Undernourishment: a look through a machine learning perspective

Introduction

Undernourishment, especially in children and mothers, is a leading risk factor for death and other health consequences. Having a diet which is both sufficient in terms of energy (caloric) requirements and diverse to meet additional nutritional needs is essential for good health (Roser & Ritchie, 2019).

The prevalence of undernourishment, as a share of the population, is the main hunger indicator used by the UN's Food and Agriculture Organization (**FAO**). It measures the share of the population which has a caloric (dietary energy) intake which is insufficient to meet the minimum energy requirements defined as necessary for a given population (Figure 1).

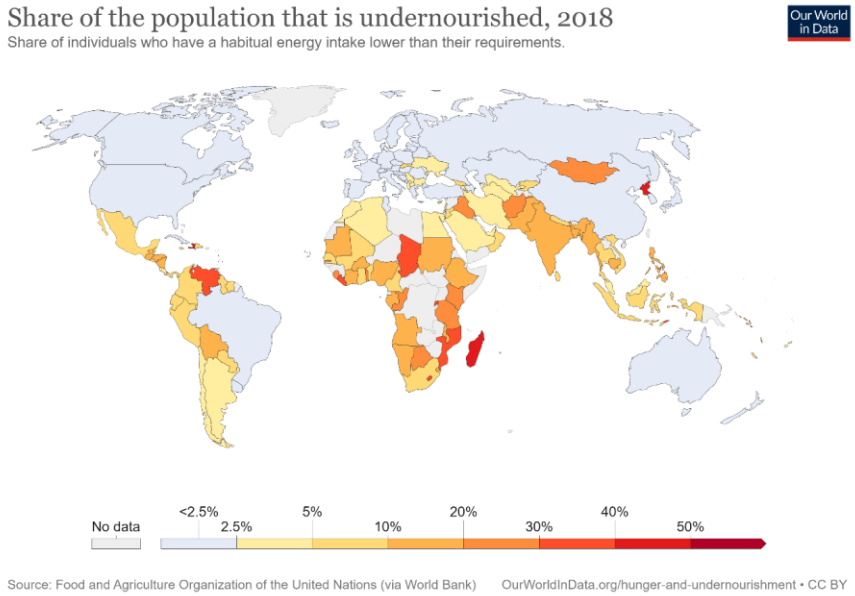

Figure 1: Share of the population that is undernourished (2018). Countries with a prevalence below 2.5% are not shown.

This concept is strongly associated with **malnutrition**: a lack of nutrients that occurs when an individual gets too few or too many nutrients, resulting in health problems. Specifically, it is "a deficiency, excess, or imbalance of energy, protein and other nutrients" which adversely affects the body's tissues and form (Hickson & Smith, 2018). Malnutrition refers then to an unbalanced diet - including excessive eating - whereas the term **undernutrition** refers more specifically to a deficiency of nutrients.

This dietary imbalance is determined by the type of macronutrients that form part of the population's diet. These are: glucids (also known as carbohydrates), proteins and fats. A good diet is therefore a good balance between these three components. According to the World Health Organization, the ideal is to eat between 45%-55% of carbohydrates, 25%-35% of fats, and 15%-20% of proteins.

Commented [RM1]: https://www.fao.org/sustainable-development-goals/indicators/2.1.1/en/

https://www.fao.org/hunger/en/#:~:text=What%20is%20food%20insecurity%3F,of%20resources%20to%20obtain%20food.

Food Security Indicators.

The indicator we use during this work is the **prevalence of undernourishment** (PoU), an estimate of the proportion of the population whose habitual food consumption is insufficient to provide the dietary energy levels that are needed to maintain a normal active and healthy life (Source: FAO). It is expressed as a percentage.

Problem

Considering the points developed above, the problem we face is to find **a good predictor** of undernourishment based on the available indicators.

