Turkish Journal of Science & Technology
18(1), 233-241, 2023

Research Paper
https://doi.org/10.55525/tjst.1248426

Predicting the Height of Individuals with Machine Learning Methods by Considering Non-Genetic Factors

Osman ALTAY[1*], Tuğba ÇELİKTEN[1], Tuba AKBAŞ[1], Hüseyin Yasin DÖNMEZ[1]
[1] Yazılım Mühendisliği, Hasan Ferdi Turgutlu Teknoloji Fakültesi, Celal Bayar Üniversitesi, Manisa, Türkiye
[*1] osman.altay@cbu.edu.tr, [1] tugba.celikten@cbu.edu.tr, [1] tubakbas1@gmail.com, [1] hyasindonmez@gmail.com

(Geliş/Received: 06/02/2023; Kabul/Accepted: 13/03/2023)

Abstract: As many parents want to know how many centimeters their child will be in the future, many people in their developmental years want to know how many centimeters their future height will be. In addition, the development of children in terms of height and weight is medically controlled from the moment they are born. As a result, height development is important for both individuals and medical professionals. In this study, it is aimed to predict the height of individuals using personal and family information and factors affecting height. In the study, the 10 most known characteristics among the factors affecting height were selected. These attributes, mother's height, father's height, economic status, jumping and weight sports status, gender, information about the child's age, history of chronic illness in the individual, the longest living region, and the individual's height were taken as input values in machine learning methods. Using these input values, the length of the individual was predicted using Linear Regression (LR) and Artificial Neural Network (ANN) from machine learning methods. In addition, three error measurement methods were used to evaluate the success of the model: mean absolute error (MAE), mean square error (MSE) and R-Square (R^2). In the R^2 evaluation metric, the method was 84.48% in LR and 81.74% in ANN.

Key words: Height of individuals, machine learning algorithms, artificial neural network, linear regression.

Genetik Olmayan Faktörler Ele Alınarak Bireylerin Boyunun Makine Öğrenmesi Yöntemleri ile Tahmini

Öz: Pek çok ebeveyn doğmuş veya doğacak olan çocuğunun gelecekte kaç santimetre boya sahip olacağını bilmek istediği gibi birçok gelişim çağındaki birey de ilerdeki boylarının kaç santimetre olacağını bilmek ister. Ayrıca tıbbi olarak çocukların gelişimleri boy ve kilo olarak ilk doğdukları andan itibaren kontrol edilmektedir. Bu yüzden boy gelişimi hem bireyler için hem de tıbbi olarak önemlidir. Bu çalışmada ise bireylerin kişisel ve aile bilgilerinden yararlanılarak ve boy uzunluğuna etki eden etmenler kullanılarak boylarının tahmin edilmesi amaçlanmıştır. Yapılan çalışmada boy uzunluğuna etki eden etmenler arasından en bilinen 10 nitelik seçilmiştir. Bu nitelikler, anne boy uzunluğu, baba boy uzunluğu, ekonomik durum, sıçrayış ve ağırlık sporları yapma durumu, cinsiyet, kaçıncı çocuk olduğu bilgisi, bireyde geçirilen kronik rahatsızlık öyküsü, en uzun yaşanılan bölge ve bireyin boyu makine öğrenmesi yöntemlerinde giriş değerleri olarak alınmıştır. Bu giriş değerleri kullanılarak bireyin boyu makine öğrenmesi yöntemlerinden Lineer Regresyon (LR) ve Yapay Sinir Ağı (YSA) kullanılarak tahmin edilmiştir. Ayrıca modelin başarısını değerlendirmek için ortalama mutlak hata (OMH), ortalama kare hata (OKH) ve R-Kare (R^2) olmak üzere üç hata ölçüm yöntemi kullanılmıştır. R^2 değerlendirme metriğinde LR yöntemi %84.48 ve YSA ise %81.74 başarım elde etmiştir.

Anahtar kelimeler: Bireylerin boyu, makine öğrenmesi algoritmaları, yapay sinir ağları, lineer regresyon.

