



CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE

DATA SCIENCE

MÉTHODOLOGIE

I. PRESENTATION DE L'IDEE D'APPLICATION

II.DESCRPTION DES DONNEES

III.VALIDATION DES DONNEES

IV.ANALYSE UNIVARIÉE DES VARIABLES

V.ANALYSE MULTIVARIÉE

VI. CONCLUSION

PRESENTATION DE L'IDEE D'APPLICATION

Nutriments Conseillés

- Fibres alimentaires
- Glucides
- Oméga 3

Objectifs

La valeur du nutrition-score nous permet de :

- Orienter les malades du **Diabète** dans leur alimentation
- Prévenir les personnes sur le risque du **Diabète** par rapport au choix de leur aliment

Nutriments Déconseillés

- Sucre
- Gras
- Acides gras saturés

DESCRIPTION DES DONNEES

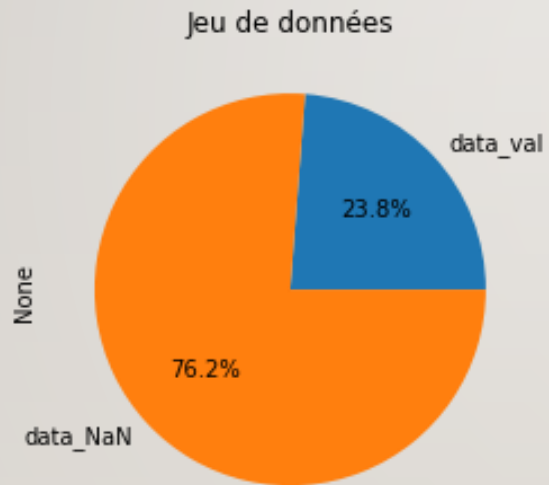
Le jeu de données est un fichier de :

- ✓ 320.772 produits (ligne)
- ✓ 162 champs ou variables (colonnes)

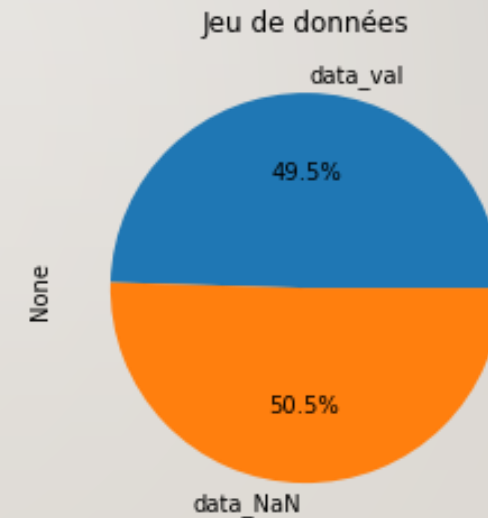
Les champs pertinents par rapport à notre projet sont vingt quatre (24)
principalement des variables de nutriment

DESCRIPTION DES DONNEES

Données Originales



Données avec les champs pertinents



VALIDATION DES DONNÉES

Identification des Erreurs

Les différents types d'erreurs sont:

- ❖ Les erreurs de format

Aucune erreur identifiée,

- ❖ Les valeurs doublons

36 lignes en doublons ont été identifiées

VALIDATION DES DONNÉES

Identification des Erreurs

❖ Les valeurs manquantes :

```
Out[14]: code                4
product_name              17762
origins                  298565
categories                236345
additives_n              71814
ingredients_from_palm_oil_n 71814
fat_100g                 76861
saturated-fat_100g       91195
alcohol_100g             316603
trans-fat_100g           177451
cholesterol_100g        176658
glucose_100g            320710
fructose_100g           320698
sugars_100g              75781
omega-3-fat_100g        319895
fiber_100g              119863
proteins_100g           60844
energy-from-fat_100g     319880
carbohydrates_100g      77164
vitamin-d_100g          313680
energy_100g              59639
nutrition-score-fr_100g  99540
nutrition_grade_fr       99540
dtype: int64
```

VALIDATION DES DONNÉES

Identification des Erreurs

❖ Les valeurs outliers :

Il y a en effet de grosses incohérences :

- ✓ des produits de 100 g qui ont une composition en des nutriments supérieure à 100g .
- ✓ certains produits ont une composition en nutriments négative

VALIDATION DES DONNÉES

Traitement des Erreurs

❖ Les Doublons :

