Concevez une application au service de santé publique

L.khellouf[1] Mentor: B.Beaufils[2]

[12]OpenClassRooms

Projet 2, Decembre 2020

Sommaire

- Problèmatique
- Idée d'application
- Présenatation des données
- Analyse des indicateurs
- Dimensionality reduction avec la PCA
- Prediction du nutriscore grade
- Evaluation et Résultats
- Conclusion

Introduction

L'agence "Santé publique France" a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation.
Le jeu de données Open Food Fact est disponible sur le site officiel: https://world.openfoodfacts.org/.
L'objectif et de pouvoir proposer une application à partir de ces données.

Idée d'application

Le calcul du nutriscore existe mais il demande d'avoir accès à tout un tas d'indicateur ce qui en pratique impossible dans la vie de tous les jours. Je propose cette application qui estime la valeur à partir de très peu d'information.

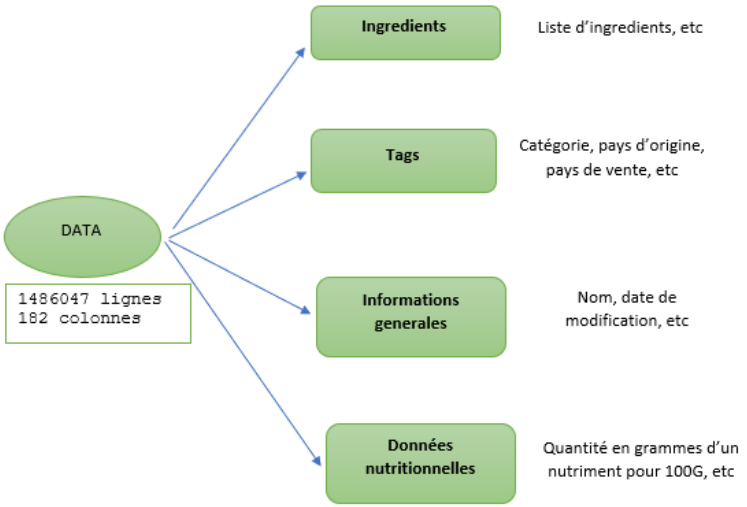
Présenatation des données

Open Food Facts est une base de données de produits alimentaires créée par tout le monde , pour tout le monde et elle utilisé pour faire de meilleurs choix alimentaires.

- Dans un premier temps, la suite de notre travail consistera par décrire le source de données, suivi du nettoyage de la donnée.
- Dans un second temps, de réaliser une analyse exploratoire, de proposer une application et enfin nous conclurons.

Présentations des données

Les champs des données sont séparés en quatre sections :

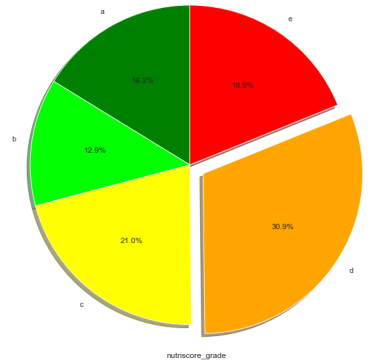


Notre analyse exploratoire sera en partie sur les données nutritionnelles.

Score final variant de **-15 (qualité nutritionnelle élevée)** à **40 (faible qualité nutritionnelle)**

Aliments (points)	Boissons (points)	Couleurs	Logo
<0	Eau	Vert foncé	
0 à 2	<2	Vert clair	
3 à 10	2 à 5	Jaune	
11 à 18	6 à 9	Orange	
19 à 40	10 à 40	Rouge	

Notre dataset contient 31.9% le nutriscore grade "e"



