Machine Learning in Crop Production: European Tomato Market Case Study

Candidate Numbers: 45450, 46047, 52240, 54454

Submitted for the Master of Science, London School of Economics, University of London

August, 2023

Executive Summary

The convergence of traditional farming practices with cutting-edge technologies is reshaping the agricultural landscape. Machine learning, in particular, stands out as a powerful tool poised to revolutionize crop production forecasting and prediction. Accurate forecasting is vital for food security, resource optimization, and climate change mitigation, particularly within Europe, where agriculture plays a central role in the economy. This study explores the application of machine learning in predicting tomato production in the European context.

Through comprehensive data analysis and predictive modelling, we have unveiled valuable insights into the determinants of tomato production in Europe. Our study not only assesses the suitability of machine learning algorithms for agricultural data but also delves into the significance of demographic and socioeconomic variables. Specifically, we aim to understand how these variables, in addition to agricultural data, can influence tomato production through their impact on demand dynamics.

Our investigation reveals that countries like the Netherlands, Spain, Germany, Poland, Italy, and the United Kingdom are influential in the European tomato trade network. Geographical and policy factors impact real-world tomato trading. Exploratory data analysis highlights Italy's dominance in both gross production value and unit yield, while the Netherlands plays a crucial role in the tomato trade network. Key factors influencing yield differences include machinery usage and export value.

In time series forecasting, LSTM models exhibit predictive power but lack interpretability compared to ARIMAX models. Supervised machine-learning models, particularly kNN, achieve impressive predictive accuracy, with kNN outperforming others in specific contexts. Model performance varies across countries, emphasizing the importance of tailored approaches.

This study underscores the potential of advanced machine learning techniques in agricultural production prediction and suggests tailored strategies for diverse agricultural landscapes and historical periods. Further exploration is recommended for complex agricultural regions, and two historical periods warrant deeper investigation: the 1990s and 2007-2009. Ultimately, this research contributes to sustainable and resilient agricultural practices in Europe.

Contents

List of Figures

List of Tables

1 Introduction

The agricultural landscape is experiencing a significant transformation, driven by the fusion of traditional farming methods with state-of-the-art technologies. Among these transformative technologies, machine learning has emerged as a potent tool with the potential to revolutionize crop production forecasting. Accurate production forecasting is critical for ensuring food security, optimizing resource allocation, and mitigating the impacts of climate change on agriculture. In the European market, where agriculture plays a central role in the economy, this takes on added significance.

This study delves into the realm of machine learning and its practical applications for forecasting tomato crop production in the European market. Utilizing a comprehensive approach encompassing data analysis, modelling, and predictive analytics, our research endeavours to offer valuable insights into the multifaceted determinants of tomato crop production in the European context. Our investigation goes beyond conventional agricultural data, aiming to discern the potential influence of demand-side factors on production. We explore the intricacies of various machine learning algorithms, conducting a rigorous comparison between these modern techniques and traditional models. In addition, our study also delves into the performance evaluation of these models, seeking to identify the most effective approach.

The remaining sections of this report are organized as follows. In Section 2, we provide an overview of relevant literature in the hybrid domain of agriculture and machine learning, highlighting both their strengths and potential weaknesses. Section 3 elaborates on our choice of Europe as the focal region and tomatoes as the subject of our study. Additionally, we detail the data sources we will leverage, present the modelling techniques and associated terminologies, elucidate the mathematical foundations. Section 4 serves as the core of our report. Here, we embark on an exploratory and descriptive journey, starting with an analysis of the European tomato trade network using graph analytics. We then delve into a comparative case study involving selected countries through exploratory data analysis. Subsequently, our focus narrows to Spain, where we aim to develop time series forecasting models. Finally, we employ supervised regression models to predict tomato production, considering a comprehensive array of factors, including climate, economic variables, and chemical data. In the concluding Section 5, we synthesize our findings and offer insights and conclusions drawn.

2 Literature Review

Different studies have already investigated the factors that affect crop production. Agriculture's vulnerability to changing climatic conditions is emphasized in research by Harkness (n.d.), which demonstrates that crops, especially in the UK, are susceptible to variations in weather patterns. These fluctuations impact not only the water availability to plants but also the soil's water balance, consequently affecting overall production. Adding to the complexity of the situation, Goodyear (2023), in a publication at Cambridge University, accentuates pests and diseases as urgent and primary threats to global food production. These challenges are increasingly destabilizing food security and livelihoods, particularly in regions that are more prone to climate change. A new combined drought indicator (CDI) is introduced in Jiménez-Donaire, Tarquis, and Giráldez (2020), which integrates rainfall trends, soil moisture tracking, and vegetation dynamics. The CDI uses a standardized precipitation index (SPI), a bucket-type soil moisture model, and satellite-based normalized difference vegetation index (NDVI) data. The CDI has four levels of increasing severity for drought monitoring. Applied to five grain-growing areas in SW Spain from 2003 to 2013, the CDI's performance was validated by comparing it with observed crop damage data. The results showed a significant correlation between the CDI and actual crop damage, with the new indicator correctly predicting major drought events affecting 70 to 95% of the total insured area. Despite our best efforts to gather the necessary data, we found that the information related to both disease occurrence and soil condition was either scarce or fragmented. The quality of the data we did find was often inconsistent, and in many cases, the data were outdated or lacked the granularity needed for our specific model.

Setayesh, Zadeh, and Bahrak (2022) explores the dynamics of the global trade network. Utilizing exponential random graph models (ERGMs), the study delves into the country-level trade network, considering structural, political, geographical, and economic features. The paper represents a significant contribution to understanding the global trade network from both temporal and static viewpoints, identifying key attributes such as GDP, diplomatic exchanges, distances, and other structural features that shape trade relations. This inspires us to consider economic factors when predicting demand levels and also compare the empirical tomato export network with random versions to verify the existence of unique structural features in the real world.

The arena of forecasting has seen the application of traditional time series methods like ARIMA, ARCH model, GARCH, and exponential smoothing for forecasting. Innovative hybrid models like ARIMA-ANN (Babu and Reddy 2014), segment-level forecasting with ARIMA (Murray, Agard, and Barajas 2018), and other enhancements have come to the fore. The proposed hybrid ARIMA-ANN model with EMD shows great promise in enhancing forecasting accuracy by intelligently blending linear and nonlinear methods and overcoming traditional assumptions. However, it is sensitive to the stationarity of the data and other potential complexities. Concurrently, the use of computational algorithms such as machine learning techniques like SVM, RF, LSTM, and kNN is becoming prominent. Studies exhibit the efficacy of these models in diverse applications, from predicting carbon price variation (Atsalakis 2016) with kNN (Martınez et al. 2019), and electricity load prediction through SVM (Yanhua et al. 2015). Sagheer and Kotb (2019) applied LSTM recurrent networks along with the genetic algorithm to forecast petroleum production. Compared with several standard models, the proposed methodology performs better and obtains accurate production prediction. Shi, Hu, and Zhang (2019) also employed the LSTM algorithm to detect and predict the abnormal change of the temperature in the gyroscope shell. Compared with SVM and BP networks, the performance indicators of the LSTM model are better. Therefore, LSTM could be used in time series forecasting.

Regional and specific studies have emphasized machine learning for crop production prediction. For example, Saadio (2022)'s report on crop yield prediction emphasizes the importance of machine learning in predicting the yield of essential crops like rice, maize, cassava, utilizing decision trees, multivariate logistic regression, and k-nearest neighbour models. The promising results underscore the potential of machine learning in tackling food security challenges in densely populated regions. Similarly, Cubillas et al. (2022)'s case study in Southern Spain demonstrates how predictive systems can aid in the early prediction of crop yield.

In summary, time series models like ARIMA and LSTM approaches tend to perform exceptionally well when ample training data is at our disposal. Regrettably, our pursuit of harnessing these advanced models was thwarted by the unavailability of the necessary database from our client partner. Confronted with this limitation, we found ourselves compelled to explore alternatives, ultimately redirecting our focus toward conventional models like kNN and Random Forest, which are better aligned with our existing constraints. Furthermore, it's noteworthy that traditional agricultural models

typically focus solely on product-level factors such as fertilizer and pesticide usage, as well as weather conditions. In contrast, our study takes into consideration a broader spectrum of demand-related factors, including the Food Production Index, urbanization, and agricultural employment trends. By innovatively incorporating these factors into our analysis, our aim is to present a more comprehensive and holistic portrayal of the economic dynamics surrounding tomatoes in the European market.

3 Materials and methods

3.1 Study Area

The choice of the study area, Europe, for our research, was driven by several key factors. Europe is a diverse and economically significant region with a rich agricultural landscape. As illustrated in Figure 1, lands coloured yellow are cultivated areas. It also encompasses a wide range of climates, from the Mediterranean to temperate zones, making it an ideal candidate for studying the impact of varying environmental conditions on crop production. Additionally, Europe has a well-established agricultural sector with a substantial contribution to global food production and trade. The European agricultural market is highly dynamic and influenced by factors such as climate change and market demand. By focusing on Europe, we aimed to provide valuable insights into the complexities of crop production forecasting in a region that is both economically important and vulnerable to agricultural challenges.

Figure 1: Europe Land Cover

Image Source: European Space Agency (ESA) 2020

3.2 Crop Selection

The selection of tomatoes as the focal crop for our study was a strategic choice driven by several compelling reasons. First and foremost, tomatoes are one of the most widely cultivated and consumed vegetables in the world, with a significant economic impact on the agriculture industry. Their versatility and use in various culinary traditions make tomatoes a staple in European cuisine, further underscoring their importance. Furthermore, tomatoes are sensitive to environmental changes, including temperature

fluctuations and precipitation patterns, which makes them an excellent candidate for studying the effects of climate variability on crop production. Tomatoes are also subject to a range of pests and diseases, adding complexity to their cultivation and making them susceptible to production fluctuations. By selecting tomatoes as our target crop, we aimed to address critical challenges faced by European farmers in ensuring consistent and sustainable tomato production.

3.3 The Dataset

Our primary data is collected from FAO. This is the value of agricultural production for tomatoes in 37 countries, covering all of the European continents, according to the definition by Food and Agriculture Organization (FAO) (2023). The data spans from 1961 to 2021 and is recorded in 1,000 USD (United States Dollars) in constant 2014-2016 dollars. The data quality is categorized as `Estimated (E)`, indicating that the values provided are approximations. Moreover, we calculated tomato unit yield by dividing this value by the total cropland area of the particular country as an additional production value indicator. Then, we gathered additional 10 categories of data from different public sources, mainly from FAO and World Bank, which are grouped into two broad categories and detailed reasons of including them are listed below. (See Appendix A for a full explanation of each feature.)

3.3.1 Agricultural and Environmental Factors (Supply-Side)

Tomato cultivation's success rests heavily on the interplay of agricultural and environmental factors, determining the conditions under which this crop thrives.

1. **Weather (Mean Temperature and Precipitation):** Weather conditions play a significant role in tomato growth and development. Optimal temperatures and sufficient precipitation are essential for germination, flowering, fruit set, and overall crop productivity, but extreme weather like droughts or floods.

2. **Pesticide:** Pesticides are used to control pests and diseases in tomato crops. Effective pest control ensures a stable supply of tomatoes by minimizing damage caused by pests and ensuring healthier plants.

3. **Fertilizer Consumption:** The application of fertilizers to arable land affects soil fertility and nutrient availability. A proper balance of nutrients in fertilizers, such as nitrogen, phosphorus, and potassium, can maximize tomato yield.

4. **Machinery:** The use of machinery, like tractors and irrigation systems, can improve farm efficiency and productivity. Mechanization can help with various farming tasks, such as planting and irrigation, allowing farmers to manage larger areas and increase production.

5. **Agricultural Workforce Indicator:** The availability and efficiency of the workforce can impact the supply of tomatoes. An adequately skilled and sufficient labour force is essential for various farming activities, including planting, harvesting, and maintaining crop health.

3.3.2 Market and Economic Factors (Demand-Side)

Apart from agricultural factors, a comprehensive understanding of the demand-side factors is pivotal, enabling growers to make strategic choices that align with market needs and economic viability, ensuring a stable and responsive tomato industry.

1. **Agriculture, Forestry, and Fishing, Value Added (% of GDP):** The percentage of GDP contributed by the agriculture, forestry, and fishing sectors indicates the economic importance of these activities in a country or region. A higher contribution may suggest a stronger demand for agricultural products and thus, pull up the supplied quantity.

2. **Food Production Index:** The food production index tracks changes in the production of various food commodities over time. A rising food production index indicates increasing food production, including tomatoes.

3. **Total Population:** The total population of a region directly reflects the domestic market size, influencing the demand for tomatoes and other food products. A larger population generally means higher consumption, leading to increased demand and thus supply, to meet the food needs of the population.

4. **Urbanization:** Urbanization represents the proportion of the population living in urban areas compared to those living in rural areas. Urbanization can impact tomato production through reduced available farmland due to land conversion for urban development, increased competition for agricultural land from real estate, and shifts in crop choices to align with urban demand.

5. **Exports and Imports:** The exports and imports of tomatoes significantly influence tomato production. High import value may indicate inadequate domestic

production value to meet the overall demand. High export prices may create incentives for increased domestic production to capture international markets, leading to expanded cultivation.

3.4 Network Analysis Methods

Network analysis is extremely useful when we identify important or influential nodes in a network. Centrality measures and equivalence classes are frequently utilized when deciding the position of a node or the relations between nodes.

Centrality Centrality is often treated as a measure of node property in a network. Several kinds of commonly used centrality include degree centrality, which is divided into in-degree and out-degree, betweenness centrality, closeness centrality, eigenvector centrality and PageRank centrality. These measures are listed and explained in the following table (table 1).

Centrality	**Calculation**	**Note**
In-degree	$k_i^{in} = \sum_{i=1}^{n} A_{ij}$	The number of adjacent edges with direction **'in'**
Out-degree	$k_i{}^{out} = \sum_{j=1}^{n} A_{ij}$	The number of adjacent edges with direction **'out'**
Eigenvector	$x_i = \frac{1}{\kappa_1} \sum_j A_{ij} x_j$, κ_1 is the largest eigenvalue of the adjacency matrix	Gives a vertex a **score proportional** to its neighbours
PageRank	$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta$, α and β are constants	Deals with nodes that have only **outgoing** edges (by adding a small constant), and weights centrality by out-degree
Closeness	$C_i = \frac{n}{\sum_j d_{ij}}$, d_{ij} is the shortest path (geodesic path) between nodes i and j	Measures the **mean distance** to all other vertices
Betweenness	$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}}$, g_{st} is the total No. of geodesic paths between nodes s and t and n_{st}^i is the No. of those passing through node i	Measures how often a node **lies on the geodesic path** between other nodes

Table 1: Centrality

Structural Equivalence Two nodes with the same centrality may occupy different positions. Structural equivalence could be used to identify the **set of relations** that comprise a role, the **set of individuals** who occupy a similar position and classify **equivalence classes** containing sets of nodes that are equivalent. **Dendrogram** plot is extremely important in network analysis because it is able to show equivalence classes clearly.

