REPUBLIC OF TURKEY
ADANA ALPARSLAN TÜRKEŞ SCIENCE AND TECHNOLOGY UNIVERSITY

INSTITUTE OF GRADUATE SCHOOL
DEPARTMENT OF INDUSTRIAL ENGINEERING

PREDICTION OF APRICOT EXPORT VOLUME USING ARTIFICIAL INTELLIGENCE

SÜMEYYE ÖLMEZOĞLU
MASTER OF SCIENCE

ADANA 2022

REPUBLIC OF TURKEY

ADANA ALPARSLAN TÜRKEŞ SCIENCE AND TECHNOLOGY UNIVERSITY

INSTITUTE OF GRADUATE SCHOOL

DEPARTMENT OF INDUSTRIAL ENGINEERING

PREDICTION OF APRICOT EXPORT VOLUME USING ARTIFICIAL INTELLIGENCE

SÜMEYYE ÖLMEZOĞLU

MASTER OF SCIENCE

SUPERVISOR

ASSOC. PROF. DR. AYŞE TUĞBA DOSDOĞRU

ADANA 2022

ABSTRACT

PREDICTION OF APRICOT EXPORT VOLUME USING ARTIFICIAL INTELLIGENCE

Sümeyye ÖLMEZOĞLU
Department of Industrial Engineering
Supervisor: Assoc Prof. Dr. Ayşe Tuğba DOSDOĞRU
November 2022, 51 pages

This study aims to determine the most appropriate predicting method for the export volume of apricot products, which is one of the essential export items in Turkey. While deciding the predicting strategy, the most appropriate prediction method is found as a result of comparing the predictions with Artificial Neural Networks (ANN), Seasonal Autoregressive Integrated Moving Average (SARIMA), and Extreme Gradient Boosting (XGBoost) methods.

In this study, 2002-2020 General Trade System (GTS) Foreign Trade monthly data is obtained from the Turkish Statistical Institute (TURKSTAT). In which seasonality is observed, Apricot data is estimated using AR, ARIMA, SARIMA, ANN, and XGBoost methods for 2020. As a result, XGBoost achieves better prediction accuracy than SARIMA and ANN, according to the impact of the performance measure.

Keywords: Agriculture in Turkey, Apricot, Foreign Trade, Export, Artificial Intelligence, ANN, XGBoost, Time Series, AR, ARIMA, SARIMA

ÖZET

YAPAY ZEKA KULLANILARAK KAYISI İHRACAT MİKTARININ TAHMİNİ

Sümeyye ÖLMEZOĞLU
Endüstri Mühendisliği Anabilim Dalı
Danışman: Doç. Dr. Ayşe Tuğba DOSDOĞRU

Kasım 2022, 51 sayfa

Bu çalışmada amaç, Türkiye de önemli ihracat kalemlerinden olan kayısı ürünün ihracat miktarı için en uygun tahmin yöntemini belirlemektir. Tahmin yöntemi belirlenirken, Yapay Sinir Ağları (YSA), Mevsimsel Otoregresif Entegre Hareketli Ortalama (SARIMA) ve Gradyan Artırılmış Ağaçlar (XGBoost) yöntemleri ile tahminlerinin karşılaştırılması sonucu en uygun tahmin yöntemi bulunmuştur.

Bu çalışmada Türkiye İstatistik Kurumu'ndan (TÜİK) aylık bazda 2002-2020 Genel Ticaret Sistemi (GTS) Dış Ticaret verileri elde edilmiştir. Mevsimsellik gözlemlenen kayısı verileri 2020 yılı için AR, ARIMA, SARIMA, YSA ve XGBoost yöntemleri kullanılarak tahmin edilmiştir. Sonuç olarak mevcut verilerle karşılaştırıldığında, performans ölçüm sonuçlarına göre en uygun yöntemin XGBoost olduğu bulunmuştur.

Anahtar Kelimeler: Türkiye’de Tarım, Kayısı, Dış Ticaret, İhracat, Yapay Zeka, Yapay Sinir Ağları, XGBoost, Zaman Serileri, AR, ARIMA, SARIMA

1. INTRODUCTION

Agriculture has a vital role in a country's economic growth, such that its place and importance are increasing daily. Agricultural production and yield vary from country to country due to geography, agricultural policies, climate, and fertilization. Agriculture is important in Turkey and each country because it has a high share in foreign trade. Besides, agricultural production may differ yearly due to some reasons, such as pollution and climate change. Therefore, the fluctuation of agricultural output affects the foreign trade volume, which is still under investigation.

Foreign trade is the international exchange of goods or services, where buying and selling transactions occur. It occurs in two ways: import and export. On the other hand, foreign trade statistics cover the trade between Turkey and other countries. Foreign trade statistics are one of the crucial indicators of the Turkish and world economy. These statistics and data are used extensively by public institutions, organizations, business circles, academic circles, and international organizations. Statistics are used in several areas, such as Turkey's balance of payments balance sheet calculated by the Central Bank of the Republic of Turkey, Gross National Product and input-output tables calculated by the Turkish Statistical Institute, development plans, medium-term programs, and annual programs.

Moreover, academic circles use related scientific studies, while private sector organizations use statistics on products within their field of activity. Foreign trade statistics produced and published by TURKSTAT are at international standards regarding definition, scope, and methodology (Turkish Statistical Institute, 2022). The movement is known as an economic transaction between countries called international trade.

Foreign trade is one of the essential levers of the economies due to the reasons such as contributing to the development of nations, increasing the communication of countries with each other, production of value-added products, and the increase it will provide in exports (Orkunoğlu Şahin, 2022).

Every country has a direct or indirect relationship between its agricultural sector and economy. Countries with high agrarian production have a high level of international agricultural trade.

Agricultural foreign trade is divided into General Trade System (GTS) and Special Trade System (STS). GTS shows the more trader-oriented regional trading. For example, if a large batch of coffee is brought from Brazil and sold to that country and the other part to neighboring countries, it is the case of first importing and then exporting to another country. On the other hand, STS does not show the number of coffee trade as import or export, which is only tax paid and entered the country, nationalized, and sold to other countries that do not enter the country. In this thesis, GTS data will be analyzed.

According to Food and Agriculture Organization (FAO), Turkey ranked first with 846.606 tons in apricot production in 2019. TURKSTAT publishes Turkey's imports and exports data in monthly periods. TURKSTAT reported that Turkey had 68 thousand tons of export volume and 305 tons of import volume in 2020. In this thesis, Turkey's export volume of fresh apricot is predicted using monthly data from TURKSTAT.

Table 1.1. HSCODE12 codes for the data of apricot

HSCODE 12	**EXPLANATION**
080910000000	Apricot (including turmeric) (fresh)
081190950012	Apricots (including turmeric) (with no added sugar or sweetener) (frozen)
081290100000	Apricots (including turmeric) (temporarily canned)
081290250000	Apricot (including turmeric), orange; provisionally preserved, not edible

The World Customs Organization suggests a code called Harmonized System Code (HS code) to classify the products worldwide. These standardized codes, which have been standardized, are taken from the Harmonized system. The HS code consists of 12 digits. The harmonized system determines the first six digits of the code. Countries are not allowed to make any changes in these first six codes. Table 1.1 HS shows the code for the apricot.

Data collection and analysis of agricultural data are essential in Turkey because Turkey is an agrarian country. Therefore, predicting future data becomes more critical. Artificial intelligence (AI) studies related to agriculture have been rising recently.

With AI, concepts such as business intelligence, data analysis, machine learning, database, artificial neural networks, and statistics have emerged. AI algorithms can be categorized as supervised learning, unsupervised learning, and reinforcement-based learning. It is broadly divided into supervised learning and unsupervised learning, where supervised learning is divided into two parts, regression, and classification, to predict the target categories. Unsupervised learning is clustering algorithms. They are also used in problems of extracting the relationship between inputs.

Dictionary meaning of the word guess is the prediction of something, such as an event, which may happen based on reason, intuition, or some data. Predicting the future or trying to predict is an element of socioeconomic development. To decide all the situations that must be present in the future. When organizations or companies make future-oriented decisions, they need guesswork. Predicting future events helps to maintain and improve their conditions. They need to find suitable solutions within the framework of a good plan. Because inaccurate predictions negatively affect the future goals of companies.

Time series aims to make predictions for the future with the help of observation values from past periods. It is widely used in many fields, such as medicine, engineering, business, economics, and finance. Different models, such as AR, ARIMA, and SARIMA, were created using various methods to make predictions with the help of time series. However, many time series encountered in real life do not contain only linear relationships. ANN and XGBoost, which can model linear and non-linear relationships due to their structure, have been alternative methods used in analyzing time series in recent years. The most crucial advantage of ANN and XGBoost is that in cases where the functional structure of the data set cannot be determined precisely, it can successfully model many different forms of the available system based on the data.

It is difficult to conclude whether the relationship in the time series used is linear or non-linear. Therefore, models created using a single method may only sometimes give the best results. This study uses AR, ARIMA, and SARIMA models from time series models and ANN and XGBoostT models from machine learning models.

This thesis predicts the apricot export amount monthly for 2020 using AR, ARIMA, SARIMA, ANN, and XGBoost methods. Finally, these methods are compared, and the most suitable model is selected.

The aim of this thesis is twofold. The first one is to accurately analyze the data of Turkey’s apricot export amount, which ranks first in world exports. Moreover, the second is to predict with models that have not yet been used in this data set and have recently become popular.

This thesis consists of 5 chapters, where the first chapter gives a general introduction to the thesis's aim and subject. Chapter 2 explains a literature review and background of related studies, and chapter 3 provides information about the dataset and the methods used in this thesis. Chapter 4 gives the performance measurement results of the proposed models. Chapter 5 presents the compared prediction results; for this dataset, the most appropriate method is selected. The thesis ends with conclusions and recommendations.

2. LITERATURE REVIEW

2.1. Review of Artificial Intelligence Techniques

In recent years, artificial intelligence methods have been used in many fields, and successful results have been obtained. This section examines the studies on artificial intelligence methods, with the used models and the obtained results from those models.

Luo, Zhang, Fu, & Rao (2021) provided information about the course of the pandemic by estimating the transmission times of the Covid19 pandemic by examining the Covid19 data in the USA using long short-term memory (LSTM) and XGboost models. LSTM has been applied to confirm the accuracy of the COVID-19 pandemic data due to its long duration and diversity. Sensitivity analysis was also performed with the XGBoost model. They used MAE, MSE, RMSE, and MAPE evaluation parameters to evaluate the results from the model. As a result, it has been observed that Covid19 infection decreases with the isolation of people who do not have the disease. They stated that it is possible to eliminate the pandemic with measures such as social distancing and contact tracing.

Osman et al. (2021) used XGBoost, ANN, and Support Vector Regression (SVR) models to predict groundwater levels in Selangor, Malaysia. They used 11 months of precipitation, temperature, and evaporation data to estimate groundwater levels. Using different evaluation techniques, they observed that the XGBoost model performed better and served as an excellent benchmark for predicting future groundwater levels.

In their thesis, Zhu et al. (2021) estimated the position of the rock heads with the hybrid model N-XGBoost. The location information of the rock head is essential for the design and construction of tunneling or underground excavations, and its accurate estimation is difficult. In this study, the XGBoost model was compared with the gradient boosting regression tree (GBRT), the light gradient boosting machine (LightGBM), the multivariate linear regression (MLR), the artificial neural network (ANN), and the support vector machine (SVM). Using different evaluation parameters, they observed that the XGboost model gave much more successful results than other models. Therefore, they mentioned that the proposed N-XGBoost probability hybrid model is reliable for estimating the rock head location.

Duan et al. (2020) estimated the strength of recycled aggregate concrete using the ICA-XGBoost hybrid model. Recycled aggregate concrete is used as an alternative material to protect the environment and achieve sustainable development. Recycled aggregate concrete is a critical parameter in terms of the compressive strength of the material, and its deficiency is a significant concern. In this study, a sociopolitical algorithm obtained four hybrid models that brought the 28-day compressive strength of recycled aggregate concrete (ICA). They investigated which performed better using the ICA-XGBoost, ICA-ANN, ICA-SVR, and ICA-ANFIS models. It has been proven that the best among these models developed according to performance indicators is the ICA-XGBoost model. The results showed that the proposed ICA-XGBoost model outperformed other models in predicting the compressive strength of recycled aggregate concrete. The ICA-XGBoost model can be used in civil engineering to achieve an adequate mechanical performance of recycled aggregate concrete and allow its safe use for building purposes.

Li et al. (2019) estimated gene expression values. They used XGBoost, the D-GEX algorithm, linear regression, and K-Nearest Neighbors (KNN) to estimate gene expression values. They observed that XGBoost gives the best results.

Yücelsan (2018) proposed a sales prediction model for the white goods industry. Forty-six months of sales data is used for dishwashers, washing machines, refrigerators, small appliances, and television products. The factors affecting sales, such as exchange rate, holidays, consumer confidence index (CPI), producer price index (PPI), and regional house sales, are used as explanatory variables. When the results obtained by ANN, ARIMA, and ARIMAX methods are compared according to the mean squared error performance criteria, it can be said that the most accurate estimates are obtained by using the ANN method.

Chen et al. (2018) analyzed the monthly mean temperature in Nanjing, China, from 1951 to 2017 using SARIMA methods. The data between 1951 and 2014 are used as the training set, and the data between 2015 and 2017 as the test set. This study has a detailed description of model selection and prediction accuracy. The results indicate that the proposed approach achieves the best prediction accuracy.

