Journal of Information Systems Engineering and Management
2025, 10(11s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Application of Machine Learning Algorithmic Models for the Authentication of Albanian Mono-Cultivar Olive Oils

Ardiana Topi[1], Daniel Hudhra[1], Dritan Topi[2]*
[1]European University of Tirana, Faculty of Engineering, Informatics and Architecture, Department of Informatics and Technology, Street Xhanfize Keko, Kompleksi Xhura, Tirana, 1000, Albania
[2]University of Tirana, Faculty of Natural Sciences, Department of Chemistry, Blvd. Zogu 1, Tirana, Albania
Corresponding author: dritan.topi@unitir.edu.al

ARTICLE INFO

Received: 10 Nov 2024
Revised: 28 Dec 2024
Accepted: 16 Jan 2025

ABSTRACT

Distinguished for its nutritional benefits and high economic value, olive oil has faced issues with adulteration and fraud. As production increases, the need to identify olive oil by specific cultivars and regions has become more pressing. Analyzing chemical data related to the origin and olive cultivars will facilitate advanced quality control and authenticity practices for Albanian olive oil, enhancing its competitiveness as an organic product. While traditional empirical methods have been relied upon to detect olive oil fraud and evaluate quality, this study pioneers a modern approach using machine learning algorithms to differentiate authentic products from counterfeits. Establishing effective mechanisms and best practices to trace product origins and quality indicators will raise awareness about the risks of adulteration to both consumers' health and the broader food industry. To enhance the precision of origin predictions, data pre-processing steps—especially the normalization process following the isolation of independent features from the target variable—are crucial for distance-based algorithms like kNN, which improve accuracy. Furthermore, performance metrics for all algorithms were evaluated, including k-Nearest Neighbors, Logistic Regression, Support Vector Machines, implementation hyperparameter tuning techniques, and the best-performing model. Applying supervised machine learning methods to categorize Albanian Olive Oils (OO) according to their chemical composition aids in identifying their geographical and cultivar origin. Our results indicate an accuracy of 88.88%, constrained by the limitations of the current dataset; however, we intend to expand the dataset in the future.

Keywords: Olive oil; Authenticity; Machine Learning, Classification, kNN, Logistic Regression, SVM, Albania

INTRODUCTION

The olive tree (Olea europaea L.) is an evergreen plant native to the Mediterranean region and is an essential crop for its agricultural economies. Its fruits and oil distinguish it from other vegetable oils. Its role in the Mediterranean Diet has given importance to economic aspects by spreading to different areas of the world, such as Australia and the Americas, due to its valuable products: olive oil and table olives (IOC, 1996; Boskou, Blekas & Tsimidou, 2006). The geography of Albania has shaped its climatic characteristics, with the western regions exhibiting a typical Mediterranean climate (Figure 1). The olive tree is present in the western and southern areas, alongside the Adriatic and Ionian Seas, two water bodies of the Mediterranean basin. Genetic studies have revealed the existence of twenty-two native olive cultivars, along with several introduced foreign olive cultivars (Topi et al., 2021), distributed strictly across six regions: Berat, Elbasan, Kruja, Lezha, Tirana, and Vlora. Among native cultivars, the most distinguished are Kalinjot, Kotruvs, Kokërrmadh Berati, Mixan, Krips, Nisjot, Ulli i zi, and Bardhi Tirana cvs. (Bianco di Tirana) (Topi, Thomai, & Halimi, 2012; Topi et al., 2019).

Olive oil is widely renowned as the main contributor to the distinguished Mediterranean Diet, globally recognized for its longevity and low incidence of cardiovascular diseases. Despite not being scientifically proven as a component of longevity among Albanians, olive oil is believed to be the key factor in its health benefits (Topi et al., 2025). However, being more expensive than vegetable oils, adulteration with cheaper or lower-quality oils provides significant economic benefits to its sellers. The most common adulterations of olive oil involve mixing it with sunflower, corn,

coconut, soybean, and even hazelnut oil (Ordukaya & Karlik, 2017). Other counterfeiting practices include mixing olive oil from different production years and adding pigments for color improvement, such as combining olive oil products with geographical designations of olive oil of unidentified origin.

According to the data, global production has reached over 9.4 million tons of olives. About 805 million olive trees, accounting for 98%, are cultivated in the Mediterranean region, producing around two million tons of olive oil annually (Ordukaya & Karlik, 2017). The world olive oil market was estimated at $14.64 billion in 2023, increasing to $18.42 billion in 2030 (Aiello, 2024).

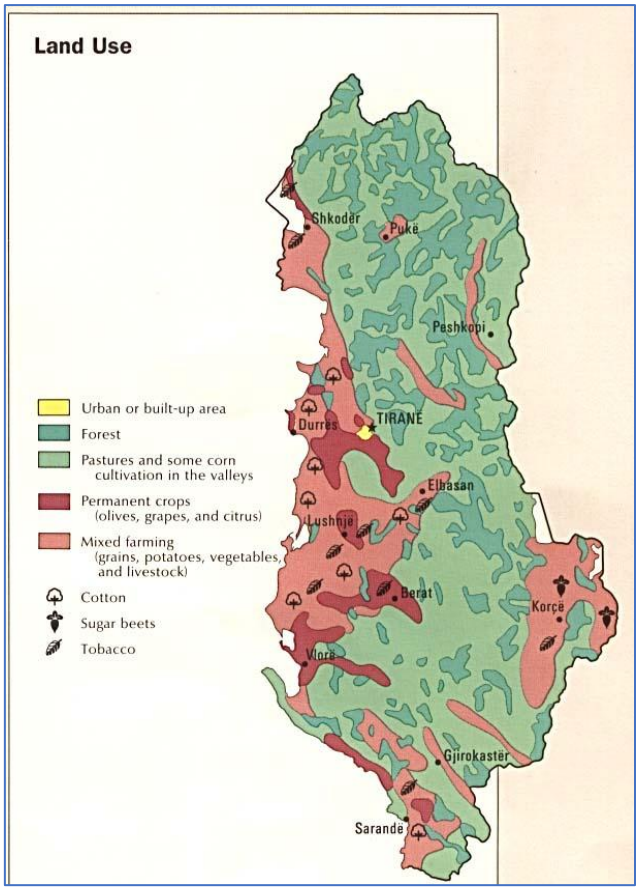


Figure 1. *Annual and permanent crops of Albania (Map of Land Use in Albania, 1990).*

Virgin olive oil (VOO) is extracted from the olive fruit solely by mechanical or other physical means under conditions that do not lead to alterations in the oil and that have not undergone any treatment other than washing, decantation, centrifugation, and filtration (IOC, 1996). OO primarily consists of triacylglycerols (~99%) and minor compounds, which include phenolic compounds as well as other lipids such as hydrocarbons, sterols, aliphatic alcohols, tocopherols, and pigments (Topi et al., 2019). Phenolic compounds belong to different families, such as classes tyrosol and hydroxytyrosol, both identified as phenylethanoid and their derivatives of 4-hydroxybenzoic, 4-hydroxyphenylacetic, and 4-hydroxycinnamic acids; as well as lignans and flavonoids (Topi et al., 2020).

Olive oil is mainly a mixture of palmitic (C16:0), palmitoleic (C16:1), stearic (C18:0), oleic (C18:1), linoleic (C18:2), and linolenic (C18:3) acids, while myristic (C14:0), heptadecanoic, and eicosanoic acids are found in trace amounts (Topi et al., 2012). The fatty acid (FA) composition may differ according to cultivars, production zones, latitude, climate, variety, and fruit maturity stages. Greek, Italian, and Spanish OOs are low in linoleic and palmitic acids while having a high percentage of oleic acid. Tunisian OOs are high in linoleic and palmitic acids and lower in oleic acid (Boskou, Blekas, & Tsimidou, 2006).

Among native olive cultivars, the most important is the Kalinjot cv., which contributes to high domestic production. It covers approximately 50% of plantations at the national level, with a denser structure—over 70% of the plantations' structure—in the southern regions of Vlora and Mallakastra (Table 3) (Velo & Topi, 2017). Under the values established by EU legislation for olive oil commodities, the primary fatty acids are oleic acid (68.03–76.83%), linoleic

acid (7.85–14.22%), and palmitic acid (8.54–13.62%). Meanwhile, there is a threshold for the linolenic acid content to be below 1% (Table 3).

Studies suggest VOO offers health benefits like reducing coronary heart disease risk, preventing certain cancers, and modifying immune responses. These benefits stem from phenolic compounds that help combat chronic diseases such as atherosclerosis and strokes (Topi et al., 2020). Research on phenolic compounds in Albanian OO from native cultivars (Kalinjot, Bardhi Tirana, Ulli i Zi, Krips Kruja, and Bardhi Kruja) revealed that secoridoids are the most abundant, followed by phenolic alcohols. Notably, 3,4-DHPEA-EDA (hydroxytyrosol) and p-HPEA-EDA (tyrosol) were dominant, particularly in Kalinjot virgin OO (Topi et al., 2020).

Albania's five most productive olive areas are Berat, Elbasan, Fieri, Vlora, and Tirana, where 90% of olive production is concentrated. Fieri is the leading region for the first three indicators: total number of plants, number of plants in production, and total production. The economic potential resulting from the main olive cultivars is very high, contributing to rural economies and regions with low development potential, such as southern and inland hilly areas (Topi, Thomaj, & Halimi, 2012).

Alongside quality, the agri-food sector faces a persistent issue regarding traceability of geographical origin. Misleading practices, such as labeling geographical origin incorrectly or using incorrect grape and olive varieties, can erode consumer trust and tarnish the reputation of wine and olive-producing regions. Unfortunately, a growing number of low-quality olive oils often find their way to our tables and are difficult to identify. Therefore, it is crucial to ensure the authenticity of olives and their geographical origins to preserve the integrity of the industry and enhance its supply chain through the implementation of traceability systems.

This study used the Random Forest (RF) and K-nearest neighbors (kNN) algorithms, among others (Sheth, 2022).

The main study contributions

i) evaluation of olive oil modeling using its physicochemical characteristics,

ii) testing different AI-based classification techniques to determine the highest accuracy,

iii) identifying the characteristics of an olive oil cultivar based on its geographical origin.