These indicators are, therefore:

- Share of dietary energy supply derived from **cereals, roots and tubers** (*kcal/cap/day*) (3-year average)
- Average **protein** supply (*g/cap/day*) (3-year average)
- Average **fat** supply (*g/cap/day*) (3-year average)

Commented [RM2]: se repite en Data, ver dónde queda mejor

Based on these, different methods, described below, will be used to try to predict undernourishment. An important point to mention is that all these indicators are numerical.

Methods

Commented [RM3]: Main method – Regression. - kernel ridge, svr(linear, poly, rbf). Kernel ridge and svr(rbf) are the ones that work.

Indicators – C, epsilon, alpha

Secondaty method – Clustering

?? Neural Net?

We start first with **linear regression**, which is nothing more than a linear model with coefficients to minimize the residual sum of squares between the observed targets in the dataset (carbs, protein, and fat), and the targets predicted (undernourishment) by the linear approximation.

Secondly, we tested with a **Support Vector Regression** (SVR). It is a *non-linear function* by employing the *kernel trick* (that allows us to operate in the original feature space without computing the coordinates of the data in a higher dimensional space). Thus, while linear regression model minimizes the error between the actual and predicted value through the line of best fit, SVR manages to **fit the best line** within threshold of values.

Third, we use the **Kernel Ridge** method. This is also a *non-linear function,* but it is typically faster for medium-sized datasets (> than 1000 samples, as the one we are using) compared to SVR, that scales better on larger datasets. On the other hand, the learned model is non-sparse and thus slower than SVR at prediction-time.

The fourth method evaluated was that of **neural networks**. Due to the nature of our database, the specific method that we decided to use was the **Multi-layer Perceptron regressor** (MLP Regressor). It utilizes a supervised learning technique called *backpropagation* for training (calculates the gradient of the error function with respect to the neural network's weights). We decided to test this method since correlation among variables does not affect the method performance compared to the previous ones.

K-means clustering – The main problem to solve in unsupervised clustering comes from taking different observations and grouping them in clusters. The data points in the clusters should be as different as possible between clusters and as similar as possible inside the

cluster. K-means clustering recursively creates k-number of centroids and calculates the distances between the different data points. Normally three criteria are used to stop the recursion. No change in the centroid position, no change of clusters within the points or a max set of iterations was reached.

Dataset

The dataset we worked with during our study is structured as follows:

As **inputs**, we have:

- **Carbs:** Share of dietary energy supply derived from cereals, roots and tubers (*kcal/cap/day*) (3-year average)
- **Proteins:** Average protein supply (*g/cap/day*) (3-year average)
- **Fats:** Average fat supply (*g/cap/day*) (3-year average)

As **output**, we have:

- **Undernourishment**: Prevalence of undernourishment (percent) (3-year average)

As an overview, we show on the Table 1 the basic statistics of what the database looks like:

	Carbs	Proteins	Fats	Undernourishment
Count	2699			
mean	46,94	79,68	85,29	10,75
std	14,34	20,29	35,69	10,82
min	12	23,2	17,3	2,5*
percentile 50%	46	79,7	80,7	6,1
max	83	143,3	170,3	67,5

Table 1: count of data, mean, standard deviation, minimum, maximum and percentile 50% values
*Countries with the label "undernourishment < 2,5%" where grouped in this category

For the study, we took data from all countries in the world where FAO has coverage (160 in total). For this reason, the dataset also includes their names, as well as the average of the three consecutive years from which both inputs and outputs were obtained, starting from 2000 to 2018.

This dataset was taken from official data published by FAO (Dataset, FAO, 2018). It should be noted that, given the composition of the dataset as described above, it is supervised data, since it relies on labelled input and output training data.

