

LAMSADE

UMR CNRS 7243

*laboratoire d'analyse et modélisation de systèmes pour l'aide à la décision*

Dauphine | PSL   
UNIVERSITÉ PARIS

# Décryptage algorithme Yuka

GHARSALLAOUI Dhia Eddine

Université Paris Dauphine, France

Dhia-eddine.Gharsallaoui@dauphine.eu



Yuka

## Table de matières

<b>Liste des figures.....</b>	<b>3</b>
<b>Chapitre 1 : Introduction .....</b>	<b>4</b>
1. L'application Yuka :.....	4
2. Fonctionnement : .....	4
3. Le score Yuka : .....	4
3.1. Nutri-Score : .....	4
3.2. Additifs : .....	5
3.3. BIO .....	6
<b>Chapitre 2 : Décryptage de l'algorithme de Yuka .....</b>	<b>7</b>
1. Construction de la base : .....	7
2. Analyse corrélation :.....	8
3. Fonction utilité Nutri-score :.....	9
4. Application Machine Learning : .....	12
4.1. Pré-processing : .....	12
4.2. Les Modèles :.....	12
4.3. Résultats avec des additifs :.....	12
4.4. Résultats sans des additifs.....	13
5. Conclusion : .....	14

## Liste des figures

Figure 1: Nutri Score .....	5
Figure 2: Exemple Liste additifs et description .....	5
Figure 3 Euro-feuille Label.....	6
Figure 4: Exemple Bio.....	6
Figure 5: Composition de la base .....	7
Figure 6 : Capture Base .....	7
Figure 7 Heatmap des corrélations .....	8
Figure 8: Exemple de calcul de $U_N$ .....	9
Figure 9: Exemple de $U_N$ des aliments Bio.....	9
Figure 10: Exemple de $U_N$ des aliments NON Bio .....	10
Figure 11: $U_N$ en fonction de Nutri-score pour les aliments BIO .....	10
Figure 12: $U_N$ en fonction de Nutri-score pour les aliments NON BIO.....	10
Figure 13: fonction Utilité $U_N$ en fonction de Nutri-score .....	10
Figure 14: Contre-exemple de la monotonie du Score Yuka .....	11
Figure 15: "One-Hot Encoding" appliqué sur additifs.....	12
Figure 16: Comparaison de l'erreur absolue moyenne pour chaque modèle .....	13
Figure 17: Comparaison de l'erreur absolue moyenne pour chaque modèle sans additif .....	13

*« Un bon croquis vaut mieux qu'un long discours. »*

*Napoléon Bonaparte*

# Chapitre 1 : Introduction

## 1. L'application Yuka :

Yuka est une application mobile pour iOS et Android, développée par Yuca SAS, qui scanne les aliments et les cosmétiques pour obtenir des informations détaillées sur les effets des produits sur la santé. Son objectif est d'aider les consommateurs à choisir des produits considérés comme bénéfiques pour leur santé et d'encourager les fabricants à améliorer les ingrédients de leurs produits.

## 2. Fonctionnement :

La lecture du code-barres d'un produit par téléphone permet à l'application d'accéder à des informations détaillées sur les ingrédients du produit et de renvoyer des commentaires dans des couleurs allant du vert au rouge et un score allant de 0 à 100. Lorsque son impact est jugé négatif, l'application peut vous recommander des produits similaires qui vous conviennent mieux.

## 3. Le score Yuka :

Ce que nous intéresse c'est la fonction d'attribution de ce score. On veut décrypter la fonction de calcul de ce score en commençant par les informations disponibles sur le site de l'application qui dit que ce score est sous la forme suivante.

$$\text{Score Yuka} = 0.6 * U_N (\text{Nutri score}) + 0.3 * U_A (\text{additives}) + 0.1 * U_B (\text{Bio}) \quad (1)$$

On peut voir que les facteurs qui déterminent la note d'un aliment allant de le plus impactant au moins sont le **Nutri Score**, les **additifs** et le critère **BIO**. On va commencer détailler les entrées de cette formule et essayer de les comprendre chacun seul sans tenir en compte leurs corrélation.