Pan (2017) estimated the hourly PM2.5 air pollutant concentration using XGBoost, random forest algorithm, multiple linear regression, decision tree regression, and support vector machine. It has been observed that the XGBoost algorithm gives the best result.

Wu et al. (2017) proposed a new prediction model based on data mining techniques to predict type 2 diabetes. This prediction model consists of two parts, the first one is an improved K-means algorithm and the second one is a logistic regression algorithm.

Valipour (2015) investigated the capabilities of the SARIMA and ARIMA models for long-term flow prediction in the United States. Their study uses the average flow of stations in each state as a dataset, whereas the data from 1901 to 2010 were used as a train set. Each US state estimates runoff for 2011. When the results were examined, it was found that the accuracy of the SARIMA model gave better results than the ARIMA model.

Şengür & Tekin (2013) estimated the graduation grades of Fırat University using machine learning algorithms. They used ANN and Decision Trees methods for the prediction.

Mombeni, Rezaei, Nadarajah, & Emami (2013) estimated monthly residential water consumption values in Iran for one year using the SARIMA model. They used monthly residential water consumption data in Iran from May 2001 to March 2010 as a dataset. They found that a three-parameter log-logistic distribution is suitable for the residual model.

Kaynar & Taştan (2009) created a hybrid model by combining the ARIMA model and the multilayer perceptron (MLP) model for estimating monthly and daily exchange rates (YTL/$). In the hybrid model, the linear component of the time series was combined with the ARIMA model and the nonlinear component with the MLP model and estimated. They measured the prediction performance of the hybrid model by comparing the prediction results obtained using the ARIMA and MLP models alone and the prediction results of the hybrid model. They observed that the hybrid model gave better results than the other two models.

Stojanova et al. (2006) predicted forest fires in Slovenia using various methods. They used different machine learning algorithms in their study. Logistic regression, decision trees, random forests, bagging, and boosting of decision trees are applied. The best results are obtained when the prediction results are examined by bagging the decision trees.

2.2. Review of Agricultural Prediction Studies

Numerous studies in the agricultural sector used artificial intelligence to predict seasonal data. This section summarizes the studies that used artificial intelligence techniques for the agricultural industry, the models they used, and the results obtained.

Akın et al. (2021) estimated Turkey's cherry production for 2020-2025 using the 1961-2019 annual production data. They found that the cherry data are not stationary. For this reason, stationary time series is obtained by taking the first differences of the series before the time series modeling has been done. They used the ARIMA model for estimation.

Aksoy, Halis, & Salman (2020) emphasized the problem of predicting plant diseases and solved this problem using artificial intelligence methods. In the study, convolutional neural network models detect leaves with scabies, black rot, and rust disease in apple plants. Among the models used, ResNet-34 is the best model for disease detection in apple plants.

Günel (2020) analyzed the prediction performances of Turkey's exports using machine learning models and classical models, Long Short-Term Memory (LSTM), Support Vector Machines (SVM), Random Forest (RF), ARIMA, and Exponential Smoothing (ETS) models. When the results are compared, it is seen that ETS gave the best prediction, while the ARIMA model is the weakest.

Güngül & Yenilmez (2019) estimated the foreign trade data in the agricultural sector for the next five years with the exponential smoothing method. The ratio of exports to imports in 2023 is calculated using the hot-winter method, considering the seasonal effects. It has been observed that there will be a decrease according to the prediction results.

Jha et al. (2019) summarized various studies for automation in the agricultural sector, such as sensors, IoT, and machine learning. In this study, studies aimed at increasing productivity in agriculture are examined. They presented various agricultural applications such as crop selection for appropriate field management, using these technologies for appropriate fertilizer selection, and crop status warning systems with sensors.

Terzi et al. (2019) explained various studies using artificial intelligence in plant and animal production. This study examines various studies such as raw milk quality, detection and identification of animal behavior, vegetative yield, soil erosion prediction, frost control, detection and removal of weeds, and disease detection in plants.

Akkaya (2007) explained the ANN method and discussed the usability of ANN in agriculture and the various applications of this method as an alternative method in solving agricultural problems. ANN is used for agricultural lands as an alternative method to solve problems and draws the attention of researchers to this method.

Dellal & Koç (2003) tried to estimate the demand elasticity of essential importers and exporters and the supply elasticities of dried apricots for the apricot market. The Fruit Bearing Trees Model is used for the demand, and the Linear Model is used for the export estimation of dried apricots. As a result, it has been observed that while fresh apricots are consumed, the demand for dried apricots will decrease.

A summary of related literature is given in Table 2.1.

Table 2.1. Selected studies from the literature review

Authors (Years)	**Methods**	**Data / Variables**
Luo, Zhang, Fu, & Rao, (2021)	LSTM and XGBoost	Transmission times of the Covid19
Osman, Ahmed, Chow, Huang, & El-Shafie, (2021)	XGBoost, ANN, and SVR	The level of groundwater in Selangor Malaysia
Zhu, et al., (2021)	XGBoost, GBRT, LightGBM, MLR, ANN, and SVM.	The position of the rockheads
Akın, et al., (2021)	ARIMA model	Cherry production
Duan, Asteris, Nguyen, Bui, & Moayedi, (2020)	ICA-XGBoost hybrid model	The strength of recycled aggregate concrete
Aksoy, Halis, & Salman, (2020)	Convolutional Neural Network (CNN)	The problem of apple plant diseases
Günel, (2020)	LSTM, SVM, RF, ARIMA and ETS models	The prediction performances on Turkey's exports

Güngül & Yenilmez, (2019)	Exponential Smoothing Method	Turkey's Agriculture Sector Foreign Trade Balance
Li, Yin, Quan, & Zhang, (2019)	XGBoost, D-GEX algorithm, Linear Regression, and KNN models.	Gen expression value
Jha, Doshi, Patel, & Shah, (2019)	Artificial Intelligence	They talk about various studies on automation such as sensors, the internet of things, and machine learning in the agricultural sector.
Terzi, Özgüven, Altaş, & Uygun, (2019)	Blurry Logic, ANN, Genetic Algorithm, Expert Systems, Ant Algorithms	They mention various studies in which artificial intelligence is used in both plant and animal production.
Yücelsan, (2018)	ANN, ARIMA, and ARIMAX methods	The white goods industry
Chen, Niu, Liu, Jiang, & Ma, (2018)	SARIMA	The average monthly temperature in Nanjing, China from 1951 to 2017
Pan, (2017)	XGBoost Algorithm	PM2.5 air pollutant concentration
Wu, Yang, Huang, He, & Wang, (2017)	K-means algorithm and logistic regression algorithm based on data mining techniques.	Type 2 diabetes
Valipour, (2015)	SARIMA and ARIMA models	Long-term runoff study in the United States
Şengür & Tekin, (2013)	ANN and Decision Trees methods.	Fırat University's graduation grades
Mombeni, Rezaei, Nadarajah, & Emami, (2013)	SARIMA model	Monthly residential water consumption in Iran
Kaynar & Taştan, (2009)	A hybrid model is created by combining the ARIMA model and the MLP model for the prediction of time series.	Monthly and daily exchange rates (YTL/$)
Akkaya, (2007)	ANN method	The various applications in the agricultural sector
Stojanova, Panov, Kobler, Dzeroski, & Taskova, (2006)	Logistic Regression and Decision Trees.	Forest fires in Slovenia
Dellal & Koç, (2003)	Using the Fruit Bearing Trees Model for demand and the Linear Model for dried apricot exports.	The supply and demand elasticity of Apricots

In this thesis, the export volume of fresh apricot, one of Turkey's essential agricultural products, is predicted and expected to contribute to the correct analysis of the sector. With the recently popular artificial intelligence methods, the export volume of apricot data in the selected HS code is predicted. To the best of our knowledge, the proposed methods have yet to be used to analyze the export volume of fresh apricot for the selected HS code in the literature. Thus, the main goal of this thesis is to compare the performance of established methods to predict the export volume of fresh apricot.

3. MATERIALS AND METHODS

This chapter mentions detailed information about the selected dataset and the methods used. The following procedure conducts the research: first, AR, ARIMA, and SARIMA time series models are explained. Then, ANN and XGBoost, selected from artificial intelligence models, are described to predict the export volume of fresh apricot.

While choosing the model, AR, ARIMA, and SARIMA models, which are the most frequently used in Time series models, are used. Since there is seasonality in the data set, it is predicted that the SARIMA model, one of the Time series models, would yield more successful results than other models. ANN and XGBoost models, popular recently and more commonly used in similar datasets, are chosen for comparison. These models are between nonlinear and machine learning models.

TURKSTAT, the most critical institution in Turkey in terms of statistical data, compiles data and information in the fields needed by the country and produces, publishes, and distributes necessary statistics. When collecting data, it uses surveys and censuses from individuals, households, and workplaces. It makes various statistical analyzes and reports with the data it collects. This statistical information is a reliable guide in the decision-making stages of all segments of society. In this thesis, the data is obtained from TURKSTAT.

This study uses the export volume of the fresh apricot dataset for Turkey. The dataset covered the period between 2002 and 2020 and was downloaded from TURKSTAT (TURKSTAT, 2022). Descriptive statistics are also given in Table 3.1.

Various analyzes are performed to understand and interpret what a data set represents. These analyzes are also called statistics. Statistics allows us to understand and interpret data and transform observations into information. In short, in a place where uncertainties exist with statistical techniques, it controls these uncertainties through the concept of probability. It makes observations and gives information and interpretation based on science.

Table 3.1. Descriptive statistics of the dataset for apricot trade volume

Measure	Values
Count	228
Sum	677,240,330
Mean	2,970,532
Std	5,950,319
Min	8,740
Max	30,185,841

According to the Count value, a total of 228 monthly apricot export volumes is observed within the dataset. The sum value shows that the sum of monthly apricot exports is 677,240,330, while the mean value gives the center of the distribution of monthly apricot exports and is calculated as 2,970,532. Std shows that the standard deviation value for monthly apricot exports is 5,950,319, whereas the min value shows that the minimum value in monthly apricot exports is found as 8,740. The Max value shows that the maximum value for monthly apricot exports is 30,185,841.

The first twelve rows of the export volume are given as an example in Table 3.2.

Table 3.2. Example Data

Date	Export Volume
1.01.2002	184,189
1.02.2002	80,840
1.03.2002	87,220
1.04.2002	111,252
1.05.2002	404,411
1.06.2002	1,435,703
1.07.2002	2,704,889
1.08.2002	1,078,801
1.09.2002	496,134
1.10.2002	233,200
1.11.2002	59,000
1.12.2002	158,000

There are vast data available today. However, with so much data to examine, it can be difficult for people to see what the data means. Data visualization helps turn all this detailed data into easy-to-understand, visually appealing, and valuable business information. Good data visualization enables people to quickly and easily see and understand relationships with structures, detecting trends that may not be noticed in tables of pure numbers only. A well-designed graphic provides information, and a powerful presentation amplifies that information, grabs attention, and engages people in a way that no spreadsheet can.

A chart should always consider its data type and purpose. Some information is better suited to display, for example, on a bar chart versus a pie chart. In most tools, however, the user has a wide selection of visual analytics, from general graphs such as line and bar charts to timelines, maps, graphs, histograms, and custom designs. A bar chart for the apricot export volume by year is given in Figure 3.1. When the data is analyzed over the years, it has been observed that the volume of exports has increased gradually (Figure 3.1).

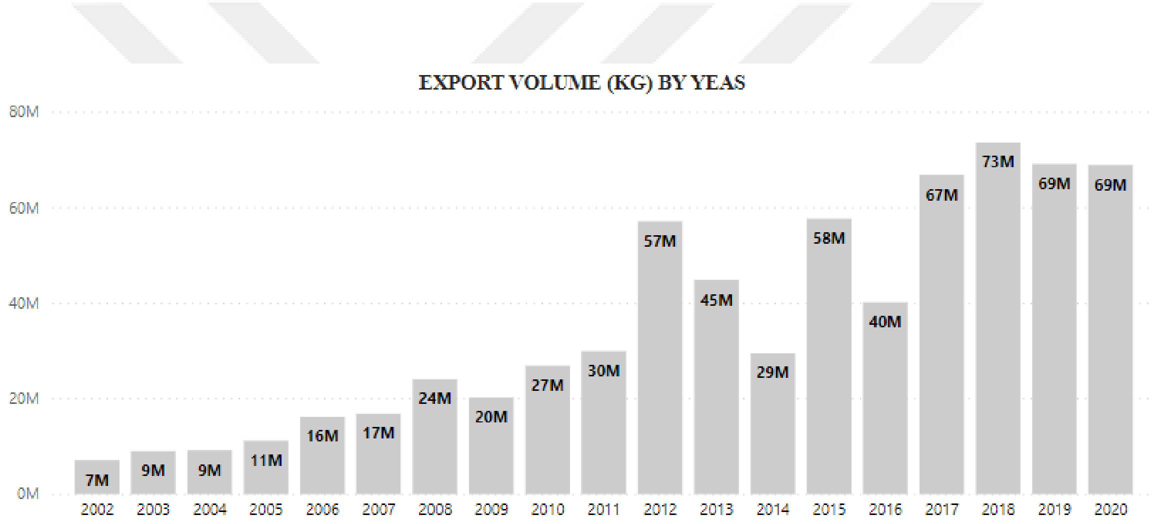


Figure 3.1. The Bar Chart of Export Volume (kg) By Years

When the data is analyzed by month, it is observed that the volume of exports in June, July, and August is higher than in other months (Figure 3.2). The figure shows that there is seasonality.