Suppression des lignes en doublons

❖ Les valeurs manquantes :

➤ Suppression de ligne

➤ Remplacement de valeurs NaN de colonne

➤ Suppression de colonne

❖ Les Outliers :

➤ Remplacement des valeurs supérieures à 100 par la valeur moyenne

➤ Remplacement des valeurs négatives par 0

VALIDATION DES DONNEES

Traitement des Erreurs

Le jeu de données nettoyé est composé de :

- ✓ 320.732 produits (ligne)
- ✓ 14 champs ou variables (colonnes)

product_name	origins	categories	additives_n	ingredients_from_palm_oil_n	fat_100g	saturated-fat_100g	sugars_100g	omega-3-fat_100g	fiber_100g	carbohydrates_100g	nutri-sc fr_'
Farine de blé noir	Aucune	Aucune	0.0	0.0	12.264135	4.827212	16.003369	3.182103	2.862013	32.073621	
Banana Chips Sweetened (Whole)	Aucune	Aucune	0.0	0.0	28.570000	28.570000	14.290000	3.182103	3.600000	64.290000	
Peanuts	Aucune	Aucune	0.0	0.0	17.860000	0.000000	17.860000	3.182103	7.100000	60.710000	
Organic Salted Nut Mix	Aucune	Aucune	0.0	0.0	57.140000	5.360000	3.570000	3.182103	7.100000	17.860000	
Organic Polenta	Aucune	Aucune	0.0	0.0	1.430000	1.430000	16.003369	3.182103	5.700000	77.140000	

ANALYSE UNIVARIEES

❑ nutrition_grade

❑ nutrition_score

Les lettres du nutrition_grade sont fonction de la la valeur du nutrition_score du produit :

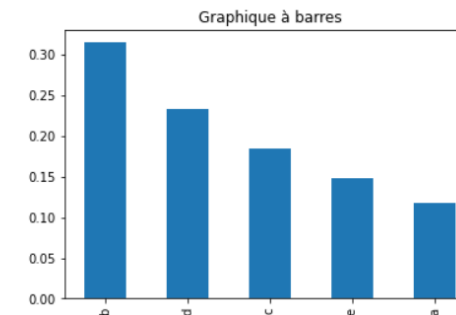
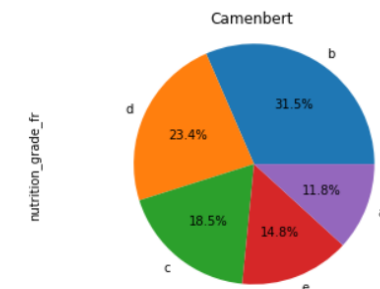
nutrition_grade	nutrition_score
a	$[-15 ; 0[$
b	$[0 ; 3[$
c	$[3 ; 11[$
d	$[11 ; 19[$
e	$[19 ; 40]$

ANALYSE UNIVARIEES

La distribution empirique de variable

❑ Variable qualitative : nutrition_grade_fr

	nutrition_grade_fr	n	f
0	b	101065	0.315107
1	d	74983	0.233787
2	c	59460	0.185388
3	e	47409	0.147815
4	a	37815	0.117902