From these results, an AI-driven system was developed and launched, which is easily accessible for individuals and organizations interested in olive oil. This system predicts the quality and origin of this product based on its physical and chemical properties. These insights highlight the capability of machine learning (ML) algorithms to extract advanced information from unstructured data. In summary, ML enhances the agri-food sector's ability to make well-informed decisions, improve product quality, adapt to shifting market trends, and boost its offerings' quality and competitiveness.

Based on our research study, we identified two approaches to data that we may need:

We judge that dataset #3, with 572 data objects, is closer to our idea, given that, based on seven input characteristics (characteristics of olive oil fatty acids, e.g., palmitoleic acid, stearic acid, oleic acid, etc.), the algorithm can predict the geographical region of the olive oil, which is represented by the output variable we want to predict (in the cited study, there are nine such regions, e.g., North Apulia, South Apulia, Calabria, etc.). Therefore, our target (output) is also categorical (Aiello et al., 2024).

Analogously, we can use similar input characteristics in the dataset being processed for our study. Each data object (olive oil) will be populated with accurate data regarding its fatty acid properties. The more data we have, the higher the prediction accuracy we expect from Machine Learning algorithms. From a confusion matrix perspective, the correct prediction can usually be interpreted as a TP (True Positive) value, i.e., if the actual data is "Kanina" oil and the predicted data is "Kanina" oil. Thus, the prediction by the algorithm was performed correctly. When we discuss classification tasks (e.g., classifying the score given to wine based on quality to predict whether the wine is good, normal, or bad), SVM algorithms (96%), RF (92%), and kNN (87%) have provided higher accuracy of test results for a balanced dataset (Zaza et al., 2023).

SVM (Support Vector Machine)

SVM is one of the most advanced and widely used methods in ML. This method divides the samples through an optimal hyperplane, which maximizes the distance between classes (Nattane *et al.*, 2021). In other words, this

algorithm aims to find a hyper-plane that can efficiently separate different classes of data points within a multi-dimensional space (Zaza *et al*., 2023). Thus, SVM represents one of the most popular supervised learning algorithms that enables the maximization of the discriminant boundary.

In our case study, being a "multiclass" classification, we can use one of the methods:

1-vs-rest. So, we use classifiers; for example, one of them is ["Kanina"] vs. ["Babice, "Rromës" "Qeparo"], etc.

1-vs-1. In this case, to generate a classifier, we use the formula $N * \frac{N-1}{2}$, Where N indicates the number of classes we are considering. So, for N=4, we will have six classifiers, e.g., "Kanina" vs. "Babice", "Kanina" vs. "Rromës" etc...

kNN (K-Nearest Neighbours)

The kNN algorithm performs the classification task by predicting a new given object based on the Euclidean distance (the distance between the training point and the test observation).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (\mathbf{1})$$

In Equation (1), n represents the counts of dimensions, whereas x_i and y_i denotes the data points.

Eq. (1) is presented as:

$$d(x_i, x_t) = \sqrt{\sum_{j=1}^{d}\left(x_{ij} - x_{tj}\right)^2} = \|x_i - x_t\| \qquad (2)$$

where x_i represents the training dataset and x_t indicates the test observation (Zaza *et al*., 2023).

The rationale behind the algorithm and its adjustment to our dataset

Suppose we need to create a new data object, such as an olive oil, for which the user can specify attributes like *oleic acid* and *linoleic acid*. In this scenario, the algorithm aims to determine if the olive oil belongs to one of the following categories (origin): Kanina, Babicë, Rromës, or Qeparo mono-cultivar OO.

Step 1. Calculate the distance based on the input characteristics (such as oleic acid, linoleic acid, etc.). Since our dataset includes multiple input characteristics, the extended formula (1) will be beneficial, allowing us to work beyond just two dimensions.

Step 2. Determining the rank. In straightforward terms, the first rank corresponds to the smallest distance identified in the initial step.

Step 3. Identify the "nearest neighbor" by selecting a specific value of k. For instance, when k=1, the top-ranked place will be chosen (e.g. Kanina). Conversely, with k=4, we examine the top four ranks. If three of these four ranks correspond to Kanina, these three data points are closest to Kanina in distance. Consequently, this leads to the prediction that the new data object, which pertains to an olive oil production origin of interest to the user, will be classified as a Kanina olive oil.

An Olive Oil Origin Predictor - Predicting the origin of the cultivar

A brief description of our dataset

The dataset includes twenty-eight samples from the same cultivar, sourced from various regions of Albania. This study aims to predict the cultivar's origin, termed a "classification task" or "multiclass classification" in our context. The key question is: how do the input features affect the output variable?

Key features will include specific fatty acids and other substances that allow us to distinguish the oil's origin, as each region's unique conditions, including soil and climate, influence its chemical makeup.

Secondly, the machine learning model is essential, enabling the classifier to learn from these values to uncover hidden patterns and forecast the region.

C16:0	C16:1	C17:0	C17:1	C18:0	C18:1n9cis	C18:1n7cis	C18:2n6c	C20:0	C18:3n3	C20:1	C22:0	C24:0	OOorigin
12.89	1.01	0.09	0.16	2.33	69.64	2.88	9.19	0.43	0.81	0.36	0.12	0.09	Oshtima
9.72	0.54	0.14	0.23	2.67	73.01	1.8	9.97	0.46	0.82	0.39	0.12	0.07	Drashovice
9.4	0.5	0.13	0.19	3.02	76.83	0.12	7.95	0.51	0.73	0.4	0.15	0.08	Tragjas
8.54	0.44	0.1	0.15	2.82	76.34	1.74	8.23	0.44	0.64	0.38	0.12	0.06	Trevllazer
9.41	0.44	0.13	0.2	2.95	72.47	1.7	10.76	0.49	0.78	0.41	0.13	0.07	Vezhdanisht
9.16	2.91	0.15	0.21	2.68	71.81	1.81	9.37	0.48	0.79	0.41	0.13	0.07	Kanina
10.75	0.91	0.04	0.07	2.63	68.87	2.26	12.81	0.47	0.63	0.34	0.15	0.07	Kanina
10.22	0.55	0.15	0.22	2.75	72.47	2.04	9.7	0.48	0.79	0.41	0.14	0.07	Panaja

Figure 2. *Displaying a part of the OO dataset, including the first 8 data objects. The last column, "Origin," represents the output variable.*

Table 1. A brief description of features.

Feature	Explanation of the feature	Value type
C16:0 (Palmitic acid)	A saturated FA critical in lipid metabolism and energy storage is commonly found in plant and animal oils.	Continuous
C16:1 (Palmitoleic acid)	A monounsaturated FA with potential antioxidative and anti-inflammatory properties.	Continuous
C17:0 (Heptadecanoic acid)	A rare FA is often used as a biomarker for specific dietary intakes.	Continuous
C18:0 (Stearic acid)	A saturated FA contributes to membrane stability and energy production.	Continuous
C18:1n9cis (Oleic acid)	A monounsaturated FA linked to cardiovascular health.	Continuous
C18:2n6c (Linoleic acid)	An essential polyunsaturated FA with a role in cellular signaling and structural integrity.	Continuous
C20:0 (Arachidic acid)	A long-chain saturated FA with roles in energy storage.	Continuous
C20:1 (Eicosenoic acid)	A monounsaturated FA is associated with metabolic processes.	Continuous
C22:0 (Behenic acid)	Found in vegetable oils, it aids in lipid processing.	Continuous
C24:0 (Lignoceric acid)	A very long-chain FA integral to neural tissue structure.	Continuous

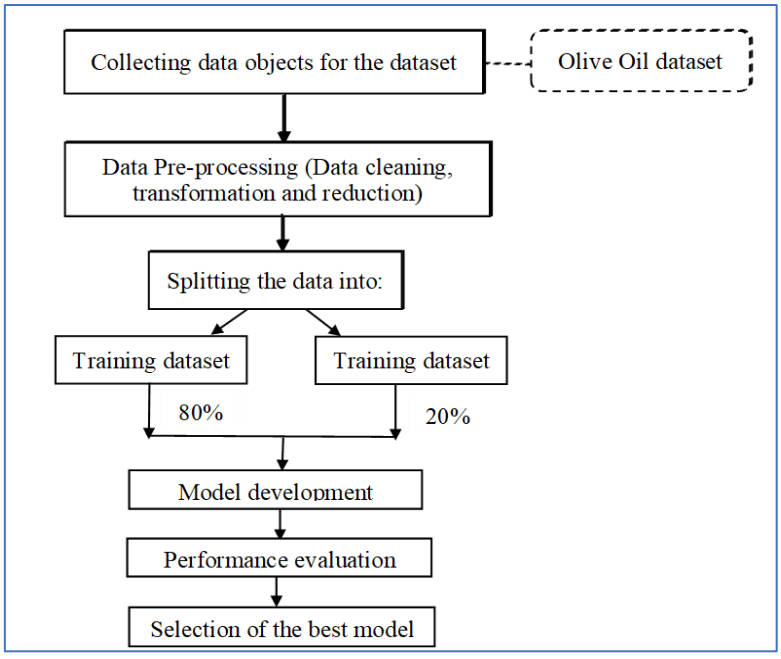


Figure 3. The research methodology according to Suleiman et al. (2022).

Managing imbalanced dataset

RandomOverSampler (ROS) was selected in this analysis to address the severe class imbalance in the target variable. While the Synthetic Minority Oversampling Technique (SMOTE) relies on generating synthetic neighbors for balancing, it requires at least two samples per minority class to function effectively. In cases where certain classes

contain only one sample, SMOTE encounters technical limitations, leading to errors. RandomOverSampler, on the other hand, duplicates existing samples without relying on neighborhood calculations, making it more robust for datasets with extremely imbalanced distributions. (Hayati et al, 2021). This approach ensures balanced class representation while maintaining data integrity, making it a better fit for the given scenario.

