Çukurova Üniversitesi Mühendislik Fakültesi Dergisi, 37(3), ss. 793-802, Eylül 2022
Cukurova University Journal of the Faculty of Engineering, 37(3), pp. 793-802, September 2022

Performance of Machine Learning Methods in Location-Based Prediction

Nuh Mehmet ÖZMERDİVENLİ[1], ORCID *0000-0002-0854-2280*
Murat TAŞYÜREK[2], ORCID *0000-0001-5623-8577*
Serhat HIZLISOY[*2], ORCID *0000-0001-8440-5539*
Bahatdin DAŞBAŞI[3], ORCID *0000-0001-8201-7495*

[1]*Kayseri University, Department of Calculated Sciences and Engineering, Kayseri*
[2]*Kayseri University, Faculty of Engineering, Architecture and Design, Department of Computer Engineering, Kayseri*
[3]*Kayseri University, Faculty of Engineering, Architecture and Design, Department of Engineering Basic Sciences, Kayseri*

Geliş tarihi: 25.05.2022 Kabul tarihi: 23.09.2022

Atıf şekli/ How to cite: ÖZMERDİVENLİ, N.M., TAŞYÜREK, M., HIZLISOY, S., DAŞBAŞI, B., (2022). Performance of Machine Learning Methods in Location-Based Prediction. Çukurova Üniversitesi, Mühendislik Fakültesi Dergisi, 37(3), 793-802.

Abstract

Thanks to the technological developments that have taken place in recent years, the number, variety and quality of the data obtained using IoT (Internet of Things) sensors have been increasing. Data obtained from IoT sensors have been used in many scientific fields such as land use, climate change, vegetation analysis and air quality forecasting. In this study, a location-based spatial analysis application was carried out using the data obtained from IoT sensors with machine learning. With this application, the average temperature information of the station was estimated with Artificial Neural Network (ANN), Random Forests (RF), and Support Vector Machines (SVM) methods using daily average humidity, average pressure, and station altitude information on real datas of Kayseri acquired from the Turkish State Meteorological Service, and then performances of the methods were compared. In the experimental evaluations, the ANN, RF and SVM methods obtained an average of 0.83, 0.75 and 0.50 R^2 values. The ANN method outperformed the RF and SVM methods in location-based temperature estimation.

Keywords: Location based prediction, Spatial data, Machine learning

[*]Corresponding author (Sorumlu yazar) : Serhat HIZLISOY, *serhathizlisoy@kayseri.edu.tr*

Konum Tabanlı Tahminde Makine Öğrenme Yöntemlerinin Performansları

Öz

Son yıllarda meydana gelen teknolojik gelişmeler sayesinde IoT (nesnelerin interneti) sensörleri kullanılarak elde edilen verilerin sayısı, çeşitliliği ve niteliği artmaktadır. IoT sensörlerinden elde edilen bu veriler arazi kullanımı, iklim değişikliği, bitki örtüsünün incelemesi ve hava kalitesi tahmini gibi birçok bilimsel alanda kullanılmaktadır. Bu çalışmada, IoT sensörleri üzerinden elde edilen verileri makine öğrenmesi yöntemi ile kullanılarak konum bazlı mekânsal analiz uygulaması gerçekleştirilmiştir. Bu uygulama ile Meteoroloji Genel Müdürlüğü'nden alınan gerçek veriler içerisinden Kayseri ilinet ait günlük ortalama nem, ortalama basınç ve istasyon rakım bilgisi kullanılarak istasyonun ortalama sıcaklık bilgileri Yapay sinir ağı (ANN), Rasgele orman (RF) ve Destek vektör Makineleri (SVM) algoritmaları ile tahmin edilerek yöntemlerin performansları karşılaştırılmıştır. Deneysel değerlendirmelerde ANN, RF ve SVM yöntemleri ortalama 0,83, 0,75 ve 0,50 R^2 değeri elde etmiştir. ANN yöntemi konum bazlı sıcaklık tahmininde RF ve SVM yöntemlerine göre daha üstün performans göstermiştir.

Anahtar Kelimeler: Konum tabanlı tahmin, Mekansal veri, Makine öğrenmesi

1. INTRODUCTION

Data mining is applied in a large-scale data group to find the information that the researcher wants to discover in the study area. The classification and regression algorithms applied to obtain the desired information, analyze the data set within certain rules. With the increase in the processing power of computers and the ease of accessing large amounts of data due to technological developments, analysis processes using machine learning algorithms have become frequently used in scientific studies [1].