Properties of Network Reciprocity and transitivity are often used to determine the relationship between nodes.

- **Reciprocity** The fraction of edges that are reciprocated, ranging from 0-1.
- **Transitivity** If i $\leftrightarrow$ j and k $\leftrightarrow$ j, then i $\leftrightarrow$ j, known as the clustering coefficient.
- **Average Path Length** Larger average path length always contributes to less efficiency of a network.

3.5 Time Series Forecasting Models

Time series forecasting is a technique for predicting future values of a variable based on its past data, considering the patterns and trends it exhibits over time. This method has broad applications, from finance and economics to weather and sales forecasting. By analyzing historical data, recognizing patterns, and understanding seasonality and trends, time series forecasting assists organizations in making informed decisions more efficiently and accurately.

$$\hat{Y}_{t+h} = f(Y_t, Y_{t-1}, Y_{t-2}, \ldots, Y_1) \tag{1}$$

3.5.1 Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX)

Time series analysis plays an important role in predicting outcomes across fields, such as economics and environmental science. While traditional methods, like ARIMA (Autoregressive Integrated Moving Average), have been widely used for analyzing time series data, it's important to recognize that real-world scenarios involve interdependencies. Events and variables from one-time series can influence others, and capturing these interactions is essential.

To address this issue the ARIMAX model (Autoregressive Integrated Moving Average

with Exogenous Variables) comes into play. It expands upon the ARIMA framework by incorporating exogenous variables that may impact the dynamics of the time series. By taking into account the relationships among these series, the ARIMAX model provides insights, into how the system behaves. The foundation of the ARIMAX model consists of four key components:

1. **Autoregression (AR):** This element captures the relationship between an observation and a number of lagged observations in the series.
2. **Integration (I):** This component addresses non-stationarity in time series data by differencing, ensuring the series being modelled are stationary.
3. **Moving Average (MA):** This focuses on the relationship between an observation and residual errors from moving average models applied to lagged observations.
4. **Exogenous Variables (X):** This addition incorporates external variables into the model, considering their influence on the time series.

Mathematically, the ARIMAX model can be described as:

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{k=1}^{r} \beta_k x_{tk} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \tag{2}$$

Source: Adapted from Matlab n.d.

In this formulation 2, y_t represents the observed series. The parameters ϕ_i and θ_j correspond to the **Autoregression (AR)** and **Moving Average (MA)** components, respectively. The term ϵ_t denotes the error component, while x_{tk} signifies the exogenous variables with their respective coefficients β_k.

For a more succinct representation using lag operators, the model is given by:

$$\phi(L)y_t = c + x_t'\beta + \theta(L)\epsilon_t \tag{3}$$

Source: Adapted from ibid.

When we combine these elements, ARIMAX not only analyzes patterns within a time series but also considers the impact of external factors. In the context of exploring the Tomato Export Network in Europe, utilizing the ARIMAX model can offer insights

into how various factors and markets impact one another over time, providing an understanding of the trading dynamics.

3.5.2 Long Short-Term Memory (LSTM)

LSTM is a recursive neural network suitable for predicting time series data with long intervals and lags in the real-world, and it is potentially useful in our context. This method was first introduced by Hochreiter and Schmidhuber (1997) and developed by many later scholars.

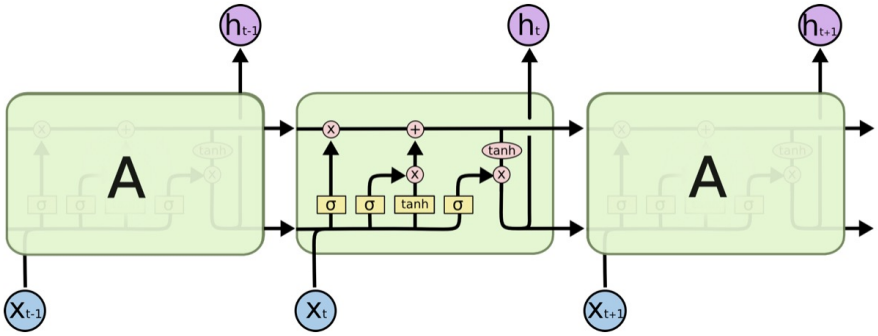


Figure 2: Structure of LSTM Neural Network

Source: Adapted from Colah 2015

Cell state is the key to LSTM which is shown as the cubic cell in Figure 2. The cell state resembles a conveyor belt with only minor linear interactions to ensure the stability of information flow. There are 3 special structures called `gates` to control the ability of removing or adding information to the cell state. The first structure is `forget gate layer` which decides what information we are going to remove from the cell state. It outputs a value ranging from 0 and 1 where 0 represents `information completely remove` and 1 stands for `information completely keep` from the cell state. The second structure including `input gate layer` and `tanh layer` decides what new information to be stored and updated in the cell state. Finally, the cell state is updated and new information is produced.

3.6 Supervised Regression

Supervised regression is a fundamental machine-learning technique used for predicting a continuous numerical outcome based on input features. In this approach, a model is trained on a labelled dataset where each data point consists of both input variables and their corresponding target values. The ultimate goal is to learn a mathematical

relationship that can accurately map the input data to the desired output, allowing for the prediction of new, unseen data points.

3.6.1 K-Nearest Neighbors (kNN)

The K-Nearest Neighbors (kNN) regression algorithm is implemented in this study. kNN is a non-parametric method that relies on the principle that similar observations are close to one another in the feature space. The kNN model makes predictions by averaging the target values of the `k` closest neighbours to a given input. kNN focuses on local similarities between instances. Various values for `k` were explored to find the optimal balance between bias and variance, and the data was normalized to ensure that all features contributed equally to the distance metric. kNN offers a more flexible, localized approach that adapts to the intricacies of the data.

3.6.2 Random Forest and XGBoost

Random Forest stands out as an ensemble learning method that harnesses the collective power of multiple decision trees to yield precise predictions. What distinguishes the Random Forest algorithm is its adeptness at addressing intricacies and complexities in the data. By aggregating insights from numerous decision trees, it excels in generating dependable predictions.

XGBoost, an optimized embodiment of gradient boosting, has risen to prominence as a widely embraced machine-learning technique. It boasts the capability to harness the collective strength of numerous modest learners, resembling decision trees, to deliver precise predictions. Notably, XGBoost excels in handling both linear and non-linear relationships between variables. Furthermore, it incorporates a plethora of regularization techniques to combat overfitting and enhance generalization, rendering it an appealing choice for tasks related to crop production forecasting.

The key distinction between these two methods lies in their approach to ensemble learning. While Random Forest relies on an ensemble of decision trees that operate independently and in parallel, XGBoost takes an iterative, sequential approach to refine the predictions of each weak learner to create a stronger model. This sequential refinement allows XGBoost to excel in capturing both linear and non-linear relationships, making it a powerful choice.

3.6.3 Multilayer Perceptron (MLP)

Adding Multilayer Perceptron (MLP) to our project is like gaining a powerful tool to understand complex patterns in our data. MLP is a type of neural network with at least three layers: one for input, one or more hidden layers, and one for output (Figure 3). Each layer has many nodes that talk to each other using weights, helping us transform data in complex ways. By using activation functions like sigmoid or ReLU, MLP can understand tricky patterns in our data. To make accurate predictions, MLP adjusts its internal settings through a process called backpropagation. This process helps MLP learn from its mistakes and get better at forecasting. MLP is more advanced than the basic perceptron, which can only handle simple problems. MLP can tackle much harder problems thanks to its layers and smart functions. To get even smarter, MLP keeps fine-tuning itself using a method called gradient descent. This helps it adapt to changes and peculiarities in our data over time. That's why MLP is great for forecasting, especially when dealing with tricky time-related patterns. Even with limited data, MLP can still use it effectively to make predictions by finding important connections.

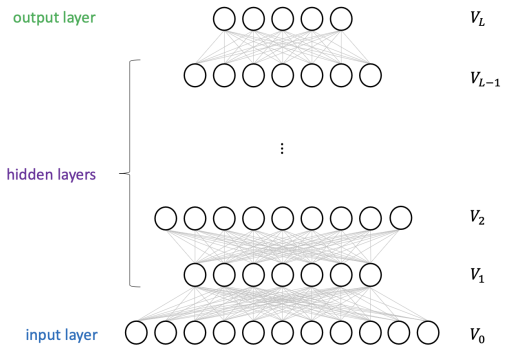


Image Source: Vojnovic 2023

Figure 3: MLP Architecture Illustration

3.7 Evaluation Metrics

During the forecasting and prediction process, we utilized three fundamental metrics to provide a comprehensive evaluation of our models, and we also introduced the coefficient of determination as an additional measure to assess the fitness of our regression.

Mean Squared Error (MSE): MSE is the average of the squared differences between the actual and predicted values and is expressed mathematically as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

Mean Absolute Error (MAE): MAE measures the average absolute differences between actual and predicted values, providing a straightforward assessment of accuracy:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

Root Mean Squared Error (RMSE): RMSE, reported in units such as kg/ha for yield and kg for gross production value, provides a measure of the square root of the average of the squared differences between actual and predicted values. A lower RMSE value indicates better model performance, as it reflects smaller deviations between actual and predicted values:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

Coefficient of Determination: It is often denoted as R-squared (R^2). Unlike RMSE, MAE, and MSE, R^2 measures the proportion of the variance in the dependent variable that is explained by the independent variables. It quantifies the goodness of fit of the model to the data, with higher values indicating a better fit and greater explanatory power. These metrics collectively provided comprehensive insights into the performance and accuracy of our predictive models.

4 Results

4.1 Network Analysis

To gain a better understanding of the European tomato trade network, we focus on European tomato export in 2021. Table 2 is part of the European tomato export trading matrix, `Reporter.Countries` are tomato importers and `Partner.Countries` are tomato exporters. For example, as is shown in the first line of Table 2, Albania exported 649 (1000 USD) tomatoes to Austria in 2021. Based on the European tomato trading matrix, we created a tomato export network to see which countries are influential.

Reporter.Countries	Partner.Countries	Value
Albania	Austria	649
Albania	Bosnia and Herzegovina	2418
Albania	Bulgaria	876
Albania	Croatia	671
Albania	Czechia	32
Albania	Denmark	9

Table 2: Example of European Tomato Trading Matrix

4.1.1 Dendrogram of Tomato Export Network in Europe

In order to explore which countries are in a similar status in the network, we plot the dendrogram and divided them into 4 clusters. As Figure 4 shows, Germany, Italy, Poland, Netherlands and Spain are in the same cluster. Therefore, these 5 countries play a similar role in the European tomato trading network.

4.1.2 Tomato Export Network in Europe

Based on the classification, we plot the directed network as shown in Figure 5, with nodes coloured by their structural classes and the width of the edge representing the export value. It is obvious that Netherlands, Spain, and Germany have wider edges. In order to identify the most influential countries in the trading network. We use PageRank and Betweenness centrality to measure the influence. Betweenness centrality measures the importance of a country in the trading network if it lies on the shortest trading path between other countries. Therefore, we think Betweenness could reflect a

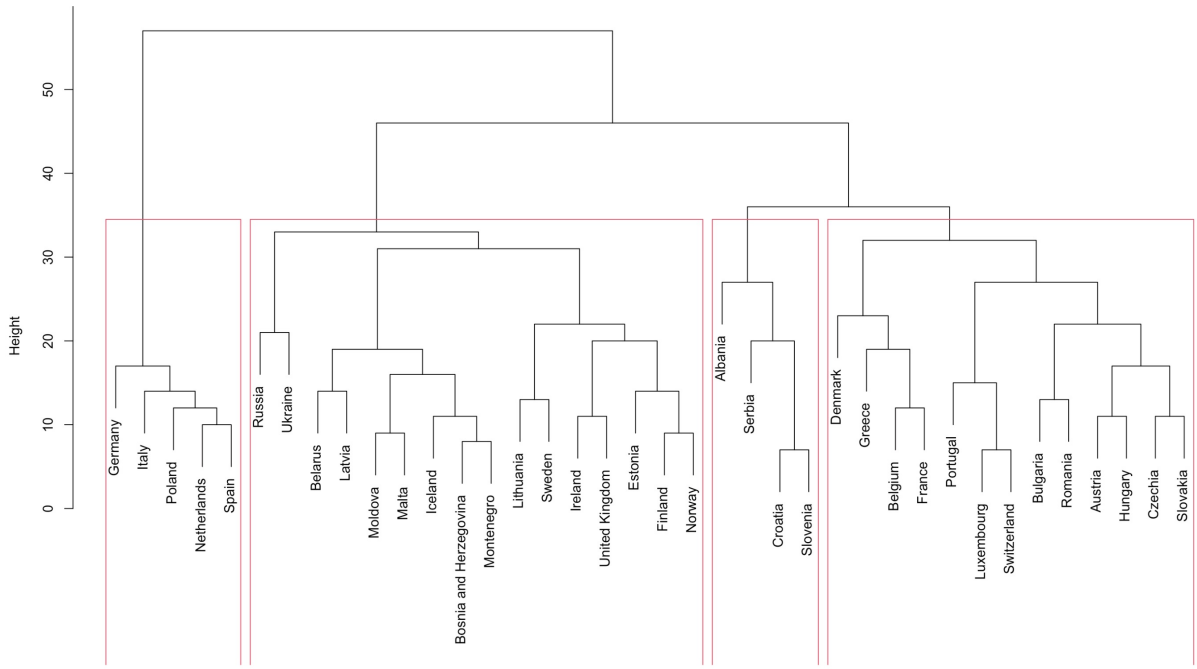


Figure 4: Dendrogram of Tomato Export Network in Europe

country's potential to control or influence the tomato trade flows. PageRank centrality could also measure the importance of a country if it is connected with other influential countries. Therefore if one country is linked with some other influential countries in the trading network, it must also be important in the network.

The overall metrics of the network give some intuition of the nature of the trading of tomatoes. As Table 3 shows, the transitivity of the empirical tomato network is 0.6764, suggesting that if two countries both export tomatoes from another country, there is a 67.64% chance that they also have trading connections but the directionality is uncertain. The reciprocity for the real-world tomato export network is 0.5885, which implies that if one country has exported tomatoes from another, there is a 58.85% possibility that the other country has as well. This means that European countries tend to have long-term tomato export relationships.

