# Makine Öğrenmesi Algoritmaları ile Çocuklarda Erken Otizm Teşhisi
# Early Autism Diagnosis of Children with Machine Learning Algorithms

Fatiha Nur Büyükoflaz
Bilgisayar Mühendisliği Bölümü
KTO Karatay Üniversitesi
Konya, Türkiye
fatiha.nur.buyukoflaz@ogrenci.karatay.edu.tr

Ali Öztürk
Bilgisayar Mühendisliği Bölümü
KTO Karatay Üniversitesi, Konya, Türkiye
Havelsan A.Ş., Ankara, Türkiye
ali.ozturk@karatay.edu.tr

***Özetçe* —Günümüzde Otistik Spektrum Bozukluğu (OSB) önemli sağlık problemleri arasına girmiş bir nöro-gelişimsel bozukluktur ve erken teşhisi hastalığın kontrol edilmesi açısından büyük öneme sahiptir. Otizmin ekonomik etkisi ve dünyadaki OSB vakalarının sayısındaki artış, kolayca uygulanan etkili tarama yöntemlerinin geliştirilmesine acil ihtiyaç olduğunu ortaya koymaktadır. Bu çalışmada UCI 2017 Autistic Spectrum Disorder Screening Data for Children veri kümesi üzerinde, Naive Bayes, IBk(k- en yakın komşu), Radyal Temel Fonksiyon Ağı (RBFN) ve Rasgele Orman(RO) olmak üzere dört farklı sınıflandırma yöntemi kullanılarak performans karşılaştırılmaları yapılmıştır. Deney sonucunda Rasgele Orman yönteminin Naive Bayes, IBk ve RBFN yöntemlerinden daha başarılı olduğu gösterilmiştir.**

***Anahtar Kelimeler—otistik spektrum bozukluğu; makine ögrenmesi; naive bayes sınıflandırıcı; IBk sınıflandırıcı; rasgele orman sınıflandırıcı; ABFN sınıflandırıcı;***

***Abstract*—Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that has become one of the major health problems, and early diagnosis has a great deal of important in terms of controlling the disease.The increase in the number of autoimmune influenza and ASD cases in the world reveals an urgent need to develop easily applied and effective screening methods In this study, performance comparisons were made using three different classification methods, Naive Bayes, IBk (k-nearest neighbors), RBFN (radial basis function network), and Random Forest, on UCI 2017 Autistic Spectrum Disorder Screening Data for Children dataset. As a result of the experiment, Random Forest method has been shown to be more successful than Naive Bayes, IBk and RBFN methods.**

***Keywords—autistic spectrum disorder; machine learning; Naive Bayes classifier; IBk classifier; Random Forest classifier; RBFN classifier;.***

## I. GİRİŞ

Otizm spektrum bozukluğu(OSB) nedeni tam olarak bilinmemekle birlikte genetik ve çevresel faktörlerin etkili olduğu düşünülen nöro-gelişimsel bir bozukluktur. Otistik spektrum bozukluğu olan bireyler toplumsal etkileşimde ve iletişimde yetersizlikler ile davranış, ilgi ve etkinliklerde sınırlı, basmakalıp ve yineleyici örüntü özelliklerini içerir. Toplumsal iletişimde kullanılan dil ya da sembolik/imgesel oyun becerilerinin en az birinde üç yaşından önce gecikmelerin ya da olağandışı bir işlevselliğin ortaya çıktığı görülür [1].

Otizm, günümüzde zihinsel yetersizlikten sonra en sık rastlanan nörolojik bozukluktur ve Hastalıkları Kontrol Etme ve Önleme Merkezi (Centers for Disease Control Prevention)'nin verilerine göre 2006 yılında Her 150 çocuktan 1'inde otizm görülürken, 2012 yılında Her 88 çocuktan 1'inde otizm görülmüştür. 2014 yılında verilen son bilgiye göre ise her 68 çocuktan 1'inde otizm görülmektedir. [1]

