Agriculture and Food

AIMS Agriculture and Food, 10(4): 989–1003.
DOI: 10.3934/agrfood.2025052
Received: 07 November 2025
Revised: 15 December 2025
Accepted: 17 December 2025
Published: 25 December 2025

https://www.aimspress.com/journal/agriculture

Research article

Classification and evaluation of cow's milk according to the place of origin based on compositional parameters utilizing Machine Learning Techniques

Theodoros Markopoulos [1,*], Sotirios Papadopoulos [2], Stavros Kontakos [3], Alexandros Tsoupras [1,*]

[1] Hephaestus Laboratory, School of Chemistry, Faculty of Science, Democritus University of Thrace, Kavala University Campus, St Lukas, 65404, Kavala, Greece
[2] Directorate General of Agricultural Economy, Region of Eastern Macedonia and Thrace, Komotini, Greece
[3] Department of Production and Management Engineering, Democritus University of Thrace, Xanthi, Greece

* **Correspondence:** Email: thamarkopoulos@yahoo.gr (T.M.); atsoupras@chem.duth.gr (A.T.).

Abstract: The increasing expectations regarding food quality in recent years have prompted a detailed investigation into the factors that define this concept. Several of these factors are strongly shaped by the production method, while in other cases, compositional characteristics are influenced by the production region itself. This creates a distinctive relationship between the product's basic characteristics and its place of origin, which can be leveraged to ensure consistently high discrimination with minimal variation. In this study, we examined the compositional parameters of 84,425 cow's milk samples collected from individual farms of neighboring regions in northern Greece, aiming to assess their relationship with the geographical area of production. Four supervised Machine Learning classification methods—k-Nearest Neighbors, Decision Tree, Random Forests, and Support Vector Machines—were employed, all suitable for Big Data analysis. The findings indicate that all four methods consistently classify the milk samples meaningfully only for two of the four regions, specifically the prefectures of Serres and Xanthi. As anticipated, the Random Forest algorithm achieved the strongest classification performance among the tested techniques.

Keywords: dairy; implementation; machine learning; classification; certification; production

1. Introduction

Quality has long been a central topic of study across numerous scientific disciplines, with particular emphasis on the food sector. The concept of quality is inherently multidimensional, encompassing both objective attributes—such as chemical composition and microbiological integrity—and subjective factors like taste, aroma, and visual appeal. Especially intriguing is how consumers perceive quality, as their perceptions are influenced by a range of determinants including personal taste, cultural background, education, and information disseminated through the media [1]. Consumer preferences regarding food are shaped by a complex network of influences [2]. Consequently, perceived quality results from the interaction between a product's objective and subjective characteristics. According to Grunert (2005) [3], consumer perception of quality is informed not only by a product's appearance, packaging, origin, and price but also by the sensory experience of consumption. In addition, modern trends toward health consciousness and environmental awareness strongly influence consumer choices, while ongoing phenomena such as climate change also affect product quality attributes [4].

Within the dairy industry, quality occupies a particularly prominent position due to the high nutritional value of milk and dairy products and their critical role in agricultural and processing activities. Consumer demand for milk and its derivatives is largely determined by quality, which in turn depends on characteristics such as fat and protein content, microbiological safety, and the absence of harmful chemical residues [5]. These features are shaped by factors including animal health and diet, as well as hygiene standards during milking, storage, and transportation. As consumers become more informed and discerning, the notion of quality expands to include dimensions such as sustainability, transparency, and connection to local production.

This study focuses on two key aspects: (a) some of the compositional characteristics of milk and (b) the region of the production. Milk classification will be examined in light of modern requirements regarding safety and nutritional value, along with the production processes that preserve or enhance these attributes in dairy products. Furthermore, the relationship between basic milk compositional parameters and its production area is analyzed to understand how regional factors contribute to product characteristics.

In the dairy sector, consumer trust is closely tied to product safety and authenticity. Recent research suggests that certified characteristics and geographical indications—such as PGI, PDO, or organic certifications—enhance consumer confidence and product appreciation [6]. Increasingly, sustainability and the connection between place and production methods have become essential elements in how consumers assess quality.

Milk quality is influenced by multiple spatially dependent factors, including animal nutrition, health, hygiene conditions, and processing techniques. Microbiological purity and the absence of toxic substances, such as antibiotics or pesticide residues, are particularly critical [5]. Processing methods such as pasteurization and homogenization ensure safety and nutrient preservation. The quality of the final product depends not only on the raw milk but also on production processes. For example, traditional cheese and yogurt—often produced using local raw materials and artisanal methods—are perceived as higher quality because of their distinctive flavor and texture [7]. Hence, locality and tradition become central determinants of perceived quality. Products associated with specific regions—such as feta and parmesan—gain added value and identity through geographical links [8]. Meanwhile, advances in automation and data-driven production introduce new opportunities for

quality assurance.

The adoption of sustainable practices, such as using renewable energy and reducing water consumption, reflects consumer demand for environmentally responsible products [9]. Ultimately, perceived quality results from a complex interaction between consumer perceptions, product properties, production techniques, and geographic origin. Maintaining a balance between innovation, tradition, and sustainability will be essential for meeting evolving consumer expectations.