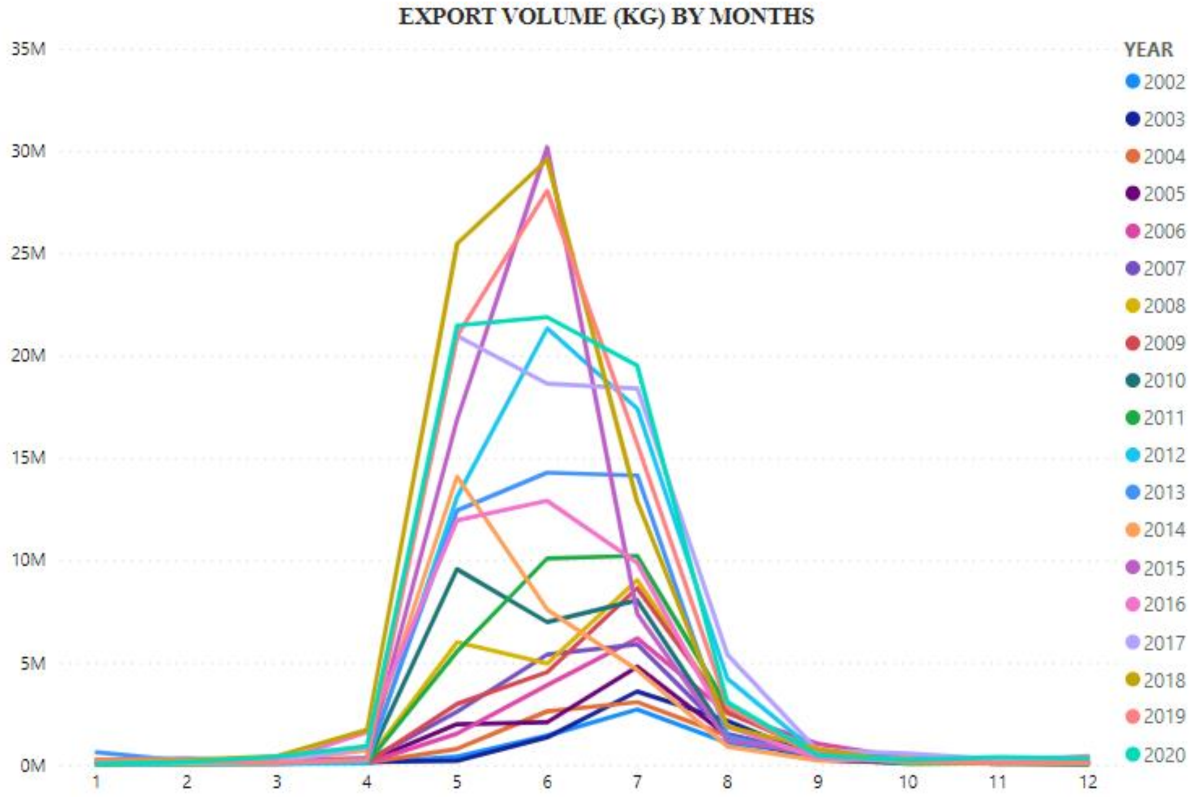


Figure 3.2. The Line Graph of Export Volume (kg) By Months

We can briefly call data science "the discovery journey of information." Data mining is using patterns in data or historical data based on statistical probabilities if predicted. As data grows, data science has become a trendy term recently. Combining data science with artificial intelligence, machine learning, and collected data is essential in gathering meaningful information.

Data mining reveals meanings from anonymous information used in many areas, such as customer management, marketing analysis, sales prediction, and many industries. Predictive, descriptive, and data visualization models are used in data science. Predictive Models can be divided into three as Classification, Regression, and Time Series Analysis.

Some of these algorithms used in data mining are ANN, Genetic Algorithms, Statistical Techniques, Decision Trees, Rule Inference, Fuzzy Sets, Situational Causality, and KNN.

This thesis will use prediction methods from artificial intelligence and machine learning techniques to predict agricultural trade data. We can make predictions by making sense of historical data by using artificial intelligence techniques. Selected methods are explained in the subsequent chapters.

Artificial intelligence is one of the most critical technologies in today's world. The concept of 'Artificial Intelligence' arouses curiosity in many people at first hearing. Human-specific qualities include finding a solution, understanding, inferring, generalizing, and learning. In the scientific world, artificial intelligence is also defined as the ability of a computer or computer-assisted machine to perform tasks related to higher logical processes of past human qualities. In doing so, it benefits from past experiences. Topics spoken around artificial intelligence and components are artificial neural networks, expert systems, fuzzy logic, and genetic algorithms. Pirim (2006) examined artificial intelligence in four groups which are given in Table 3.3.

Table 3.3. Artificial Intelligence Insights

Think like a human	Think wisely
Act humanely	Act wisely

The cornerstone of artificial intelligence is algorithms. An algorithm is simply a way of solving a mathematical problem consisting of specific iterations and functions. So it is a problem-solving step. In the programming field, it can be defined as creating an output depending on the available inputs. As it is known, there is more than one solution to a problem. The value of the algorithm is directly related to the value's approach to the problem, its performance, and the accuracy, scope, economy, and speed of the results (Arslan, 2020).

In studies on artificial intelligence, it is claimed that everything related to human beings in the universe operates within the framework of an algorithm. Accordingly, consciousness is the result of a mathematically complex algorithm. For most AI thinkers today, the brain is a structure that derives its functions from the laws of the physical world. Furthermore, as a result, its operations can be imitated with a mathematical formation, that is, a well-designed algorithm which means that artificial intelligence shows a mental feature (Arslan, 2020).

Machine learning is seen as a sub-branch of artificial intelligence. Machine learning is about letting computers learn what to do with algorithms rather than giving every possible step to solving a problem upfront. It means the development of algorithms that find solutions to a problem that may arise dynamically, using many previously defined or learned contents rather than programs that follow a predetermined path. (Arslan, 2020)

Machine learning algorithms create a pattern or a model using data. This model or pattern is software that uses it to predict new situations it may encounter. For example, it can be considered machine learning that a machine that monitors the market in the past or instantaneously analyzes the future market. Machine learning can be regarded as learning or using the process in which the steps of analyzing data, creating a model, and recognizing using the model are developed by using iteratively. (Arslan, 2020)

3.1. Box-Jenkins Method (AR / ARIMA / SARIMA)

The Box-Jenkins method is used in time series analysis in the study. This technique is based on batch, linear stochastic processes. Box-Jenkins Methods are generally divided into two parts: non-seasonal, given in section 3.1.1, and seasonal, given in section 3.1.2.

A time series is a series of data points. It consists of many consecutive measurements over some time. Different structures to consider when analyzing time series, such as autocorrelation, trend, or seasonal variation. Time series analysis, on the other hand, describes the structure of data points taken over time. The data in equal time intervals consisting of apricot export volume data in a specific period form a time series data. These data are then analyzed using time series techniques. Today, statistical techniques of time series analysis occupy a prominent place in the literature and are still being discussed. Time series analyses are used in various studies involving time-dependent data.

Time series models were first introduced in 1960 by Box and Jenkins. For this reason, the Box-Jenkins Model consisted of their names. Initially, the Box and Jenkins methodology consisted of three iterative steps. This included model selection, parameter estimation, and model control. Later, two more stages were added. These are the preliminary stage of data preparation and the estimation stage, which is the final stage of model application. The studied data requires it to meet the Box and Jenkins assumptions before modeling. In this sense, data preparation includes transformations and variances.

On the other hand, model selection involves using different graphs and model selection tools to determine the most suitable model for the data. Parameter estimation is a statistic when a statistic is used to determine an unknown parameter and to find the values of the model coefficients that best fit the data used. Model checking means testing assumptions to identify areas of model error. It is used for prediction after the model has been selected, estimated, and checked.

3.1.1. Non-Seasonal Box-Jenkins Models

If the data set is not affected by the seasonality condition, Non-Seasonal Box-Jenkins models are selected (Duru, 2007).

3.1.1.1. Autoregression Models (AR)

AR models express the observation value of a time series data in a specific time period as the linear calculation of the observation value and the error term in a certain number of periods before the same series.

3.1.1.2. Moving Average Models (MA)

MA models are a linear combination of data in each time period of a time series. These are the error terms of the same time period and the error terms of a given number of past time periods.

3.1.1.3. Autoregressive Moving Average Models (ARMA/ARIMA)

The ARMA model is used to model stationary time series and to understand their future value. This model is a linear combination of time series data for a specific time period, a certain number of previous data, and the error term. The ARMA model combines AR models with p terms and MA models with q terms. It is denoted as ARMA(p,q) and has the term p+q.

3.1.2. Seasonal Box-Jenkins Models

These models are selected if the data set is affected by the seasonality condition (Can, 2009).

3.1.2.1. Seasonal Autoregression Models (SAR)

Applying the autoregressive moving average method in the series containing the seasonal effect (weekly, monthly and seasonal data) becomes more difficult with adding a seasonal trend that repeats at regular intervals (s).

3.1.2.2. Seasonal Moving Average Models (SMA)

The seasonal moving average model is in question by expressing the current value of a stationary time series as a linear combination of the error term and the error term s periods ago.

3.1.2.3. Seasonal Autoregressive Moving Average Models (SARMA/SARIMA)

Three trend parameters are common to the ARIMA and SARIMA models and require configuration.

Trend Parameters are given as follows:

p: Trend autoregression degree.

d: Trend difference degree.

q: Trend moving average degree.

There are only four seasonal parameters that need to be configured, which are the SARIMA parameters; these are given as follows:

P: Seasonal autoregressive degree.

D: Seasonal difference degree.

Q: Seasonal moving average degree.

m: Number of periods in a seasonal season.

Below is the formulation of a general ARIMA(p,d,q) model.

$$Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \cdots + \Phi_p Z_{t-p} + \delta + a_t - \Theta_1 a_{t-1} - \Theta_2 a_{t-2} - \cdots - \Theta_q a_{t-q} \quad (3.6)$$

Inequality; Z_t, Z_{t-1},…, Z_{t-p} d-degree observation values, Φ_1, Φ_2,….. Φ_p Coefficients for d-degree observations, δ constant value, a_t, a_{t-1},…, a_{t-q} error terms and coefficients for Θ_1, Θ_2,…, Θ_q error terms.

Together, the notation for an SARIMA(p,d,q)(P,D,Q)m model is specified as:

$$\Phi_p(B^s)\Delta_s^D y_t = \Theta_q(B^s)\varepsilon_t \tag{3.7}$$

Δ_s in the model represents the seasonal difference operator, and S represents the seasonal period, which is taken as S=12 for monthly data in this thesis. The Δ^D operator in the model means that the seasonal difference of the data is taken D times. Non-stationary data and Δ_s^D are expressed as stationary data after the difference-taking process. In the model, Φ_p represents the seasonal autoregression (SAR) parameter and seasonal moving average (SMA) parameter, and y_t represents the non-stationary series.

3.2. Artificial Neural Networks (ANN)

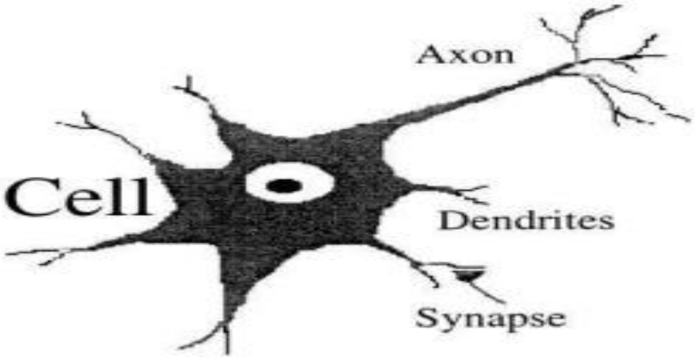


Figure 3.3. Biological Representation of a Nerve Cell (Agatonovic & Beresford, 718-719)

Similar to the human brain, the ANN consists of nerve cells. The general representation of a nerve cell consisting of dendrites and axons is given in Figure 3.3.

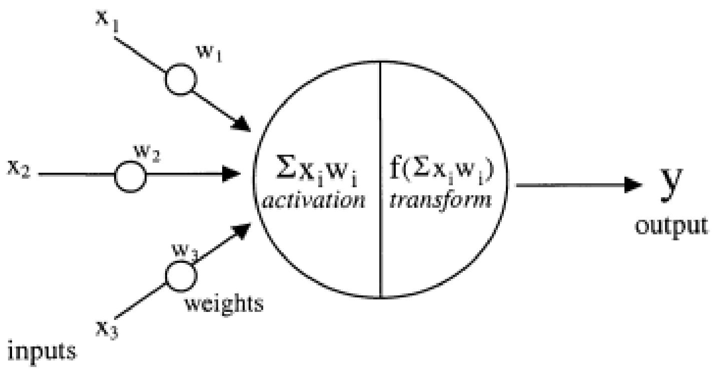


Figure 3.4. Mathematical Model of a Nerve Cell (Agatonovic & Beresford, 718-719)

The general mathematical model of a human nerve cell can be shown in Figure 3.4. Mathematically, the nerve cell, which is the most basic processing unit of the neural network, If we want to model the electrical signals from each dendrite, we first calculate a weight (w). We need to multiply by the value. These values are then passed into a sum function. The total value found is sent to an activation function, and the activation function will decide whether to transmit the electrical signal to the following nerve cells. The result of the activation function is transferred to other cells to which the nerve cell is connected, and the same processes are repeated. The total value sent to the activation function is obtained with the formula in 3.1 (Keçeci, 2020).

$$Y=w*x+b \tag{3.1}$$

y: It is a dependent variable since it depends on the value of x. Gives the score for the entry.

x: argument, input.