Dünya genelinde OSB vakalarının sayısındaki hızlı artış, davranışsal özelliklere ilişkin veri setleri gerektirir. Bu çalışmada otizm taramasının ileri analizinde, özellikle etkili otistik özellikleri tanımlamak ve OSB vakalarının sınıflandırılmasını iyileştirmek için 20 özellikli bir veri seti kullanılmıştır. Kullanılan bu veri seti üzerinde, Naive Bayes, IBk, RBFN ve Rasgele Orman yöntemleri ile hastalık tahmin oranları hesaplanmıştır ve bu yöntemlerin performansları karşılaştırılmıştır. Rasgele Orman sınıflandırıcısıyla elde edilen sonuçların diğerleriyle karşılaştırıldığında daha etkili sonuçlar verdiği görülmüştür.

Literatürde çocuklarda otizmle ilgili yakın zamanda yapılmış değişik çalışmalar bulunmaktadır. Otizmli çocukların duygusal çöküş anlarını tahmin etmek üzere derin öğrenme tekniklerinin kullanıldığı çalışmada [2] ebeveynlerin ya da bakıcıların kısa sürede duruma müdahale edebileceği bir sistem önerilmiştir. OSB'li çocuklar için mobil tabanlı bir öğrenme uygulamasının geliştirildiği çalışmada [3] matematik konuları, iletişim, okuma ve yazma becerilerinin geliştirilmesi ilgi çekici bir kullanıcı arayüzü ile sağlanmıştır.

Ayrıca çocukların kaybolma ihtimaline karşı konum izleme özelliği de uygulamada yer almaktadır. Akıllı telefonlardaki hareket algılayıcı veriler, kamera ve ses kayıtları ile bir erken uyarı sisteminin geliştirildiği çalışmada [4] ise otizmli

978-1-5386-1501-0/18/$31.00 © 

[1] https://www.montessorietkinlikler.com/otizm-spektrum-bozuklugu/

çocukların davranışlarının ve bulundukları ortamın tahmini için Weka kütüphanesindeki J.48 sınıflandırıcı kullanılmıştır. Destek Vektör Makinelerinin (DVM) OSB'li hastaların sınıflandırması [5] için kullanıldığı çalışmada rasgele DVM kümeleme yönteminin çok iyi sonuç verdiği görülmüştür. Bu çalışmada 46 sağlıklı ve 61 OSB rahatsızlığı bulunan insana ait veriler Otizm Beyin Görüntüleme Veri Değişimi veritabanından elde edilmiştir.

## II. MATERYAL VE YÖNTEMLER

### A. Veri Kümesi

Bu çalışmada kullanılan veri seti kümesi 2017 Tıp ve Sağlık Bilişimi Uluslararası Konferansında sunulan [6] UCI 2017 Autistic Spectrum Disorder Screening Data for Children veri kümesidir [7]. Bu veri kümesi üzerinde makine öğrenmesi yöntemlerinin uygulandığı bir çalışma henüz bulunmamaktadır [8]. Veri kümesinde 4-11 yaş arasındaki 290 çocuğa ait 20 farklı özellik bulunmaktadır. Bu özelliklerden on tanesi davranış bilimlerindeki kontrollerden OSB vakalarını tespit etmekte etkili olduğu ispatlanmış davranışsal özellikler (AQ-10-Çocuk) diğer on tanesi ise çocukların kişisel özellikleridir. Bu veriler Weka programı kullanılarak 244 tanesi eğitim verisi 58 tanesi ise test verisi olarak iki ayrı gruba ayrılmıştır [9].

### B. IBk

IBk sınıflandırıcısı, minimum uzaklık tabanlı K en yakın komşu (KNN) algoritması olarak da adlandırılır. K-en yakın komşu algoritması örnek tabanlı bir öğrenme yöntemidir. KNN, her örneği n-boyutlu bir uzayda bir noktayla ilişkilendirir ve bir dizi nitelik olarak, yani $< a_1(x), a_2(x), ..., a_n(x) >$ (n niteliklerin sayısı) olarak tanımlar. $X_i$ ve $x_j$ örneği arasındaki mesafe, formül (1) ile hesaplanır [10]-[11].

$$d(x_i, x_j) \equiv \sqrt{(\sum_{r-1}^{n} (a_r(x_i) - a_r(x_j))^2)} \quad (1)$$

Yeni bir $x$ örneğini sınıflandırmayı amaçlayan KNN, eğitim veritabanında en yakın örnekleri $x$'i örneklemek için seçer ve x örneğinin sınıfını belirlemek için k örneklerini kullanır. KNN'nin çalışma prosedürü aşağıdaki gibidir.