Milk and dairy product quality is fundamentally linked to safety, nutrition, and consumer acceptance. Milk is a key dietary staple, valued for its proteins, fats, vitamins, and minerals. Its quality depends on physicochemical composition, microbiological integrity, and the absence of chemical residues—whether deliberate or accidental [5]. The quality of dairy derivatives is further influenced by raw material quality and production practices. In Greece, traditional production methods and high-quality raw ingredients have contributed to the global reputation of products such as feta and Greek yogurt [10,11].

For consumers, dairy quality is associated with nutritional value, safety, and taste. Particular attention is paid to product origin and certification schemes like PDO, which signal authenticity and sustainability [6]. Greek studies have also emphasized authenticity as a vital factor—consumers tend to prefer dairy products that reflect specific regional and traditional production methods [12,13].

Analyzing cow's milk characteristics involves assessing several parameters that affect both nutritional value and safety. Protein composition—especially casein and whey proteins—is critical for cheese and dairy manufacturing. Fat content influences taste and texture, while microbiological safety depends on the absence of pathogenic microorganisms [14]. Residues of antibiotics and pesticides remain significant concerns that affect consumer confidence [5]. Research in Greece has also highlighted that milk quality is tied to animal health and feeding practices, with organic farming producing superior results [15,16].

The geographical area of production significantly affects the physicochemical and microbiological properties of milk. In Greece, such variations are attributed to differences in feed, climate, and farming practices across regions [10]. Globally, milk quality varies widely due to environmental and genetic factors [17]. The association between milk and its place of origin enhances perceived quality, as seen in renowned regional products like feta and Parmigiano Reggiano [8].

Studies have shown that regional characteristics—such as topography, climate, and flora—directly influence milk composition. In Greece, milk from mountainous areas tends to have higher fat and protein content, whereas lowland milk yields are larger but less nutrient-dense [12]. Internationally, similar variations are observed: northern European milk tends to have higher fat levels, while tropical regions exhibit greater variability in microbiological quality due to storage and transport conditions [17]. Overall, milk and dairy quality emerge as multifactorial phenomena shaped by physicochemical, microbiological, environmental, and geographical influences. The literature reveals notable research gaps, particularly regarding local variations and the relationship between quality parameters and production area, highlighting the need for further investigation.

To date, there is a lack of research examining cow's milk quality in Eastern Macedonia and Thrace, specifically aiming to differentiate between the four subregions—Drama, Kavala, Serres, and Xanthi—using machine learning techniques. This study seeks to lay the groundwork for future research by providing methodological insights and preliminary findings that can serve as a basis for deeper exploration.

The primary objectives of this study are twofold. First, it aims to determine whether variables

such as pH, freezing point, fat, protein, lactose, non-fat dry extract (NFDE), somatic cell count, and total bacterial count can effectively discriminate among milk samples from the four regions. Second, it evaluates how the application of machine learning methods on a large dataset can generate valuable insights, both for comparing the effectiveness of these techniques and for predicting future outcomes.

2. Materials and methods

2.1. Data and software

For the statistical analysis, cow's milk data were used which were provided to us by ELOGAK and in particular its local branch, whose jurisdiction extends to a specific area: 84,425 cow's milk samples were used, from individual farms of the Regional Units of Kavala, Drama, Serres and Xanthi (Figure 1 shows the total number of cows in each region), to classify milk based on unique farm-level characteristics (e.g., feed, soil, practices) before any mixing occurred. ELOGAK refers to the Greek Organization for Milk (ΕΛΟΓΑΚ) or the Greek Agricultural Organization ΔΗΜΗΤΡΑ's milk-related services, including the "Artemis" online platform for dairy industry data. It is associated with the collection, management, and online submission of data related to milk production, processing, and distribution in Greece. More specifically, ELGO-DIMITRA manages registers of milk buyers, processors, and cooperatives, and handles data on raw milk production and distribution for the Ministry of Rural Development and Food and other government bodies.

