



# Food Data HYbrid Prediction and Enrichment (Food Data HYPE)

Nicolò Ferretti\*,

[nicolo.ferretti@wur.nl](mailto:nicolo.ferretti@wur.nl)

Hannelore Heuer\*,

[hannelore.heuer@wur.nl](mailto:hannelore.heuer@wur.nl)

Jim Hoekstra\*,

[jim.hoekstra@wur.nl](mailto:jim.hoekstra@wur.nl)

Sander van Leeuwen\*

[sander.vanleeuwen@wur.nl](mailto:sander.vanleeuwen@wur.nl)

\*Food Informatics, Wageningen Food and Biobased Research (WFBR)

## Introduction

Taste is an important attribute of food, as it is one of the key drivers of food acceptability. It could be an indicator for specific nutrients present in food products and vice versa. It is also a main factor in food design, for example in sugar reduction. Worldwide, several databases exist on food composition, and very few on the taste of food products. Currently, there exist a food taste database from INRAE and a sensory database from WUR. These two databases contain food products that were scored on their **sweetness, sourness, bitterness, umami taste, saltiness, and fattiness**. They base their scores (**0-10**) on the Spectrum™ scale (0-15).

It is costly to make such a data set due to the effort of composing and training a taste panel. Therefore, predicting taste using existing knowledge and machine learning techniques can be much more efficient. The Dutch Food Composition Database (NEVO) contains many food products, which can be enriched with a taste. Using the sensory database, the Human Nutrition and Health group of Wageningen Food & Biobased Research developed a model predicting taste based on the average taste of products in the same NEVO food group. This model is used as a baseline for this project.

## Objectives

- Improve on the baseline model using statistical methods, based on the characteristics of a food product.
- Determine whether it is possible to reduce the prediction error by combining the NEVO and INRAE datasets for training using an ontology.

## Method

We base our prediction on the product category, the food forms, and a selection of the micro and macro nutrients of the food products, which can be found in NEVO (2019). The nutrients we use are selected based on their availability for all food products and on their theoretical relevance to taste, judged by a nutrition expert.

We applied one-hot encoding to the product category variable, to transform it to a numeric variable. The food form can have one of three values (solid, semi-solid, liquid), which we label-encoded to the values 0, 1, 2, respectively. When training the neural network, the features were normalised to be between 0 and 1. For all other models, the original, unnormalized values were used.

As a performance metric we use the Mean Squared Error (MSE). Since the distribution of the target variables is skewed (see Figure 1), MSE seems to be a good choice as it "punishes" large errors more due to the fact that the errors are squared. This prevents the model from always predicting zeroes.

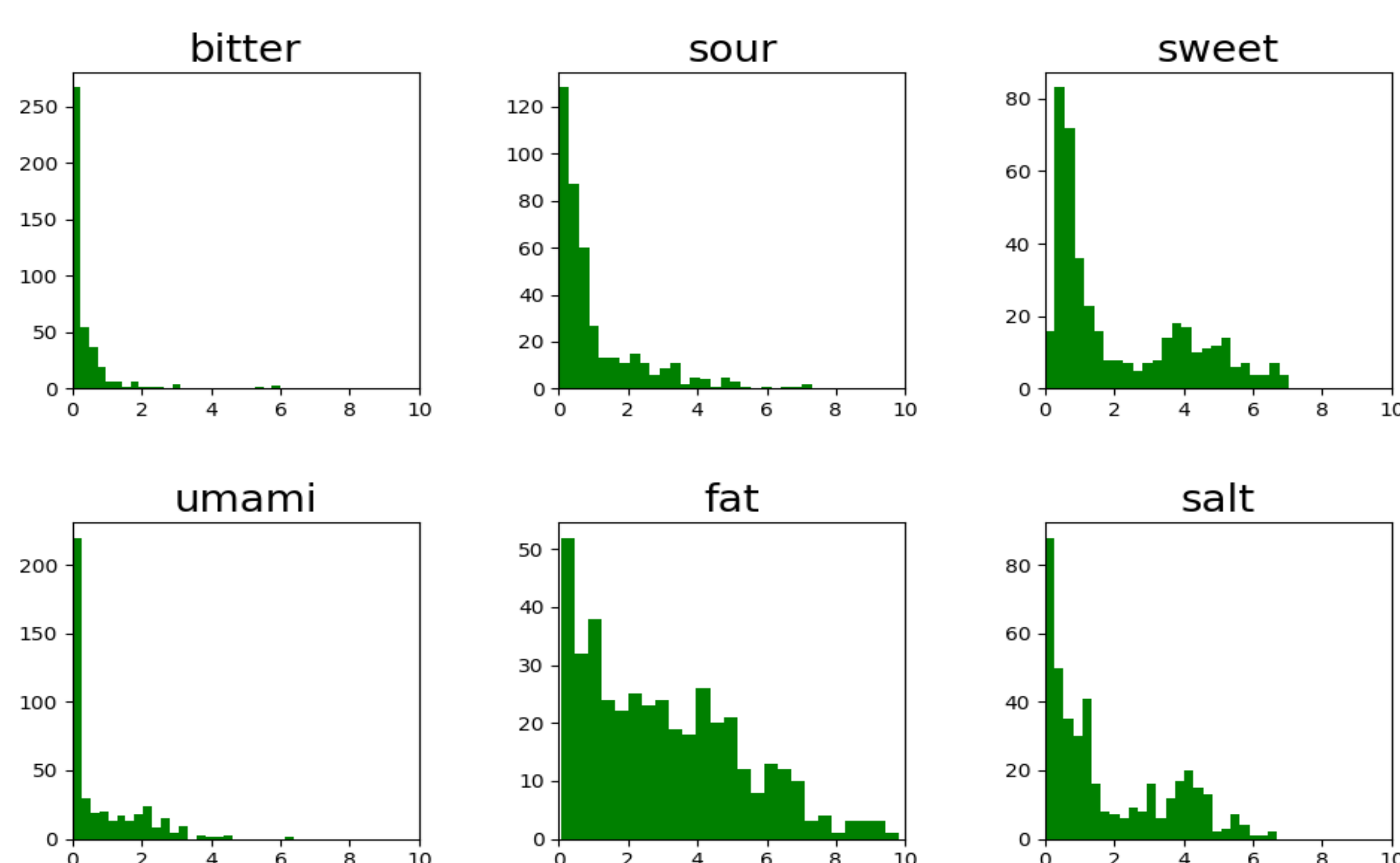


Figure 1. A histogram of the target variables of both the Sensory and INRAE data combined.

