# Explore Open Food Facts: Nutrition Facts and Food Health Ranks

1896548 CHIH-YUAN YANG

***Abstract* – With industrialization, nowadays foods are easy to prepare and acquire in developed economies, but do we really "know" our foods? In this study, we'll explore the nutrients information of the things we consume, using dataset acquired from Open Food Facts project, which has more than 700,000 entries. We'll explore the dataset to verify some common beliefs about foods, such as 'Do products with higher fiber contents necessary healthier?', and 'Are products with "Organic" labels having higher health ranks?' After the exploratory questions, we'll also model some aspects of foods. One of the models focuses on estimating the cocoa content of a chocolate products, another one aims to conduct binary classifications to estimate the product to be healthy or not with main nutrients information.**

## I. Introduction

### *A. What do we know about the food we purchase?*

A survey conducted by FDA in 2014 [1] asked participants "How often do you use the Nutrition Facts label when deciding to buy a food product?", and the responses showed that 16% out of 1244 participants answered "Always", 34% answered "Most of the time", and 27% answered "Sometimes". The survey showed that at least half of the population would try to understand the contents of the product they'll buy.

In the same survey, participants were also asked: "People have different reasons for not using the nutrition information on the food label ("Nutrition Facts" table, shown in *Figure 1*). Please say whether you strongly agree, somewhat agree, somewhat disagree, or strongly disagree with each of the following reasons for not using the food label". 59% of the surveyees answered with Strongly/Somewhat agree on "The information is hard to understand". Clearly, most people find the nutrient information necessary but the product labels hard to understand.

In this study, we're going to explore Open Food Facts dataset acquired from Open Food Facts project. The project started in around 2012 with the goal to make information of food products available to everyone, and all the data are gathered by thousands of contributors from around the world. The dataset includes various information found on packaging of food products, from basic data such as product name or amount of protein contained (found in Nutrition Facts, as shown in *Figure 1*) to particular content such as amount of cocoa/zinc/magnesium contained.

**Nutrition Facts**
Serving Size 2/3 cup (55g)
Servings Per Container About 8

**Amount Per Serving**
**Calories** 230 — Calories from Fat 72

| | % Daily Value* |
|---|---|
| **Total Fat** 8g | **12%** |
| Saturated Fat 1g | **5%** |
| *Trans* Fat 0g | |
| **Cholesterol** 0mg | **0%** |
| **Sodium** 160mg | **7%** |
| **Total Carbohydrate** 37g | **12%** |
| Dietary Fiber 4g | **16%** |
| Sugars 12g | |
| **Protein** 3g | |
| Vitamin A | 10% |
| Vitamin C | 8% |
| Calcium | 20% |
| Iron | 45% |

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on your calorie needs.

| | Calories: | 2,000 | 2,500 |
|---|---|---|---|
| Total Fat | Less than | 65g | 80g |
| Sat Fat | Less than | 20g | 25g |
| Cholesterol | Less than | 300mg | 300mg |
| Sodium | Less than | 2,400mg | 2,400mg |
| Total Carbohydrate | | 300g | 375g |
| Dietary Fiber | | 25g | 30g |

**Nutrition Facts**
8 servings per container
**Serving size** **2/3 cup (55g)**

**Amount per serving**
**Calories** **230**

| | % Daily Value* |
|---|---|
| **Total Fat** 8g | **10%** |
| Saturated Fat 1g | **5%** |
| *Trans* Fat 0g | |
| **Cholesterol** 0mg | **0%** |
| **Sodium** 160mg | **7%** |
| **Total Carbohydrate** 37g | **13%** |
| Dietary Fiber 4g | **14%** |
| Total Sugars 12g | |
| Includes 10g Added Sugars | **20%** |
| **Protein** 3g | |
| Vitamin D 2mcg | 10% |
| Calcium 260mg | 20% |
| Iron 8mg | 45% |
| Potassium 235mg | 6% |

* The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

FIGURE 1
"NUTRITION FACTS" TABLE ON FOOD PRODUTS

During the exploratory stage, we'll be using this dataset to answer some interesting exploratory questions about foods. The goal is to check if some of our common beliefs can hold true using the huge dataset we have.

In the modelling stage, we'll firstly try to estimate cocoa contents through available information on Chocolate product. Next, we'll try to estimate if a product is healthy or not using only the information included in "Nutrient Facts" table.

### *B. Evaluate the health values of foods by Nutri-Score*

There have been several methods to label the healthy values of food products. Currently in UK, most major food producers are using 'Traffic Lights' as front-of-pack label (introduced in II. Background) for nutrient guidelines. However, the issue with 'Traffic Lights' is that it's just a shorter version of "Nutrition Facts" table and doesn't provide an intuitive way to categorize food products.