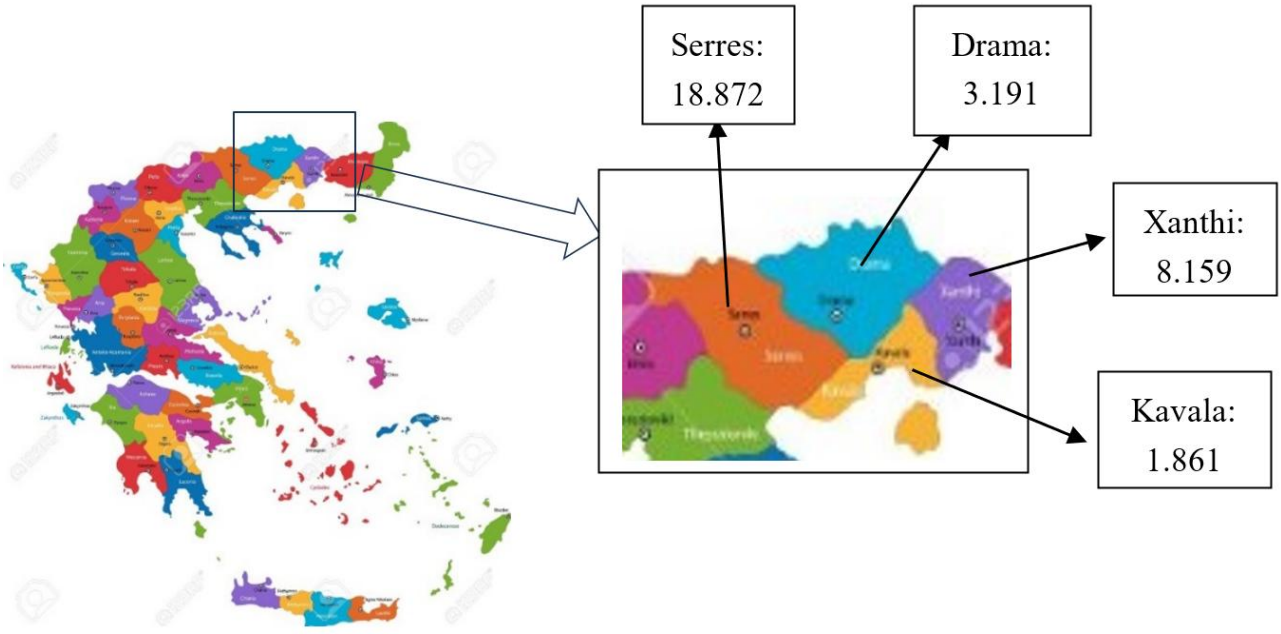

Figure 1. The area of the study with the number of cows for each Region.

The samples for this study were collected over a five-year period, from 2019 to 2023, and for each of them, the following characteristics were measured: pH, freezing point, fat content, proteins, lactose, non-fat dry extract (NFDE), somatic cells (number of) and total bacterial (number of). The presence or absence of antibiotics was also recorded. The number of samples that collected and

analyzed from each investigated region in each of the research years are presented in Table 1.

Table 1. Number of samples that collected and analyzed from each investigated region in each of the research years.

		Year					Total
		2019	2020	2021	2022	2023	
Region	Drama	2118	2160	1968	1806	327	8379
	Kavala	461	519	495	440	79	1994
	Serres	6942	6149	5738	11296	2103	32228
	Xanthi	9101	9590	14209	7741	1183	41824
Total		18622	18418	22410	21283	3692	84425

The existence of a large datasets (Big Data) led to their processing based on techniques suitable for this type of data (data mining algorithms) in order to draw safer and more convincing conclusions. The R software (R 4.4.1, The R Foundation for Statistical Computing, R Development Core Team, (2008) [18]) was used for the analysis.

2.2. Method of analysis

In particular, Classification Analysis was carried out using methods that fall within the scope of Big Data Analysis and Machine Learning. Specifically, four classification methods were used: the k-Nearest Neighbors (kNN) method, the Decision Tree (DT) method, the Random Forest (RF) method and the Support Vector Machines (SVM) method. On the one hand, these methods aim to differentiate the respective milks originating from the Regional Units of Kavala, Drama, Xanthi and Serres, based on some quantitative features, with the ultimate goal of prediction, while on the other hand, the different samples are compared with each other in terms of the application and the best performance of the model they express.

Generally speaking, classification is the process of assigning data to specific groups based on their similarity in terms of some properties. Classification is a supervised learning method, since the categories are known from the outset, i.e. they are given with the set of the other variables used in the research. Following this, some theoretical elements will be mentioned briefly about the classification methods used in this particular research. Much more detailed information about these methods, as well as other similar ones, can be found in the scientific literature [19–21], including more information about regression in general [22–24].

At the heart of all learning machine methodologies is the distinction between training and test sets in order to enable model validation. In order to validate the models applied to our data, the total dataset (84,425 cases) was divided by the 50-50 rule. Two subsets were randomly created (one with 42,213 and the other with 42,212 cases). The first will be the train set and the second the test or control set for the creation and the validation of the models, respectively.

2.2.1. k-Nearest Neighbors (kNN)

This particular methodology is a classification methodology that uses the minimization of distances from clusters or groups. For each element, the k closest to it are checked and the specific

element is assigned to the category that has the majority among the k neighbors. The number k of neighbors as well as the percentage of the necessary "majority" are determined by the users themselves. The distance measure used to determine the neighbors is the Euclidean Distance. The Euclidean distance between two vectors $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})$ and $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \dots, x_{kn})$ in n-dimensional space is defined as:

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_k) = \sqrt{\sum_{j=1}^{n}(x_{ij} - x_{kj})^2} \quad (1)$$

A basic condition for the application of the method in order to avoid misclassification due to the distances used is the normalization of the variables. Thus, the variables in the data set were normalized by range, i.e each observation of a variable is transformed based on its difference from the minimum value of the specific variable and this difference is divided by the variable's range. Moreover, the distance weighting was used, giving the greater weight to the nearest neighbors, achieving as much as possible a balance sampling.