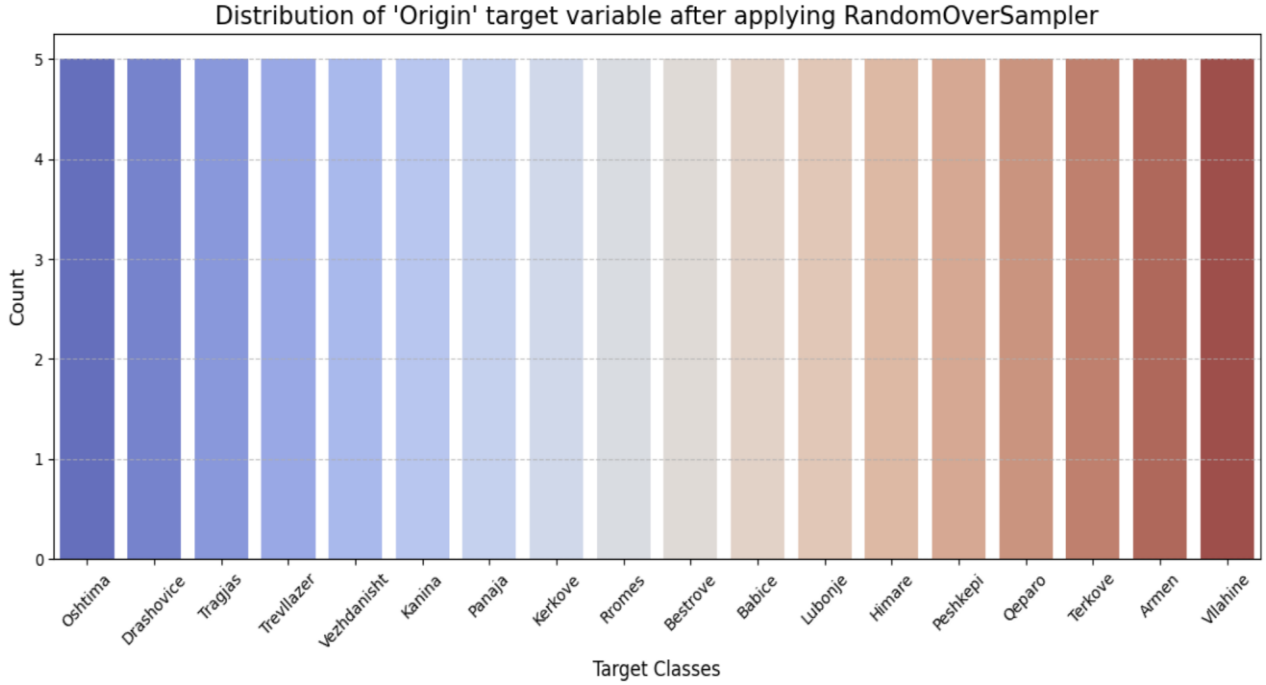


Figure 4. *Class distribution after ROS. All the target classes have equal division.*

Our dataset consists of samples from the same cultivar sourced from various locations (Oshtima, Drashovice, Tragjas, and so on). To balance the dataset, we utilized ROS to replicate instances of the minority classes. Therefore, we confirmed that the distribution of instances across classes became uniform, increasing minority class instances, e.g., from 1 to 5 in the case of Oshtima, as illustrated in Figure 4.

Splitting the Olive Oil dataset into training and test dataset

To include 20% of the data for testing and the remaining 80% used for training the model (also shown in Figure 3), we have used the *scikit-learn* library in Python.

In our training dataset, the model was trained on X_train and Y_train. The prediction will be performed using the X_test (in the unseen samples), and we must compare the predicted results of our models with those of the *Y_test*.

Feature selection and feature extraction (PCA) after splitting the datasets

In the Olive Oil dataset, feature scaling is important as our chosen ML algorithms must calculate distances between data. If we do not perform feature scaling (i.e., the features are not scaled), the features with a higher value range will dominate when distances are calculated.

Scikit-learn is a Python module that integrates the newest machine-learning algorithm for supervised and unsupervised problems (Kothawade, 2021). We intend to make our data standardized, meaning it will have a μ=0 and a a σ=1. The upper and lower values can vary (they do not need to be in range from 0 to 1).

We use the following formula for standardization to calculate the standard score of a sample x:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Where x_i is each value, μ is the mean of the training samples, and σ is the standard deviation of the training samples.

```
X_train.describe().round(1)
✓ 0.0s
```

	1	2	3	4	5	6	7	8	9	10	11	12	13
count	72.0	72.0	72.0	72.0	72.0	72.0	72.0	72.0	72.0	72.0	72.0	72.0	72.0
mean	0.0	0.0	-0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0
std	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
min	-1.3	-0.8	-2.4	-3.0	-2.1	-2.0	-1.9	-1.4	-2.1	-2.8	-2.5	-1.1	-2.0
25%	-0.7	-0.6	-0.2	-0.9	-0.5	-0.5	-0.3	-0.6	-0.6	-0.6	-0.7	-1.1	0.3
50%	-0.5	-0.4	0.4	0.2	0.1	-0.1	0.1	-0.0	0.0	-0.0	0.2	-0.2	0.3
75%	0.6	0.3	0.6	0.5	0.8	0.7	0.4	0.4	0.6	0.4	0.8	0.7	0.6
max	2.0	5.4	1.1	1.3	2.2	1.7	1.7	3.9	1.9	2.4	1.4	4.3	0.9

Figure 5. Statistical data (mean and standard deviation) for the thirteen features post-standardization.

Showing that min-max values might differ from normalization values, necessitating scaling from 0 to 1. Our number of target classes is eighteen, each containing five instances. Thus, the number of instances in the training set X_train $= \frac{80}{100} \times (18 \times 5) = 72$ instances.

Conversely, Principal Component Analysis (PCA) is a dimensionality reduction method that lowers the number of input features while preserving as much information from the original dataset as possible. PCA converts the thirteen original features into smaller principal components, each representing a sizeable portion of the dataset's variance (Da Costa, 2021).

For instance, rather than utilizing all thirteen features, PCA can condense them into a few components that accurately summarize the same information. This technique reduces computational requirements during model training and enhances resource efficiency while maintaining the dataset's essential characteristics. By integrating feature scaling with PCA, the dataset is effectively positioned for optimal Machine Learning Analysis.

Support Vector Machines (SVM) and Multinomial Logistic Regression can classify the PCA-reduced data. SVM maximizes the margin between data classes using a hyperplane. The decision boundary can be expressed as $w^T x + b = 0$, where w is the weight vector, and b is the bias. Logistic Regression applies the SoftMax function. $\sigma(z) = \frac{e^{z(i)}}{\sum_{j=0}^{k} e^{z(j)}}$, where $z = Xw + b$, to estimate probabilities for each class. Our code's output plots decision boundaries by evaluating model predictions across a grid of points in the reduced feature space, enabling clear visualization of classification performance.

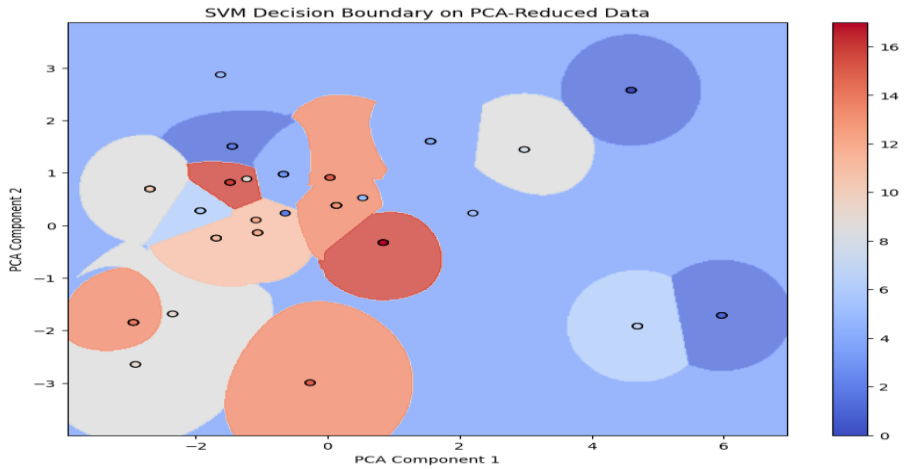


Figure 6. *This visualization demonstrates how the SVM classifier separates different regions (or classes) in a reduced 2D PCA space. By mapping categorical class labels to numeric values, the decision boundary becomes interpretable, representing the regions predicted by the SVM model. This approach is essential for visualizing non-numeric, multiclass classifications.*

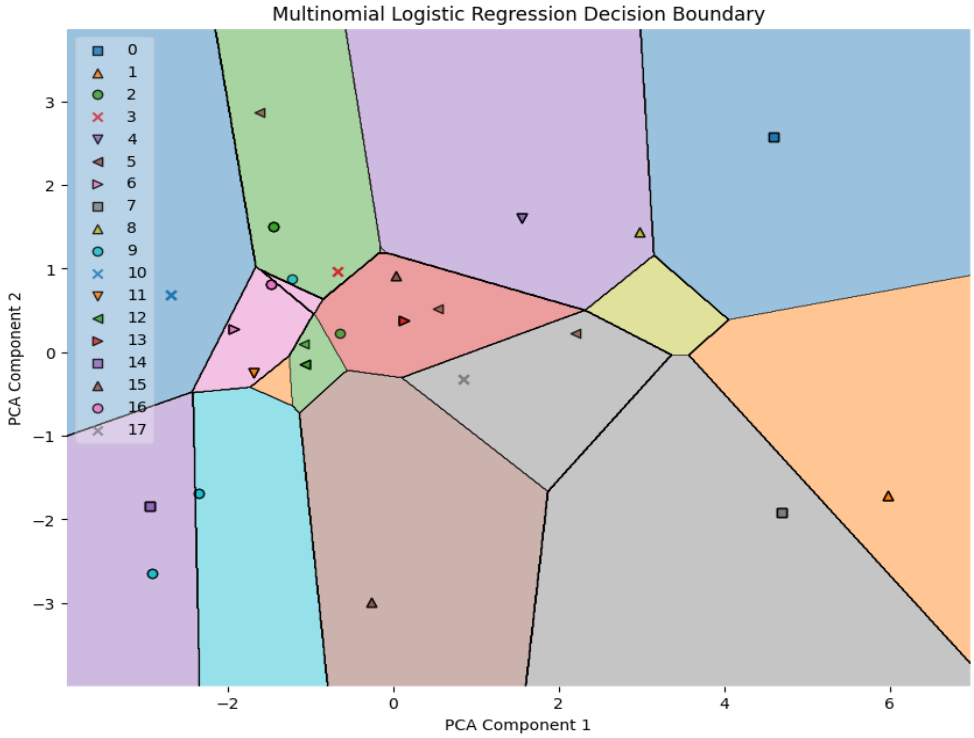


Figure 7. *Showcasing the decision regions of a multinomial logistic regression model in the reduced 2D PCA space. Each color represents a different region or class, indicating the probabilities assigned to each class by the logistic regression model.*

