Extra Virgin Olive Oil Price Prediction from Multi-source Variables and Machine Learning

Juan J. Cubillas
Dept. Information and Communication Technologies applied to Education.
International University of La Rioja
Logroño, Spain
e-mail: {juanjose.cubillas}@unir.net
Ángel Calle
Dept. Computer Science.
University of Jaen
Jaen, Spain
e-mail: {acalle}@ujaen.es
M.Isabel Ramos, Ruth Córdoba
Dept. Cartographic, Geodetic and Photogrammetric Engineering.
University of Jaen
Jaen, Spain
e-mail: {miramos}@ujaen.es

***Abstract*—This research underscores the vital need for accurate Extra Virgin Olive Oil (EVOO) price prediction, especially in Andalusia, Spain, given its significant economic and social impact on inflation, trade, and stability. Anticipating price fluctuations benefits producers, distributors, consumers, and governments for improved planning. The complexity arises from diverse influencing factors like climate, global markets, energy costs, and policies, highlighted by recent price surges due to adverse conditions. The study aims to develop a Machine Learning (ML) approach using historical and current data from official sources, processed with ML algorithms and Oracle Data Mining. The promising results demonstrate the feasibility of enhancing prediction accuracy, potentially stabilizing markets, optimizing distribution, and improving agricultural budgeting. Furthermore, this work contributes to advancing predictive modeling research within the agricultural sector.**

Keywords-EVOO Price; Machine Learning Algorithms; Multi-source Data.

I. INTRODUCTION

The close relationship between the economy and the food industry is evidenced by macroeconomic indicators that directly affect the food supply chain, and vice versa, fluctuations in food prices influence price stability and purchasing power, in turn affecting macroeconomic indicators through inflation. In addition, recent global events such as the pandemic, the war in Ukraine and climate change have generated significant disruptions in global fuel and food prices, underlining the critical importance of food stability for economies and societies [1] and [2]. Predicting food prices is a crucial economic objective, as fluctuations affect inflation, trade and economic stability. Forecasting stabilises markets, enables informed decisions for producers and consumers, and facilitates the formulation of government policies on trade, subsidies and food security. It also helps mitigate food crises and plan distribution in emergencies, allowing consumers to manage their budgets.

Predictive modelling, an application of Machine Learning (ML), is revolutionizing price prediction and economic behaviour. Using algorithms and historical data, these models identify patterns and make predictions without explicit programming, applying to a wide range of commodity prices [3], [4]. Generally, the price of food is directly related to crop production and the behavior of markets. Specifically, these factors are weather and climate behaviors, global trade of commodities, market trends and speculation, energy and phytosanitary prices, government policies, and even natural disasters or international conflicts. The impact of ML techniques for price forecasting in different types of food is widely represented in the literature [5], [6]. The increase in olive oil prices is attributed to a combination of complex factors. Adverse weather conditions and declining crop yields are primary causes. Added to this are high energy costs, market speculation, low stock levels and disruptions caused by the Russian-Ukrainian war.

In addition, there is a change in consumer behaviour, with consumers showing an increasing preference for healthier fats, strengthening the demand for Extra Virgin Olive Oil (EVOO), which is recognised for its beneficial properties. This trend suggests that consumers are willing to pay a premium price for a high quality product with functional benefits [7]. Olive oil price prediction has already been studied in the literature using soft computing techniques [8]. However, ML and deep learning techniques are currently the most widely used. Most of these methodologies use regression, the supervised learning technique used to understand the relationship between one dependent variable (olive oil price) and one or more independent variables (historical price series, weather, fuel prices, etc.).