2.2.2. Decision tree and random forests

Decision tree is a method that can be used both for classifying categorical variables and for predicting continuous variables. Decision trees are graphs with nodes and edges that are related to learning rules deduced from the data; they aim to result in 'leaves' that are as consistent as possible with respect to the categories they include. In order to perform the classification, a separation rule is necessary and it's defined by minimizing the following:

$$I(q_1, q_2, \dots, q_K) = \sum_{k=1}^{K} \frac{Q_k}{Q} I(q_k) \quad (2)$$

where $I()$ is one of the impurity indexes (i.e Gini) and the rule separates the examples contained at node q into K descendant nodes $\{q_1, q_2, \dots, q_K\}$ each containing Q_k instances. Q is the total number of instances at q. (More analytically the method is described in Vercellis, (2009) [25]).

A generalization of decision trees becomes a random forest. In this case, a series of trees – and not a single tree – is created; this series of trees aims to reduce the dispersion between the decision trees and thereby increase the accuracy of the model. Moreover, the class weights were used, where the weights are automated as inversely proportional to the classes, to account for the imbalance of the samples.

2.2.3. Support vector machines

This particular methodology was developed to solve the problem of the overfitting of data that occurs in the problems associated with supervised learning. Overfitting is the problem of a model working correctly only for the data set from which it was derived, so that it does not have the ability to generalize and is therefore not stable. Support vector models are more stable, but they lack efficiency. The main idea is the following: Let $\boldsymbol{w}$ denotes the vector of a hyperplane coefficients used to separate a class and b the intercept. Then $\|\boldsymbol{w}\| = \sum_j w_j$ and the method is an optimization problem regarding the minimization of the quantity $\frac{1}{2}\|\boldsymbol{w}\|^2$ subject to $y_i(\boldsymbol{w}'x_i - b) \geq 1$ where y_i is the class of the i-

th observation and x_i is the i-th observation. (More analytically the method is described in Vercellis, (2009) [25]). As in kNN method, the variables in the data set were normalized by range. Also, as in Random Forests, the class weights were used to account for imbalance.

3. Results and discussion

The collected cow's milk sample data were analyzed with each one of the aforementioned classification techniques. The aim was both to seek for differences in classification of the milk samples between the four Prefectures and to compare the four methodologies to the way they discriminate the samples. The results for each different methodology are presented in the following discussion:

3.1. k-Nearest Neighbors

The analysis of the classification for cow's milk for the four prefectures based on the k nearest neighbor analysis showed that this model is moderately satisfactory for predicting the prefecture of origin of cow's milk based on quantitative characteristics (pH, freezing point, fat content, proteins, lactose, non-fat dry extract, somatic cells count and total bacterial count). For the total set of cow's milk data per prefecture that was present in the test data, Table 2 shows where they were finally classified according to this method.

Table 2. Classification results of the control set samples based on the k-Nearest Neighbours (kNN) classification method

	Classification made by the kNN model			
Actual Origin	Drama	Kavala	Serres	Xanthi
Drama	3112	179	733	300
Kavala	151	354	149	277
Serres	664	164	9234	1442
Xanthi	311	318	1483	13968
Accuracy: 0.812, 95%CI: (0.807, 0.816), Kappa: 0.696				

It is quite clear from the results that the largest percentage of cow's milk samples that were correctly classified are among the data that originated from the Prefecture of Serres and the Prefecture of Xanthi. The number of neighbors (hyperparameter k) was chosen to be k = 1 because as it can be shown in Figure 2, the increase in the ARI index (Adjusted Rand Index), with respect to the increase in the number of neighbors, takes its highest value when the number of neighbors is one, and this value does not improved at all (on the contrary, it remains constant) if the number of neighbors required to turn the classification into a category increase. The other hyperparameter regarding the minimum number of neighbors needed in order to classify an observation, was set equal to unity (l = 1). The ARI index for this method is moderate, with a value of 0.494 (on a scale of 0–1). Last, there are 9,373 samples that have not been assigned to any category.

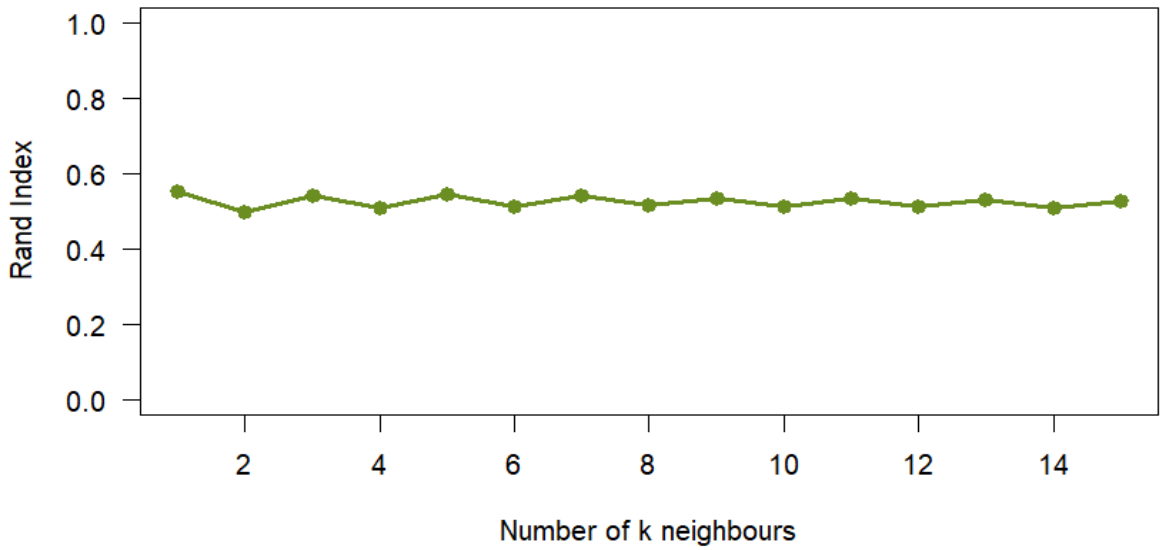

Figure 2. Variation in the ARI index with respect to the number of neighbors used for kNN classification.