Feature importance using Random Forest

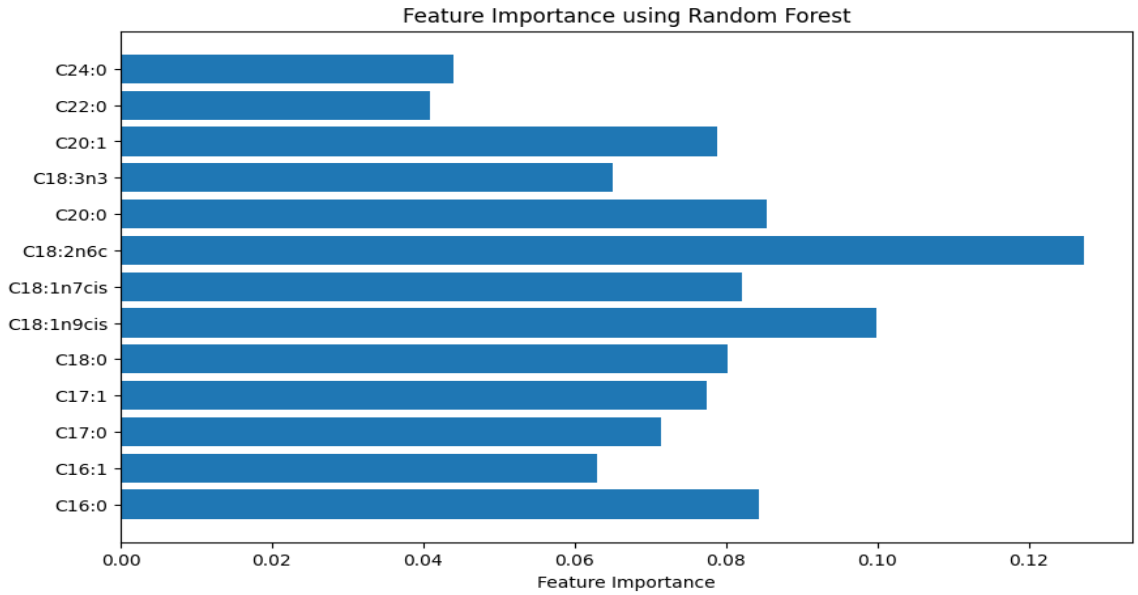


Figure 8. *The feature importance of "Random Forest" according to the Olive Oil dataset.*

A random forest is an ensemble model combining multiple decision trees. A multiclass classification problem assigns the final class based on majority voting among the predictions of all trees (Sharma et al., 2020).

The feature importance analysis of the Random Forest classifier (Figure 8) reveals critical insights into the chemical properties of olive oils from the Kalinjot cultivar. Oleic acid (C18:1n9cis) stood out as the most impactful feature. This result is significant, as oleic acid content directly correlates with environmental and geographical factors influencing olive growth. The high importance of oleic acid emphasizes its role in distinguishing Albanian OO origins.

Secondary contributors, such as palmitic acid (C16:0) and linoleic acid (C18:2n6c), underscore the importance of regional soil and climatic conditions in shaping the fatty acid profile. These findings validate the application of machine learning in agro-food studies, particularly for regional authenticity in mono-cultivar olive oils. The novelty lies in quantifying the individual contributions of these fatty acids in the specific context of Kalinjot olives. This research highlights the chemical fingerprinting of olive oil as a non-invasive method for geographic authentication, laying the groundwork for further exploration into different products across Albania (e.g., wine, dairy foods, etc.).

Evaluation metrics

To check if the predictions are correct or incorrect, there are four ways:

- True Positive: Number of samples that are predicted to be positive that are truly positive.
- False Positive: Number of samples predicted to be positive and truly negative.
- False Negative: Number of samples that are predicted to be negative that are truly positive.
- True Negative: Number of samples predicted to be negative and truly negative.

Moreover, several metrics help assess a model's ability to classify the OO correctly according to its origin. According to Niyogisubizo and coauthors (2024), several classification metrics are used to evaluate the results, such as:

Accuracy: indicates the number of correctly classified instances over the total number.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

After performing ROS, our target class is well balanced, so accuracy will be a good metric that measures how often the classifier predicts correctly.

Precision: expressed by the proportion of positive instances predicted, which are predicted as positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall: calculated as the proportion of current positive instances that are precisely predicted as positive.

$$Recall = \frac{TP}{TP + FN}$$

F1 score: calculated as the balanced mean of recall and precision.

$$F1\ score = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

RESULTS AND DISCUSSIONS

Evaluation Metrics Analysis

The classification results provide critical insights into the performance of various machine learning models across geographically defined classes. *Logistic Regression*, achieving an accuracy of 83.33%, leverages the SoftMax activation function to effectively classify data points in a reduced feature space, ensuring robust multiclass discrimination suitable for our research. However, despite achieving a high weighted average F1-score of 0.79, performance discrepancies persist, as seen in classes like 'Bestrova', highlighting the challenges of effectively modeling certain classes even after balancing the dataset with 5 instances per target class.

SVM, yielding a 77.78% accuracy, emphasizes margin maximization for decision boundaries. The SVM model exhibits higher precision and recall in well-represented classes but struggles with sparse class distributions. Models like *Random Forest* and *Gradient Boosting* match Logistic Regression's accuracy, suggesting their ensemble approach captures complex feature interactions. However, *kNN*, with an accuracy of 61.11%, demonstrates limited generalizability in sparse feature spaces due to its reliance on local data distributions.

Table 2. Logistic Regression Model.

Accuracy: 83.33 %				
Classification Report:				
Regions (villages)	Precision	Recall	F1-score	Support
	0.769	0.833	0.79	
Armen	1.00	1.00	1.00	2
Babicë	1.00	1.00	1.00	1
Bestrova	0.00	0.00	0.00	1
Drashovica	1.00	1.00	1.00	2
Himara	1.00	1.00	1.00	1
Kanina	0.00	0.00	0.00	2
Lubonja	1.00	1.00	1.00	1
Oshtima	1.00	1.00	1.00	3
Qeparo	1.00	1.00	1.00	1
Rromës	0.50	1.00	0.67	1
Trevllazër	1.00	1.00	1.00	1
Vezhdanisht	1.00	1.00	1.00	1
Vllahina	0.33	1.00	0.50	1
Accuracy			0.83	18
Macro avg	0.76	0.85	0.78	18
Weighted avg	0.77	0.83	0.79	18

Table 3. Support Vector Classifier Model.

Accuracy: 77.78 %				
Classification Report:				
Regions (villages)	Precision	Recall	F1-score	Support
	0.713	0.778	0.731	
Armen	1.00	1.00	1.00	2
Babica	1.00	1.00	1.00	1
Bestrova	0.00	0.00	0.00	1
Drashovica	1.00	1.00	1.00	2
Himara	1.00	1.00	1.00	1
Kanina	0.00	0.00	0.00	2
Lubonja	1.00	1.00	1.00	1
Oshtima	1.00	1.00	1.00	3
Qeparo	1.00	1.00	1.00	1
Rromës	0.50	1.00	0.67	1
Kërkova	0.00	0.00	0.00	0
Trevllazër	0.00	0.00	0.00	1
Vezhdanisht	1.00	1.00	1.00	1
Vllahina	0.33	1.00	0.50	1
Accuracy			0.78	18
Macro avg	0.63	0.71	0.65	18
Weighted avg	0.71	0.78	0.73	18

Table 4. The kNN Model.

Accuracy: 61.11 %				
Classification Report:				
Regions (villages)	Precision	Recall	F1-score	Support
	0.507	0.611	0.536	
Armen	0.40	1.00	0.57	2
Babica	1.00	1.00	1.00	1
Bestrova	0.00	0.00	0.00	1
Drashovica	1.00	1.00	1.00	2
Himara	1.00	1.00	1.00	1
Kanina	0.00	0.00	0.00	2
Lubonja	1.00	1.00	1.00	1
Oshtima	0.00	0.00	0.00	3
Qeparo	1.00	1.00	1.00	1
Rromës	1.00	1.00	1.00	1
Kërkova	0.00	0.00	0.00	0
Trevllazër	0.00	0.00	0.00	1
Vezhdanisht	1.00	1.00	1.00	1
Vllahina	0.33	1.00	0.50	1
Accuracy			0.61	18
Macro avg	0.55	0.64	0.58	18
Weighted avg	0.51	0.61	0.54	18

Table 5. The Decision Tree Model.

Accuracy: 83.33 %				
Classification Report:				
Regions (villages)	Precision	Recall	F1-score	Support
	0.796	0.833	0.806	
Armen	1.00	1.00	1.00	2
Babica	1.00	1.00	1.00	1
Bestrova	0.00	0.00	0.00	1
Drashovica	1.00	1.00	1.00	2
Himara	1.00	1.00	1.00	1
Kanina	0.00	0.00	0.00	2
Kërkova	0.00	0.00	0.00	0
Lubonja	1.00	1.00	1.00	1
Oshtima	1.00	1.00	1.00	3
Qeparo	1.00	1.00	1.00	1
Rromës	1.00	1.00	1.00	1
Trevllazër	0.33	1.00	0.50	1
Vezhdanisht	1.00	1.00	1.00	1
Vllahina	1.00	1.00	1.00	1
Accuracy			0.83	18
Macro avg	0.74	0.79	0.75	18
Weighted avg	0.80	0.83	0.81	18

Table 6. The Random Forest Model.

Accuracy: 83.33 %				
Classification Report:				
Regions (villages)	Precision	Recall	F1-score	Support
	0.778	0.833	0.796	
Armen	1.00	1.00	1.00	2
Babica	1.00	1.00	1.00	1
Bestrova	0.00	0.00	0.00	1
Drashovica	1.00	1.00	1.00	2
Himara	1.00	1.00	1.00	1
Kanina	0.00	0.00	0.00	2
Lubonja	1.00	1.00	1.00	1
Oshtima	1.00	1.00	1.00	3
Qeparo	1.00	1.00	1.00	1
Rromës	0.33	1.00	0.50	1
Trevllazër	0.50	1.00	0.67	1
Vezhdanisht	1.00	1.00	1.00	1
Vllahina	1.00	1.00	1.00	1
Accuracy			0.83	18
Macro avg	0.76	0.85	0.78	18
Weighted avg	0.77	0.83	0.79	18

Table 7. The Gradient Boosting Model.