To have an insight into the real-world tomato exporting structural properties, we build two random networks including the Erdős–Rényi model and the configuration model based on the empirical network. It is obvious that the empirical network has both higher transitivity and reciprocity compared to both models which are generated randomly, which means that real-world tomato export networks are more clustered than random models. In the real world, countries have a higher chance of being interconnected or building mutual connections. This could be explained by the potential long-term relationship or friendship between countries in the real world. For the average

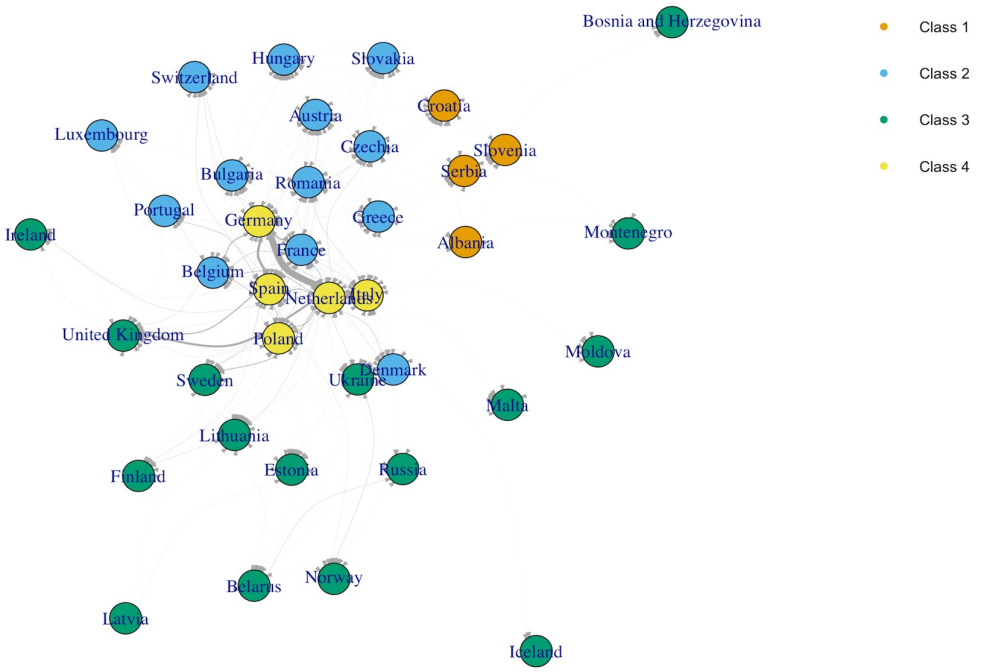


Figure 5: Tomato Export Network Showing Structural Equivalence

path length, the empirical network has a slightly higher average path length than the E-R model but a lower equivalent value compared to the configuration model, which implies that the tomato export network is more efficient in transmitting information than the configuration model but less efficient compared to E-R model.

Overall, these comparisons suggest that empirical networks have some similar characteristics to random models but also have some unique structural features in the real world such as long-term trade relationships among countries, geographical locations, tariffs and potential trading limitations.

Network	Average Path Length	Transitivity	Reciprocity
Empirical Tomato Network	1.7074	0.6764	0.5885
Erdős–Rényi Model	1.6358	0.6029	0.3769
Configuration Model	1.8282	0.5622	0.3147

Table 3: Comparison of Empirical and Random Networks

4.1.3 Top Influential Trading Countries in Europe

Centrality measures including degree, betweenness, and PageRank are frequently used when identifying the most influential nodes in a network. The higher values often imply larger expected influences. To identify which countries play more important roles in

the tomato trading network, we plot the centrality ordered by betweenness which is shown in Figure 6. Besides, we calculated the PageRank and Betweenness values and identified the top 5 countries using these two centrality measures. As shown in Table 4 and 5, the top 5 PageRank centrality countries are the Netherlands, Spain, the United Kingdom, Germany, and Poland. The top 5 Betweenness centrality countries are the Netherlands, Italy, Spain, Poland, and Germany. Therefore, we focus on the tomato market in these 6 countries in later sections.

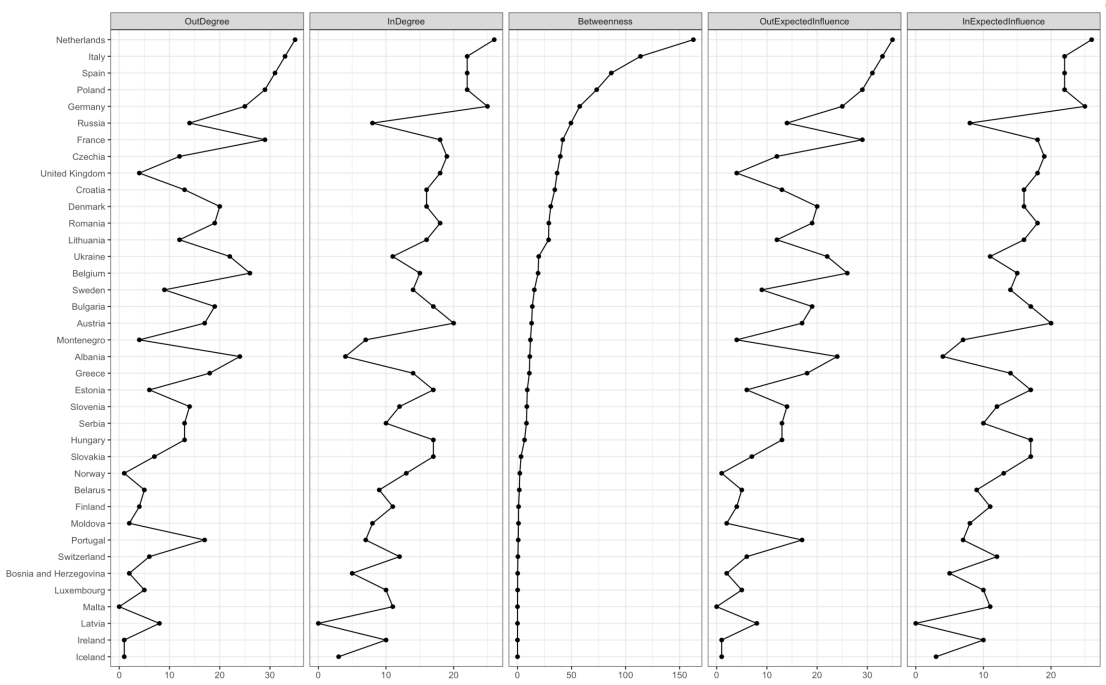


Figure 6: Centrality Plot Ordered by Betweenness

Country	**PageRank**
Netherlands	0.0546
Spain	0.0493
United Kingdom	0.0487
Germany	0.0446
Poland	0.0444

Table 4: Top 5 Countries - PageRank

Country	**Betweenness**
Netherlands	162.73
Italy	113.76
Spain	86.82
Poland	73.27
Germany	57.54

Table 5: Top 5 Countries - Betweenness

4.2 Explanatory Data Analysis

Utilizing the insights gained from PageRank and Betweenness analyses in our previous trade network study, we have singled out six key countries – Italy, Netherlands,

Spain, United Kingdom, Germany, and Poland – for an in-depth case study aimed at enhancing our understanding of the European tomato market. This section comprises two primary segments. Firstly, we focus on yield and total production, highlighting disparities among these countries and identifying potential temporal patterns. Subsequently, we conduct a comprehensive correlation analysis to shed light on the factors influencing country-level variations in both yield and total production.

4.2.1 Overall trend in production and yield

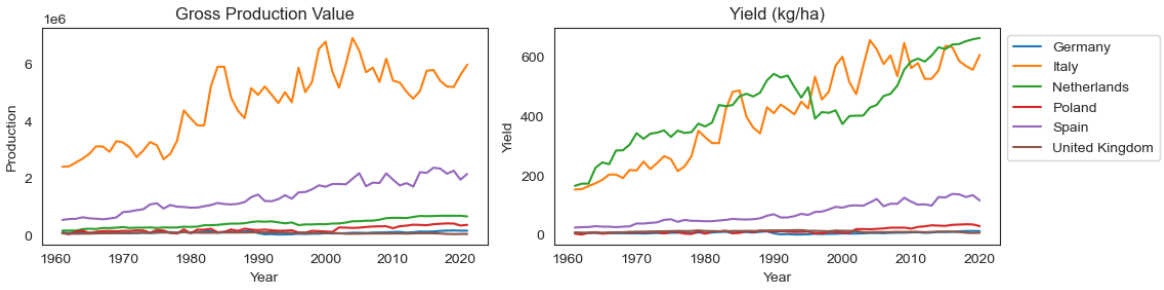


Figure 7: Production and Yield for Tomatoes Plot by Country

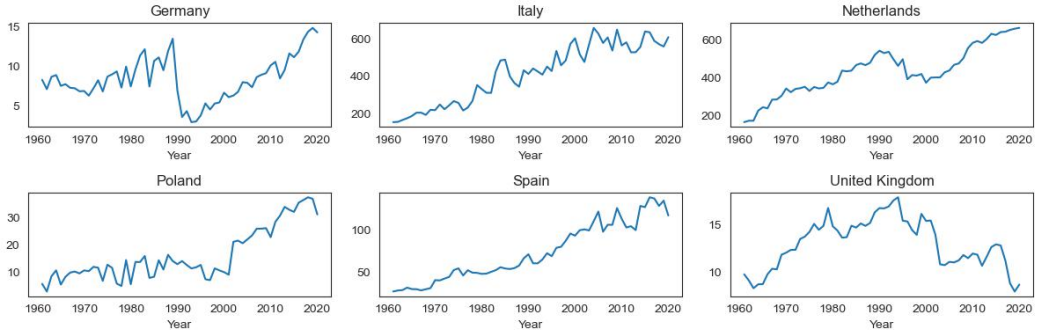


Figure 8: Unit Yield by Country

Figure 7 underscores Italy's preeminence as the largest country in terms of both gross production value and unit yield within the context of tomato cultivation. Although the Netherlands exhibits a relatively lower production value, it remarkably excels in unit yield. This intriguing contrast can primarily be attributed to the Netherlands' strategic pursuit of an agricultural export-oriented approach, which will be elaborated on in a subsequent section (4.2.2.3). Moreover, based on Figure 8, we can further group those countries based on the similarities highlighted:

1. **Poland, Italy, and Spain:** They have displayed an increasing trend. Italy and Spain exhibit a consistent upward trajectory, with Italy showing more fluctuations. Poland initially maintained stability surged around 2000.

Table 6: CAGR Values by Country

Country	**CAGR (%)**
Poland	2.37
Spain	2.31
Netherlands	2.30
Italy	1.53
Germany	0.79
United Kingdom	-0.43

2. **Germany, the United Kingdom, and the Netherlands:** They share a pattern of growth before 1990, followed by a decline. Notably, the Netherlands experienced the least stagnation. While the Netherlands and Germany swiftly entered recovery phases, the United Kingdom's continuous decline has persisted even into recent years.

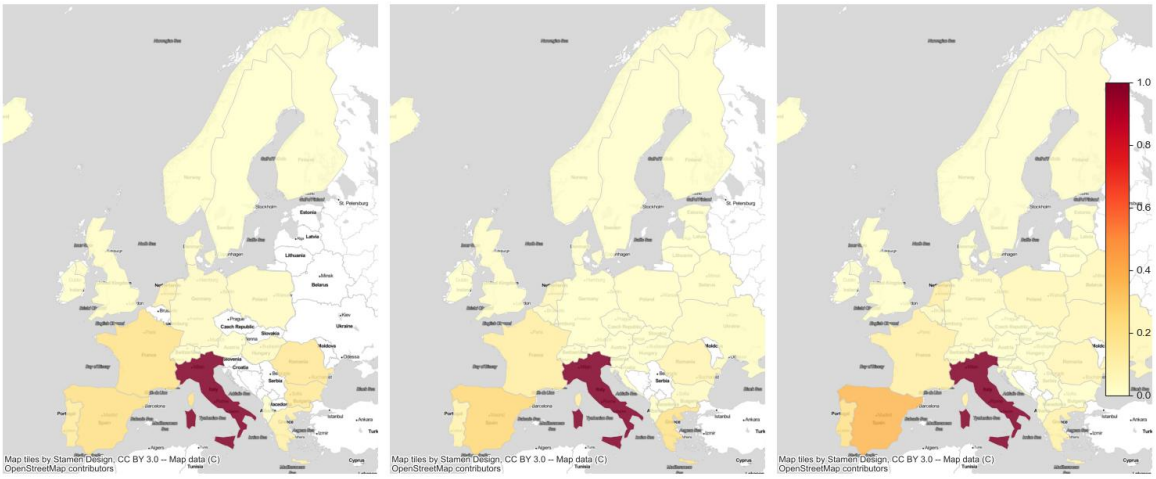


Figure 9: Geographical Visualization of Production for the Year 1965, 2000 and 2021

Figure 9 selects 1965, 2000, and 2021 as exemplary, it is clear from this coloured map that Italy consistently maintains its stature as the top producer within the group, while Spain stands out with a notable increase in production. The Calculated Compound Annual Growth Rate (CAGR) further emphasizes Spain's prominence in growth among the selected countries, standing at an impressive 2.31%, ranking second among six selected countries (Table 6). Moreover, Spain is found to maintain an unyielding position as the second-largest tomato producer among all countries in Europe (in terms of total

production value), solidifying itself as the steadfast runner-up, consistently positioned directly behind Italy in the European tomato production hierarchy. A meticulous time series analysis will be exclusively devoted to Spain in a subsequent section.