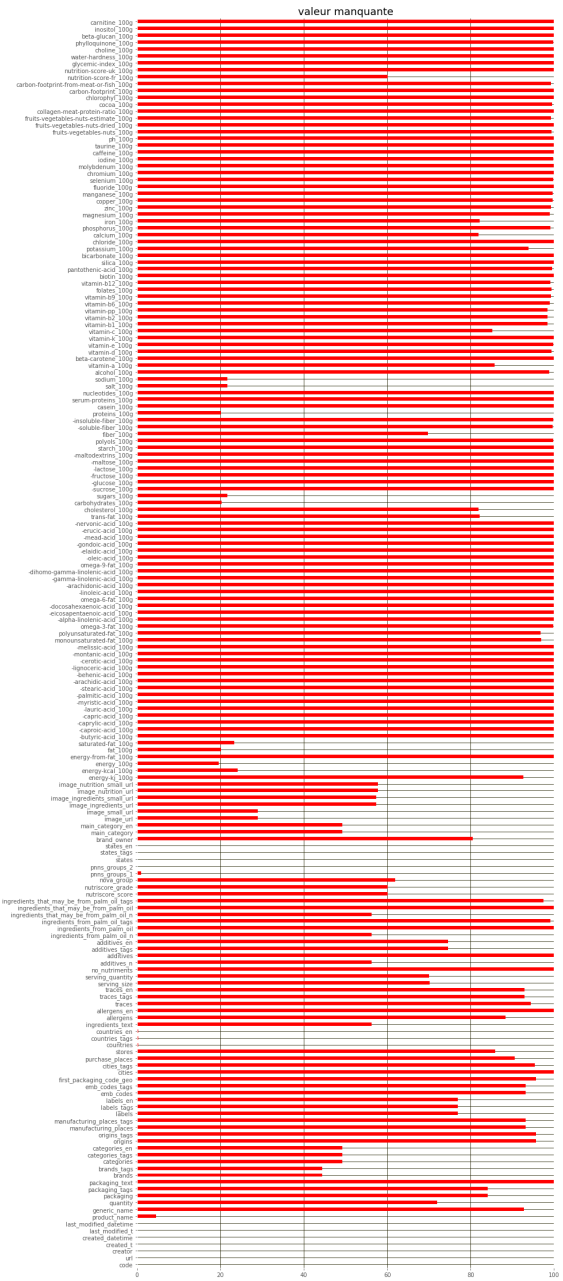
Analyse Exploratoire

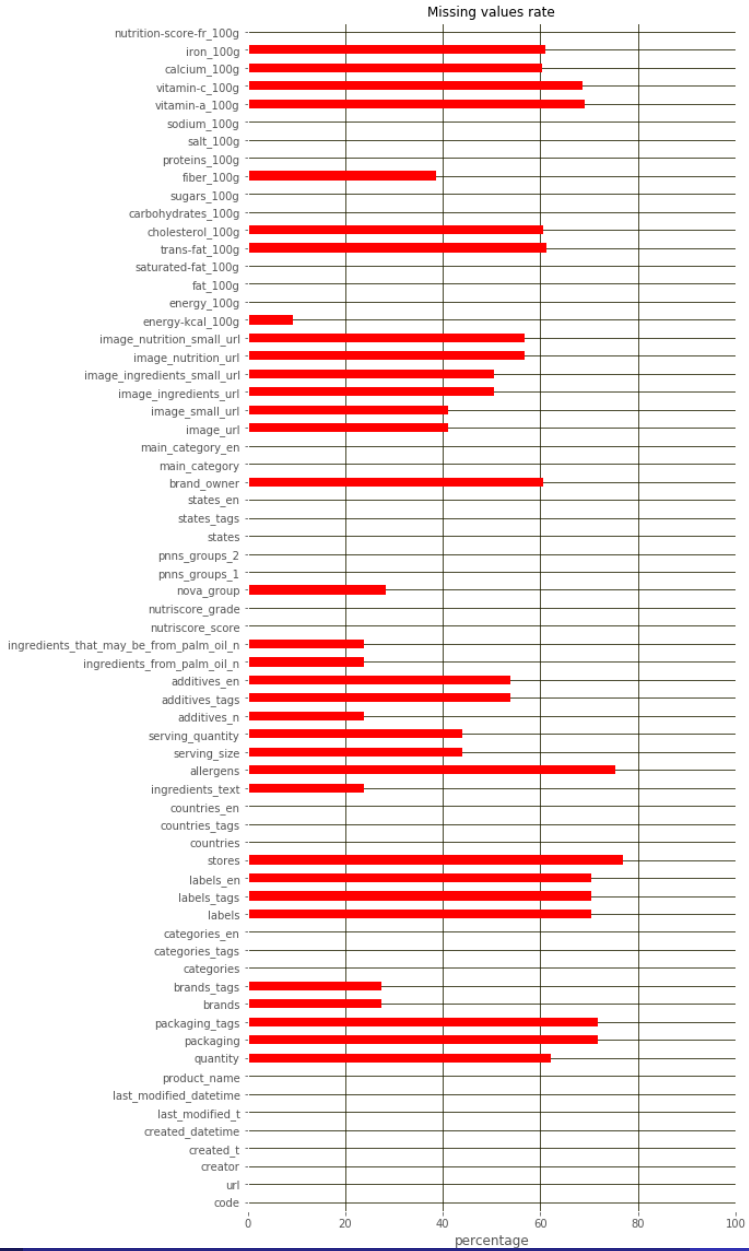
Cette analyse consistera à sélectionner uniquement les indicateurs nutritionnel et le grade nutritionnel avec la suppression des colonnes avec 80% de nan, des lignes dupliquées et les lignes non pertinents.