[*] Corresponding Author: osman.altay@cbu.edu.tr. ORCID Numbers of Authors: 0000-0003-3989-2432, 0000-0001-7480-4026, 0000-0002-6431-7520, 0000-0002-5583-8375

1. Introduction

Along with the development of technology throughout history, the factors affecting the development of human beings are also being investigated. It has emerged as a result of research that the environment, heredity, and hormones have an effect on human development and health [1]. The development of an individual occurs as a result of the complex interaction of environment and heredity in two directions. The environment is the natural environment that includes the factors affecting living things and the living things themselves. Heredity is the ability or trait that is passed on from parents to their children through genetic pathways. Genes passed down through the generations in the family consist of deoxyribonucleic acid (DNA). Individuals are separate and special from each other as a result of the environment in which they live and the genes that come from their family. As a result, the development of each individual may be different from that of other individuals [1]. Hereditary traits and diseases are passed on to future generations through DNA. Hereditary diseases occur due to chromosomal or gene defect inherited from parents to the individual. Some of these diseases are down syndrome, albino, hemophilia, fish scale etc. are diseases. Some of the hereditary features can be counted as many features, such as the individual's height, eye color, blood group, skin color, facial features, the shape of our fingers, and the shape of their hair. Environmental factors, on the other hand, can be given as examples of parents' parenting style, nutrition, unrest in the family, the mother's age, the geographical region, the number of siblings, and the parents' education level. It has been shown in the literature that hormones also have an effect on development. For example, an imbalance in the endocrine glands affects the development of the organism [1].

Artificial intelligence is a broad discipline with roots in philosophy, mathematics, and computer science that aims to understand and develop systems that demonstrate the characteristics of intelligence. The development of artificial intelligence still continues today and is used in many areas. Many areas, such as health services, education, banking, agriculture, the economy, and the military, can be given as examples. Some of the artificial intelligence applications used in healthcare diagnosis and prediction include applications such as prediction of hospitalization for heart disease, cardiovascular risk estimation, diagnosis of pneumonia from a lung x-ray, and identification of benign and malignant tumors [3]. Between 2000 and 2006, many researches and studies were carried out in the field of artificial intelligence. As a result of these studies, artificial intelligence has become an indispensable part of daily life and for the first time, a vacuum cleaner called Roomba and a product that uses artificial intelligence has managed to enter our homes. After 2006, companies such as Twitter, Facebook, Amazon, and Google realized the power of artificial intelligence, and many companies gave more importance to the studies in the field. These studies have succeeded in making artificial intelligence technologies more popular and have contributed to the development of artificial intelligence by further increasing the acceleration of artificial intelligence [4]. Although artificial intelligence is a basic science field, it includes machine learning and deep learning sub-branches.

Machine learning (ML) is a sub-science of artificial intelligence that aims to create a mathematical model by processing the data given to it by machines [5]. Machine learning strategies are grouped under three main headings. These include supervised, unsupervised, and reinforced learning. Supervised learning is the most widely used algorithm among machine learning algorithms. After the data is labeled, the model is trained, and predictions are made with the data that the model does not see. Regression and classification techniques are used in supervised learning. Unsupervised learning is another branch of machine learning. There is no labeled data here. The algorithm makes inferences by taking unlabeled data as input. Clustering techniques are used in unsupervised learning. In clustering algorithms, data are grouped according to their similarity [6]. Using clustering algorithms, applications are being developed on topics such as voice and image processing, keyword searching, call center records, speech recognition, and grouping customer purchasing behavior [7]. Reinforcement learning is a machine learning algorithm that automatically evaluates the most appropriate behavior in a given context or environment to increase the efficiency of software tools and machines. This type of learning is based on reward or punishment. It is a tool used to train artificial intelligence models that can help increase automation or optimize the efficiency of complex systems such as robotics, autonomous driving, and supply chain logistics [8].

As a result, in this study, height prediction was made using the machine learning algorithms ANN and LR. The results of the study showed that the height of people, which shows their development and is followed from birth, can be predicted by machine learning methods. The general organization of the study is as follows: In the second section, "Materials and Methods," the creation of the data set, the statistical distribution of the attributes of the data set, and the machine learning methods are explained in detail. In the third section, the results obtained from machine learning methods are given and examined. Finally, the article is completed with conclusions and recommendations.