4.2.2 Correlation Analysis

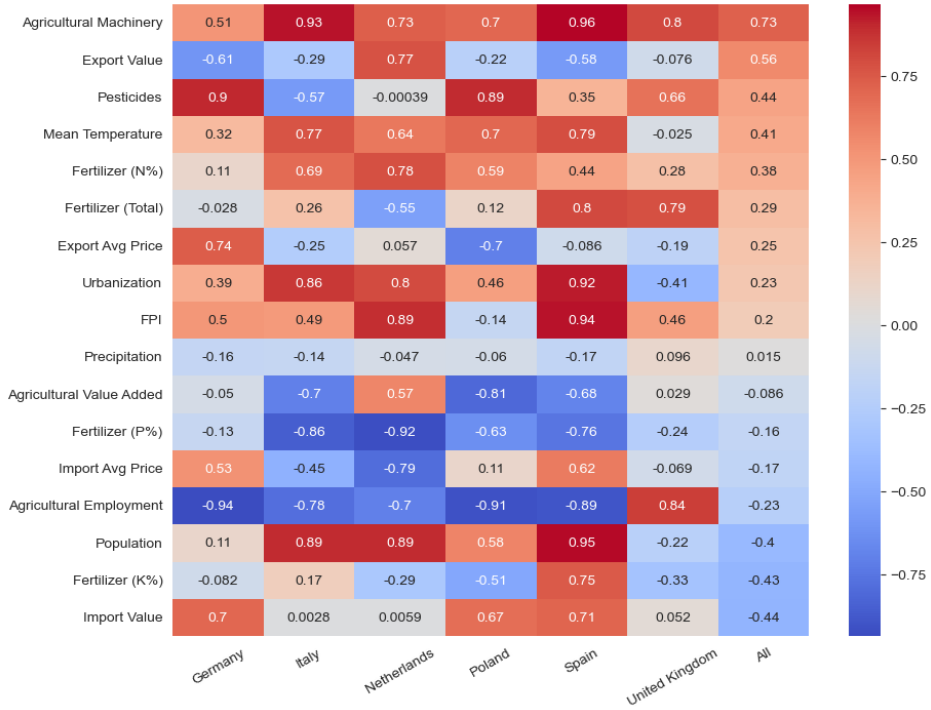


Figure 10: Correlation between Yield and Different Explanatory Variables by Country

4.2.2.1 Climate Factors (Temperature and Precipitation) Figure 10 shows that in most countries, there's a positive link between tomato yield and average temperature. For instance, Italy has a strong connection with a score of 0.77, while Spain's correlation is even higher at 0.79. When you look at all six countries together, there's a significant 0.41 correlation between average temperature and tomato production. On the other hand, there's a small negative relationship (-0.015) between tomato yield and rainfall. A further understanding of the patterns emerges from Figure 11. This visualization underscores that mean temperature exerts a more significant influence on tomato production compared to precipitation based on the fact that both Spain and Italy maintain higher mean temperatures while sustaining moderate levels of precipitation. This lines up with the fact that tomatoes tend to thrive in warmer climates.

4.2.2.2 Agricultural Inputs (Fertilizers and Pesticides) Mixed patterns were observed from both the usage of pesticides and fertilizers. For pesticides, correlations

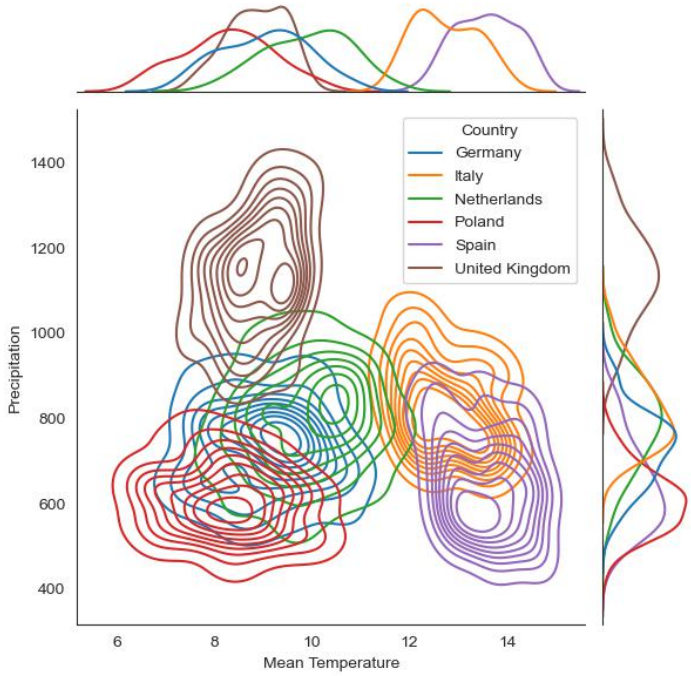


Figure 11: Mean Temperature versus Precipitation

between tomato yield and its usage vary significantly across different countries. In Spain, the correlation between tomato yield and pesticides reaches as high as 0.9, while in Italy, it drops down to -0.57.

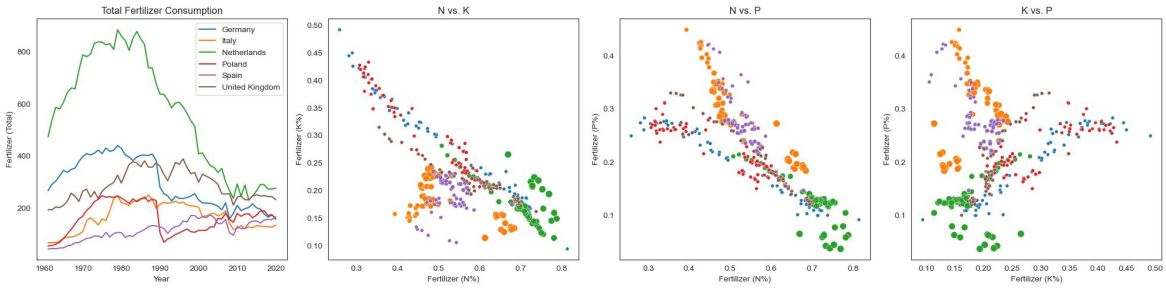


Figure 12: Overall Fertilizer Trend and Comparison Between Nutrition

For fertilizers, it's notable from Figure 12 that there was a significant reduction in fertilizer usage around 1990 across all countries, which is aligning the fact that tomato yield also declined in this period (Figure 8). Among the leading tomato-yielding countries such as the Netherlands and Italy, the Netherlands stands out with the most extensive application of fertilizers. Both countries share a preference for high nitrogen (N) usage and relatively lower utilization of potassium (K), indicating a common trend in nutrient management. However, when examining phosphorus (P), a divergence becomes evident. The Netherlands relies less on phosphorus compared to Italy, highlighting differing strategies in addressing nutritional requirements. Adding to the

complexity, Italy's nutritional consumption data reveals an insightful observation. It shows a dual-cluster distribution, suggesting an evolution in the country's agricultural nutritional practices. This intriguing pattern implies that Italy might be undergoing a transformation, adapting its nutrient application strategies, and potentially realigning its agricultural approach.

4.2.2.3 Exports and Imports Export-related factors also has a significant impact on tomato production. For the Netherlands, an export-oriented strategy finds support through a substantial correlation between the yield and Export Value (0.77) (Figure 10), positioning it as a critical tomato supplier. In contrast, Figure 13 further reveals distinct contrast in the trade strategies among countries. For Germany and the United Kingdom, it is evident that both of them heavily rely on import values, indicating a significant portion of their tomato supply is sourced externally. For Italy, although it is also a high-yielding country, export is significantly lower than the Netherlands, suggesting higher domestic demand and self-sufficiency.

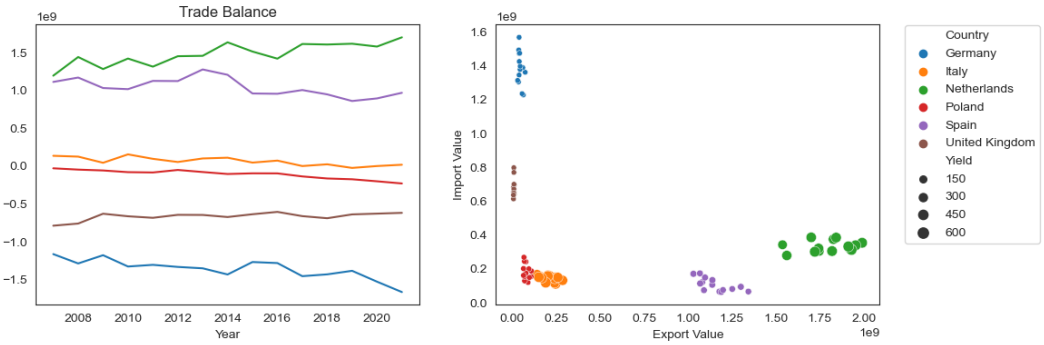


Figure 13: Exports and Imports

4.2.2.4 Urbanization and Population In most countries, positive correlations exist between tomato yield and urbanization, as well as population (Figure 10). The connection is obvious. Urbanization often leads to greater access to markets, which provides farmers with better selling channels. A larger population means greater demand, which in turn prompts farmers to increase their cultivation.

4.2.2.5 Agricultural Employment and Machinery The relationship between tomato yield and agricultural employment is generally negative across countries, which is anti-intuition. However, this could be possibly explained by an overall 0.73 positive correlation between agricultural machinery and tomato production. Modern agriculture has improved farming techniques, which allow farmers to achieve higher yields with

fewer labour inputs and agricultural practices become more efficient and mechanized. Except for the United Kingdom, where a positive correlation is observed (0.84) (Figure 10). One possible explanation is that the United Kingdom is a low-yield country, which may still have traditional agricultural practices where labour-intensive practices or specialized techniques are employed in tomato cultivation.

4.3 Time Series Model

In this dedicated section, we turn our attention to Spain, selecting it as a special case for its relatively comprehensive dataset and the consistent upward trend in tomato production over time. Our primary aim is to uncover temporal patterns and gain valuable insights from the analysis. To achieve this, we have developed a range of time series forecasting models. These include traditional methods like ARIMAX, as well as more contemporary approaches like LSTM. We will thoroughly assess and compare their forecasting performance specifically for Spain.

4.3.1 Exogenous Variables Selection via mRMR

The mRMR (Minimum Redundancy Maximum Relevance) method is an algorithm that systematically identifies variables of utmost relevance while simultaneously avoiding redundant ones. In the context of time series forecasting, the selected variables from mRMR can serve as the most relative variables and can be used as input variables for our LSTM and potential exogenous variables for our ARIMAX model.

The mRMR methodology was applied to several variables related to economic, environmental, and agricultural factors, potentially impacting our dependent variable, **Gross Production Value of Tomato**. An initial inspection of the dataset included variables such as `Total Population`, `FPI`, `Urbanization`, and various agricultural measures like `Cropland (1000 ha)` and specific fertilizer consumption rates. Based on the scores from the mRMR selection process, the variables were ranked:

From Table 7, `Population` emerged as the most influential exogenous variable followed by `FPI` and `Urbanization`. The lesser ranks were occupied by variables like `Mean Temperature` and specific fertilizer components. These scores provide an objective measure of each variable's relevance to the dependent variable.

By leveraging these insights, the first seven variables (`Population, FPI, Urbanization, Cropland (1000 ha), Fertilizer (Total), Mean Temperature, Fertilizer (K2O)`)

Variable	mRMR Score
Population	528.261164
FPI	450.737345
Urbanization	388.863462
Cropland (1000 ha)	329.473779
Fertilizer Total	133.290824
Mean Temperature	91.023472
Fertilizer K2O	63.534910
Fertilizer N	55.585354
Fertilizer P2O5	3.457196
Precipitation	1.983453

Table 7: mRMR Scores for the Selected Exogenous Variables

were initially chosen from the mRMR process for integration into the ARIMAX model. We will next conduct the stationarity test to ensure the reliability of these variables.

4.3.2 Time Series Characteristics

4.3.2.1 Stationality For time series models, ensuring the data is stationary is paramount for enhancing forecasting accuracy and for providing a comprehensive understanding of the influencing factors. A stationary series is one in which its properties do not depend on the time at which the series is observed. If the series is stationary, its mean and variance will remain consistent over time, ensuring the series' structure remains unchanged. Failing to account for stationarity might result in unreliable and spurious results, potentially leading to poor understanding and inaccurate forecasts.

To verify the stationarity of our variables, we initially examined the moving average and standard deviation for each series. As illustrated in Figure 14, the evident variations in both the mean (represented by the red line) and the standard deviation (depicted by the black line) over time signify the presence of a trend effect in the time series, leading to the conclusion that the series are non-stationary.

Further, we proceed with the Augmented Dickey-Fuller (ADF) test, a type of unit root test, to make validation. Due to the nature of our annual data, the potential seasonality might not be a concern. The ADF test equation is represented as:

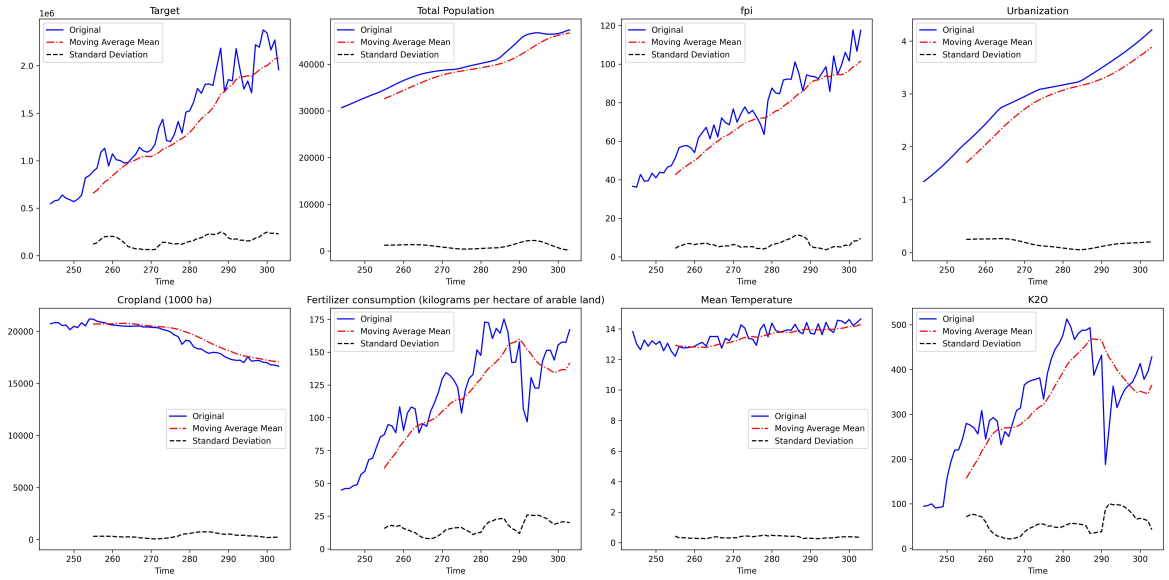


Figure 14: Moving Average and Standard Deviation

$$\Delta y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \Delta Y_{t-1} + \cdots + \phi_p \Delta Y_{t-p} + \varepsilon_t \tag{7}$$

Source: Adapted from Prabhakaran 2019

In the equation 7, y_{t-1} denotes the series lagged by one period, and Δy_{t-p} represents the difference of the series at time $t - p$.

The primary focus is on α - unit root: if it is equal to 0, the series possesses a unit root and is non-stationary; otherwise, it's considered stationary. With the hypothesis test conducted at a 95% confidence level, we will compare the `Test Statistic` with the `Critical Value (5%): -2.915` to determine the stationarity. If the `Test Statistic` is less than the critical value, we reject the null hypothesis and conclude the series is stationary.