W: weight parameter

b: bias value

β: coefficient

P: variable

H: number of layers

$$h_k(x) = g\left(\beta_{ok} + \sum_{j=1}^{P} X_j \beta_{jk}\right) \tag{3.2}$$

$$g(u) = \frac{1}{1+e^{-u}} \tag{3.3}$$

$$f(x) = \gamma_0 + \sum_{k=1}^{H} \gamma_k h_k \tag{3.4}$$

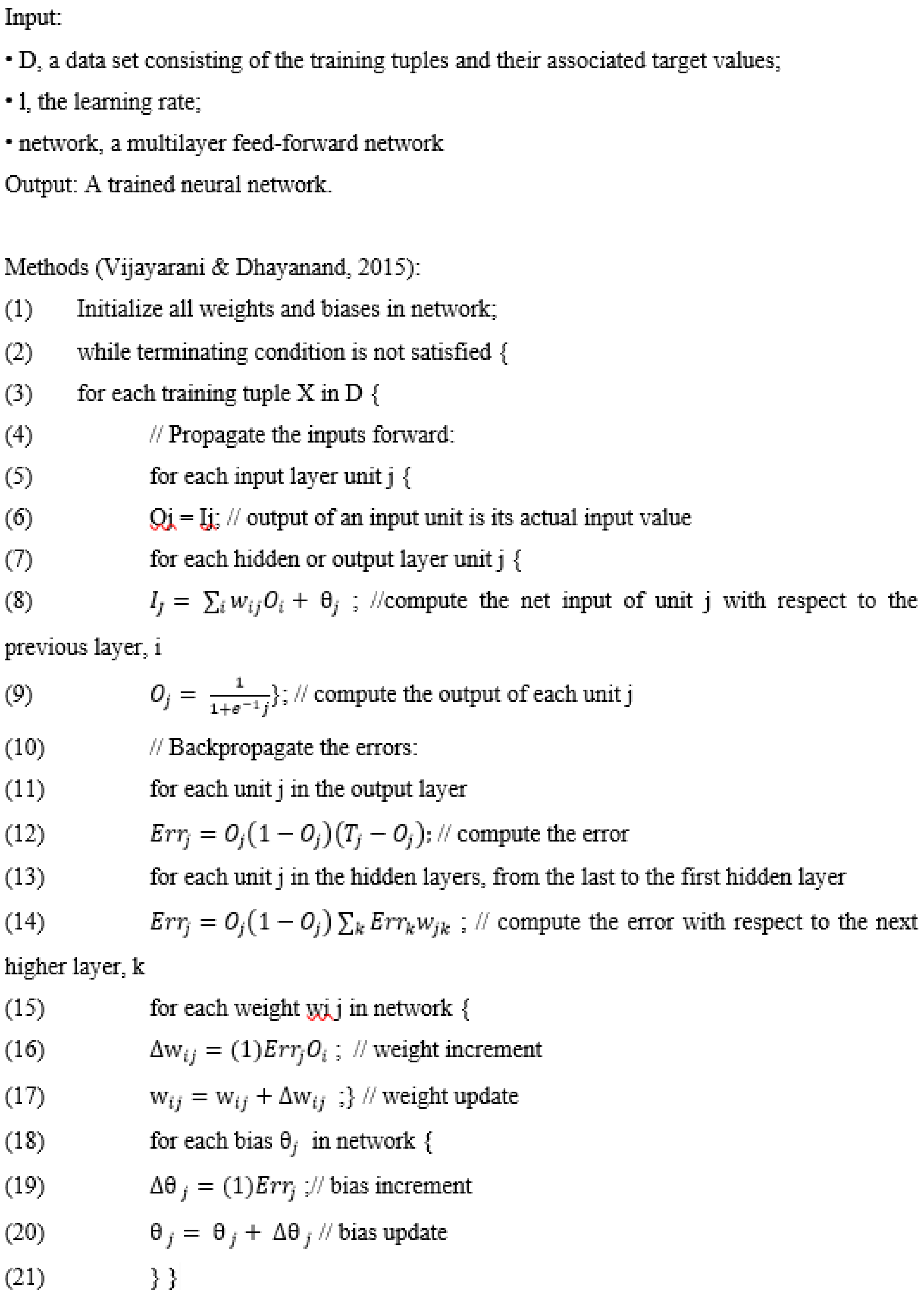
$$\sum_{i=1}^{n} \left(y_i - f_i(x)\right)^2 + \lambda \sum_{k=1}^{H} \sum_{j=0}^{P} \beta_{jk}^2 + \lambda \sum_{k=0}^{H} \gamma_k^2 \tag{3.5}$$

Equation 3.2, a linear function, is transformed by a nonlinear function. Another linear combination in Equation 3.4 connects these layers to the output. Equation 3.5 is a function of the penalty. Penalty terms are added to the regression coefficients, thus reducing the effects of large values. λ usually takes values between 0 and 1.

An artificial neural network consists of neurons for which calculations are made each time. There are three types of layers in which these neurons are specialized. These are the input layer, hidden layer, and output layer.

Backpropagation is compared with the estimated value for each network group with the known actual target value. Iteratively processes a dataset of training groups. The target value can be the known class label of the training bundle or a continuous value. The weights for each training group are calculated to minimize the mean squared error between the prediction of the network and the actual target value. These modifications are made backward from the output layer to the first hidden layer. That is why it is called backpropagation. Though not sure, the weights will eventually converge, and the learning process will stop. The steps involved in this process are inputs, outputs, and errors, and the algorithm is summarized below. (Vijayarani & Dhayanand, 2015)

The ANN algorithm's definition, given above, is shown in Figure 3.5. below.



Input:

- D, a data set consisting of the training tuples and their associated target values;
- l, the learning rate;
- network, a multilayer feed-forward network

Output: A trained neural network.

Methods (Vijayarani & Dhayanand, 2015):

```
Initialize all weights and biases in network;
while terminating condition is not satisfied {
for each training tuple X in D {
        // Propagate the inputs forward:
        for each input layer unit j {
        Oj = Ij; // output of an input unit is its actual input value
        for each hidden or output layer unit j {
        $I_j = \sum_i w_{ij} O_i + \theta_j$ ; //compute the net input of unit j with respect to the previous layer, i
        $O_j = \frac{1}{1+e^{-1}{}_j}$}; // compute the output of each unit j
        // Backpropagate the errors:
        for each unit j in the output layer
        $Err_j = O_j(1 - O_j)(T_j - O_j)$; // compute the error
        for each unit j in the hidden layers, from the last to the first hidden layer
        $Err_j = O_j(1 - O_j) \sum_k Err_k w_{jk}$ ; // compute the error with respect to the next higher layer, k
        for each weight wi j in network {
        $\Delta w_{ij} = (1) Err_j O_i$ ; // weight increment
        $w_{ij} = w_{ij} + \Delta w_{ij}$ ;} // weight update
        for each bias $\theta_j$ in network {
        $\Delta\theta_j = (1) Err_j$ ;// bias increment
        $\theta_j = \theta_j + \Delta\theta_j$ // bias update
        } }
```

Figure 3.5. The Steps of the ANN algorithm

3.3. Extreme Gradient Boosting (XGBoost)

There are three basic types of ensemble learning methods in the literature. These are named bagging, stacking, and boosting in Figure 3.6. In this thesis, boosting algorithms will be explained in detail, which is why the algorithms are used as one of the leading models in this study. XGBoost is an ensemble learning algorithm based on gradient boosting, which is also an optimization model that combines a linear model with an increment tree model. It also uses the loss function's first and second derivatives for the second-order derivative (Yu and Others, 2019).

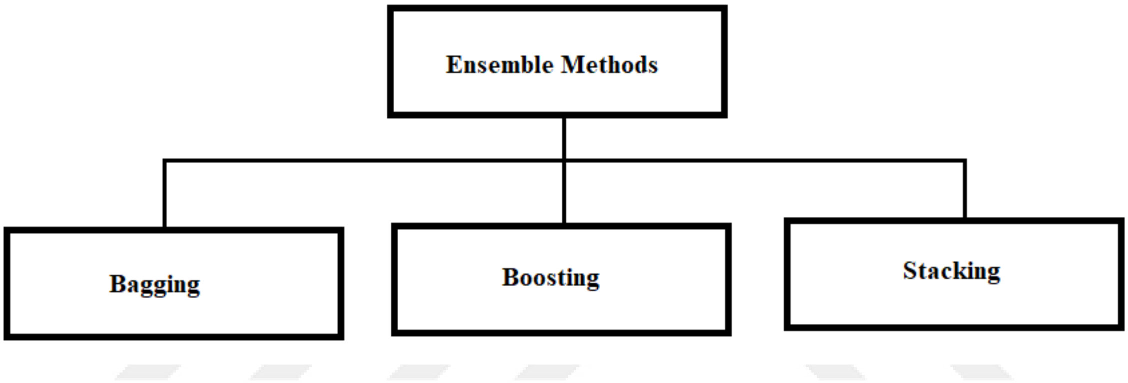


Figure 3.6. Basic types of ensemble methods

XGBoost is a scalable machine-learning algorithm for tree boosting (Chen & Guestrin, 2016). The algorithm's four most important features are obtaining high predictive power, preventing over-learning, managing open data, and doing them more quickly.

Equation 3.8 combines the tree model with together supplement method, uses K additive functions to predict the output, and uses F to show the simple tree model (Li, Yin, Quan, & Zhang, 2019).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \qquad (3.8)$$

The objective function is given as follows:

$$L = \sum_i \; l(\hat{y}_i, y_i) + \sum_k \; \Omega(f_k) \qquad (3.9)$$

where L is the loss function indicating the error between the prediction data and the test data; Ω is the indicating function for regularization to avoid overfitting and calculated as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2 \tag{3.10}$$

The T value represents the number of leaves per tree, while the w value represents the weight of each tree's leaves.

First, the quadratic Taylor expansion of the objective function and other calculations detailed in the Supplementary Material is made. It can then obtain the information gained of the objective function after each division.

Then, the quadratic Taylor expansion of the objective function and other calculations detailed in the Supplementary Material is made. Then, the information gain of the objective function after each division is calculated as follows:

$$Gain = \frac{1}{2}\left[\frac{\left(\sum_{i\epsilon I_L} g_i\right)^2}{\sum_{i\epsilon I_L} h_i+\lambda} + \frac{\left(\sum_{i\epsilon I_R} g_i\right)^2}{\sum_{i\epsilon I_R} h_i+\lambda} + \frac{\left(\sum_{i\epsilon I} g_i\right)^2}{\sum_{i\epsilon I} h_i+\lambda}\right] - \gamma \tag{3.11}$$

The introduction of the XGBoost regular loss function outperforms other tree-boosting methods due to the weights of each new tree. The fact that a specific constant n can scale down reduces the effect of a single tree on the final score and column sampling, which works similarly to random forests (Pan, 2017).

XGBoost has many advantages, some of which are given below.

- It is highly compatible.
- It can achieve high predictive power.
- It takes advantage of parallel processing.
- It can prevent excessive learning.
- Ability to manage missing data and make them fast.
- After each iteration, the user can perform cross-validation.

3.4. Performance Measures

One widely accepted criterion in choosing various forecasting models is that the model fits well with the data. The predictive success of the model should be high. For example, when the prediction successes of the two models are compared, the model that provides better prediction accuracy is preferred. In this context, various statistics are used to compare the predictive accuracy of the models.

Performance measurements are made with the data estimated by the test data after the estimation is made. There are many different performance measures. The most common of these used in this thesis are MSE, RMSE, MAE, R^2, and MAPE. Comparisons are made using these functions. Performance measurements are available for each installed model. Moreover, using the measurement results, the estimation results are evaluated. In this way, the best prediction result is determined.

According to Willmott & Matsuura (2005), the model error is calculated as e_i, i = 1, 2,…, n, assuming there are 1 to n samples.