3.2. Decision trees and random forests

As mentioned in the introduction, decision trees are a classification technique that can be used both for the classification and for the prediction of continuous variables. In this work, a decision tree was used, which ranks the Prefectures of the origin of the cow's milk by their quantitative properties, also determining the importance they have in the classification. The adaptation of the decision tree for cow's milk yielded the classification shown in Figure 3 below.

As can be seen in Figure 3, the first basic distinction is based on proteins, which the model considers as differentiating all the samples from Serres and Xanthi. From the total set (100%) of the Serres samples, those with a protein value above 3.4 are classified in Serres (32%), and below 3.4 in Xanthi (68%). The next most basic differentiating variable is the freezing point (FRP) for Xanthi with a threshold value of 0.52, and the pH for Serres with a threshold value of 6.7. The samples from Kavala seem not to participate in this method, because the model probably failed to classify this particular class due to the small number of samples. The evaluation of the model is relatively satisfactory based on the Accuracy indices with a value of 0.664, and the Kappa index with a value of 0.421.

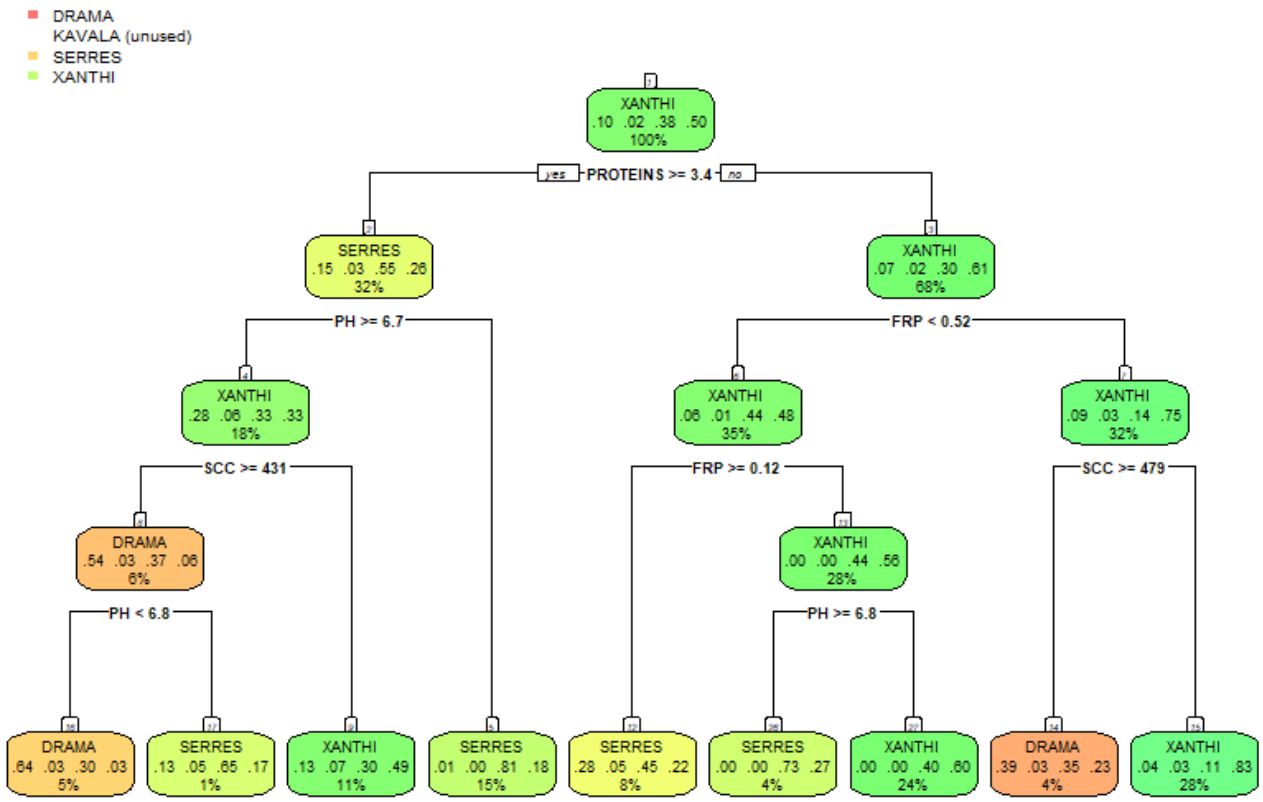

Figure 3. Classification of the region of origin of the cow's milk based on the decision tree (DT) model.

