<Title of the Project>Food Grades and Categories in the Open World Food Facts Dataset

# Introduction

First, provide the context of the problem and then state the problem (your main research question). Second, write briefly that what are you proposing to solve this problem (don’t write details of the solution here).

The research problem and question should be clearly specified here

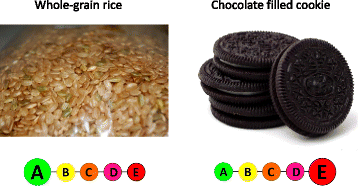
# Introduction

Provide the context of the problem and then state the problem (your main research question). Second, write briefly that what are you proposing to solve this problem (don’t write details of the solution here). (You can use part of your abstract here)

This data for the study comes from the [Open Food Facts database](http://world.openfoodfacts.org/data), a free, open, and collaborative database of food products around the world.

It contains a nutrition grade for every entry, graded from A to E decreasing as exemplified in the image below.

**“Discriminating nutritional quality of foods using the 5-Color nutrition label in the French food market: consistency with nutritional recommendations” [1]** is a study that assesses the performance of the 5 color grade label and concludes that the 5 color grade label displays a high performance in assessing nutritional quality of foods across the dataset.



(Image from http://nutritionj.biomedcentral.com/articles/10.1186/s12937-0)

**Problem 1**: I will try to predict the food grade of an entry based on nutrition info provided and main food category of that entry.

**Problem 2**: I will try to predict the food category of an entry in the dataset based on the nutrition info and nutrition grade of that entry.

**Problem 3**: I will give a visual of the correlation between nutrition information and a chart of the main categories of food produced/consumed in different geographic regions.

# Literature Review

Write summary of the related papers that you reviewed here. Write the summary in your own words—don’t use the technical jargon from the paper that you don’t understand. Keep this section short—a short paragraph or few sentences about each paper you reviewed should be sufficient.

# Dataset

Give the description of the dataset that you are using along with the individual attributes you will or will not use in your analysis. Also mention the source of the dataset (where did you get it from). In case the data is curated and created by you please explain the details.

The data for the study comes from Kaggle’s [Open Food Facts database](http://world.openfoodfacts.org/data), an open database of food products around the world.

It contains a nutrition grade for every entry, graded from A to E decreasing with an associated color code:

Nutrition grade:

[](http://in.openfoodfacts.org/contributor/openfoodfacts-contributors/nutrition-grade/c)

The open foods dataset classifies its entries in 10 big categories:

* Beverages
* Cereals and potatoes
* Fruits and Vegetables
* Composite Foods
* Fats and sauces
* Fish Meat Eggs
* Milk and Dairy products
* Salty Snacks
* Sugary Snacks

Each big category in its turn is divided in sub categories – for example *Composite foods* are divided in *One dish meal*, *Pizza and quiche*, and *Sandwich* subcategories.

Each entry in the dataset contains detailed nutrition info, the ones that I retained for my study are (all per 100 grams):

* additive number
* fiber, protein
* carbohydrates
* sugar
* fat
* saturated fat
* salt
* sodium content
* energy

Also retained for my analysis:

* main food category: pnns\_groups\_1
* subcategory pnns\_groups\_2
* country of origin
* nutrition grade (nutrition\_grade\_fr)

( ex: pnns\_group\_1 = Beverages and pnns\_group\_2 = Fruit Nectars)

I have removed the other features in the dataset (open world foods dataset has close to 200 attributes) describing packages, stores where to buy, brands, minerals , vitamins and amino acids for either lacking data or not relevant to my analysis.

A full description of the dataset is available at: [www.kaggle.com/openfoodfacts/world-food-facts](http://www.kaggle.com/openfoodfacts/world-food-facts)

# Approach

Create a block diagram for the steps of your approach to clearly provide an overview. For example, if you first scrapped twitter, second applied NLP techniques to extract keywords, third labelled the tweets as positive and negative using a set of keywords, and fourth build a classifier, then you should create a box for each of the steps with arrows connecting one step to the next one. A sample block diagram is shown below.

Visuals

Once this is done, explain each of the steps in detail. What are you planning to do in each step or have already done. For example, in the above case you would create subheadings for each of the steps.

