Forecast of Turkey's Carbon Emissions Within the Framework of the European Union's Green Deal

Mustafa Terzioglu
Akdeniz Universitesi
Mehmet KAYAKUŞ (mehmetkayakus@akdeniz.edu.tr)
Akdeniz University: Akdeniz Universitesi https://orcid.org/0000-0003-0394-5862
Dilsad ERDOGAN
Akdeniz Universitesi

Research Article

Keywords: Carbon dioxide emission, environment, Turkey, machine learning

Posted Date: April 6th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2580959/v1

Abstract

The most important of such efforts is the Paris Climate Agreement signed in 2015 and the t's Green Deal, implemented by the European Union (EU) within the framework of this Agreement. The targets stated in Grean Deal include measures affecting not only the EU countries but also third countries with which the EU has foreign trade links. For this purpose, ın this study, the carbon emission of Turkey, which has serious commercial relations with the EU, was tried to be estimated using machine learning techniques and an estimate was made for the year 2030 on the basis of the results obtained. These results were evaluated in line with the targets of the Green Deal. The R^2 value of Support Vector Regression (SVR), which is applied in the model as one of the machine learning techniques, was found to be 98.4% and it was found to have the highest predictive power. This technicque is followed by Multiple Linear Regression (MLR) with a 97.6% success rate and Artificial Neural Network (ANN) with 95.8% success rate, respectively. According to the estimates made with the most successful model, SVR, Turkey's carbon emissions are expected to be 723.97 million tons (mt) of CO_2 in 2030, the target year set by the EU. This level is 42% more compared to the the target that needs be achieved given the data existing in 2019. In terms of the results obtained from the study, it is thought that the study could be an exemplary model for other countries that have commercial ties with the EU.

1. Introduction

The increasing world population given its limited resources, as well as the increase in the diversity and in the level of per capita consumption, are rapidly changing the natural structure of the planet. It is the atmosphere that is most affected by this change.

The negative affect caused by greenhouse gas emissions is shown as the most important cause of global climate change. Industrialization and urbanization processes increase the volume of greenhouse gas emissions. In addition, fossil fuels, waste, insufficient use of renewable energy and uncontrolled consumption of natural resources are other factors that trigger the increase in emissions. Carbon dioxide is most important driving factor of greenhouse gas, which is composed of compounds with heat retention in the atmosphere, and causing global warming, The activities carried out by individuals, countries or organizations result, at the end, in carbon emissions into the nature (Plassmann et al., 2010).

The increasing negative effects of climate change have caused the international organizations and the countries to take important steps in this regard in recent years. The first awakening in this sense started with the Stockholm conference in 1972. These efforts were later reinforced by the Brutland Commission Report, the Kyoto Protocol and the Paris Climate Agreement signed in 2015. The Paris Climate Agreement is historically significant as it is the world's first comprehensive climate agreement. The agreement, signed by 175 countries responsible for at least 55% of global greenhouse gas emissions, aims to raise socioeconomic awareness on climate change. The increasing environmental awareness with the Paris Agreement has caused countries to adopt roadmaps for climate resilience. Having taken action for this purpose, the European Union determined a road map under the title of 'Green Deal' in 2019 to ensure

cooperation in mutual trade and economic goals and to fulfill the requirements of the Paris Agreement in sustainable economic transformation. The aim of this document is to reduce greenhouse gas emissions by 50% in 2030 and to achieve zero carbon emissions by 2050.

In this framework, the EU reshaped its entire economic policy and determined the actions and the sanctions it will impose on other countries, with which it has foreign trade relations, along with its member countries. In this context, in order to prevent carbon emissions and to ensure implementation by commercial partners, it designed the Carbon Border Adjustment Mechanism and planned to introduce additional customs duties at the consumption level (Uçak & Villi, 2021). The Green Deal, which serves as a guide for the transition to the circular economy, binds, with its zero carbon target, not only the member countries but also the third countries with which EU has trade relations. With this document, the EU published the sanctions to be applied for noncooperation and clearly stated the rules with the Carbon Border Adjustment Mechanism. Carbon Border Adjustment Mechanism is an application that establishes foreign connection in parallel with the trading system established for reducing greenhouse gas emissions in manufacturing industry sectors such as iron and steel, aluminum and fertilizer. For this purpose, the European Union Green Reconciliation Action Plan has been published by Turkey, taking into account the 2023 development goals. With the plan published in 2021, Turkey determined its own roadmap in green transformation.

The purpose of this study is to predict the carbon emission of Turkey by 2030 with a model already designed in parallel to some of the existing research in the literature for the purpose of evaluating the distance of Turkey to the target set in the EU Green Deal as well as to present a model proposal for the countries that have commercial relations with the EU. This makes the study important for decision makers and as well as policy practitioners. For this purpose, the predictive power of the model designed in the study was tested by using three of the machine learning methods: artificial neural networks, support vector regression and multiple linear regression. The study differs from other studies on carbon emission estimation as it evaluates these models within the framework of the EU's Green Deal.

2. Literature Review

In the research made on the literature related to the subject, forward-looking estimation methods used in the calculations for carbon emission, greenhouse gas emission, carbon footprint and ecological footprint, which are seen the causes of global warming, were examined. Making accurate predictions and assessments on climate change will be an important step in finding effective solutions to the problem and in implementing the necessary measures.

In his study, Baareh (2013) examined the effect of artificial neural networks model on carbon emission estimation. In this study, where four inputs were used (the consumtion of global oil, natural gas, coal as the primary energy sources) 1982–2000 period was chosen for training set and 2003–2010 period for testing set. Manhattan distance, Euclidean distance and average size of relative error were compared with predicted and actual values, and high performance of artificial neural network models was observed. The

study emphisized the fact that accurate predictions about climate change could be a useful tool in solving future problems (Baareh, 2013).

Radojevic et al. (2013) made an estimation of greenhouse gas emissions for Serbia using the artificial neural network method. In the study conducted to guide decision makers in ensuring sustainable development, data from 1999–2001 period were used for training and data from 2002–2007 for test purposes. In the relevant years, the greenhouse gas emissions were chosen as output variable and Gross Domestic Product (GDP), share of renewable energy sources, gross energy consumption and energy density were taken as input parameters. The findings were evaluated with R2, and a strong correlation was found between the estimat and the actual results. In the research, the strong prediction feature of the artificial neural network method was confirmed, It was emphasized that this could be a guiding method for future-oriented policies (Radojević et al., 2013).

Abdullah and Pauzi (2015) examined the methods used in carbon emission estimation. The aim of the study was to review the literature on methods of estimation used for this purpose. In the study covering the 2003–2013 period, models related to artificial neural networks, gray model, computer simulation, intergovernmental climate change panel (IPCC) modeling, optimal growth model and fuel analysis were examined, and the most powerful and reliable results were determined. At this study, 40 different factors were examined and it was determined that the most frequently used variables were energy consumption, GDP, fuel use, population, vehicle use, cement production, agricultural growth, cultivated area size and workforce variables and the most preferred methods was found as artificial intelligence methods. This research was described as a guide for possible future studies (Abdullah & Pauzi, 2015).