2. Materials and Methods

2.1. Data Acquisition

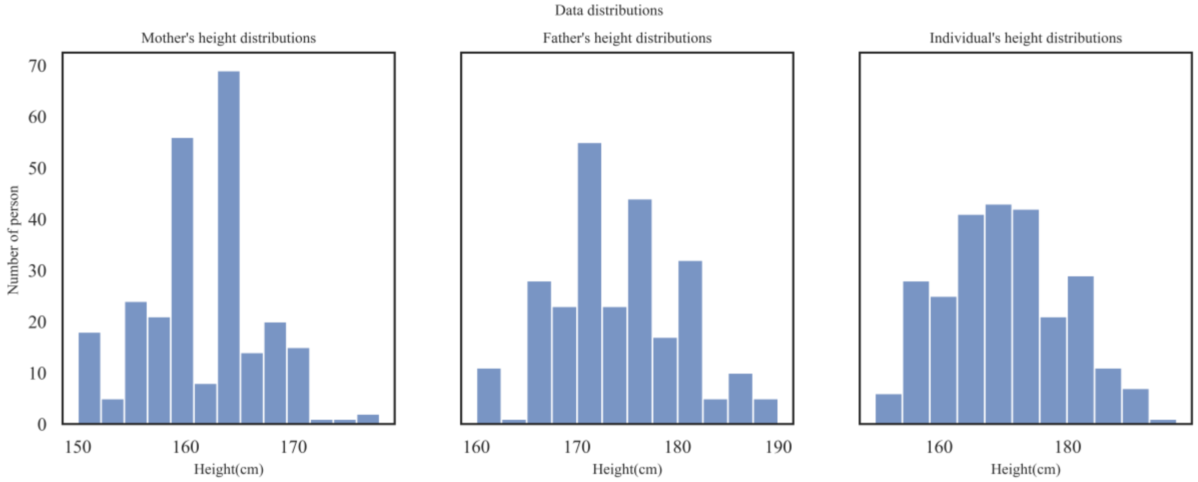
281 samples of data were collected from individuals aged 18 and over living in different regions of our country. In the data set, values less than 150 centimeters for the mother's height, values less than 160 centimeters for the father's height, and values greater than 6 for the number of children in the family were not taken. Meaningless data were also cleaned from the data set and analyzed with a 254-sample data set.

In the data set, mean maternal height is 161 centimeters, median is 161 centimeters, variance is 30, standard deviation is 5.5, shortest height is 150 centimeters, longest is 178 centimeters, range is 28, and the most repeated value is 160 centimeters. The mean father's height was 173 centimeters, the median was 173 centimeters, the variance was 40, the standard deviation was 6.3, the shortest height was 160 centimeters, the longest was 190 centimeters, the interval was 30, and the most repeated value was 170 centimeters. The average height of the individuals was 170 centimeters; the median was 170 centimeters; the variance was 91; the standard deviation was 9.5; the shortest height was 150 centimeters; the longest was 197 centimeters; the interval was 47; and the most repeated value was 170 centimeters. There are 145 female and 109 male individuals in the data set. There are three different economic situations. The general economic situation of the families is moderate. In the data set, these levels and examples are Bad (19), Average (206), and Good (29). The data set includes 7 geographical regions in total, and the Aegean Region is the largest sample in the data set from these regions. There are 27 samples of data from the Mediterranean Region, 16 from the Eastern Anatolia Region, 111 from the Aegean Region, 15 from the Southeast Anatolian Region, 24 from the Central Anatolia Region, 9 from the Black Sea Region, and 52 from the Marmara Region. Detailed information is given in Table 1.