### 3.1. Nutri-Score :

Le Nutri-Score a été développé pour faciliter la compréhension des informations nutritionnelles par les consommateurs et ainsi de les aider à faire des choix éclairés d'où la lutte contre les maladies cardiovasculaires, l'obésité et le diabète.[1] Ce score se calcule à la base d'une évaluation de 100g ou 100mL d'un produit. Cette évaluation se divise sur deux comportements. Le premier c'est l'augmentation avec la quantité des nutriments favorable comme les fibres, les légumes, les fruits et les protéines. Le deuxième c'est la diminution avec les nutriments défavorables comme l'énergie, les acides gras saturés, les sucres et le sel.

Ce score est une valeur comprise entre -15 et +40. Mais il se manifeste sous la forme d'un logo apposé en face avant des emballages qui informe sur la qualité nutritionnelle des produits sous une forme simplifiée. La représentation simplifiée se décompose d'une lettre avec une couleur les lettres allant de A à E allant du meilleur au plus mauvais.



Figure 1: Nutri Score

- A correspondant à une valeur comprise entre -15 et -2
- B correspondant à une valeur comprise entre -1 à +3
- C correspondant à une valeur comprise entre +4 à +11
- D correspondant à une valeur comprise entre +12 à +16
- E correspondant à une valeur comprise entre +17 à +40

### 3.2. Additifs :

L'application extrait les additifs existants dans un aliment et puis il fait les évaluer en se basant sur une base qui est formé par des études indépendantes et des avis de l'EFSA, de l'ANSES et du CIRC. L'évaluation consiste à les classer sous 4 groupes :

- Sans risque (pastille verte)
- Risque limité (pastille jaune)
- Risque modéré (pastille orange)
- Risque élevé (pastille rouge)

Les informations sur les risques associés à chaque additif, ainsi que les sources scientifiques correspondantes sont affichés après la classification.

← List of additives

**E250**  
Sodium nitrite  
Preservative  
High risk

**E407a**  
Processed Eucheuma Seaweed  
Texturizing agent  
Moderate risk

**E262**  
Sodium acetates  
Preservative  
No risk

**E316**  
Sodium erythorbate  
Antioxidant  
No risk

**E326**  
Potassium lactate  
Antioxidant  
No risk

← Additive

**Sodium nitrite**  
High risk

**Preservative**  
Extends the product's shelf life

When combined in the stomach with certain foods, nitrites may contribute to the development of nitrosamines, compounds classified by IARC as "probably carcinogenic to humans."

Nitrites may also increase the risk of blood disorders, particularly in at-risk individuals.

[LEARN MORE](#)

[SCIENTIFIC SOURCES](#)

Figure 2: Exemple Liste additifs et description

Ce critère compte de 30% du score Yuka.

### 3.3.BIO

Il s'agit d'une bonification accordée aux produits considérés comme biologiques. Les produits considérés comme biologiques portent un label biologique européen (Euro-feuille).




	Sugar Low impact	17 g ● ▼
	Energy Low impact	342 kcal ● ▼
	Organic Natural product	✓

Figure 4: Exemple Bio



Figure 3 Euro-feuille Label

## Chapitre 2 : Décryptage de l'algorithme de Yuka

Dans ce chapitre on va parler des travaux réalisés pour comprendre mieux l'algorithme de Yuka et le décrypter en partant de l'hypothèse que c'est un modèle additif.

### 1. Construction de la base :

Afin d'atteindre notre objectif on a besoin de préparer une base de données qui contient à la fois tous les éléments utilisés pour calculer le score Yuka et le score Yuka même.