In this study, we'll be using a food rank method called 'Nutri-Score' (shown in *Figure 2*), which is also available in Open Food Facts dataset. In short, 'Nutri-Score' is a colour-coded label system to categorize food products into one of the five categories from A to E. The formula used to compute the score for each category is made publicly available, which taking the main nutrients (i.e. calories, fats, sodium, sugars and proteins) into account, plus some other variables.

In our analyses, we'll rely on 'Nutri-Score' as an evaluation criterion for healthy values.


FIGURE 2
"NUTRI-SCORE" LABELLING SCHEME

## II. Background

### A. Open Food Facts project

Open Food Facts [2] is a free, online and crowdsourced database of food products from around the world. [3] The project was launched in 2012 by French blogger Stéphane Gigande, with the goal to gather data on food products from around the world and make them openly available to everyone.

For each item, the database stores its several features, including generic name, quantity, type of packaging, brand, category, manufacturing or processing locations, countries and stores where the product is sold, list of ingredients, any traces (for allergies, dietary laws or any specific diet), food additives and nutritional information.

Almost all of the data is added by contributors, since 2012, 7,000+ contributors have added 700,000+ products from 144 countries.

### B. Labelling Scheme by UK, *"Traffic Lights"*

Since 2013, UK has adopted a new food labeling scheme nicknamed as "Traffic Lights", named after its colour-coded nature (shown in *Figure 3*). The label tries to provide easier to understand information by using "Green" as low, "Yellow" as medium and "Red" as High.

### C. Nutri-Score Labelling Scheme

Since 2017, France has implemented a newly developed 5 colour Front-of-Pack (FOP) nutrition labelling system named "Nutri-Score" [4]. Nutri-Score is derived from UK Food Standard Agency nutrient profiling system (FSA score), which is computed taking into account nutrient content per 100g for food and beverages. It allocates positive points for "un-favourable" content: energy (kJ, 0-10 points), total sugar (g, 0-10 points), saturated fatty acids (g, 0-10 points) and sodium (mg, 0-10 points). Negative points are allocated for "favourable" contents: fruits, vegetables and nuts (0-5 points), fibers (0-5 points) and proteins (0-5 points).

The total from positive (0-40 points in total) and negative (0-15 points) points is computed, yielding a global score ranging from -15 for the healthiest foods (0 positive points and 15 negative points) to +40 for less healthy foods (40 positive points and 0 negative points). *Figure 4* show the computation formula.

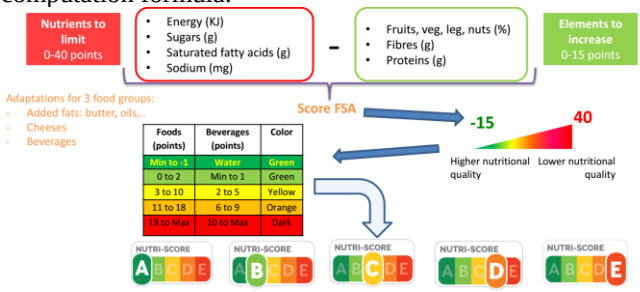


FIGURE 4
"NUTRI-SCORE" RANK METHODS [5]

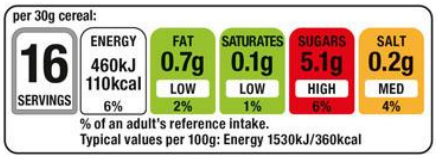


FIGURE 3
"TRAFFIC LIGHTS" LABELLING SCHEME

## III. Dataset

The dataset used for this report can be downloaded directly from Open Food Facts [6], which is provided under Open Database License [7]. The website also provides a search engine allowing visitors to select and download only the content they want.

In this study, we'll be using the full dataset with preprocessings to extract information needed for analysis. The full dataset contains 173 attributes and 710,013 rows, most of the attributes contain NA values. There're several possible interpretations for NA values, one is that the lack of information on product packages causing contributors not filling the cell, e.g. for alcoholic drinks, there's usually limited nutrient information available; another is that the product doesn't contain that kind of nutrient content specified by the attribute, e.g. for non-chocolate product, attribute "cocoa_100g" is likely to be left blank.

In the following analysis, we'll be assuming the possible reasons for certain NA values and then processing the data accordingly. Since the data are added by various contributors, there're also several anomalous outliers in the dataset, which will be dealt with during preprocessing.