- Eğitim süresi: her bir eğitim örneğini $< x, f(x) >$, eğitim veritabanına ekler.
- Sınıflandırma periyodu: yeni bir örneği sınıflandırmak için, eğitim veritabanında k en yakın örnekler seçerilir. Sonra formül (2) 'de gösterildiği gibi $y$ sınıflandırmanın sonucunu döndürür [11].

$$f(y) \leftarrow argmax_{v \epsilon V} \sum_{i=1}^{k} \delta(v, f(x_i)) \quad (2)$$

### C. Naive Bayes

Naive Bayes (NB), Bayes teoreminin istatistiksel teorisine dayanan bir sınıflandırmadır. Sınıfın özelliklerine göre varsayımlanan Bayes Kuralı'na dayanan bu algoritma koşullu olarak bağımsızdır [12]. Literatürde, NB algoritması, metin sınıflandırması, arama motoru kalitesinin geliştirilmesi, görüntü işleme, arıza tahmini ve tıbbi teşhisler gibi çeşitli uygulamalarda uygulanmıştır.

NB sınıflandırıcısı aşağıdaki gibi çalışır: eğitim seti $X$=$[x_1, x_2, ..., x_n]$ eğitim eğitim kümesindeki belirli sınıf etiketine (C) aittir. Her bir sınıfın olasılığı, tahmini nitelikler için verilen değerler göz önüne alındığında formül (3) kullanılarak bulunabilir:

$$P(Y_j \mid X) = \frac{P(Y_j) \cdot P(X_j \mid Y)}{\sum_i^c P(Y_i) \cdot P(X_i \mid Y)}, j = 1, 2, ..., c \quad (3)$$

Burada $P(X_i)$, $Y_i$ sınıfının öncelikli olasılığı ve $P(Y_j \mid X)$ sınıf koşullu olasılığın yoğunluk fonksiyonudur. Koşullu bağımsızlık, veri kümesindeki özelliklerin koşullu olarak birbirinden bağımsız olduğunu varsaymaktadır. Basit bir şekilde test durumlarını hesaplamak ve eğitim verilerini tahmin etmek aşağıdaki formül (4) kullanılarak gerçekleştirilebilir.

$$P(X \mid Y_j) = \prod_{i-1}^{n} P(X_i \mid Y_j), j = 1, 2, ..., c \quad (4)$$

Burada $X_i$, $i$'deki $i$'inci özniteliğin değeri ve $n$ ise özniteliklerin sayısıdır. Sınıfların sayısı $k$, $C_i$ $i$'inci sınıf olduğu varsayılırsa, özellikler kümesindeki olasılık dağılımı aşağıdaki formül (5) kullanılarak hesaplanır [13].

$$P(X) = \prod_{i-1}^{k} P(C_i) \cdot P(X \mid C_i) \quad (5)$$

### D. Rasgele Orman

Rastgele Orman algoritması, sınıflandırma algoritmaları arasında en iyi olanlardan biridir , büyük miktarda veriyi doğru bir şekilde sınıflandırabilir. Rasgele Orman algoritması , karar ağacı sınıflandırıcıları $\{h(x, \Theta k)$, $k$=$1, 2, 3, ..., K$ kümesinden oluşan entegre bir sınıflandırmadır. Burada $x$ bir giriş vektörü , $\Theta k$ tek bir ağacın büyüme sürecini belirleyen bağımsız olarak eşit dağılıma sahip rasgele bir vektördür. $K$ ise rastgele bir ormandaki karar ağaçlarının sayısını temsil eder.

