ESTUDAM Bilişim Dergisi
Cilt 3, Sayı 2, 23-30, 2022

Journal of ESTUDAM Information
Volume 3, Issue 2, 23-30, 2022

DOI:https://doi.org/10.53608/estudambilisim.1071335

(Geliş Tarihi / Received Date: 11.02.2022, Kabul Tarihi/ Accepted Date: 24.03.2022)

(**Research Article**)

Improving Iris Dataset Classification Prediction Achievement by Using Optimum k Value of kNN Algorithm

Ahmet ÇELİK[*1]

[1]Kütahya Dumlupınar Üniversitesi, Tavşanlı Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, 43300, Tavşanlı/Kütahya, ORCID No : http://orcid.org/0000-0002-6288-3182

Keywords:
Classification,
Prediction
kNN,
Data mining,
Machine learning

Abstract: Machine learning methods are widely used in automated technologies. Classification prediction is a machine learning based on data mining. Today, many technological devices can make new predictions by getting experience from past data with machine learning methods. Machine learning is widely studied in two types, supervised and unsupervised. The limits of the objectives in supervised learning are predetermined. In unsupervised learning, there are no predetermined targets. In this learning, the machines are required to determine the targets automatically. Prediction process is one of the basic components of machine learning. Machines need to use some algorithms in order to perform the prediction process on the basis of data mining. k nearest neighbor (kNN), Naive Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM) algorithms are used mostly. k nearest neighbor (kNN), Naive Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM) algorithms are used mostly. These algorithms can be applied with the help of some tools on data sets. In this study, kNN algorithm was used to estimate Iris data set classification using Orange tool. The success of the KNN algorithm depends on using the correct attribute and changing the optimum k value. As a result of the tests, it was determined that when the k neighbor value was selected as 15, it was the most suitable k neighbor value, providing 98.67% classification prediction success in the Iris dataset.

(**Araştırma Makalesi**)

kNN Algoritmasının Optimum k Değerini Kullanarak İris Veri Seti Sınıflandırma Tahmin Başarısının İyileştirilmesi

Anahtar Kelimeler:
Sınıflandırma,
Tahmin,
kNN,
Veri madenciliği
Makine öğrenmesi

Özet: Otomatik çalışan teknolojilerde, makine öğrenmesi yöntemleri olarak yaygın kullanılmaktadır. Sınıflandırma tahmini, veri madenciliği temeline dayanarak gerçekleştirilen bir makine öğrenmesidir. Makine öğrenmesi, makinelerin geçmiş verilerden tecrübe elde ederek yeni tahminlerde bulunmasına olanak sağlamaktadır. Makine öğrenmesi yaygın olarak denetimli ve denetimsiz olarak iki tür olarak incelenmektedir. Denetimli öğrenmede hedeflerin sınırları önceden belirlenmiştir. Denetimsiz öğrenmede ise önceden belirlenmiş hedefler yoktur. Bilgisayarların hedefleri otomatik belirlemesi istenmektedir. Tahmin işlemi makine öğrenmesinin temel bileşenlerinden birini oluşturmaktadır. Bilgisayarlar tahmin işlemini gerçekleştirebilmek için veri madenciliği temelinde bazı algoritmaları kullanması gerekmektedir. En çok k en yakın komşu (kNN), Naive Bayes (NB), Karar Ağacı (DT) ve Destek Vektör Makinesi (SVM) algoritmaları kullanılmaktadır. Bu algoritmalar bazı araçlar kullanılarak veri setleri üzerinde uygulanabilmektedir. Bu çalışmada Orange aracı kullanılarak İris veri seti üzerinde kNN algoritmasıyla tahmin işlemi gerçekleştirilmiştir. kNN algoritmasının başarısı doğru öznitelik kullanmaya ve optimum k değerinin kullanılmasına bağlıdır. Yapılan testler sonucunda, k komşu

*İlgili yazar/corresponding author: ahmet.celik@dpu.edu.tr

değeri 15 seçildiğinde İris veri setinde %98,67 sınıflandırma tahmin başarısı sağlayarak, en uygun k komşu değeri olduğu belirlenmiştir.