Accuracy: 83.33 %				
Classification Report:				
Regions (villages)	Precision	Recall	F1-score	Support
	0.764	0.833	0.791	
Armen	1.00	1.00	1.00	2
Babica	1.00	1.00	1.00	1
Bestrova	0.00	0.00	0.00	1
Drashovica	1.00	1.00	1.00	2
Himara	1.00	1.00	1.00	1
Kanina	0.00	0.00	0.00	2
Lubonja	1.00	1.00	1.00	1
Oshtima	1.00	1.00	1.00	3
Qeparo	1.00	1.00	1.00	1
Rromës	0.50	1.00	0.67	1
Kërkova	0.00	0.00	0.00	0
Trevllazër	1.00	1.00	1.00	1
Vezhdanisht	1.00	1.00	1.00	1
Vllahina	1.00	1.00	1.00	1
Accuracy			0.83	18
Macro avg	0.75	0.79	0.76	18
Weighted avg	0.81	0.83	0.81	18

Note: The last column in each of the following tables (named "support") provides us the number of samples in each class.

Confusion Matrix Analysis

The confusion matrices across models provide a comparative perspective on classification performance. *Logistic Regression*, as a baseline linear model, struggles to define boundaries between regions due to the inherent complexity of the dataset. *SVM* improves upon this by introducing kernel-based transformations; however, some misclassifications persist due to overlapping class distributions.

Decision trees exhibit overfitting tendencies, which is evident from their perfect predictions for certain classes. *As ensemble methods, Random Forest and Gradient Boosting* mitigate this issue by aggregating multiple weak learners, demonstrating superior performance. These models effectively capture complex patterns within the data, highlighting their robustness for multiclass classification tasks, like the one in our study.

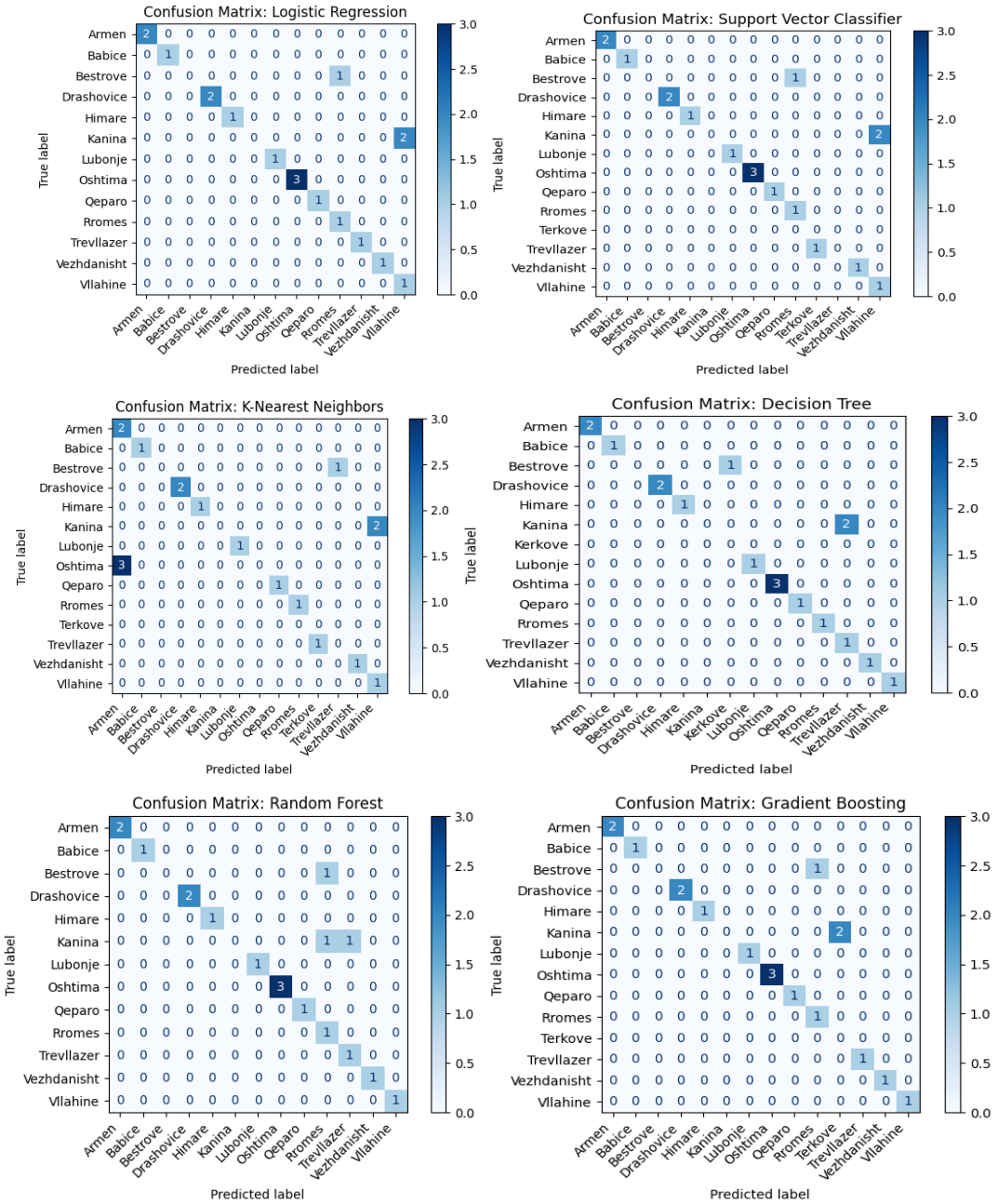


Figure 9. *Visualization of confusion matrices created using the ConfusionMatrixDisplay class from the sklearn—metrics library.*

Hyperparameter Tuning for each classifier

This study utilizes various machine learning algorithms to classify the origin of olive cultivars based on fatty acid composition derived from Kalinjot samples in Albania. Fatty acid profiles are biochemical markers for determining geographical origin and cultivar authenticity. The predictive models were evaluated using *accuracy, precision, recall,* and *F1 score* to assess their performance in distinguishing between potential origins. Furthermore, in our analysis, we systematically tuned the hyperparameters of various models to optimize their performance.

kNN

The main goal is to find the K-nearest neighbors to a given data point. In this case, the metric used is **Euclidean distance**, which is the distance between two points in the hyperplane. This metric, along with Manhattan distance are special cases of Minkowski distance[24]:

$$d(x,y) = \left(\sum_{i=1}^{n}(x_i - y_i)^p\right)^{\frac{1}{p}}$$

When the power parameter for the Minkowski metric (p) is set to 2, we get the formula for the Euclidean distance, which we are considering in the code.

According to the documentation of *scikit learn*[25], the value of the other main hyperparameter of kNN (k, number of neighbors) is set to 5 by default. This value does not guarantee the highest accuracy; thus, some experiments with k values up to 28 (also shown in Figure 10) highlight that the optimal value of k is equal to 2, which gives an accuracy of 83,3%.

Value of k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Accuracy	0.83	0.83	0.83	0.61	0.61	0.61	0.44	0.28	0.11	0.11	0.11	0.11	0.11	0.11	0.06

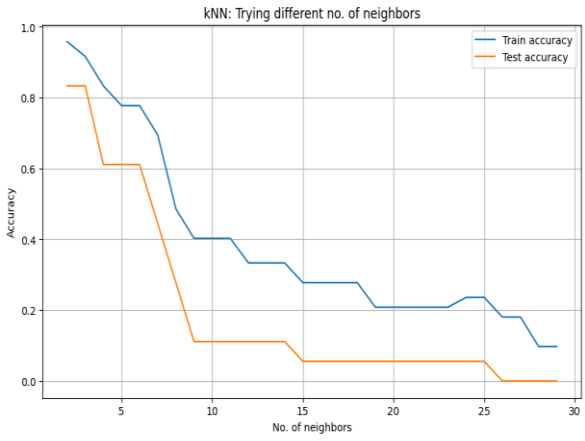


Figure 10. A number of nearest neighbors and their corresponding accuracy.

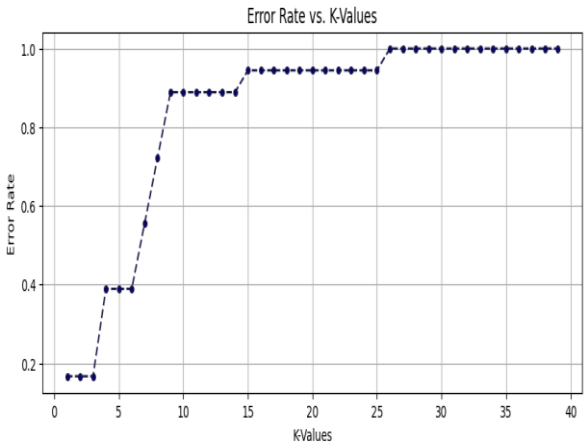


Figure 11. Number of nearest neighbors and the corresponding error rate,

Another way to find the optimal value of k is to use a graph that describes the error rate for each of our k values in the specified range in Figure 11. From the graph, we have chosen the value of k = 2 because the lowest point (error) represents the best or most optimal value of k (Figure 11). Therefore, after hyperparameter tuning of the kNN classifier, we received an accuracy of 83,33%, a precision score of 76,85%, a recall of 83,3%, and an F1 score of 78,7%.

Tuning k = 2, the accuracy and the F1 score achieved for kNN are comparable to Logistic Regression. While kNN is advantageous for its simplicity and instance-based learning, its sensitivity to noise and reliance on distance metrics pose challenges for datasets like this one, where feature scaling significantly impacts performance.

Hyperparameter search

The process of Grid Search involves the creation of a grid of hyperparameters and searching for the combination that produces the highest validation score. In the case of the kNN classifier, we used it to optimize the neighbors' parameter, which controls how many neighbors influence the classification decision. For example, we defined a parameter grid (*param_grid = {'n_neighbors': np. range (2, 30, 1)}) to test neighbor* values between 2 and 29. Using 5-fold cross-validation (*KFold(n_splits=5)*), the data was split into training and validation sets to evaluate each candidate value. The process ensured that the model was assessed on unseen data during each iteration, reducing the risk of overfitting (that usually happens for low values of k).