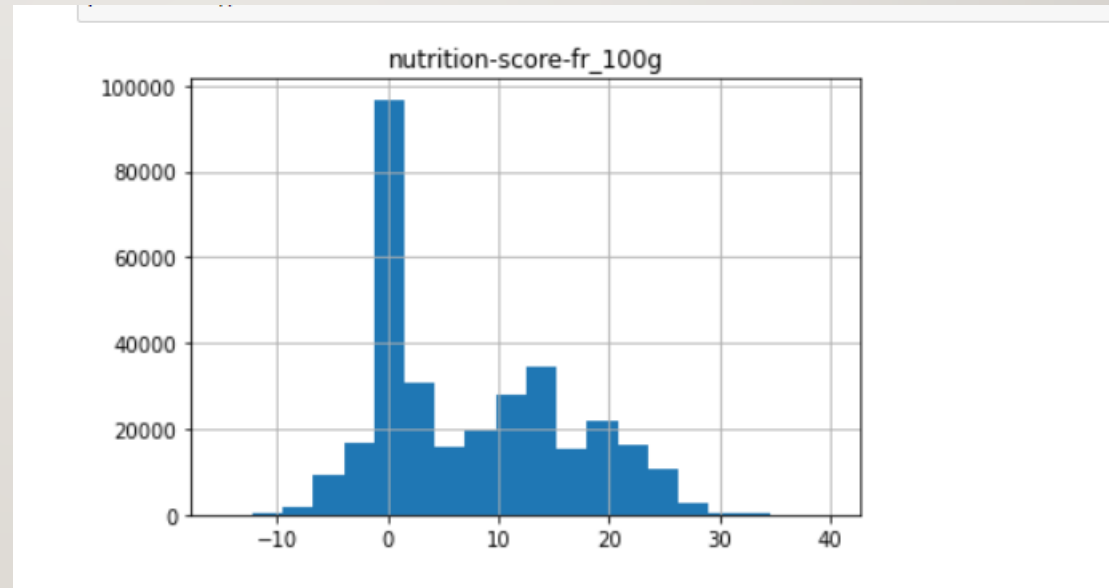


ANALYSE UNIVARIEES

Les mesures de tendance centrale

❑ Variable quantitative : nutrition-score-fr_100g

- Le mode : 0
- La médiane : 7,45
- La moyenne : 5



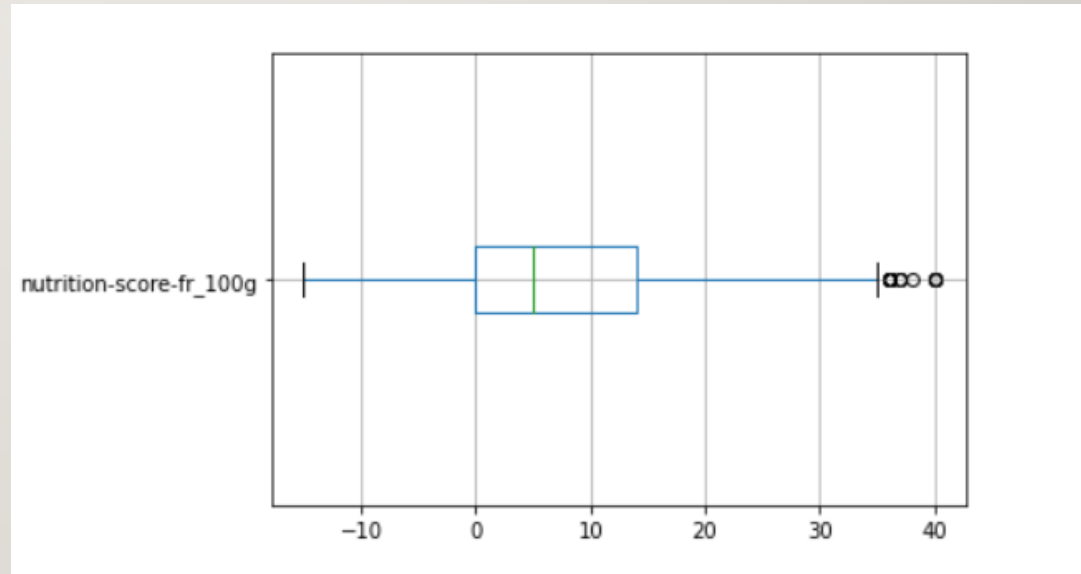
ANALYSE UNIVARIEES

Les mesures de dispersion

❑ **Variable quantitative** : nutrition-score-fr_100g

➤ Le Variance : 75,44

➤ L'écart-type : 8,68



ANALYSE UNIVARIEES

Les mesures de dispersion

❑ **Variable quantitative** : nutrition-score-fr_100g

- Le Skewness : 0,48
- Le Kurtosis : 8,68

les valeurs du Skewness empirique et Le Kurtosis empirique sont conforme à ce que nous avons observé sur l'histogramme :

- ✓ la distribution est étalée à droite
- ✓ la distribution est plus aplatie.

ANALYSE MULTIVARIÉE

Analyse Bivariée

Le Coefficient de corrélation

Après le calcul des coefficients de corrélation entre la variable "nutrition-score-fr_100g" et les variables de nutriments :

- aucune corrélation.
- Valeur de corrélation un peu élevée pour les variables :
 - ✓ saturated-fat_100g
 - ✓ fat_100g
 - ✓ sugars_100g.

