Yayın Geliş Tarihi (Submitted): **17/06/2022**

Yayın Kabul Tarihi (Accepted): **16/12/2022**

Makele Türü (Paper Type): **Araştırma Makalesi – Research Paper**

Please Cite As/Atıf için:
Gazeloğlu, C. (2022), Otizm spektrum bozukluğunda bulanık kaba küme özellik seçimi kullanılarak lojistik regresyon ile sınıflandırılması, *Nicel Bilimler Dergisi*, 4(2), 176-189. *doi:10.51541/nicel.1132140*

OTİZM SPEKTRUM BOZUKLUĞUNDA BULANIK KABA KÜME ÖZELLİK SEÇİMİ KULLANILARAK LOJİSTİK REGRESYON İLE SINIFLANDIRILMASI

Cengiz Gazeloğlu[1]

ÖZ

Otizm Spektrum Bozukluğu (OSB), doğuştan gelen ve genel olarak sosyal ilişkilerde ve iletişim kurmada sıkıntı yaşama durumudur. Bu durum aslında bazı uzmanlar tarafından nöro gelişimsel bir bozukluk veya psikolojik durum spektrumu olarak da tanımlanabiliyor. Her hastalıkta olduğu gibi bu rahatsızlıkta da erken tanı çok önem arz etmektedir. Bu çalışmanın temel amaçlarından biri, OSB rahatsızlığını, lojistik regresyon algoritmasını kullanarak bireylerde bu bozukluğun olup olmadığını doğruluk oranı yüksek bir şekilde sınıflandıra bilmektir. Diğer amaç ise öne sürülen sınıflandırma modeli ile alanda çalışan doktorlara hata yapmamaları anlamında hem yardımcı olmak hem de teşhis yöntemini daha hızlı hale getirerek zamandan ve maliyetten tasarruf etmektir. Çalışma verileri WEKA programı yardımı ile analiz edilmiştir. Sınıflandırma algoritması olarak lojistik regresyon algoritması kullanılmıştır. Algoritmanın daha hızlı ve doğru çalışması adına bulanık kaba küme yöntemi ile özellik seçimi yapılmıştır. Algoritmanın veri ezberleme durumu ortadan kaldırmak adına 10 döngülü çapraz doğrulama yapılmıştır. Sonuçların değerlendirilmesi için TP ve FP oranları hesaplanmıştır. Hesaplanan sonuçlara göre TP oranı özellik seçimi yapılmadan önce 0,947 iken özellik seçimi yapıldıktan sonra 0,974 olarak hesaplanmıştır. Benzer şekilde FP oranları ise sırasıyla 0,043 ve 0,028 olarak tespit edilmiştir. Bu sonuçlara göre algoritmanın OSB'yi sınıflandırmada başarılı olduğu söylenebilir. Ek olarak özellik seçimi yapılmadan önce ve sonraki sonuçları karşılaştırmak için ROC analizi yapılmıştır. Analiz sonucuna göre ROC eğrisinin altında kalan alanın 0,99 olarak

[1]Sorumlu yazar, Doç. Dr., İstatistik Bölümü, Fen Edebiyat Fakültesi, Süleyman Demirel Üniversitesi, Isparta, Türkiye. ORCID ID: https://orcid.org/0000-0002-8222-3384

hesaplanmış olması özellik seçimi yapılmasının doğru bir karar olduğunun göstergesidir. Ayrıca özellik seçimi yapıldıktan sonra doğru sınıflandırma oranı %95,205'ten %96,575'e çıkmıştır.

Anahtar Kelimeler: Bulanık Kaba Küme, Lojistik Regresyon, Otistik Spektrum Bozukluğu, Özellik Seçimi

CLASSIFICATION OF AUTISM SPECTRUM DISORDER BY LOGISTIC REGRESSION USING FUZZY ROUGH SET FEATURE SELECTION

ABSTRACT

Autism Spectrum Disorder (ASD) is a congenital and general problem in social relations and communication. This condition can actually be defined by some experts as a neurodevelopmental disorder or a spectrum of psychological states. As in any disease, early diagnosis is very important in this disease. One of the main purposes of this study is to classify ASD with a high accuracy rate, using the logistic regression algorithm. The other purpose is to help doctors working in the field not to make mistakes with the proposed classification model, and to save time and cost by making the diagnosis method faster. Study data were analyzed with the help of WEKA program. Logistic regression algorithm was used as classification algorithm. In order for the algorithm to work faster and more accurately, feature selection was made with the fuzzy coarse set method. In order to eliminate the data memorization situation of the algorithm, 10-cycle cross validation was performed. TP and FP ratios were calculated to evaluate the results. According to the calculated results, while the TP ratio was 0.947 before feature selection, it was calculated as 0.974 after feature selection. Similarly, FP rates were determined as 0.043 and 0.028, respectively. According to these results, it can be said that the algorithm is successful in classifying ASD. In addition, ROC analysis was performed to compare the results before and after feature selection. The fact that the area under the ROC curve was calculated as 0.99 according to the analysis result indicates that the feature selection is the right decision. In addition, after the feature selection was made, the correct classification rate increased from 95.205% to 96.575%.