In their study, Pabuççu and Bayramoğlu (2016) made an estimation of CO2 emissions for Turkey with the artificial neural network method. In the study,the estimations for greenhouse gas emission for EU-28 countries were compared with the that Turkey as an candidate country. For this purpose, the population, GDP, energy production, energy consumption, energy use for transportation variables of the EU-28 countries and Turkey were used as inputs in the five-year models for the 1990–2030 period, and carbon equivalent greenhouse gas emissions were estimated for the years 2020, 2025 and 2030. The long range of the data set and the estensive coverage of the countries increased the reliability of the study, and the estimates were evaluated with least squares R2 and MSE. According to the highly reliable results, Turkey's carbon emissions were estimated to be 740.33 million tons (mt), 1039.32 mt and 1244.13 mt for 2020, 2025 and 2030, respectively. In addition, the study emphasized that the findings obtained in the research were well above the value Turkey committed for the year 2030 in the Paris climate agreement (Pabuççu & Bayramoğlu, 2016).

Garip and Oktay (2018) serached for an robust estimation method in calculating future carbon emissions. The data set of the study, in which random forest and support vector methods, as the machine learning methods, were compared, data set was applied for 1965–2014 period.. 1965–2003 period was taken as the basis for training and 2004–2014 period for testing purposes. In this study, the variables of oil, natural gas, coal, hydroelectricity, renewable energy and population, which are thought to affect the

CO2 emission, were determined as the inputs of the model. The obtained findings were evaluated with mean absolute error (MAE) and mean absolute percent error (MAPE), and it was observed that the support vector machine method achieved better results (Garip & Oktay, 2018).

Appiah et al. (2018) carried out the carbon emission estimation for four developing countries with artificial neural networks method. In the research conducted for China, India, Brazil and South Africa, seven variables were used as input: GDP, crop production index, animal production index, fossil fuel consumption, renewable energy consumption, import and export amounts. The data set of the study was created for the period 1971–2013, and the estimation performance was tested with the mean square error (MSE), and a high value, 0.0003345, was obtained. As a result, the predictive efficiency of the artificial neural network was proved (Appiah et al., 2018).

Acheampong et al. (2019) made a carbon emission estimation for Australia, Brazil, China, India and America using artificial neural networks. In the study, The quarterly data, for the period 1980–2015, were used for the variables of population, economic growth, energy consumption, R&D, financial development, foreign direct investments, foreign trade openness, industrialization and urbanization, which are thought to be important factors affecting carbon emission. In the study, it was determined that the estimations made for each country reached very high R-square values, hence the artificial neural network method could be an effective method with low error in the calculation of carbon emissions of these countries. In addition, the study revealed that the models developed and the results achieved can guide international organizations and decision makers in policies to be followed against climate change (Acheampong & Boateng, 2019).

In his study, Shabri (2019) searched for the model with the best forecasting performance in short-term carbon emission estimation for Malaysia. In the study where Group Method of Data Handling algorithm (GMDH), artificial neural network method and gray model were compared, the models were created to predict one year ahead between 2000–2016. The performance of the models were evaluated with least squares and least absolute shrinkage and selection operator (LASSO) methods. According to the results obtained, the LASSO-GDMH model showed the best performance in the short-term annual carbon emission analysis for Malaysia, and it was stated that the artificial neural network estimation method could be effective in longer-term analyzes (Shabri, 2022).

Çeşmeli and Pençe (2020) made a greenhouse gas emission estimation for Turkey using machine learning methods. In their study, the data set covering the years 1967–2017 was taken as time series and tested. In the research, using Poisson Regression, linear regression (LR), artificial neural networks (ANN), ANFIS and LSTM algorithms, greenhouse gas emissions were estimated for the period 2018–2031. 10-fold cross validation was applied to the results of the research and the results were evaluated with RMSE, MAPE and R2 methods. According to the findings, the highest predictive value in the mentioned period was obtained with the long-short-term memory (LSTM) algorithm. It has been stated that the estimated emission values are at a high level, and recommendations were made regarding the necessary measures to be taken (Çeşmeli & Pençe, 2020).

In his study, Özhan (2020) estimated the CO2 emissions in Turkey with time series using artificial neural networks and exponential smoothing method. In the study, the data set for the years 1960–2014, which included the greenhouse gas emission (CO2 equivalent) values of Turkey, was used. This period is divided into two: 1960–2004 as the period for training set of the data and 2005–2014 as the period for test set. Holt linear trend method and artificial neural network method were applied to both sections and the results were evaluated with RMSE and MAPE. It was observed that the model obtained from artificial neural networks had given more successful results than the Holt linear trend method, one of the exponential smoothing methods. According to the estimated values until 2021, it was underlined that carbon emissions were in a fluctuating along with a tendency to increase (Özhan, 2020).

In their study, Roumani and Modifi (2021) realized the ecological footprint estimation using machine learning methods. Different macro variables were used and a data set was created for G-20 countries in relation to 1999–2018 period. In the research, ecological footprint and its share in total and per individual biocapacity were taken as dependent variables while population of countries, birth rate, agricultural production, GDP, gross fixed capital formation, renewable energy consumption, total energy consumption, rural population, carbon emissions, renewable energy consumption. consumption, rural population, particulate matter pollution, degrees of freedom of personal and political rights and degrees of civil freedom were included in the model as independent variables. At the same time, in the study in which the improved regression and artificial neural network methods were compared, the findings were evaluated with R-square and root mean square deviation (RMSE), and it was observed that the artificial neural network method had given better results. In addition, it was emphasized that machine learning methods could give realistic results in projections estimating ecological footprints in the future (Quenard & Roumanie, 2021).

Jena et al. (2021) carried out their work on carbon emission estimation for for 17 countries, which play a key role in the world economy and have the highest emissions, using artificial neural networks method. In the study, where data set for 2017–2019 period was used, GDP, rural population ratio and foreign trade openness rates were taken as variables affecting carbon emissions. A prediction accuracy of 96% was determined for the results obtained. It was observed that the predictions made with the artificial neural network method were more effective than the predictions that were made earlier with the linear statistical models. The results showed that the countries with high emissions such as China, India, Iran, Indonesia, Saudi Arabia would reach higher values in the near future and the countries with low emission levels such as Mexico, South Africa, Turkey and South Korea would follow an increasing trend, while the emiisions in countries such as America, Japan, England, France, Italy, Australia and Canada would decrease. In addition, the study highlighted that that such forecasts could guide the countries in the transition process to the green economy (Jena et al., 2021).

Akyol and Uçar (2021) made a carbon footprint estimate for Turkey by using time series data mining methods in their study. This study aimed to predict greenhouse gas emissions for 2030 and to predict their effects on the economy, by comparing the algorithms most often as methods such as linear regression (LR), multi-layer perceptron (MLP), limited minimum optimizationand the support vector

machine for regression (SMOreg). 1990–2017 values variables that are thought to affect the carbon footprint, that is population, GDP, energy production and energy consumption were used as inputs for 2018–2030 estimations. The findings were evaluated with MSE and MAPE statistics, each estimate was compared with the actual values between 2009 and 2017, and it was founf that the closest and most reliable estimation algorithm was SMOreg. According to the results, the greenhouse gas emission amount of Turkey in 2030 was determined as 728,301 metric tons of CO2. The comparison was made with the values targeted in the climate protocols, and recommendations were made with regard to the passage to renewable energy (Akyol & Uçar, 2021).

Qader et al. (2022) carried out their study for Bahrain, where they estimated CO2 emissions with different methods. In the research using data from 1933–2018, nonlinear autoregressive, Gaussian process regression, Holt estimation method and artificial neural networks were applied. The performance evaluations of the methods were made with the root mean square error (RMSE) and the artificial neural network model had the lowest level of errors. The findings showed that the artificial neural network model was the most effective method among others in estimating carbon emissions (Qader et al., 2022).