ANALYSE MULTIVARIÉE

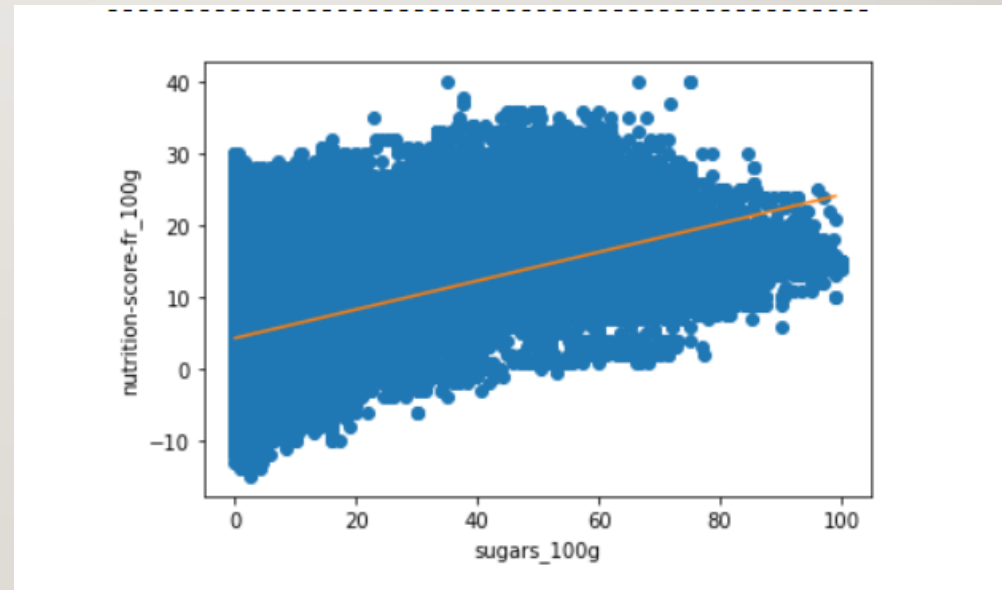
Analyse Bivariée

Analyse par régression linéaire

□ nutrition-score-fr_100g et sugars_100g

La droite de régression :

$$y = 4.3 + 0.2x$$

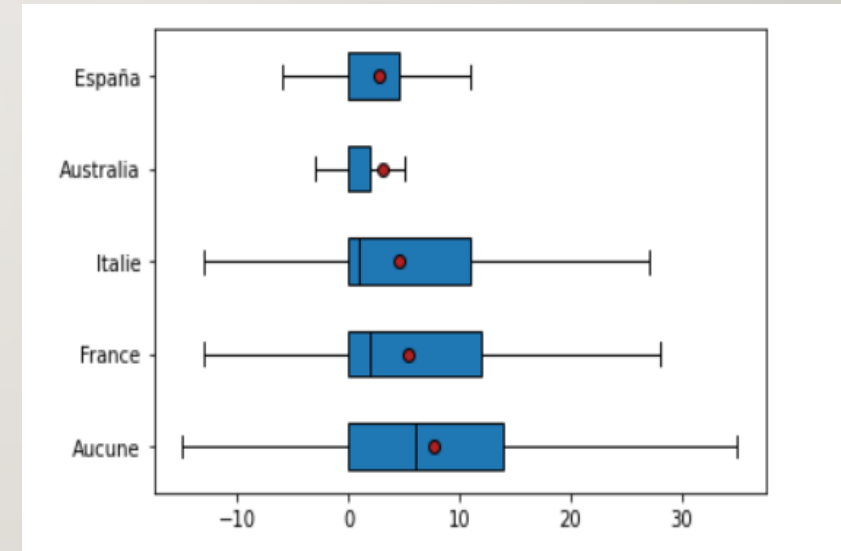


ANALYSE MULTIVARIÉE

Analyse Bivariée

Analyse par ANOVA

- nutrition-score-fr_100g et origins
 - ✓ Les nutrition-score-fr_100g des produits d'origine : "Aucune", France" et "Italie" sont très dispersés.
 - ✓ Les produit originaire de l'Australie sont de meilleurs qualités