Etapes	Suppression et set.nan	Dataset avant	Dataset après
Lignes pertinentes	Nutriscore et nutrigrade est nan	1486047*lignes*	591485*lignes*
Colonnes pertinentes	80% de nan et 45 non_pertinents	182 colonnes	66 colonnes
Valeurs aberrantes	$val_e nergie > 1000kcal$	22216 outliers	_
Valeurs aberrantes	$prot\acute{e}ine + glucide + gras > 100, 0 > val\_100g > 100$	22216 outliers	_
Boissons	139880 boissons	(591485, 66)	(451605, 66)
Lignes dupliquées	482 dupliquées	241 lignes supprimé	_
Dataset shape	$(nbre\_produits * nbre\_features)$	(1486047, 182)	(451605, 21)

Table: Les étapes de l'analyse exploratoire

Analyse Exploratoire



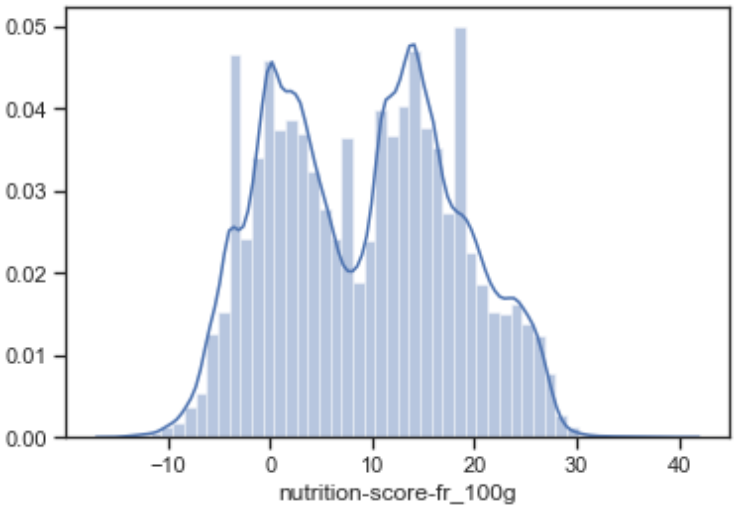


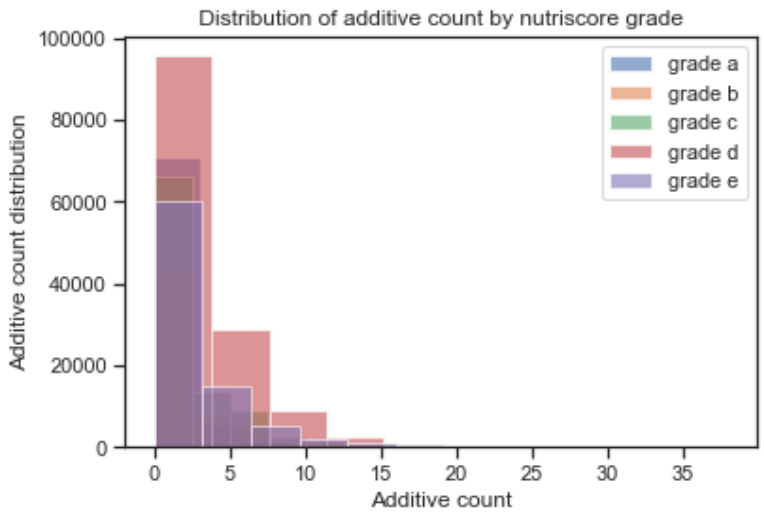
SimpleImputer est une classe scikit-learn qui est utile pour gérer les données manquantes dans l'ensemble de données du modèle prédictif. Il remplace les valeurs NaN par un espace réservé spécifié. Il est implémenté par l'utilisation de la méthode SimpleImputer () qui prend les arguments "median" comme stratégie. Cette méthode à été appliqué sur les données de Test et les données de Training séparément.

skewness: c'est un coefficient d'asymétrie correspant à une mesure de l'asymetrie de la distribution d'une variable aléatoire réelle puisque skewness=0.00 donc la distribution est symétrique

kurtosis: coefficient d'aplatissement. kurtosis=-0.99

- le grade prédominant du nutriscore grade ici est le grade 3 = grade "d" ce qui signifie que beaucoup de produits ont une note inférieure à la moyenne





Kolmogorov–Smirnov Test

- En statistiques, le test de Kolmogorov-Smirnov est un test d'hypothèse utilisé pour déterminer si un échantillon suit bien une loi donnée connue par sa fonction de répartition continue, ou bien si deux échantillons suivent la même loi.

```
Features:  additives_n KstestResult(statistic=0.5405888887175947, pvalue=0.0)
Features:  ingredients_from_palm_oil_n KstestResult(statistic=0.5, pvalue=0.0)
Features:  energy-kcal_100g KstestResult(statistic=0.985368131157875, pvalue=0.0)
Features:  energy_100g KstestResult(statistic=0.9965767540674306, pvalue=0.0)
Features:  fat_100g KstestResult(statistic=0.8951303399573969, pvalue=0.0)
Features:  saturated-fat_100g KstestResult(statistic=0.8303033723166535, pvalue=0.0)
Features:  trans-fat_100g KstestResult(statistic=0.5, pvalue=0.0)
Features:  cholesterol_100g KstestResult(statistic=0.5, pvalue=0.0)
Features:  carbohydrates_100g KstestResult(statistic=0.9448859809247595, pvalue=0.0)
Features:  sugars_100g KstestResult(statistic=0.8804484450664306, pvalue=0.0)
Features:  fiber_100g KstestResult(statistic=0.8301894192347086, pvalue=0.0)
Features:  proteins_100g KstestResult(statistic=0.9137139034611078, pvalue=0.0)
```