For purpose of easier presentation and description, here we roughly categorize all 173 attributes into 5 types as following:

- **Basic Product Features**:

  Attributes in this type are the basic information of food products, such as its unique code, its product name, brand, its category (of food type), its packaging style, serving quantity, its ingredients in text, its labels for anything (e.g. label for organic product), its processing stats, URL link to its product image, datetime when it's added …etc.

  Of all the attributes, we'll be using the "category" to identify certain kind of food, the unique "code" to identify the unique items in the dataset, and the "label" to find products with some special features. Each cell in "category" contains several possible categories for a product, it may include its main category, sub-category, or even sub-sub-category. During analysis, we'll extract its category type depending on the analysis question.

- **Geography Features**:

  This type includes the geographical information of a food product, i.e. its origin, the countries where it's sold, the country where it's made, the store where its bought…etc.

  One notable thing about geography feature in Open Food Data is that it's distributed across 144 countries but centered around European ones. Roughly at least one

third of the data is about food products from France or its neighboring countries, and data from France usually has the most completed information. Part of the reason is the project starter himself is from France, and the project is known to many in France.

Since the dataset is biased toward Europe, in this study we will focus on using the dataset as a whole, instead of using the geography features for cross-country analyses.

- **Main Nutrient Features**:

  Main nutrition features are those that are almost always available on all food product packaging, i.e. "Calories/Energies", "Total Fat" (including Saturated Fat and Trans Fat), "Sodium/Salt", "Proteins" and "Sugars". These attributes are easily available from table of "Nutrition Facts", as shown in *Figure 1*. In the dataset, all the nutrient attributes are normalized to per 100 grams of the food product, e.g. 'energie_100g' is 'Out of 100 grams of the product, how much energy does it have?'.

  Because of their high availability, it would be uncommon if the data have missing values on any of these attributes. Therefore, in the preprocessing, any data have missing value in any of these attributes will just be removed.

- **Other Nutrient -related Features**:

  This includes the nutrient contents that are not always labelled on a product and the additives listed in product ingredients.

  For the nutrient contents, most of the cells could be NA, only for certain category of foods would have values. For example, "cocoa_100g" will have value only if it's chocolate-related products; "alcohol_100g" will have value only if it's alcoholic beverages; "-montanic-acid_100g" will have value only on some special product type.

  For the additives, the dataset includes additives in text and its total count. But only a small percentage of the dataset has included this information.

  In the analyses, we'll be using these data depending on the question proposed, and each NA value of nutrients will be removed or fill with 0 conditional on the type of analysis conducted.

- **Nutrition Rating Features**:

  There're 3 types of food rating available on Open Food Facts: "nutrition-score-fr_100g" and "nutrition-score-uk_100g" are both scores based on numeric values, each has its own formula to derive the scores, the scores ranging from -15 to 40; "nutrition_grade_fr" uses the same categorization method as "Nutri-Score" described in Background section, and it has categoric values from 'a' to 'e' with 'a' being the most healthiest and 'e' being the least.

  In the analyses, we'll be using 'nutrition_grade_fr' as a reference for the health value of a food product.

## IV. Tools for Preprocessing

All the preprocessings and analyses are done with Python. Python packages used are listed below.

- sklearn (sci-kit learn): for regression, classification and clustering modelling algorithms, also for some data processing tools.
- pandas: for reading, manipulating, and split-apply-combine dataset operations.
- numpy (Number Python): for algebraic manipulations.
- matplotlib, seaborn: for creating plots and graphs.

## V. Dataset Preprocessing

### *A. Subset on full dataset*

With 173 attributes, there're only some of them we'll be interested in this study. Firstly, data image URL and data status information are all removed. Secondly, attributes of trace elements are also removed, since only some of the products have values for them. After these two steps, there're still 119 attributes left.

In the third step, we decided that for this study, we would be only interested in following types of attributes: "Main Nutrients" and "related nutrients", including 'fiber_100g' (Amount of Fiber), 'cocoa_100g' (Amount of Cocoa), 'alcohol_100g' (Amount of Alcohol), 'caffeine_100g' (Amount of Caffeine) and 'n_additives' (Number of additives). Other product-related attributes, such as "Product Category", "Product Labels", "Product Code", are also kept. This left us a dataset with 710,013 rows and 25 attributes.

### *B. Handling Missing Values*

As stated earlier, we assume that "all main nutrients should be available on every product". Thus, if there're missing values in the main nutrients, the data will be removed. Among the attributes, we identify that these are the attributes should not have NA: 'sodium_100g', 'salt_100g', 'sugars_100g', 'fat_100g', 'proteins_100g', 'energy_100g', 'product_name' and 'code'. Any data is these attributes have missing values are removed.