In general, the mean model-estimation error can be written as follows

$$\underline{e}_{\gamma} = \left[\frac{\sum_{i=1}^{n} w_i |e_i|^{\gamma}}{\sum_{i=1}^{n} w_i}\right]^{\frac{1}{\gamma}} \qquad (3.12)$$

MSE (Mean Squared Error): MSE is found by taking the sum of the squares of the difference between the actual values and the predicted values in the data set and dividing the result by the number of samples. It can be said that the prediction model performs better when the MSE value is close to zero (Yurttabir & Sen, 2021).

$$MSE = \ [n^{-1} \sum_{i=1}^{n} |e_i|^2] \qquad (3.13)$$

RMSE (Root Mean Square Error): The RMSE, which is the standard deviation of the prediction errors, indicates how spread these errors are, and an RMSE value of zero means that the model made no errors (Yurttabir & Sen, 2021).

$$RMSE = \ \lfloor n^{-1} \sum_{i=1}^{n} |e_i|^2 \rfloor^{\frac{1}{2}} \qquad (3.14)$$

MAE (mean absolute error): This measure is called MAE (Willmott & Matsuura, 2005).

$$MAE = \lceil n^{-1} \textstyle\sum_{i=1}^{n} |e_i| \rceil \tag{3.15}$$

R^2 Score: It can be defined as the ratio of the variance in the dependent variable that can be predicted from the independent variables (Chicco, Warrens, & Jurman, 2021).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{3.16}$$

MAPE (Mean Absolute Percentage Error): Percentile errors have the advantage of being unitless. Therefore, they are often used to compare prediction performances between dataset fields (Yurttabir & Sen, 2021).

$$MAPE = \frac{100}{n} \sum_{j}^{n} \frac{|e_j|}{|A_j|} \tag{3.17}$$

4. ANALYSIS AND COMPUTATIONAL RESULTS

This thesis uses Apricot Export Volume 2002-2019 (included) data as input. Apricot Export Volume 2020 data is chosen as the test data. Apricot export volume data taken from TURKSTAT are used as input data in all proposed models. This study estimates the apricot export amount using AR, ARIMA, SARIMA, ANN, and XGBoost models. Five models are used, and the results of these methods are compared. The models are AR, ARIMA, SARIMA, ANN, and XGBoost models.

Korelogram (Auto-correlation function (ACF)) gives the relationship between the autocorrelation coefficient of the series and the lag values. The partial autocorrelation function (PACF) gives the delays between the two lags; the other lags are unimportant in the partial correlation (Hyndman, 2014). ACF and PACF graphs are used to understand whether there is seasonality in a data set.

ACF and PACF graphs are given in Figure 4.1 and Figure 4.2, respectively. As can be seen in Figure 4.1, the series does not have a trend. Because the top 4 most considerable lag on the ACF graph is out of bounds simultaneously, the series does not have a trend, and this series is stationary. The second most significant lag is 12 values after the first most considerable lag in the ACF, indicating a seasonality with a period of 12. That is why we examine the PACF graph. As can be seen in Figure 4.2, High autocorrelation is observed in the data because positive and negative values are observed after a spike in the first lag. Here, the second largest value is the next bar, indicating that this data is seasonal.

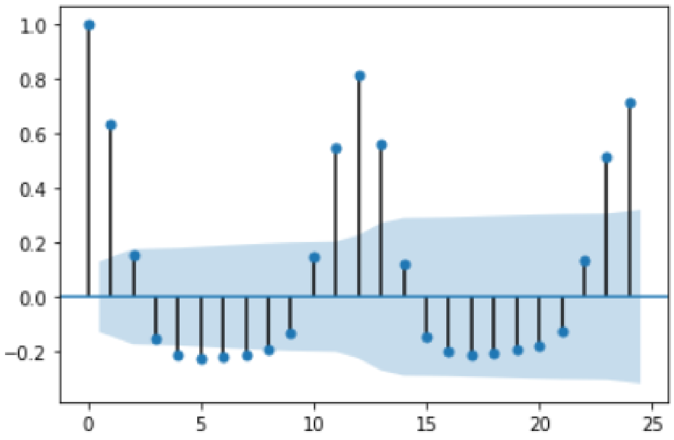


Figure 4.1 Autocorrelation plot of the dataset

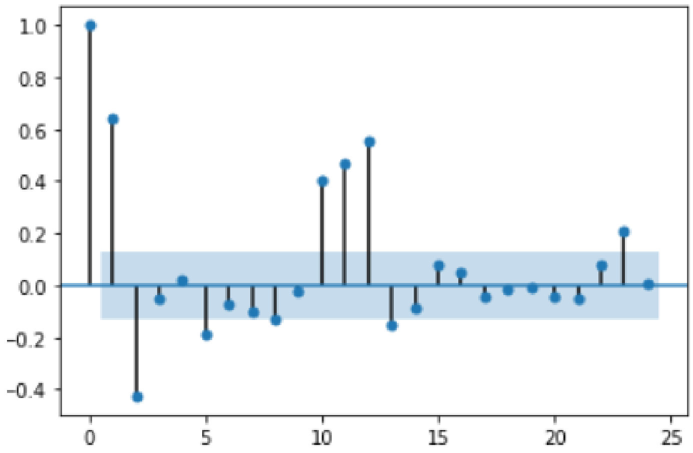
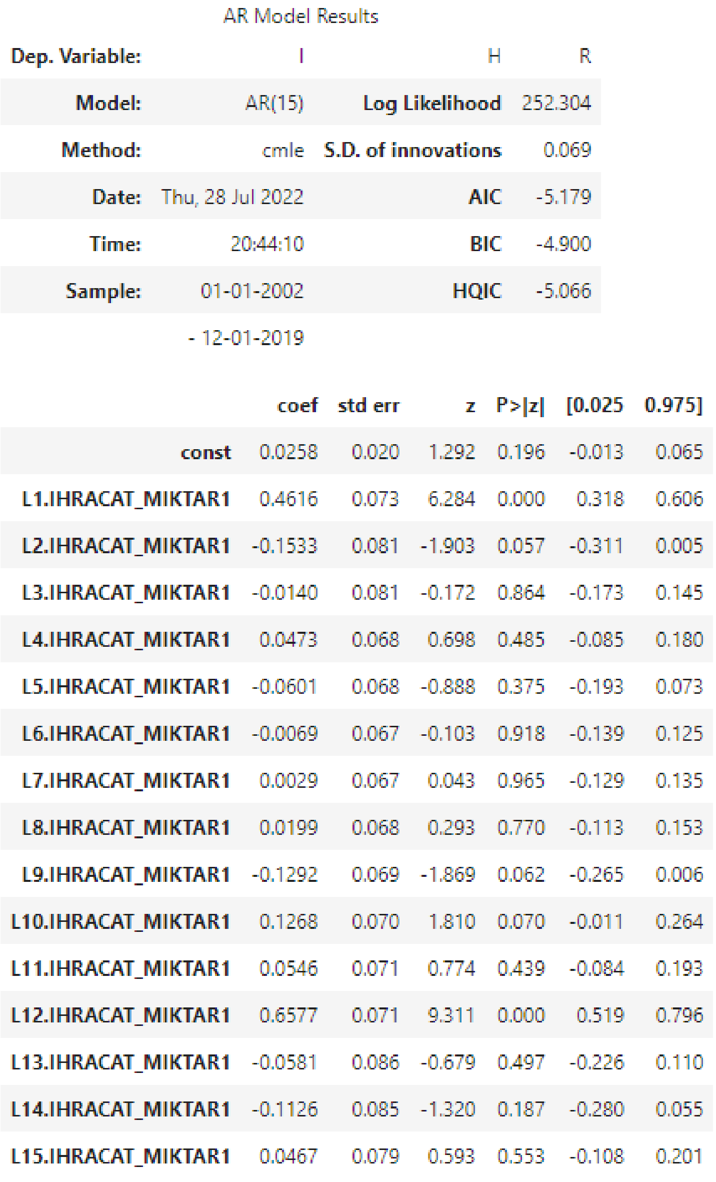


Figure 4.2 Partial Autocorrelation plot of the dataset

Proposed models are coded by using Python 3.8 via JupyterLab on Windows 10 Pro, 64-bit (x64) platforms, Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz.

4.1. Results of the AR Model

The statsmodels.tsa.ar_model model from the statsmodels library is used for the AR model. First, since there is seasonality in the data, the data is scaled by the min-max scaler and AR is modeled by using the training set. Then, the AR model is fit to the training data set. After training, prediction value is obtained for the year 2020. AR Model results are given in Figure 4.3. From AR (15) model, it can be understood that the lower the AIC and BIC values, the better our model is.



AR Model Results

Dep. Variable:	I	H	R
Model:	AR(15)	**Log Likelihood**	252.304
Method:	cmle	**S.D. of innovations**	0.069
Date:	Thu, 28 Jul 2022	**AIC**	-5.179
Time:	20:44:10	**BIC**	-4.900
Sample:	01-01-2002	**HQIC**	-5.066
	- 12-01-2019		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	0.0258	0.020	1.292	0.196	-0.013	0.065
L1.IHRACAT_MIKTAR1	0.4616	0.073	6.284	0.000	0.318	0.606
L2.IHRACAT_MIKTAR1	-0.1533	0.081	-1.903	0.057	-0.311	0.005
L3.IHRACAT_MIKTAR1	-0.0140	0.081	-0.172	0.864	-0.173	0.145
L4.IHRACAT_MIKTAR1	0.0473	0.068	0.698	0.485	-0.085	0.180
L5.IHRACAT_MIKTAR1	-0.0601	0.068	-0.888	0.375	-0.193	0.073
L6.IHRACAT_MIKTAR1	-0.0069	0.067	-0.103	0.918	-0.139	0.125
L7.IHRACAT_MIKTAR1	0.0029	0.067	0.043	0.965	-0.129	0.135
L8.IHRACAT_MIKTAR1	0.0199	0.068	0.293	0.770	-0.113	0.153
L9.IHRACAT_MIKTAR1	-0.1292	0.069	-1.869	0.062	-0.265	0.006
L10.IHRACAT_MIKTAR1	0.1268	0.070	1.810	0.070	-0.011	0.264
L11.IHRACAT_MIKTAR1	0.0546	0.071	0.774	0.439	-0.084	0.193
L12.IHRACAT_MIKTAR1	0.6577	0.071	9.311	0.000	0.519	0.796
L13.IHRACAT_MIKTAR1	-0.0581	0.086	-0.679	0.497	-0.226	0.110
L14.IHRACAT_MIKTAR1	-0.1126	0.085	-1.320	0.187	-0.280	0.055
L15.IHRACAT_MIKTAR1	0.0467	0.079	0.593	0.553	-0.108	0.201

Figure 4.3. AR Model Results

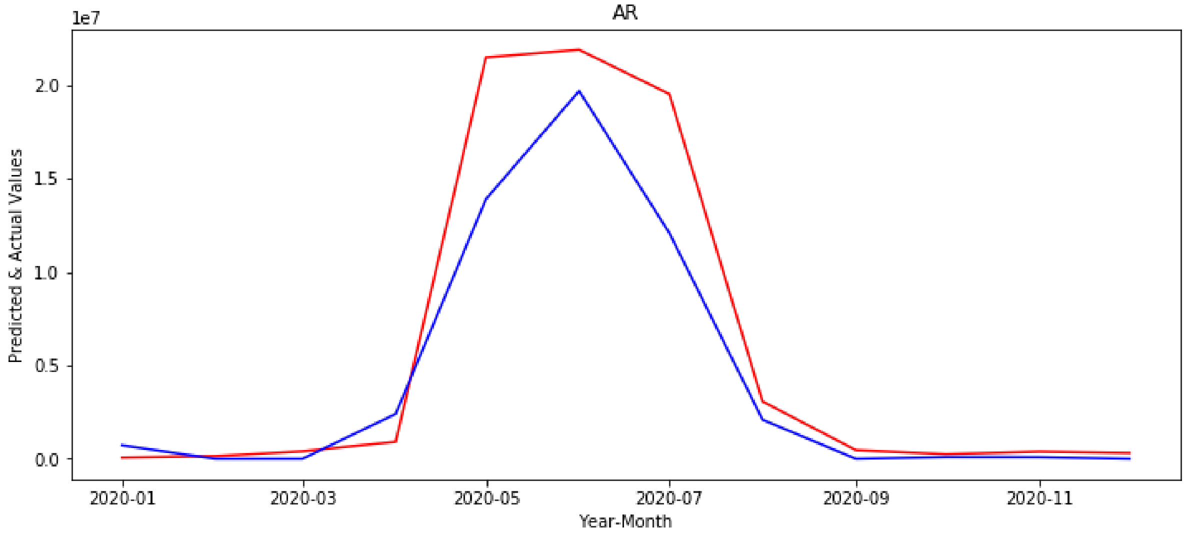


Figure 4.4. Comparison of Predicted and Actual Values for AR Model (Red =Actual, Blue = Prediction)

Predicted and actual values are shown in Figure 4.4. In the figure, actual values are shown in red, and estimated values are in blue. This study uses MAE, MSE, RMSE, MAPE, and R^2 to measure the AR Model's prediction accuracy. The results are given in Table 4.1. When the results are examined, the AR model reasonably estimates this data set.

Table 4.1. Performance metrics of AR Model

Performance measurement	**Result**
MAE	1,844,787.34
MSE	10,159,197,550,470.90
RMSE	3,187,349.90
R^2	0.87
MAPE	152.18

According to the MAE value, the monthly apricot export volume can be explained by 1,844,787. MSE value shows that the mean squared error in monthly apricot exports is 10,159,197,550,470. RMSE value gives the root mean squared error in monthly apricot exports is 3,187,349. R^2 score, which is the agreement of the prediction with the actual values in monthly apricot exports, is calculated as 0.87. MAPE value, the mean absolute error percentage, is 153% for the monthly export volume of apricot.

4.2. Results of the ARIMA model

The statsmodels.tsa.arima_model model from the statsmodels library is used for the ARIMA model. First, the data is scaled by the min-max scaler since there is seasonality in the data. After, ARIMA is modeled by using a training set. Then, the ARIMA model is fit to the training dataset.