Sürekli öznitelik A ve bir düğüm üzerindeki örnek kümelerin sayıs m ise A, üzerindeki rasgele orman algoritması şu şekilde işlem görür.

- Düğüm üzerindeki örnekler sürekli öznitelik A'nın somut değerine göre küçükten büyüğe sıralanır ve öznitelik değerinin $\{A_1, A_2, ....A_m\}$ dizisi elde edilir.
- Değer dizisinde $m - 1$ bölme noktası üretilir. $J(0 < j < m)$ bölme noktasının değeri, düğümdeki örnek kümesini $\{s \mid s \epsilon S, A(S) W_j\}$ ve $\{s \mid s \epsilon S, A(S) > W_j\}$ alt kümelerine bölen $W_1 - (A_j + A_{j+1})/2$ formülü ile ayarlanır.
- $m - 1$ bölünmüş noktaların Gini katsayıları folmül (6)'de gösterildiği gibi hesaplanır ve örnek kümeyi bölmek için en küçük Gini katsayısına sahip olan noktalar seçilir.

$$Gini(S) = 1 - \sum_{n=1}^{n} {p_i}^2 \quad (6)$$

Burada S bir veri kümesi ve $\mid S \mid$ örneklerin sayısıdır. Tüm örneklerde $n$ farklı öznitelik $C_i$ bulunur. S veri kümesinde $C_i$ sınıfına ait örnek sayısı $\mid C_i \mid$'dir. $P_i$ olasılığı ise $\frac{|C_i|}{|S|}$'dir [14].

### E. Radyal Temel Fonksiyon Ağı

Radyal temel fonksiyon ağı (RBFN) lineer çıkış ünitelerine tam olarak bağlı tek bir gizli birim katmanı olan ileri bir besleme ağıdır. Çıkış birimleri, gizli katman düğümleri tarafından hesaplanan temel işlevlerin doğrusal bir kombinasyonunu oluşturur [15]. Gizli katmanın aktivasyon fonksiyonu olarak genellikle Gauss fonksiyonu tercih edilmektedir. Gauss fonksiyonu formül (7) kullanılarak bulunabilir.

$$\psi = \exp\left[\frac{-\parallel x - c_i \parallel^2}{2 \cdot {\sigma_i}^2}\right], i = 1, 2, ..., N \quad (7)$$

Burada $x$ giriş vektörü, $C_i$ $x$'inci temel işlevin merkezidir. $Y$ merkezin çevresindeki işlevin genişliğini belirleyen $x$'inci temel fonksiyonunun ölçek faktörü, $N$ gizli katman düğümlerinin sayısı ve $\parallel . \parallel$ kıtasal normdur. RBF ağının çıkışı ise formül (8) kullanılarak hesaplanır [16].

$$y_k = \sum_{i=1}^{N} \omega_{ik} \cdot \phi_i(x), i = 1 \ldots N, k = 1 \ldots M \quad (8)$$

Burada $M$ çıkış düğümlerinin sayısı, $w_{ik}$ gizli katmandaki RBF ağının çıktı ağırlıklarıdır.

## III. KULLANILAN METRİKLER VE DENEYSEL SONUÇLAR

Bu çalışmada belirlenen veri seti %80 eğitim verisi %20 test verisi olmak üzere iki kısma ayrılmıştır. Veri seti üzerinde Weka yazılım paketi içerisindeki bulunan Rasgele Orman, Naive Bayes, RBFN ve IBk sınıflandırıcılarının performansı ve hastalık tahmini hesaplanmıştır.

### A. Kullanılan Metrikler

Performans analizi yapılırken confisuion matris precision, Recall, F-measure ve Accuracy metrikleri kullanılmıştır. Kullanılan metrikler aşağıda tanımlanmıştır. Metriklerde bulunan parametreler ise Tablo I'de gösterilmiştir.