Yaglikara (2022) examined the effects of economic, political and social globalization on the ecological footprint of five member countries of the Association of Southeast Asian Nations. The study, where panel cointegration, expanded mean group (AMG) estimator and Dumitrescu-Hurlin panel causality tests were used, revealed four independent variables that are tought to be affecting the ecological footprint. According to the findings obtained in the study, in which energy consumption per capita, economic globalization index, political globalization index and social globalization index were taken as inputs, it was determined that energy consumption increased the ecological footprint; and a one-way causality was found to be existing between ecological footprint and political and social globalization. In addition, a two-way causality was found between energy consumption and political & social globalization and a one-way causality between energy consumption and economic globalization (Yağlıkara)

3. Material And Method

In the study, the estimation of the carbon footprint of Turkey in 2030, a developing country, was carried out using machine learning methods. For this purpose, annual data in 1990–2020 period were used. Artificial neural networks, support vector regression and multiple linear regression methods were utilized in the study.

3.1. Data Set

In the study, annual data from 1990 and 2020 period were used. The data were obtained from the open data sharing platform of the central data distribution system (Biruni) of the Turkish Statistical Institute (TUIK). The model was designed by examining previous studies in the literature and adding variables that were thought to be related to the carbon emissions. The output variable of the model is the estimated tons of carbon emissions. The input variables of the model consist of demographic, economic, energy,

agricultural and animal variables. The first input variable of the model is the population which was also used in the studies carried ou by Garip and Oktay (2018), Acheampong et al. (2019), Roumani& Modifi (2021) and Akyol&Uçar (2021).

The second variable is Gross Domestic Product (GDP), which shows the economic volume of the country. This variable was also used Radojevic et al. (2013), Pabuççu & Bayramoğlu (2016), Appiah et al. (2018), Acheampong et al. (2019), Roumani& Modifi (2021), and Akyol&Uçar (2021), as the input variable. Another economic variable used as imput in the model is the industrial production index. Although filter chimney systems with carbon capturing capability are in operation, fossil-based energy consumption in the industry is common in developing countries such as Turkey. Since the development potential of these countries also focuses on industrial production, this variable was added to the model during the design phase. Acheampong et al. (2019) added a similar variable, the industrialization rate, to their model.

Another variable group is related to energy sources and consumption. The ratio of renewable energy in total energy production is included in the model in a similar fashion as in the studies made by Radojevic (2013), Garip & Oktay (2018), Appiah et al. (2018) and Roumani & Modifi (2021), In addition, the variables related to fossil fuel consumption used by Bareh (2013) and Garip & Oktay (2018) are taken into account in this study as ratios within the total energy shares. As energy consumption, per capita electricity consumption by years is included in the model as it was done in the study made by Yaglikara (2022. In their literature study on carbon emissions, Abdullah & Pauzi (2015) stated that agricultural indicators were frequently used variables and that they increased increase the predictive power of the model. Based on this study, the area of agricultural lands and the number of cattle were added as input variables. As is known, forests are one of the biggest carbon absorbers. At the same time, destruction of forests for the purpose of creating agricultural land reduces carbon absorption. In order to reflect this situation in the model, the surface area of forest land was chosen as an input variable. Finally, the number of vehicles with internal combustion engines used over the years has been added to the model. Although the engine and exhaust systems of the vehciles with internal combustion engines (especially diesel vehicles) using fossil fuels are produced according to national/international carbon emission standards, the increase in the levels of the carbon emissions continue in parallel to the with the increase in the number of vehicles in traffic. Table 1 presents all the variables of the model together.

Table 1
Variables of the Model

Input Variables	Output Variable
Population	Carbon emissions (tonnes)
Gross domestic product ($)	
Electricity consumption per person (kWh)	
Share of Renewable Energy in total energy (%)	
Share of coal in total energy (%)	
Share of natural gas in total energy (%)	
Share of liquid fuels (Gasoline, Diesel etc.) in total energy (%)	
Number of internal combustion engine vehicles	
Industrial production index (2015 = 100)	
Number of cattle	
Agricultural area (hectares)	
Forest area (hectares)	

3.2. Artificial Neural Networks

Artificial neural networks are mathematical models inspired by the functions of biological neural networks. Thanks to the learning feature of ANNs, they can provide solutions to problems that are too complex for traditional techniques. Thanks to irs learning ability, by using known examples, it can make generalizations about situations that have not been encountered. Artificial neural networks, which are used for purposes such as estimation, classification and shape completion using numerical information, can be applied in many areas from financial issues to engineering and medical science, from production applications to fault detection and analysis (Ağyar, 2015).

The artificial neural network algorithm is arranged in multiple layers in order to establish the correlation between inputs and outputs. The first layer is the input layer. The output layer is the last layer. The other layers in between are called the middle layer(s) or hidden layer(s). A network can have more than one hidden layer. The input layer is the terminals through which individual data is delivered to the network. The number of neurons in this layer is equal to the number of data inputs, and each input neuron receives one data. The data here passes to the next layer, the hidden layer, without being processed. Weights are components that show the importance and effect, for the artificial nerve cell, of the information coming to that artificial nerve cell. Each entry has its own weight. The large or small value of a weight does not indicate that the input is important or unimportant for the neural network. Weights can take on variable or

fixed values. The hidden layer is the one that performs the basic function of the network. Some implementations may have more than one hidden layer in the network. The number of hidden layers and the number of neurons in the hidden layer depend on the problem. This layer receives the data from the input layer and processes the it with a function suitable for the problem by multiplying it with weights and transmits it to the next layer. The addition function calculates the net input received by a cell. Here, the value of each incoming input is multiplied by its own weight and added. Thus, the net input to the network is found.

$$\mathrm{NET} = \sum_{\mathrm{i}}^{\mathrm{n}} \mathrm{x_i w_i} \; 1$$

Here x shows input, w weight value and n show the total number of inputs to the neuron.

Another important component of the artificial neuron is the activation function. This function outputs the information from the aggregation function. The activation function is also called the threshold function. Usually linear function, step function, sigmoid function and hyperbolic tangent function are used as an activation functions (Aygören et al., 2012). The most frequently used activation function is the sigmoid activation function given in Eq. 2.

The output layer is the most extreme layer of the network. It processes the data it receives from the hidden layer with the function used by the network and outputs it. The number of neurons in the output layer is equal to the number of outputs of each data that are delivered to the network. The values obtained from this layer are the results obtained from the ANN for the existing problem (Var & Türkay, 2014).

$$\mathrm{f\,(NET)} = \frac{1}{1+\mathrm{e}^{-\mathrm{NET}}} \; 2$$

As an architectural structure, ANN is divided into two parts: one with feedforward and the other with feedback. In a feedforward architecture, there is a one-way flow of information. Cells are in the form of regular layers from the entrance to the exit. The information coming to the network passes to the input layer, then it is processed in the hidden layers and the output layer, respectively, and reaches the output/exit of the network. In a feedback neural network, the output of a cell is not given as an input not only to the next layer. But it can be given as an input to any cell in the previous layer or in its own layer.

In the artificial neural networks, information is held in the weights of the connections of the neurons in the network. Therefore, it is important how the weights are determined. Since the information is stored in the entire network, the weight value of a node does not mean anything by itself. The weights in the entire network should take optimal values. The process to reach these weights is called "training the network". Accordingly, in order for a network to be trainable, the weight values must be changeable in a dynamic way under a certain rule.