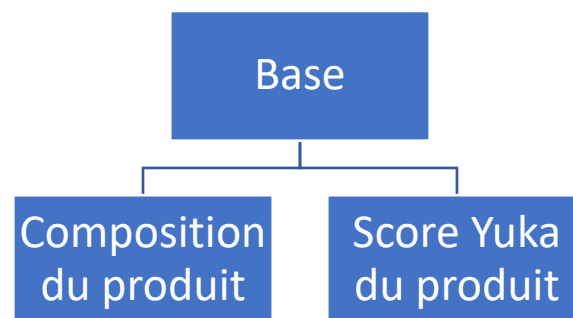


Figure 5: Composition de la base

- Composition du produit :

On a pris les ingrédients à partir de « OpenFoodFacts ». OpenFoodFacts est une base de données sur les produits alimentaires faite par tout le monde, pour tout le monde. Cette base contient presque la composition détaillée de chaque aliment.

- Score Yuka

Pour les scores de Yuka, on est sorti au supermarché. On a scanné plus que 500 produits. En faisant la combinaison de la base OpenFoodFacts et les scores Yuka des produits scannés, on a eu une base complète pour commencer les analyses.

product_name	additives_n	additives_tags	nutrition_grade_fr	nova_group	energy_100g	saturated-fat_100g	sugars_10g	fiber_100g	proteins	sodium_1g	nutrition-score_fr	nutrition-score_uk	Score Yuka
Couscous grain moyen	0		a		1506	0.3	2		12	0.004	-1	-1	90
Mix cereales	0		a	1	1586	1.3	3.7	9.8	24	0.016	-5	-5	90
Fideo Entrefino	0		a		1458	0.5	3.5	3	12	0.012	-4	-4	90
Tostada crujiente centeno & sÃ©sa	0		a	3	1631	1.4	3.1	16	11	0.18	-4	-4	90
Tiburones tricolor	0		a		1457	0.5	4	5	12	0.02	-6	-6	90
Macarrones integrales	0		a		1464	0.5	3	8	12	0.012	-6	-6	90
Arroz Basmati	0		a		1492	0.2	0.5	1.9	9.7	0	-2	-2	90
Stylelle Copos de arroz, trigo integr	0		b	3	1603	0.6	16	5.2	7.9	0.32	1	1	75
Stylelle Frutos Rojos	0		b	3	1573	0.6	18	5.5	7.9	0.288	1	1	75
Pan de centeno con semillas de sÃ©samo y amapola			c		1268	1.2	2.4		11	0.4	3	3	60
Pan de Molde Integral	6	en:e200,en:e270,en:e28	a	4	970	0.8	3	6	9.1	0.51	-2	-2	60
Copos de cereales chocolateados			d		1632	1.4	29		9.5	0.092	12	12	57
Corn Flakes	0		d	3	1584	0.7	1.3	3.5	8.3	1.08	11	11	54

Figure 6 : Capture Base

La base préparée est disponible sur ce lien : <https://www.kaggle.com/dhiagharsallaoui/yuka-base1>

## 2. Analyse corrélation :

Dès que la base est prête on a commencé l'exploitation. Tout d'abord on a essayé de voir la corrélation entre tous les variables et le score Yuka. Le Heatmap ci-dessous montre les degrés de corrélation.