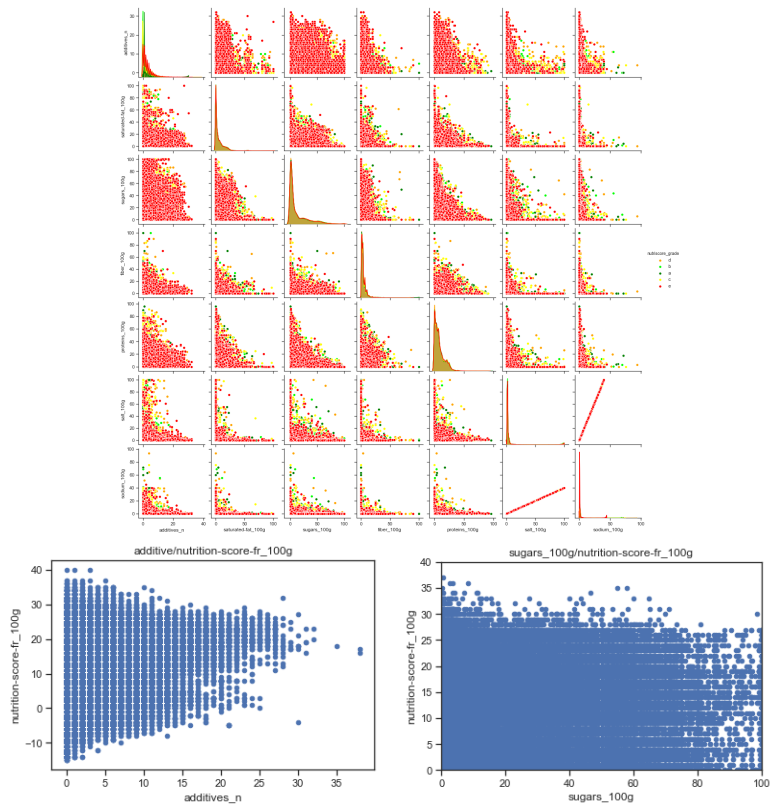
Les échantillons ne suivent pas la loi normal car p_value=0.0

Analyse des indicateurs

Analyse du Pair Plot

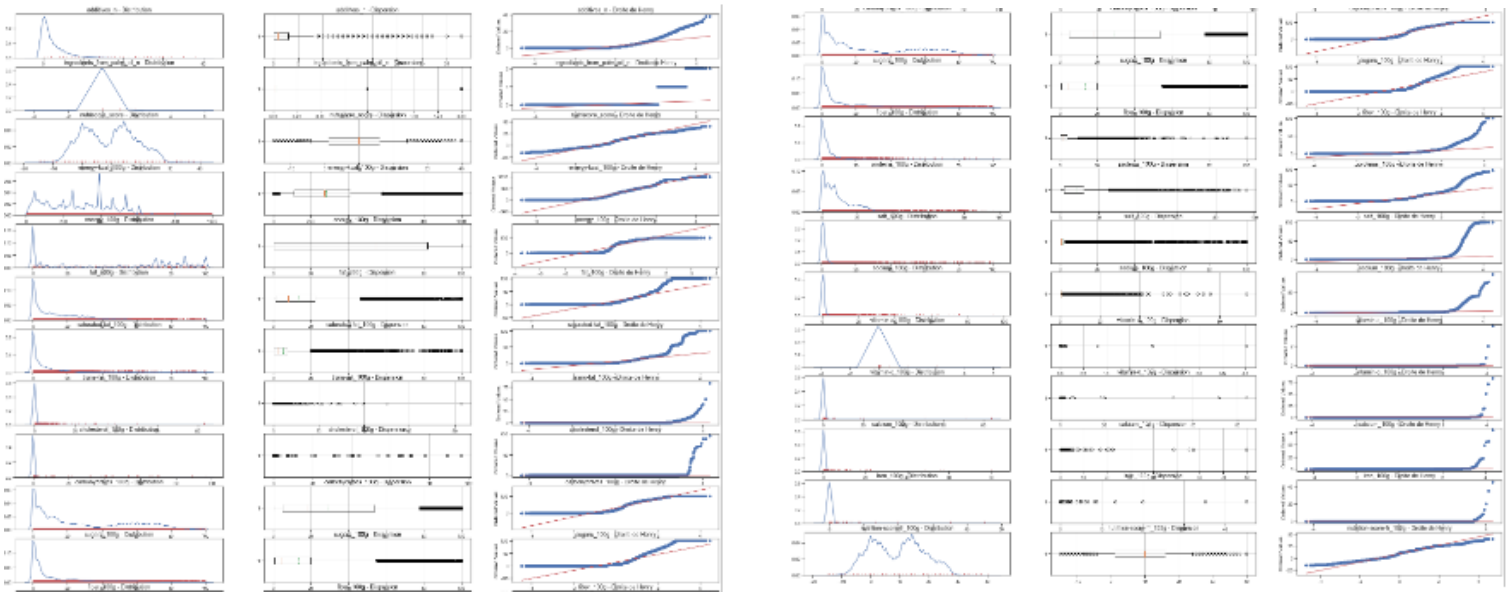
Interprétation du pairplot:

- Plus une representation tend vers une droite, plus les deux variables composants le graphique sont des doublons pour nos analyses:
- Pour cela nous pouvons supprimer les indicateurs suivants: salts et sodium
- Ce pairplot nous permet aussi d'analyser quel indicateur est correlé au grade nutritionnel, grace à la couleur représenatant chaque lettre de l'echelle.



Analyse des indicateurs

Analyse des distributions

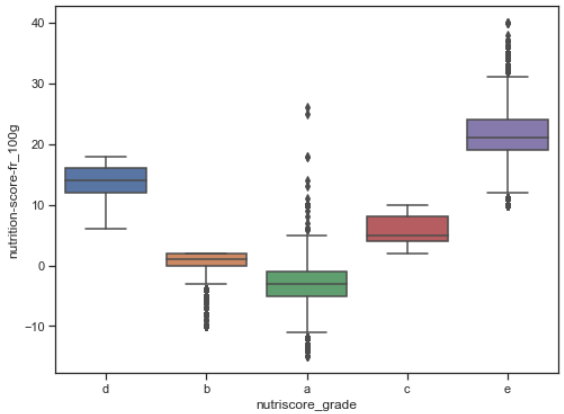


- La plupart des distributions ont une courbe ressemblant à des distributions Gaussiennes asymétriques à gauche. - La droite de Henry permet de vérifier visuellement qu'une distribution est normale. Le principe est simple. Si les points suivent la droite, alors la distribution est Gaussienne. Dans notre cas, nous pouvons constater que les distributions ne sont pas Gaussienne.

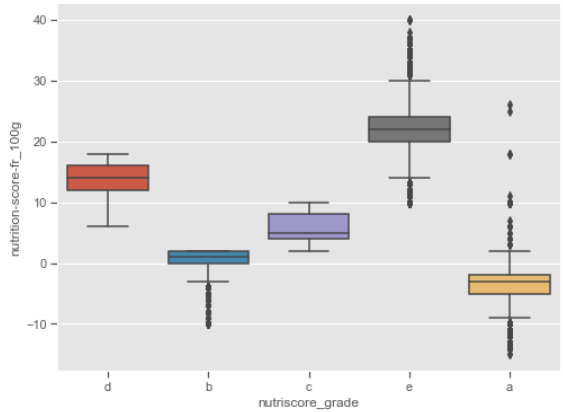
incohérence entre nutriscore et nutrigrade

Le calcul du nutriscore est différent pour les boissons et Aliments.

Le Scatter plot ci_dessous montre l'incohérence entre le nutriscore et le nutrigrade avant la suppression des boissons



La distribution du nutri_grade après la suppressions des boissons.