3.3. Support Network Regression

Support Vector Machines (SVM) is a supervised learning method developed by Vapnik and one of the supervised learning methods used for classification and regression (Vapnik, 1999). Compared to other traditional learning methods, this method has much better performance and ability to solve nonlinear problems. Adaptation of SVM for regression, which is widely used for classification problems, was proposed by Smola et al. (Smola & Schölkopf, 2004). The main purpose of support vector regression (SVR) is to draw a line to separate points placed on a plane. This line is intended to be at the maximum distance for the points of both classes. It is suitable for complex but small to medium datasets.

SVR is divided into two as linear and nonlinear regression methods. In the case of linear separability, these bivalent data can be separated directly by an extreme plane. This extreme plane is called the separating hyperplane. The purpose of SVR is to ensure that this extreme plane is equidistant to the sample groups in two separate classes (Yakut et al., 2014). Linear SVR is shown in Fig. 1.

In cases where the data cannot be separated linearly, nonlinear classifiers can be used instead of linear classifiers. In this context, for nonlinear feature space, it may be possible to obtain linear classifiers in a new space, by transforming the observation vector xi ϵ Rn into a vector z in a higher-order space. Figure 2 shows the nonlinear SVR.

The nonlinear SVR tries to find a regression function expressed as f(x) = wTϕ(x) + b in hyperspace. This function is obtained using the "ϵ-insensitive" loss function. Nonlinear SVR can be obtained by solving the following Quadratic Programming Problem (KPP):

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\left(e^T\xi + e^T\xi^*\right)$$

$$(\phi(A)w + eb) - Y \leq e\epsilon + \xi, \xi \geq 0e$$
$$Y\text{-}\ (\phi(A)w + eb) \leq e\epsilon + \xi^*, \xi^* \geq 0e$$

3

Here, C is a predetermined parameter and is an adjustment parameter that balances the adaptation of errors and the flatness of the regression function. ξ and ξ* are dummy variables that indicate whether the samples enter the ϵ-tube, and e is the unit vector.

We obtain the regression function using the Lagrangian multipliers α and α∗.

$$f(x) = \sum_{i=1}^{n} \left(\alpha^* - \alpha\right) K(x_i, x) + b \quad 4$$

Here, K(xi, x) = (ϕ(xi) · ϕ(x)) represents the kernel function and gives the dot product in hyperspace. α and α∗ are Lagrangian multipliers and they satisfy $\alpha i \alpha i*$ = 0, i =1,2, …, n. The f(x) function is determined only by samples (support vectors) with Lagrangian multipliers $\alpha i \neq 0$ or $\alpha i* \neq 0$. It also shows the inputs and outputs of the A = (x1, x2, ... ,xn) and Y = (y1, y2, ... ,yn) training set, respectively (İnce & İmamoğlu, 2016).

3.4. Multiple Linear Regression

Linear regression analysis is used to estimate the value of one variable relative to the value of another variable. The variable you want to predict is called the dependent variable. The variable you use to predict the value of the other variable is called the independent variable. There are two types of linear regression analysis: simple regression and multiple regression. While a single explanatory variable is used in simple regression; multiple regression uses a large number of explanatory variables.

Multiple linear regression examines the linear relationship between two or more independent variables and one dependent variable. In other words, it is an analysis to reveal the relationship between a dependent variable and a series of independent variables that are related to it.

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{X}_{\mathrm{i1}} + \beta_2 \mathrm{X}_{\mathrm{i2}} + \beta_3 \mathrm{X}_{\mathrm{i3}} + \ldots . + \beta_{\mathrm{k}} \mathrm{X}_{\mathrm{ik}} + \epsilon_{\mathrm{ij}} \quad 5$$

It is defined as i = 1,2,3,…,n and j = 1,2,3,…,n.

Here, Xij, j. represents the value of independent variable at i. level, Bj, j. represents the regression coefficient, eij represnts the error term, and k represents the number of independent variables. The coefficient β expresses the amount of change that will occur in Y in terms of its own unit, in response to 1 unit change in X in its own unit.

4. Result And Discussion

In the study, 30 data SETS taken annually between the years 1990 and 2020 were used. A data set consisting of 12 independent variables was prepared to estimate Turkey's annual carbon footprint. Three different machine learning methods were used in the study, namely: artificial neural networks, support vector regression and multiple linear regression.

The basic working structure of the study is shown in Fig. 3.

In machine learning, training of data can be made more efficient by pre-processing steps and thus the data is transformed into a better form. By applying the normalization process to the raw data, a suitable data set for training is prepared. The speed and success rate of the system is directly proportional to the level of preliminary data processing.

The data set was checked in the preprocessing stage of the data set, which is the first stage of the study. Inconsistent and erroneous data in the data set is called noise. To clean the noise in the data, records containing missing values can be discarded, a fixed value, can be assigned instead of missing values. This fixed value can be found as a result of a suitable estimation (regression) based on the existing data.

In our study, data cleaning was performed to avoid noisy data.

The second stage of the study is the feature selection stage. At this stage, the effect of the independent variables determining the dependent variable is found, and the independent variables that have much less effect are removed from the data set.

Selection of the attributes is the process of selecting and finding the most useful attributes in the data set. This process greatly affects the performance of the machine learning model. Feature selection; this is the selection of subsets of features without sacrificing the accuracy. It aims to reduce the data level by deleting irrelevant and unnecessary data. Its purpose is to find accuracy by removing unnecessary features. Feature selection is a requirement for any data mining product. A dataset can contain many unnecessary features. Among the existing features, it is necessary to determine the features that affect the result the most, that is, the features that are related to the result. Feature selection works by calculating a score for each attribute and selecting the attributes with the best scores. Most feature selection methods can be divided into three main categories: filter methods, wrapper methods, and embedded methods. Filter method takes into account the relationship between the feature and the target variable in order to calculate the importance of the features. As a result of the operations, we filter our data set and create a subset by selecting the relevant features.

The most commonly used filter methods are Pearson Correlation and Chi-Square Method. In the Pearson Correlation method, the most widely used correlation measure is the degree of relationship between linearly related variables.

Another phase of the study is the normalization of the data. Normalization is one of the pre-processes used to make the data set ready for artificial intelligence applications. In the study, the data were normalized between the range of -1 and + 1 using the decimal scaling method. In this method, numbers are divided by 10 or by powers of 10 to make numbers less than 1.

As a result,, the range is set between − 1 and + 1.

$$\mathrm{A}' = \frac{\mathrm{A}}{10^{\mathrm{j}}}\ 6$$

Here, the real value of A can be defined as the smallest value that makes normalized A^{I} data, j, and A value less than one. First of all, the j values that make the maximum value less than 1 in the columns are found, and new values are obtained by division of all the elements of the relevant column by the j value.

The data used in the models are divided into two clusters for training and testing purposes. The creation of a model over the training set is based on the examinaiton of the performance (learning level) of this model over the test set. In this way, new unlabeled samples get predictable labels with the help of the model.

Although there is no definite rule about the separation rate of the data, it is usually decided by trial and error method. As a result of the tests performed, the data are divided into two with the rates that provide

the highest success rate. As a result of the tests carried out in this study, it was decided to use 70% of the data for training and 30% for testing purposes.

Take from top, linear sampling and draw randomly are some of the data selection methods that can be used. In the study, linear sampling method was preferred in data selection in order to compare the results of the two models.