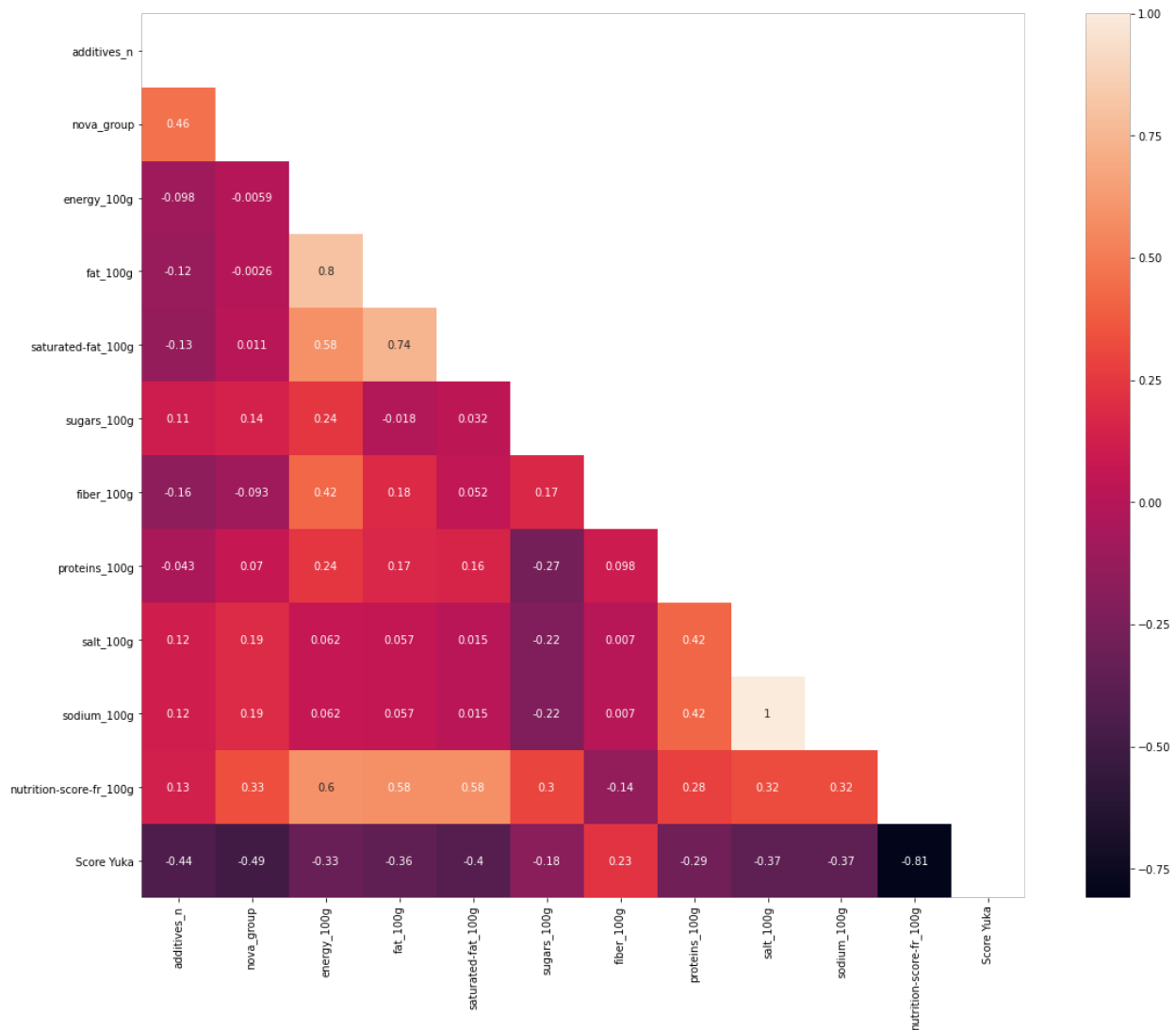


Figure 7 Heatmap des corrélations

Ce Heatmap confirme les informations données par Yuka. On peut voir que le **Nutri score** est le plus corrélé avec le score Yuka avec -0.81. Cette corrélation est négative ce qui est logique car le nutri score diminue avec l'augmentation de la qualité de nutriment. Et après on l'**additives\_n** qui représente le nombre d'additive existant dans l'aliment qui est corrélé négativement aussi et enfin le **nova\_group** qui communique une information sur les transformations utilisées pour avoir le produit final.

Vu que le Nutri score est le plus impactant dans le score Yuka on va concentrer à la suite de l'étudier en étudiant la fonction utilité  $U_N$  attribué à ce dernier dans la formule



$$\text{Score Yuka} = 0.6 * U_N (\text{Nutri score}) + 0.3 * U_A (\text{additives}) + 0.1 * U_B (\text{Bio}) \quad (1)$$

### 3. Fonction utilité Nutri-score :

Pour analyser mieux l'effet de Nutri score on a éliminé les effets des additifs. Par la préparation d'une sous base qui contient seulement les aliments qui sont sans additives. La contribution vient de ce critère est le maximum qui est égale à 30.