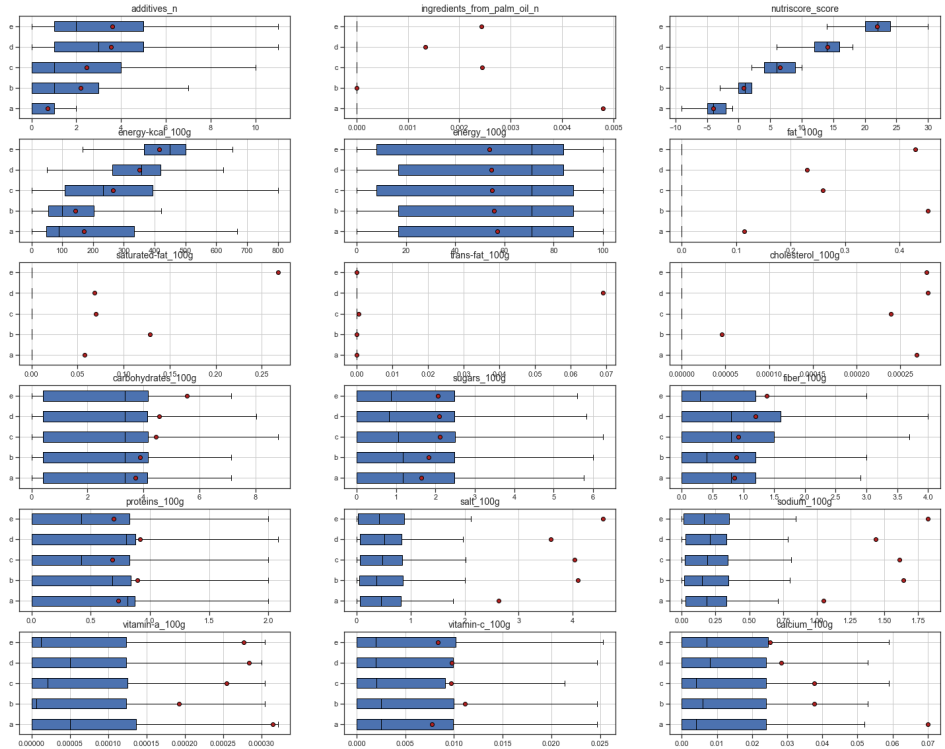
Rappel coefficient linéaire : C'est un coefficient compris entre -1 et 1. Plus la valeur est proche de 1 ou -1, plus la corrélation est vraie et la relation est linéaire. Si la valeur est proche de 0, il n'y a pas de corrélation linéaire entre les deux variables. Un modèle de prédiction linéaire n'est donc pas adapté pour faire de la prédiction.

Cependant il existe d'autres méthodes pour analyser la corrélation. Nous allons approfondir cette analyse avec deux autres méthodes.

ANOVA : Une analyse de la covariance entre deux variables (quantitative et qualitative)

Analyse des indicateurs

Analyse multivarié:

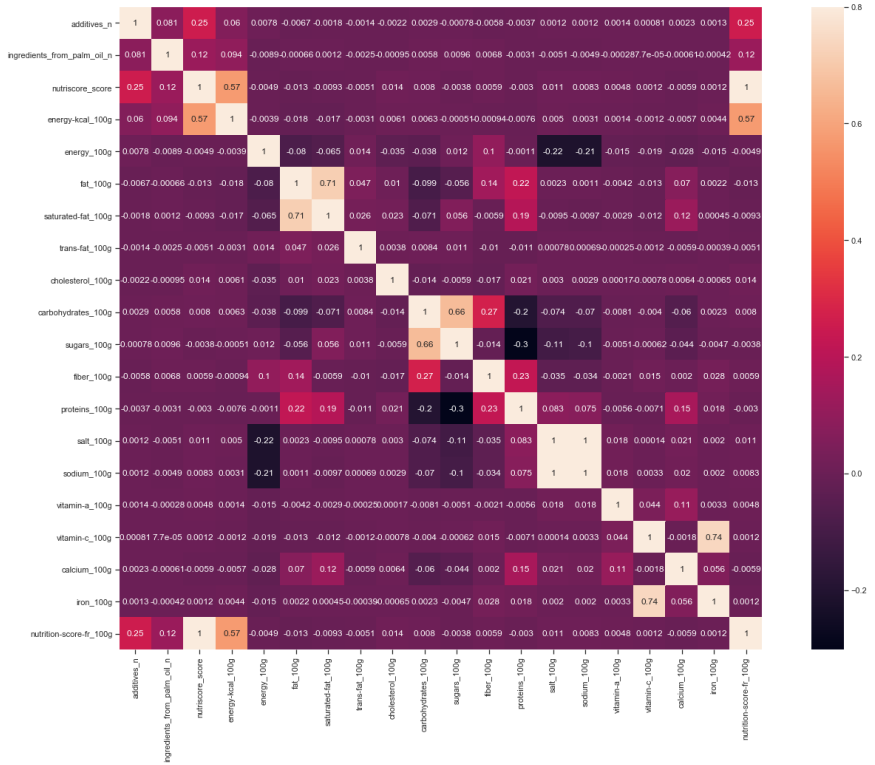


- 3 indicateurs sont corrélés par rapport au nutriscore grade. car plus le grade est mauvais plus leur valeur augmente:
- nombre d'additive
- nutrision score "logique"
- Energy_kcal
- nutriscore grade ne dépend pas des protéines et du sel