Table 1. Statistical analysis of the data set

Attributes	Standard Deviation	Mean	Median	Variance	Minimum	Maximum	Range	Mod
Mother's height	5.532	161.618	161.0	30.608	150	178	28	160
Father’s height	6.392	173.393	173.0	40.864	160	190	30	170
Individual's chronic discomfort	0.375	0.169	0.0	0.141	0	1	1	0
gender	0.495	0.570	1.0	0.245	0	1	1	1
General economic status of the family	0.433	1.039	1.0	0.188	0	2	2	1
jumping sports status	0.472	0.334	0.0	0.223	0	1	1	0
Status of doing weight sports	0.392	0.188	0.0	0.153	0	1	1	0
Region where individual has lived the longest	1.930	2.897	2.0	3.728	0	6	6	2
Which child of the family	1.075	1.972	2.0	1.157	1	6	5	1
Height of individuals	9.556	170.122	170.0	91.324	150	197	47	170

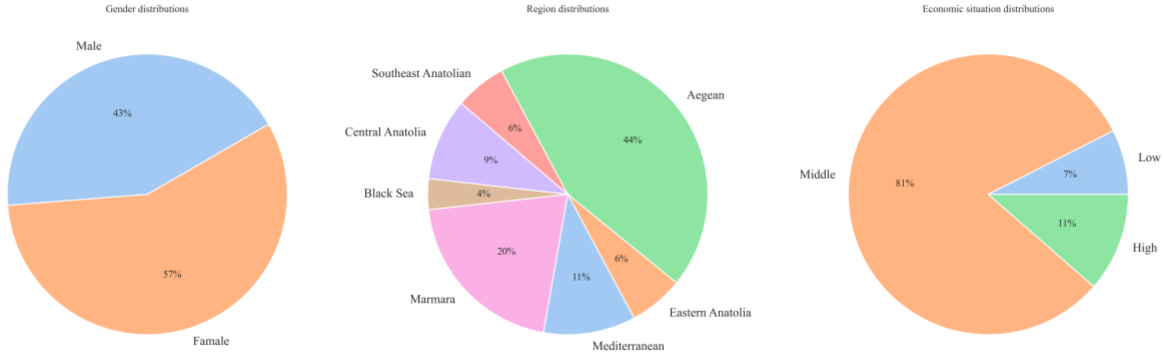
Height distributions of mothers, fathers and individuals in the data set are shown in Figure 1.



a) b) c)

Fig. 1. Data distribution of heights, a) Mother's height distribution b) Father's height distributions c) Individual's height distributions

The economic situation, gender and region distributions are shown in Figure 2.



a) b) c)

Fig. 2. The economic situation, gender and region distributions a) Gender distributions b) Region distributions c) Economic situation distributions

2.2. Machine Learning Algorithms

2.2.1. Artificial Neural Networks

The first practical use of neural networks was made when the perceptron network was released in the late 1950s. Neural networks based on brain structure originally developed in the 1940s [9]. Modeling nonlinear functions has been accomplished with the use of artificial neural networks (ANNs). They are capable of making predictions about various nonlinear functions that are accurate enough or almost so. Because they can anticipate nonlinear functions with such ease, neural networks are very useful for processing data [10].

Artificial neurons (ANN) have been patterned after their natural counterparts in the human brain and nervous system. Weighted connections link neurons together. The network consists of layers, or groups of neurons, with each layer's output feeding into the next. In this way, ANNs may be trained to solve regression issues and make predictions about their results [10]. In general, an ANN model has been made up of an activation function, weights, the sum of computed weights, input and output neurons, and a learning function. Weights indicate the connection strength of w_{ij} neurons. The b value indicates the deviation value. Net n_i represents the input neurons. The following equation has been used to calculate a rudimentary neural network model:

$$(\mathrm{n})_{\mathrm{j}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{w}_{\mathrm{ij}}\mathrm{x}_{\mathrm{i}} + \mathrm{b} \quad (1)$$

All neurons in feedforward networks are independent of one another and are only linked to one another via the neurons in the layer below them. The input of one layer becomes the output of the next layer. The connection between the layers takes place using weights. A feedforward ANN consists of data nodes that act as input neurons in the input layer, spreading the data to the hidden layer(s) and output layer via weights [10].

The input and output layers can have more than one neuron, depending on the result. The number of hidden layers or neurons that will provide the best results cannot be determined with certainty. Therefore, constructing an ANN architecture and making the optimal adjustment for the given problem requires experience and experimentation [11].