Et dans ce cas pour calculer la fonction Utilité de Nutri-score on a utilisé la formule suivante :

$$U_N (X) = \frac{\text{Score Yuka}(X) - (30 + 10 * IBIO)}{60} \quad (2)$$

Avec **IBIO** représente l'indice Bio qui est égale à 1 si l'aliment porte la label biologique et 0 sinon.

	nutrition-score-fr_100g	BIO	Score Yuka	utility
30	-10	0	90	1.000000
98	14	0	51	0.350000
117	12	0	38	0.133333
23	-4	0	90	1.000000
100	12	1	49	0.150000

Figure 8: Exemple de calcul de  $U_N$

Après ça on a divisé ce tableau en deux en séparant les aliments avec indice Bio égale à 1 et ceux avec indice Bio égale à 0. On a fait ça pour voir le comportement de chaque famille indépendamment d'où on limite les effets de critère Bio.

	nutrition-score-fr_100g	BIO	Score Yuka	utility
4	-4	1	100	1.000000
48	2	1	82	0.700000
2	-5	1	100	1.000000
105	13	1	47	0.116667

Figure 9: Exemple de  $U_N$  des aliments Bio

	nutrition-score-fr_100g	BIO	Score Yuka	utility
64	1	0	75	0.750000
15	-6	0	90	1.000000
21	-9	0	90	1.000000
34	-4	0	90	1.000000
121	13	0	37	0.116667

Figure 10: Exemple de  $U_N$  des aliments NON Bio

Après préparer les tableaux ce dessus on à essayer de tracer les fonctions utilité  $U_N$  en fonction des Nutri-score. Les graphes représentant ces fonctions sont ce dessous nous aide à interpréter plus d’informations sur leurs variations.