## Step 1: <Data Preprocessing>

The data set is available for download as an excel spreadsheet.

It has around 200 attributes per record. I have kept 12 relevant attributes stated in dataset description.

* After removing irrelevant columns I have imported the data and replaced missing values.
* For nutrition info missing value was replaced with the mean for that particular food category.
* Missing food grades were replaced with the most common value for that food category.
* Missing food categories are listed in the dataset as unknown. I have included under unknown the entries with blank food categories and subcategories. For the cases where subcategory was available the main category was inferred from subcategory (ex: Tripe Products was included under Fish Meat Eggs)

Link to R file **– foodfacts.r** at:

<https://github.com/Darianne/fact-foods---ryerson-capstone>

## Step 2: <Logistic regression for classification>

Write details of the step 2. If there is any source code that you’d like to share then provide the link of the Github.

**Problem 1**: predict the food grade of an entry based on nutrition info provided and main food category of that entry.

* Formula is: food grade in relation with nutrition components and food categories

nutrition\_grade\_fr ~ saturated\_fat\_100g + sugars\_100g + fat\_100g + additives\_n + salt\_100g +

fiber\_100g + pnns\_groups\_1 + carbohydrates\_100g + energy\_100g +

sodium\_100g + proteins\_100g

* Model for logistic regresion – use multinom function in nnet package: multinom(formula, data)
* Split in test and training, Fit model on training data: multinom (formula,data=train)
* Predict nutrition grade on test and check performance with confusion matrix and statistics

predict(model, type="class", newdata=test)

**Problem 2**: predict the food category of an entry in the dataset based on the nutrition info and nutrition grade of that entry

Repeat steps in Problem 1 but with formula changed to predict food categories:

pnns\_groups\_1 ~ saturated\_fat\_100g + sugars\_100g + fat\_100g + additives\_n + salt\_100g + fiber\_100g

+ carbohydrates\_100g + energy\_100g + sodium\_100g + proteins\_100g +

+ nutrition\_grade\_fr

Link to R file **– foodfacts.r** at:

<https://github.com/Darianne/fact-foods---ryerson-capstone>

## Step 3: <Test performance of the model and explain results>

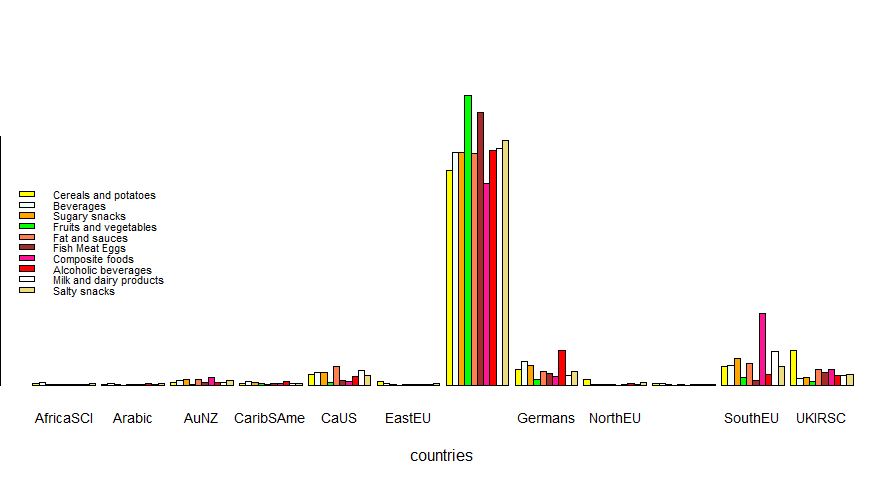
Write details of the step N. If there is any source code that you’d like to share then provide the link of the Github.