The aggregation of the data considers one row or data point per country and year. For example, Ivory Coast 2001 and Ivory Coast 2002 are two separate data points (Table 2):

Area	Year Code	carbs	proteins	fats	undernourishment
Czechia	20142016	31.0	87.3	130.0	2.5
Czechia	20152017	29.0	87.0	134.0	2.5
Czechia	20162018	28.0	86.7	140.7	2.5

Ivory Coast	20002002	65.0	51.0	54.0	20.4
Ivory Coast	20012003	64.0	52.0	55.0	21.3
Ivory Coast	20022004	64.0	53.0	55.3	21.2
Ivory Coast	20032005	65.0	54.0	55.7	20.6
Ivory Coast	20042006	65.0	54.7	55.0	20.2

Table 2: example of a portion of the database

Experiment

The first part of the work was to **obtain** and **clean** the data. As mentioned before, the data were downloaded from FAO, specifying the indicators we needed, as well as the countries and years available.

The raw data had to be **grouped by country and year**, and then we went ahead to **clean** the data to drop samples with no information (where values were equal to zero).

Once the dataset is clean and ordered, we defined the groups of **predictors** and **targets**:

- **Predictors**: carbohydrates, proteins, and fats.

- **Target**: undernourishment

One of the first tasks was to make a **correlation matrix**, to get an idea of the intercollinearity between the variables we worked with.

Feature selection was applied to the data set using **Select Best K** to select the two-most significant variables to reduce the overall complexity of the problem and create a simpler model. As a second alternative, we selected only one variable. Since the model to be deployed is a regression, we used the f-regression selector as our feature selection parameter.

Standardization was applied to the dataset to fit the regressions models better.

We decided to apply to the data a **k-means clustering** on the data, and then the methods described above to each group separately, in search of better results.

At this moment, we already have the clean data standardized and the predictor and target variables defined, so we apply different models to six different combinations:

- The **three** predictor variables together (*carbs, proteins, and fats*)

- The **two** most significant variables according to K Best (*proteins* and *fats*)

- The **two** variables that have the least intercollinearity (*proteins* and *carbs*)

- The remaining **two** variables (*fats* and *carbs*)

- The **most significant** variable according to K Best (*proteins*)

- To the **clusters** that resulted from the k-means

As a first model, we applied a **simple regression** to get a first approximation of the possible outputs, using the sklearn package.

Secondly, we continued with a **Grid Search** to find the best possible combination of parameters for them. These parameters were C value and gamma value, which apply to the **SVR methods** (*rbf, linear* and *polynomial*)**,** whereas alpha and gamma were for **Kernel Ridge** method. Once the best parameters were determined, the next step was to train the methods on the different combinations of datasets mentioned above.

Finally, we decided to apply **neural networks** with **MLP Regressor** using a **Grid Search** as well, searching for the best hidden layers parameters. Then, we set a sample size of 20%, a maximum of 5000 iterations, and with the solver "*adam*", since we found it was the best fit considering our data.

Results and discussion

The first result observed was the **correlation matrix** (Figure 2). As we can see in the matrix, there is a high degree of collinearity between the variables, mainly between fats and carbohydrates. On the contrary, proteins and carbohydrates present the lowest correlation value.

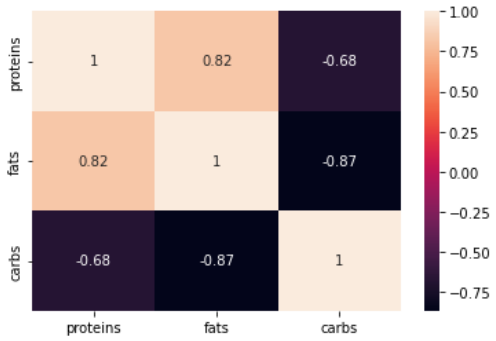

Figure 2: Correlation matrix

Secondly, and as a first model, we got the results from the **lineal regression**. The score was **0.558** for the coefficient of determination (R^2) for the best variable (*proteins*), which in comparison with the models we will see below, is a low value.