All data containing location information can be called spatial data and configured in vector, line, and point formats and raster (image) types. Raster data are mainly obtained from remote sensing system sensors. With remotely sensed systems, data sets can be created in areas such as forestry, agriculture, geology, natural resources, land cover detection, land management plans, illegal structure detection, destruction of forests, and natural protected areas. In addition, meaningful information can be obtained from these datas with classification methods.

Another important subject of the remote sensing discipline is to produce maps representing different earth features with the help of the datas obtained [2]. In addition, in the last 20 years, plant species and plant production areas have been successfully estimated with the help of various classification and estimation techniques using the data obtained by remote sensing [3]. With the increasing use of information systems, the need for up-to-date spatial data has emerged, and data obtained from sensors by remote sensing method has become more preferred than traditional methods [4]. There is a great increase in the amount of data produced and collected for spatial applications. For all these reasons, the remote sensing discipline needs machine learning algorithms to analyze multi-dimensional data.

In the literature, there are many studies in which machine learning algorithms are used to classify spatial data and for location-based prediction [4,6-12].

When the studies in the literature are examined as a location-based estimation method, Zolfaghari et al. [13] estimated Atterberg limits and indices to examine the use of soil and environmental data in 113 spatial locations using artificial neural network

(ANN) models at the western Iran basin scale. Hong et al. [14] calculated the landslide susceptibility indices by using the landslide inventory data containing 282 landslide locations with support vector machine (SVM) to create a landslide susceptibility map in Luxi city, Jiangxi province, China. Dharumarajan et al. [15] used the random forest model (RF), referencing 116 different spatial points, to predict the spatial variation of major soil features in the Bukkarayasamudrum Mandal in Anantapur district, India.

In this study, first of all, random forest algorithms, support vector machines, and artificial neural networks, which are some of the machine learning algorithms used for spatial data analysis, are introduced. Then, the performance of these methods was examined by using real data from the Meteorological Service. Finally, by using the daily average humidity, average pressure and station altitude data, the average temperature information of the station was estimated by ANN, RF, and SVM methods and their performances were compared.

2. MACHINE LEARNING ALGORITHMS FREQUENTLY USED IN LITERATURE FOR SPATIAL DATA ANALYSIS

In this section, some machine learning algorithms used for spatial data analysis are introduced.

2.1. Random Forest Algorithm

The Random Forest (RF) algorithm is a machine learning algorithm that creates decision trees by dividing the data into multiple subgroups. It is frequently used in the literature because it is fast and easy to apply in classifying spatial data [16]. With this method, multiple decision trees are created, and then predictions are produced for each decision tree. The classification process is performed by using the output result of the majority of the decision trees created. The data sets used in this method are randomly selected from the data set, and each is a subset of the data set. It is extreme against the over-fitting problem, as different and multiple data sets are used with the RF algorithm [17].

In the Random Forest algorithm, the user has to determine the number of trees (N) and the number of variables (m) used when creating tree structures [18].

After determining the relevant variables, samples are created from the training data, and tree formation is started for each sample. The best branching is determined with m randomly selected variables at each node.

If there is no separate data set for the test, 2/3 of the training data set is used as training data (in the bag), and 1/3 is used as test data (Out-of-Bag (OOB)) [19].

$$\sum\sum_{i=j}\left(\frac{f(C_i,T)}{|T|}\right)\left(\frac{f(C_j,T)}{|T|}\right) \quad (1)$$

CART (Classification and Regression Tree) algorithm is used for this process. The CART algorithm uses the GINI index given in Equation (1) to determine the best branch [8].

In Equation (1), T represents the training dataset, C_i represents the class to which the pixel belongs, and $\frac{f(C_i,T)}{|T|}$ represents the probability that the selected pixel belongs to the C_i class. The purpose of using the GINI index is to determine the homogeneity of the samples at each node. The algorithm selects the variable with the smallest GINI index calculated according to the randomly selected variables at each node, passes to the other node, and ends the branching if this index is zero [19]. In short, nodes are divided into branches, and tree structures are created according to the division criteria determined using the training data.