Keyword: Fuzzy Rough Set, Logistic Regression, Autistic Spectrum Disorder, Attribute Selection

1. GİRİŞ

Otizm, nerobiyolojik araştırmalarda beynin yaşam boyu statik gelişimsel bozukluğu olarak tanımlanmaktadır (Rapin ve Katzman, 1998).

İlk olarak Kanner (1943) tarafından tıp yazınına kazandırılan otizm; kısıtlanmış, yinelenen davranış örüntüleri, toplumsallaşmada sözlü ve sözel olmayan iletişimde bozukluk gibi çekirdek belirtileri olan süregelen bir bozukluktur. Kanner'in, otizmi tanımlamasının ardından, bugüne dek yapılan biyolojik, psikolojik ve klinik araştırmalar sonucunda, hastalığa bakış açısı epey değişikliğe uğramıştır. Önceleri otizmin, anne ve babanın tutumu, sevgi yoksunluğu ya da sosyal ilişki kurma konusunda duyulan korkudan kaynaklandığı sanılmaktaydı. Son 20 yıldır otizmin, çocuğun yetiştirilme biçimi ya da geçmiş yaşantısıyla ilintili olmadığı, nörobiyolojik bir etiyolojiye sahip olduğu görüşü ağırlık kazanmıştır (Bodur ve Soysal, 2004).

Günümüzde birçok tanı sistemi, otizm tanısını koymaya yönelik olarak kullanılmaktadır. Bu sistemlerin ortak özelliği, otizm tanısı koymak için üç yeti alanında eksiklik olması gerektiğini vurgulamalarıdır. Bu alanlar; (Bodur ve Soysal, 2004).

1. İletişim ve toplumsal gelişim alanlarında bozukluğun olması.
2. Yineleyici, sınırlayıcı ilgi ve davranışlar.
3. Bu alanlardaki bozuklukların 30 ay öncesinden görülmesi

Otizm Spektrum Bozukluğu (OSB) ise erken dönemde ortaya çıkan, sosyal etkileşim ve sosyal iletişimde bozukluk, sosyal etkileşim ve toplumsal ilişki geliştirmede sorunlar, basmakalıp ve yineleyici davranışlar ile ilgi alanlarında sınırlılık olarak karakterize olan ve bu sınırlılıkların zihinsel yetersizlik veya gelişimsel gerilik ile açıklanamadığı bir bozukluktur (Amerikan Psikiyatri Birliği, 2013). Bu bozukluk psikiyatrik bozuklukların başında gelen rahatsızlıklardan biridir (Thabtah, 2017). Bu rahatsızlıklarla ilgili çarpıcı rakamlar ilgili literatür taraması yapıldığı zaman göze çarpmaktadır. Bu rakamlarla ilgili olarak;

Hastalıkları Kontrol Etme ve Önleme Merkezi'nin verilerine göre, 2006 yılında her 150 çocuktan 1'inde ve 2012 yılında her 88 çocuktan 1'inde OSB görülürken, 2014 yılında her 68 çocuktan 1'inde görülmektedir. Ayrıca erkeklerde kızlardan 3-4 kat daha fazla görüldüğü bilinmektedir. Fakat erkeklere oranla kızlarda daha ağır seyrettiği ve zekâ geriliğinin daha fazla eşlik ettiği de bilinmektedir (Halk Sağlığı, 2019).

OSB'nin ülkeler bazında yaygınlığı hakkında net olarak bir bilgi bulunmamaktadır. Ancak OSB'nin tüm ırklarda ve toplumlarda görüldüğü bilinmektedir. Tam olarak sebebi ve bulunduğu coğrafyanın bilinmemesi bilim insanlarının bu konu ile alakalı çalışmalara yönlenmesine sebep olmuştur. Çalışmaların bir kısmı hastalığın tıbbi yöntemler ile tanısı ve tedavisi üzerine yapılırken bir kısmı ise bilgisayarlı ortamlarda makine öğrenmesi yöntemleri kullanılarak bu hastalığı sınıflandırıp tıbbi alanda çalışan uzmanlara yardımcı olması yönünde ilerlemektedir. Ayrıca OSB'nun son zamanlarda görülme sıklığının giderek artan bir rahatsızlık haline gelmesi birçok ülkede bu konuya yönelik eylem planları geliştirilmesine de neden olmaktadır. Bu eylem planlarının birinci basamağı erken tanıdır. Her rahatsızlıkta olduğu gibi erken tanı OSB'de de büyük önem arz etmektedir. Bu çalışmanın amaçları arasında yer alan erken tanı ve doğru teşhis konusunda bu alandaki uzman kişilere yardım etme anlamında büyük bir katkı sağlayacaktır.

Tablo 1'de OSB rahatsızlığı ile ilgili bazı makine öğrenmesi algoritmaları ile ilgili çalışmalar yer almaktadır. Tabloda çalışmalarda kullanılan algoritmaların hangi bilgisayar programlarında yapıldığı ve algoritmaların doğru sınıflandırma oranları gibi bilgilere yer verilmiştir.