This study proposes a ML approach to predict the price of EVOO, incorporating key variables identified in the literature. Time series of olive oil prices from Spanish and international

markets are used, together with prices of other vegetable fats. Energy prices, especially fuels, and critical climatic factors such as drought are also included. The resulting dataset is processed with Oracle Data Mining, where various ML algorithms are evaluated. The objective is to approximate the function that relates the input variables to the price of EVOO. The study addresses the prediction of the price of EVOO in Jaén, highlighting in Section I its economic importance and complexity due to multiple factors, and proposing a ML approach. Section II, Methodology, describes the acquisition of historical data (2009-2023) of economic, climatic and production variables, and the application of several ML algorithms. The Results, Section III, show non-linear relationships and that Gradient Boosting and Random Forest are more accurate in cross-validation. The Conclusions, Section IV, confirm the success of the ML model and the effectiveness of non-linear models in capturing market complexity.

II. METHODOLOGY

A. Data acquisition

This initial phase focuses on obtaining quality data from official web sources. The relevant variables for the model, related to the factors that influence the price of EVOO, are precisely defined. In this case, extensive historical information is sought from 2009 to 2023. The variables considered include economic and agronomic factors:

- *Base price*. The base price of EVOO is obtained from the European Union's olive oil price website, specifically from the API which provides weekly data in JSON format by province, taking Jaén as a reference.This price represents the value of EVOO in the month prior to the calculation of the forecast [9].
- *Month*.Seasonality influences demand and supply. The values of all variables in each of the twelve months are considered.
- *Diesel prise*. The price of diesel and EVOO are interconnected by global economic factors and by the dependence on diesel in agricultural machinery for olive oil production, which implies that an increase in the price of diesel can increase the production costs of EVOO. Data on the average monthly price of diesel in the province of Jaén, obtained from the Spanish Ministry for Ecological Transition and the Demographic Challenge [10].
- *Accumulated rainfall*. The accumulated rainfall during the last 24 months is considered crucial for predicting the price of EVOO, as it directly influences the production, quality and costs of the oil, due to the biannual cycle of the olive tree. Data from the Andalusian Agroclimatic Information Network (RIA), which has more than twenty stations in Jaén [11].
- *Average level of reservoirs*. The level of reservoirs has a significant influence on the price of EVOO, as the availability of irrigation water directly affects the quantity and quality of olives. Historical data on Spanish reservoirs, available through the Ministry for Ecological Transition and the Demographic Challenge, allow this relationship to be analysed [12].
- *Consumer Price Index (CPI)*. The Consumer Price Index, CPI, which reflects inflation and directly affects the price of EVOO, as increases in the CPI generate inflationary pressure in agriculture and alter consumer purchasing power, influencing demand. The CPI data, obtained from the National Statistics Institute (INE), allow this economic relationship to be analysed [13].
- *World olive oil production*. The price of EVOO in Spain is closely linked to world prices due to the globalisation of the market and Spain's dominant role as a producer and exporter. Fluctuations in global production, influenced by producing countries, are reflected in Spanish prices, as Spain competes in both local and export markets. World production data are obtained from the International Olive Oil Council (IOC).[14].
- *World production of other types of oil*. Data on the world production of other vegetable oils, obtained from FAOSTAT [15], are essential to understand the price dynamics of EVOO, as these oils are substitutes in the global market. Fluctuations in their prices directly impact the demand and competitiveness of EVOO, especially in key export markets. The price of sunflower oil, within vegetable oils, is crucial due to its strong substitution effect on EVOO, as consumers may opt for one or the other depending on its relative price. Moreover, the interconnectedness of the oil market implies that fluctuations in the price of sunflower oil affect the overall supply and demand dynamics, indirectly influencing EVOO.
- *Early prediction value of olive crop yield*. The olive crop yield is a crucial factor determining the supply of raw material for EVOO. High yields can lower prices due to abundance, while low yields raise prices due to scarcity. It also influences production costs and market strategies, with predictions based on climatic variables and satellite vegetation indices [16].
- *Early prediction value of olive oil production*. This variable provides an early estimate of the quantity of olive oil that will be available on the market, thus capturing the direct relationship between supply and price. This input value is obtained following the workflow described in the article Ramos et al. [16].
- *Price of fertilisers*. The price of fertilisers is a key variable in the prediction of EVOO prices due to its direct impact on production costs and crop yields. Fertiliser prices, obtained from the Ministry of Agriculture, Fisheries and Food (MAPA) and its Price and Market Information Service (SIPMA), influence the health and productivity of olive trees, as well as global economic trends affecting the olive oil sector.