The activation function that the ANN utilizes for the model structure greatly influences the capability and effectiveness of finding a beneficial solution to a specific issue. The network's speed is greatly influenced by the activation function selection. Depending on the task at hand, several activation functions may be used in ANN models [12] and many different activation functions can be used. ReLU was employed as the activation function in this investigation, while the linear function was used in the bottom layer. Equation 2 specifies the linear activation function used in this research, and Equation 3 describes the ReLU activation function applied in this research:

$$f(n)_j = (n)_j \tag{2}$$

$$f(n)_j = \begin{cases} 0, & (n)_j < 0 \\ (n)_j, & (n)_j \geq 0 \end{cases} \tag{3}$$

One of the most well-known ANNs used to solve engineering challenges is the multilayer perceptron, which has successfully been shown to anticipate nonlinear correlations in a variety of applications [10]. To make a forecast, the ANN model must be trained. There are many different ways to train neural networks, and the way they work depends on the data set. Back propagation is a common and effective algorithm used to train multilayer sensor networks [13]. Backpropagation ANN sends the weight values of all the neurons in each layer to Equation 1, then moves on to the next layer. Then, the error for each layer is passed back to the layer below it. This is called the process of training or learning. During the training process, the network is shown a pair of templates that match the inputs and the outputs that are wanted. ANN uses weights and model thresholds to figure out the real outputs. By sending the error back over the network, the actual output is matched to what the network predicted. The weight values in each layer are changed to reduce the error in the output layer. The main goal of this process is to reduce the difference between the predicted output and the actual output [14]. ANNs need a large number of test cases in the training data set to be very accurate [15].

2.2.2 Linear Regression

LR includes a response variable y and a single predictive variable x. This is the simplest form of regression, and y is used as a linear function of x. It has been stated as Equation (4) [16]:

$$y = b + wx \tag{4}$$

It is assumed that the y -difference is constant and that b and w' are regression coefficients that determine the y-intercept and slope, respectively. Therefore, the regression coefficients b and w' can be weighted as follows [16]:

$$y = w_0 + \ w_1 x \tag{5}$$

The least squares approach, which calculates the best suited straight line with the smallest difference between the actual data and the line's predict, may be used to solve these coefficients. Variable D is the set of values for some population, x is the response variable, and y is the set of values associated with the response variable x. The training set contains (x_1, y_1), (x_2, y_2), …, (x_D, y_D), data points. The regression coefficients can be predicted using equations (6) and (7) [16].

$$\frac{\sum_{i=1}^{D}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{D}(x_i-\bar{x}\,)^2} \tag{6}$$

$$\mathrm{w}_0 = \bar{\mathrm{y}} - \mathrm{w}_1\bar{\mathrm{x}} \tag{7}$$

Here $x_1, x_2, \ldots, x_D$ is the mean of $\bar{x}$ and $y_1, y_2, \ldots, y_D$ is the mean of $\bar{y}$. The coefficients w_0 and w_1 generally provide good approximations to complex regression equations.

3. Results and Discussion

In the study, two different machine learning methods were used to predict height using 10 different features. These are the ANN and LR methods. In the evaluation of the designed models, the data set consisting of 254 samples was randomly divided into training and test data. Training data consists of 203 samples (80%) and test data consists of 51 samples (20%). The activation functions of the ANN layers were determined as ReLU and Linear, respectively. For the training of the created ANN, Adam was determined as the optimizer parameter, and MSE was determined for the loss parameter. For the training of the model, the epoch number is set to 150 and the batch size value is set to 16. All coding was done in Python using NumPy, Keras, Pandas, TensorFlow, Sklearn, Seaborn and Matplotlib libraries. In the study, R^2, Mean Absolute Error (MAE) and Mean Square Errors (MSE) were used in the performance evaluation of ANN and LR. R^2 is the coefficient that decides the accuracy of the model, so it is expected to give high results. It is expected to yield low results since MAE and MSE are the error criteria used for continuous variables [17]. MAE is calculated by taking the absolute value of the difference between the true value and the predicted value. It is expressed in MAE Equation 8, with y_i being the actual value and $\hat{y}_i$ being the predictive value [18], [19].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

MSE is expressed as the difference between the predicted and actual value. It provides a value indicating how close the fit line is to the data points. To return negative values to positive values, the value is squared. The smaller the MSE value, the better the performance of the model. It is calculated by taking the difference and averaging the squared value. The MSE is expressed in Equation 9 [20].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (9)$$

R^2 is using for evaluate how well the model is performing. It is expressed as the square of the correlation coefficient containing the actual and predicted values. R^2 value close to 1 indicates that the model has achieved good performance. R^2 is expressed in Equation 10 [21].