In Table 8, the results of the ADF test on the original data series revealed all the series were non-stationary which is consistent with Figure 14, hence, achieving stationarity, we applied a first-order difference to the series and retested.

In Figure 15, the moving average fluctuates around the value of 0 for most of the series except **Total Population** and **Urbanization**, and the variance displays minimal fluctuation over time. This suggests that these differenced series appear to be stationary. Subsequently, to further validate this observation, we employed the Augmented Dickey-Fuller (ADF) test.

As shown in Table 9, all variables reached stationarity at their first-order difference

Variable	Test Statistic	Result
Production	-0.566	Non-stationary
Total Population	-1.236	Non-stationary
FPI	-0.784	Non-stationary
Urbanization	-0.463	Non-stationary
Cropland (1000 ha)	0.78	Non-stationary
Fertilizer (Total)	-1.791	Non-stationary
Mean Temperature	-0.501	Non-stationary
Fertilizer (K2O)	-2.252	Non-stationary

Table 8: Stationarity test results for original data series

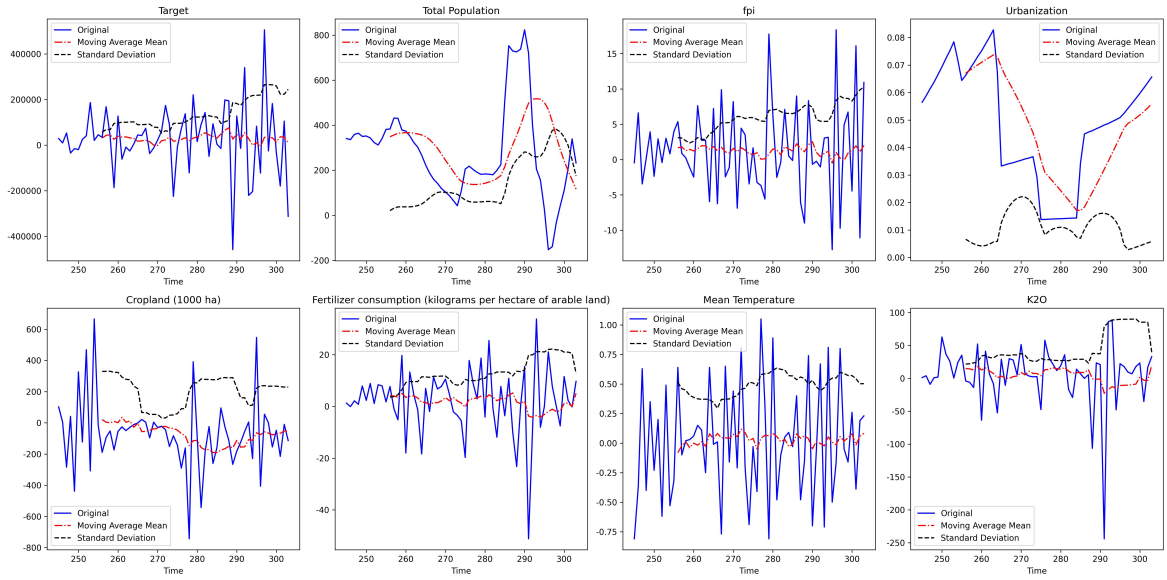


Figure 15: Moving Average and Standard Deviation at First Order Difference

except for **Urbanization**. Based on the evidence from Figure 15, we decided to exclude **Urbanization** from our next modelling steps to ensure model accuracy.

4.3.2.2 White Noise Test Following the stationarity tests, we further investigated the nature of our time series data with the white noise tests. A series that exhibits white noise characteristics implies it is a sequence of random numbers and is inherently unpredictable. Therefore, validating our series is not white noise is crucial before proceeding.

We employed the Ljung-Box test, a type of white noise test that checks the autocorrelations of a time series to determine if they are significantly different from zero. In other words, it helps to test the randomness of our time series data. And it's formulated as:

Variable	Test Statistic	Result
Production	-4.268	Stationary
Total Population	-3.594	Stationary
FPI	-3.757	Stationary
Urbanization	-1.722	Non-stationary
Cropland (1000 ha)	-10.647	Stationary
Fertilizer (Total)	-3.95	Stationary
Mean Temperature	-3.271	Stationary
Fertilizer (K2O)	-4.018	Stationary

Table 9: Stationarity test results at first-order differencing

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \tag{8}$$

Source: Adapted from Ljung and Box 1978

In this equation 8, Q is the Ljung-Box test statistics, $\hat{\rho}_k$ is the sample autocorrelation at lag k. n, h are the number of observations and the number of lags being tested.

Under the null hypothesis, the test statistic Q follows a chi-squared distribution with h degrees of freedom. If the computed Q is significantly greater than the critical value from the chi-squared distribution, then at least one of the autocorrelations is significantly different from zero, indicating the series is not white noise. However, in practice, we often refer to the corresponding p-value of the Q statistic. A low p-value suggests that it's not white noise.

We utilized a lag value of 40 (i.e., $lag = 40$) and set the significance level at $\alpha = 0.05$ for the first-order differenced data to process the test. The presented results below in the table are based on the p-values associated with the Q statistic for each series.

Clearly, as evident from Table 10, none of the variables tested exhibit white noise behaviour. Each variable has a p-value below the threshold of 0.05, meaning there are underlying patterns or structures in these series that can be explored and modelled. With these insights in hand, our next step is to determine and select the most fitting models for our time series data, optimizing our chances for accurate forecasting.

Variable	p-value	White Noise?
Production	0.005149	No
Total Population	0.0	No
FPI	0.0	No
Cropland (1000 ha)	0.006356	No
Fertilizer (Toal))	0.018143	No
Mean Temperature	0.000034	No
Fertilizer (K2O)	0.002592	No

Table 10: White Noise Test Result with $lag = 40$ and $\alpha = 0.05$

4.3.3 Model Building

4.3.3.1 ARIMAX

ACF & PACF Analysis To determine the appropriate order of autoregressive (AR) and moving average (MA) components for our model, we made the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of our data.

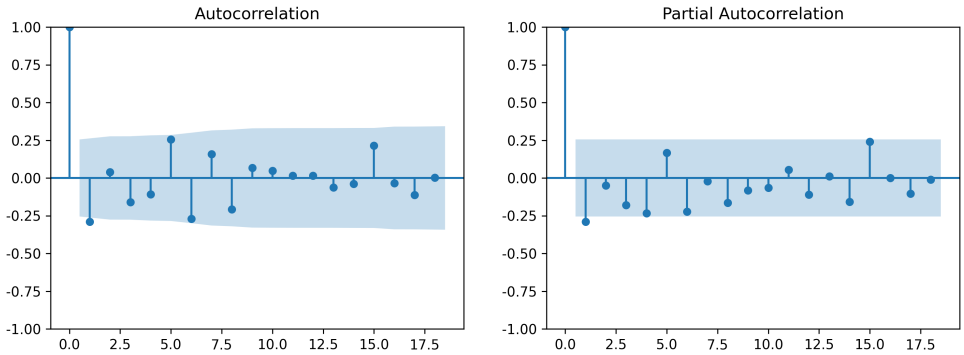


Figure 16: ACF and PACF

From the ACF and PACF plot above, Figure 16, we observed that the autocorrelation sharply cuts off after the first lag, implying that there is a significant correlation between an observation and its immediate predecessor. Beyond this lag-1, the ACF values hovered within the 95% confidence interval bounds, oscillating between approximately -0.25 and 0.25. In addition, the PACF plot mirrored these findings. The partial autocorrelation was significant at lag-1 and then remained within the confidence bounds for the subsequent lags.

Considering the abrupt cutoff in the PACF at lag-1, it suggests the possibility of an

AR(1) component in our time series data. Similarly, the significant autocorrelation at lag-1 in the ACF implies a potential MA(1) component. Given that both the ACF and PACF display patterns reminiscent of white noise after the first lag, it is logical to first attempt an ARMA(1,1) model. Consequently, an ARIMA(1,1,1) model, which incorporates one autoregressive term ($p = 1$), one level of differencing ($d = 1$), and one moving average term ($q = 1$), emerges as a potential fit for our data. However, further validation is essential before finalizing this model choice.

Grid Search for Optimal Model Using graphical methods to determine the optimal parameters for the ARIMA model is not straightforward. The process is highly subjective and time-consuming. Moreover, to validate the appropriateness of the ARIMA(1,1,1) model derived from the ACF and PACF analysis, a more objective method is necessary. Therefore, we further consider employing a grid search with Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) approaches to systematically select the optimal parameter values.

The **Akaike Information Criterion (AIC)** and the **Bayesian Information Criterion (BIC)** are two measures used to find the most appropriate model that strikes a balance between model fit and complexity:

$$AIC = 2k - 2\ln(L) \tag{9}$$

$$BIC = k\ln(n) - 2\ln(L) \tag{10}$$

Source: Adapted from Kumar 2023

where k denotes the count of estimated parameters, within the model, n is the number of data points utilized for model fitting and L refers to the likelihood estimation.

In this step, we split our data into training and testing sets. The training set has observations from 1961 to 2010, while the testing set contains data points ranging from 2011 to 2020.

As a result of our grid search with training data and exogenous variable, the ARIMAX(1,1,1) model with the lowest AIC value of 1277.62 and the lowest BIC value of 1294.46 was identified as the best-fit model. Further, it corresponds with the previous ACF and PACF analysis, hence, we will evaluate with the ARIMAX(1,1,1).

Model Diagnosis When fitting the ARIMAX model, model diagnosis is important before using the model, which ensures that no assumptions made by the model are violated. Specifically, our primary concern revolves around whether the residuals of the model are correlated and if they adhere to a zero-mean normal distribution.

p_value	White Noise?
0.597514	Yes

Table 11: White Noise Test Results for Residuals

We employed the White Noise test(Ljung-Box test) on residual, and from Table 11, it turns out that the p-value for our model's residuals is 0.597514, which is greater than the common significance level of 0.05. Hence, the residuals are a white noise series. Therefore our model ARIMAX(1,1,1) has adequately extracted the information.

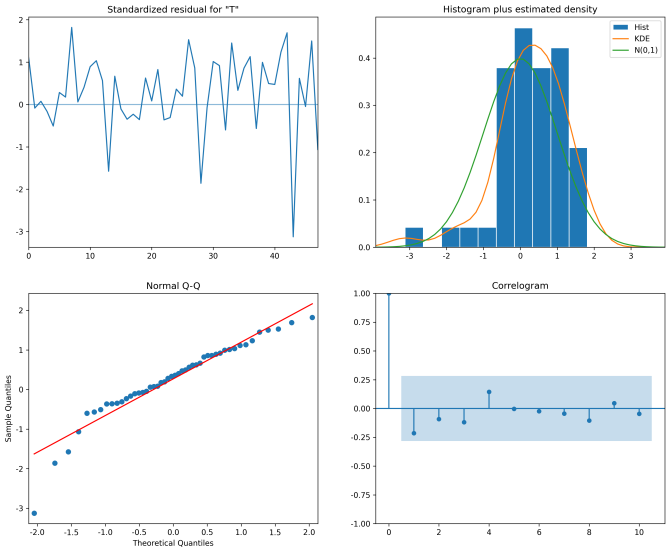


Figure 17: The Result of Model Diagnosis

From Figure 17, the residuals' time series plot is consistently stable, with no significant fluctuations as time progresses. The normal distribution graph in the top right corner demonstrates that the orange KDE line stays closely with the N(0,1), and this pattern strongly indicates a normal distribution of the residuals. Further, it is confirmed by the Q-Q plot located in the bottom left corner. In addition, the ACF plot in the bottom right corner shows the absence of any significant autocorrelation within the residuals, suggesting that they represent a white noise series. Therefore, we conclude that the

ARIMAX(1, 1, 1) model is a good fit for our time series data, and it ensures a robust understanding of the original time series data and provides us with predictive ability.

4.3.3.2 LSTM (Long Short-Term Memory)

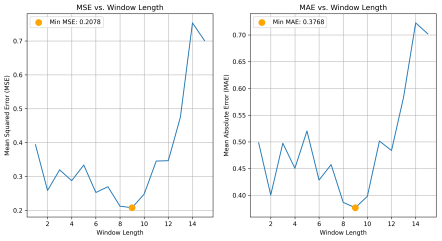
To construct an effective LSTM model for time series prediction, a selection of hyper-parameters plays a pivotal role in achieving optimal model performance. Each chosen hyper-parameter is crucial for the model's behaviour and its ability to capture intricate patterns within the data.

Model Configuration The best-selected hyperparameters for the LSTM model are as follows: the activation function is set to `tanh`, implying the use of the hyperbolic tangent activation function; `dropout` is set at 0.1, indicating that during training, 10% of neurons in each of the two LSTM layers are randomly deactivated to prevent overfitting; `num_layers` is set to 2, signifying the inclusion of two LSTM layers in the model; the optimizer is `rmsprop`, indicating the use of the Root Mean Square Propagation optimization algorithm for efficient learning; and units is set to 128, determining the number of neurons in each LSTM layer, allowing for the capture of intricate patterns. These hyperparameters were determined through a hyperparameter search process that aimed to optimize the model's performance in terms of forecasting accuracy, measured by metrics like Mean Squared Error (MSE) and Mean Absolute Error (MAE). With this model architecture, it is then compiled onto 100 epochs and a `batch_size` of 32.

Year	Production	Fertilizer (N)	Fertilizer (K2O)	Fertilizer (P2O5)	
1961	547341.0	327.2	94.7	307.9	
1962	576366.0	346.0	96.0	311.0	X_1
1963	585814.0	333.0	100.0	314.0	Y_1
1964	638755.0	364.0	91.0	312.0	
1965	604179.0	474.0	92.0	306.0	

Year	Production	Fertilizer (N)	Fertilizer (K2O)	Fertilizer (P2O5)	
1961	547341.0	327.2	94.7	307.9	
1962	576366.0	346.0	96.0	311.0	
1963	585814.0	333.0	100.0	314.0	X_2
1964	638755.0	364.0	91.0	312.0	Y_2
1965	604179.0	474.0	92.0	306.0	

(a) An illustration of how the rolling window works when the length is 2



(b) The performance of LSTM with respect to different window length and optimum

Figure 18: Caption for both figures

Rolling Window In addition to the previously mentioned configurations, we conducted fine-tuning of the time step, employing a rolling window (as depicted in Figure 18a), to evaluate the model's performance across various window lengths. Figure 18b presents an overview of the average performance metrics, specifically high-

lighting Mean Squared Error (MSE) and Mean Absolute Error (MAE), for the model across different window lengths. Notably, as the window length ranged from 1 to 15, the model's performance displayed fluctuations, initially decreasing until reaching a minimum and subsequently increasing. Particularly, a window length of 9 emerged as a standout, yielding one of the lowest average MSE and MAE values. This observation suggests that considering the previous 9 data points or time periods when making predictions leads to accurate forecasts, effectively striking a balance between capturing historical context and adapting to changing patterns in the time series data.