Figure 1 shows the level of influence of the variables considered in the prediction of the EVOO price using a linear regression model. The most relevant variables include the base (historical) EVOO price, the olive crop yield prediction and

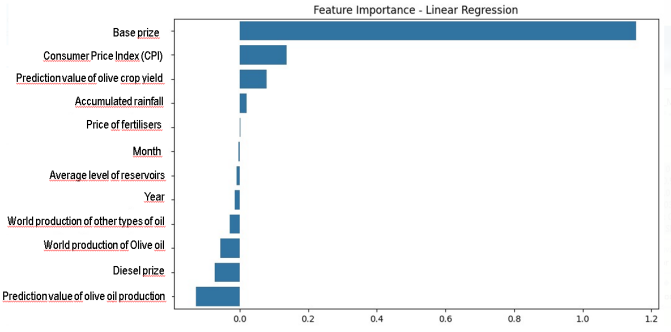

Figure 1. Feature Importance of variables in predicting EVOO price using Linear Regression.

the Consumer Price Index (CPI), which reflect the importance of price history, raw material supply and inflation. Overall, the figure highlights that the model considers both local factors (climate, costs) and global factors (world production, inflation), providing a comprehensive view of the dynamics affecting the price of EVOO.

B. Maching Learning algorithms

Data preparation for ML algorithms is essential to ensure compatibility, capture complex relationships and improve accuracy. This process includes transforming data into numerical, categorical or binary formats, handling outliers and missing values, and applying techniques such as temporal aggregation, spatial selection and seasonal categorisation.

Both linear and non-linear models have been selected in order to consider different types of relationships between attributes and target. As confirmed in the previous section, the variables considered have different influences on the target and even their seasonality is key in the predictive model. In this study, algorithms analysed are: Linear regression, Support Vector Machines, Neural Networks, Random Forest, Gradient Boosting and K-Nearest Neighbors.

III. Results

The attributes considered in this study have different weights on the target attribute and the relationship between them does not follow a linear pattern. The level of accuracy of each of the algorithms used in this study can be analysed from the scatter plot, Figure 2. The figure displays six scatter plots, each evaluating a regression model by comparing actual (x-axis) and predicted (y-axis) values. A dashed red line $y{=}x$ represents perfect prediction. Closer points to this line indicate higher model accuracy, while deviations signify errors. The vertical distance from the line shows the absolute error. The dispersion of points reveals the model's fit (related to R^2, Systematic over or underestimation is visible by points clustering above or below the line. The models' handling of extreme values, like the point near 8, indicates their generalizability. These plots offer a robust visual method for model comparison, outlier identification, and prediction fidelity assessment.

The quality of each model is evaluated using the cross-validation method. This consists of generating a predictive model using data from all months of each year except the month to be tested. Then, the data for that excluded month (the last month of the set of all months of all years) is used to assess the accuracy of the model by comparing the prediction obtained with the actual price data for that month. This procedure is repeated for each month of the historical data set, excluding it from the training set and using it for validation. In this way, the model is tested against the actual value of several months independently. Finally, once the models have been evaluated, a final model is generated using all months of all years as training data. The accuracy of the algorithms varies significantly when predicting the price of EVOO. Linear Regression and Support Vector Machine (SVM) show lower accuracy due to their difficulty in modelling non-linear relationships. Random Forest, Gradient Boosting and K-Nearest Neighbors offer higher accuracy, with Gradient Boosting standing out for its accuracy. Gradient Boosting, when combining weak models and correcting errors, shows the best performance. The Neural Network is also accurate, but inferior to the ensemble models. If we analyse Figure 2 in detail, clearly the value of 8 reached in one of the months could be interpreted as an outlier. However, although it is an outlier, it is a real value which cannot be eliminated. As the volume of training data increases, the model will adjust to these oil price fluctuations.