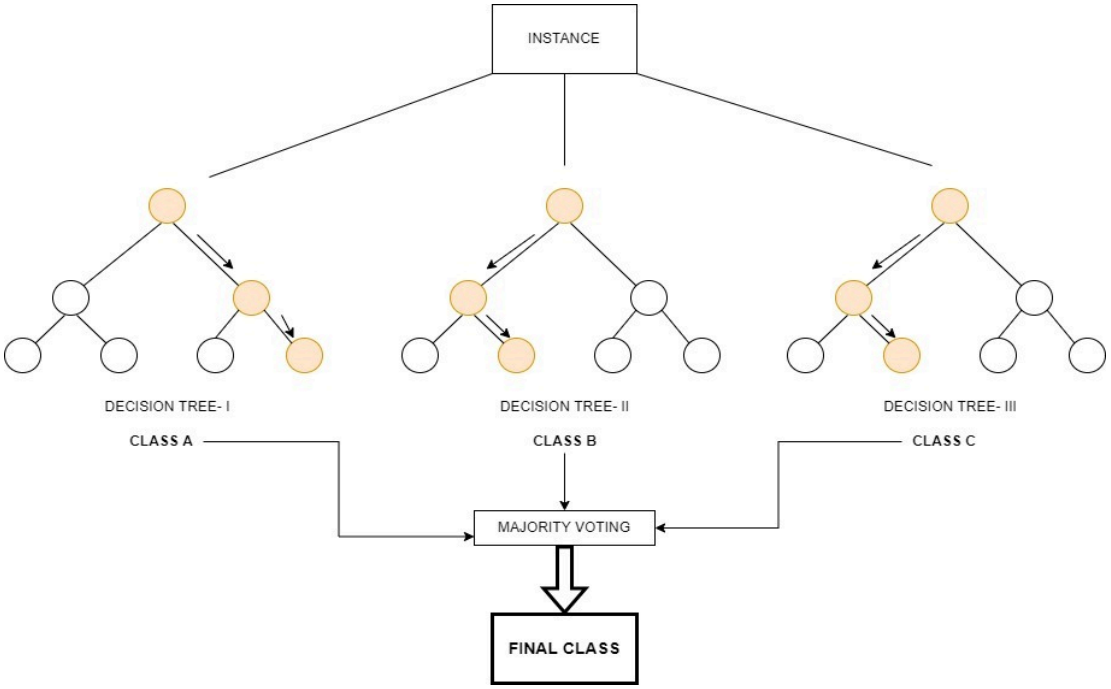


Figure 1. Diagram of random forest algorithm

2.2. Support Vector Machine

Support Vector Machine (SVM) algorithm is a machine learning algorithm used to separate data belonging to different classes from each other [20]. It is frequently used in spatial studies because it can classify high-dimensional hyperspectral data using limited training data [16]. Classification of both linear and non-linear data is possible using Support Vector Machines. For linear data, firstly, the boundary that will separate the two classes from each other is determined. Then, the decision limits on both sides of this border, called margin, are determined. The purpose of this algorithm is to make the distance between the decision boundaries as high as possible.

The classification process in nonlinear data is done by making the data set linear using kernel functions and then determining the most appropriate hyperplane. Kernel functions allow separating nonlinear separable support vectors using a linear plane [21].

Kernel functions are used to determine hyper-planes for the classification of remotely sensed datas with non-linear data characteristics with Support Vector Machines. Although many kernel functions have been defined in the literature, the radial basis function (RTF) is the most preferred kernel function due to its efficiency in problem-solving and its high classification accuracy [22].

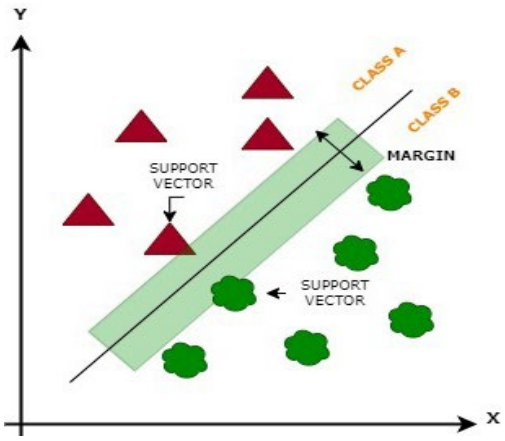


Figure 2. General representation of the SVM

Table 1. Kernel Functions that are frequently used for SVM in the literature

Kernel Functions	Formula	Description
Linear Kernel	$K(x,y)=xy$	-
Polynomial Kernel	$K(x,y)=((x.y)+1)^d$	d, Polynomial Degree
Radial Based Function Kernel	$K(x,y)=e^{-\gamma\|(x-x_i)\|^2}$	γ , Size of Gauss Kernel
Sigmoid Kernel	$K(x,y)=tanh(b(x.y)+r)$	b,r Kernel Parameters