A generalization of the above model is the random forest model. In random forests, a single decision tree is not created; instead, a series of different trees are developed which are derived through random selection. In this way, the dispersion is reduced and we get better results. The adaptation of the random forest model in the case of cow's milk (with the number of trees being 1000) yielded the results shown below in Table 3. Specifically, Table 3 presents the samples as they have been classified based on the model in the test subset of the data. The hyperparameter which denotes the number of the variables used for the discrimination was set equal to three (mtry = 3).

Table 3. Classification results of the control set samples based on the random forest (RF) classification.

	Classification made by the RF model				
Actual Origin	Drama	Kavala	Serres	Xanthi	Classification error
Drama	3579	239	532	133	0.201
Kavala	10	213	20	31	0.222
Serres	436	177	12813	3154	0.227
Xanthi	172	380	2692	17631	0.155
Accuracy: 0.811, 95%CI: (0.811, 0.814), Kappa: 0.682					

It can be observed that with this model, the best estimate is given by the samples of cow's milk from Xanthi, which have a classification error of just 0.155, or 15.5%. The samples from Drama follow with an error rate of 20.1%, followed by the samples from Kavala with an error rate of 22.2%.

The overall accuracy of the model is very satisfactory, since the Accuracy and Kappa indices

have values of 0.811 and 0.682 respectively. The variables were then ranked according to the average reduction of the Gini coefficient. The Gini coefficient is a measure of the dissimilarity of the nodes in each tree. Large values correspond to nodes where all categories are represented. The goal of the trees is to reduce the value of the coefficient. The average reduction and the ranking of the variables are presented in Table 4 and in Figure 4.

Table 4. Average decrease in the Gini index for the importance of the classification variables based on the random forest (RF) model.

	Variable	Average Gini decrease
7	pH	7818.046
6	TBC	7125.209
5	FRP	6204.305
4	Proteins	6158.981
3	FAT	5265.883
2	NFDE	4756.811
1	Lactose	4102.166

From both Table 4 and Figure 4, we notice that the most important differentiating variable among the cow's milk samples included in the study is pH (id = 7), which also presents the largest average decrease in the Gini index (7818.046). The next most important variable is TBC (id = 6) with an average decrease of 7125.209. The variable SCC, i.e. the number of somatic cells of the samples does not appear in Table 4 and Figure 4 as the methodology judged not to be significant for the analysis. Specifically, the SCC appears to have very small variability. Therefore, the variance of the SCC variable is very small and consequently, the Gini coefficient for SCC is close to null and this variable does not appear in the specific analysis.

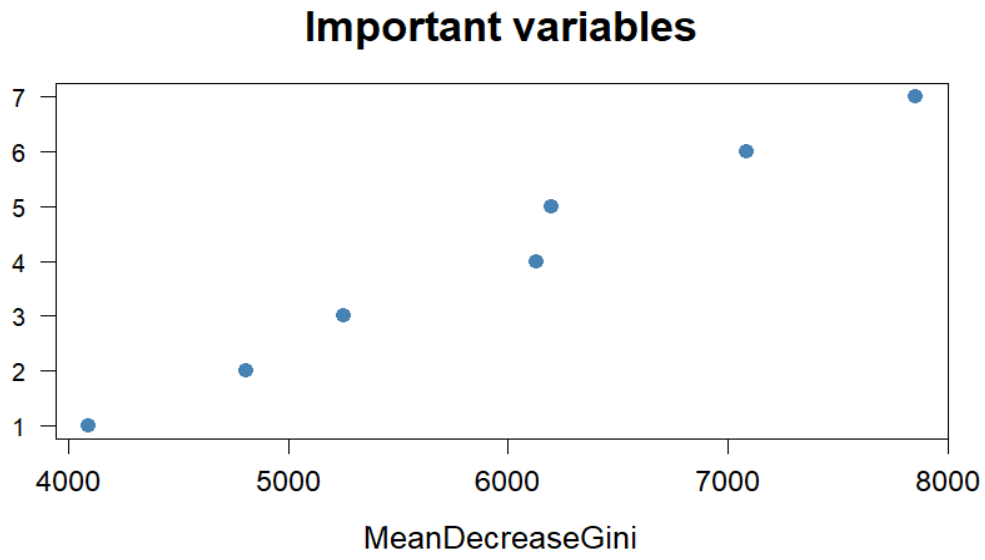

Figure 4. Average decrease of the Gini index for each variable (1 = Lactose, 2 = NFDE, 3 = FAT, 4 = Proteins, 5 = FRP, 6 = TBC, 7 = pH).

As could be seen, the random forest method gives different results compared to the decision tree one, about the most important variable in the discrimination. This is not odd, if someone has in mind that Random Forests are superior to decision trees in terms of the level of prioritization of the variables. As they incorporate information from a very large number of trees, the order in which variables are used "is cleaned" by any noise and other asymmetric trends, and therefore the ranking is more accurate [25].

3.3. Support vector machines

The classification of the cow's milk samples using the support vector machines yielded the results in Table 4, with the classification of the samples in relation to their initial origin. For the analysis, the model was tuned in order to specify the best values of the parameters *cost* and *gamma*. The cost hyperparameter defines the "penalty" if a sample was classified in a wrong class and was set equal to 10. The gamma coefficient which denotes the degree in which elements located in the same space are assigned to the same group. This parameter was set equal to 0.1.