IV. Conclusions and Future Work

This study developed a ML model to predict the price of EVOO in Jaén, using a wide range of economic, climatic and cost variables. The Gradient Boosting and Random Forest models proved to be the most effective in capturing the complex and non-linear relationships in the market, suggesting that the EVOO market is influenced by multiple interconnected factors. The high accuracy of the models indicates that the input variables adequately reflect the market dynamics in Jaén and that the data sources and data processing were suitable for building predictive models.

References

[1] T. Ulussever, H. M. Ertugrul, S. Kılıç Depren, M. T. Kartal, and O. Depren, "Estimation of Impacts of Global Factors on

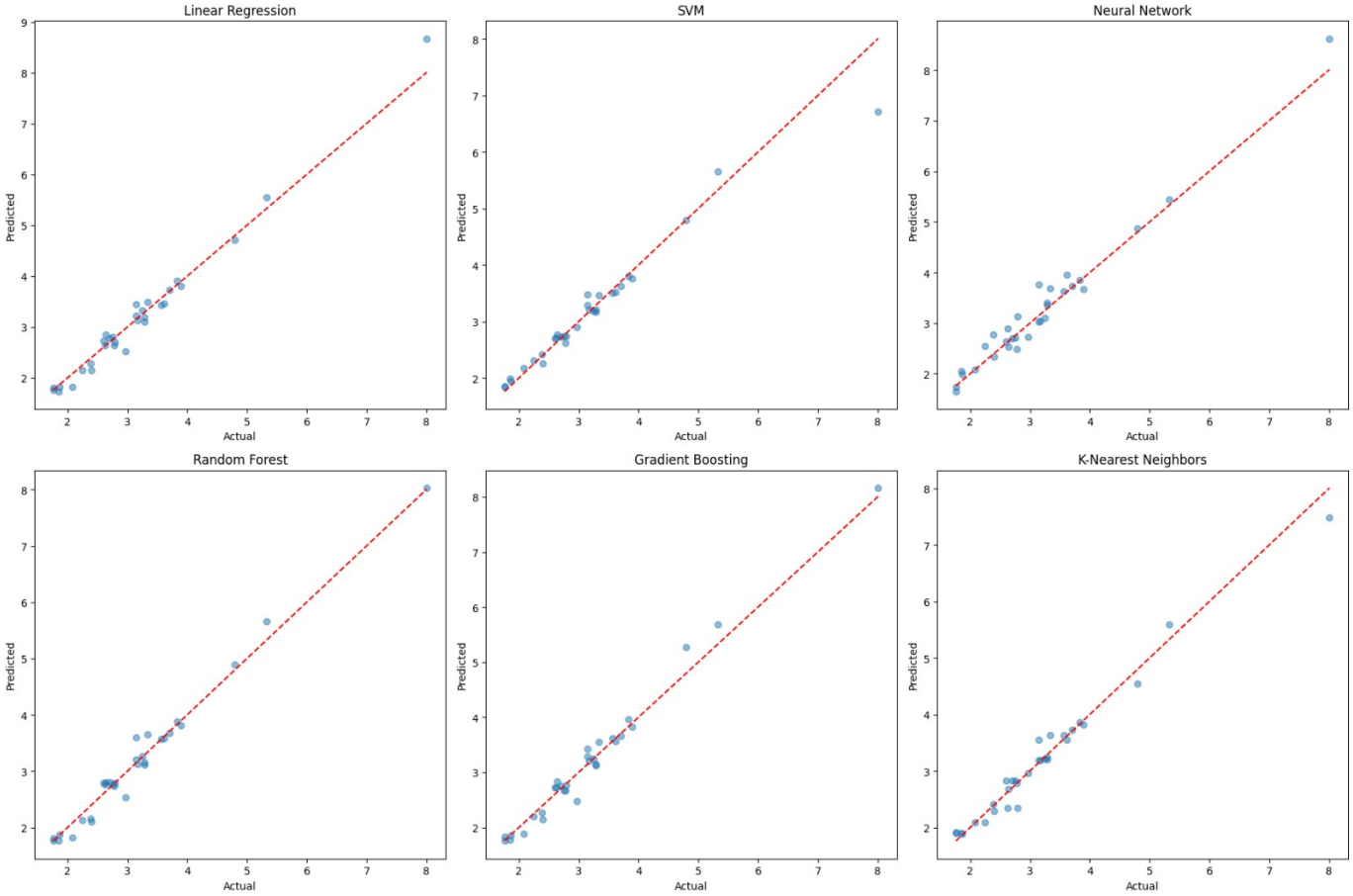

Figure 2. Scatterplots of Actual vs. Predicted Values per each algorithm.