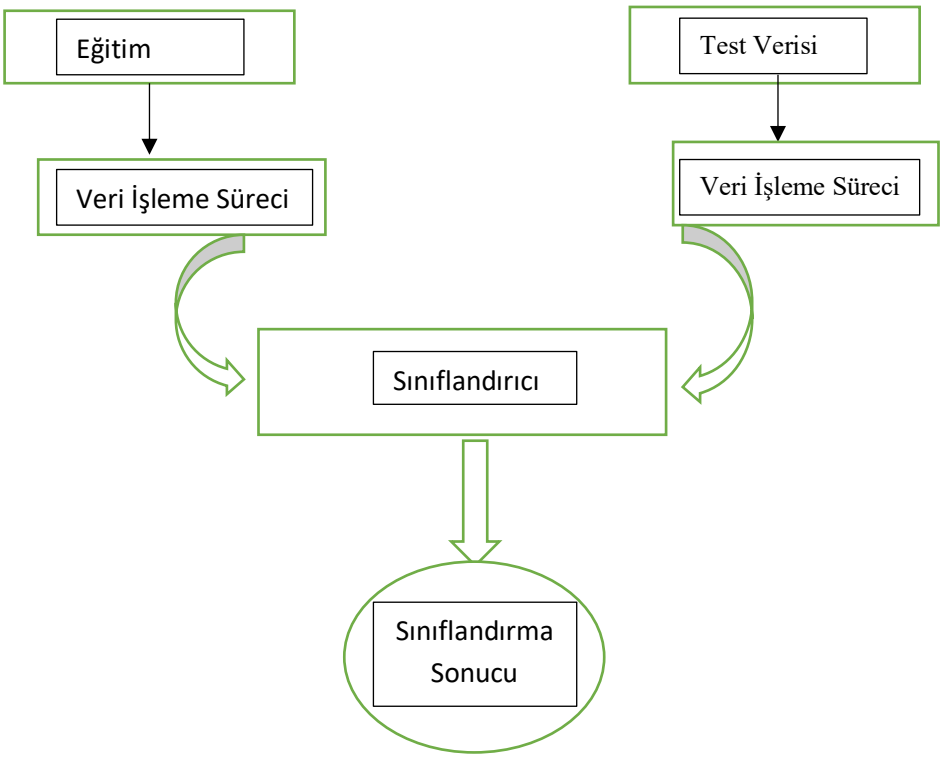
Tablo 1. OSB rahatsızlığının makine öğrenmesi algoritmaları ile sınıflandırılmasına yönelik çalışmalar (thabtah, 2019)

Yıl	**Özellik Seçimi**	**Kullanılan Makine Öğrenmesi Algoritmaları**	**Kullanılan Bilgisayar Programı**	**Veri Özellikleri**	**En İyi Algoritma**	**Doğruluk Oranı %**
2012	Yok	SVM, LG, Tree, Olasılık ve çeşitleri)	Weka	29 değişken, 612 kişi OSB, 15 kişi OSB değil	ADTree	%99,70
2012	Yok	SVM, LG, Tree, Olasılık ve çeşitleri)	Weka	93 değişken, kişi OSB 891, 75 kişi OSB değil	ADTree	100%
2014	Yok	Naive Bayes, SOM, Neural Fuzzy, LVQ Nueral Network, Kmeans, Fuzzy C Mean	Developed	16 değişken ve 100 kişi	SOM ve Naïve Bayes	100%
2015	Var	SVM, LR, DT, Olasılık ve çeşitleri	R, Weka	28 değişken 3885 kişi OSB, 655 kişi OSB değil	SVM LR	%98,27 %97,66
2016	Var	SVM, LR, DT	Scikitlearn	65 değişken, 2775 kişi OSB, 150 kişi OSB değil	SVM	YOK
2016	Var	SVM	LibSVM	65 değişken, 1264 kişi OSB, 462 kişi OSB değil	SVM	YOK

2. YÖNTEM

2.1. Sınıflandırma

Sınıflandırma kavramı temel olarak şu anlama gelmektedir. Bir veri setini bazı kurallara göre daha önceden tanımlanmış gruplara ayırmasıdır. Literatürde birçok sınıflandırma yöntemi bulunmaktadır. Burada önemli olan veri yapısına uygun doğru sınıflandırma yöntemini tayin etmek ve bu tayin edilen yöntem veya yöntemlerin doğru sınıflandırma oranının yüksek olmasıdır.



Şekil 1. Makine öğrenmesi akış şeması

Şekil 1'de bir sınıflandırma algoritmasına bağlı olarak makine öğrenmesi algoritmasının nasıl işlediğine dair temel bir akış diyagramı yer almaktadır. Akış şemasına göre veri seti iki ayrı grup olarak belirlenmektedir. Birinci grup ile eğitim yapılmakta iken ikinci grup da algoritmanın başarı oranını belirlemek için test işlemi yapılmaktadır. Ayrıca veri setleri için gerekli ise bir ön işleme tabi tutularak düzenleme yapılmaktadır. Daha sonra sınıflandırma algoritması belirlenerek algoritma çalıştırılır ve sonuçlar elde edilir.