We decided to perform a **grid search** for the different regression models in order to find the regression model that best fits the data and the hyperparameters inside that regression model that will fit our data the best (Figure 3). As we mentioned before, these parameters were C, gamma for SVR (left) and alpha and gamma for Kernel Ridge (right):

Commented [RM4]: Usando 1 variable (prot)
-rbf = 0,69
- Kernel = 0,69
- NN = 0.73

Usando 2 variables (fat y prot)
-rbf = 0,71
- KR = 0,70
- NN = 0,78

Usando 2 variables (carbs y prot)
-rbf = 0,70
- KR = 0,70
- NN = 0,72

Usando 2 variables (fat y carbs)
-rbf = 0,53
- KR = 0,56
- NN = 0,60

Usando 3 variables
-rbf = 0,72
- KR = 0,73
- NN = 0.83

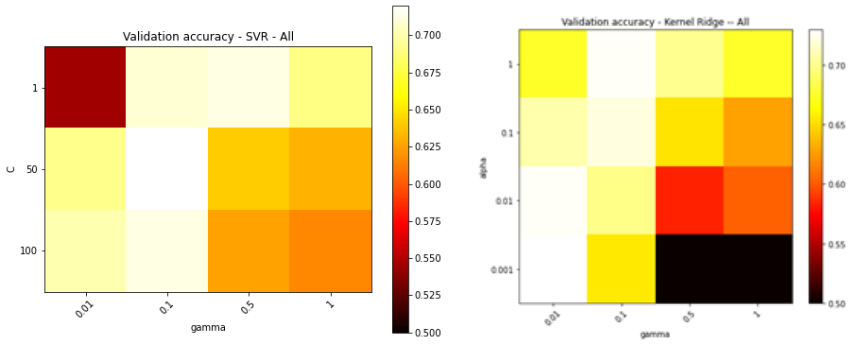

Figure 3: Grid search for the SVR parameters (left) and KR parameters (right)

As we can see, the most suitable parameters for SVR were a **C of 50** with a **gamma of 0.1**, while for Kernel Ridge it was an **alpha of 1** and a **gamma of 0.1**.

Based on the parametric results, we set out to train the different methods described above for different combinations of different variables. Figure 4 shows a graphical representation of them:

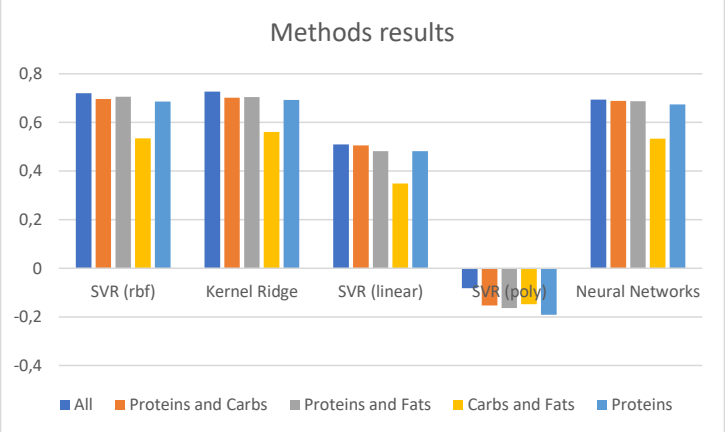

Figure 4: Results achieved for the different methods and combination of variables

As we notice from the figure above, the SVR (*rbf*), Kernel Ridge and Neural Networks methods were the ones that stood out, with a tiny difference in favor of **Kernel Ridge**, which obtained the highest scores overall.

On the other hand, when the ***Protein*** variable was analyzed separately, the results in all the models had almost the same value as when all the variables were analyzed together, thus denoting its high contribution to the model. Likewise, when it was removed from the analysis and *Carbohydrates* and *Fats* were used together, the scores of all models dropped significantly, showing the preponderance of *Protein* for the models.

Finally, it is worth noting the low values obtained when performing the **polynomial SVR**, which are extremely low in comparison with the others, thus explaining the unsuitability of the model for this dataset. For more information on the results, a table with the outcomes is attached in the appendix (Table 3).