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (10)$$

The MAE value was found to be 2.720 for ANN and 2.780 for LR. The MSE value was found to be 12,754 for ANN and 10,857 for LR. With R^2 values, a success rate of 81.74% and 84.48% was obtained in ANN and LR, respectively. According to these measurement results, it was seen that the model designed with LR was more successful than the model designed with ANN. Measurement methods and results for ANN and LR are shown in Table 2.

Table 2. Results of machine learning algorithms

	Algorithms	
Evaluation Metrics	ANN	LR
MAE	2.720	2.780
MSE	12.754	10.857
R^2	0.817	0.844

The relationship between the predicted height value and the actual height value for ANN and LR has been shown in Figure 3 and Figure 4, respectively, with a line graph.

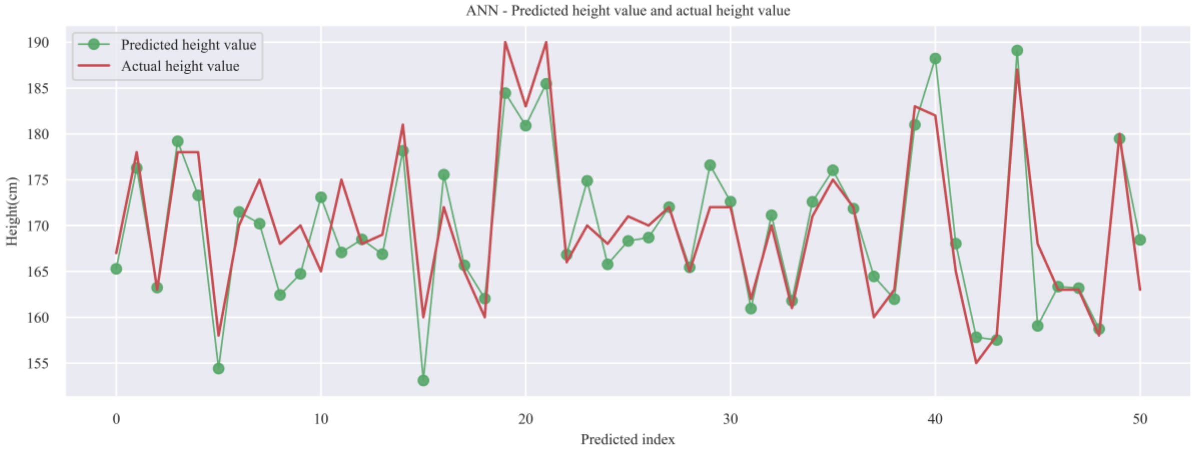


Fig. 3. Predicted height value and actual height value (ANN)