Table 5 shows that by using this particular method, the cow's milk samples from Serres are those that are correctly classified to the greatest extent, followed by the samples from Xanthi. It should not be overlooked that the error rate is relatively high, but as previously mentioned, the specific method is often considered stable even though it is not always efficient. The Accuracy coefficient has a value of 0.665, while the Kappa index has a value of 0.404, which shows a moderate fit of the specific model. These values are similar to the corresponding fit of the decision tree model.

Table 5. Classification results of the control set samples based on the classification made by the Support Vector Machines model.

	Classification made by the SVM model				
Actual origin	Drama	Kavala	Serres	Xanthi	Classification error
Drama	1298	239	706	323	0.494
Kavala	102	320	96	100	0.481
Serres	1496	202	9295	2348	0.303
Xanthi	1275	554	6872	18284	0.322
Accuracy: 0.665, 95%CI: (0.660, 0.669), Kappa: 0.404					

3.4. Integration and comparison of results

Speaking of cow milk, the results obtained from the different classification models clearly show that milk's characteristics like pH, freezing point, fat content, proteins, lactose, non-fat dry extract, somatic cells, and total bacterial count are of great importance in the discrimination of cow milk. The analysis of those properties shows that the milk samples coming from Prefectures of Serres and Xanthi are better discriminated compared to the samples originating from other locations, which also underlines the important role of the geographical origin on the determination of milk quality.

Scientific literature confirms the link between the geographical area and the compositional characteristics of dairy products. More specifically, the work of Listiasari et al. (2024) [26], analyzes the legal protection afforded to geographical indications and their relation with milk product quality in Indonesia, showing that climate and geography influence the qualitative characteristics of milk.

Table 6 compares the different machine learning classification methodologies. In order to avoid bias due to unbalanced samples from the four regions, metrics robust to imbalance are used. These metrics are presented for each method and for each class.

Table 6. Comparison of classification methods for the cow's milk samples used in the study.

		Region			
Model	Metric	Drama	Kavala	Serres	Xanthi
kNN	Balanced Accuracy	0.845	0.665	0.844	0.874
	Precision	0.719	0.380	0.802	0.868
	Recall	0.734	0.348	0.796	0.873
	F1 score	0.726	0.363	0.799	0.871
RF	Balanced Accuracy	0.914	0.604	0.827	0.845
	Precision	0.798	0.777	0.772	0.844
	Recall	0.852	0.211	0.798	0.841
	F1 score	0.824	0.332	0.785	0.843
SVM	Balanced Accuracy	0.638	0.501	0.686	0.729
	Precision	0.502	0.401	0.677	0.675
	Recall	0.311	0.215	0.528	0.872
	F1 score	0.384	0.369	0.593	0.761

Table 6 shows that the most reliable, in terms of metrics accounted for imbalanced data, is Random Forests. Moreover, all metrics agree that Xanthi has the best discrimination followed by Serres. The latter confirms the possibility of using advanced machine learning techniques in milk quality assessment according to the study by Alvanou et al. [27], dealing with the use of microbiological tools for dairy products traceability.

This is also underlined in the importance of geographical origin and linking characteristics with the place of production, which is well reflected in the study by Mwungu et al. [28], which shows how geographical differentiation impacts agricultural practices and product quality within developing countries.

4. Conclusions

The cow's milk produced in the regions of Serres and Xanthi seem to be most distinguishable, due to local environmental and climatic factors. Implementing certification linking specific milk's characteristics with the region of origin is a very good way to introduce transparency and always keep the consumers informed about the origin of the product, since this would also enhance the competitiveness of local products. FAO also confirms the importance of geographical differentiation in determining nutritional and quality traits in foods, including milk [29].

Although there were some restrictions regarding the heavy unbalanced data from the four regions, this approach can be used as a guide to apply machine learning techniques in big datasets from many aspects of agricultural economy. Thus, new opportunities could be developed with a view to strengthening sustainable development in the agricultural sector and consolidating consumers' confidence and trust in local products.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

Special thanks are due to the ELOGAK office of East Macedonia, and especially to its Directorate of Milk and Meat Control Management of the General Directorate for Quality Assurance and Competitiveness of Agricultural Products of ELGO DIMITRA of ELOGAK, which provided us with the relevant data, as well as to the School of Chemistry of the Faculty of Sciences of the Democritus University of Thrace for their Continuous Support.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, All Authors; methodology, S.K.; software, S.K.; validation, A.T.; formal analysis, S.K.; investigation, All Authors; resources, S.P. and A.T.; data curation, T.M., S.P. and S.K.; writing—original draft preparation, All Authors; writing—review and editing, A.T.; visualization, T.M.; project administration, T.M. and A.T.; All authors have read and agreed to the published version of the manuscript.