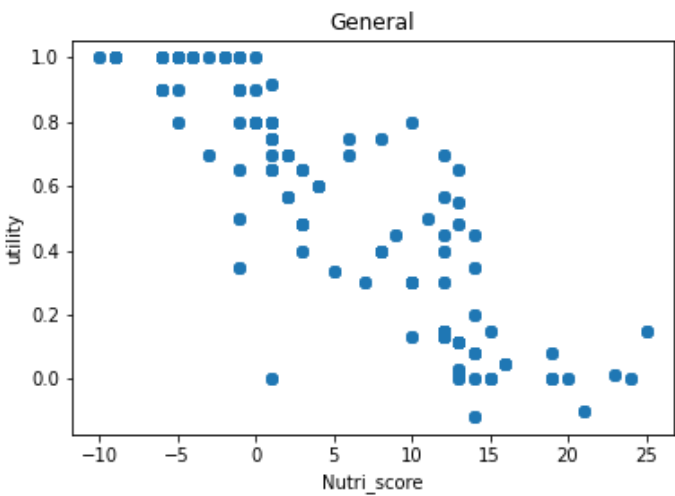


Figure 13: fonction Utilité  $U_N$  en fonction de Nutri-score

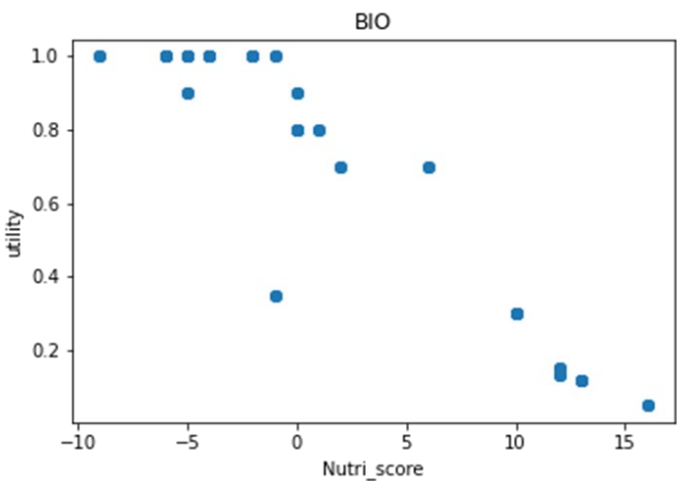


Figure 11:  $U_N$  en fonction de Nutri-score pour les aliments BIO

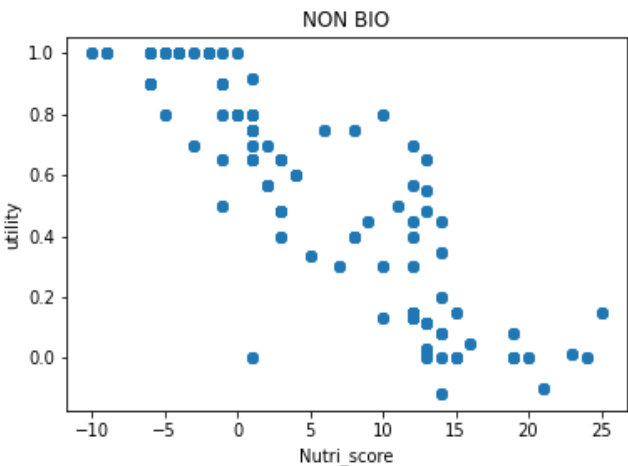


Figure 12:  $U_N$  en fonction de Nutri-score pour les aliments NON BIO

En regardant les graphes ci-dessus, on peut voir qu'il y a des utilités différentes pour la même valeur de Nutri-score. D'où la fonction utilité ne respecte pas la définition d'une fonction. Et par suite elle ne peut pas être une fonction.

Ces résultats nous donnent l'intuition d'investiguer si ce score Yuka est croissant avec le Nutri score comme il est déclaré ou non. Et pour faire ça, j'ai parcouru ma base pour trouver s'il y a un contre-exemple de monotonie. Et par le contre-exemple, je veux dire deux aliments qui ont le même critère BIO avec le même nombre d'additifs qui est zéro et ayant des Nutri score différents ne vérifiant pas la monotonie. C'est-à-dire l'aliment, avec le plus grand Nutri score, a le plus grand score Yuka.

			
<b>Produit</b>	Arachides Coques		Ananas en tranches
<b>BIO</b>	Non Bio		Non Bio
<b>Additives</b>	0		0
<b>Nutri Score</b>	10		-3
<b>Yuka score</b>	<b>78</b>		<b>72</b>

Figure 14: Contre-exemple de la monotonie du Score Yuka

Comme il montre le contre-exemple ci-dessus, le score Yuka n'est pas monotone. Aussi ce Score ne peut pas être un modèle additif d'où cette hypothèse est rejetée.

D'où j'ai décidé de faire un recours sur des modèles non additifs et je commencer à faire des modèles de Machine Learning.

## 4. Application Machine Learning :

Après l'élimination de l'hypothèse que score Yuka est sous cette forme additive :

$$\text{Score Yuka} = 0.6 * U_N (\text{Nutri score}) + 0.3 * U_A (\text{additives}) + 0.1 * U_B (\text{Bio}) \quad (1)$$

On a pensé à faire des essais avec les modèles de Machine Learning.

### 4.1. Pré-processing :

Afin d'avoir des bons résultats, on a commencé à préparer les entrées de modèles. Pour atteindre notre objective on a utilisé une technique qui s'appelle « One-Hot Encoding ». Cette technique nous permet de transformer une variable catégorique en plusieurs variables binaires. On l'a utilisé pour la variable

	additive :en:e100	additive :en:e120	additive :en:e131	additive :en:e1400	additive :en:e141	additive :en:e141i	additive :en:e141ii	additive :en:e1442	additive :en:e14xx	additive :en:e150a	additive :en:e150b	additive :en:e150c	additive :en:e150d	additive :en:e150e	additive :en:e150f	additive :en:e150g	additive :en:e150h	additive :en:e150i	additive :en:e150j	additive :en:e150k
0	0	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0

5 rows x 101 columns

Figure 15: "One-Hot Encoding" appliqué sur additifs

« **additifs** » en la transformant à 101 variables binaires. Le résultat se trouve dans la figure ci-dessous.