* Create confusion matrix for both models/formula and analyze statistics. (see Results)
* Describe where the the most errors occur and come up with a metric for errors between classes

Link to R file **– foodfacts.r** at:

<https://github.com/Darianne/fact-foods---ryerson-capstone>

## Step 4: Visuals



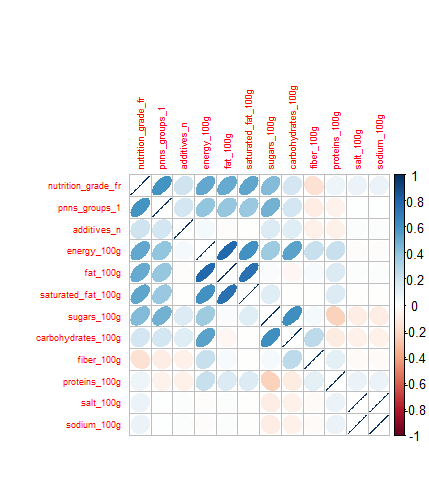
Barplot of foods produced in geographical regions is consistent with the fact that 32000 entries out of 43.000 entries in the dataset are from France alone.

But I can see that the nutrition pattern in France/Belgium/Netherlands/Luxembourg consists of:

1. Fruits and Vegetables
2. Meat Fish Eggs
3. Salty Snacks.

Or that Canada/US #1 category is Fats and Sauces (subcategories: Dips and Dressings, Fats) .

Southern Europe (Italy, Greece, Portugal, Spain) cuisine # 1 category is Composite Foods (subcategories: One dish meal, Pizza and Quiche, Sandwich)



Correlation between nutritional facts shows that:

* Sodium/salt are independent of the other ingredients – but confirms they are strongly correlated with each other
* Protein and sugar do not agree (inverse correlation)
* Energy is given first of all by fat and then by sugars and carbohydrates
* Nutrition grade is strongly correlated with the main food category

# Results

Explain your results here. Consider that you need to communicate your results to executives in an organization. For example:

1. Insert tables and/or charts showing the results
2. Write description of the tables and charts, such that they show the usefulness for an organization
3. Identify the evaluation measures, such as accuracy, precision, recall, etc.

**Problem 1**: predict the food grade of an entry based on nutrition info provided and main food category of that entry.

#Confusion Matrix and Statistics

# Overall Statistics

**# Accuracy : 0.6954**

# 95% CI : (0.6863, 0.7045)

# No Information Rate : 0.261

# P-Value [Acc > NIR] : < 2.2e-16

#

# Kappa : 0.6153

# Mcnemar's Test P-Value : < 2.2e-16

#

# Statistics by Class:

#

# Class: a Class: b Class: c Class: d Class: e

# Sensitivity 0.8337 0.58533 0.6499 0.7037 0.6951

# Specificity 0.9596 0.92683 0.8702 0.9022 0.9549

# Pos Pred Value 0.8229 0.59538 0.5975 0.7176 0.7610

# Neg Pred Value 0.9625 0.92396 0.8935 0.8961 0.9381

# Prevalence 0.1837 0.15536 0.2287 0.2610 0.1713

# Detection Rate 0.1531 0.09094 0.1486 0.1837 0.1191

# Detection Prevalence 0.1861 0.15274 0.2487 0.2560 0.1565

# Balanced Accuracy 0.8967 0.75608 0.7601 0.8030 0.8250

Confusion Matrix:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Prediction Reference | a | b | c | d | e |
| a | 1519 | 278 | 34 | 8 | 7 |
| b | 263 | 902 | 331 | 15 | 4 |
| c | 33 | 290 | 1474 | 506 | 164 |
| d | 1 | 3 | 370 | 1822 | 343 |
| e | 6 | 68 | 59 | 238 | 1181 |

Confusion matrix shows that most of the errors occur between “neighboring” food grades. (ex A predicted as B = 278). Error at a “distance” more than one food grade are less. (ex: D predicted as A = 1)

**Problem**: What is the accuracy if we allow confusion between neighboring classes? 95% !

**#2619 1 class difference error**

E1<-conf[1,2]+ conf[2,1]+ conf[2,3]+ conf[3,2]+ conf[3,4]+ conf[4,3]+ conf[5,4]+conf[4,5]

**#308 2 class difference error**

E2<-conf[1,3]+conf[3,1]+conf[2,4]+conf[4,2]+conf[3,5]+conf[5,3]

**#81 3 class difference error**

E3<-conf[1,4]+conf[4,1]+conf[2,5]+conf[5,2]

**#13 4 class difference error**

E4<-conf[1,5]+conf[5,1]

ALL TRUE<-conf[1,1]+ conf[2,2]+ conf[3,3]+ conf[4,4]+ conf[5,5] #6898 (diagonal)