A feedback model has been developed in the ANN method. The number of hidden layers and neurons in the model was decided by trial and error method. It was observed that the structure with two hidden layers and two neurons in each layer produced more successful results. The structure of the model that was developed is shown in Fig. 4.

1000 iterations were performed to get the best result in the model. The error curve of the ANN model based on iterations is shown in Fig. 5.

R2 (Coefficient of determination), RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error) techniques were used to analyze and interpret the study results.

Nonlinear SVR was used for support vector regression, which is another machine learning method used in the study. Polynomial, hyper tangent, radial basis function (RBF) was tested for the kernel function and RBF was preferred in the study since it was determined as the most successful model. The other parameters of the model were chosen as follows: the overlapping penalty:100 and RBF sigma value: 0.1.

In order to measure the effects of the variables on the system in the multiple linear regression, a significance value was determined first. The variable with the current highest p-value (probability value) was determined and if P > SL, the variable was removed from the system. The model was rebuilt and then this step was repeated.

When P < SL for all variables, elimination is terminated. Since there were no independent values below 0.05 for p values in the designed model, the model was found to be significant.

R2 is a measure of how well the data fit a linear curve. R2 is the coefficient of determination, also known as the coefficient of determination. It measures the percentage level, that is, how much the independent variable x explains the dependent variable y with the regression model. R2 takes values between 0 and 1 (0 < R2 < 1). If R2 equals 1, this indicates that the experimental data provides a perfect linear curve: higher R2, better the fit of the regression model.

$$\mathrm{R}^2 = 1 - \frac{\mathrm{UnexplainedVariation}}{\mathrm{TotalVariation}} \quad 7$$

RMSE is a quadratic metric that measures the magnitude of error of a machine learning model, which is often used to find the distance between the predictor's predicted values and the true values. RMSE is the standard deviation of the estimation errors (residues). That is, residuals are a measure of how far the regression line is from the data points; RMSE is a measure of how widespread these residues are. In other

words, it tells how dense the data is around the line that best fits the data. RMSE formula is shown in Eq. 8: Here n represents the number of data and e is the error value.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{j=1}^{n} e_j^2}{n}} \quad 8$$

MAPE statistics eliminates the disadvantages that may arise when comparing models with different unit values. Among the listed criteria, "Mean Absolute Percent Error" (MAPE) is considered to be superior to other criteria, since it expresses the prediction errors as a percentage, so that it has a meaning on its own. Models with MAPE < 10% are classified as “very good”, modelswith 10% <MAPE < 20% as “good”, models with 20% <MAPE < 50% as “acceptable” and models with 50%< MAPE as “inaccurate and classified faulty”. The MAPE formula is shown in Eq. 7:

$$\mathrm{MAPE} = \frac{100}{n} \sum_{j}^{n} \frac{|e_j|}{|A_j|} \quad 9$$

The success and error values of the models according to the ANN, SVR and MLR methods are provided in Table 2.

Table 2
Comparison the Success and the Failure of the Models

	ANN	SVR	MLR
R^2	0,958	0,984	0,976
RMSE	0,019	0,013	0,016
MAPE	0,063	0,024	0,051

Figure 6 shows the graphical representation of the error and success values of the models.

An R2 value of 1, which indicates how well the data fit a linear curve, indicates that the test data provided a linear curve. As a result of the study, the R2 value was 95.8% for ANN, 98.4% for SVR, 97.6% for MLR and it was found to be very close to the ideal value. If the RMSE value, which is used to find the distance between the estimated values and the actual values, is zero, it means that the model does not make any mistakes. Therefore, it is desired that the RMSE value is close to zero. In the study, it was observed that the RMSE value was 0.019 for ANN, 0.013 for SVR and 0.016 for MLE, which was close to the ideal value. Models with a MAPE value which is below 10 percent are considered very good. In the study, it was seen that this value was 6.3% for ANN, 2.4 for SVR and 5.1% for MLR. It is accepted that the MAPE value is very good in all three models.

When the error and success values are examined, it is seen that the most successful and least erroneous models are SVR, MLR and ANN, respectively. Since the results of the study are normalized, they are expressed between values between 0 and 1. In order to adapt the results to real values and make sense of them, denormalization is required. Table 3 shows the actual values and the estimated values as a result of denormalization.

Table 3
Real Values and Estimated Values

Years	Real data	Predicted denormalized data		
		ANN	SVR	MLR
1993	240635808	234546234	228985654	224423254
1996	267662004	244524652	255885655	249365742
1999	277879815	278238452	275634552	280020365
2002	285698774	290165447	276998742	299698742
2005	336986354	351273684	337231583	358964523
2008	387867898	389246521	392112365	391742365
2011	428492478	444356743	437885423	424056485
2014	459365743	465204873	474876325	440555634
2017	528311866	521549256	539122744	524423687
2019	508078171	516743526	500805324	504889652

Looking at the results in Table 1, the mean error was 2.20% for SVR; 2.33% for ANN and 2.87% for MLR. It is seen that these values are close to ideal values and all are acceptable.

5. Conclusion

Carbon emissions, which is one of the most important causes of global climate change, is an important problem that needs to be dealt with. Especially with the acceleration of coal use after the Industrial Revolution that started in the 18th century and, with an increased the use of fossil fuel derivatives together with the increasing use of internal combustion land vehicles in the 19th century the amount of carbon dioxide in the atmosphere rapidly increased. Since the 1990s, in line with increase, many national/international organizations started to develop policies that reduced carbon emissions, as a precaution against changes in the climate. At the forefront of these efforts is the Paris Climate Agreement. With this agreement, many countries have made commitments to reduce their future carbon emissions. One of the first concrete actions taken within the framework of this agreement is EU's Green

Deal. This Agreement imposes some binding obligations not only on EU countries, but also on countries that have commercial activities with EU countries. These countries, which have commercial ties with the EU, need to reduce their carbon emissions by half by 2030. Otherwise, such countries trading with EU countries will face practices such as high carbon taxation at the border.

The aim of this study is to estimate the carbon emissions, using machine-teaching techniques, of Turkey in 2030, as a country that has a high trade volume with the EU (approximately $178 Billion as of 2021),. In addition, it is also aimed to determine the distance of Turkey to the target determined by the EU, and to offer a practical and effective model for other countries that have commercial relations with the EU.

While designing the model, the literature was reviewed, and variables used in the existing literature were examined in the literature review. Takimg into account these studies, the carbon emission of Turkey was tried to be estimated by using 12 input variables. Different variables that were thought to affect carbon emissions in case of Turkey were also included. In the study, ANN, SVR and MLR machine learning techniques, were used to measure the predictive power of the model, as they are assumed to be more effective than other classical statistical techniques. The R2 value of SVR, was found to be 98.4% and it was determined that it had the highest predictive power.

This model was followed by MLR with a 97.6% success rate and by ANN with a success rate of 95.8%, respectively. The results of this study are in confirmity with the success predictions made with the machine learning techniques in some of the existing work in the literature such as Radyojevic (2013), Abdullah & Pauzi (2015), Garip & Oktay (2018), Appiah et al. (2019), Rounmani & Modifi (2021), Jena et al. (2021), Akyol & Uçar (2021) and Oader (2022).

In this study, in which three different machine learning techniques are used, SVR, MLR and ANN, unlike the exiting studies in literature, it is found that SVR is more successful in terms of the measurement of carbon emissions.

Based on the success of this this carbon emission measurement model model, it is thought that it could be a viable model to to reach the Green deal targets for the decision makers in developing countries that commercial relations with the EU just like Turkey.