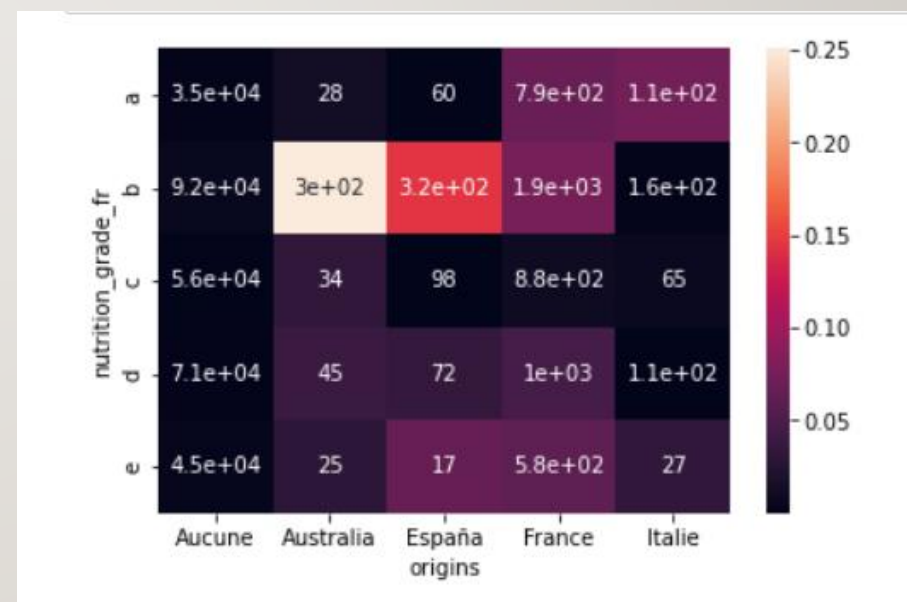
ANALYSE MULTIVARIÉE

Analyse Bivariée

Analyse par Khi carré

□ nutrition-grade_100g et origins

Les produits originaires de l'Australie
ou de l'Espagne ont souvent de meilleurs
nutrition-grade_100g



ANALYSE MULTIVARIÉE

Analyse Exploratoire

Standardisation des données

Elle a consisté à :

- ❖ Supprimer les colonnes non-numérique ;
- ❖ Garder uniquement les variables de nutritions ;
- ❖ Centrer et réduire les données.

ANALYSE MULTIVARIÉE

Analyse Exploratoire

Standardisation des données

	0	1	2	3	4	5	6
count	320732.00	320732.00	320732.00	320732.00	320732.00	320732.00	320732.00
mean	0.00	0.00	0.00	-0.00	0.00	-0.00	0.00
std	1.00	1.00	1.00	1.00	1.00	1.00	1.00
min	-0.82	-0.66	-0.88	-11.09	-0.79	-1.28	-2.59
25%	-0.79	-0.66	-0.74	-0.00	-0.57	-0.89	-0.86
50%	-0.13	-0.23	-0.13	-0.00	0.01	0.02	-0.28
75%	0.20	0.03	0.02	-0.00	0.01	0.67	0.75
max	6.08	13.09	4.71	197.97	26.78	2.78	3.75

ANALYSE MULTIVARIÉE

Analyse Exploratoire

Variances expliquées

Les deux axes du premier plan factoriel expliquent plus la variance des données (**56%**).

]:

	Dimension	Variance expliquée	variance expliquée	cum. var. expliquée
0	Dim1	2.267868	32.0	32.0
1	Dim2	1.638209	23.0	56.0
2	Dim3	1.101184	16.0	72.0
3	Dim4	0.995921	14.0	86.0
4	Dim5	0.378747	5.0	91.0
5	Dim6	0.335231	5.0	96.0
6	Dim7	0.282863	4.0	100.0