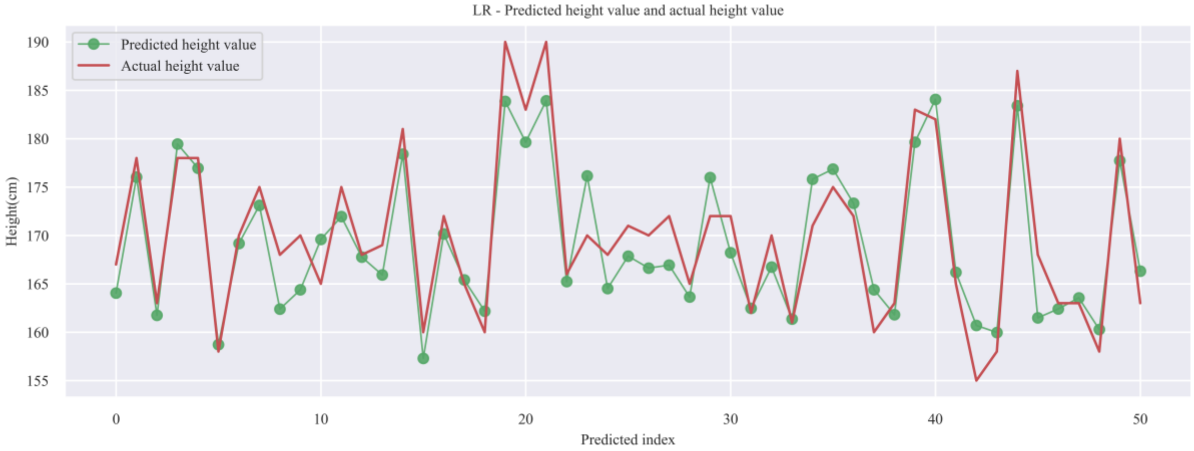


Fig. 4. Predicted height value and actual height value (LR)

The scatter plots between the predicted height value and the actual height value for ANN and LR have been shown in Figure 5.

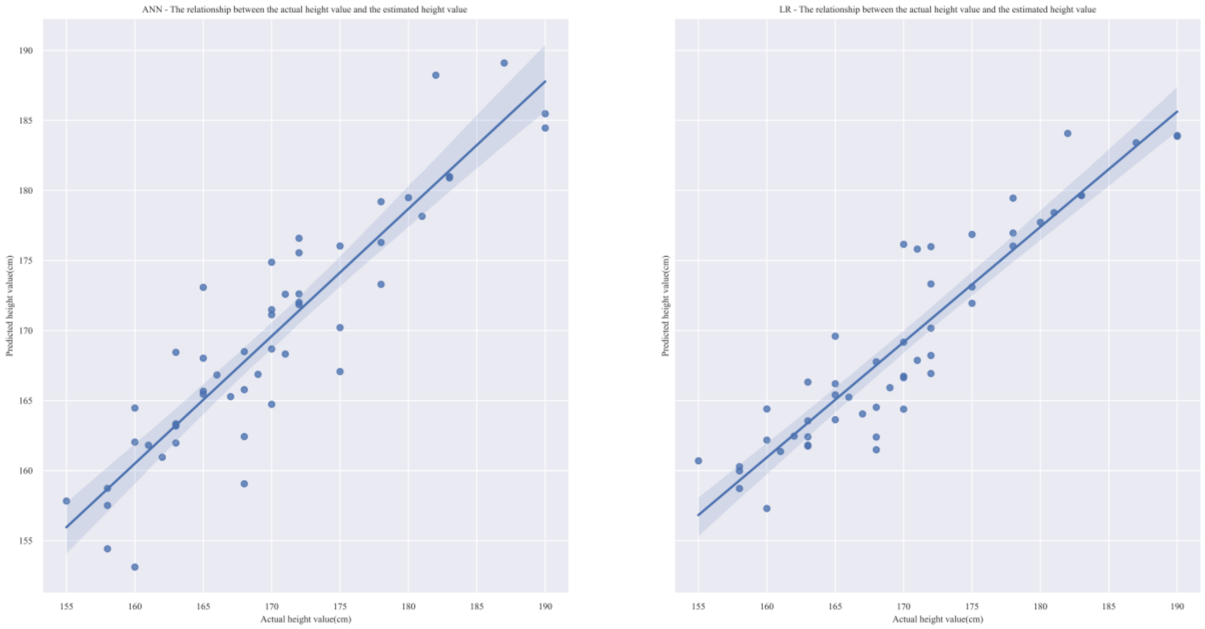


Fig. 5. Scatter plots of ANN and LR

4. Conclusion

The height of individuals was predicted by machine learning methods using 10 different non-genetic features. For this purpose, data were collected through questionnaires and made suitable for machine learning methods. Among the machine learning methods, the LR and ANN methods, which are the most frequently used in the literature, were preferred. It has been observed that the LR method can predict the height of individuals. As a result of the study, the LR method was able to perform approximately 3.3 percent better than ANN at R^2 value. Thus, it has been shown that the height of people, which has been tracked since the day they were born, can be predicted by machine learning methods in terms of how many centimeters it will be in adulthood. It is thought that the study will be beneficial for individuals and health centers. In future studies, performance can be improved with different methods by expanding the data set.