2.2. Lojistik Regresyon, ROC, TP ve FP Analizi

Regresyon yöntemleri, herhangi bir veri setinde bir bağımlı değişken ve bir ya da daha fazla bağımsız değişkenler arasındaki ilişkiyi bulmaya çalışan bir analiz yöntemleridir (Hosmer ve Lemeshow, 2000). Lojistik regresyon ise gözlem değerlerini belirli bir kurala göre bir sınıfa atayan regresyon yöntemlerinden bir tanesidir.

Lojistik regresyon analizi, sınıflama ve atama işlemi yapmaya yardımcı olan bir regresyon yöntemidir. Normal dağılım varsayımı, süreklilik varsayımı önkoşulu yoktur. Bağımlı değişken üzerinde açıklayıcı değişkenlerin etkileri olasılık olarak elde edilerek, risk faktörlerinin olasılık olarak belirlenmesi sağlar (Özdamar, 2002).

$$P = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}{1+ e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}} \quad (1)$$

P: Olayın gözlenme olasılık değeri

β_0 : Bağımsız değişkenler sıfır değerini aldığı zaman bağımlı değişkenin değerini başka bir ifade ile sabiti

$\beta_1\beta_2 \ldots \beta_k$: Bağımsız değişkenlerin regresyon katsayıları

$X_1X_2 \ldots X_k$: Bağımsız değişkenler

k: Bağımsız değişken sayısı

e: 2.71 sabit katsayı

Lojistik regresyon analizi Berkson (1944) tarafından ilk olarak biyolojik deneylerde analiz yöntemi olarak önerilmesinin yanı sıra, Berkson'un önermiş olduğu modeli çeşitli uygulamalar yaparak literatüre Cox (1970) tarafından katkı sağlanmıştır. Lojistik regresyonun popüler hale gelmesi ise katsayı tahmin işlemlerinde diskriminant fonksiyonunu kullanan Cornfield (1962) sayesinde olmuştur.

İstatistiksel analizler yapılırken parametrelerin tahmin edilme aşamasında bir ok yöntem kullanılmaktadır. Bu yöntemlerin başında momentler yöntemi, en çok olabilirlik yöntemi gelmektedir. Bu çalışmada da en çok olabilirlik tahmin yöntemi kullanılmıştır.

En çok olabilirlik yöntemi;

Tanım: $X_1, X_2, \ldots, X_n$ örnklemi olasılık yoğunluk fonksiyonu $f(X_i;\theta)$ olan kitleden alına bir örneklem olmak üzere θ parametresi için olabilirlik fonksiyonu

$L(\theta; x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$ şeklindedir. Bu fonksiyonu max yapan değer θ parametresinin en çok olabilirlik tahmin edicisidir.

Receiver Operating Characteristic (ROC) analizi temel anlamda gerçekte doğru ve test sonucunda da doğru olarak kabul edilenlerin oranının gerçekte yanlış ama test sonucunda doğru diye kabul edilenlere oranı olarak bilinmektedir.

$$ROC = \frac{TP}{FP} \tag{2}$$

ROC eğrisi yöntemi aşağıda belirtilen hususlar dâhilinde kullanılabilir (Muhammad ve Amir, 2018).

- Kurulan modelin sınıflandırma gücünde
- Model performanslarının karşılaştırılması
- Eşit değerinin belirlenmesinde
- Model sonuçlarının kalite takibi
- Uygulamacı(ların) gelişim takibi
- Farklı uygulamacı(ların) karşılaştırılmasında

ROC analizi sınıflandırma algoritmalarının sonuçlarını değerlendirmede kullanılan bir analiz türüdür (Takıcı, 2018). Bu analiz algoritmaların performanslarını değerlendirmede kullanılan en yaygın yöntemlerden biridir.

TP rate; Bu oran gerçekte doğru olan bir durumun test sonucunda doğru diye sınıflandırılması olarak tanımlanabilir.

Duyarlılık, test tarafından doğru bir şekilde tanımlanan gerçek pozitiflerin oranıdır (Altman ve Bland, 1994).

$$Duyarlılık = TP/(TP + FN) \tag{3}$$

FP Rate; Gerçekte yanlış olan bir durumun test yapıldıktan sonra doğru olarak karar verilmesidir.