After evaluating all combinations, GridSearchCV identified the optimal number of neighbors, which was employed to train the final kNN model on the entire training dataset. This approach led to a significant improvement in accuracy, thus improving the model performance.

The main reasons we proceeded with *Grid Search* rather than *Randomized Search* are that we are dealing with a small number of hyperparameters and training a simple model, although it consumes considerable computational resources. We might train too many models in the future, and the number of hyperparameters and their values will increase.

Multinomial Logistic Regression

For this multiclass classification task, the multiclass is "*multinomial*." Thus, the SoftMax function is crucial to estimating the predicted probability of each class[28].

This model achieved an accuracy of 83.33%, a precision of 76.85%, a recall of 83.33%, and an F1 score of 78.7%. These values suggest a moderate performance, with the model excelling in recall but underperforming in precision. The *L2 regularization* minimizes overfitting, while the *lbfgs solver* ensures computational efficiency. However, logistic Regression's linear decision boundaries limit its ability to model the complex relationships inherent in fatty acid profiles.

SVM

With hyperparameters C=1 and γ=1, the SVM model outperformed all other algorithms, delivering an accuracy of 88.88%, precision of 94.44%, recall of 88.88%, and an F1 score of 90.74%. The SVM's ability to map input features into a higher-dimensional space using the radial basis function (RBF) kernel enables it to effectively capture non-linear patterns, making it the most suitable algorithm for this dataset.

The model achieved accuracy scores of 0.875, 0.958, and 0.875 across three folds, indicating consistent performance. Each fold's evaluation completed almost instantly, as reflected by the total time of 0.0 seconds per fold (Figure 12).

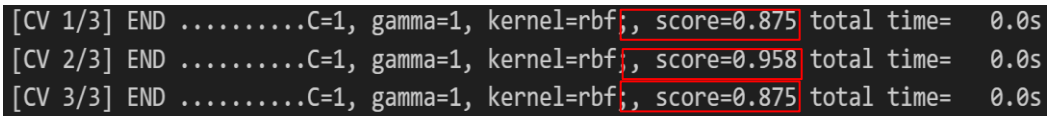


Figure 12. *Results of the SVM model using cross-validation with an RBF kernel, where the hyperparameters were set to C=1 and gamma=1.*

Decision Tree

The decision tree model achieved an accuracy of 83.33% and an F1 score of 80.55%, benefiting from its ability to model non-linear relationships. Using the Gini impurity criterion, the model effectively partitioned the feature space. However, decision trees are significantly prone to overfitting when hyperparameters like maximum depth and minimum samples per split are not restricted.

Random Forest

Regarding the RF algorithm, if we adjust the maximum depth to 9 (considering that the number of estimators is set to 100 by default), the accuracy will almost remain the same (84%). However, the precision, recall, and F1 scores will slightly increase (76.85%, 83.33%, 78.7%). Thus, by aggregating predictions from 100 trees and constraining the maximum depth to 9, the model reduced overfitting and captured more complex interactions between fatty acid features.

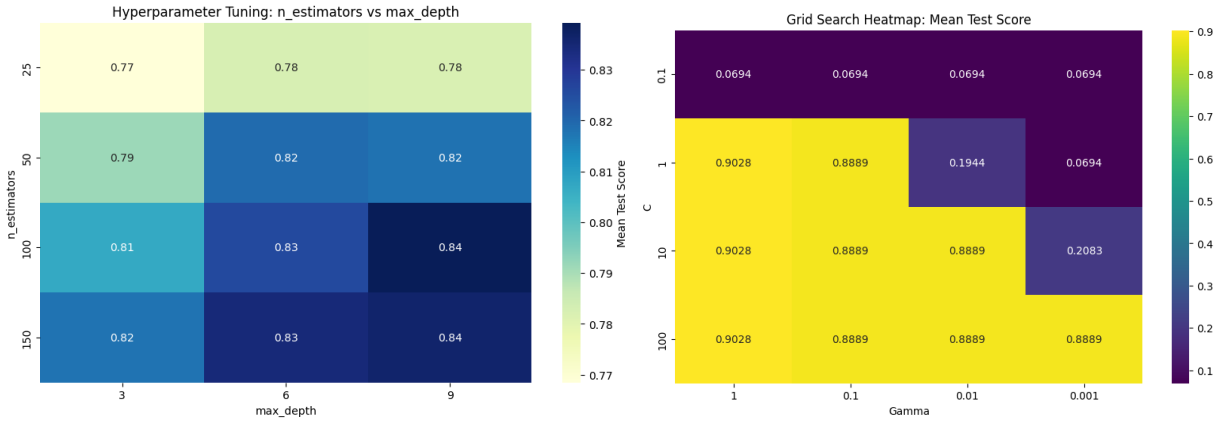
On the other hand, another observation was that setting the maximum depth to three and the number of estimators to twenty-five led to very low accuracy, precision, recall, and F1 scores (61,11%, 50%, 61,11%, and 53,15%, respectively).

Gradient Boosting

The gradient boosting model achieved an accuracy of 83.33% and an F1 score of 81.48%, slightly outperforming other ensemble methods. With a learning rate of 0.01 and 50 estimators, the model demonstrated the potential for further optimization. Thus, Gradient Boosting's iterative nature and ability to minimize residual errors make it a robust algorithm for datasets with subtle feature interactions.

Table 8. Algorithmic Models among different mono-cultivar Oos.

Algorithmic model	**Hyperparameter(s)**	**Evaluation Metrics**			
		Accuracy	**Precision score**	**Recall**	**F1 score**
Multinomial Logistic Regression	C = 1 Multi_class = 'multinomial' Penalty = 'l2' Solver = 'lbfgs'	83.33 %	76.85%	83.33%	78.7%
SVM	C = 1 Gamma = 1	88.88%	94.44%	88.88%	90.74%
kNN	k = 2 (number of neighbors)	83.33%	76.85%	83.3 %	78.7%
Decision Tree	Criterion='gini', Max_depth=None, Min_samples_leaf= 1, Min_samples_split= 2	83,33 %	79,62%	83,33%	80,55%
Random Forest	Maximum of depth = 9 No. of estimators = 100 (by default)	84%	76.85%	83.3%	78.7%
Gradient Boosting	Learning rate = 0.01 No. of estimators = 50	83.33 %	80.55%	83.33%	81,48%



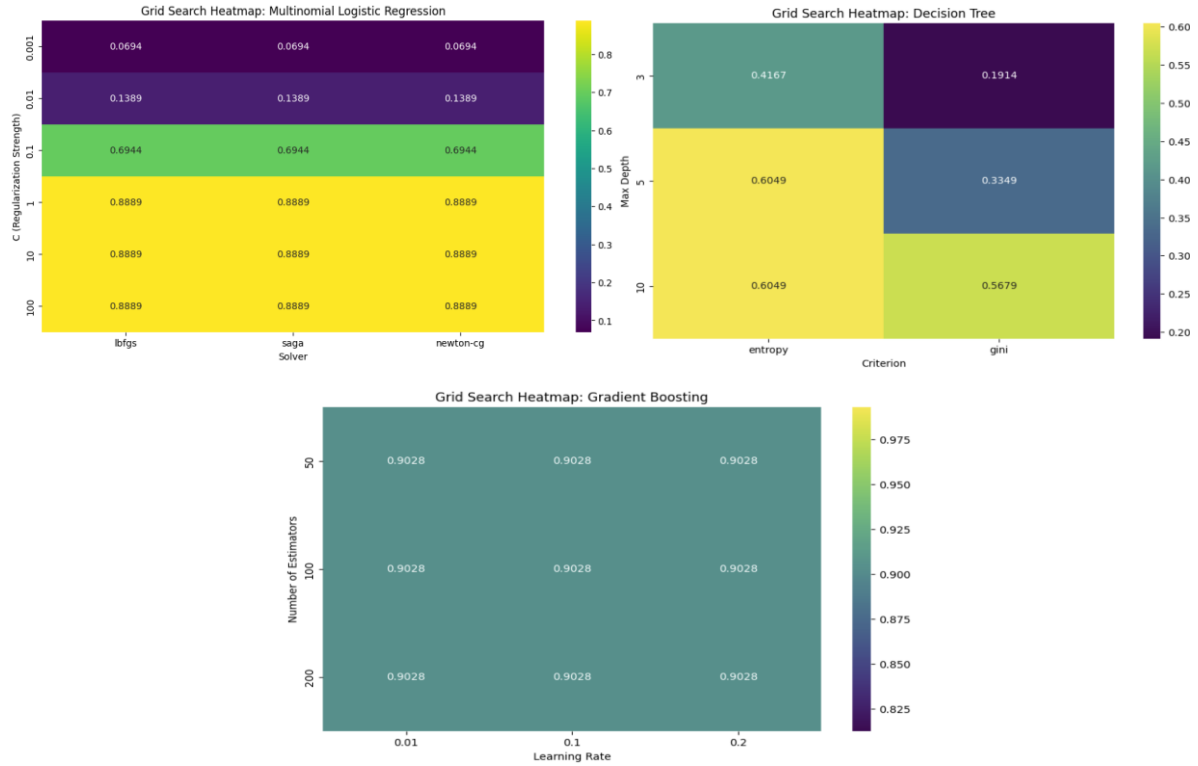


Figure 13. *Grid Search Heatmap for RF, SVM, Multinomial Logistic Regression, Decision Tree, and Gradient Boosting Algorithms.*

Best-Performing Model

Based on the metrics, SVM (C=1, Gamma=1) is the best-performing model with the highest accuracy (88.88%) and F1 score (90.74%). Its high precision (94.44%) indicates minimal false positives, and its high recall (88.88%) ensures robust classification of the cultivar origin. This performance reflects SVM's suitability for handling non-linear relationships in fatty acid data.

GUI of Olive Oil Origin Predictor

The Olive Oil Origin Predictor's graphical user interface (GUI) is designed to classify the origin of mono-cultivar Albanian olive oil samples based on their chemical composition. It provides input fields for entering various fatty acid concentrations and oil composition parameters.

The user enters the values, and by clicking the "Predict Origin" button, the application employs a pre-trained machine learning model and a scaler to prepare the inputs and determine the *origin* of the olive oil sample. The predicted region appears below the button in real-time. The interface is designed to be straightforward and user-friendly, dynamically updating the prediction while maintaining responsiveness through computations that run in a separate thread.