Analyse des indicateurs

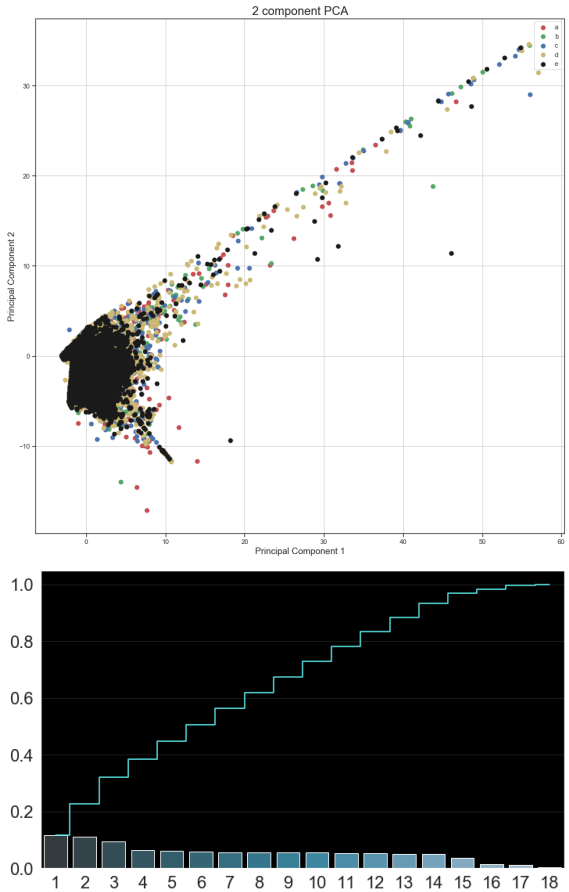
Analyse multivarié:

Matrice de corrélation: Cette matrice nous donnera un aperçu de la corrélation linéaire entre nos variables.
Le coefficient de corrélation linéaire permet de chiffrer le lien de corrélation entre 2 variables. Sa valeur varie en -1 et 1. Si elle est proche de 0, il n'y a pas de corrélation entre les variables. En revanche, si elle est proche de -1 ou 1, cela veut dire qu'une corrélation existe. Par contre, ce coefficient n'indique pas à quel point les variables sont liées entre elles.



Dimensionality reduction avec la PCA

- L'ACP est une technique qui transforme un ensemble de données de nombreuses caractéristiques en composants principaux qui ≪résument≫ la variance qui sous-tend les données
- Chaque composante principale est calculée en trouvant la combinaison linéaire de caractéristiques qui maximise la variance, tout en assurant également une corrélation nulle avec les composantes principales précédemment calculées

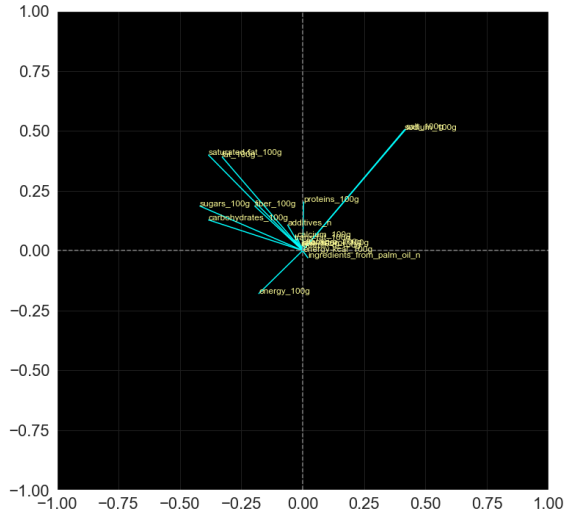


Dimensionality reduction avec la PCA

Analyse de la PCA:

- Les variables selt et sodium sont corrélées positivement à P2
- Les variables fat , saturated_fat et protiens sont corrélées négativement à P1
- Les variables carbohydrates, sugar sont corrélées négativement à P1.
- l'enrgy_100g est corrélée négativement à la fois P1 et P2;
- Les autres variables sont faiblement corrélées et mal représentées à la fois P1 et P2.

L'angle qui sépare "fat et saturated_*fat*" , "selt et sodium" et "sugars et carbohydrate" est proche de 0° donc leur coefficient de corrélation est proche de 1 : Ces variables sont très corrélées.



Après l'analyse de la matrice des corrélations analyse des cercles corrélations ACP on supprime les variables suivantes: fat_100g , selt_100g ,carbohydrates_100g.

Prediction du nutriscore

les points importants pour bien cerner le problème :

- Les données d'entraînement sont étiquetés donc la tache est dite supervisée.
- Nous charchons une valeur pour le nutrition-score-fr_100g, donc c'est un problème de régression.
- Notre dataset contient plusieurs indicateurs donc c'est une régression multivariée.
- Nous cherchons ensuite le nutriscore_grade, qui est un problème de classification.

Prediction du Nutriscore

La dataset à été partitionnée en 70% données d'entrainement (Training set) et 30% données de test (testing set)

Regression lineaire:

Le modèle de régression linéaire est un modèle de prédiction simple qui est très efficace quand la corrélation entre les variables est très forte. Plus les points (produits) sont resserrés autour de la droite de régression linéaire, plus les prédictions seront performantes.