1. INTRODUCTION

Today, automated computer tools and using machine learning algorithms have become necessary. These technologies make easy to analyze complex and big data [1]. However, machine learning is not limited to computers. Many devices using embedded systems it also uses machine learning algorithms now. Learning methods are widely used even in mobile devices that have a very important presence in human life. Machine learning is based on data mining analysis.

Machine learning is a method of learning from historical data and making predictions with the help of algorithms. There is no programming in this learning method. In Machine Learning, performance measurement while using existing data sets and making improvements when necessary will contribute greatly to technological developments. The most used learning method is supervised and unsupervised learning methods: Supervised learning is learning that consists of known feedback and response pairs when training examples are tested. In this method, a mathematical function is used to predict future results by minimizing errors. Unsupervised learning is a learning method that is deprived of relevant target values within the data set. In this method, the dataset is an array that validates how likely it is to detect a particular object in the future [2].

Machine Learning is based on predicting invisible data or test data. A number of algorithms are needed to perform tasks in machine learning. The features chosen for evaluating machine learning algorithms are very important. Feature selection also affects the performance value of machine learning algorithms. Therefore, using different features and choosing the algorithm are important parameters. Classification problems are the most common type of machine learning problem. Therefore, there are numerous features that can be used to evaluate predictions for these problems. Classification accuracy is the accuracy of all predictions made. This is the most common assessment measure for classification problems. With an equal number of observations in each class, the accuracy measurement of the prediction classification can be realized [3]. kNN, is one of the simplest classification algorithms available for supervised learning. However, kNN has less computational intensity. The basis of this algorithm is to search for the nearest neighbor. When the kNN closest algorithm is tested on very large data blocks, it takes a long time to find the closest neighbors [4].

Thirunavukkarasu et al. [3], worked on supervised learning in writing. They used K-Nearest Neighbors (kNN) classification algorithm on the iris data set. In this study, they aimed to develop a model that can automatically recognize iris types. For this, they used Numpy, Pandas, Matplotlib tools.

Yigit [5], proposed a weighting method for the k nearest neighbor (kNN) algorithm in his study. In this study, they found the optimal weights with the Artificial Bee Colony (ABC) algorithm. This algorithm mimics the foraging behavior of honey bees in search of quality food. Since it is an ABC algorithm, it is used in many applications. Four UCI data sets (Iris, Haberman, Breast Cancer and Zoo) were used in the study. Results show that it improves its correct performance in the Iris, Haberman and Breast Cancer data set. However, when applied to the Zoo data set, there was a decrease in performance. In this study, it has been shown that ABC algorithm can be applied together with kNN algorithm.

Kanju et al. [6], in their study, tested the kNN algorithm, Naive Bayes, Decision Tree, Support Vector Machines, Neural Network, and Random Forest machine learning algorithms for predict phishing sites. Many experienced users know how to spot a phishing site, but because of the pace of life, users ignore the ambiguities in domain names. That's why they unknowingly fill in many forms of the secret phishing website. Thus, they cause their personal information to be leaked.

Agrawal et al. [7], in their study, used feature reduction method on a system consisting of GPU(Graphics processing unit) to reduce kNN Regression and Random Forest Regression estimation times. In the study, the performance of the proposed system to predict the operating time of the processors was found to have an accuracy of 98.46% after using feature reduction techniques.

Lahmer et al. [1], proposed a new approach that aims to classify two types of genes in their study. Generally, gene classification is based on processing numerical data. However, image data were used to be more efficient in the study. SVM and kNN classification algorithms were used on the image data they produced for each gene. In the study, it was seen that it achieved 14% higher success than the latest technology [1].

Alkhatib et al. [8], in their study, applied the k-nearest neighbor algorithm and nonlinear regression approach to predict the stock prices of six major companies listed on the Jordan stock market to help investors, management, decision-makers and users make the right decisions. The kNN algorithm has achieved successful results with a high error rate. Stock price prediction is an interesting and challenging subject of research. Stock Markets with large and dynamic sources of information are considered as a suitable environment for data mining and business researchers.

Goel et al. [9], in their study; they created a forecast model by using revenue, usage and company graph data to identify Telekom Corporate Customers. Their accuracy has been tested by changing the k value from 3

to 17. The highest accuracy rate is obtained when the neighbor value of k is 5.