Tablo I: KARIŞIKLIK MATRİSİ

| | Beklenen Pozitif | Beklenen Negatif |
|---|---|---|
| Gerçek Pozitif | TP | FN |
| Gerçek Negatif | FP | TN |

- TP (True Pozitif): Doğru şekilde tanımlanan pozitif vakaların oranıdır.
- FP (False Negatif): Yanlışlıkla pozitif olarak sınıflandırılan negatif vakaların oranıdır.
- TN (True Negatif): Doğru sınıflandırılmış negatif vakaların oranıdır.
- FN (False Negatif): Yanlışlıkla negatif olarak sınıflandırılan pozitif vakaların oranıdır.

Bir sistemin değerlendirmesini açıklamak için en basit yol, kesinlik (precision) ve duyarlılık (recall) olarak bilinen iki metriği kullanmaktır.

$$precision = \frac{TP}{TP + FP} \quad (9)$$

$$recall = \frac{TP}{TP + FN} \quad (10)$$

F ölçüsü (f-measure) ise kesinlik ve hassasiyet parametlerinin harmonik ortalamasıdır.

$$f - measure = 2 \cdot \frac{(precision * recall)}{(precision + recall)} \quad (11)$$

Doğruluk (accuracy) sınıflandırıcının doğruluğunu ve yeteneğini ifade eder. Ayrıca yeni bir veri için öngörülen öznitelik değerinin ne kadar iyi tahmin edilebileceğini belirtir. Accuracy doğru olarak sınıflandırılan örneklerin toplam örneklere oranıdır.

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (12)$$

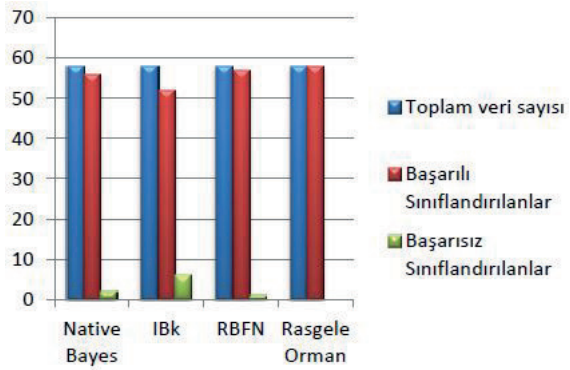
### B. Deneysel Sonuçlar

Tanımlanan metrikler kullanılarak elde edilen üç farklı sınıflandırıcının veri seti değerleri Tablo II' de gösterilmiştir.

Tablo II: VERİ SETİ SONUÇLARI

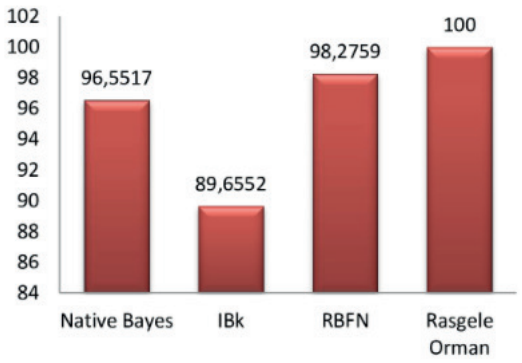
| | Rastgele Orman | IBk | RBFN | Naive Bayes |
|---|---|---|---|---|
| TP Oran | 1,000 | 0,897 | 0,983 | 0,966 |
| FP Oran | 0,000 | 0.103 | 0,017 | 0,034 |
| Kesinlilik | 1,000 | 0,898 | 0,983 | 0,966 |
| Duyarlılık | 1,000 | 0,897 | 0,983 | 0,966 |
| F-ölçüsü | 1,000 | 0,896 | 0,983 | 0,966 |

Yapılan çalışmada kullanılan sınıflandırıcıların performans başarı grafiği Şekil 1'de verilmiştir.




Şekil 1: Sınıflandırıcıların Başarı Grafiği

Şekil 2'de ise çalışmada kullanılan sınıflandırıcıların başarı yüzdeleri verilmiştir.