4.3.4 Model Performance

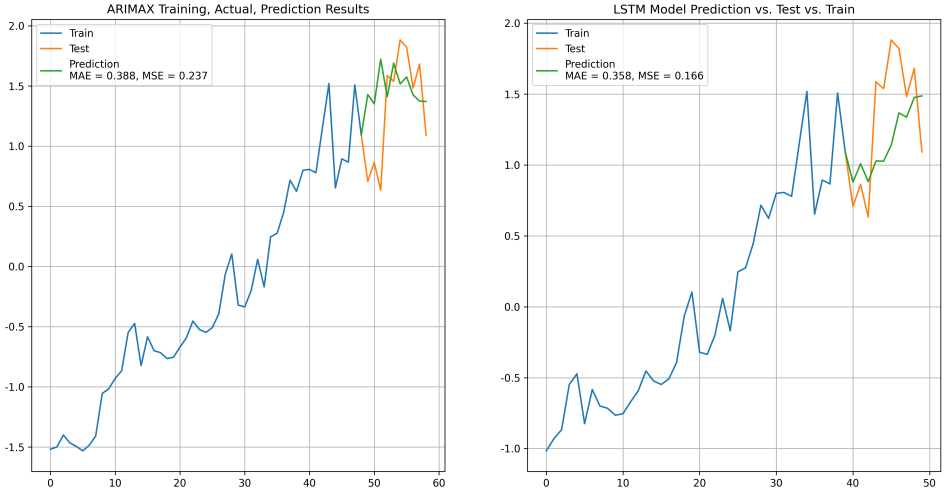


Figure 19: Training, Actual, Prediction Results Comparison for ARIMAX and LSTM

Figure 19 displays the prediction performance of both ARIMAX and LSTM models, with orange and green lines representing the true and predicted values, respectively. Upon comparison, it becomes evident that ARIMAX predictions exhibit fluctuations around the true values, while LSTM keeps all the trends but forecasts seem to deviate to a greater extent. Statistical assessment reveals that ARIMAX yielded a mean absolute error (MAE) of 0.388 and a mean squared error (MSE) of 0.237, both of which exceeded the corresponding metrics for LSTM, which achieved an MAE of 0.358 and an MSE of 0.166. This indicates that LSTM outperforms the traditional method.

LSTM's superior performance can be attributed to several factors. Firstly, the traditional method relies on moving averages and autoregressive techniques, resulting in forecasts that tend to gravitate toward historical mean values. In contrast, LSTM

employs a neural network architecture that allows it to capture and store relevant information while discarding auto-correlations within the data. Additionally, real-world data often exhibit intricate patterns with long intervals and lags, influenced by numerous complex exogenous factors. Recursive neural network algorithms like LSTM excel at identifying these patterns and making accurate predictions. Secondly, real-world data frequently exhibit significant oscillations that do not tightly correlate with historical values, favouring LSTM's ability to be more precise.

4.4 Supervised Regression

In this section, our primary goal is to build predictive models that can effectively capture the relationship between input features and the target variable, namely tomato yield across Europe. In the context of supervised regression, our initial inclination often leans towards linear models due to their simplicity. However, upon a closer examination of the scatter plot, an intriguing discovery emerges. Even after applying log transformation to address positive skewness, the plot, as depicted in Figure 20, fails to reveal a strong linear relationship between the selected features and tomato yield, suggesting that we may need to explore alternative modelling approaches capable of handling the more intricate and potentially non-linear data patterns.

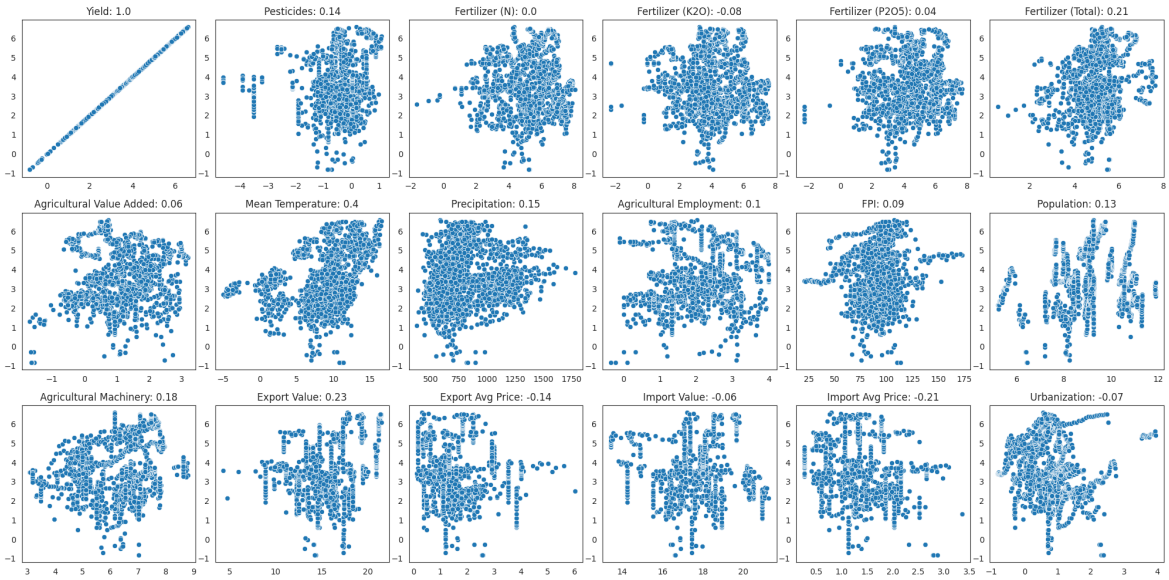


Figure 20: Scatter plot between Yield and Each Feature (After Log-transformation)

We propose employing a diverse set of models, namely Random Forest (RF), XGBoost, K-Nearest Neighbors (kNN) Regressor, and Multilayer Perceptron (MLP). Each of these models represents a distinct class of modelling techniques. RF and XGBoost

belong to the ensemble methods category, which excels at capturing complexity in data. kNN is representative of non-parametric models, known for its interpretability and adaptability. Lastly, MLP is renowned for its ability to handle non-linear data, making it well-suited to our specific problem and dataset. By harnessing the strengths of these diverse models, we aim to unravel the complexities of tomato yield prediction and achieve superior results.

Our predictive modelling procedure predominantly capitalizes on the pre-existing machine learning algorithms provided by the `scikit-learn` and `TensorFlow` libraries. Moreover, to enhance the efficacy of our models by refining their hyperparameters, we embrace the `GridSearchCV` library and develop our custom functions. These components seamlessly integrate the cross-validation process into our methodology.

4.4.1 Feature Engineering

1. **Handling missing values:** To effectively address missing values within our dataset, we meticulously partitioned the data into subsets based on each country, year, and feature. To handle missing data, we employed a twofold strategy. Firstly, we utilized backwards filling to propagate the first available non-NaN value to previous NaN data points. Secondly, we employed forward filling with the historical mean to maintain statistical coherence. A special case arose with Serbia's data for pesticide usage, where no data was available. In this instance, we used Romania's pesticide data as a proxy for Serbia's, as their tomato yield value had the highest correlation at 0.83.

2. **Dealing with categorical features:** For categorical features (i.e., Year and Country), we drop them out during the modelling process. While notable country-by-country differences and year-to-year fluctuations exist within the data (which has been analyzed through the last sections), in this section we are more concerned with identifying direct factors that could influence tomato yield. Initially, 17 features were passed to modelling.

3. **Split into train and test:** In the context of forecasting tomato yield, our approach involves a randomized division of the dataset into an 80% training set and a 20% testing set, utilizing the `train_test_split` function from the `scikit-learn` library. The final training data used is consist of 1,383 observations and 346 for the test dataset.

4. **Scaling:** The `StandardScaler` from `scikit-learn` was used to normalize our data. This process is essential for ensuring that all features are on the same scale, which helps in preventing certain models, particularly those sensitive to differences in feature scales like k-NN, from dominating the training process due to larger magnitude features.

4.4.2 Model Evaluation

Models	**Train RMSE (kg/ha)**	**Train R^2**	**Test RMSE (kg/ha)**	**Test R^2**
XGBoost	1.579	99.99%	23.552	95.36%
Random Forest	8.997	99.53%	20.109	96.59%
K-Nearest Neighbours	22.393	97.05%	23.433	95.49%
MLP	52.176	80.71%	49.493	75.64%

Table 12: Model performance with default configuration

Table 12 displays the performance metrics for the models on the training and test datasets. It shows that XGBoost and Random Forest consistently demonstrate strong predictive performance on both the training and test datasets. On the training dataset, the XGBoost model stands out with the lowest RMSE of 1.579 kg/ha and an R-squared value of 99.99%, suggesting an excellent fit to the training data. The Random Forest model, while having slightly higher RMSE values at 8.997 kg/ha, still demonstrates strong performance on the training data with an R-squared value of 99.53%. Moving to the test dataset, XGBoost and Random Forest perform slightly lower than in the training phase, indicating robustness and generalizability. The default K-Nearest Neighbours (n_neighbours = 5) exhibits relatively higher RMSE compared to the previous models but maintains good predictive performance with an R-squared value of 97.05%. It maintains consistent performance between the training and test datasets. In contrast, the default MLP (1 hidden layer with 100 units, compiled by Adam optimizer, 100 epochs, and batch size 32) has the highest RMSE among the models, indicating larger prediction errors. Its R-squared value suggests moderate performance on the training data and a notable drop in performance on the test data, indicating potential challenges to new data and a need to fine-tune the architecture and parameters.

4.4.3 Hyper-parameters Tuning

4.4.3.1 Feature selection To identify the most influential predictors for tomato yield and remove any irrelevant features which may hinder the model performance and generality, we employed **Lasso Regression** and **Feature Importance Analysis**.

Lasso Regression is a technique that introduces a penalty term based on the absolute values of the coefficients of the predictors. With a regularization parameter (α) set to 0.001, both `Agricultural Employment` and `Export Avg Price` exhibited zero coefficients, indicating their limited impact. This implies that these two predictors have little to no influence on tomato yield and can be safely excluded from our model. This decision not only streamlines model complexity but also guards against overfitting, enabling the model to focus on the most impactful predictors.

Concurrently, Feature Importance Analysis using a Random Forest algorithm ranked these two features as having low influence on the prediction of tomato yield (Figure 21). The Random Forest algorithm evaluates the importance of each feature by measuring the decrease in model accuracy when a feature is removed. Features like `Agricultural Employment` and `Export Avg Price` with low importance scores contribute less to predicting tomato yield compared to other features. Thus, these findings corroborate our decision to exclude them from the model. It is also noteworthy that `Export Value` ranks first in this analysis, highlighting its significant impact on the production value.

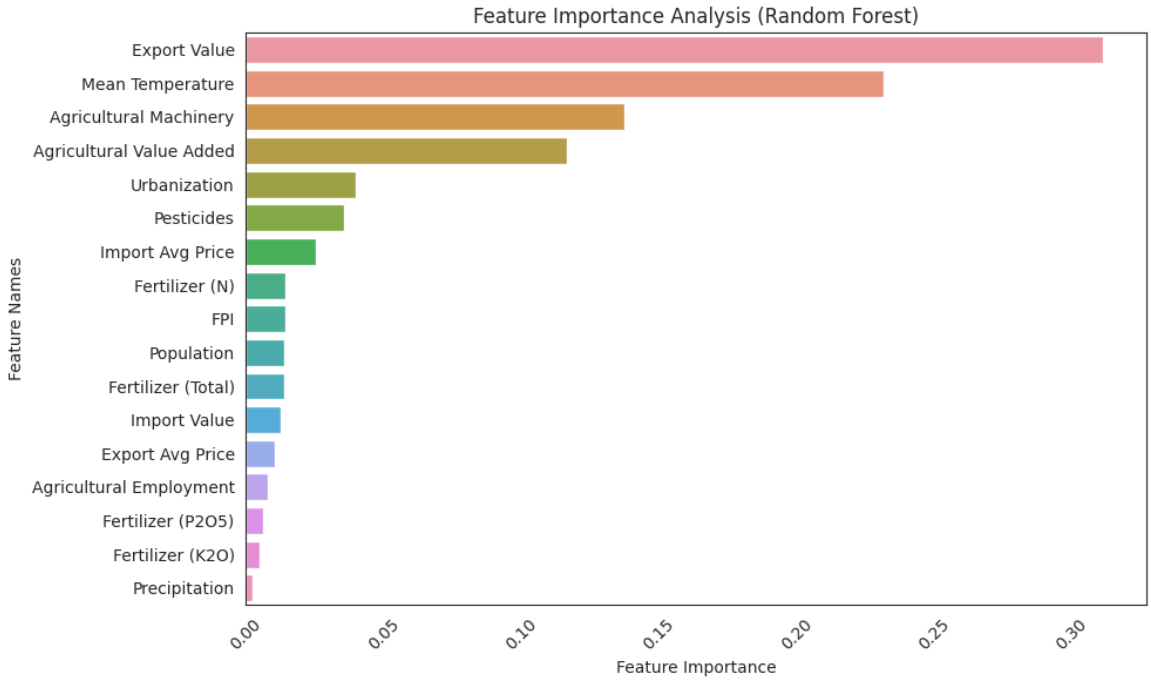


Figure 21: Feature Importance Analysis (Random Forest)

To prevent overfitting and improve the model generalization ability, we systematically fine-tuned our model configuration, architecture and training methods. We mainly utilized GridSearchCV, to define our model, the hyperparameter grid, the scoring metric (negative mean squared error), and the number of cross-validation folds (5-fold cross-validation). GridSearchCV meticulously explored various hyper-parameter combinations through cross-validation and identified the best-performing set. After fitting it to our training data, we extracted the optimal hyper-parameters, which minimize the mean squared error during cross-validation on the training data.

4.4.3.2 Random Forest (RF) We set the maximum tree depth (`max_depth`) to `None`, allowing the trees in our Random Forest to grow freely. This flexibility is essential due to our dataset's complexity, where intricate and non-linear feature relationships require deeper trees to capture nuances effectively. By setting the minimum samples per leaf node (`min_samples_leaf`) to 1, we enable the model to create nodes for individual data points, aligning with our goal of fine-grained predictions. This accommodates subtle factors that can impact tomato yield. For the minimum samples required to split an internal node (`min_samples_split`), we use 2 to strike a balance between complexity and overfitting risk, ensuring the model captures essential patterns while maintaining generalizability. Lastly, we select 300 trees (`n_estimators`) in the forest.