2.3. Artificial Neural Networks

The artificial Neural Networks (ANN) model, inspired by biological neural networks, is frequently used in science and engineering problems. Neural networks consist of artificial units called neurons that work together to solve complex problems. The most basic task of the Artificial Neural Network is to determine an output set that can correspond to a given input set.

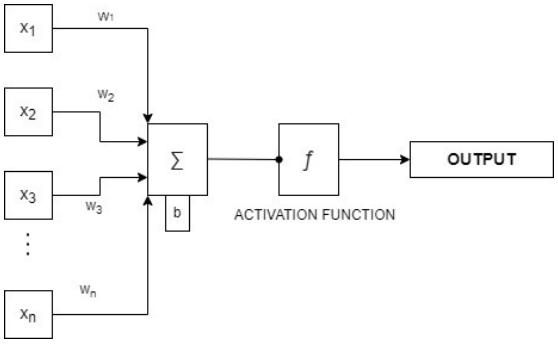


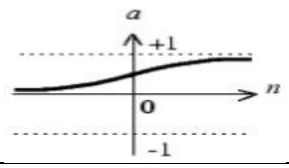
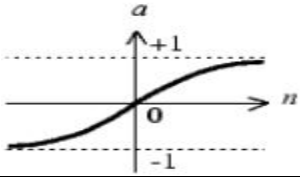
Figure 3. Basic steps of artificial neural network

In the ANN model shown in Figure 3; ($x_1,x_2,\ldots,x_n$) represent inputs, ($w_1,w_2,\ldots,w_n$) represent the weights of these inputs, b represent bias value ve f represents the decision function.

The weight values are automatically changed according to the specified learning rule by giving output values against a given set of inputs. In Artificial Neural Network architecture, the activation function is an important parameter that affects the accuracy of the system. The most used activation functions in the classification of remotely sensed datas are Log-Sigmoid and Hyperbolic Tangent [23].

An artificial neural network structure consists of 3 layers: input layer, hidden layer, and output layer. The flow of data entering the network is towards the output layer. The information transmitted from the input layer to the hidden layer is weighted here and transferred to the output layer. In the output layer, the result values are reached [23]. The most widely used method for training networks is backpropagation learning [24]. During the training, the weights are updated until the stopping criterion is reached, enabling the cost function of the data set to achieve the best result. In this method, the errors progress from the input to the output by decreasing.

Table 2. The most used activation functions in classification

Activation Function	Formula	Graphical Display
Log-Sigmoid	$y=\frac{1}{1+e^{-x}}$	
Hyperbolic Tangent	$y=\frac{1-e^{-2x}}{1+e^{2x}}$	