According to the estimates made with SVR over the model, the carbon footprint of Turkey is expected to be 723.97 million tons (mt) of CO2 in 2030, the target year determined by the EU. This rate is 42% more than the target rate that should be achieved according to the data existing in 2020. This estimated amount of carbon coincides with other studies made on Turkey covering different variables and periods.

Pabuççu and Bayramoğlu (2016) estimated 740.33 million tons (mt) of CO2 using ANN and Akyol & Uçar (2021) estimated 728.301 million tons (mt) of CO2 using SMOreg.

According to the results obtained, when the fossil-based energy sources used in the model are examined, it is thought that Turkey will not be able to reach both the 2030 and 2050 carbon targets if the current use continues and the renewable energy sources are not increased. It is necessary to diversify the incentives

for renewable energy sources that will accelerate Turkey's transition to a sustainable economy in a short time without compromising the country’s targets for economic growth.

In addition, in line with these goals, it is recommended that, taking into account sustainable development goals, environmental policies and economic policies be redesigned in harmony with each other.

Declarations

Ethical Approval

Not applicable

Consent to Participate

Not applicable

Consent to Publish

Not applicable

Authors Contributions

Mustafa Terzioglu, experiment, data analysis, writing-original draft; Mehmet Kayakus, conceptualization, writing-review and editing; supervision,;Dilsad Erdogan, writing-original draft

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors declare no competing interests.

Availability of data and materials

The data can be available on request.

References

1. Abdullah, L., & Pauzi, H. M. (2015). Methods in forecasting carbon dioxide emissions: a decade review. *Jurnal Teknologi*, *75*(1).
2. Acheampong, A. O., & Boateng, E. B. (2019). Modelling carbon emission intensity: Application of artificial neural network. *Journal of Cleaner Production*, *225*, 833-856.

3. Ağyar, Z. (2015). Usage areas of artificial neural networks and an application. *Engineer and Machine*, *56*(662), 22-23.
4. Akyol, M., & Uçar, E. (2021). Carbon footprint forecasting using time series data mining methods: the case of Turkey. *Environmental Science and Pollution Research*, *28*(29), 38552-38562.
5. Appiah, K., Du, J., Appah, R., & Quacoe, D. (2018). Prediction of potential carbon dioxide emissions of selected emerging economies using artificial neural network. *Journal of Environmental Science and Engineering A 7*, *14*, 321-335.
6. Aygören, H., Saritaş, H., & Morali, T. (2012). Forecasting ISE 100 Indice Using Artifıcial Neural Networks And Newton Numerical Search Models *International Journal of Alanya Business Faculty*, *4*(1), 73-88.
7. Baareh, A. K. (2013). Solving the Carbon Dioxide Emission Estimation Problem: An Artificial Neural Network Model. *Journal of Software Engineering and Applications*, *6*, 338-342.
8. Çeşmeli, M. Ş., & Pençe, İ. (2020). Forecasting of Greenhouse Gas Emissions in Turkey using Machine Learning Methods. *Academic Platform-Journal of Engineering and Science*, *8*(2), 332-348.
9. Garip, E., & Oktay, A. B. (2018). *Forecasting CO2 Emission with Machine Learning Methods* 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey
10. Jena, P. R., Managi, S., & Majhi, B. (2021). Forecasting the CO2 Emissions at the Global Level: A Multilayer Artificial Neural Network Modelling. *Energies*, *14*(19), 6336.
11. Özhan, E. (2020). Estimation Of CO2 Equivalent Greenhouse Gas Emissions In Turkey By Artificial Neural Networks And Exponential Smoothing Method. *European Journal of Science and Technology*(19), 282-289.
12. Pabuçcu, H., & Bayramoğlu, T. (2016). CO2 Emissions Forecast with Neural Networks With: The Case Of Turkey. *Gazi University Journal of Faculty of Economics and Administrative Sciences*, *18*(3), 762-778.
13. Plassmann, K., Norton, A., Attarzadeh, N., Jensen, M., Brenton, P., & Edwards-Jones, G. (2010). Methodological complexities of product carbon footprinting: a sensitivity analysis of key variables in a developing country context. *Environmental Science & Policy*, *13*(5), 393-404.
14. Qader, M. R., Khan, S., Kamal, M., Usman, M., & Haseeb, M. (2022). Forecasting carbon emissions due to electricity power generation in Bahrain. *Environmental Science and Pollution Research*, *29*(12), 17346-17357.
15. Quenard, S., & Roumanie, M. (2021). A simple method for a protective coating on stainless steel against molten aluminum alloy comprising polymer-derived ceramics, oxides and refractory ceramics. *Materials*, *14*(6), 1519.
16. Radojević, D., Pocajt, V., Popović, I., Perić-Grujić, A., & Ristić, M. (2013). Forecasting of greenhouse gas emissions in Serbia using artificial neural networks. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, *35*(8), 733-740.
17. Shabri, A. (2022). Forecasting the annual carbon dioxide emissions of Malaysia using Lasso-GMDH neural network-based. IEEE 12th Symposium on Computer Applications & Industrial Electronics

(ISCAIE), Penang Island, Malaysia.

18. Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*(3), 199-222.
19. Uçak, S., & Villi, B. (2021). Possible effects of the European green deal on the steel industry. *Journal of Empirical Economics and Social Sciences*, *3*(2), 94-113.
20. Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.
21. Var, H., & Türkay, B. E. (2014, 27 – 29 Kasım 2014). *Short Term Electric Load Forecasting Using Artificial Neural Networks* Electrical – Electronic – Computer and Biomedical Engineering Symposium, Bursa, Turkey.
22. Yağlıkara, A. Effects Of Economic, Political and Social Globalization On Ecological Footprint: The Case Of ASEAN-5 Countries. *Fiscaoeconomia*, *6*(2), 656-676.
23. Yakut, E., Elmas, B., & Yavuz, S. (2014). Predicting Stock-Exchange Index Using Methods of Neural Networks And Support Vector Machines. *Suleyman Demirel University The Journal of Faculty of Economics and Administrative Sciences*, *19*(1), 139-157.