ANALYSE MULTIVARIÉE

Analyse Exploratoire

Représentation des variables

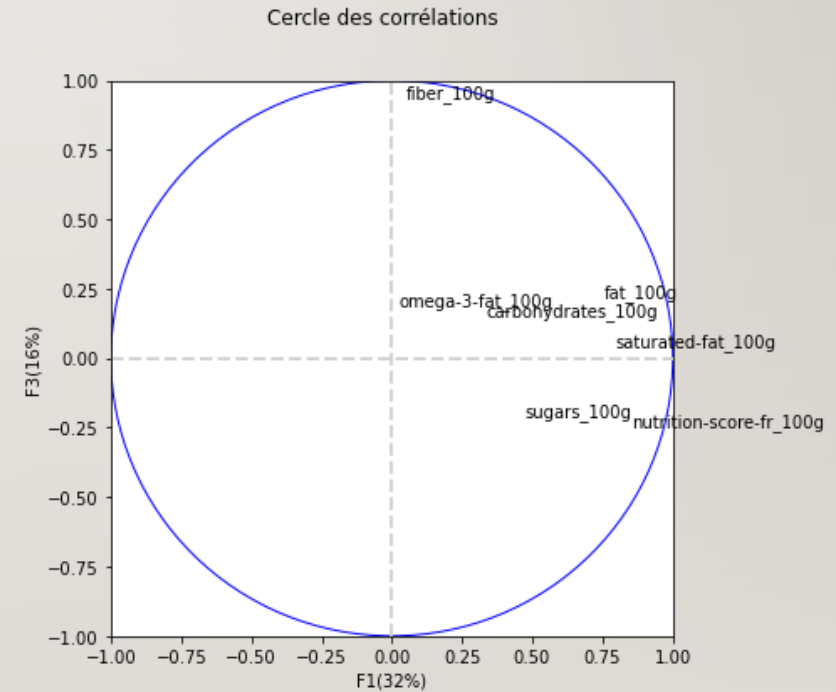
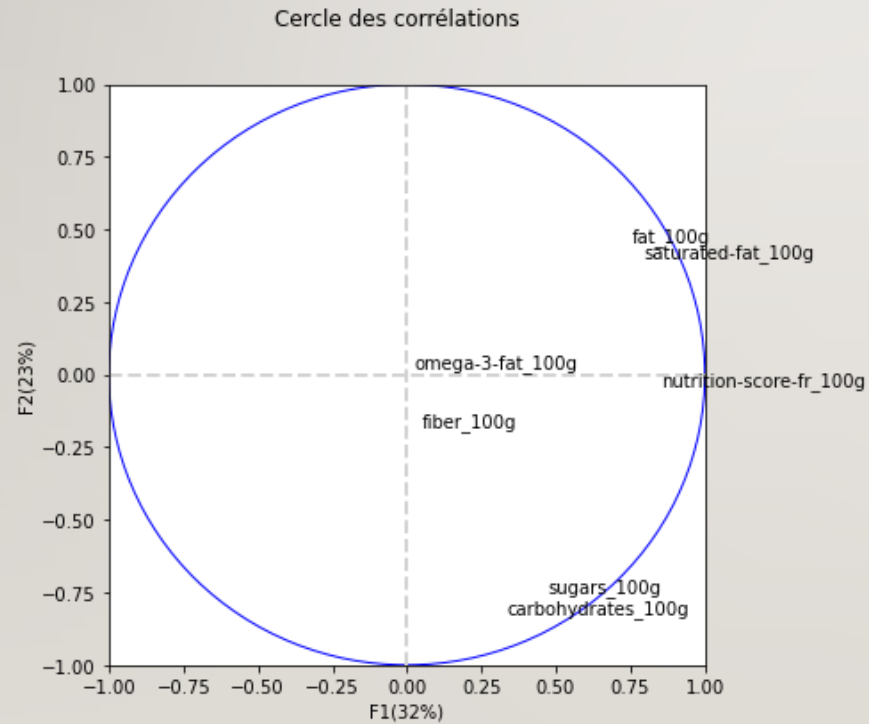
Les trois (3) premiers axes expliquent **(72%)** de la variance des données.
Nous allons poursuivre, notre analyse sur :

- Le plan constitué par les axes F1 et F2 (premier plan factoriel).
- Le plan constitué par les axes F1 et F3

ANALYSE MULTIVARIÉE

Analyse Exploratoire

Représentation des variables




CONCLUSION

Au vu de nos différentes analyses, il ressort les informations suivantes :

Les produits en provenance de l'Australie ou l'Espagne ont plus de chance d'avoir un meilleur nutrition-score que ceux originaires d'autres pays ou d'origine inconnue.

Le nutrition-score ne répond pas efficacement à notre problématique alimentaire des personnes souffrantes ou voulant se prévenir du diabète; car il :

CONCLUSION



Permet de prévenir les personnes diabétiques sur les risques de la consommation de certains produits.

Ne permet d'orienter les malades du diabète dans le choix d'aliments riches en nutriments favorables aux personnes diabétique.



MERCI

OPENCLASSROOM