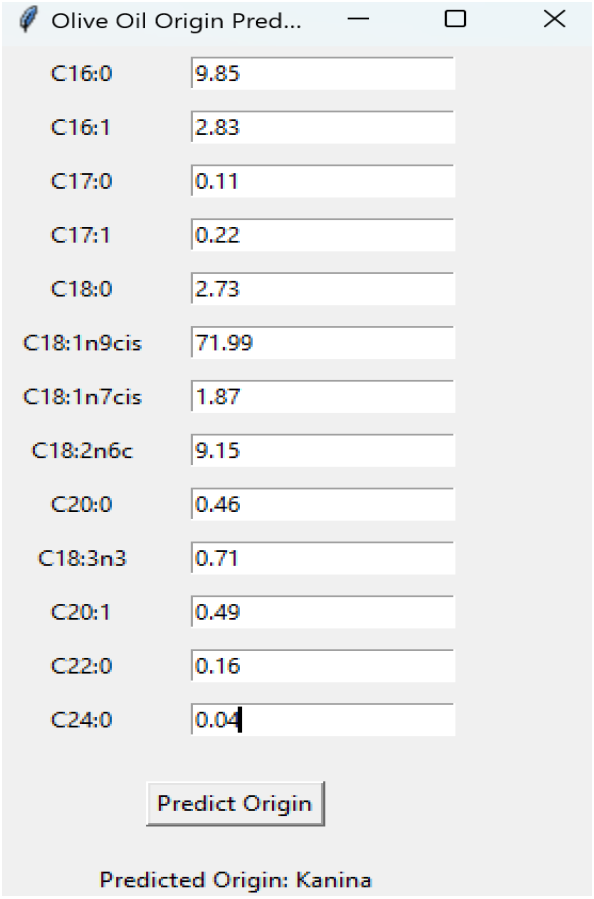


Figure 14. *Implementing the GUI.*

Exploration of dataset #2 ("Cultivar" dataset)

Olive oil's chemical composition varies significantly across cultivars, influencing its nutritional value, stability, and market appeal. So, our additional case study focuses on analyzing the Kalinjot cultivar, a prized olive variety in Albania, examining its fatty acid profile, and comparing it to other local cultivars.

16:00	16:1(n-9)	16:1(n-7)	17:00	17:1 (n-7)	18:00	18:1(n-9)ci	18:1(n-7)	18:2(n-6)tr	18:2 (n-6)c	20:00	18:3 (n-3)	20:1 (n-9)	22:00	Target
10.92	0.09	0.48	0.04	0.14	2.31	75.11	1.88	0	7.56	0.36	0.72	0.31	0.08	Kalinjot (Mallakastra)
10.88	0.07	0.35	0.13	0.19	2.83	74.61	1.53	0	8	0.43	0.58	0.28	0.07	Bardhi Tirana
12.38	0.08	0.51	0.12	0.17	2.88	71.91	1.76	0.01	8.77	0.46	0.51	0.25	0.18	Mixan
10.41	0.13	0.61	0	0	2.1	76.16	2.2	0	6.92	0.4	0.67	0.33	0.12	Kokerrmadh Berati
12.1	0.15	0.62	0.12	0.24	2.17	66.24	2.19	0	15.19	0.3	0.5	0.19	0.07	Krips Kruje
9.41	0.1	0.26	0.13	0.18	2.99	74.59	1.3	0	9.8	0.4	0.56	0.29	0.03	Kalinjot (Marikaj)

Figure 15. *Showcasing the first 6 data objects of the "Cultivar" dataset (referred to as dataset #2).*

The visualizations below reveal a distinctive biochemical profile for the Kalinjot cultivar compared to others. The boxplot highlights Kalinjot cv. significantly higher and more consistent oleic acid, 18:1(n-9)-cis, content, guided by its reputation for high-quality olive oil.

PCA and t-SNE confirm the cultivar distinctiveness, with its samples clustering tightly in a separate region of the feature space, suggesting homogeneity in fatty acid composition. K-Means clustering reinforces this observation, isolating Kalinjot samples predominantly into one cluster. The correlation heatmap uncovers strong relationships between *oleic acid* and other *fatty acids*, such as 16:0 and 16:1(n-7), which are key in determining oil stability and health benefits. These findings position Kalinjot as a cultivar with unique and desirable traits, supporting its differentiation in scientific and commercial contexts.

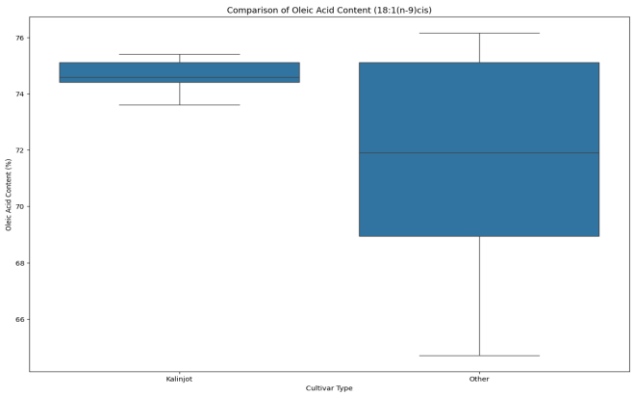


Figure 16. *Comparison of oleic acid content in Kalinjot cultivar and the other remaining cultivars.*

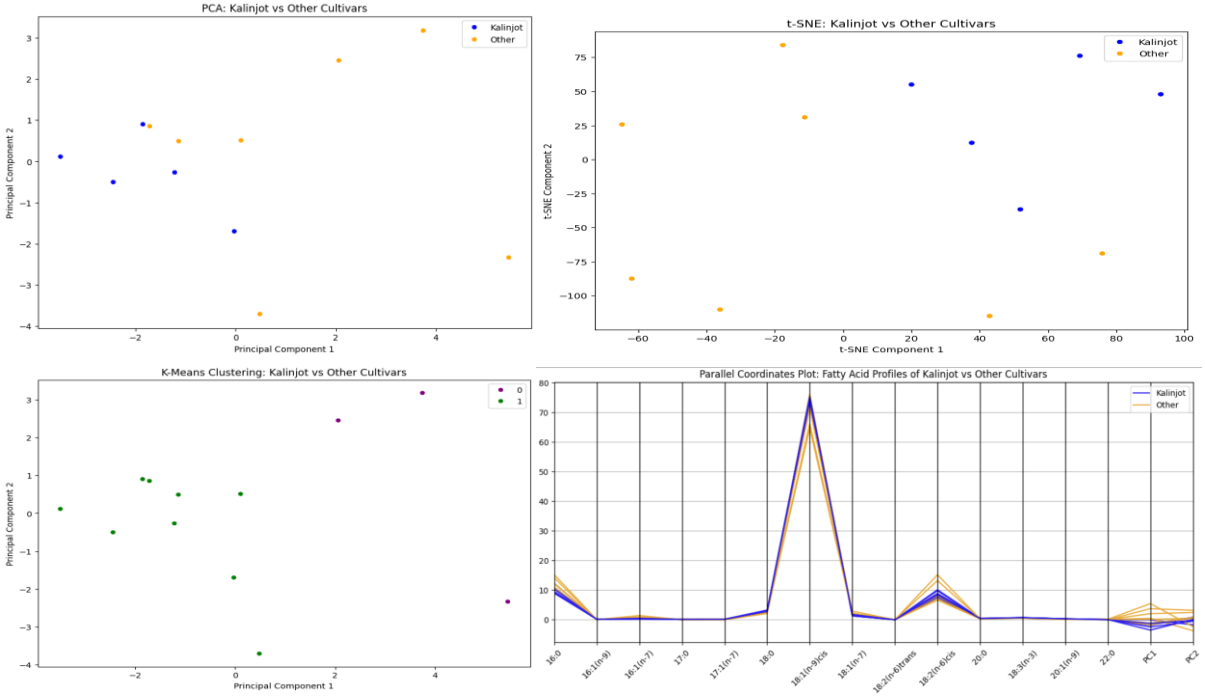


Figure 17. *Comparison of Kalinjot CV vs other cultivars, applying PCA, t-SNE, K-means Clustering and parallel coordinates plot for their fatty acid profiles.*

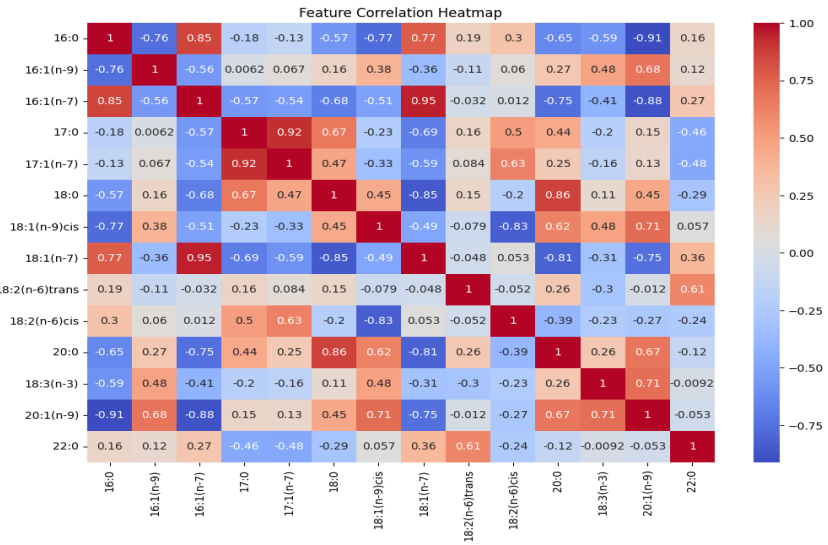


Figure 18. *Feature Correlation Heatmap (involving all 14 features considered in dataset #2).*

Following the procedure demonstrated in the research methodology (Figure 3) similar to our first dataset, dataset #2 illustrates a high level of distinguishability between cultivars, as evidenced by the perfect classification performance of several models, including *SVM*, *kNN*, and *Decision Tree*, achieving 100% accuracy, precision, recall, and F1 score (Table). This indicates that the dataset is highly separable, likely due to distinct fatty acid profiles that effectively discriminate between cultivars. However, such results also raise concerns about potential *overfitting*, particularly for kNN (with $k = 1$) and Decision Tree, as these algorithms are prone to memorizing the training data rather than generalizing it to unseen samples.