One mentionable change in data removal is the 'alocohol_100g' attribute. Before removal, there're 6,216 rows have values in 'alcohol_100g', but data removal reduced the number to 389 rows. This is due to the lack of information on most alcoholic beverages, most of the alcoholic data don't have information on main nutrients.

After data removal, we have a new dataset having 560,536 rows and 25 attributes.

### *C. Handling Outliers with KMeans and boxplot*

With KMeans, we fit a model with 5 clusters on the dataset, the initial result (as in *Table 1*) shows unbalanced counts in each label, with the first label has almost all the data points. Also, the boxplots on numeric attributes also show some attributes have values greater than 10,000. Both of the results telling us there're outliers in the dataset.

Outliers removal is done with some logical assumptions: for attribute 'energy_100g' and 'energy-from-fat_100g', their unit is in kilojoule, thus the value should be between 0 and 10,000; for other nutrients, since thery're based on "every 100 grams", the values should be between 0 and 100. All the outliers met above conditions are removed. After the removal, an extra step is added to convert the unit of 'energy_100g' and 'energy-from-fat_100g' from kilojoule to kcal (kilocalorie).

We have also verified our outliers processing with KMeans and boxplots. KMeans returns a more balanced count in each cluster (as shown in *Table 1*) and boxplot shows a more evenly distributed boxplots ('energy_100g' and 'energy-from-fat_100g' are not plotted, since they have higher ranges and compress other boxplots). After outliers are removed, our dataset now has 559,935 rows and 25 attributes.

| | KMeans Lables | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| Initial Run | 560507 | 1 | 3 | 3 | 22 |
| After outliers removal | 92882 | 175441 | 126851 | 10148 | 154613 |

TABLE 1
KMEANS CLUSTERING RESULTS


FIGURE 5
BOXPLOTS AFTER OURLIERS ARE REMOVED

## VI. Research Questions

With cleaned dataset, we can start exploring some interesting questions about foods and modelling on some aspects of food products.

*Part1. Exploratory Questions*

A. *What are the correlations between the main nutrients?* And the correlations between nutrients and each Nutri-Score rank? Dose higher content in 'Fiber' indicate the product being healthier?
B. *Are all sweets (sugary snacks) unhealthy?* What is the sub-category of sweets that is the healthiest?
C. *What is the distribution of cocoa content among all chocolate products?* Some products state having more than 70% or 80% cocoa content, some only label their contents on the back-side in 'Ingredients', which are more common?
D. *Are products with 'Organic' label necessary healthy?* The definition of 'Organic' does not state it being healthier, but people tend to assume the correlation between these two concepts, what can we discover about organic foods from this dataset?

*Part2. Modelling Questions*

A. *Can we estimate cocoa content using regression model?* Although almost all governments require producers to label the main nutrients of food products in "Nutrition Fact" table, cocoa is not on the list. In fact, in the dataset, there're 5,795 products under the category of "Chocolates", but only 3,072 of them have values in the attribute of "cocoa_100g", i.e. only 53% chocolate products have labelled their cocoa content. We'll explore this information using regression models, and test their estimation accuracy.
B. *Can we classify the food products into binary category: healthy or unhealthy, instead of 5?* Using the rank value from Nutri-Score, we'll treat rank 'a' and 'b' as healthy and 'c', 'd' and 'e' as unhealthy, and then build binary classifiers using Logistic Regression, Nearest Neighbor and Decision Tree algorithms.

## VII. Exploratory Analysis

*A. Nutrients Correlations and Health Ranks*

We're interested in the correlations between nutrients and ranks of Nutri-Score. We can get a visualization of these correlations by heatmap as show in *Figure 6*. The categorical variable "nutrition_grade_fr" (rank for "Nutri-Score") has been transformed into dummy variables, i.e. each label "nutrition_grade_fr_[a-e]" in heatmap is denoting each category in "Nutri-Score".

By reading the color pattern of heatmap, we find that there're positive correlation between "Sugar" and rank of "e" in Nutri-Score and negative correlation between "Sugar" and rank of "a", i.e. more sugar, more likely to be ranked as 'e'. Another notable pattern is the correlation between "Fiber" and each rank. "Fiber" is positively correlated with rank of 'a', having correlation 0.21, but negatively correlated with rank of 'e', having correlation -0.1.