Özgüllük, test tarafından doğru bir şekilde tanımlanan gerçek negatiflerin oranıdır (Altman ve Bland, 1994).

$$Özgüllük = TN/(FP + TN) \tag{4}$$

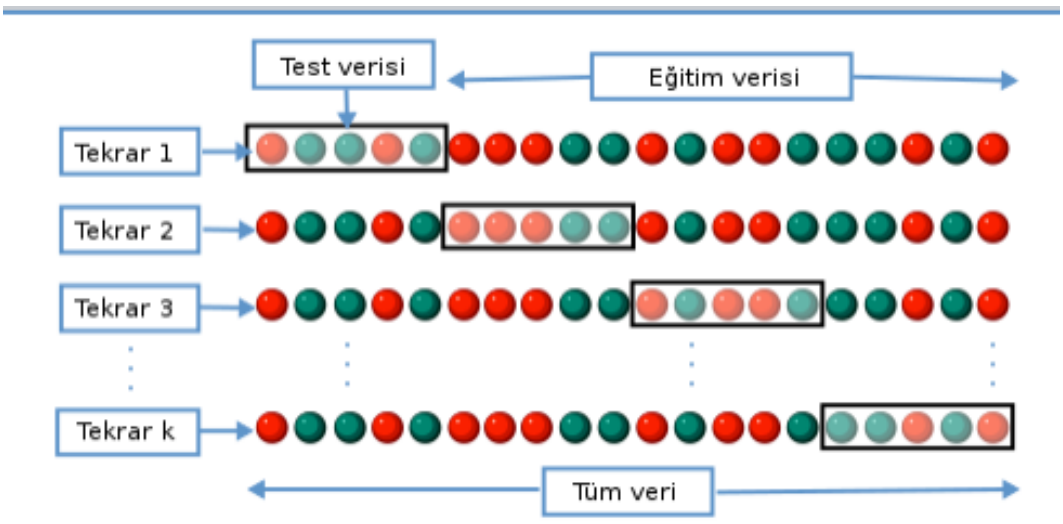
2.3. Özellik Seçimi

Özellik seçimi, makine öğrenmesi sistemlerinde kullanılan yöntemlerin doğru sınıflandırma oranlarını ve performanslarını geliştirerek daha tutarlı sonuçlar elde etmek için kullanılan önemli tekniklerden biridir. Bu çalışmada özellik seçimi yöntemleri arasından bulanık kaba küme teorisi kullanılmıştır.

Bulanık kaba kümeler, kesin bir bulanık kümenin yaklaşım uzayından türetilen kaba bir kümenin genelleştirilmesidir. Bu koşullu öznitelik değerlerinin kesin ve bulanık olduğu durumuna karşılık gelmektedir. Bulanık kaba küme teorisinin ana odak noktası bulanık kümenin evreni denklik nedeni ile pürüzlü hale geldiğinde ya da denklik ilişkisini benzer bulanık ilişkiye dönüştürüldüğünde söz konusu kümenin alt ve üst yaklaşımını tanımlamaktır (Kumar ve Yadav, 2015).

2.4. Çapraz Doğrulama

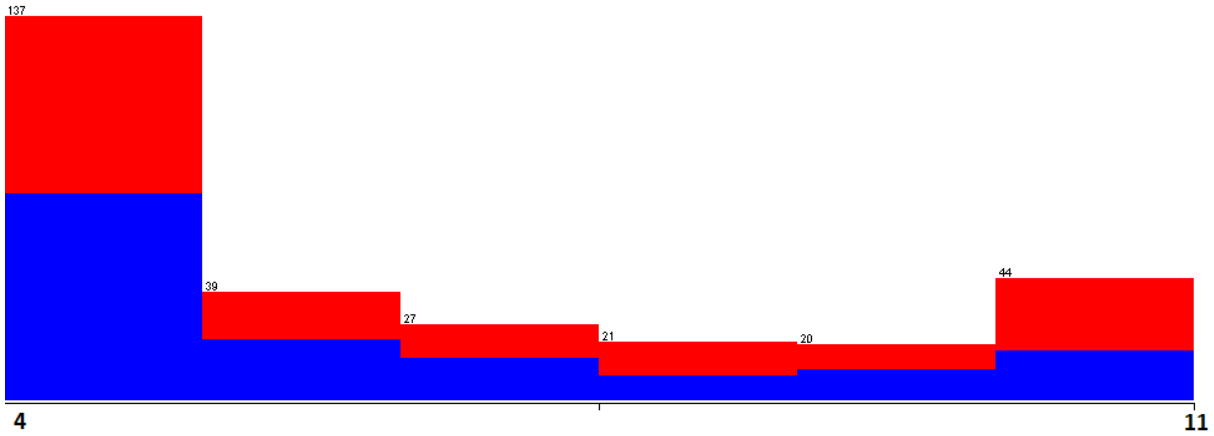
Çapraz doğrulama, makine öğrenmesi algoritmalarında, derin öğrenmede, yapay sinir ağları gibi birçok alanda verinin eğitilmesi ve testinde çok fazla kullanılan bir algoritmadır. Bu algoritmanın temelinde veri seti verilen k sayısı kadar eşit parçalar halinde bölünüyor. Bu parçaların bir tanesi test için kullanılırken geri kalan kısmı ise veri senin eğitiminde kullanılıyor. Bu döngü k sayısı kadar yapılıyor. Ancak her iterasyondaki bölünen gruplar diğer iterasyonlardan farklı olmak zorundadır. Bu sayede sistemin veri setini ezberlemesi engellenmiş oluyor. Şekil 2'de k'nın n olduğu bir çapraz doğrulama örneği görülmektedir.