Table 13: Best Hyperparameters for Random Forest Regressor

Hyperparameter	**Value**
Maximum Tree Depth (`max_depth`)	None
Minimum Samples per Leaf Node (`min_samples_leaf`)	1
Minimum Samples to Split an Internal Node (`min_samples_split`)	2
Number of Trees in the Forest (`n_estimators`)	300

4.4.3.3 XGBoost In configuring the XGBoost model, we employed a set of best hyperparameters tailored to optimize its performance. Notably, learning rate (`learning_rate`) is set to 0.1, striking a balance between rapid convergence and accurate predictions. The choice of a maximum tree depth (`max_depth`) of 5 provides a controlled level of complexity, ensuring the model doesn't overfit while still capturing intricate relationships within the data. With 300 estimators (`n_estimators`), the model benefits from an ensemble of trees, enhancing predictive accuracy. Lastly, we introduced subsampling

(`subsample`), which is set at 0.8, allowing each tree to learn from a random subset of the data, adding robustness to it.

Table 14: Best Hyperparameters for XGBoost Regressor

Hyperparameter	**Value**
Learning Rate (`learning_rate`)	0.1
Maximum Tree Depth (`max_depth`)	5
Number of Estimators (`n_estimators`)	300
Subsample (`subsample`)	0.8

4.4.3.4 K-Nearest Neighbors (kNN) The process of selecting the best value of k for the k-Nearest Neighbors (kNN) algorithm involves a systematic evaluation of various k values to strike the right balance. We begin by defining a range of k values, typically from 1 to a chosen upper limit with all values odd (in our case, 13), representing the number of nearest neighbours. For each k value within this range, we train a kNN regressor model using the training dataset, enabling it to learn how to make predictions based on the specified k. Subsequently, we calculate MSE for both the training and testing datasets to measure how well the model's predictions align with actual values. By plotting the MSE values against the corresponding k values, we gain a visual understanding of how model performance varies with k (Figure 22). Here k = 3 is identified to minimize the testing MSE, indicating the optimal balance between bias and variance in the kNN model. Moreover, distance measurement is adjusted from Euclidean distance to Manhattan distance to accommodate the data characteristics better.

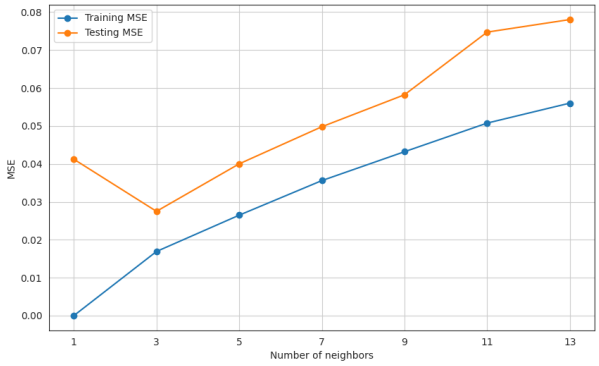


Figure 22: MSE for different number of K for K-Nearest Neighbors

4.4.3.5 Multilayer Perceptron (MLP)

To identify the optimal MLP architecture, we explored multiple configurations. This exploration involved varying the (`num_hidden_laye`
from 1 to 3, adjusting the size of each (`hidden_layer_size`) to either 50 or 100 neurons, and setting the(`drop_out`) to 0, 0.2, or 0.4. We employed the `Adam` optimizer with (`learning_rate`) of 0.001, 0.01, and 0.1. These hyper-parameters formed a parameter grid for systematic tuning. We iteratively created and trained the model for various combinations from this grid. During training, the MSE served as the loss function, and to minimize randomness, we repeated each three times, averaging the MSE.

The selected architecture consists of 21,901 parameters, employs 3 hidden layers, omits dropout layers and uses `ReLU` as the activation function. This final model is compiled with the Adam optimizer, specifically with a learning rate of 0.01. Subsequently, the model undergoes training on the provided dataset for 50 epochs, updating weights after processing batches of 32 samples. A 10% validation split is employed.

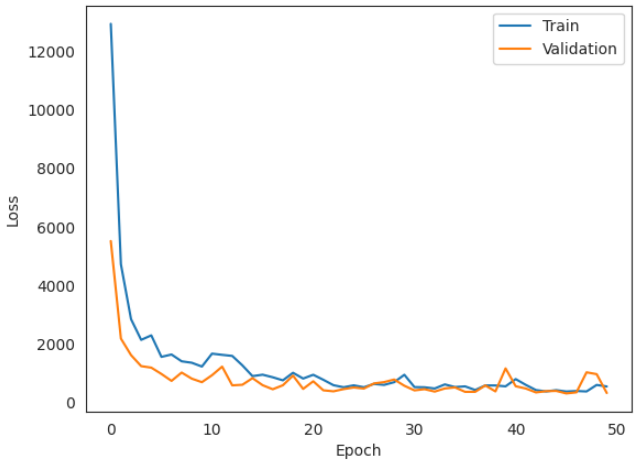


Figure 23: MLP training and validation Loss with epochs

Figure 23 monitors model performance across epochs. Notably, it illustrates that the model begins to converge at approximately epoch 10, where both the `val_loss` and `loss` curves start to exhibit convergence behaviour. Around epoch 25, they approach an equilibrium, with `val_loss` and `loss` nearly converging to the same values.

4.4.4 Performance Discussion

Figure 24 presents the model's performance on both the training and test datasets. Among the ensemble models, XGBoost and Random Forest exhibit excellent fits on the training data, with XGBoost performing slightly better. However, on the test

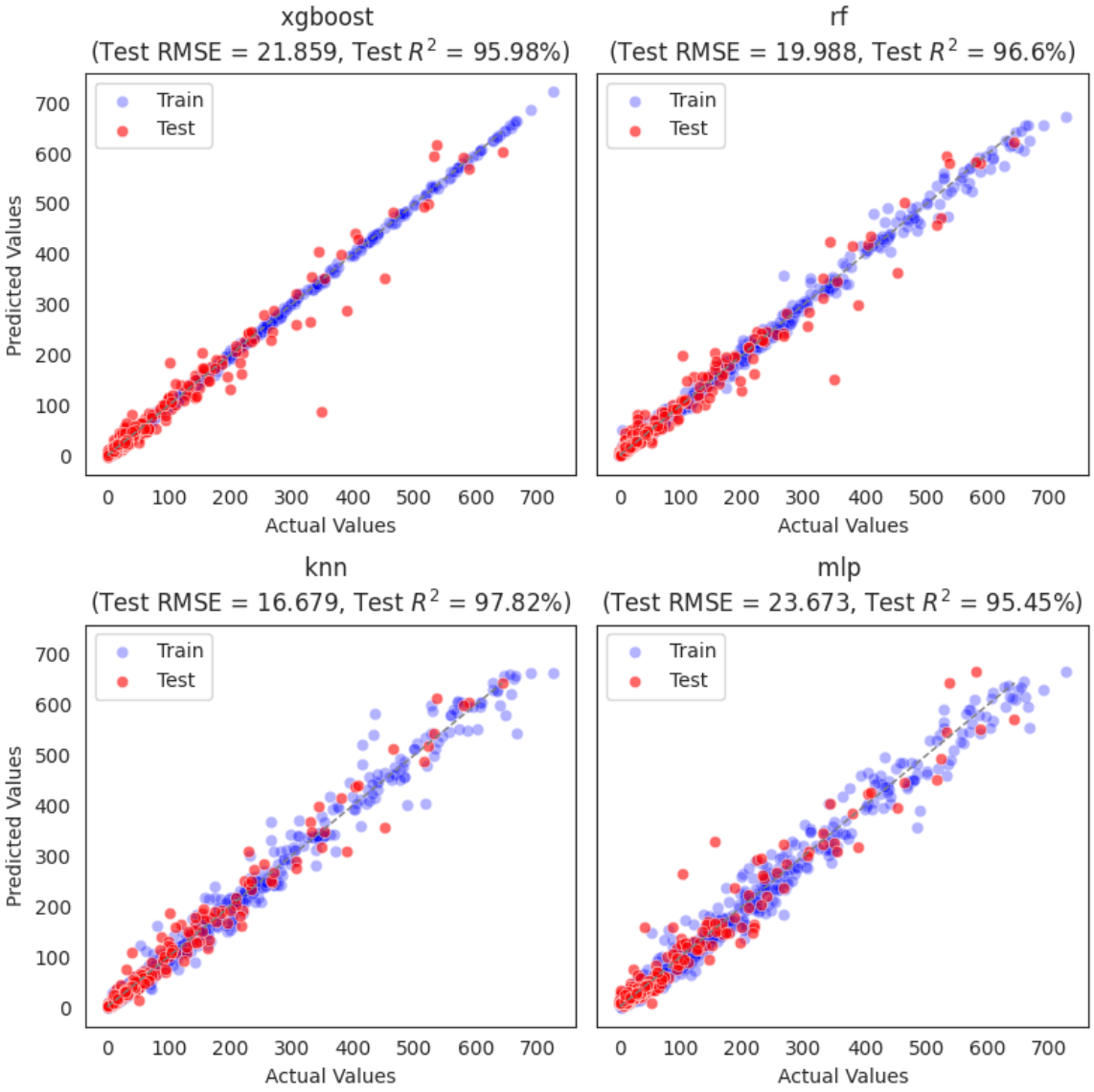


Figure 24: Actual value vs. predicted value by model

data, they struggle to handle outliers, resulting in several data points falling below the prediction line, indicating underestimations. In contrast, MLP tends to predict values higher than the actual ones, as evident from data points lying above the line.

The standout performer among the four models is K-Nearest Neighbors (kNN), boasting an impressive R^2 of 97.82% and an RMSE of 16.679 kg/ha on the test data. While it may not exhibit the tightest fit on the training data, kNN excels in generalizing to unseen data, maintaining robust performance on the test set.

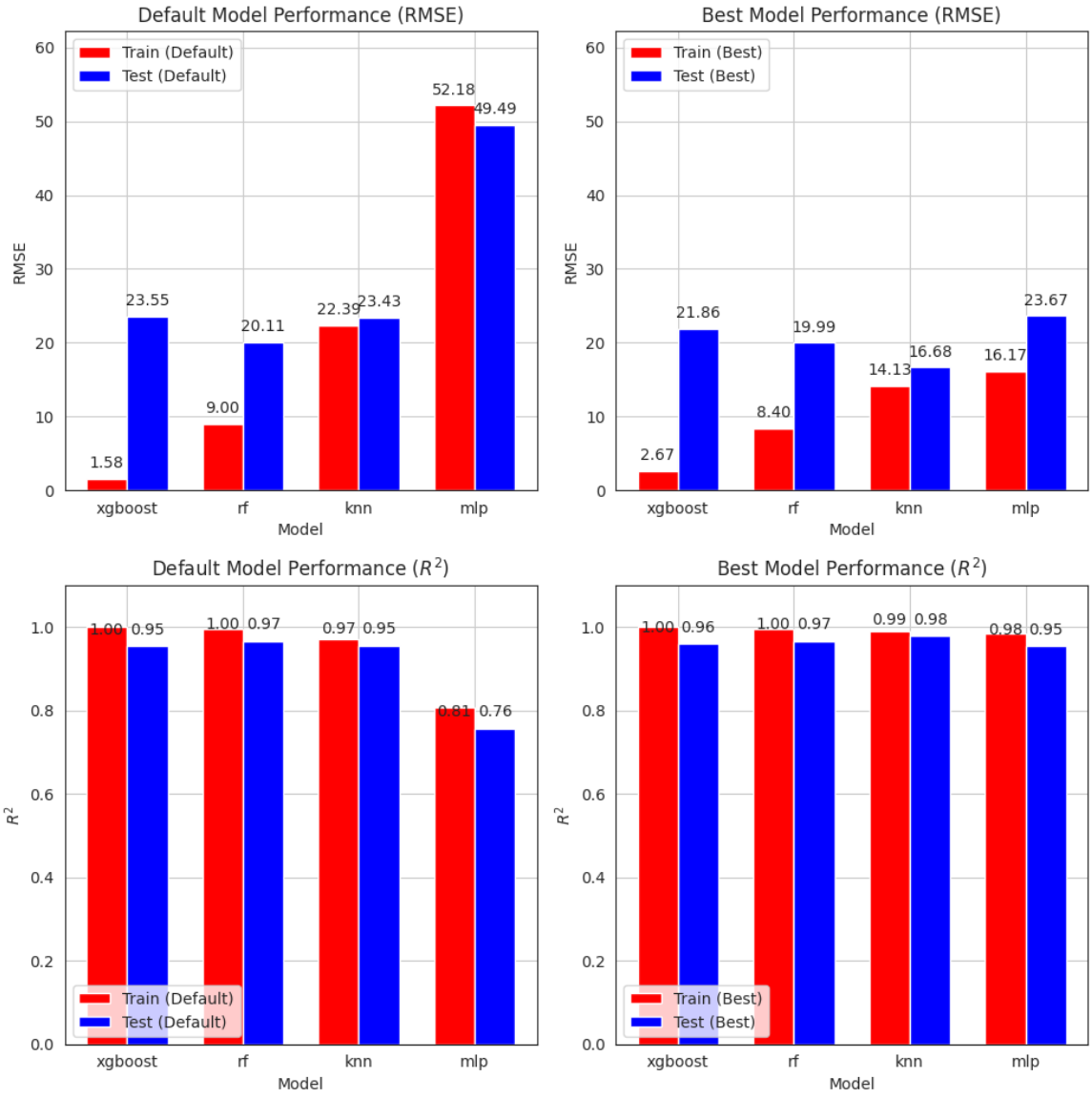


Figure 25: Model Model performance before and after hyper-parameter tuning

Figure 25 provides a visual representation of how our model tuning process has improved the model's performance by mitigating overfitting and enhancing overall predictive accuracy. The most substantial enhancement is observed in the MLP model, where the training R^2 has surged from 81% to an impressive 98%, and the test R^2 has risen

from 76% to a solid 95%. These improvements were achieved through adjustments to the model's complexity and training methods. On the other hand, the ensemble models, particularly XGBoost, exhibit relatively minor differences before and after tuning. In fact, XGBoost's Train RMSE even increased slightly after configuration adjustments, which can be attributed to its inherent robustness and ability to handle complex data. This demonstrates that ensemble models can provide strong predictive performance with minimal fine-tuning. Among these four models, our best-performing model (selected by test MSE), K-Nearest Neighbors (kNN) shows a moderate level of change before and after tuning. It boasts a 3% increase in test R^2 and a reduction of 6.75 kg/ha in test RMSE, reflecting its stable performance throughout the tuning process. kNN's effectiveness lies in its simplicity and ability to capture local patterns in the data, which may explain why it requires less extensive tuning compared to the MLP. In summary, our model tuning has yielded significant improvements in accuracy, especially for the MLP, while also showcasing the robustness of ensemble models.