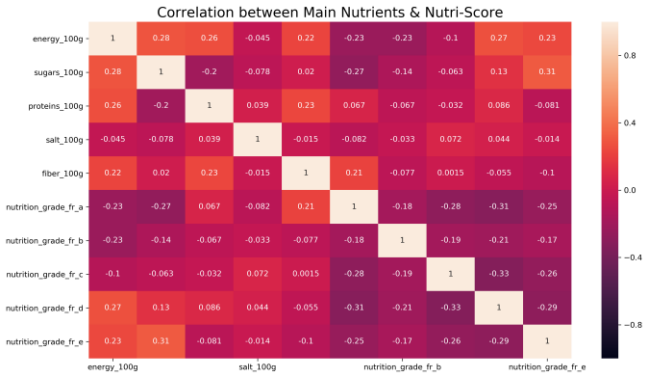


FIGURE 6
HEATMAP OF CORRELATIONS BETWEEN NUTRIENTS

*B. Are sweets (sugary snacks) all unhealthy?*

In this study, whether a food product is healthy is judged by its Nutri-Score rank. Here we want to understand the rank of all of the products in category 'Sugary Snacks'.

Products of 'Sugar Snacks' is firstly sliced from all dataset by regular expression string matching in attribute 'categories_tags', percentage of each Nutri-Score rank is computed (as show in *Table 2*, 'combined'). Next, each product in sub-category 'Sugary Snacks' is also matched in attribute 'categories_tags', the percentages are also shown in *Table 2*.

| | Nutri-Score rank (in Percentage) | | | | |
|---|---|---|---|---|---|
| **Sub-Category** | A | B | C | D | E |
| Bars | 2.53% | 2.65% | 29.71% | 51.83% | 13.27% |
| Biscuits-and-Cakes | 0.55% | 0.81% | 6.15% | 36.19% | 56.30% |
| Chocolates | 0.08% | 0.74% | 0.93% | 13.64% | 84.61% |
| Confectioneries | 0.59% | 6.59% | 7.23% | 40.69% | 44.90% |
| Popcorn | 4.06% | 1.11% | 15.87% | 54.98% | 23.99% |
| Viennoiserie | 0.19% | 0.31% | 5.39% | 55.05% | 39.06% |
| Sugary Snacks (combined) | 0.54% | 2.14% | 6.10% | 34.34% | 56.89% |

TABLE 2
NUTRI-SCORE RANK COUNTS IN PERCENTAGE
(SUB-CATEGORY OF SUGARY SNACKS)

From *Table 2*, we may conclude that indeed almost all sweets are less healthy, particularly, 'Chocolates' has over 80% ranked as 'e'. But if we're looking for a healthier option among sugar snacks, perhaps 'Bars' or 'Popcorn' are good alternatives.

*C. Cocoa content in Chocolate products*

Using the dataset from previous exploration (*Are sweets (sugary snacks) all unhealthy*), we can further extract products of sub-category "Chocolates". We then plot histograms with observations having value in attribute 'cocoa_100g'. The plot is shown in *Figure 7*, and each bin is the frequency count of chocolate products with certain number of cocoa contents in it.

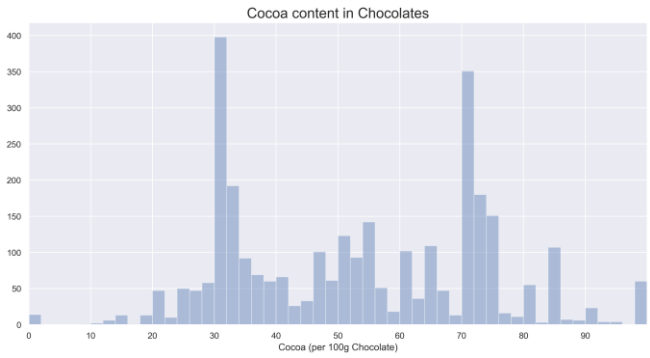


FIGURE 7
HISTOGRAM OF COCOA CONTENT IN CHOCOLATES

The above plot shows that the cocoa content distribution is bimodal, with most products on the market have 30% or 70% cocoa content.

*D. Organic Labelled food products*

By definition, *Organic food products* means products that are grown under a system of agriculture without the use of harmful chemical fertilizers and pesticides with an environmentally and socially responsible approach [8].

Due to its emphasis on NOT using 'chemical fertilizers and pesticides', people tend to correlate organic products with being 'healthier'. Here, we want to analyze this correlation with the dataset.

Data of organic products are extracted from the dataset by regular expression matching in attribute 'labels_tags' with string 'en:organic' or 'en:eu-organic'. There're 19,505 products with organic label.