We created an ontology to combine the dataset from INRAE and the sensory database and to make a connection to the NEVO database. The ontology has been developed in RDF/OWL and made use of the existing Food Item Ontology and the Ontology of Units and Measure.

After combining the datasets, if multiple products had a connection to the same NEVO code, an average of the taste scores was taken. In the end, we used 384 food products from the sensory database and 50 from the INRAE database for the models. The taste scores from the sensory database were divided by 10 to standardize the scores.

## Results

Table 2. The validation MSEs, comparing models trained on sensory data only or sensory and INRAE data and the average result of the Random Forest model on the test set.

Model	Sensory	Sensory + INRAE data	Test set
Baseline	2.48	2.63	
Random Forest	0.79	0.79	0.72
XGBoost	0.76	0.81	
Artificial Neural Network	1.61	1.36	

The different machine learning models were fitted to the data using 5-fold cross-validation. The validation MSEs of all models are shown in Figure 2. Every box is generated from 30 values (five cross-validation runs times six distinct taste values). The mean validation MSE for every model is also shown in Table 2, for easy comparison.

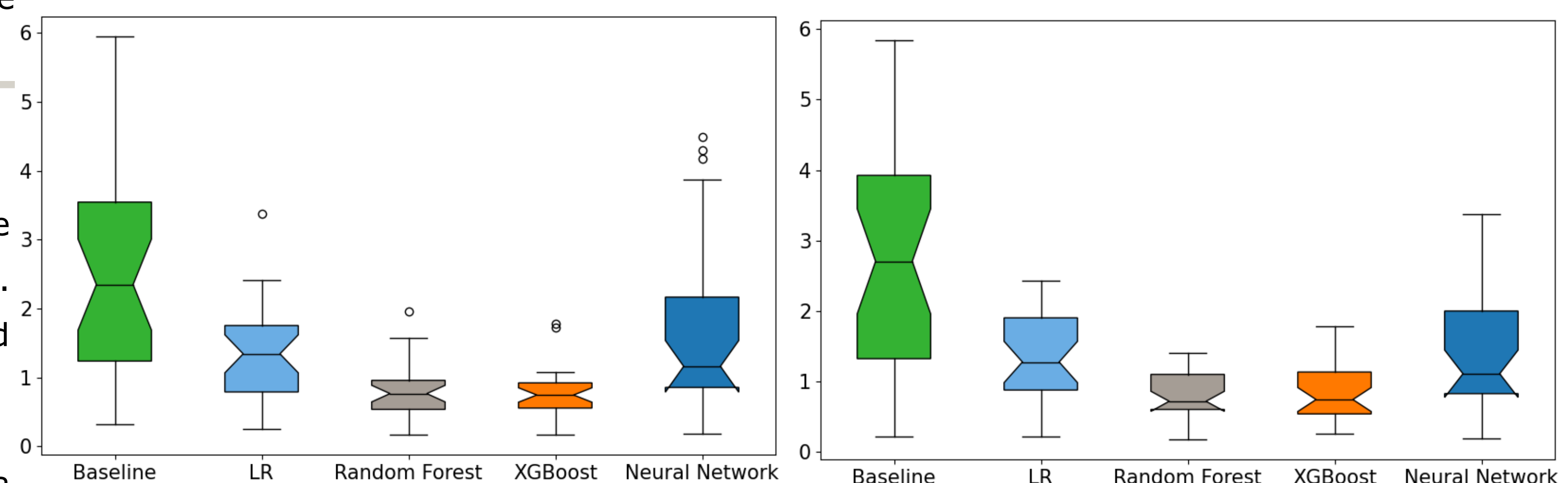


Figure 2. Comparison of the baseline model vs. a selection of the models we tried: linear regression, random forest, XGBoost and neural network. The y-axis shows the mean squared error. In 2(A) INRAE data is not included in the training/validation data. In 2(B) it is.

## Conclusions

- The prediction of tastes can be improved using machine learning models. The Random Forest and XGBoost perform the best.
- The difference between adding or not adding the INRAE data does not result in a significant difference in results. This can be explained by the fact that there are not many data points available in the INRAE database.

Ways to improve the quality of the predictions:

- Use a global food product database instead of the Dutch NEVO
- Generalisability: add more data from other countries to ontology
- Use actual branded product information with real nutrients. For product development it is important to zoom into small differences.

## Acknowledgements

We want to thank dr.ir. Monica Mars and Claudia Tang for providing us with the data set and the overall help during the project. We also want to thank Julian Bianco Martinez for all of the inspiration and for teaching us a lot about machine learning.

## References

Martin et al. (2014). Creation of a food database using an in-home "taste" profile method. *Food Quality and Reference*, 36, 70-80. *Nederlands Voedingsstoffenbestand (NEVO)* | RIVM. (n.d.). Retrieved October 18, 2022, from <https://www.rivm.nl/nederlands-voedingsstoffenbestand>