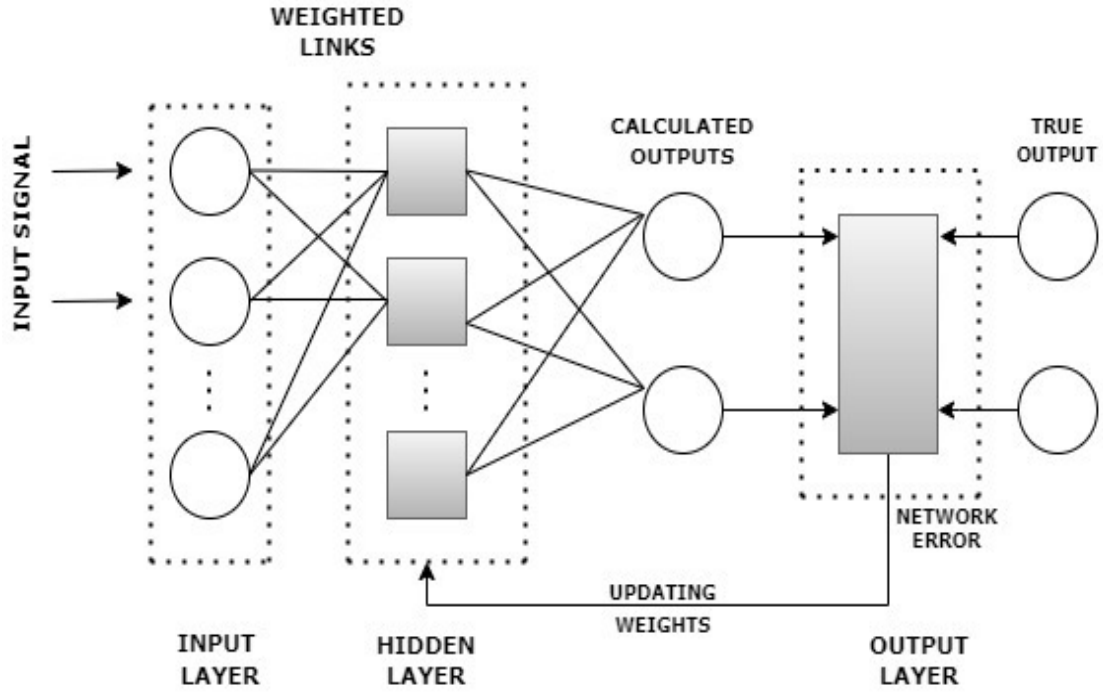


Figure 4. A feed forward backpropagation artificial neural network

3. APPLICATION

In this part, the performance of the methods is examined by using real data obtained from the Meteorological Service. The average temperature information of the station was estimated by ANN, RF, and SVM methods using daily average humidity, average pressure, and station altitude information. Each model was run separately for each station. In this way, a special formula has been produced for each location (the station's location where the station is located). The information on the meteorology stations used within the scope of this study is presented in Table 3. The view of the locations in the real world on the two-dimensional map plane is shown in Figure 5.

Table 3. Dataset for each station used in the application

Station No	Station Name	Altitude	Latitude	Longitude
17836	Develi	1204	38.3744	35.4797
17195	Kayseri Erkilet Airport	1053	38.7730	35.4908
17196	Kayseri Region	1094	38.6870	35.5000
18148	Kocasinan / Yamula Dam	1075	38.9028	35.2695
17840	Sarız	1599	38.4781	36.5035
18207	Yesilhisar	1141	38.3408	35.0875
17837	Tomarza	1402	38.4522	35.7912
18149	Melikgazi /Erciyes Ski Center	2210	38.5428	35.5244
17802	Kayseri / Pinarbasi	1542	38.7251	36.3904

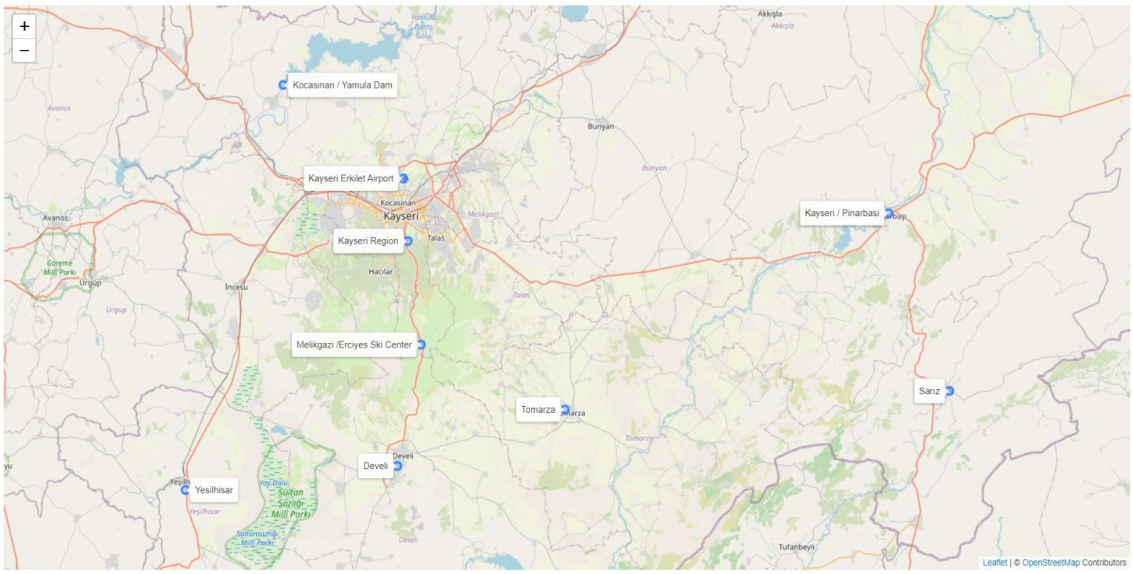


Figure 5. Locations of stations on the map

In the experimental evaluations, the temperature information of the station was estimated by using the altitude, pressure, and humidity information for each station using ANN, SVM, and RF methods. For each station, approximately %80 of the data in the dataset was used as a training set and %20 as the test set.