Bütüner et al. [10], in their study; they used kNN, RseslibKnn and A1DE machine learning methods in the diagnosis of Parkinson's disease and compared their success. According to the results they obtained, RseslibKnn was found to be the most successful method.

Tosunoğlu et al. [11], in their study; they used Support Vector Machines (SVM), Adaptive Boosting (AdaBoost), K-Nearest Neighbours (KNN) and Random Forest machine learning algorithms algorithms for predicting monthly streamflow the Coruh river.

Çimen et al. [12], in their study; they used SVM, KNN, Bayesnet, NavieBayes, J48 and Random Forest algorithms for classification NSL-KDD dataset using WEKA interface. They obtained the best performance rate of 98.1237% with kNN algorithm.

Ozkan et al. [13], in their study; they used advanced kNN method for clasification pistachios. They 2148 pistachios images used in the dataset. They obtained performance rate of 94.18% with advanced kNN algorithm.

Ozmen-akyol & Gulbandilar [14], in their study; they showed that iris flower type can be predicted with very little error by fuzzy artificial neural network (ANFIS) and artificial neural network (ANN) methods.

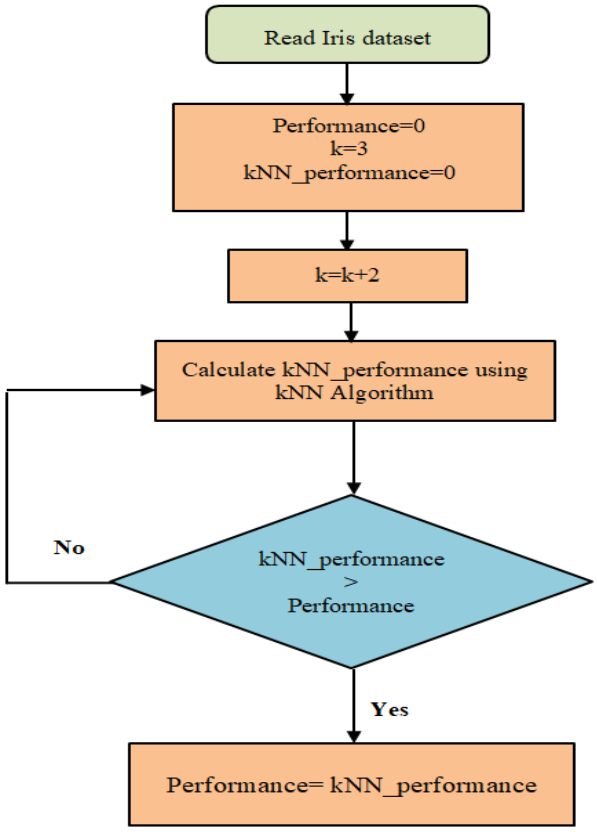

Figure 1. Flowchart of the method that detects high performance.

In this study; while classifying using the kNN neighbor algorithm, the effect of the selected k closest neighbor value on performance success has been shown and the optimum k value has been found. The Iris data set is most commonly used in machine learning based classification estimates. In this study, the most appropriate k value was determined to be 15 in the classification on the Iris data set data. When this value is chosen, the highest success rate was found to be 98.67%.

2. MATERIALS AND METHODS

In this study kNN (k nearest neighbors) algorithm is used. k Neighboring values with the highest success rate were found on the Iris data set. A flowchart is shown on Figure 1.
Performance comparison is made depending on the change of k value in the kNN algorithm in the flow chart and the highest success performance is determined.

2.1. k Nearest neighbor algorithm (kNN)

The k-Nearest Neighbor (kNN) algorithm was developed in 1968 by Cover and Hart [15]. The kNN algorithm is a supervised learning method. This algorithm requires a well-tagged data training dataset. The Euclidean distance between the new data point and the labeled data Point is calculated. The class of the new entry is determined by taking the nearest k samples [1].

This algorithm is a simple and easily applicable machine learning algorithm. This algorithm new incoming sample is included in the category (class) to which most of the k close samples belong. Traditional classification algorithms are shown as Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM) and K-Nearest Neighbor (kNN). Among these, kNN classification algorithm has the advantages of simple algorithm principle, mature theory structure, high classification accuracy and easy application. Therefore, it is widely used in many areas. However, the calculation of the kNN classification algorithm takes a long time [15].