Şekil 2: Sınıflandırıcıların Başarı Yüzdeleri

Elde edilen sonuçlara göre 58 adet veri üzerinde sırasıyla Naive Bayes 56 adet, IBk 52 adet, RBFN 57 adet, Rasgele Orman 58 adet veriyi doğru şekilde sınıflandırmayı başarmıştır. Kullanılan veri seti üzerinde en düşük doğruluk oranı %89,65 ile IBk sınıfındırıcısına ait iken en yüksek doğruluk oranı ise %100 ile Rasgele Orman sınıflandırıcısına aittir.

## IV. SONUÇ

Bu çalışmada makine öğrenme algoritmalarının UCI 2017 Autistic Spectrum Disorder Screening Data for Children veri seti üzerinde gösterdikleri hastalık tahminine dayalı başarı oranları analiz edilmiştir. Kullanılan veri seti 244 tanesi eğitim verisi 58 tanesi test verisi olacak şekilde iki kısma ayrılmıştır. Test verileri üzerinde IBk, Naive Bayes, RBFN ve Rasgele Orman sınıflandırıcıları kullanılmıştır. Kullanılan sınıflandırıcıların performans analizinden elde edilen sonuçlara göre veri seti üzerinde IBk %89.65, Naive Bayes %96.55, RBFN %98.27, Rasgele Orman ise %100 başarı oranı göstermiştir.

## KAYNAKÇA


[1] Diken, İ. H., Otistik bozukluğu olan öğrenciler. İ. H. Diken, (Ed.), Özel Eğitime Gereksinimi Olan Öğrenciler Ve Özel Eğitim İçinde (411-444). Ankara: Pegem Akademi, 2010.

[2] Sindhoor P.V., Feba T.G., Kiran G. and Abhishek V., "Deep Learning Based Recognition of Meltdown in Autistic Kids", IEEE International Conference on Healthcare Informatics, 391-396, 2017.

[3] Raafat A., Fadi A., Anam M., Kamil K. and Suad A., "AutiAid: A Learning Mobile Application for Autistic Children", IEEE 19th International Conference on e-Health Networking, Applications and Services, 1-6, 2017.

[4] Chuah M. and Diblasio M., "Smartphone Based Autism Social Alert System", 8th International Conference on Mobile Ad-hoc and Sensor Networks, 6-13, 2012.

[5] Bi X.A., Wang Y., Shu Q., Sun Q and Xu Q., "Classification of Autism Spectrum Disorder Using Random Support Vector Machine Cluster", Frontiers in Genetics, 6:9-18, 2018.

[6] Thabtah, F., "Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment", Proceedings of the 1st International Conference on Medical and Health Informatics, Taichung City, Taiwan, ACM, pp.1-6, 2017.

[7] Lichman M., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2013.

[8] Thabtah, F., "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward", Informatics for Health and Social Care, 13:1-20, 2018.

[9] Eibe Frank, Mark A. Hall, and Ian H. Witten. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[10] Aha D W, Kibler D, and Albert M K, "Instance-based learning algorithms", Machine Learning, 6: 37- 66, 1991.

[11] B. Sun, J. Du, and T. Gao, "Study on the improvement of K-nearest-neighbor algorithm," Int. Conf. Artif. Intell. Comput. Intell. AICI, vol. 4, pp. 390–393, 2009.

[12] Webb, G., Naive Bayes . In Encyclopedia of MachineLearning (pp. 713-714). Springer. 2010.

[13] D. Mittal, M. Bala, "Hybrıd Feature Selectıon Aapproach Usıng Bacterıal Foragıng Algorıthm Guıded by Naıve Bayes Classıfıcatıon" IEEE – 40222, 2017.

[14] Y. Xu, "Research and Implementation of Improved Random Forest Algorithm Based on Spark," pp. 499–503, 2017.

[15] C. Panchapakesan, M. Palaniswami, D. Ralph, and C. Manzie, "Effects of moving the centers in an RBF network," IEEE Trans. Neural Networks, vol. 13, no. 6, pp. 1299–1307, 2002.

[16] S. Zhou and H. Lin, "Function approximation based on self-adaptive RBF neural network with combined clustering algorithm," Intell. Control Inf. Process. (ICICIP), 2010 Int. Conf., no. 1, pp. 435–438, 2010.