Maintenant les entrées sont prêts et on peut commencer l'apprentissages.

### 4.2. Les Modèles :

Pour l'apprentissages on a décidé d'utiliser des modèles de différents comportements et architecture. Comménçant par la régression linéaire pour confirmer le rejet de l'hypothèse qu'il est additif. Et après des modèles d'architecture Arbre et enfin un Modèle de Deep-learning qui est le réseau de neurones. En résumant les modèles utilisé sont comme suit : régression linéaire, arbre de décision, Random Forest et Réseau de neurones.

Afin d'avoir le plus d'information, on a décidé de faire deux modèles de chaque type l'un en utilisant les additifs et l'autre en éliminant ce critère par prendre que les aliments avec zéro additif.

### 4.3. Résultats avec des additifs :

En entrainant nos modèles sur la base complète on a eu des résultats convenants. Ces résultats sont présentés sur le graphe ci-dessous *figure 17* où on a le « Erreur absolue moyenne » pour chaque modèle utilisé.

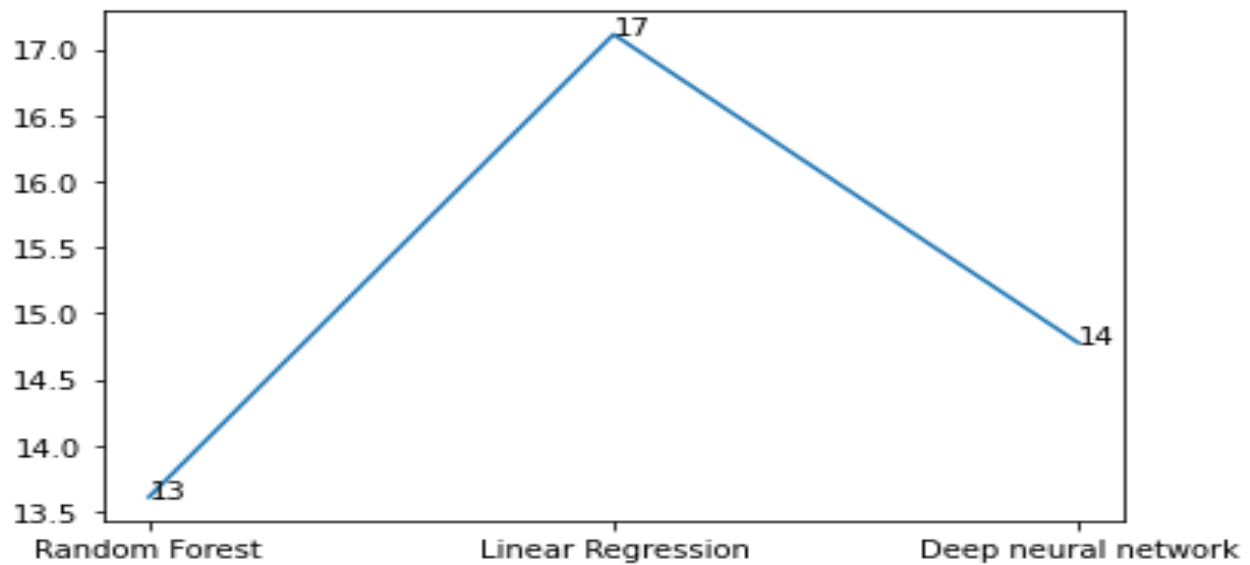


Figure 16: Comparaison de l'erreur absolue moyenne pour chaque modèle

Comme il montre la figure le pire c'est le modèle de régression linéaire ce qui est bien attendu après qu'on a montré que le modèle est non additif. En contrepartie le Random Forest qui est un modèle d'architecture arbre est le meilleur. Ces résultats sont bien attendus et se convient avec notre logique.