However, by paralleling the kNN algorithm on the Hadoop platform, data sets can be classified quickly [16]. The K-nearest neighbor technique is a machine learning algorithm that is considered simple to implement. The distance of the newly entered data to each data in the training set is determined by the Euclidean distance algorithm. A majority vote is made among the k records selected to determine the class label, and then the new sample is added to the specified class [6, 17]. The Euclidean distance vector calculation is shown on equation 1.

$$d_{Euclidean}(X_i, Y_i) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (1)$$

X_i shows the i'th sample value, which is entered from the outside and whose class will be determined. Y_i is the sample in the data set whose class is specific. n is the

number of attributes. $d_{Euclidean}(X_i, Y_i)$, is the distance value between X_i and Y_i. [15, 18].

2.2. Iris dataset

UCI (University of California, Irvine) and Scikit-learn online machine learning data libraries are widely used. These libraries are available as open source. Many data sets can be accessed that support supervised and unsupervised learning in these libraries. There are also tools for classification estimation [19].

The Iris dataset include classification data and is well known. The dataset contains the species classes of the iris plant. The iris plant is also known as the rainbow plant. The data set contains 150 data samples. These classes are called Iris-setosa, Iris-versicolor and Iris-virginica. Figure 2 shows examples of the three classes.

a)

b)

c)

Figure 2. Classes in Iris dataset [3]: a) Iris-setosa, b) Iris-versicolor and c) Iris-virginica

Classes are made according to the width and length attributes of the petal and the sepal. These attributes consist of four different numerical values. The values in the data set are all real. Table 1 shows the properties of the Iris dataset. One class can be linearly separated from the other two in the Iris data set. However the other two cannot be separated linearly [20].

Table 1. Iris dataset information

Classes	Attributes	Feature	Number of Samples
Iris-setosa	Sepal length	Real	150
Iris-versicolor	Sepal width		
Iris-virginica	Petal width		
	Petal length		

3. FINDINGS AND DISCUSSION

Separating the data into the Iris-setosa class can be performed linearly in the Iris data set. So the attribute values are quite different from other Iris-versicolor and Iris-virginica classes. However, it is very difficult to distinguish between Iris-versicolor and Iris-virginica classes so nonlinear classification can be performed. Figure 3 shows these classes of Iris. In this study, it is aimed to perform classification Iris data set automatically with kNN algorithm on the basis of machine learning. The high success of this classification will reveal its reliability. The classification with the highest success rate is realized when the neighbor value of k is selected as 15. Figure 4 shows this classification.

It can be seen that the two values marked (circled) are incorrectly classified in the figure. One of the wrong classifications is made in the Iris-versicolor class. The other was made in the class Iris-virginica. Iris-setosa classification is completely correct.

It is k parameter that should be optimized in the kNN algorithm. Since the neighbor value k changes for different data, It should be selected in the most appropriate way according to the classification accuracy. It must be large enough and small enough that other class instances are not included [9]. In this study, classification success rates were determined according to different k neighbor values.

