Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Statistisk sentralbyrå
Statistics Norway

Estimating missing nutritional values with Natural Language Processing and Machine Learning

Annabelle Redelmeier

Norwegian Computing Center

August 23rd, 2022

NSM 2022

Background

SSB publishes statistics on the diets of Norwegians

- What does a typical family consume per month?
- How much is spent on processed foods?
- Are certain regions healthier than others?
- How are diets changing over time?

vet du at .no

NSM 2022

Data

- 117,000 unique products
- 11 interesting nutritional values (protein, salt, carbohydrates, fat…)

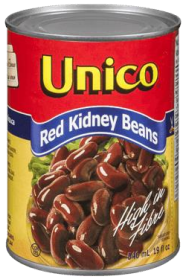
	Product	Product Code	Energy (kJ)/100gr	Protein/100gr	Fat/100gr	Carbohydrates/100gr	Fiber/100gr
	Red kidney beans	10789345	448 kJ	8.8 g	0.7 g	12.7 g	7 g
	Banana, Dole	78145391	354 kJ	0.5 g	0.3 g	18.3 g	2 g
	Chicken, drumstick	78978899	623 kJ	17.2 g	9 g	0 g	0 g

- Nutritional values are *optionally* provided to vetduat.no by producers…

- 1.1 million = 82% values missing!

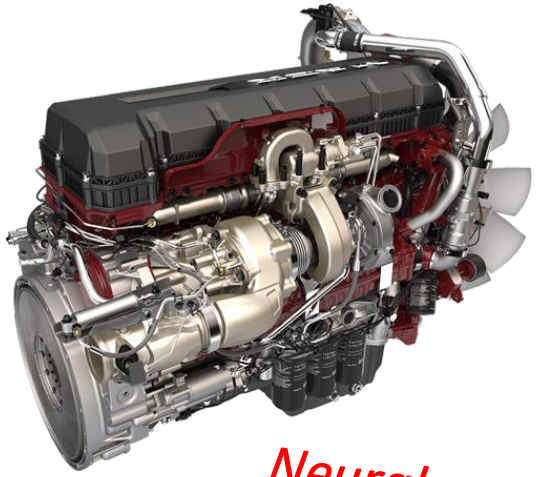
Frame as a supervised task

Input



Features?

Prediction model



Support vector machine

Linear regression

Neural network

Decision tree

Output

Carbohydrates per 100 grams

Protein per 100 grams

Energy content per 100 grams

Fat content per 100 grams

Salt content per 100 grams

Features?

Product name

COICOP = food classification

	COICOP group name	Percent
	Meat, fresh, chilled or frozen	22%
	Bread and bakery products	11%
	Meat, dried, salted, in brine or smoked	8%
	Ready-made food	5%
	Cheese	3%
	Fish, dried, salted, in brine or smoked	4%

70 groups

Important observations

1. The distribution of missing values is **not uniform**.
2. Grocery stores tend to stock several *types* of the same product.

Features?

Input

Product name

COICOP = food classification

Match

Carbohydrates

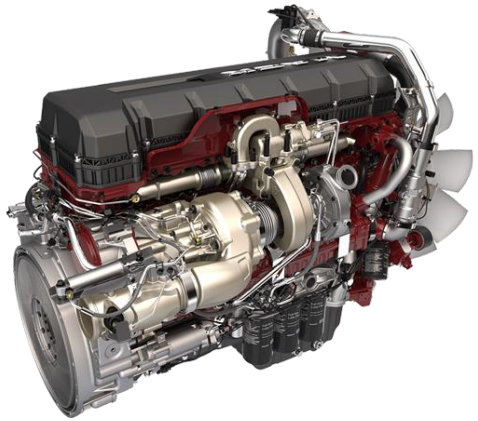
Protein

Energy

Fat

Salt

Prediction model



Output

Carbohydrates per 100 grams

Protein per 100 grams

Energy content per 100 grams

Fat content per 100 grams

Salt content per 100 grams

Matching: 2 steps

Create a set of products with maximal similarity.

Step 1: Calculate pair-wise similarity measure (**Jaccard similarity**) between all pairs of products:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