On the other hand, we observed that ensemble models like *Random Forest* and *Gradient Boosting* displayed moderate performance with 80% scores across all metrics. This is likely due to hyperparameter settings, such as the limited tree depth in Random Forest and a small learning rate in Gradient Boosting, which prioritize generalization overfitting the training data. *Logistic Regression* also achieved an 80% accuracy, reflecting its simplicity and the potential linear separability of the dataset.

Table 9. Algorithmic Models and Evaluation metrics of different mono-cultivar OOs.

Algorithmic model	**Evaluation Metrics**			
	Accuracy	**Precision score**	**Recall**	**F1 score**
Logistic Regression *(C=1)*	80%	80%	80%	80%
SVM *(C=1, gamma=1)*	100%	100%	100%	100%
kNN *(k=1)*	100%	100%	100%	100%
Decision Tree	100%	100%	100%	100%
Random Forest *(Max_depth=3, N_estimators=25)*	80%	80%	80%	80%
Gradient Boosting *(Learning_rate=0.01, N_estimators=50)*	80%	80%	80%	80%

Note: Evaluation metrics using the same algorithmic models as the first dataset.

Future Dataset Expansion

Ensemble methods, while less accurate here, may offer better performance in real-world applications with noisy or imbalanced data. Future research could focus on validating these findings using techniques like cross-validation, predicting the quality of the olive oil, or even exploring additional features that might influence cultivar differentiation.

The dataset utilized in this research is currently small, consisting of limited samples from the Berat, Fier, and Vlora regions. However, as the study progresses, we anticipate a significant increase in the dataset size. This expansion will result from gathering more samples, potentially covering additional regions, and incorporating seasonal variations. We suppose that the RF algorithm is particularly suited for this type of expansion as it can scale effectively and manage large datasets without sacrificing performance. Furthermore, as the dataset grows, the model's capacity to generalize and provide precise predictions will be enhanced, leading to improved accuracy and reliability in determining the origin of OO.

CONCLUSIONS

By transforming the input characteristics of olive oil, as discussed during development, into predictions about its geographical origin, we can validate the product's authenticity and minimize the risk of adulteration. This process is crucial for building consumer trust and improving food safety standards. Furthermore, identifying the origin of the olive oil—be it a specific region in Albania or elsewhere—enhances labeling transparency and aids in preventing consumer deception related to the quality standards associated with specific origins.

The research results reveal that the machine learning algorithms applied, including Logistic Regression, kNN, and SVM, have demonstrated high accuracy in determining the origins of olive oil, evidenced by strong performance metrics such as accuracy, precision, and recall. This improvement enhances "True Positive" rates for accurately

identifying origin classes. Additionally, our findings underscore the superior oleic acid content and cohesive chemical profile of the Kalinjot cv., marking it as a high-quality cultivar.

By employing these accurate classifications, the model enhances quality control efforts and supports the efficient traceability of Albanian mono-cultivar olive oils. Ultimately, this study aims to help the olive oil industry ensure product authenticity, improve quality assurance protocols, and protect consumers from fraud.

REFERENCES

[1] Aiello, G. (2024). An Artificial Intelligence-based tool to predict "unhealthy" wine and olive oil, *Journal of Agriculture and Food Research*, 16, 101179, at https://doi.org/10.1016/j.jafr.2024.101179

[2] Aminu, S., Eviwiekpaefe, A.E., Abdulhadi, Y., Garba, U., & Zainab, Y. (2022). Evaluation of Some Selected Breast Cancer Classification Algorithms in Nigeria. *Journal of Biochemistry, Microbiology and Biotechnology*. 10. 29-34. Doi: 10.54987/jobimb.v10i2.754.

[3] At https://scikit-learn.org/1.5/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[4] At https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[5] Boskou, D., Blekas, G. & Tsimidou, M. (2006). Olive Oil Composition, in: Olive Oil Chemistry and Technology 2nd Edition (Ed. Dimitrios Boskou), AOCS Press, Champaign, Illinois, USA, pp.41-72.

[6] Clarin, J.A. (2022). Comparison of the Performance of Several Regression Algorithms in Predicting the Quality of White Wine in WEKA. *International Journal of Emerging Technology and Advanced Engineering*, 12(7), 20-26, DOI: 10.46338/ijetae0722_03.

[7] Da Costa, N.L., Valentin, L.A., Castro, I.A., Barbosa, R.M. (2021). Predictive modeling for wine authenticity using a machine learning approach. *Artificial Intelligence in Agriculture*, 5, 157-162.

[8] Hayati, M., Mutmainah, S. & Ghufran, S. (2021). Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification. International Journal of Artificial Intelligence Research. 4. 86. 10.29099/ijair.v4i2.152.

[9] International Olive Council (IOC), (1996). World Olive Encyclopaedia; International Olive Oil Council: Madrid, Spain, p. 488. ISBN 9788401618819.

[10] Kothawade, R. D. (2021). Wine Quality Prediction Model Using Machine Learning Techniques, Retrieved at https://www.diva-portal.org/smash/get/diva2:1574730/FULLTEXT01.pdf

[11] Nanou, E., Pliatsika, N., & Couris, S. (2023). Rapid Authentication and Detection of Olive Oil Adulteration Using Laser-Induced Breakdown Spectroscopy, *Molecules*, 28(24), 7960, Retrieved at https://doi.org/10.3390/molecules28247960

[12] Niyogisubizo, J., Ninteretse, J. de D. ., Nziyumva, E., Nshimiyimana, M., Murwanashyaka, E. & Habiyakare, E. (2024). Towards Predicting the Quality of Red Wine Using Novel Machine Learning Methods for Classification, Data Visualization and Analysis. *Artificial Intelligence and Applications*, 0(0), 1-12, Retrieved at https://doi.org/10.47852/bonviewAIA42021999

[13] Ordukaya, E. & Karlik, B. (2017). Quality Control of Olive Oils Using Machine Learning and Electronic Nose. *Journal of Food Quality*, 9272404, 7, at https://doi.org/10.1155/2017/9272404

[14] Patkar, G. S. & Balaganesh, D. (2021). Smart Agri Wine: An Artificial Intelligence Approach to Predict Wine Quality. *Journal of Computer Science*, 17(11), 1099-1103, Retrieved at https://doi.org/10.3844/jcssp.2021.1099.1103

[15] Ranaweera, R.K., Osmond, *G.*, Gilmore, A.M., Capone, D.L., Bastian, S.E.P., Jeffery, D.W. (2021), Authenticating the geographical origin of wine using fluorescence spectroscopy and machine learning. IVES Conference Series, Infovine, 2021.

[16] Retrieved at https://www.geeksforgeeks.org/k-nearest-neighbours/

[17] Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2020). Machine Learning Applications for Precision Agriculture: A Comprehensive Review. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.3048415.

[18] Sheth, V., Tripathi, U. & Sharma, A. (2022), A Comparative Analysis of Machine Learning Algorithms for Classification Purpose, *Procedia Computer Science*, 215, 422-431, doi: 10.1016/j.procs.2022.12.044.

[19] Topi, D., Amanpour, A., Kelebek, H., & Selli, S. (2019). Screening of aroma profiles in Albanian cvs. *Kalinjot* and *Bardhi Tirana* olive oils using purge and trap extraction technique. *RISG Rivista Italiana Sostanze Grasse*, 96(2), 101-108.

[20] Topi, D., Guclu, G., Kelebek, H., & Selli, S. (2020). Comparative elucidation of phenolic compounds in Albanian olive oils using LC-DAD-ESI-MS/MS. *Journal of Liquid Chromatography & Related Technologies*, 43(5–6), 203–212. doi:10.1080/10826076.2019.1711117.

[21] Topi, D., Guclu, G., Kelebek, H., & Selli, S. (2021). Olive Oil Production in Albania, Chemical Characterization, and Authenticity. In Olive Oil-New Perspectives and Applications (Eds. Akram, M.). IntechOpen. Rijeka, Croatia. doi:10.5772/intechopen.96861.

[22] Topi, D., Risto, J. & Osmani-Lataj, L. (2025). Evaluating Olive Oil Quality and Food Safety Compliance in Tirana, Albania, Retail Markets. *South Eastern European Journal of Public Health*. SEEJPH-418.

[23] Topi, D., Thomaj, F., Halimi, E. (2012): Virgin Olive Oil Production from The Major Olive Varieties in Albania. *Agriculture and Forestry*, 58(2), 87-95.

[24] Topi, D., Topi, A., Guclu, G., Selli, S., Uzlasir, T. & Kelebek, H. (2024). Targeted analysis for the detection of phenolics and authentication of Albanian wines using LC-DAD/ESI–MS/MS combined with chemometric tools. *Heliyon*, 10, 11, e31127.

[25] University of Texas-UTEXAS (1990). Map of Land Use in Albania. Retrieved at http://www.lib.utexas.edu/maps/atlas_east_europe/albania-landuse.jpg

[26] Vega-Márquez, B., Nepomuceno-Chamorro, I., Jurado-Campos, N. &Rubio-Escudero, C. (2020). Deep Learning Techniques to Improve the Performance of Olive Oil Classification. Frontiers in Chemistry. 7, 929. doi: 10.3389/fchem.2019.00929

[27] Velo, S. & Topi, D. (2015). Study of Kalinjoti Extra Virgin Olive Oils, Fatty Acids Profiles, and trans-Isomers. *Journal of Hygienic Engineering and Design*, 12, 129-133.

[28] Velo, S. & Topi, D. (2017). Characterization of Kalinjot and Nisioti Monocultivar Virgin Olive Oils produced in Albania. *Asian Journal of Chemistry*, 29(6), 1347-1350.

[29] Yakar, Y. & Karada, K. (2022). *Identifying Olive Oil Fraud and Adulteration Using Machine Learning* Algorithms. *Química Nova*, 45 (10), Retrieved at http://dx.doi.org/10.21577/0100-4042.20170948

[30] Zaza, S., Atemkeng, M., Hamlomo, S. (2024). Wine Feature Importance and Quality Prediction: A Comparative Study of Machine Learning Algorithms with Unbalanced Data. In: Tchakounte, F., Atemkeng, M., Rajagopalan, R.P. (eds) Safe, Secure, Ethical, Responsible Technologies and Emerging Applications. Retrieved at https://doi.org/10.1007/978-3-031-56396-6_20