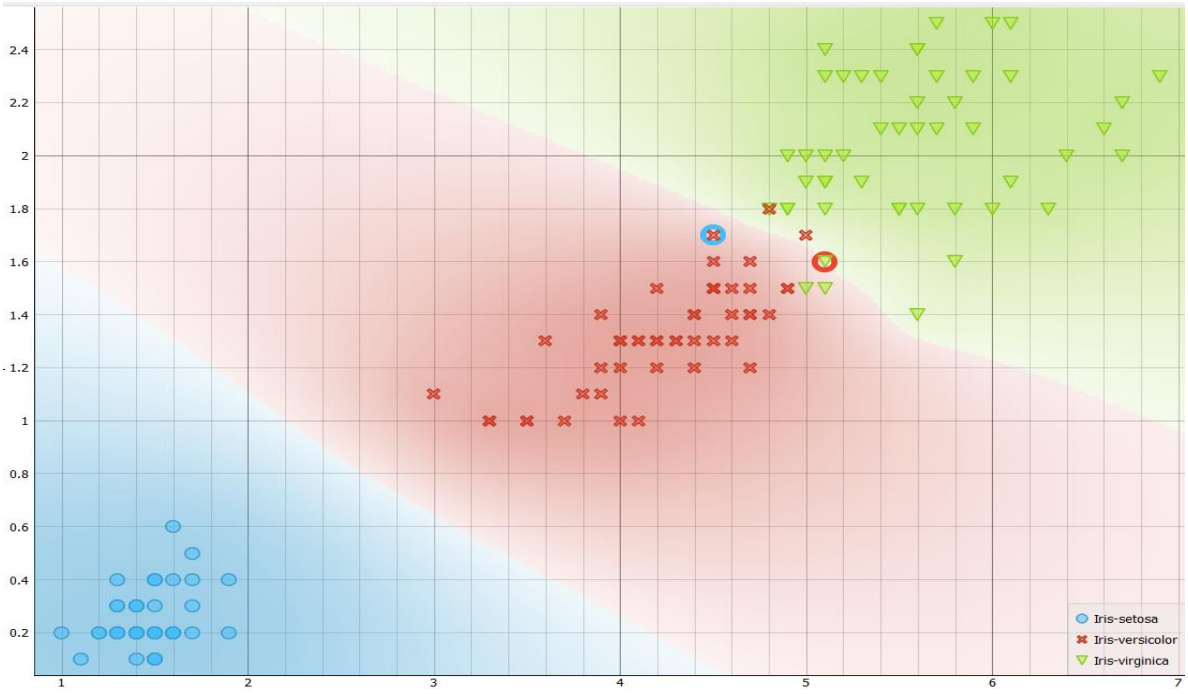

Figure 3. Actual class values of Iris dataset

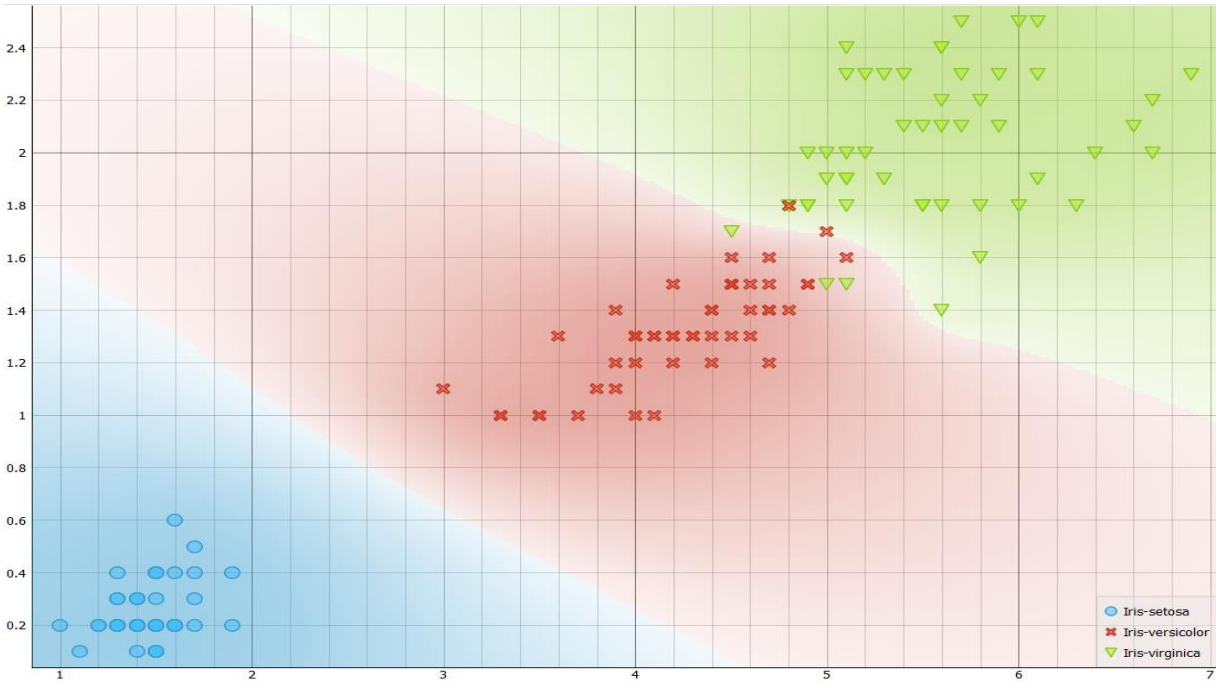

Figure 4. Class prediction result when kNN algorithm is applied on Iris dataset.