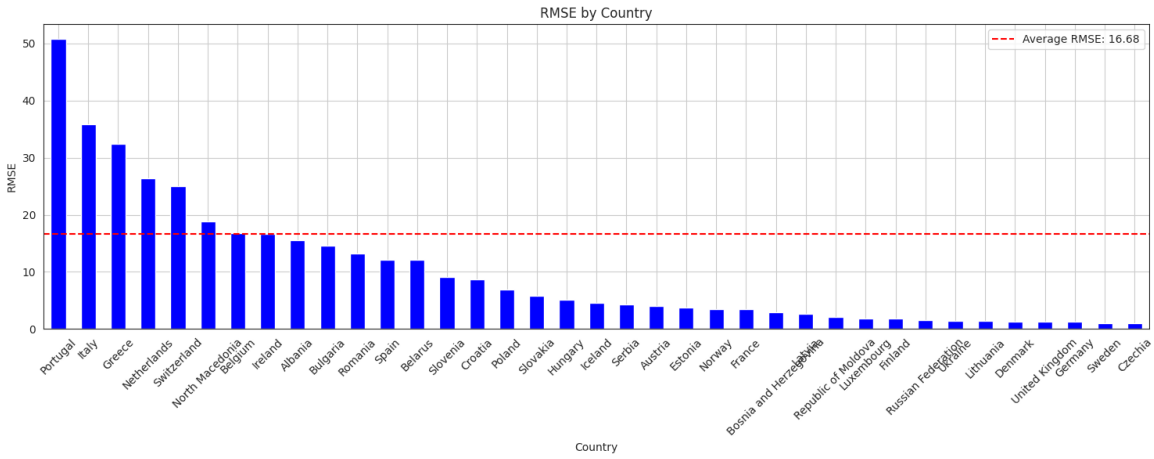


Figure 26: RMSE by different countries

The model's performance across various agricultural regions is examined in Figure 26. Czechia and Sweden stand out as exemplars of precision, boasting remarkably low RMSE values of 0.9537 kg/ha and 1.0336 kg/ha, respectively. In stark contrast, high-yield countries like Portugal (50.8160 kg/ha), Italy (35.9044 kg/ha), and the Netherlands (26.3739 kg/ha) pose challenges, resulting in significantly higher RMSE values.

This examination of the model's performance at the country level reveals distinct patterns. As shown in Table 15 and 16, which list the top and bottom-ranked countries in terms of annual average tomato yields, our model excels in regions characterized by mid to low-level agricultural yields. This proficiency might be attributed to the stability

Country	Yield (kg/ha)
Netherlands	435.61
Italy	414.27
Portugal	266.63
Greece	240.66
Belgium	223.42
North Macedonia	183.85
Switzerland	159.18
Albania	80.31
Spain	73.50
Romania	72.80
Bulgaria	70.35

Table 15: High-Yield Countries

Country	Yield (kg/ha)
United Kingdom	12.99
Denmark	12.24
Republic of Moldova	8.47
Germany	8.42
Estonia	8.20
Sweden	8.08
Latvia	7.35
Ukraine	6.86
Czechia	5.76
Luxembourg	3.68
Lithuania	2.98

Table 16: Low-Yield Countries

and relatively simpler agricultural conditions found in these areas, where agricultural production might be supplemented by imports to meet demands, such as Germany and the United Kingdom, as discussed earlier (see Section 4.2). Conversely, a different scenario unfolds in high-yield countries such as Portugal, Italy, and the Netherlands. Here, notably elevated RMSE values indicate substantial disparities in the model's forecasts. These discrepancies may arise from the intricate and multifaceted factors influencing agriculture in these regions, which the model may not entirely encompass. Improving accuracy in high-yield countries requires considering additional unlisted variables, such as soil conditions. By doing so, the model could potentially better adapt to the diverse agricultural landscapes in these regions, ultimately providing more precise predictions.

Figure 27 illustrates the RMSE trends across different years. A notable observation is that RMSE is consistently lower in recent periods, particularly after 1990. This trend may be attributed to several factors, such as improvements in data collection methods, thus data quality, or advancements in agricultural practices. Notably, there are two distinct peaks in RMSE values. The first surge occurs from 1990 to 1998, coinciding with a period when several countries experienced a significant decrease in tomato yields, including Germany, the United Kingdom, and the Netherlands, as discussed earlier in Section 4.2. The second surge is observed from 2007 to 2009, aligning with the subprime mortgage crisis and a global economic recession. While the timing

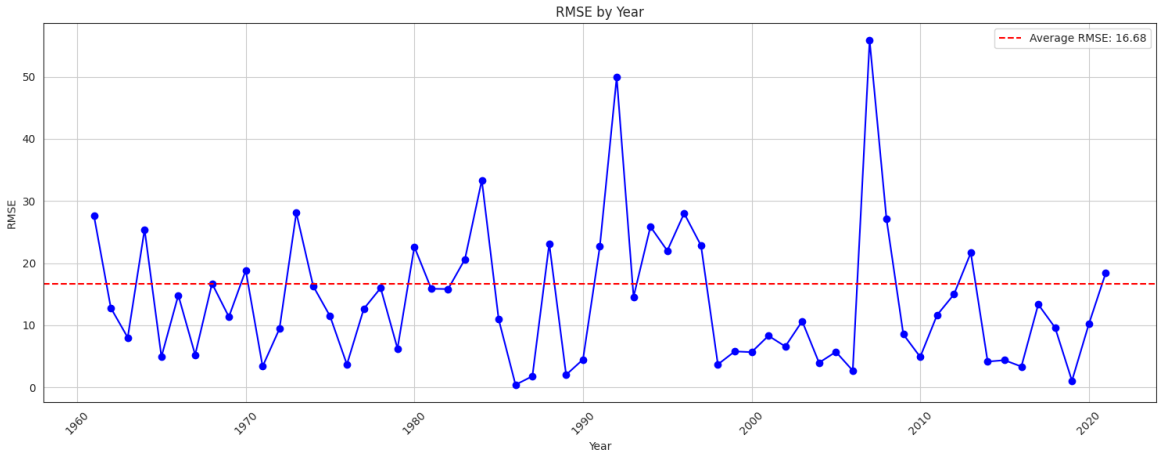


Figure 27: RMSE trend by year

of this surge and the economic events is notable, it's important to exercise caution when attributing causality. Economic recessions can have ripple effects across various sectors, including agriculture, which might explain the observed RMSE increase during this period. However, establishing a direct causal relationship would necessitate more extensive research and data analysis. If a causal relationship were to be established, it would validate the assumption that the model performs much better during periods of economic stability. Exploring these temporal patterns during those two periods may provide valuable insights into the complex interplay between economic factors and agricultural outcomes, warranting in-depth investigation.

5 Conclusion

In this study, we employed various data science and machine learning tools to conduct analysis of the European tomato market. Our approach encompassed network analysis, exploratory analysis, time series forecasting, and supervised regression.

Through network analysis, we gained valuable insights into the roles of various countries in the tomato export market. Notably, countries like the Netherlands, Spain, Germany, Poland, Italy, and the United Kingdom emerged as influential players, marked by high PageRank and Betweenness centrality values. Moreover, comparing the empirical network with random models revealed that, in agricultural product trade, countries tend to form enduring relationships, evident in higher transitivity and reciprocity in the empirical network.

During exploratory data analysis (EDA), we acquired a foundational understanding of how this market has evolved over time and identified significant differences at the country level. For instance, Italy stood out as a dominant force in terms of both gross production value and unit yield. The Netherlands, while not as prominent in gross production, exhibited high unit yield and played a pivotal role in the tomato trade network. We further delved into the data using graphical representations and correlation analyses to elucidate potential factors contributing to these temporal and regional disparities. This exploration led to the identification of key influential variables, such as machinery usage and export value.

Then, in time series forecasting, we selected Spain as a particular country we wish to investigate more for its continuous growth over time. We employed a dual approach, incorporating both traditional time series models and the more advanced LSTM model, in our analysis. The results made it evident that LSTM while wielding considerable predictive power, came at the cost of some interpretability compared to the ARIMAX.

Extending our analysis further, we developed supervised machine-learning models using data from all European countries. Among these models, XGBoost and Random Forest exhibited exceptional performance, showcasing impressive predictive accuracy and compatibility. And we also witnessed the importance to tailor the design of neural networks to bring the best prediction outcome. kNN outshone all other contenders, achieving a remarkable Root Mean Square Error (RMSE) of 16.68 kg/ha and an outstanding goodness-of-fit, as reflected by $R^2 = 98\%$, when assessed on previously unseen

datasets. Our study also showed that export value is ranked as first in the feature importance analysis, validating our previous discussion in Section 4.1 and Section 4.2 that trade has a significant impact on the production of tomatoes. Our investigation also delved into the intricacies of our model's performance, revealing its excellence in predicting yields for countries historically characterized by lower yields, potentially owing to the relative simplicity of agricultural markets and landscapes in these regions. In contrast, higher-yield countries exhibited greater variability in predictions. To enhance our model's precision, we recommend further exploration in countries harbouring complex factors, such as diverse soil conditions and frequent data updates. Additionally, our model showcased superior accuracy and stability when applied to recent-year data compared to earlier stages. We pinpointed two specific periods, the 1990s and 2007-2009, as promising avenues for deeper research. In sum, our study not only underscores the potential of advanced machine learning techniques in agricultural yield prediction but also highlights opportunities for tailored approaches across diverse agricultural landscapes and historical periods.

Appendix A Variable and Data Source

Variable	Data Source	Unit/Description
Gross Production Value of Tomatoes	FAO	1000 USD, constant 2014-2016
Agriculture, forestry and fishing, value added	FAO	% of GDP
Food Production Index	World Bank	2014-2016 = 100
Population (Total, Rural and Urban)	FAO	1000 No
Pesticides	FAO	kg/ha, Use per area of cropland
Fertilizer Consumption: Nutrient Breakdown by N, P and K	IFA (International Fertilizer Association)	1000 tonnes of nutrients
Fertilizer Consumption (Total)	FAO	kg/ha, Use per area of cropland
Cropland	FAO	1000 ha
Mean Temperature	World Bank CCKP (Climate Change Knowledge Portal)	°C
Precipitation	World Bank CCKP (Climate Change Knowledge Portal)	mm
Agricultural machinery	World Bank	tractors per 100 km2 of arable land
Share of employment in agricultural, forestry, and fishing in total employment	FAO	% (ILO Modelled Estimates)
Export and imports value of tomatoes	UN Comtrade Database	1000 USD

References

Atsalakis, G. S. (2016). "Using computational intelligence to forecast carbon prices". In: *Applied Soft Computing* 43, pp. 107–116.

Babu, C. N. and E. Reddy B (2014). "A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data". In: *Applied Soft Computing* 23, pp. 27–38.

Colah (Aug. 2015). *Understanding LSTM Networks*. URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Cubillas, J. J. et al. (Aug. 2022). *A machine learning model for early prediction of crop yield, nested in a web application in the cloud: A case study in an olive grove in southern Spain*. URL: https://www.mdpi.com/2077-0472/12/9/1345.

European Space Agency (ESA) (2020). *Europe Land Cover Mapped in 10 m Resolution*. Accessed on [Insert Date]. ESA. URL: https://www.esa.int/ESA_Multimedia/Images/2020/03/Europe_land-cover_mapped_in_10_m_resolution.

Food and Agriculture Organization (FAO) (2023). *FAOSTAT: Food and Agriculture Organization Corporate Statistical Database*. https://www.fao.org/faostat/en/#definitions. Accessed: August 20, 2023.

Goodyear, C. (Feb. 2023). *Predicting threats to food security*. URL: https://www.cam.ac.uk/stories/predicting-threats-to-food-security.

Harkness (n.d.). *Adverse weather conditions for UK wheat production under climate change*. URL: https://pubmed.ncbi.nlm.nih.gov/32184532/.

Hochreiter, S. and J. Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Jiménez-Donaire, M., A. Tarquis, and J. V. Giráldez (Jan. 2020). *Evaluation of a combined drought indicator and its potential for agricultural drought prediction in southern Spain*. URL: https://nhess.copernicus.org/articles/20/21/2020/nhess-20-21-2020-discussion.html.

Kumar, A. (May 2023). *AIC amp; BIC for selecting regression models: Formula, examples*. URL: https://vitalflux.com/aic-vs-bic-for-regression-models-formula-examples/.

Ljung, G. and G. E. P. Box (1978). "On a Measure of Lack of Fit in Time Series Models". In: *Biometrika* 66, pp. 67–72.

Martınez, F. et al. (2019). "A methodology for applying k-nearest neighbor to time series forecasting". In: *Artificial Intelligence Review* 52.3, pp. 2019–2037.

Matlab (n.d.). *Create ARIMA Models That Include Exogenous Covariates*. MATLAB Simulink. URL: https://www.mathworks.com/help/econ/arimax-model-specifications.html.

Murray, P. W., B. Agard, and M. A Barajas (2018). "Forecast of individual customer's demand from a large and noisy dataset". In: *Computers & industrial engineering* 118, pp. 33–43.

Prabhakaran, S. (Nov. 2019). *Augmented Dickey Fuller Test (ADF Test)*. URL: https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/.

Saadio, C. (Apr. 2022). *Crops yield prediction based on machine learning models: Case of West African countries*. URL: https://www.sciencedirect.com/science/article/pii/S2772375522000168.

Sagheer, A. and M. Kotb (2019). "Time series forecasting of petroleum production using deep LSTM recurrent networks". In: *Neurocomputing* 323, pp. 203–213.

Setayesh, A., Z. S. H. Zadeh, and B. Bahrak (June 2022). *Analysis of the global trade network using exponential random graph models - applied network science*. URL: https://appliednetsci.springeropen.com/articles/10.1007/s41109-022-00479-7.

Shi, H., S. Hu, and J. Zhang (2019). "LSTM based prediction algorithm and abnormal change detection for temperature in aerospace gyroscope shell". In: *International Journal of Intelligent Computing and Cybernetics* 12.2, pp. 274–291.

Vojnovic, M. (2023). *Milan Vojnovic ST456 Deep Learning Lecture 2*.

Yanhua, C. et al. (2015). "A hybrid application algorithm based on the support vector machine and artificial intelligence: An example of electric load forecasting". In: *Applied Mathematical Modelling* 39.9, pp. 2617–2632.