A = {Unico, Red, Kidney, Beans}

B = {Bush's, Best, Organic, Baked, Beans}

Matching: 2 steps

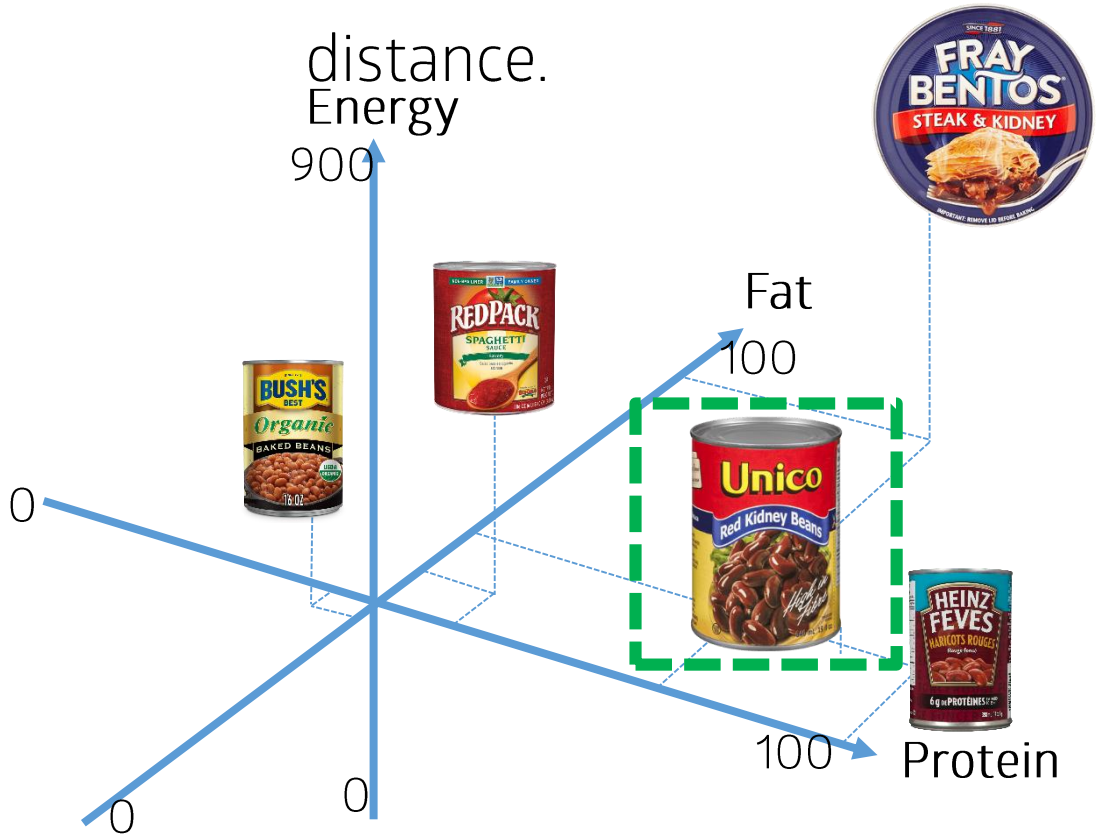
Step 2: Narrow matches down using “nearest neighbor” approach.

Calculate the Euclidean distance between the nutritional values:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

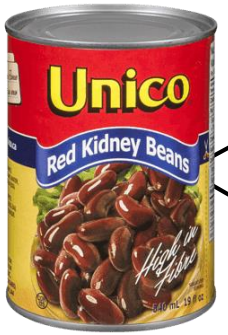
p: product
q: *potential match*

Define the best match as the product with the smallest Euclidean distance.



Prediction model

Input

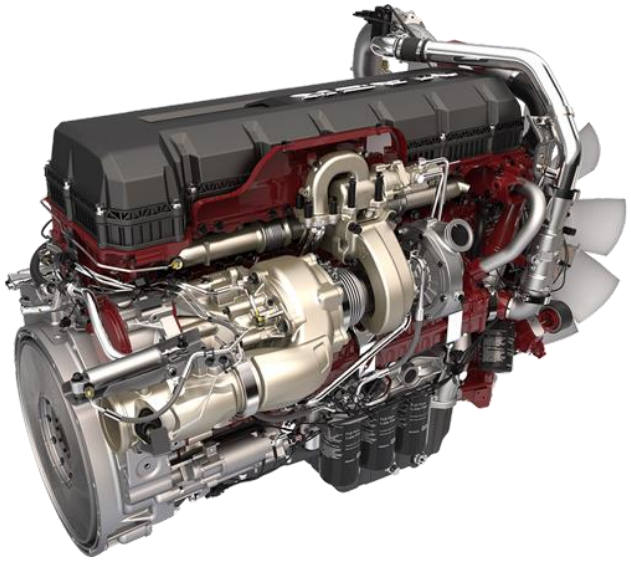


Match from SSB

Protein
Carboh
Jaccard
Euclidean

Protein
Carboh
Jaccard
Euclidean

Match from Food Composition Table



Random forest

Split 70%/30% training/test.