The R^2 in Equation (2) was used to evaluate the performance of the methods.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \tilde{y})^2} \quad (2)$$

Here, n is the number of samples in the dataset, y_i is the actual value, $\hat{y}$ is the value predicted by the model and $\tilde{y}$ is the average of the values predicted by the model. The fact that the R^2value is close to 1 indicates that the proposed model finds the relationship between the input data and the output data at the maximum level. If this value is close to 0, it is understood that the proposed method cannot model the relationship between input data and output data. The R^2 value is calculated with the above formula. Performances according to the R^2 the value obtained by ANN, SVM, and RF methods are presented in Table 4 and Figure 6.

Table 4. Performances of estimation methods

Station	Total Data	Training Set	Test Set	ANN	RF	SVM
Develi	6274	5019	1255	**0.74**	0.69	0.69
Kayseri Erkilet Airport	6033	4826	1207	**0.95**	0.91	0.67
Kayseri Region	6286	5029	1257	**0.76**	0.67	0.50
Kocasinan / Yamula Dam	1602	1282	320	**0.79**	0.71	0.57
Sarız	6290	5032	1258	**0.86**	0.74	0.58
Yesilhisar	1578	1262	316	**0.76**	0.65	0.48
Tomarza	6272	5018	1254	**0.98**	0.86	0.53
Melikgazi /Erciyes Ski Center	1506	1205	301	**0.74**	0.65	0.55
Kayseri / Pinarbasi	6302	5042	1260	**0.88**	0.86	0.50
			Mean R^2 values	**0.83**	0.75	0.56