ALL<-ALL TRUE+E1+E2+E3+E4 #9919

**ACCURACY**<-ALL TRUE/ALL  **#0.69543307** (Same as: see Overall Statistics above)

**Calculate Accuracy if we allow confusion between adjacent nutrition grades**

ALL TRUE 1 <- ALL TRUE + E1 #9517

**ACCURACY** **1**<- ALL TRUE 1 /ALL **#0.9594717**

**Problem 2**: predict the food category of an entry in the dataset based on the nutrition info and nutrition grade of that entry

# Confusion Matrix and Statistics = with correctly predicted shown in bold

# Reference

# Prediction Beverages Cereals and potatoes Composite foods Fat and sauces Fish Meat Eggs Fruits and vegetables

# Beverages **702**  2 4 12 0 15

# Cereals and potatoes 2 **842** 36 1 2 36

# Composite foods 16 107 **673** 39 63 90

# Fat and sauces 5 4 22 **479**  8 26

# Fish Meat Eggs 9 39 44 10 **960** 5

# Fruits and vegetables 47 287 103 28 55 **1018**

# Milk and dairy products 74 7 151 88 117 3

# Salty snacks 2 39 9 6 0 1

# Sugary snacks 67 96 1 5 141 21

# Reference

# Prediction Milk and dairy products Salty snacks Sugary snacks

# Beverages 76 0 92

# Cereals and potatoes 2 27 41

# Composite foods 12 11 44

# Fat and sauces 5 5 2

# Fish Meat Eggs 169 29 13

# Fruits and vegetables 84 4 7

# Milk and dairy products **1236** 90 31

# Salty snacks 3 **324**  21

# Sugary snacks 25 28 **2050**

#

# Overall Statistics

#

# **Accuracy : 0.7497**

# 95% CI : (0.7415, 0.7577)

# No Information Rate : 0.2082

# P-Value [Acc > NIR] : < 2.2e-16

#

# Kappa : 0.7116

# Mcnemar's Test P-Value : < 2.2e-16

#

# Statistics by Class:

#

# Class: Beverages Class: Cereals and potatoes Class: Composite foods Class: Fat and sauces Class: Fish Meat Eggs

# Sensitivity 0.75974 0.5917 0.64525 0.71707 0.71322

# Specificity 0.98015 0.9847 0.96183 0.99258 0.96723

# Pos Pred Value 0.77741 0.8514 0.63791 0.86151 0.75117

# Neg Pred Value 0.97812 0.9423 0.96298 0.98199 0.96050

# Prevalence 0.08362 0.1288 0.09439 0.06045 0.12181

# Detection Rate 0.06353 0.0762 0.06090 0.04335 0.08688

# Detection Prevalence 0.08172 0.0895 0.09548 0.05032 0.11566

# Balanced Accuracy 0.86995 0.7882 0.80354 0.85482 0.84023

# Class: Fruits and vegetables Class: Milk and dairy products Class: Salty snacks Class: Sugary snacks

# Sensitivity 0.83786 0.7667 0.62548 0.8909

# Specificity 0.93747 0.9406 0.99231 0.9561

# Pos Pred Value 0.62339 0.6878 0.80000 0.8422

# Neg Pred Value 0.97908 0.9594 0.98178 0.9709

# Prevalence 0.10995 0.1459 0.04688 0.2082

# Detection Rate 0.09213 0.1119 0.02932 0.1855

# Detection Prevalence 0.14778 0.1626 0.03665 0.2203

# Balanced Accuracy 0.88766 0.8537 0.80890 0.9235

# Conclusions

Give a short summary (one to two paragraphs) of your analysis and conclude the discussion by defining the usefulness of your analysis.

Based on confusion matrices and overall statistics, food category and nutrition grade can be predicted with good accuracy based on key nutritional ingredients.

Accuracy increases significantly if we are tolerant to errors between adjacent food grades.

Analysis of food categories and geographical regions are not very useful because from 43000 entries, 32000 are from France alone. But it offers an image of “star food products” inside a geographical region

**Notes:**

**[1]** http://nutritionj.biomedcentral.com/articles/10.1186/s12937-0