i>energy_100g</i> | Quantité de kilocalories pour 100g |
| <i>fat_100g</i> | Quantité de matières grasses pour 100g |
| <i>saturated-fat_100g</i> | Quantité de graisses saturées pour 100g |
| <i>trans-fat_100g</i> | Quantité de graisses transformées pour 100g |
| <i>cholesterol_100g</i> | Quantité de cholestérol pour 100g |
| <i>carbohydrates_100g</i> | Quantité de carbohydrates pour 100g |
| <i>sugars_100g</i> | Quantité de sucres pour 100g |
| <i>fiber_100g</i> | Quantité de fibres pour 100g |
| <i>proteins_100g</i> | Quantité de protéines pour 100g |
| <i>salt_100g</i> | Quantité de sel pour 100g |
| <i>sodium_100g</i> | Quantité de sodium pour 100g |
| <i>vitamin-a_100g</i> | Quantité de vitamine A pour 100g |
| <i>vitamin-c_100g</i> | Quantité de vitamine C pour 100g |
| <i>calcium_100g</i> | Quantité de calcium pour 100g |
| <i>iron_100g</i> | Quantité de fer pour 100g |
| <i>nutrition_grade_fr</i> | Notre variable cible qui est décrite par des lettres allant de A à E exprimant la qualité nutritionnelle du produit. |

Tableau 1 : Description des variables utilisées.

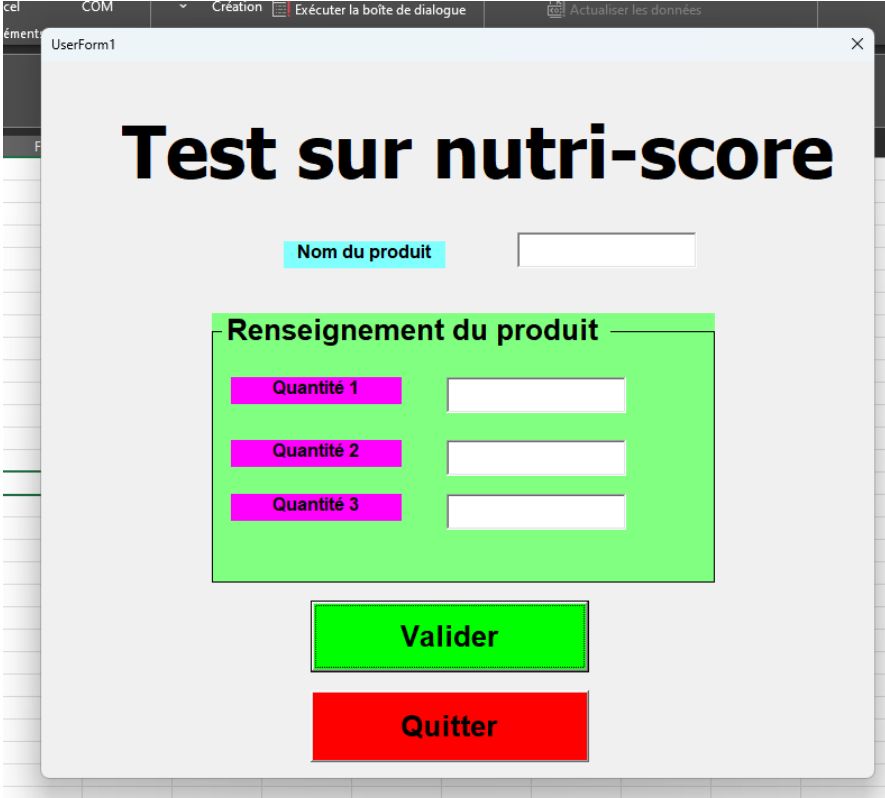
Cette base est présente sur le site de l'open-data du gouvernement¹. Grâce à cela, nous allons mobiliser des outils d'apprentissage statistique pour qu'un consommateur, en entrant les informations du produit, puisse évaluer de la pertinence du nutri-score d'un produit (ou alors connaître le son nutri-score dans le cas où il n'y a pas accès).

¹ <https://www.data.gouv.fr/fr/datasets/open-food-facts-produits-alimentaires-ingredients-nutrition-labels/>

1. Premier sprint (du 06/10 au 20/10)

1.1. Interface utilisateur

Pour notre première interface, nous avons opté pour la simplicité en ne mettant à disposition que 3 renseignements (figure 1). Le but de cette première interface est de préparer le terrain pour la suite. En effet, on ne sait pas combien de variables, nous allons décider de conserver.



The image shows a software interface titled "Test sur nutri-score". It is a dialog box with a title bar "UserForm1". The main content area has a light gray background. At the top, there is a text input field labeled "Nom du produit". Below this, there is a green-bordered box titled "Renseignement du produit". Inside this box, there are three rows, each with a pink label ("Quantité 1", "Quantité 2", "Quantité 3") and a white text input field. At the bottom of the dialog box, there are two buttons: a green "Valider" button and a red "Quitter" button.

Figure 1 : Première interface.

1.2. Préparation des données

Pour la préparation des données, en plus des variables que nous avons énumérées, il existe d'autres variables dans notre base de données. Cependant, celles-ci présentent soit des caractéristiques inutiles, soit une complexité qui les rend difficilement exploitables, ou encore une faible pertinence pour un utilisateur lambda. En conséquence, nous avons pris la décision de les exclure de notre analyse. De plus, notre choix de focaliser notre attention sur les valeurs nutritionnelles renforce notre justification pour leur suppression.