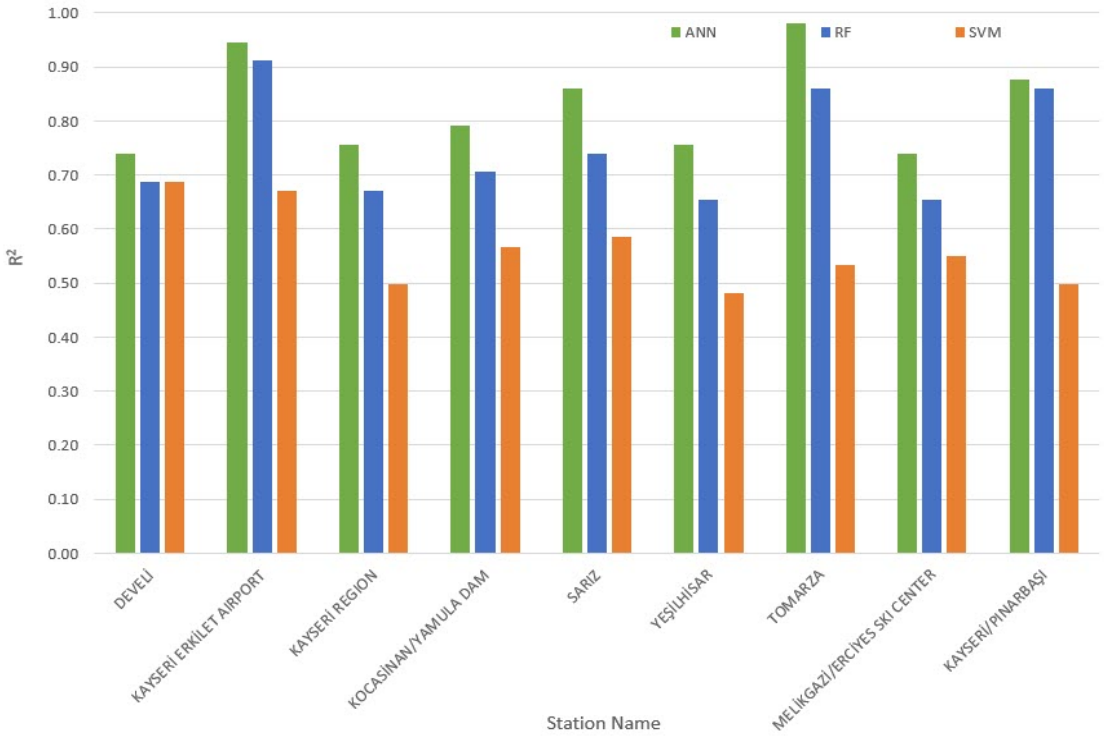


Figure 6. Performance graph of estimation methods

When Table 4 and Figure 6 are examined in detail, the ANN method performed better than RF and SVM methods in temperature estimation of the Meteorological Service stations. With the ANN method, the highest 0,98 R^2 value and the lowest 0,74 R^2 value was obtained. With an R^2 value of 0.74, it was seen that the ANN method modeled the relationship between temperature, pressure, humidity, and altitude superior to other models. The highest 0.91 R^2 value and the lowest 0.65 R^2 the value was obtained with the RF method. On the other hand, the SVM method obtained an R^2 value of 0.69 in the best case and 0.48 in the worst case. Since the SVM method could not find the R^2 value at a good level, it was able to model the relationship between temperature, pressure, humidity, and altitude at a lower rate than other models. In general, when the performance of the methods is examined according to the R^2 value, the ANN method has achieved superior performance to the RF and SVM methods. On the other hand, the RF method achieved superior performance compared to the SVM method.

4. CONCLUSION

Considering the developments in IoT techniques, computer hardware, and software, many studies have been carried out in the literature in recent years in remote sensing using different classification methods. One of the most important parameters affecting classification accuracy is the quality of the data to be used in the study field. On the other hand, when the recent studies are examined, it is observed that RF, SVM, and ANN methods are frequently used.

In this study, the average temperature information of the station was estimated with ANN, RF, and SVM methods using daily average humidity, average pressure, and station altitude information on actual data obtained from the Meteorological Service, and then performances of the methods were compared. In the experimental evaluations, R^2 values for ANN, RF, and SVM models were examined. ANN, RF and SVM models obtained

0.98, 0.91 and 0.69 in good condition, 0.74, 0.65, 0.48 in bad condition, respectively. The ANN method modeled the location-based temperature, pressure, and humidity relationship better than the RF and SVM models in the best and worst cases.