Şekil 2. k=n için çapraz doğrulama modeli (Sanjay, 2018)

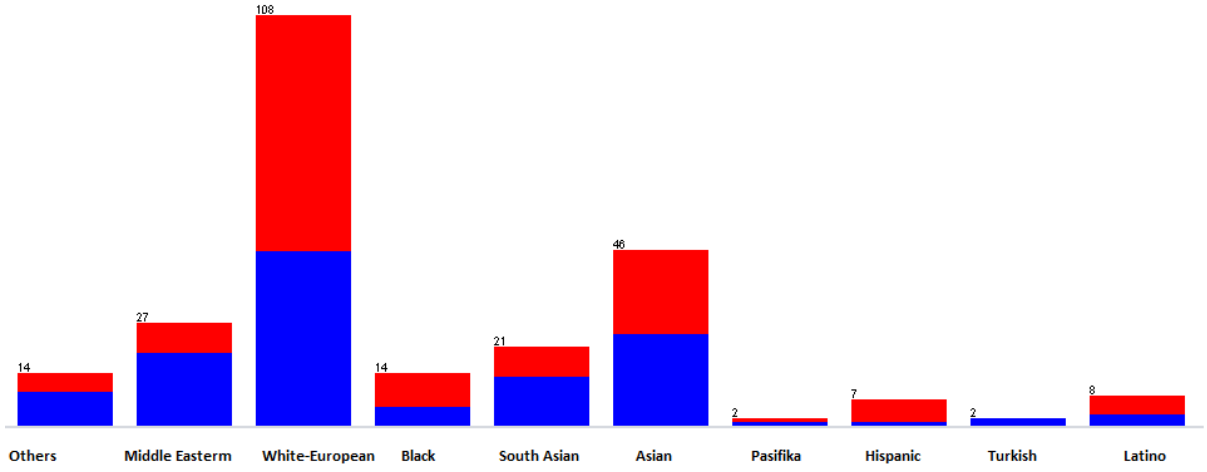
3. BULGULAR

Çalışmada kullanılan veriler toplamda 21 değişkenden oluşmaktadır. Bu değişkenlerin 20 tanesi OSB'nin belirlenmesi ile ilgili değişkenlerden oluşurken bir tanesi de OSB'nin olup olmadığı yani sınıfını belirtmektedir. Çocuklar üzerinden toplanan bu veriler 292 adettir. Bu veri ile ilgili alarak bazı detaylı bilgiler şekil 3 ve 4'de verilmiştir. Veri ile ilgili daha fazla bilgi almak isteyen araştırmacılar (Thabtah, 2017) kaynağına bakabilirler.



Şekil 3. Veri setinin yaş dağılımı ve rahatsızlık durumları

Şekil 3'de kişilerin yaş dağılımları ve rahatsızlık durumları yer almaktadır. Kırmızı ile gösterilen alanlar rahatsızlığın olduğu mavi ile gösterilen alan ise rahatsızlığın olmadığının bir göstergesidir. Ayrıca çocukların yaş dağılımı en az 4 iken en fazla 11'dir. Yaş değişkeni 6 farklı grubu ayrılmıştır. 1. Grup 137 kişiden, 2. Grup 39 kişiden, 3. Grup 27 kişiden, 4. Grup 21 kişiden 5. Grup 20 kişiden ve 6. Grup 44 kişiden oluşmaktadır. Renklerin gruplar içerinde dağılımlarına bakıldığı zaman hemen hemen aynı oranda oldukları görülmektedir.



Şekil 4. Veri setinin etnik grup dağılımı ve rahatsızlık durumu

Şekil 4 veri setinin etnik grup dağılımları ve rahatsızlığın olup olmama durumunu göstermektedir. Şekil 4'e göre kırmızı renk rahatsızlığın, mavi renk ise rahatsızlığın olmadığını göstermektedir. Ayrıca Latino'da 8, Türkiye'de 2, Hispanic 7, Pasifika 2, Asian 46, Sounth Asian 21, Black 14, White-European 108, Middle Eastern 27 ve other 14 kişiden oluşmaktadır. Renk dağılımları incelendiğinde Türkiye'den ele alınan 2 kişinin her ikisi de rahatsızlığı olmayan kişilerden oluştuğu, White-European'da rahatsızlığı olan kişiler biraz daha fazla olduğu görülmektedir. Ayrıca Middle-Easetem'de ise rahatsızlığı olmayanların sayısı daha fazladır. Bunların dışında diğer etnik gruplarda hemen hemen rahatsızlık olup olmama durumlarının aynı olduğu şekil 4'de görülmektedir.

Fuzzy Rough Set ile özellik seçimi yapıldıktan sonra söz konusu rahatsızlığın sınıfı dahil 6 değişkene indirgeme yapılmıştır. Bu değişkenlerin isimleri aşağıdaki tablo 2'de verilmiştir.