These success rates are shown in Table 2. k neighbor values, iris classes, Iris-virginica, Iris-versicolor and Iris-setosa correct classification detection values and success rates of each classification are shown in the table.

Classification was made for Iris-virginica and Iris-versicolor with the highest 98% success rate. The lowest classification success rate was 94% for these two classifications.

Table 2. Success rates of class determination according to the change of the k value.

k value of kNN	Class	False	True	Iris-virginica (%)	Iris-versicolor(%)	Iris-setosa (%)	Average Perfomance (%)
3	Iris-virginica	3	47	0.94	0.94	1.00	96.00
	Iris-versicolor	3	47				
	Iris-setosa	0	50				
5	Iris-virginica	3	48	0.96	0.94	1.00	96.67
	Iris-versicolor	2	47				
	Iris-setosa	0	50				
7	Iris-virginica	3	49	0.98	0.94	1.00	97.33
	Iris-versicolor	1	47				
	Iris-setosa	0	50				
9	Iris-virginica	2	49	0.98	0.96	1.00	98.00
	Iris-versicolor	1	48				
	Iris-setosa	0	50				
11	Iris-virginica	2	48	0.96	0.96	1.00	97.33
	Iris-versicolor	2	48				
	Iris-setosa	0	50				
13	Iris-virginica	2	49	0.98	0.96	1.00	98.00
	Iris-versicolor	1	48				
	Iris-setosa	0	50				
15	Iris-virginica	1	49	0.98	0.98	1.00	98.67
	Iris-versicolor	1	49				
	Iris-setosa	0	50				
17	Iris-virginica	2	49	0.98	0.96	1.00	98.00
	Iris-versicolor	1	48				
	Iris-setosa	0	50				
19	Iris-virginica	2	49	0.98	0.96	1.00	98.00
	Iris-versicolor	1	48				
	Iris-setosa	0	50				
21	Iris-virginica	2	49	0.98	0.96	1.00	98.00
	Iris-versicolor	1	48				
	Iris-setosa	0	50				
23	Iris-virginica	2	49	0.98	0.96	1.00	98.00
	Iris-versicolor	1	48				
	Iris-setosa	0	50				

In addition, the average success performance value is also shown for each of the neighboring values. Accordingly, the classification was made with 100% accuracy for Iris-setosa in all k values. Classification was made for Iris-virginica and Iris-versicolor with the highest 98% success rate. The lowest classification success rate was 94% for these two classifications.

In addition, the average success performance value is also shown for each of the neighboring values. Accordingly, the classification was made with 100% accuracy for Iris-setosa in all k values.

It was seen that the performance value remained constant after the neighboring value of k was chosen as 15. The success rates are calculated by taking these values into account since it is known that the sample data of the three Iris classes are equal and 50.

In this study, the graph of the average performance values obtained by changing the k neighbor values is shown in figure 5.

The average success rate, for all k neighbor values were highest 98.67%. However, the lowest average success rate was 96%.

The graph shows that k neighbor values vary between 3 and 23. The lowest success rate was 96%, when the neighbor value of k was chosen as 3. The highest success rate was 98.67%, when the neighbor value of k was chosen as 15 In addition, after reaching remained constant at 98% for k neighboring values above 15.

The purpose of this study; showing the effect of using optimum k neighboring value on performance while the highest success value it was observed that the success rate remained constant at 98% despite increasing k neighbor values.