5. REFERENCES

1. Yan, X., Ai, T., 2018. Analysis of Irregular Spatial Data with Machine Learning: Classification of Building Patterns with a Graph Convolutional Neural Network. ArXiv Preprint ArXiv:1809.08196.
2. Torunlar, H.M.G., Tuğaç, M.G., Duyan, K., 2021. Nesne Tabanlı Sınıflandırma Yönteminde Sentinel-2A Uydu Görüntüleri Kullanılarak Tarımsal Ürün Desenlerinin Belirlenmesi; Konya - Karapınar Örneği. Türkiye Uzaktan Algılama Dergisi, 3(2), 36-46.
3. Li, L., Zheng, X., Zhao, K., Li, X., Meng, Z., Su, C., 2020. Potential Evaluation of High Spatial Resolution Multi-spectral Images Based on Unmanned Aerial Vehicle in Accurate Recognition of Crop Types. Journal of the Indian Society of Remote Sensing, 48(11), 1471-1478.
4. Kavzoğlu, T., Çölkesen, İ., Şahin, E.K., 2015. Obje Tabanlı Yaklaşımda Makine Öğrenme Algoritmalarının Sınıflandırma Performansının Analizi. TUFUAB VIII. Teknik Sempozyumu, 344-349.
5. Sabek, I., Mokbel, M.F., 2020. Machine Learning Meets Big Spatial Data. In 2020 IEEE 36th International Conference on Data Engineering (ICDE), 1782-1785.
6. Shang, X., Chisholm, L.A., 2013. Classification of Australian Native Forest Species Using Hyperspectral Remote Sensing and Machine-Learning Classification Algorithms. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7(6), 2481-2489.
7. Cracknell, M.J., Reading, A.M., 2014. Geological Mapping Using Remote Sensing Data: A Comparison of Five Machine Learning Algorithms, Their Response to Variations in the Spatial Distribution of Training Data and the Use of Explicit Spatial Information. Computers & Geosciences, 63, 22-33.
8. Pal, M., Mather, P.M., 2005. Support Vector Machines for Classification in Remote Sensing. International Journal of Remote Sensing, 26(5), 1007-1011.
9. Dilek, K.S., 2013. Kentsel Alanların WorldView-2 Uydu Görüntülerinden Makine Öğrenme Algoritmaları Kullanılarak Tematik Haritalanması. Jeodezi ve Jeoinformasyon Dergisi, (107), 71-80.
10. Özdarıcı, O.A., Akar, Ö., Güngör, O., 2011. Rastgele Orman Sınıflandırma Yöntemi Yardımıyla Tarım Alanlarındaki Ürün Çeşitliliğinin Sınıflandırılması. TUFUAB 2011 VI. Teknik Sempozyumu, 1-7.
11. Ntouros, K.D., Gitas, I.Z., Silleos, G.N., 2009. August. Mapping Agricultural Crops with EO-1 Hyperion Data. In 2009 First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing. IEEE, 1-4.
12. Tunca, E., Köksal, E., 2021. Sentinel 2 Uydu Görüntülerinden Bitki Türlerinin Makine Öğrenmesi ile Belirlenmesi. ÇOMÜ Ziraat Fakültesi Dergisi, 9(1), 189-200.
13. Zolfaghari, Z., Mosaddeghi, M.R., Ayoubi, S., 2015. ANN-based Pedotransfer and Soil Spatial Prediction Functions for Predicting Atterberg Consistency Limits and Indices from Easily Available Properties at the Watershed Scale in Western Iran. Soil Use and Management, 31(1), 142-154.
14. Hong, H., Pradhan, B., Jebur, M.N., Bui, D.T., Xu, C., Akgun, A., 2016. Spatial Prediction of Landslide Hazard at the Luxi Area (China) Using Support Vector Machines. Environmental Earth Sciences, 75(1), 1-14.
15. Dharumarajan, S., Hegde, R., Singh, S.K., 2017. Spatial Prediction of Major Soil Properties Using Random Forest Techniques-A Case Study in Semi-arid Tropics of South India. Geoderma Regional, 10, 154-162.
16. Üstüner, M., Şanlı, F.B., 2019. Çok Zamanlı Polarimetrik SAR Verileri ile Tarımsal Ürünlerin Sınıflandırılması. Jeodezi ve Jeoinformasyon Dergisi, 7(1), 1-10.
17. Watts, J.D., Lawrence, R.L., 2008. Merging Random Forest Classification with an Object-oriented Approach for Analysis of Agricultural

Lands. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 37(B7).

18. Breiman, L., 2001. Random forests. Machine learning. Kluwer Academic Publishers, 45(1), 5-32.

19. Erdem, F., Derinpınar, M.A., Nasirzadehdizaji, R., Selen, O.Y., Şeker, D.Z., Bayram, B., 2018. Rastgele Orman Yöntemi Kullanılarak Kıyı Çizgisi Çıkarımı İstanbul Örneği. Geomatik, 3(2), 100-107.

20. Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random Forests for Land Cover Classification. Pattern Recognition Letters, 27(4), 294-300.

21. Kumar, P., Gupta, D.K., Mishra, V.N., Prasad, R., 2015. Comparison of Support Vector Machine, Artificial Neural Network, and Spectral Angle Mapper Algorithms for Crop Classification Using LISS IV Data. International Journal of Remote Sensing, 36(6), 1604-1617.

22. Yu, J.H., Ge, L., Rizos, C., 2011. Digital Elevation Model Generation Using Multibaseline Advanced Land Observing Satellite/phased Array Type L-band Synthetic Aperture Radar Imagery. Journal of Applied Remote Sensing, 5(1), 053510.

23. Kavzoglu, T., Colkesen, I., 2009. A Kernel Functions Analysis for Support Vector Machines for Land Cover Classification. International Journal of Applied Earth Observation and Geoinformation, 11(5), 352-359.

24. Atasever, Ü.H., Özkan, C., 2012. Arazi Örtüsünün Belirlenmesinde Torbalama Karar Ağaçları Yönteminin Kullanılması. UZAL-CBS, Zonguldak.

25. Rokach, L., 2010. Pattern Classification Using Ensemble Methods. Series in Machine Perception and Artificial Intelligence, World Scientific, 75.