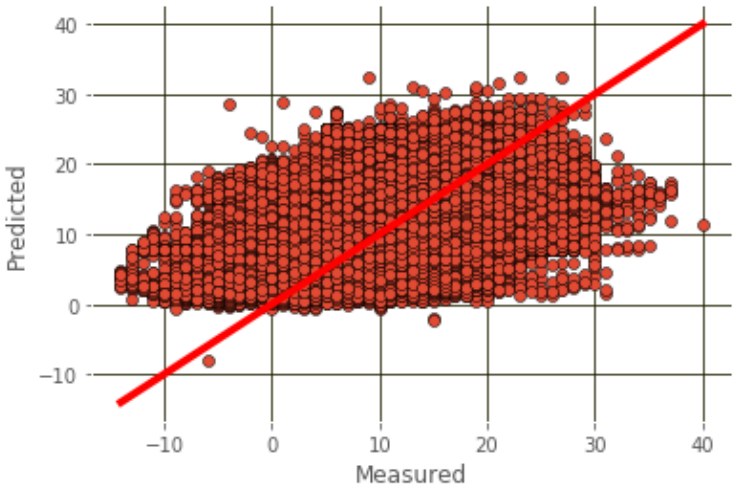
K-Nearest Neighbors

L'algorithme KNN est très mature en théorie. Ses idées simples et faciles à comprendre et sa bonne précision de classification en font un outil largement adopté. Le processus spécifique de l'algorithme comprend principalement les quatre étapes suivantes: Préparation des données, Calculer la distance, Rechercher des voisins, Classification de décision.

Prédiction du Nutriscore

Regression pour la prédiction du nutrition-score-fr_100g

1. La performance du modèle Regression lineaire sur la base de test
   1. Lerreur quadratique moyenne est 7.27. Cela veut dire que pour une prédiction, notre modèle se trompera de 7.19 sur le nutrition-score-fr_100g.
   2. le score R2 est 0.37



1. La performance du modèle knn regressor sur la base de test
   1. Avec k= 9 ,le score R2 est: 0.42

Product	Actual	Predicted
846232	13.0	9.555556
944372	0.0	7.888889
368386	15.0	14.666667
898404	14.0	15.111111
648061	23.0	9.555556

Table: Exemple de prédiction .

Prédiction du Nutriscore

Classification pour la prédiction du nutriscore_grade

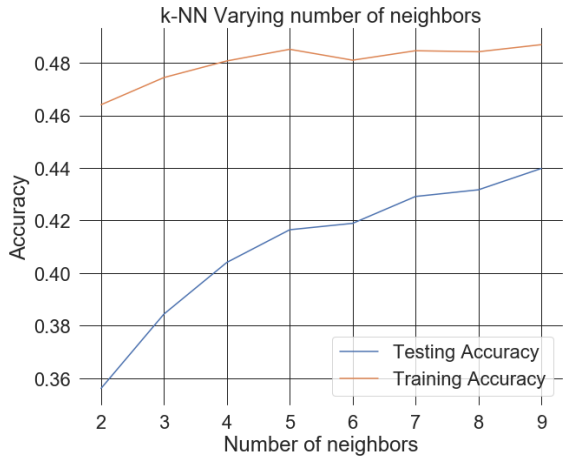
La matrice de confusion

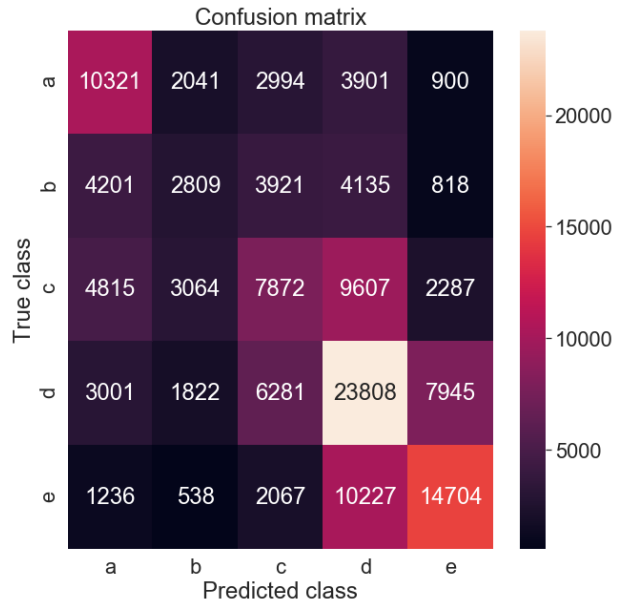
est un tableau souvent utilisé pour décrire les performances d'un modèle de classification (ou ≪classificateur≫) sur un ensemble de données de test dont les vraies valeurs sont connues.

Le coefficient de corrélation de Matthews

est utilisé dans l'apprentissage automatique comme une mesure de la qualité des classifications binaires (à deux classes)

Avec k=9 on ontient : knn_ accuracy_score: 0.43 knn_ MCC : 0.29





Merci pour votre attention