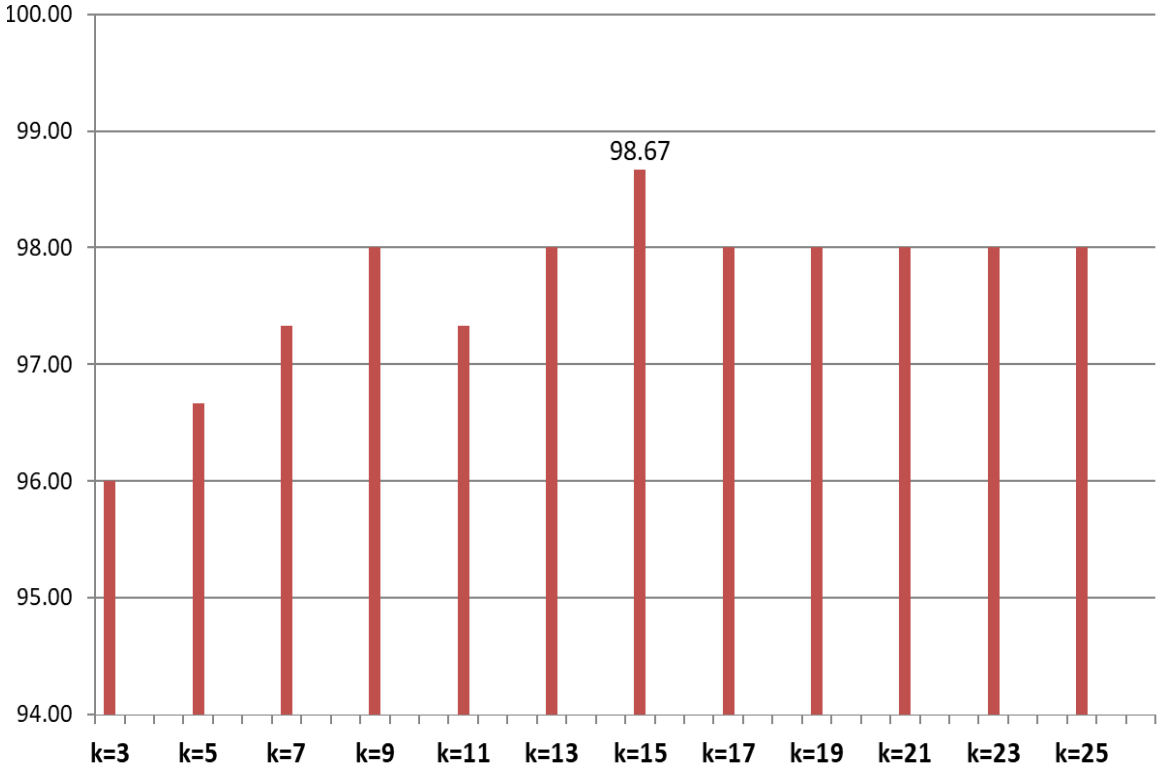

Figure 5. The highest performance class determination (k=15).

4. RESULTS

In this study, estimation process was carried out based on data mining. In machine learning, predictions can be made about new data by gaining experience from previous data. In this study, classification estimation has been made on the widely used Iris data set. This data set is classified according to the petal and sepal characteristics.

In this study, the effect of k neighbor value selected in kNN supervised learning algorithm on performance is shown. If the neighbor value of k is chosen correctly, it affects the class prediction performance of the kNN algorithm positively. However, the success rates can also be affected by the distance vectors used. In this study, the widely used Euclidean distance vector is used.

In this study, the neighbor values of k are chosen from consecutive odd numbers between 3 and 23. So it is wanted to prevent the situation of equality in the majority voting stage. In these tests, the lowest correct classification success rate was obtained when the k neighbor value was 3. The highest correct classification success rate was 98.67%, when the neighbor value of k was chosen as 15. Correct classification success rate making classification prediction with kNN machine learning algorithm. This study will form the basis for the applications that can choose the most suitable k value by making automatic performance comparison.

REFERENCES

[1] Lahmer, H., Oueslati, A. E., Lachiri, Z., "DNA Microarray Analysis Using Machine Learning to Recognize Cell Cycle Regulated Genes", 2019 International Conference on Control, Automation and Diagnosis (ICCAD); 2-4 July 2019; Grenoble, France. 2019; pp. 1-5.

[2] Sasikala, B. S., Biju V. G., Prashanth, C. M., "Kappa and accuracy evaluations of machine learning classifiers". 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT); 19-20 May 2017 ; Bangalore, India, 2017; pp. 20-23.