Tablo 2. Bulanık kaba küme özellik seçimi yapıldıktan sonra kalan değişkenler

Değişken 1
Değişken 6
Değişken 8
Değişken 13
Değişken 20

Tablo 3'de özellik seçimi kullanmadan önce lojistik regresyon analizinin doğru sınıflandırma soncu, TP, FP ve ROC analizi sonuçları yer almaktadır. Bu sonuçlara göre lojistik regresyon analizi OSB rahatsızlığının sınıflandırılmasında %95.205 oranında bir başarı sağlamıştır. Yani analiz sonucunda 292 çocuktan sadece 14 tanesini yanlış sınıflandırmıştır. Bu

sınıflandırmaların 8 tanesi OSB rahatsızlığı yok iken algoritma sonucundan rahatsızlık olduğunu belirlenmişken 6 tanesinin ise rahatsızlık var iken olmadığı tespiti yapılmıştır. Bu durum tablo 4'de açıkça görülmektedir. Ayrıca bu tabloya göre rahatsızlığı bulunmayan 151 kişinin 143 tanesi doğru sınıflandırarak rahatsızlığı yoktur olarak belirlenmiştir. Yani gerçekte rahatsız değil iken test sonucu rahatsız değildir olarak belirlenmesi TP oranını vermektedir. Bu oran 0,947 olarak hesaplanmıştır. Buna karşın rahatsızlığı bulunan 141 kişinin 135 tanesi rahatsızlığı vardır olarak sınıflandırılmıştır. Burada ise gerçekte rahatsızlığı olduğu halde analiz sonucunda rahatsızlığı yoktur olarak sınıflandırılan 6 kişidir. ROC analizi ise bir testin ayırt etme gücünün belirlenmesinde çeşitli tekniklerin karşılaştırılmasında ve uygun eşik değerinin belirlenmesinde kullanılan bir analizdir. Bu analiz sonucuna göre ROC alanı 0,992 olarak hesaplanmıştır. Bu değerin 1'e ne kadar yakın olması hastalığın tespitinden kullanılan değişkenlerin uyumunun o kadar iyi olduğunu göstermektedir.

Tablo 3. Özellik seçimi yapılmadan önce algoritma sonuçları

Lojistik Regresyon	
Doğru Sınıflandırma Oranı (%)	95,205
TP Oranı	0,947
FP Oranı	0,043
ROC Alanı	0,992

Tablo 4. Özellik seçimi olmadan önce confusion matrisi

a	**b**	**Sınıfı**
143	8	a = Hayır
6	135	b = Evet

Tablo 5'de bulanık kaba küme özellik seçimi kullanılarak değişken azaltmasına gidilerek lojistik regresyon analizi ile elde edilen sonuçlar yer almaktadır. Bu sonuçlara göre algoritmanın rahatsızlığı doğru sınıflandırma oranı yaklaşık olarak %97 olarak hesaplanmıştır. Tablo 6'da de bu sınıflandırmaların nasıl dağıldığı gösterilmektedir. Tablo 6'ya göre toplamda rahatsızlığı olmayan 151 kişinin 5 tanesini yanlış geri kalan 146 tanesini doğru sınıflandırmıştır. Benzer şekilde rahatsızlığı olan 141 kişinin 136 tanesini doğru sınıflandırırken geri kalan 5 tanesini rahatsızlığı yoktur diyerek yanlış sınıflandırmıştır. Bu sayılara göre TP oranı 0,974 iken FP oranı ise 0,028 olarak hesaplanmıştır. Yani gerçekte rahatsız olmayan kişilere analiz sonucunda rahatsız değildir diye belirleme oranı yaklaşık olarak %98 iken rahatsız olanları rahatsız değildir diye sınıflandırma oranı %2'dir.

Tablo 5. Fuzzy rougt set ile özellik seçimi yapıldıktan sonra algoritma sonuçları

Lojistik Regresyon	
Doğru Sınıflandırma Oranı (%)	96,575
TP Oranı	0,974
FP Oranı	0,028
ROC Alanı	0,996

Tablo 6. Bulanık kaba küme ile özellik seçimi yapıldıktan sonra karışıklık matrisi

a	**b**	**Sınıfı**
146	5	a = Hayır
5	136	b = Evet

4. TARTIŞMA VE SONUÇ

Elde edilen bulgulara göre OSB ilk aşamada, yani hastalığın belirlenmesinde kullanılan 20 değişkene hiçbir müdahale yapıldığında lojistik regresyon analizi ile analiz edildiğinde söz konusu rahatsızlık %95 oranında doğru sınıflandırılmıştır. Ancak değişkenler bulanık kaba küme algoritması ile azaltılarak bu oran yaklaşık olarak %97'ye çıkarılmıştır. Özellik azaltmadaki temel amaçlardan biride ilgili algoritmanın başarı performansını artırmaktır. Bu sayede söz konusu algoritmanın doğru sınıflandırma oranında artış sağlayarak bir başarı elde edildiği açık bir şekilde görülmektedir. Ayrıca özellik seçiminden sonra elde edilen TP, FP ve ROC analizlerinde de bir başarı elde edilmiştir. Özellik seçimi yapılmadan önce bu oranlar sırasıyla 0,947, 0,043 ve 0,992 iken seçim yapıldıktan sonra TP 0,974'e, ROC 0,996'ya yükselirken FP ise 0,028'e düşmüştür. Özellik seçiminden sonra algoritmanın doğru sınıflandırma oranındaki başarıya benzer şekilde performans değerlendirme ölçütlerinde de görülmektedir.

