Foods 2024 Conference

The 5th International Electronic Conference on Foods

28–30 October 2024 | Online

MDPI

Machine learning and isotopic composition of foods: determination of the geographical origin of milk and eggs

Ángela Fernández Gómez[1], Anton Soria-Lopez[1], Maria Garcia-Marti[2], Juan C. Mejuto[1], Gonzalo Astray[1]

[1]Universidade de Vigo, Departamento de Química Física, Facultade de Ciencias, 32004 Ourense, Spain.
[2]Universidade de Vigo, Departamento de Química Analítica y Alimentaria, Facultade de Ciencias, 32004 Ourense, Spain

INTRODUCTION & AIM

In recent years, consumers have become more aware and sensitive to food safety and quality, as well as to the concepts of healthy eating and nutrition, making the detection of fraud in the food chain of vital importance. **Determining the geographical origin of food** products, such as milk and eggs, is crucial to optimize traceability processes, verify their authenticity and quality and prevent fraud such as adulteration or counterfeiting of products. The isotopic composition of food products varies depending on the agroclimatic conditions of origin, and many studies analyse this isotopic composition as a starting point[1]. **Stable isotopes** are generally used to define the place of origin of agricultural products, such as altitude differentiation ($\delta^{2}H$ and $\delta^{18}O$), others for the identification of the kind of grazing vegetation ($\delta^{15}N$) and for the revelation of the type of animal feed ($\delta^{13}C$)[2].

Considering the information mentioned above, this work[1] focused on determining the **origin of milk and eggs** using isotopic compositions as a data source for machine learning models. These tools included algorithms such as **random forest** (RF), **support vector machine** (SVM) and **artificial neural network** (ANN) that would allow the geographical origin of these products (Figure 1). For this purpose, several digital sources were consulted to obtain the necessary data, resulting in the final selection of the published studies of Kalpage *et al.* (2022)[3] for milk and Geist *et al*. (2024)[4] for eggs.

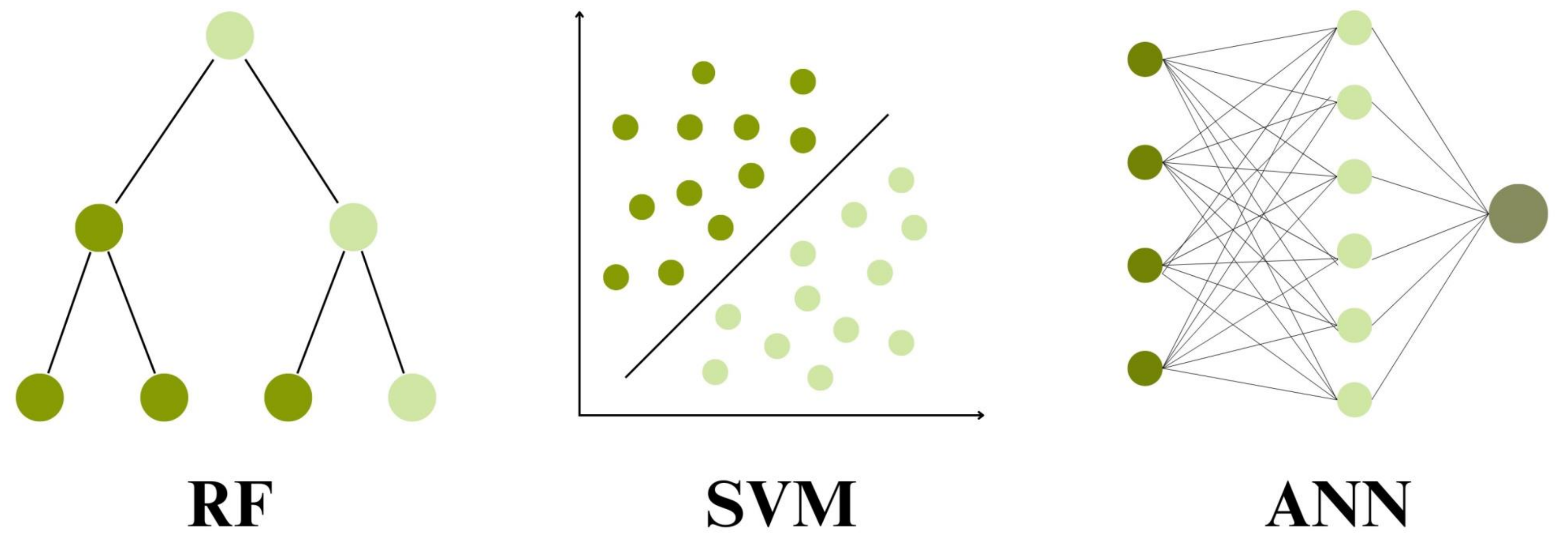

Figure 1. Machine learning algorithms; random forest (RF), support vector machine (SVM) and artificial neural network (ANN).