As a last step, three **clusters** were used based on the scree-plot validation done to decide K, as shown in the Figure 5:

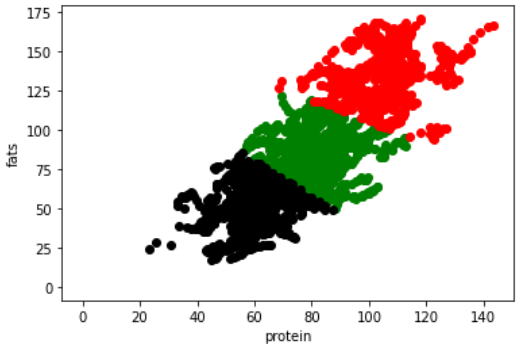

Figure 5: Clusters done with KMeans

After applying the regressions to every individual cluster, we found that the score of the models was lower (do not reach more than 0.45) and decided to keep the regressions done on the unclustered data (Figure 6). As before, a table with the values collected in detail is attached (Table 4).

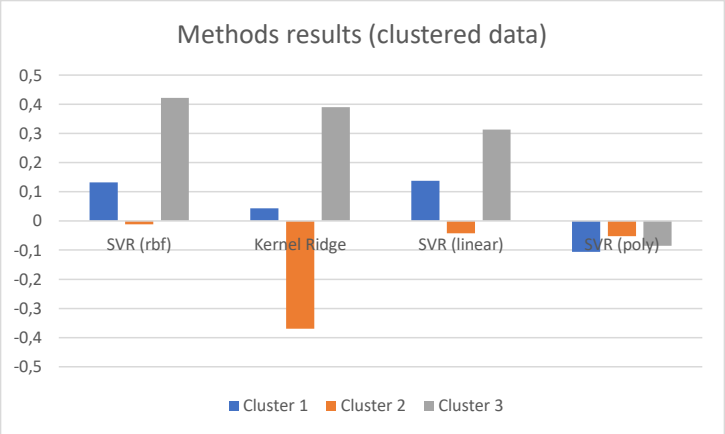

Figure 6: Results achieved for the different methods and clusters

Conclusions

Based on the above results, we can conclude the following statements:

- Average protein supply provides sufficient information to the model to predict undernourishment with 69% reliability.
- Due to the high correlation between the variables (proteins, carbohydrates and fats), the scores that best fit the data (SVR (rbf) and Kernel Ridge) don't differ drastically.
- Clustering does not help obtain a better fit of the model, probably due to the low amount of data.
- Neural Networks, SVR(rbf) and Kernel Ridge are the models that best predict the data. Kernel Ridge being the most effective one overall. The best fit for Kernel Ridge is done when low gamma and low alpha values are used.

Appendix

Table 3: Results of the different methods on the different datasets:

Method	All	Proteins and Carbs	Proteins and Fats	Carbs and Fats	Proteins
SVR (rbf)	0,72	0,697	0,706	0,535	0,686
Kernel Ridge	0,727	0,702	0,704	0,561	0,692
SVR (linear)	0,509	0,506	0,482	0,349	0,482
SVR (poly)	-0,082	-0,153	-0,164	-0,148	-0,191
Neural Networks	0,694	0,689	0,687	0,533	0,674

Table 4: Results of the different methods on the different clusters:

Method	Cluster 1	Cluster 2	Cluster 3
SVR (rbf)	0,13	- 0,01	0,42
Kernel Ridge	0,04	- 0,37	0,39
SVR (linear)	0,14	- 0,04	0,31
SVR (poly)	- 0,11	- 0,05	- 0,08

Bibliography

Dataset, FAO, 2018.

Hickson, M. and Smith, S., 2018. Advanced nutrition and dietetics in nutrition support.

Roser, M., & Ritchie, H., 2019. Hunger and undernourishment. Our World in Data.

Scikit-learn, Machine Learning in Python.