Output

- Protein content per 100 grams
- Carbohydrates per 100 grams
- Sugar content per 100 grams
- Fat content per 100 grams
- Salt content per 100 grams

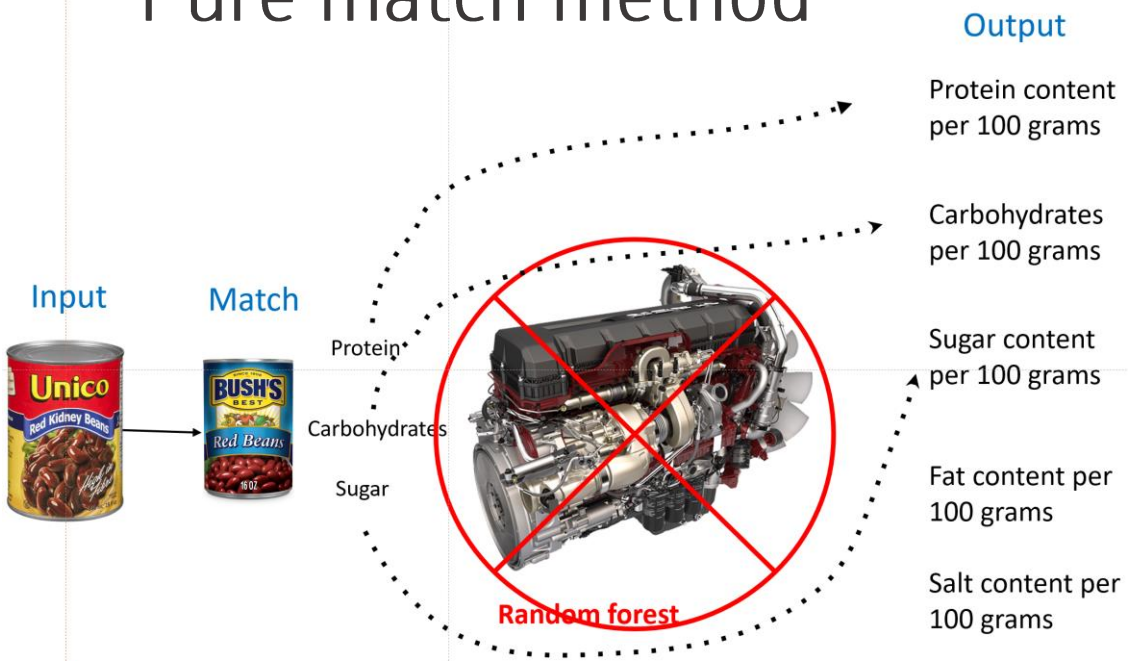
Comparison with two other methods

Is this really the best way to estimate the nutritional values?

Mean imputation by COICOP group

Replace missing value using the average energy/protein/sugar in the same COICOP group

Pure match method



Results

$$\mathrm{RMSE} = \sqrt{\frac{1}{n_{test}}\sum_{i=1}^{n_{\mathrm{test}}}(\mathrm{truth}_i - \mathrm{prediction}_i)^2}$$

$$\mathrm{MAE} = \frac{1}{n_{test}}\sum_{i=1}^{n_{\mathrm{test}}}|\mathrm{truth}_i - \mathrm{prediction}_i|$$

Nutrition	RMSE			MAE		
	Mean imputation	Pure Match	Random Forest	Mean imputation	Pure Match	Random Forest
Energy (kcal)/100g	101.24	71.51	**52.38**	67.16	26.92	**21.18**
Energy (kJ)/100g	421.82	297.12	**217.66**	279.54	111.90	**87.79**
Monounsaturated fat/100g	5.44	4.06	**3.48**	2.37	1.23	**0.85**
Polyunsaturated fat/100g	2.83	1.93	**1.71**	1.02	0.47	**0.45**
Saturated fat/100g	4.37	2.70	**2.48**	2.28	0.89	**0.87**
Total fat/100g	9.72	5.71	**4.88**	6.02	2.10	**1.98**
Fiber/100g	5.47	3.48	**2.98**	2.37	1.06	**1.05**
Carbohydrates/100g	13.21	8.96	**6.94**	7.85	3.26	**2.90**
Protein/100g	5.12	3.35	**3.03**	2.85	1.26	**1.20**
Salt/100g	5.98	**2.62**	2.70	1.51	**0.44**	0.57
Sugar/100g	11.14	6.01	**5.18**	5.76	**1.90**	1.91

Last observations

1. Generalization: More data sets?
2. Prediction approach versus pure match approach: Other features
3. Extension of Jaccard similarity:

$$\frac{|A \cap B|}{|A \cup B|} = \frac{|c_1, c_2|}{|c_1, c_2, c_3, c_4|} = \frac{1+1}{1+1+1+1} = 0.5 \rightarrow \frac{1+0.5+1}{1+0.5+1+1+1} = 0.55$$

"**nuts** loose weight"
"chili **fudge**"
"**carrots** bagged"
"**peas** yellow"

4. K-Nearest-Neighbour: Used for more than narrowing matches down?