#### 4.4. Résultats sans des additifs

En investiguant plus, on a entraîné les modèles sur une base des aliments avec zéro additif. Le résultat de ces modèles est représenté dans la figure ci-dessous.

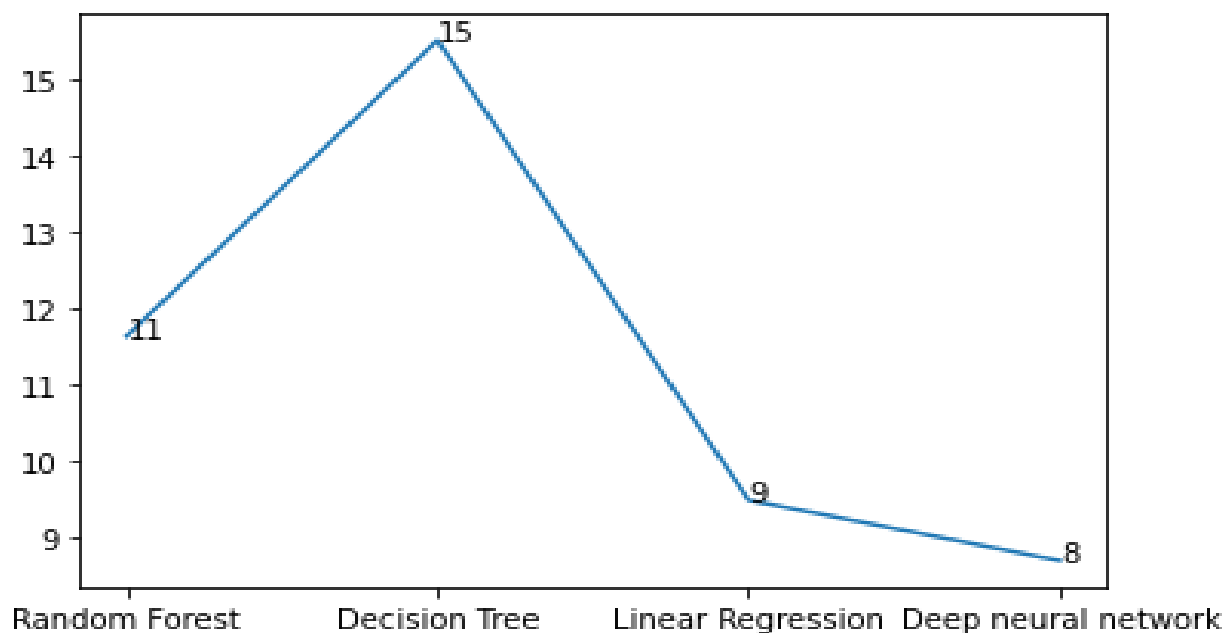


Figure 17: Comparaison de l'erreur absolue moyenne pour chaque modèle sans additif

En éliminant le critère des additifs la régression performe mieux mais n'a pas encore le meilleur. Dans tous les cas ces performances sont inutiles car on a éliminé une grande part des informations en supprimant les additifs.

## 5. Conclusion :

Le score de Yuka est un moyen développé pour aider les consommateurs à choisir les bons aliments pour leur santé. Mais le code source de ce score n'est pas partagé et l'attribution de chacun est communiquée jusqu'à maintenant. En faisant notre étude on a constaté que le score n'est pas additif. Ça ne pose pas un grand problème mais le fait que ce n'est pas monotone avec le Nutri-score pose plusieurs questions. Sachant que le Nutri-score est développé par des organismes publics en France et utilisé avec l'accord du gouvernement. D'où c'est un peu difficile d'accepter un tel score lorsqu'il ne convient pas avec le Nutri-score. Par suite il faut soit clarifier ce point par les responsables ou réviser l'algorithme de calcul et ajuster ce point.

Tout le travail réalisé est disponible sur ce lien : <https://www.kaggle.com/dhiagharsallaoui/final-yuka>