The first thing to check is the frequency of each rank in Nutri-Score. We compare the percentage of Organic products with Overall products (as show in *Figure 8*), and it's noticeable that Organic products have higher percentage at rank 'a' and lower percentage at rank 'd'.

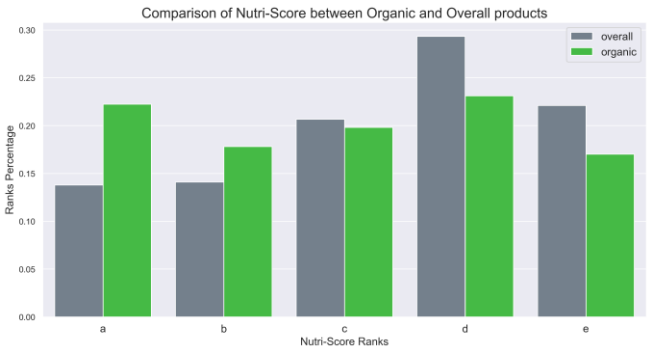


FIGURE 8
COMPARE NUTRI-SCORE RANKS OF ORGANIC FOOD PRODUCTS
WITH OVERALL FOOD PRODUCTS

Another view on organic products are comparing its nutrients with Overall products, using boxplots. The boxplots are plotted in *Figure 9*, excluding the outliers in each attribute. Some of the observations from the figures are:

- Organic has far fewer number of additives
- Not much differences on energies, proteins
- Organic has fewer salts and sugars
- Organic has higher fiber contents

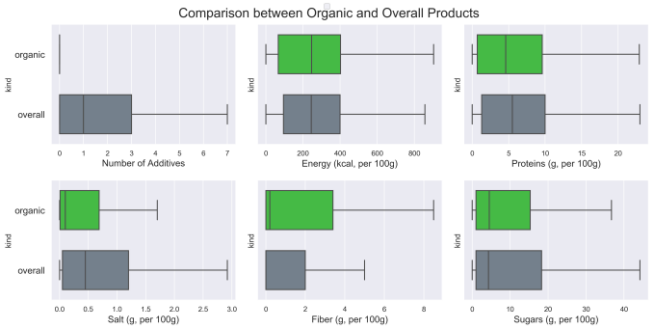


FIGURE 9
COMPARE NUTRIENTS OF ORGANIC FOOD PRODUCTS
WITH OVERALL FOOD PRODUCTS

## VIII. Modeling Analysis

### *A. Estimate Cocoa content with Regression*

With 47% chocolate products missing data on cocoa content, we want to use regression model to estimate the cocoa content of the remaining chocolate products.

The observations of 'Chocolate' category are split by 80/20 into 2,304 rows in training set and 768 rows in testing set. The attributes in the datasets contain both main nutrients ('energy_100g', 'fat_100g', 'sugars_100g', 'fiber_100g', 'proteins_100g', 'salt_100g') and other available information ('additives_n', 'saturated-fat_100g', 'carbohydrates_100g').

To find the best performance model using linear regression, we combine different modelling strategies, including "Data Normalization", "Polynomial Features" and "Regularization", the details are listed as below.

1. Data Normalization
   - Data is normalized (standardized) by Z-score
   - Z-score = $\frac{(x-mean)}{standard\ deviation}$, where x is each data
2. Polynomial Features
   - Each of the 6 attribute is squared, multiply by each other and multiply together, using sklearn function [9].
   - Resulting dataset has 66 attributes (excluding label).
3. Regularization
   - With LASSO regularization, α=0.1.
   - With Ridge (Tikhonov regularization), α=0.1.

We fit different set of features to the three types of linear regression model, and then evaluate model performance using 10-fold cross-validation with coefficient of determination, $R^2$.

The results are shown in *Table 3*. Highest $R^2$ score is achieved by Linear Regression using Lasso regularization with polynomial features and normalized data, it achieved 79.35% accuracy on testset.

| Estimate Cocoa content scores | | | |
|---|---|---|---|
| **Features** | Linear Regression | Ridge Regression | Lasso Regression |
| 1, Original | 69.86% | 69.87% | 69.87% |
| 2, 1 + Normalized | 69.87% | 69.87% | 69.88% |
| 3, Polynomial | 72.59% | 74.10% | 76.28% |
| 4, 3 + Normalized | 72.60% | 77.48% | 79.35% |

TABLE 3
LINEAR REGRESSION MODELLING RESULTS

Another view on cocoa is its correlation with other nutrients, as show in *Figure 10*. An interesting observation from the figure is that cocoa content is negatively correlated with number of additives in the product, i.e. higher cocoa content, fewer number of additives. And cocoa content is strongly negatively correlated with sugars and carbohydrates with a -0.8 correlation.