Trend Elements for the ARIMA model are given as follows:

p: Trend autoregression order.

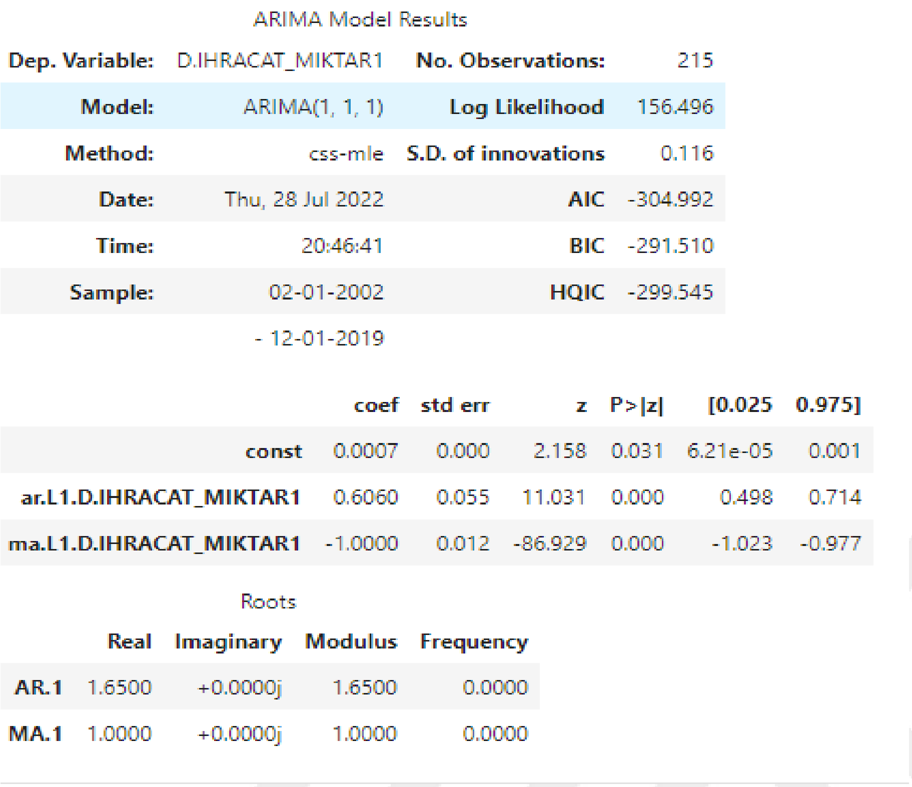
d: Trend difference order.

q: Trend moving average order.

All values between 0 and 5, used as parameters, are tested on the model. Among these parameters, the lowest AIC value is obtained by ARIMA (1,1,1) model as 7200,137947360675

After model training, the model is tested by using data for 2020. ARIMA model results are given in Table 4.2. It is seen that the AIC and BIC values are higher than the AR model. For this reason, it is observed to give worse results than the AR model. As seen from the results, this data set is not suitable for the ARIMA model since seasonality is observed in the data.

Table 4.2. ARIMA Model Results



ARIMA Model Results

Dep. Variable:	D.IHRACAT_MIKTAR1	**No. Observations:**	215
Model:	ARIMA(1, 1, 1)	**Log Likelihood**	156.496
Method:	css-mle	**S.D. of innovations**	0.116
Date:	Thu, 28 Jul 2022	**AIC**	-304.992
Time:	20:46:41	**BIC**	-291.510
Sample:	02-01-2002	**HQIC**	-299.545
	- 12-01-2019		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	0.0007	0.000	2.158	0.031	6.21e-05	0.001
ar.L1.D.IHRACAT_MIKTAR1	0.6060	0.055	11.031	0.000	0.498	0.714
ma.L1.D.IHRACAT_MIKTAR1	-1.0000	0.012	-86.929	0.000	-1.023	-0.977

Roots

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	1.6500	+0.0000j	1.6500	0.0000
MA.1	1.0000	+0.0000j	1.0000	0.0000

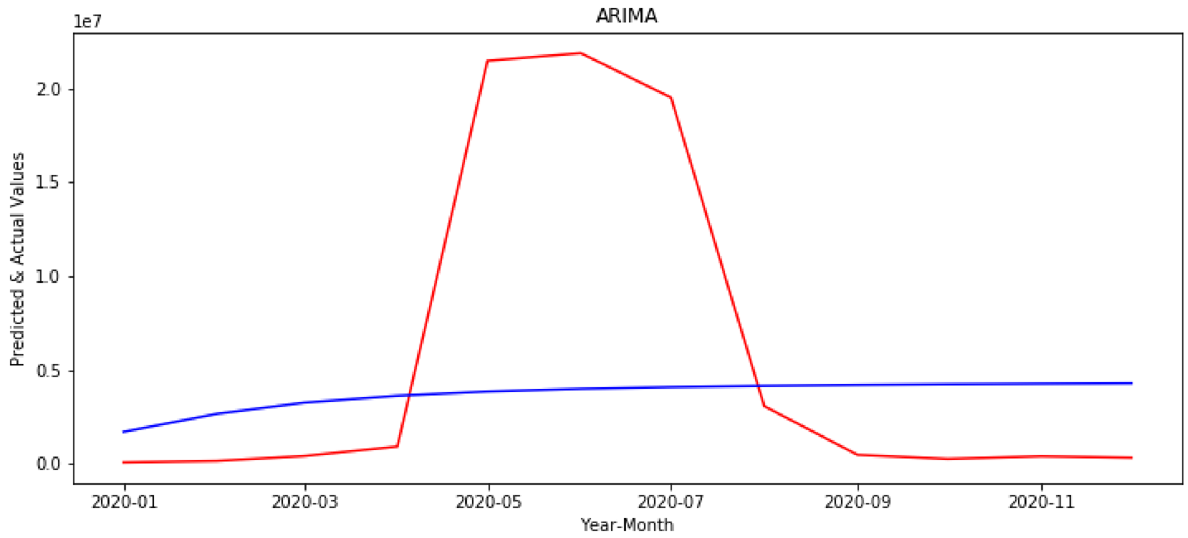


Figure 4.5. Comparison of Predicted and Actual Values for ARIMA Model (Red =Actual, Blue = Prediction

The actual values and the prediction values are given in Figure 4.5. This figure shows actual values in red and estimated values in blue. When the previously published studies are examined, it is determined that the ARIMA model is not suitable for seasonal models. Hence, a more unsuccessful result is obtained compared to other models (Table 4.3).

Table 4.3. Performance metrics of ARIMA Model

Performance measurement	**Result**
MAE	6,439,413.88
MSE	79,584,073,642,548.41
RMSE	8,920,990.62
R2	-0.02
MAPE	847.92

According to the MAE value, the monthly apricot export volume can be explained by 6,439,414. MSE value shows that the mean squared error in monthly apricot exports is 79,584,073,642,548. RMSE value gives the root mean squared error in monthly apricot exports is 8,920,991. R^2 score shows the agreement of the prediction with the actual values in monthly apricot exports is -0.02. MAPE value indicates that the monthly apricot exports' mean absolute error percentage is 848.

4.3. Results of the SARIMA model

For the SARIMA model, the statsmodels.tsa.statespace.sarimax model from the statsmodels library is used. First, since there is seasonality in the data, the data is scaled by the min-max scaler. After, SARIMA is modeled by using the training set. Then, the SARIMA model is trained.

Trend Elements
p: Trend autoregression order.
d: Trend difference order.
q: Trend moving average order.

Seasonal Elements
P: Seasonal autoregressive order.
D: Seasonal difference order.
Q: Seasonal moving average order.
m: The number of time steps for a single seasonal period.

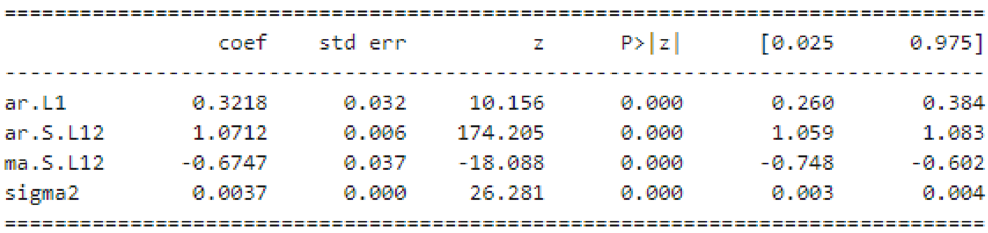
Integer values between 0 and 2 for parameters of P, d, q P, D, Q, and m are tested on the model. Within these parameters, the SARIMA (1, 1, 1) x (1, 1, 1, 12)12 model is chosen as the best model with the lowest AIC value. The lowest AIC value is obtained by SARIMA (1, 1, 1) x (1, 1, 1, 12)12 model as 6136,780935334777. In addition, the factors of stationarity and invertibility are set to False.

Table 4.4. Best Parameters Used in SARIMA Model

Best Parameters
order=(1,0,0)
seasonal_order=(1,0,1,12)
enforce_stationarity=False
enforce_invertibility=False

The model is trained by using the best parameters given in Table 4.4. After training the model, 2020 data predictions are used for testing. Considering the results of the comparison with the test set, it is observed that the SARIMA model gave better results than AR and ARIMA models (Table 4.5).

Table 4.5. SARIMA Model Results



```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3218      0.032     10.156      0.000       0.260       0.384
ar.S.L12       1.0712      0.006    174.205      0.000       1.059       1.083
ma.S.L12      -0.6747      0.037    -18.088      0.000      -0.748      -0.602
sigma2         0.0037      0.000     26.281      0.000       0.003       0.004
==============================================================================
```

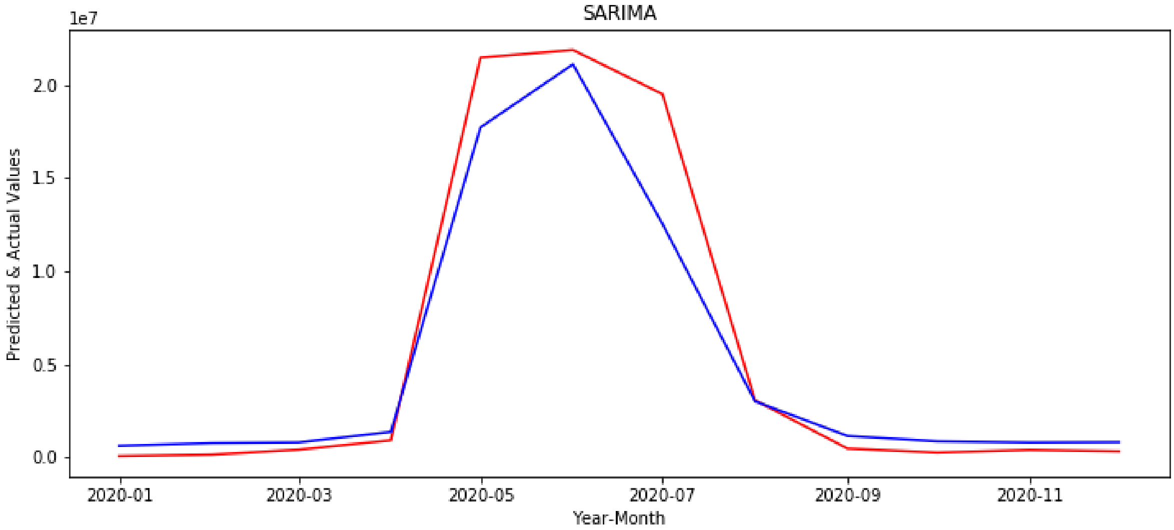


Figure 4.6. Comparison of Predicted and Actual Values for SARIMA Model (Red =Actual, Blue = Prediction

The actual values and the prediction values are given in Figure 4.6. In this figure, actual values are shown in red, and estimated values are in blue. The results are given in Table 4.6. When the results are examined, it can be said that the SARIMA model makes a more successful prediction than AR and ARIMA models from Time Series Models in this data set.

Table 4.6. Performance metrics of SARIMA

Method	**Result**
MAE	1,313,442.73
MSE	5,468,445,848,577.23
RMSE	2,338,470.84
R2	0.93
MAPE	177.41

MAE sums the absolute error value. It is a more direct representation of the sum of error terms. According to the MAE value, the monthly apricot export amount is 1,313,443. One cannot interpret much insight from a single result, but MSE gives an actual number to compare with other model results and helps to choose the best predictive model. The MSE value shows 5,468,445,848,577 of the average square error in monthly apricot exports. RMSE is the square root of MSE. The RMSE value gives the average squared error of 2,338,471 monthly exports of apricots. R^2 score shows that the agreement of the prediction with the actual values in monthly apricot exports is 0.93. It should be noted that the R^2 value is an indicator of the strength of the relationship between dependent and independent variables, and the correctness of the model is expected to increase with higher R^2 values. MAPE value is calculated as 177 for the export volume of monthly apricot by SARIMA.