Bu alanda çalışan uzmanlara rahatsızlığın doğru tespiti ve hızlı karar vermede yardımcı olmak adına yapılan bu çalışmada uzman kişilere söz konusu algoritmanın özellik seçimi yapıldıktan sonra daha başarılı olduğu tespit edilmesinden dolayı bulanık kaba küme özellik seçimi kullanmaları önerilmektedir. Ayrıca çalışmadan elde edilen sınıflandırma sonuçları incelendiğinde %1'lik bir doğru sınıflandırma oranında artış olduğu görülmektedir. Bu artış, insan sağlığı üzerine yapılan çalışmalar düşünüldüğünde ciddi bir artış olduğu görülecektir. Bu çalışmayı diğer çalışmalardan ayıran en önemli yanı, bulanık kaba küme teorisi ile birlikte lojistik regresyon analizinin bir arada kullanılarak analiz edilmesidir. Özellik seçim yöntemleri

arasında bulanık kaba küme teorisi çok sık kullanılan bir yöntem olmaması ve bu çalışmada kullanılmış olması ile birlikte elde edilen sonuçlar literatüre önemli katkılar sağlamaktadır.

Sınıflandırma problemlerinde kullanılan birçok algoritma olduğu bilinmektedir. İlerleyen zamanlarda literatürde yer alan diğer algoritmalarda kullanılarak sonuçların karşılaştırılması planlanmaktadır. Benzer şekilde literatürde var olan diğer özellik seçim algoritmaları kullanılarak hangisinin daha etkili olduğu karşılaştırması yapılarak en iyi sınıflandırma algoritması ve en iyi özellik seçimin hangisinin olduğuna dair tespitler düşünülmektedir. Son olarak yazılım aracılığı ile bir telefon uygulaması yapılarak ilgili değişken değerleri girildikten sonra kişinin OSB rahatsızlığı olup olmadığı tespiti yönünden çalışmalar planlanmaktadır.

ETİK BEYAN

"Otizm Spektrum Bozukluğunda Bulanık Kaba Küme Özellik Seçimi Kullanılarak Lojistik Regresyon İle Sınıflandırılması" başlıklı çalışmanın yazım sürecinde bilimsel, etik ve alıntı kurallarına uyulmuş; toplanan veriler üzerinde herhangi bir tahrifat yapılmamış ve bu çalışma herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiştir.

KAYNAKÇA

Altman, D.G. ve Bland, J.M. (1994), Diagnostic tests. 1: Sensitivity and specificity*, British Medical Journal,* 308, (6943): 1552.

Amerikan Psikiyatri Birliği (2013), DSM-V-R Tanı Ölçütleri Başvuru Kitabı, Ertuğrul Köroğlu (çeviri editörü), Ankara: HYB Yayıncılık.

Berkson, J. (1944), Application of the logistic function bio-assay, *Journal of the American Statistical Association*, 39, 35-365.

Bodur, Ş. ve Soysal, A.Ş. (2004), Otizmin erken tanısı ve önemi, *STED Dergisi,* 13, 394-398.

Cornfield, J. (1962), Joint dependence of the risk of coronary heart disease on serum cholesterol and sistolic blood pressure: A diskrimant function analysis, *Federation Proceedings*, 21: 58-61.

Cox, D. R. ve Snell, E. S. (1989), Analysis of binary data, London.

Halk Sağlığı, (2019), *https://lk.tc/rCw7S*, Erişim Tarihi: 22.08.2019.

Hosmer, D.W. ve Lemeshow, S. (2000), Applied logistic regression, Second edition, *A Wiley-Interscience Publication*.

Kanner, L. (1943), Autistic disturbances of affective contact, Nervous Child, 2,217-250

Kumar, M. ve Yadav, N. (2015), Fuzzy rough sets and its application in data mining field, *Advances in Computer Science and Information Technology (ACSIT)*, 2, 237-240.

Muhammad, Z.A. ve Amir, A. (2018), Performance evaluation of supervised machine learning classifiers for predicting healthcare operational decisions, *Wavy AI Research Foundation*: Lahore, Pakistan.

Rapin, I. and Katzman, R. (1998), Neurobiology of autism. *Ann Neurology*, 43, 7–14.

Sanjay, M. (2018), *https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f*, Erişim Tarihi: 22.08.2019.

Takıcı, H. (2018), Improvement of heart attack prediction by the feature selection methods, *Turkish Journal of Electrical Engineering Computer Sciences*, 26, 1-10.

Thabtah F. (2019), Machine learning in autistic spectrum disorder behavioral research: A review and ways forward, *Informatics for Health & Social Care*, 44, 278–297.

Thabtah, F. (2017), Autism spectrum disorder screening: Machine learning adaptation and DSM-5 fulfillment, *Proceedings of the 1st International Conference on Medical and Health Informatics 2017*, 1-6.

Thabtah, F.F. (2017), *https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult* , Erişim Tarihi: 20.07.2019.