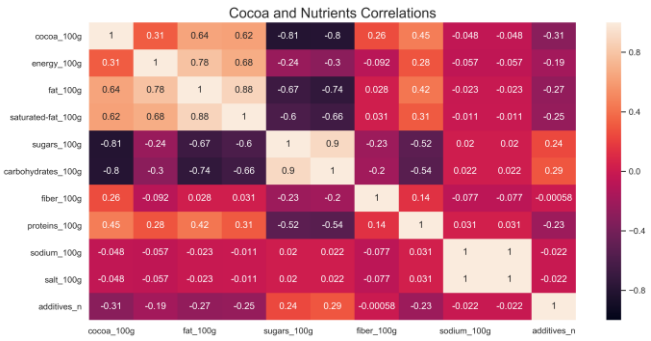


FIGURE 10
HEATMAP OF CORRELATION BETWEEN COCOA AND OTHER NUTRIENTS

### *B. Classify food products as healthy or unhealthy*

To build binary classifiers, firstly we subset on the main dataset to get the data having values on attribute 'nutrition_grade_fr'. The resulting dataset has 146,220 rows and 8 attributes, including main nutrients (i.e. 'energy_100g', 'fat_100g', 'sugars_100g', 'fiber_100g', 'proteins_100g', 'salt_100g'), 'code' and 'nutrition_grade_fr''.

Next, we create a new label 'is_healthy' having value 1 when 'nutrition_grade_fr' has 'a' or 'b', value 0 when the attribute has 'c', 'd', or 'e'. This transformation turns our dataset into imbalanced data, with 105,426 rows as '0' and 40,794 rows as '1'.

Before dataset splitting, since attribute 'energy_100g' has larger value ranges (kcal, ranging from 0 to a few thousands) than other attributes (gram, ranging from 0 to 100), we perform 'Data Normalization' before fitting dataset to model.

Train and test split are done by 90/10 ratio and randomly shuffled, this returned a trainset of 131,598 rows and a testset of 14,622 rows.

We fit three different classification models with different algorithm, including "Logistic Regression", "Nearest Neighbor" and "Decision Tree". Parameters for each mode is listed as below.

1. Logistic Regression:
   - With L2 penalty for regularization
   - Regularization strength C is set to 1.0
   - Tolerance level is set to 0.0001
   - Max number of iterations is set to 100
2. Nearest Neighbor:
   - Number of neighbors K is set to 5
   - Distance is calculated by Euclidean Distance
3. Decision Tree:
   - Use Gini index to split

The resulting accuracy of each algorithm is listed in *Table 4*, along with 'Training Time', 'Precision Score', 'Recall Score' and 'F1 Score'.

| Classification Accuracy | | | |
|---|---|---|---|
| **Features** | Logistic Regression | Nearest Neighbor | Decision Tree |
| Accuracy | 85.73% | 94.33% | 93.78% |
| Training Time | 456 ms | 14.7 s | 709 ms |
| Precision Score | 0.75 | 0.91 | 0.89 |
| Recall Score | 0.74 | 0.89 | 0.89 |
| F1 Score | 0.75 | 0.90 | 0.89 |

TABLE 4
CLASSIFICATION MODELLING RESULTS

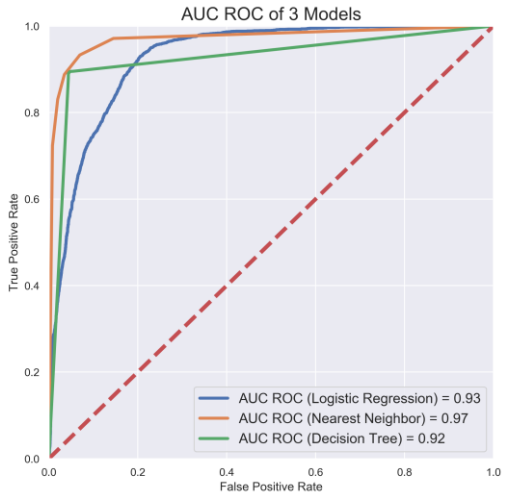


FIGURE 11
AUC ROC CURVES OF CLASSIFICATION MODELS

Intuitively, in this classification task, 'Precision' is how good the classifier is at NOT to label a product as healthy while the product is actually unhealthy, i.e. with precision, we want to minimize the number of false positive; 'Recall' is how good the classifier is at finding all the healthy products, i.e. with recall, we want to minimize the number of false negative.