[3] Thirunavukkarasu, K., Singh, A. S., Rai P., Gupta, S., "Classification of IRIS Dataset using Classification Based KNN Algorithm in Supervised Learning", 2018 4th International Conference on Computing Communication and

Automation (ICCCA); 14-15 Dec. 2018; Greater Noida, India. 2018; pp 1-4.

[4] Goel, A., Mahajan S., "Comparison: KNN & SVM Algorithm". International Journal for Research in Applied Science & Engineering Technology (IJRASET) 2017; 5(13): 165-168.

[5] Yigit, H., "A weighting approach for KNN classifier", 2013 International Conference on Electronics, Computer and Computation (ICECCO); 7-9 Nov. 2013; Ankara, Türkiye. 2013; pp. 228-231.

[6] Kunju, M. V., Dainel, E., Anthony H. C., Bhelwa, S., "Evaluation of Phishing Techniques Based on Machine Learning", 2019 International Conference on Intelligent Computing and Control Systems (ICCS); 15-17 May 2019; Madurai, India. 2019; pp. 963-968.

[7] Agrawal, S., Bansal A., Rathor, S., "Prediction of SGEMM GPU Kernel Performance Using Supervised and Unsupervised Machine Learning Techniques", 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 10-12 July 2018; Bangalore, India. 2018; pp. 1-7.

[8] Alkhatib, K., Najadat, H., Hmeidi, I., Shatnawi, M. K., "Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm". International Journal of Business, Humanities and Technology 2013; 3(3): 32-44.

[9] Goel, M., Tiwari, A. K., Patil, H. S., "Recommendation Engine for B2B Customers in Telecom By Customizing KNN Algorithm". JournalNX- A Multidisciplinary Peer Reviewed Journal 2018; 4(4): 5-7.

[10] Bütüner, İ., Kaplan, B., Adem, K., "RSESLIBKNN Makine Öğrenmesi Yöntemi Kullanarak Parkinson Hastalığının Tanısı". Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi. 2020; 9(2): 715-721.

[11] Tosunoğlu, F., Hanay, S., Çintaş, E., Özyer, B., "Monthly Streamflow Forecasting Using Machine Learning". Erzincan University Journal of Science and Technology 2020; 13 (3) , 1242-1251.

[12] Çimen, F. M., Sönmez, Y., İlbaş, M., "Performance Analysis of Machine Learning Algorithms in Intrusion Detection Systems". Düzce Üniversitesi Bilim ve Teknoloji Dergisi 2021; 9(6): 251-258.

[13] Ozkan, I. A., Koklu, M., Saraçoğlu R., "Classification of Pistachio Species Using Improved K-NN Classifier". Progress in Nutrition 2021; 23(2):1-9.

[14] Özmen-akyol, S., Gülbandılar E., "İris Çiçeği Türünün YSA Yöntemleri ve ANFIS ile Tahmini", Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi 2020; 1(1) : 5-11.

[15] Du, S. and Li, J., "Parallel Processing of Improved kNN Text Classification Algorithm Based on Hadoop", 2019 7th International Conference on Information, Communication and Networks (ICICN); 24-26 April 2019; Macao, Macao, 2019; pp. 167-170.

[16] Zhao, Y., Qian Y. and Li, C., "Improved KNN text classification algorithm with MapReduce implementation", 2017 4th International Conference on Systems and Informatics (ICSAI); 11-13 Nov. 2017; Hangzhou, China. 2017; pp. 1417-1422.

[17] Silahtaroğlu, G., "Veri madenciliği (Kavram ve algoritmaları)". 3. Basım, İstanbul, Türkiye: Papatya Yayıncılık Eğitim; 2016. pp. 118-120.

[18] Balaban, M. E., Kartal, E., "Veri madenciliği ve makine öğrenmesi temel algoritmaları ve R Dili ile Uygulamalar". 2. Basım, İstanbul, Türkiye: Çağlayan Kitap & Yayıncılık & Eğitim; 2018; 48-72.

[19] Pedregosa, F., Varoquaux G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research 2011; 12: 2825-2830

[20] Dua, D, Graff, C., "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]". Irvine, CA: University of California, School of Information and Computer Science; 2019.