World Food Prices: A Comparison of Machine Learning Algorithms and Time Series Econometric Models," *Foods*, vol. 12, no. 4, 2023, DOI: 10.3390/foods12040873.

[2] E. Breslin, A. Freedman, C. Huston, G. Marrero-Garcia, and T. Mossburg, "Ukraine Food Crisis: Understanding the Impacts of War on the Global Supply Chain and Applying to Future Events," in *2023 Systems and Information Engineering Design Symposium, SIEDS 2023*, Type: Conference paper, 2023, pp. 149–153. DOI: 10.1109/SIEDS58326.2023.10137902.

[3] X. Xu and Y. Zhang, "Price forecasts of ten steel products using Gaussian process regressions," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106 870, 2023, ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2023.106870.

[4] D. Ubilava, "A comparison of multistep commodity price forecasts using direct and iterated smooth transition autoregressive methods," *Agricultural Economics*, vol. 53, no. 5, pp. 687–701, Sep. 2022. DOI: 10.1111/agec.12707.

[5] P. Sharma, P. Dadheech, N. Aneja, and S. Aneja, "Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning," *IEEE Access*, vol. 11, pp. 111 255–111 264, 2023, DOI: 10.1109/ACCESS.2023.3321861.

[6] G. Murugesan and B. Radha, "An extrapolative model for price prediction of crops using hybrid ensemble learning techniques," *International Journal of Advanced Technology and Engineering Exploration*, vol. 10, no. 98, pp. 1–20, 2023, DOI: 10.19101/IJATEE.2021.876382.

[7] R. Zanchini *et al.*, "Eliciting consumers' health consciousness and price-related determinants for polyphenol-enriched olive oil.," *NJAS: Impact in Agricultural and Life Sciences*, vol. 94, no. 1, pp. 47–79, 2022, ISSN: 2768-5241.

[8] A. J. Rivera *et al.*, "A study on the medium-term forecasting using exogenous variable selection of the extra-virgin olive oil with soft computing methods," *Applied Intelligence*, vol. 34, no. 3, pp. 331–346, 2011, DOI: 10.1007/s10489-011-0284-1.

[9] European Comission, *Olive oil prices*, https://agriculture.ec.europa.eu/data-and-analysis/markets/price-data/price-monitoring-sector/olive-oil_en. Accessed Feb. 2025.

[10] Ministry for the ecological transition and the memographic challenge https://www.miteco.gob.es/es/energia/servicios/consultas-de-carburantes.html. Accessed. Feb. 2025.

[11] *Department of Agriculture and Fisheries. Government of Spain*, https://www.mapa.gob.es/es/. Accessed Feb. 2025.

[12] Hydrographic Confederation. Spain https://www.chguadalquivir.es/inicio. Accessed Feb. 2025.

[13] Statistical National Institute. *Spain*, https://www.ine.es/ Accessed Feb. 2025.

[14] International Olive Council, *IOC-Olive Oil dashboard*, https://www.internationaloliveoil.org/. Accessed Feb. 2025.

[15] FAO. United Nations, *FAO*, https://www.fao.org/home/es. Accessed Feb. 2025.

[16] M. I. Ramos, J. J. Cubillas, R. M. Córdoba, and L. M. Ortega, "Improving early prediction of crop yield in Spanish olive groves using satellite imagery and machine learning," en, *PLOS ONE*, vol. 20, no. 1, e0311530, 2025, Publisher: Public Library of Science, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0311530.