1.3. Algorithme de prévision

Pour le choix de l'algorithme de prévision, nous hésitons entre plusieurs modèles : la régression linéaire et la régression logistique multi-groupe ordinale. Pour le premier, nous le connaissons bien et savons comment l'utiliser. De plus, nous avons pensé qu'il nous suffirait de convertir le Nutri-Score en chiffres (au lieu de A, B, C, D et E, nous aurions 1, 2, 3, 4 et 5). Ainsi, lorsque l'utilisateur renseigne les informations, la réponse sera comprise entre 1 et 5, que nous convertirons en lettres. Par exemple, si la réponse est 3, nous renverrons à l'utilisateur que le Nutri-Score est C. Dans le cas où la réponse de l'algorithme renvoie un chiffre non entier, nous avons pensé à arrondir pour conclure.

D'un autre côté, la régression logistique multi-groupe ordinale nous semble la plus appropriée pour notre objectif. En effet, il s'agit de groupes distincts plutôt que de simples valeurs quantitatives, et chaque groupe transmet un message spécifique à l'utilisateur. De plus, ces différents groupes suivent une logique ordinale (A est meilleur que B, B est meilleur que C, etc.), ce qui nous pousse à privilégier ce modèle plutôt que la régression linéaire. Ainsi, la réponse de l'algorithme fournira à l'utilisateur une probabilité associée à chaque groupe que nous visualiserons pour obtenir une vue d'ensemble de chaque groupe.

2. Second sprint (du 21/10 au 27/10)

2.1. Data preparation

L'objectif de cette seconde quinzaine c'est de préciser notre algorithme de prédiction. Plus clairement, nous voulons à partir de toutes les variables explicatives que nous avons à disposition, en réduire leur nombre pour conserver celles qui sont vraiment utiles.

Pour atteindre cet objectif nous avons décidé d'utiliser les algorithmes AIC et BIC pour la sélection des données. Ces algorithmes nous ont donné chacun leur modèle optimal (celui qui réduit au maximum l'erreur de classification). Ensuite, nous avons comparé ainsi leur erreur de classification respectif pour ainsi choisir lequel entre le modèle optimal au sens de l'AIC et le modèle optimal au sens du BIC est le meilleur (celui qui minimise l'erreur de classification).

Il faut savoir que ces algorithmes ont certes un objectif commun (réduire au mieux l'erreur de classification), mais pas de la même façon. En effet l'AIC conserve au mieux l'information contenue dans les variables. Le BIC pénalise les modèles complexes en étant plus sévère dans sa sélection : il privilégie les modèles simples (avec le moins de variables possibles).

Au final, ces deux algorithmes ont sélectionné exactement les mêmes variables qui apparaissent dans le formulaire (figure 2).

2.2. Interface utilisateur

La sélection des variables est importante car elle permet de réduire le nombre d'informations que l'utilisateur entre dans le formulaire. En effet, moins il doit entrer d'informations, plus il aura tendance à l'utiliser fréquemment et de manière plus ludique. Ces variables sélectionnées seront justement les informations que l'utilisateur devra rentrer dans le formulaire (figure 2).

Test 1 formulaire nutri score

Calcul nutri-score

Nom du produit

Renseignement en quantité du produit

| | |
|------------------------|----------------------|
| Energie | <input type="text"/> |
| Matières grasses | <input type="text"/> |
| Matière grasses saturé | <input type="text"/> |
| Cholestérol | <input type="text"/> |
| Sucre | <input type="text"/> |
| Glucides | <input type="text"/> |
| Fibres | <input type="text"/> |
| Sel | <input type="text"/> |
| Vitamine C | <input type="text"/> |
| Fer | <input type="text"/> |

Calculer Nutri-Score **Quitter**

Figure 2 : Seconde interface avec les variables du premier modèle optimal.

2.3. Algorithme de prédiction

Nous allons utiliser ce que les algorithmes de sélection nous ont retourné. Autrement dit, nous allons effectuer une première régression logistique ordinale avec le meilleur modèle que nous avons trouvé avec les différents algorithmes de sélection de variables.

CONCLUSION

Tapez votre conclusion.

ANNEXES

TABLE DES ANNEXES

| | |
|----------------------------------|----|
| Annexe 1. Titre de l'annexe..... | 12 |
| Annexe 2. Titre de l'annexe..... | 13 |
| Annexe 3. Titre de l'annexe..... | 14 |

Annexe 1. Titre de l'annexe

Annexe 2. Titre de l'annexe

Annexe 3. Titre de l'annexe

TABLE DES FIGURES

| | |
|------------------------------------|---|
| Figure 1 : Première interface..... | 5 |
|------------------------------------|---|

TABLE DES TABLEAUX

| | |
|---|---|
| Tableau 1 : Description des variables utilisées. | 3 |
|---|---|

TABLE DES MATIERES

| | |
|---|----|
| Introduction | 3 |
| 1. Première quinzaine (du 06/10 au 20/10) | 5 |
| 1.1. Interface utilisateur | 5 |
| 1.2. Préparation des données | 5 |
| 1.3. Algorithme de prévision | 6 |
| 2. Seconde quinzaine | 7 |
| Conclusion..... | 9 |
| Annexes..... | 10 |
| Table des annexes | 11 |
| Annexe 1. Titre de l'annexe | 12 |
| Annexe 2. Titre de l'annexe | 13 |
| Annexe 3. Titre de l'annexe | 14 |
| table des figures..... | 15 |
| Table des tableaux..... | 16 |
| Table des matières..... | 17 |