4.4. Results of the ANN model

For the ANN model, the sklearn.neural_network _model MLPRegressor model from the sklearn library is used. First, the data is scaled by the min-max scaler since there is seasonality in the data. After, ANN is modeled by using the training set. The ANN model's Parameters given in Table 4.7 are selected to determine the best parameters. GridSearchCV is used to determine the best parameter using negative MSE as the objective function. According to GridSearchCV, the best negative MSE value is found as -0,01402061580943719. The best parameters found by GridSearchCV are {'activation': 'relu', 'alpha': 0.1, 'hidden_layer_sizes': (50, 100, 150)}

Table 4.7. Considered Parameter Values to Choose the Best Parameter for ANN Model

Parameter Name	**Value**
alpha	0.005, 0.01, 0.1
hidden_layer_sizes	(20,20,20), (50,100,150), (150,200,300)
activation	'relu', 'logistic'
Cv	10
Scoring	neg_mean_squared_error
n_jobs	-1
Verbose	2

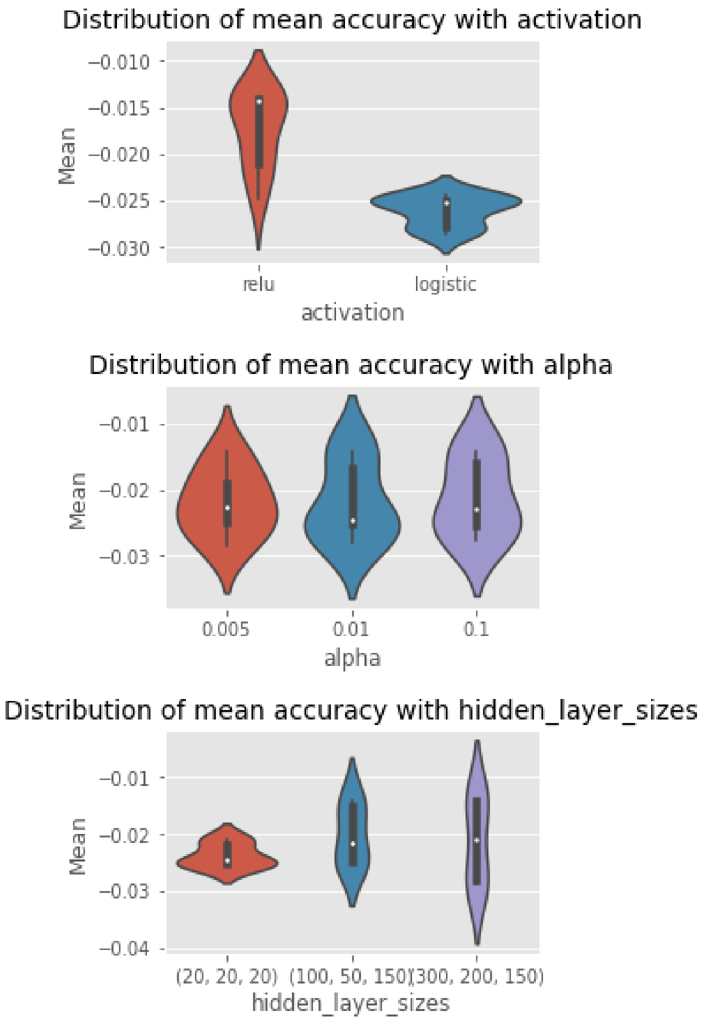


Figure 4.7. Violin Plot of Apricot Export Volume Hyperparameters for the ANN Model

The parameters of the ANN model are represented by scatter plots. Hyperparameters of each analysis in the ANN model are visualized with a violin plot. The violin plot from Python's Seaborn library is used to visualize it. Hintze and Nelson (1998) mention the violin plot as follows: the violin plot shows the combination of the box plot of the data and the probability density curve on a single plot. If we look at the structure of a violin chart, unlike a boxplot, the median is not a straight line. The white dot on the violin represents the median, allowing easy comparison of alternatives. The vertically elongated thin black line represents 95% confidence intervals. The width of the violin corresponds to the frequency of the data points in each region. The ends of the box in the middle of the violin represent the first and third quarters. The tail of the violin shows the outer points. The fiddle chart provides quick and meaningful insights into statistics about how the values in the data are distributed for each piece of data. The graphs of all hyperparameter alternatives evaluated in the grid search are given above. Figure 4.7. visualizes each hyperparameter performed for apricot export volume in the ANN model.

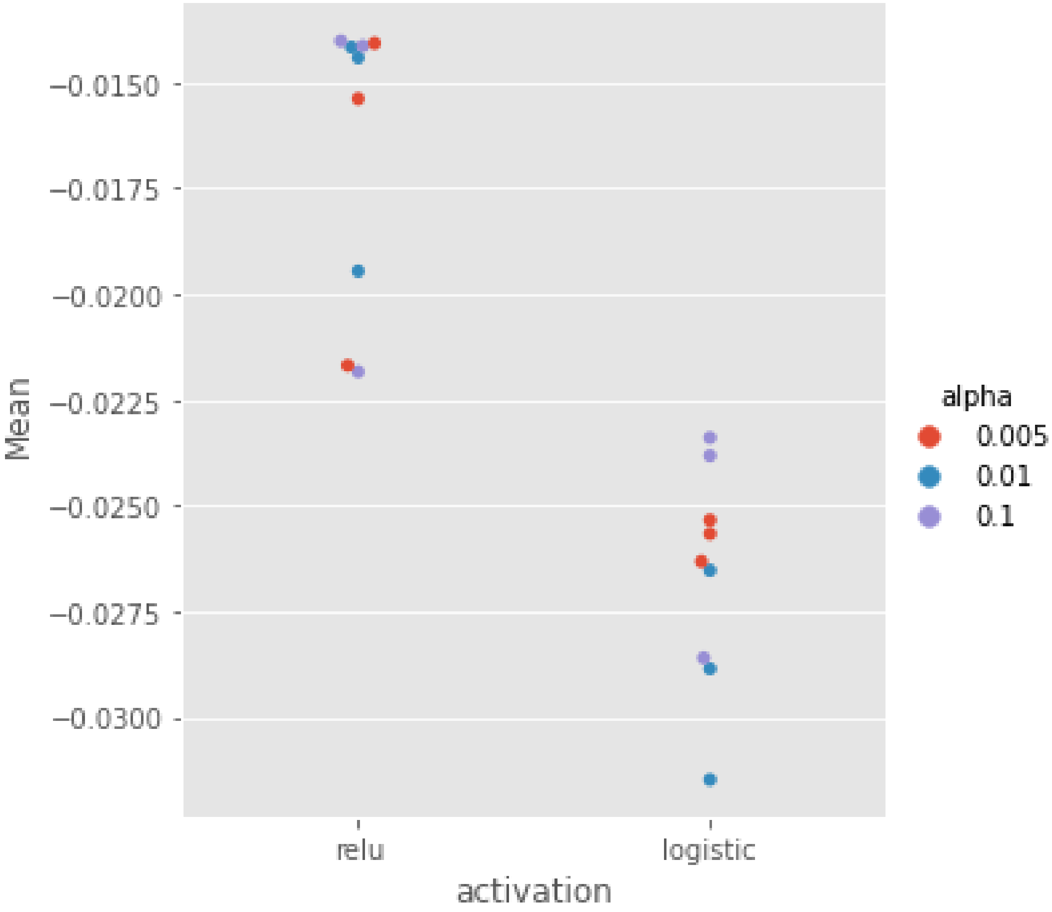


Figure 4.8. Catplot of Alpha Hyperparameters Optimization for the ANN Model

Graphs play essential roles in understanding and interpreting the relationship between variables. These graphs vary depending on whether the data is categorical or numerical. Many representations in the literature vary depending on the data type or the feature to compare. When these graphics are read and interpreted correctly, they are both catchy with their visuals and help to summarize much information effectively. The hyperparameters of the ANN model are optimized with the grid search algorithm, which has activation and alpha categorical data types. The data range of activation is selected as relu and logistic. Also, the alpha data values are 0.1, 0.01, and 0.005. To show the relationship of categorical data between activation and its location relative to the mean on the alpha y-axis was made with catplot (Figure 4.8). It is plotted with the catplot plot in the Python seaborn library.

After the ANN model is trained using the best parameters, 2020 data is used for the test.

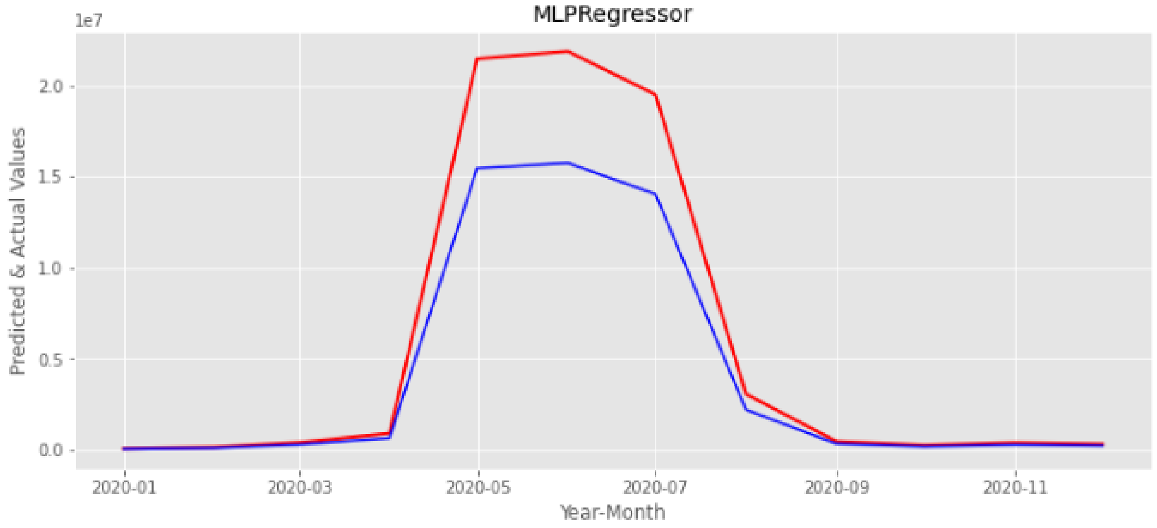


Figure 4.9. Comparison of Predicted and Actual Values for ANN Model (Red =Actual, Blue = Prediction)

Test results are shown in Figure 4.9. In this table, actual values are shown in blue, and estimated values are in red. The results are given in Table 4.8. When the results are examined, it can be said that the ANN model makes a much better prediction in this data set than all other models.

Table 4.8. Performance metrics of ANN Model

Method	Result
MAE	1,607,515.58
MSE	8,704,172,587,837.48
RMSE	2,950,283.21
R^2	0.88
MAPE	28.01

According to the MAE value, the ANN model calculated the monthly apricot export amount as 1,607,516. The MSE value, which helps choose another best forecasting model, shows 8,704,172,587,837 of the mean squared error in monthly apricot exports. RMSE is the square root of MSE. The RMSE value gives the average squared error of 2,950,283 monthly exports of apricots. R2 score, which is the agreement of the prediction with the actual values in monthly apricot exports, is calculated as 0.88. MAPE value, the mean absolute error percentage, is 28% for the monthly export volume of apricot.

4.5. Results of the XGBOOST model

For the XGBoost model, the XGBRegressor model from the XGBoost library is used. First, the data is scaled by using the min-max scaler since there is seasonality in the data. Then, the XGBoost model is trained. The XGBoost model's Parameters given in Table 4.9 are selected to determine the best parameters. GridSearchCV is used to determine the best parameter. According to GridSearchCV, the best parameter result is obtained as -0,01678724130506497. The best parameters are {'colsample_bytree': 0.1, 'learning_rate': 0.1, 'max_depth': 2, 'n_estimators': 50} by GridSearchCV (Table 4.9).

Table 4.9. Considered Parameters to Determine the Best Parameters for XGBoost Model

Parameter name	**Values**
colsample_bytree	0.05, 0.5, 0.1, 0.2, 0.3
n_estimators	50, 80, 100, 200
max_depth	2,3,4,5,6
learning_rate	0.5, 0.1, 0.01
Cv	10
scoring	neg_mean_squared_error
n_jobs	-1
verbose	2

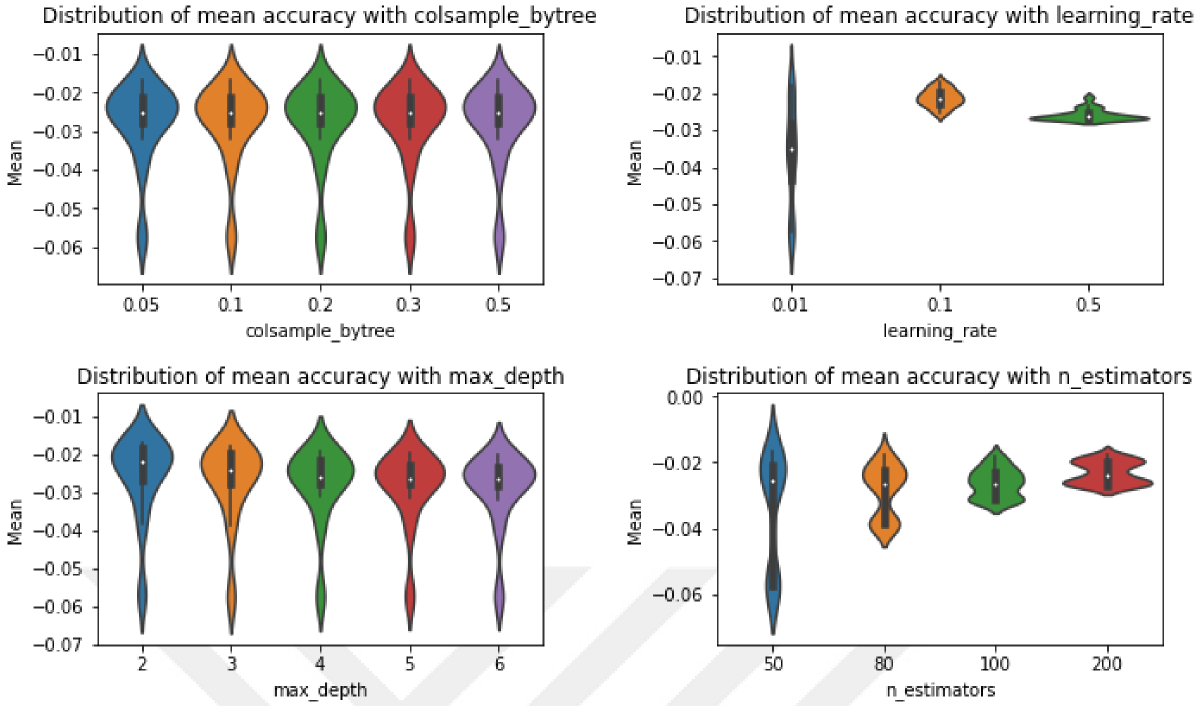


Figure 4.10. The Violin Plot of Apricot Export Volume Hyperparameters for the XGBoost Model

The violin plot provides quick and meaningful insights into statistics about how the values in the data are distributed for each piece of data. The graphs of all hyperparameter alternatives evaluated in the grid search are given above. Figure 4.10. visualizes each hyperparameter performed for apricot export volume in the XGBoost model. The hyperparameters used are colsample_bytree, learning_rate, max_depth, and n_estimators. A comparative graph of the values assigned to these hyperparameters is shown above. The best parameters obtained using GridSearchCV are 'colsample_bytree': 0.1, 'learning_rate': 0.1, 'max_depth': 2, and 'n_estimators': 50.