References

[1] Ummanel A, Dilek A. Gelişim ve öğrenme. Öğr İlke ve Yönt; (2016): 35-52.

[2] Uzun S. Yaşlılarda, kadınlarda ve adölasanlarda kişilik algısı değişimi ve nedenleri. J Humanit Soc Sci 2020; 3 (1): 431-449.

[3] Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. J Glob Health 2018; 8(2).

[4] Coşkun F, Gülleroğlu HD. Yapay zekânın tarih içindeki gelişimi ve eğitimde kullanılması. Ankara Univ J of Fac of Educ Sci (JFES) 2021; 54(3): 947-966.

[5] Ersöz F, Çınar Y. Veri madenciliği ve makine öğrenimi yaklaşımlarının karşılaştırılması: Tekstil sektöründe bir uygulama. Avrupa Bilim ve Teknoloji Dergisi 2021; (29): 397-414.

[6] Aytekin HT. Makine öğreniminin araştırmacıların veri analizi bağlamında potansiyel önemi. Ufuk Üniversitesi Sosyal Bilimler Enstitü Dergisi 10(19): 85-106.

[7] Atalay M, Çelik E. Büyük veri analizinde yapay zekâ ve makine öğrenmesi uygulamalari-artificial intelligence and machine learning applications in big data analysis. Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitü Dergisi 2017; 9(22): 155-172.

[8] Sarker IH. Machine learning: Algorithms, real-world applications and research directions. SN Comput Sci 2021; 2(3): 160.

[9] Esfe MH, Ahangar MRH, Rejvani M, Toghraie D, Hajmohammad MH. Designing an artificial neural network to predict dynamic viscosity of aqueous nanofluid of TiO2 using experimental data. Int Commun Heat Mass Transfer 2016; 75: 192-196.

[10] Ulas M, Altay O, Gurgenc T, Özel C. A new approach for prediction of the wear loss of PTA surface coatings using artificial neural network and basic, kernel-based, and weighted extreme learning machine. Friction 2020; 8: 1102-1116.

[11] Mukherjee A, Biswas SN. Artificial neural networks in prediction of mechanical behavior of concrete at high temperature. Nucl Eng Des 1997; 178(1): 1-11.

[12] Yu X, Ye C, Xiang L. Application of artificial neural network in the diagnostic system of osteoporosis. Neurocomputing 2016; 214: 376-381.

[13] Simpson PK. Artificial neural systems: foundations, paradigms, applications, and implementations. McGraw-Hill, Inc., 1991.

[14] Momeni E, Armaghani DJ, Hajihassani M, Amin MFM. Prediction of uniaxial compressive strength of rock samples using hybrid particle swarm optimization-based artificial neural networks. Measurement 2015; 60: 50-63.

[15] Dreyfus G. Neural networks: methodology and applications. Springer Science & Business Media 2005.

[16] Altay O, Gurgenc T, Ulas M, Özel C. Prediction of wear loss quantities of ferro-alloy coating using different machine learning algorithm. Friction 2020; 8: 107-114.

[17] Gültepe Y. Makine öğrenmesi algoritmaları ile hava kirliliği tahmini üzerine karşılaştırmalı bir değerlendirme. Avrupa Bilim ve Teknoloji Dergisi 2019; (16): 8-15.

[18] Iqbal N, Khan AN, Rizwan A, Ahmad R, Kim BW, Kim K, Kim DH. Groundwater level prediction model using correlation and difference mechanisms based on boreholes data for sustainable hydraulic resource management. IEEE Access 2021; 9: 96092-96113.

[19] Altay O, Varol Altay E. A novel hybrid multilayer perceptron neural network with improved grey wolf optimizer. Neural Comput Appl 2023; 35(1): 529-556.

[20] Gurgenç T, Altay O. St37 çeliğinin tornalanmasında yüzey pürüzlülüğünün destek vektör regresyonu kullanılarak tahmini. Firat Univ J of Eng Sci 2022; 34(2).

[21] Gurgenc T, Altay O. Surface roughness prediction of wire electric discharge machining (WEDM)-machined AZ91D magnesium alloy using multilayer perceptron, ensemble neural network, and evolving product-unit neural network. Mater Test 2022; 64(3): 350-362.