METHOD

The database of Kalpage *et al.* (2022)[3] contains different stable isotopes ($\delta^{13}C_{VPDB}$, $\delta^{15}N_{AIR}$, $\delta^{18}O_{VSMOW}$ and $\delta^{2}H_{VSMOW}$) to determine the agroclimatic origin of **milk** samples collected in four regions of Sri Lanka, including six farms across three different climatic regions, and seven farms from the last region, reaching a total of **142 samples**. Additionally, the data to determine the eggs' geographical origin were obtained from the OpenAgrar repository in which Geist *et al*. (2024)[4] deposited a dataset including measurements of stable isotope values ($\delta^{13}C$, $\delta^{15}N$ and $\delta^{34}S$) for each egg albumin from **180 samples**, belonging to two geographical locations, and collected over 15 months from various supermarkets in Germany.

Each one of the data matrices was divided into three groups: training (**T**), validation (**V**) and testing (**Z**). During the **training phase**, the model is trained to identify and learn the relationships present in the data. Then, in the **validation phase**, the hyperparameters are adjusted and the best model is selected from those developed in the previous stage. Finally, in the **testing phase**, the generalization power is measurement in previously unseen data, and therefore its performance and predictive power are assessed in a real scenario.

Different hyperparameters were used to evaluate the performance of the different classification models. The accuracy was used to show the percentage of correct predictions, and the kappa statistic value was used to determine whether the classifications were correct due to chance[1]. The different prediction model were developed with RapidMiner Studio Educational 10.2.000 de RapidMiner GmbH.

RESULTS & DISCUSSION

Models were developed using the three algorithms using different hyperparameter combinations. The selected algorithms presented different behaviours in all phases, however, the criterion stablished to identify the best developed model for each algorithm was the highest accuracy value in the validation phase (V). Accordingly, the following table (Table 1) shows the best model to determine the geographical origin of milk and eggs.

Table 1. Adjustments for the best machine learning models developed to classify milk and eggs

		T		V		Z	
Sample	**Model**	**Accuracy**	**Kappa**	**Accuracy**	**Kappa**	**Accuracy**	**Kappa**
Milk	**ANN**	0.986	0.981	0.977	0.969	0.929	0.905
Eggs	**RF**	0.934	0.815	0.906	0.738	0.917	0.769

Among all the models developed to determine the **geographical origin of milk**, it was concluded that the artificial neural network (**ANN**) was the model that obtained the highest accuracy value in the validation phase (97.7%, with a kappa value of 0.969). These high accuracy and kappa values, together with good adjustments in the training phase, confirm that the predictive model is robust and can be applied in real scenarios. This fact can be confirmed by observing the good adjustments data obtained by the artificial neural network model in the testing phase, where the model obtains an accuracy value of 92.9% with a high kappa value of 0.905, both values confirming that the model is suitable and works properly.

For the second group of models developed to determine the **geographical origin of the eggs**, the model that obtained the best results was a random forest model (**RF**). In this case, the RF model obtained an accuracy of 90.6% for the validation phase with a more discrete layer value of 0.738. These values were improved in the training phase, where the model also obtained good results. Although this statistical parameters are not as strong as those achieved for milk's origin , they are still highly reliable, as can be confirmed by the testing phase where the model obtains accuracy around 91.7%.

In view of these results, it would be valuable to continue research with new data to improve the machine learning algorithms aiming to ensure traceability and quality of milk and eggs within the food chain. The models presented here could also be improved by trying to combine different parameters, using different ranges for these hyperparameters, or analyzing different data distributions in each of the model implementation phases.

CONCLUSION

The results obtained in this research **proved the capacity of these machine learning algorithms to ensure the authenticity of milk and eggs**, showing accuracy values above 91.0% in a real scenario. Despite all this, further research is necessary to improve the developed models.

REFERENCES

1.Fernández-Gómez, A. (2024). Aprendizaje automático y composición isotópica: determinación del origen geográfico de productos alimentarios (leche y huevos). Master's Thesis, Universidade de Vigo.

2.Zhao, Ruting, et al. "Chemical analysis combined with multivariate statistical methods to determine the geographical origin of milk from four regions in China." *Foods* 10.5 (2021): 1119.https://doi.org/10.3390/foods10051119.

3.Kalpage, Maheshika, et al. "Stable isotope and element profiling for determining the agroclimatic origin of cow milk within a tropical country." *Foods* 11.3 (2022): 275. https://doi.org/10.3390/foods11030275.

4.Geist, J., Molkentin, J., & Döring, M. (2024). Dataset: Egg authentication using stable isotopes [Data set]. OpenAgrar-Repository. https://doi.org/https://doi.org/10.25826/Data20240229-152143-0.