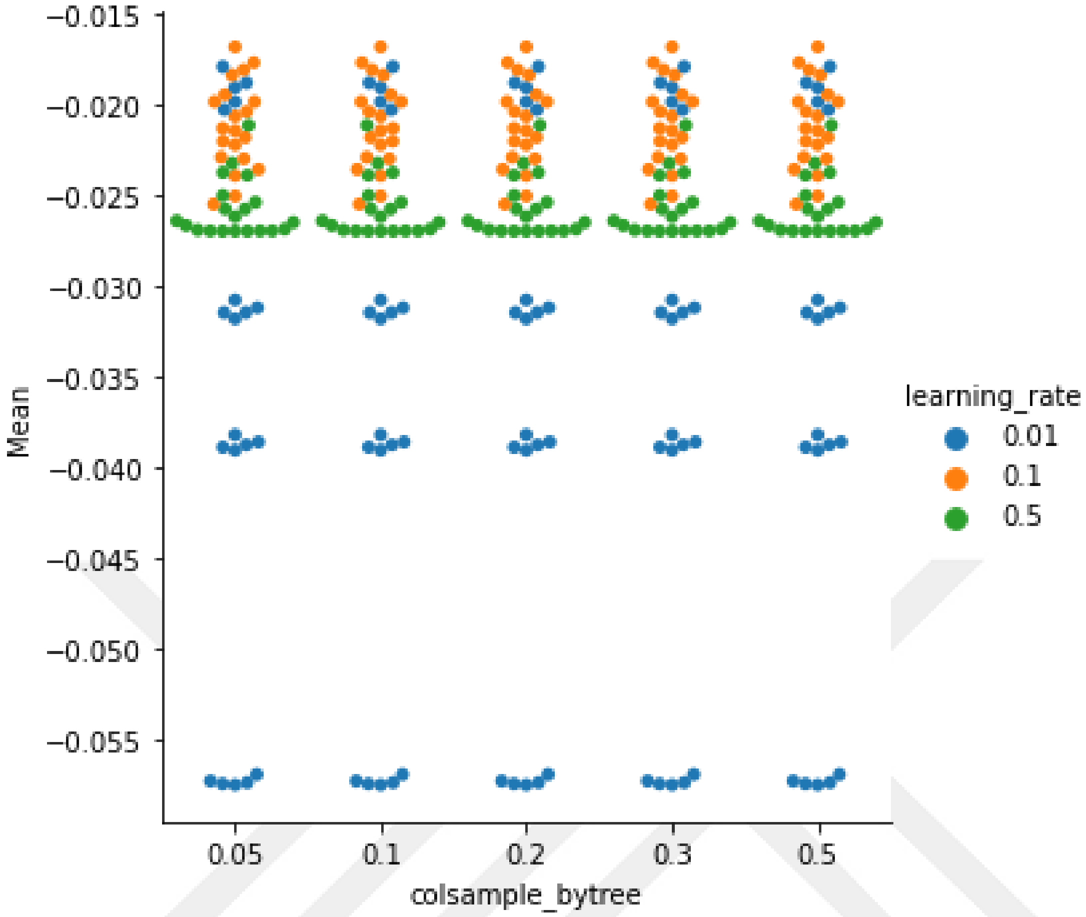


Figure 4.11. The Catplot of Alpha Hyperparameters Optimization for the XGBoost Model

The hyperparameters of the XGBoost model are optimized with the grid search algorithm. It has activation and alpha categorical data types from hyperparameters. The data range of activation is relu and logistic. Also, the alpha data values are 0.1, 0.01, and 0.005. To show the relationship of categorical data, activation and its location relative to the mean on the alpha y-axis were made with catplot. It is plotted with the catplot plot in the Python seaborn library. When the learning rates in the graph are examined, it is observed that the 0.01 learning rate is very scattered. It is observed that it is much more clustered and aggregated at a learning rate of 0.1. A learning rate of 0.5 is observed to be clustered but has a lower average.

After the XGBoost model is trained by using the best parameters, 2020 data is used for test

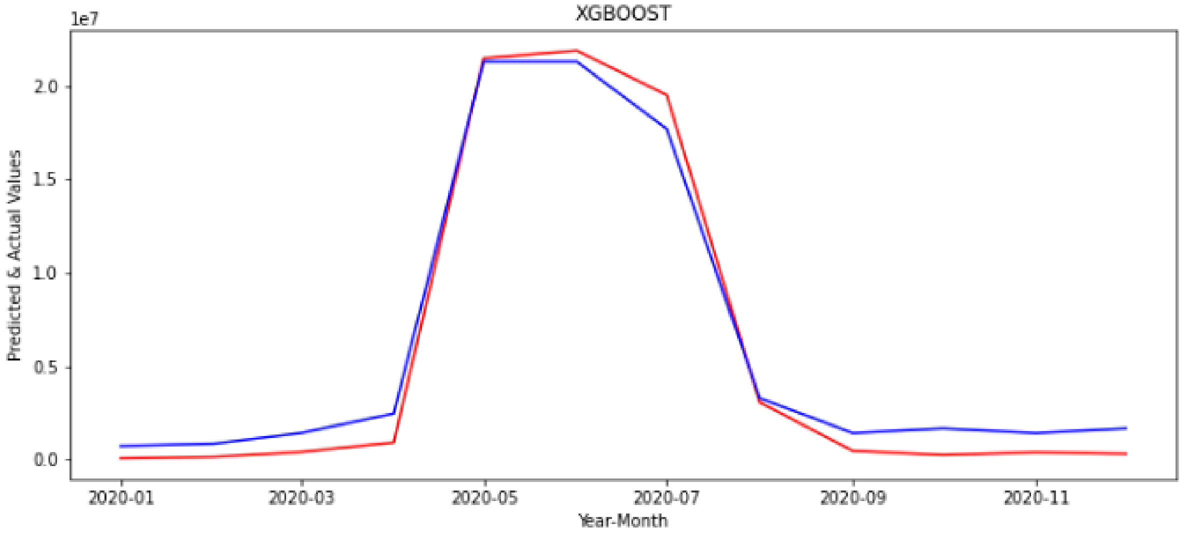


Figure 4.12. Comparison of Predicted and Actual Values for XGBoost Model (Red =Actual, Blue = Prediction)

Test results are shown in Figure 4.12. In this figure, actual values are shown in blue, and estimated values are in red. Also, performance metrics of the XGBoost are given in Table 4.10.

Table 4.10. Performance metrics of XGBoost Model Results

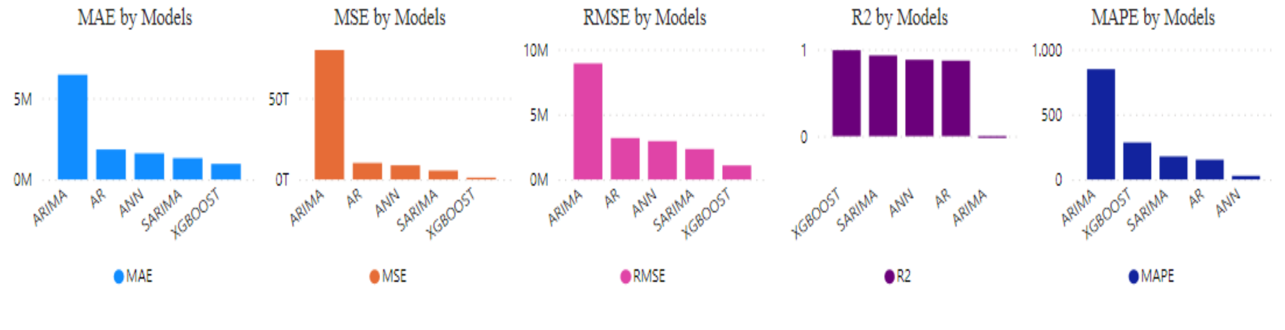
Method	**Result**
MAE	959,249.10
MSE	1,162,081,959,332.23
RMSE	1,077,999.05
R^2	0.99
MAPE	284.30

According to the MAE value, the ANN model calculated the monthly apricot export amount as 959,249. The MSE value, which helps to choose another best forecasting model, is shown as 1,162,081,959,332 of the mean squared error in monthly apricot exports. RMSE is the square root of MSE. The RMSE value gives the average squared error of 1,077,999 monthly exports of apricots. R^2 score, which is the agreement of the prediction with the actual values in monthly apricot exports, is calculated as 0.99. MAPE value, the mean absolute error percentage, is 284% for the monthly export volume of apricot.

4.6. Comparative Analysis

In this section of the thesis, the prediction results of the apricot export amount obtained by AR, ARIMA, SARIMA, ANN, and XGBOOST methods are compared, and their performances are evaluated. The data between 2002 and 2019 is used for four models; comparisons are made based on predictions for the 12 months of 2020.

Table 4.11. Comparative Model Results



	AR	ARIMA	SARIMA	ANN	XGBOOST
MAE	1,844,787.34	6,439,413.88	1,313,442.73	1,607,515.58	959,249.10
MSE	10,159,197,550,470.90	79,584,073,642,548.41	5,468,445,848,577.23	8,704,172,587,837.48	1,162,081,959,332.23
RMSE	3,187,349.90	8,920,990.62	2,338,470.84	2,950,283.21	1,077,999.05
R2	0.87	-0.02	0.93	0.88	0.99
MAPE	152.18	847.92	177.41	28.01	284.30

The current data for 2020 and the results of the models are compared in Table 4.11. As seen in Table 4.11, it is observed that the lowest error rates are obtained by the XGBoost model. Among the models compared, the XGBoost is the most appropriate model in the seasonal apricot export volume data.

For the R^2 score, sklearn.metrics.r2_score from the sklearn library is used. A negative value is found in the ARIMA model. The best possible score is 1.0, which can be negative because the model can be arbitrarily worse. Except for MAPE, XGBoost is more successful on other performance measures examined.

5. CONCLUSIONS AND RECOMMENDATIONS

Agriculture is an important sector that directly or indirectly affects almost every area of the economy. This situation does not change anywhere in the world, including Turkey. The role of the agricultural sector is enormous both for the factors of production and the foreign trade balance. For this reason, it is of great importance to understand the agricultural sector's potential and to analyze the sector correctly in terms of its impact on economic factors.

In this study, the apricot export volume of Turkey, which is an agricultural country and ranks first in the world in apricot exports, is predicted by using artificial intelligence techniques. Apricot export data monthly between 2002 – 2020 are obtained from TURKSTAT. This thesis aims to get information about Turkey's agricultural export volume and predict how it will develop in future periods. Thus, managers of this sector have preliminary information about whether there will be a foreign trade deficit.

This study makes comparisons using recently popular time series models such as SARIMA and artificial intelligence models such as ANN and XGBoost. The same train set and test set are selected for all models to compare model results more accurately. As a result of this study, the XGBoost model gave better results in almost all performance metrics except MAPE. Note that MSE is selected as the objective function for all models.

When the previous studies are examined, these algorithms have yet to be used before for Turkey's fresh apricot export volume data in the literature, and there is no such comprehensive study. In this context, it is accepted that this study is original for containing new information that has not been previously explored.

This thesis, which we conducted using Turkey's data, is quite comprehensive and pioneering. There are many studies about predictions in the literature. However, artificial intelligence algorithms have become famous for using prediction day by day. In this thesis, a satisfactory study using artificial intelligence algorithms is presented. By taking this study as a reference, the study can be expanded, and a predictive analysis can be made about the state of the export volume of fresh apricots by using the XGBoost algorithm. Also, more detailed hyper-parameter optimization can be handled by using metaheuristics.