Figures

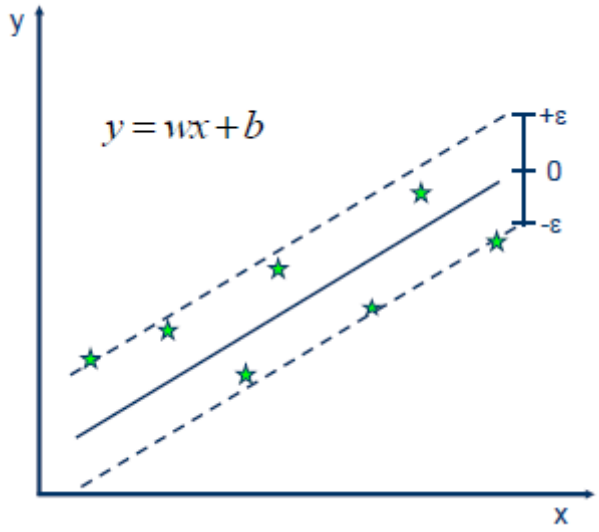


Figure 1

Linear SVR

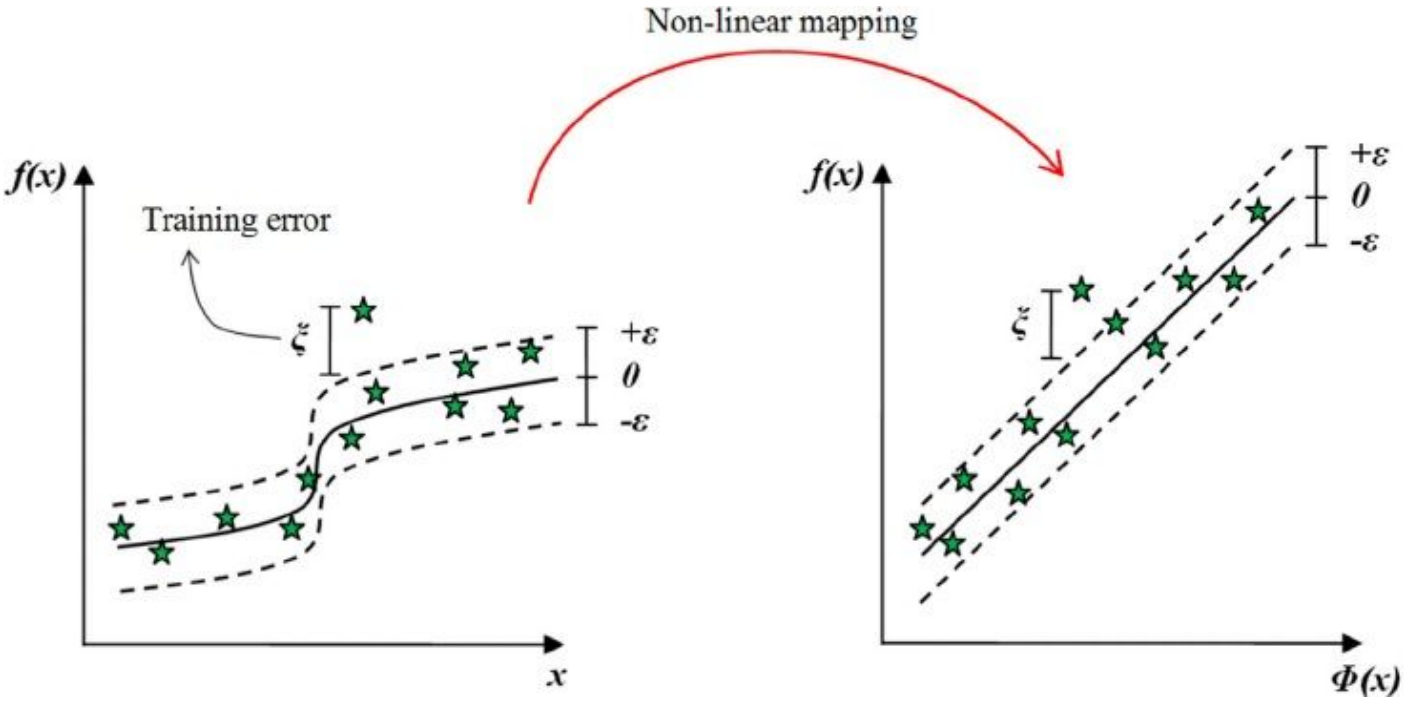


Figure 2

Non-linear SVR

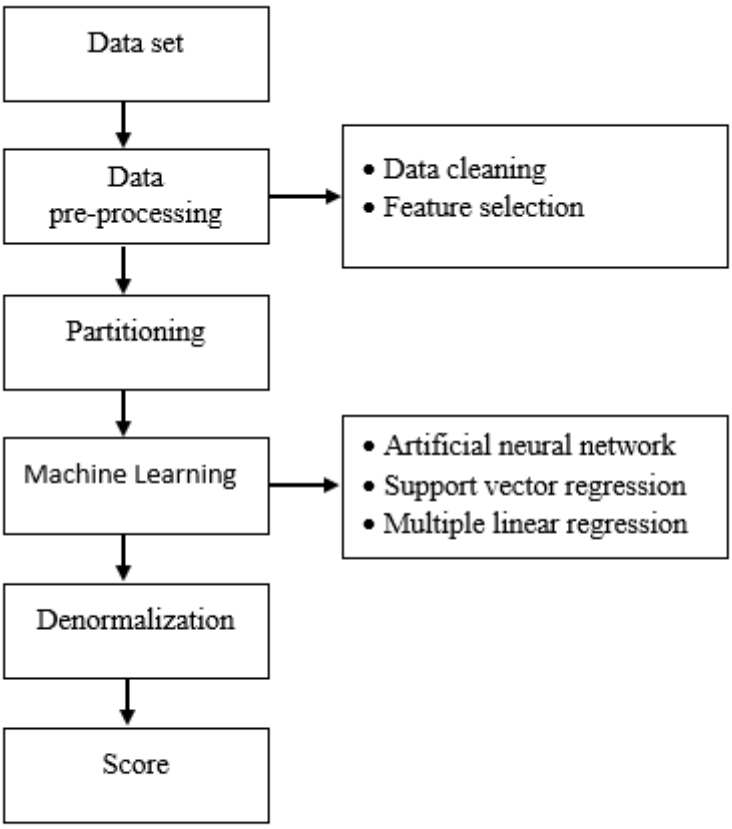


Figure 3

Working Structure of the Study

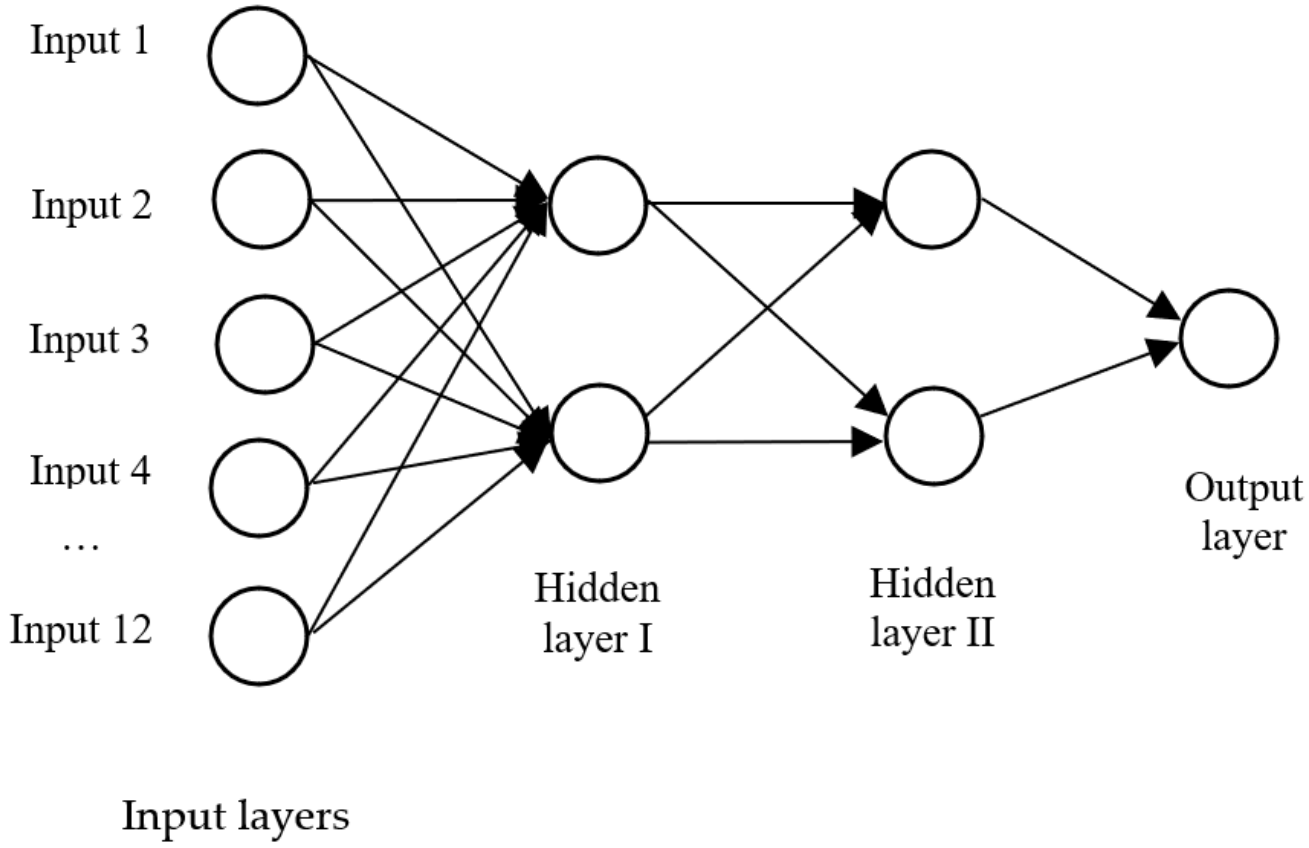


Figure 4