References

1. Steenkamp JBEM (1990) Conceptual model of the quality perception process. *J Bus Res* 21: 309–333. https://doi.org/10.1016/0148-2963(90)90019-A
2. Brankov T, Markopoulos T, Kontakos S (2020) Long-term trends in food consumption: Comparison between Serbia and Greece. *Ekonomika poljoprivrede* 66: 975–988. https://doi.org/10.5937/ekoPolj1904975B
3. Grunert KG (2005) Food quality and safety: Consumer perception and demand. *Eur Rev Agric Econ* 32: 369–391. https://doi.org/10.1093/eurrag/jbi011
4. Markopoulos T, Stougiannidou D, Kontakos S, et al. (2023) Wine quality control parameters and effects of regional climate variation on sustainable production. *Sustainability* 15: 3512. https://doi.org/10.3390/su15043512
5. Claeys WL, Cardoen S, Daube G, et al. (2014) Raw or heated cow milk consumption: Review of risks and benefits. *Food Control* 31: 251–262. https://doi.org/10.1016/j.foodcont.2012.09.035
6. Napolitano F, Braghieri A, Piasentier E, et al. (2010) Effect of information about organic production on consumer expectations and liking of organic beef. *Meat Sci* 85: 445–449. https://doi.org/10.1016/j.foodqual.2009.08.007
7. Fox PF, Uniacke-Lowe T, McSweeney PLH, et al. (2015) *Dairy Chemistry and Biochemistry.* Springer. https://doi.org/10.1007/978-3-319-14892-2

8. Pacciani A, Belletti G, Marescotti, A, et al. (2001) The role of typical products in fostering rural development and the effects of regulation (EEC) 2081/92. In: *Policy Experiences with Rural Development in a Diversified Europe,* 73rd EAAE Seminar, Ancona.
9. Noordhuizen JPTM, Metz JHM (2005) Quality control on dairy farms with emphasis on public health, food safety, animal health and welfare. *Livest Prod Sci* 94: 51–59. https://doi.org/10.1016/j.livprodsci.2004.11.031
10. Zervas G, Chatzi P (2014) Milk quality: Factors affecting it. *Vet Sci Rev* 47: 33–45.
11. Lu X, Long M, Zhu Z, et al. (2024) Comprehensive genetic analysis and predictive evaluation of milk electrical conductivity for subclinical mastitis in Chinese Holstein cows. *BMC Genomics* 25: 1230. https://doi.org/10.1186/s12864-024-11157-6
12. Karamanlis S, Siamani M (2021) The effect of geographical origin on milk quality. *Hell Agric Sci* 14: 54–66.
13. Li J, Zhu F (2024) Protein factors affecting the quality of infant formula: Optimization, limitations, and opportunities. *Curr Opin Food Sci* 62: 101264. https://doi.org/10.1016/j.cofs.2024.101264
14. Haug A, Høstmark AT, Harstad OM (2007) Bovine milk in human nutrition—A review. *Lipids Health Dis* 6: 25. https://doi.org/10.1186/1476-511X-6-25
15. Makris P, Spyropoulos D (2020) Organic livestock farming and milk quality. *Livest Rev* 22: 19–27.
16. Simões ARP, Bankuti FI, Borges JAR, et al. (2024) Dairy farmers' satisfaction with the price paid by processors in competitive markets. *J Dairy Sci* 108: 2315–2323. https://doi.org/10.3168/jds.2024-25737
17. De Marchi M, Dal Zotto R, Cassandro M, et al. (2009) Milk coagulation ability of five dairy cattle breeds. *J Dairy Sci* 92: 3543–3554. https://doi.org/10.3168/jds.2008-1163
18. R Development Team (2008) R Language definition. Comprehensive R Archive Network. Available from: http://cran.rproject.org.
19. Breiman L (2001) Random Forests. *Mach Learn* 45: 5–32. https://doi.org/10.1023/A:1010933404324
20. Mitchell TM (1997) Machine Learning. New York: McGraw-Hill.
21. Irizarry RA, Love MI (2016) *Data analysis for life sciences with R.* (1st ed.), Chapman and Hall/CRC. https://doi.org/10.1201/9781315367002
22. Flach P (2012) *Machine Learning: The Art and Science of Algorithms That Make Sense of Data.* Cambridge University Press. https://doi.org/10.1017/CBO9780511973000
23. Ho TK (1995) Random Decision Forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, 14–16.
24. Quinlan RJ (1986) Induction of decision trees. *Mach Learn* 1: 81–106. https://doi.org/10.1023/A:1022643204877
25. Vercellis C (2009) *Business Intelligence. Data Mining and Optimization for Decision Making.* Wiley, London. https://doi.org/10.1002/9780470753866
26. Listiasari FR, Kuntari W, Hastati DY (2024) Geographical Indications and Legal Protections for Indonesian Livestock Products: A Critical Analysis of Policy and Enforcement. *DiH: Jurnal Ilmu Hukum* 2024: n.pag. https://doi.org/10.30996/dih.v0i0.11346
27. Alvanou MV, Loukovitis D, Melfou K, et al. (2024) Utility of dairy microbiome as a tool for authentication and traceability. *Open Life Sci* 19: 20220983. https://doi.org/10.1515/biol-2022-0983

28. Mwungu C, Kiprop C, Kirwa L, et al. (2024) Impact evaluation of shamba shape up weather and farming news on smallholder farmers in Kenya. Accelerating Impacts of CGIAR Climate Research for Africa (AICCRA) Impact Evaluation Technical Report, 90 p.
29. FAO (2021) The nutrition and health potential of geographical indication foods. Rome. Available from: https://doi.org/10.4060/cb3913en.

AIMS Press