From *Table 4*, we can note that Nearest Neighbor classifier has highest Precision Score, but Decision Tree has a better average between Precision and Recall. In terms of training time, Nearest Neighbor took significantly longer time, comparing to other 2 models. As for F1-Score, Nearest Neighbor and Decision Tree have almost similar score.

To better understand our models, we can plot AUC (Area Under Curve) ROC (Receiver Operating Characteristics) curve, shown in *Figure 11*. Sci-kit learn has built-in functions help us compute ROC scores for different thresholds on a model, and then we plot the scores on by False Positive Rate (FPR) and True Positive Rate (TPR).

After ROC curves are plotted, we calculate the area under ROC curves, with each model, the values are:

- Logistic Regression: 0.9276
- Nearest Neighbor: 0.9710
- Decision Tree: 0.9248

From both *Table 4* and *Figure 11*, we can conclude that the Nearest Neighbor classifier has the best performance in predicting the label of food products as 'Healthy' or 'Unhealthy'.

We can use our Nearest Neighbor classifier to classify the remaining dataset that don't have Nutri-Score ranks. After normalization and prediction, the returned classification has 103,227 healthy and 310,352 unhealthy products.

## IX. Extensions

### *A. Data on Alcoholic drinks*

One of the unanswered exploratory questions is about alcoholic drinks. The question was to explore the relationship between alcohol concentration and nutrient facts. However, almost all of the alcoholic drinks in dataset has missing values on the main nutrients.

This is due to the fact that producers of alcoholic drinks are not required by governments to label nutrient facts of their products. To perform this analysis, data about alcoholic drinks have to be gathered from other sources.

### *B. Classification into 5 classes*

In the modeling analysis, we have performed binary classification on food products. Another approach would be to classify all products into 5 categories.

### *C. Clustering and Comparison*

While Nutri-Score rank the food products by fixed formula, it may be interesting to compare the result of clustering algorithm (e.g. DBSCAN) with Nutri-Score ranks. As more new food products are added to the market, clustering algorithms may be more 'flexible' than fixed formula.

### *D. Combined analysis*

One of the possible extensions is to combine Open Food Facts data with business data, e.g. supermarket food product sales data for a period. We could combine the food facts with sales volume, to get an understanding of what consumers are buying, or what kind of marketing strategy (e.g. product label on packages) is meaningful.

## X. Conclusion

In this study, we explored the dataset from Open Food Facts, found some interesting correlations between nutrients and tried to verify some common beliefs about food products. We found that:

- Energy and Sugars are negatively correlated with health ranks, i.e. more of those, less healthy
- Protein and Salt have almost no correlation with health ranks
- Fiber has positive correlation with health rank 'a', but no or low correlation with other health ranks.
- More than 80% of sugary products are unhealthy
- Most chocolates contain 30% or 70% cocoa
- Cocoa in chocolates is negative correlated with number of additives added, i.e. higher cocoa content, fewer additives

- Fat and protein are positively correlated with cocoa, but Sugar is negatively correlated with cocoa
- Organic products are always healthy (in terms of health rank), but have better average rank than overall products
- Organic products, comparing to overall products, tend to have fewer salt and additives, but higher fibers in its content.

In the modelling analysis, we built models to estimate cocoa contents in chocolates, achieved highest $R^2$ at 79.35% with Linear Regressing using LASSO regularization. Also, we built classifiers to classify food products as healthy or unhealthy, achieved 94.33% accuracy using Nearest Neighbor algorithm.

## REFERENCES


[1] Chung-Tung Jordan Lin, Yuanting Zhang, Ewa D. Carlton, Serena C. Lo, *"2014 FDA Health and Diet Survey"* 7-8, 33-34

[2] Open Food Facts, https://world.openfoodfacts.org/

[3] Wikipedia, *"Open Food Facts"*, https://en.wikipedia.org/wiki/Open_Food_Facts

[4] Chantal Julia, Serge Hercberg, "*Nutri-Score: Evidence of the effectiveness of the French front-of-pack nutrition label*"

[5] Michel Chauliac, Dr, "NUTRI-SCORE : NUTRITION LABELLING SCHEME" 4

[6] Open Food Facts data, https://world.openfoodfacts.org/data

[7] *"Open Database License"*, https://opendatacommons.org/licenses/odbl/1.0/

[8] Wikipedia, *"Organic Product"*, https://en.wikipedia.org/wiki/Organic_product

[9] Scikit-Learn, "PolynomialFeatures", https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html