ANN Model that was develeoped

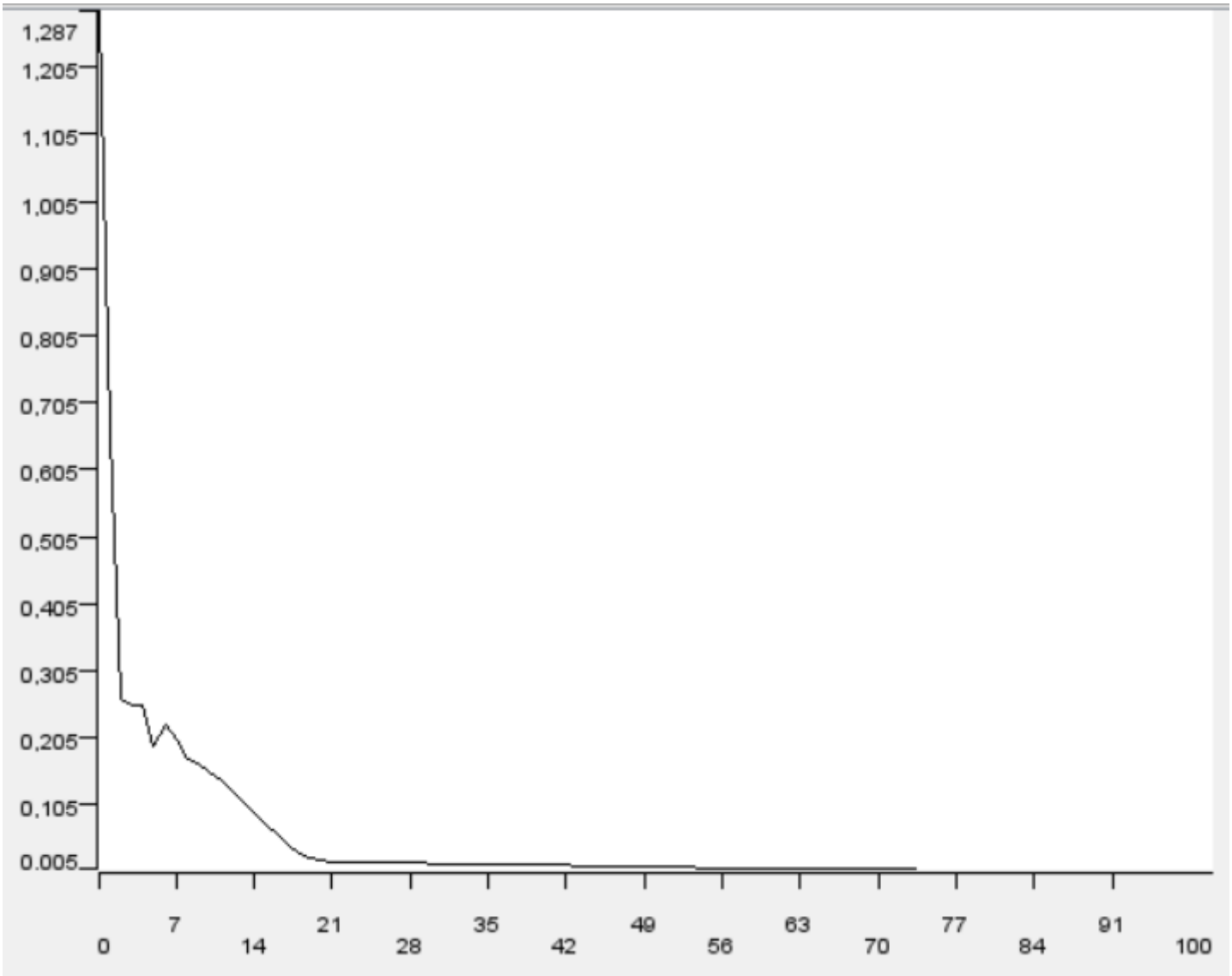
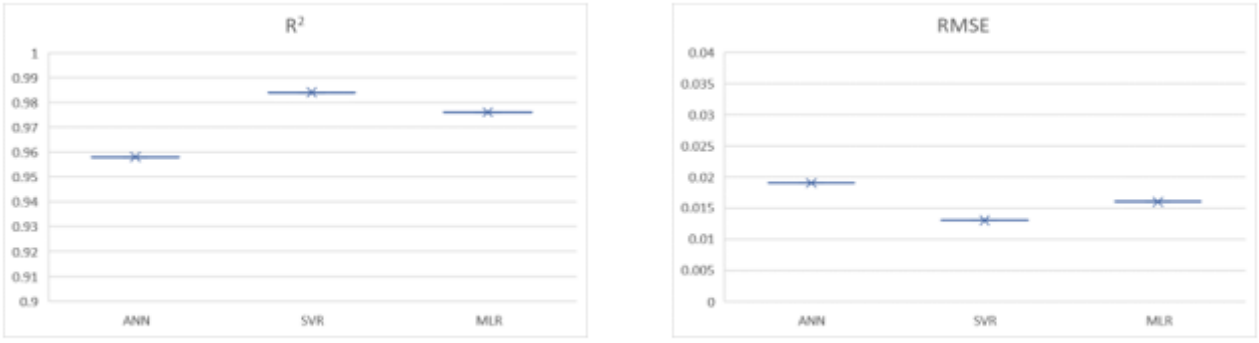


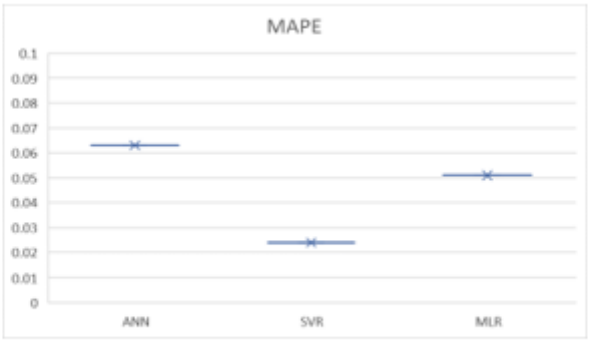
Figure 5

Error Curve at ANN Training Phase



a) R^2

b) RMSE



c) MAPE

Figure 6

Errors and Success Values of the Models