All Sciences Proceedings

3rd International Conference on Innovative Academic Studies

September 26-28, 2023 : Konya, Turkey

ICIAS

All Sciences Proceedings
http://as-proceeding.com/

https://www.icias.net/

Makine Öğrenimi Algoritmaları ile Otistik Spektrum Bozukluğu Tanısı Koyma

Mustafa SU[1*], Hidayet TAKCI [2]

[1]*Bilgisayar Mühendisliği /Fen Bilimleri, Sivas Cumhuriyet Üniversitesi, Türkiye*
[2]*Bilgisayar Mühendisliği / Fen Bilimleri, Sivas Cumhuriyet Üniversitesi, Türkiye*

[*]*(su.mustafa@hotmail.com)*

Özet – Otizm kişilerin çevreleriyle normal ilişkiler kurmakta zorlandığı gelişimsel bir bozukluktur. Erken teşhis otizmli kişilerin eğitilip normal ilişkiler kurmasında büyük önem taşımaktadır. Bu çalışmada, otizmin erken teşhisine yardımcı olup otizmli bireylerin gelişimine katkı sağlayabilmek için Naive Bayes, Lojistik Regresyon, K-en Yakın Komşu ve Rastgele Orman algoritmalarıyla otizme erken teşhis konulması hedeflenmiştir. Çalışmanın gerçekleştirilmesi için otizmli kişilerin bebeklik, ergenlik ve yetişkinlik dönemlerinin olduğu veri setleri kullanılmıştır. Ham veri kullanılarak herhangi bir optimizasyon yapılmadan elde edilen modeller ile veriler üzerinde aykırı verilerin temizlenmesi, eksik verilerin ortalama değerle doldurulması, öznitelik seçimi ve parametre optimizasyonları sonrasında elde edilen modellerin başarı durumları karşılaştırılmıştır. Yapılan çalışmalar sonucunda veri ön işleme ve parametre optimizasyonu yapılmadan elde edilen sonuçlara oranla; bebeklik veri setinde Naive Bayes %3.78, Lojistik Regresyon %10.34, K-en Yakın Komşu %0.92 ve Rastgele Orman algoritması %11.02, ergenlik veri setinde Naive Bayes %7.01, Lojistik Regresyon %25.97, K-en Yakın Komşu %7.25 ve Rastgele Orman algoritması %16.13, yetişkinlik veri setinde Naive Bayes %2.27, Lojistik Regresyon %10.43, K-en Yakın Komşu %1.13 ve Rastgele Orman algoritması %5.69 performans artışı göstermiştir. Bu çalışma, veri ön işleme ve parametre optimizasyonları sonrasında elde edilen modellerin başarı oranlarının ham veri seti ile herhangi bir optimizasyon ve veri ön işleme adımı uygulamadan elde edilen modellere göre arttığını göstermektedir.

Anahtar Kelimeler – Otizm, Makine Öğrenimi, Veri Ön İşleme, Parametre Optimizasyonu, Erken Teşhis

I. GİRİŞ

Otizm, genellikle bebeklik ve çocukluk yıllarında başlayan ve ölene kadar devam eden, bireyin çevresiyle iletişimini doğrudan etkileyen ve hem sözel hem de sözel olmayan şekilde uygun ilişki kurmasını engelleyen bir gelişimsel bozukluktur. Bu gelişim bozukluğu bireylerin farklı alanlarda sosyalleşme, iletişim ve davranış konularında sorunlar yaşamasına sebep olur. Bu sorunlar sebebiyle genellikle sözlü iletişim ve arkadaşlık kurma konusunda sıkıntıları bulunur. Otizm kaynaklı gelişim bozukluğu sonucunda birey kendini ifade etmekte zorluk yaşayabilir ve duygusal durum değişiklikleri gösterebilir. Kendilerini ifade etmekte zorlanan bu bireyler bu sebeplerden ötürü öfkelenebilir ve sakinleşmekte zorluk yaşayabilirler. Devrik ve karmaşık cümleler ile günlük hayatta kullanılan mecazi ifadeleri anlayamazlar. Bu zorluklarla başa çıkmak için ise bilinçli ebeveyn desteğine, etkili ve bilinçli eğitim desteğine ve sosyal etkileşim içinde olmalarına ihtiyaçları vardır. Bu bireylerin eğitimlerine öncelik verilmesi ve erken başlanması bireylerin gelişimi için çok önemlidir. Bu ortamın sağlanması ve gelişim bozukluğunun yavaşlatılarak engellenmesi için en önemli adım bu bozukluğun erken teşhis edilmesidir. Bu bozukluğun teşhisi ise zor ve uzun süre almaktadır. Hastalığın teşhisi sürecindeki gecikmeleri ve belirsizlikleri engellemek için son

dönemde sıklıkla makine öğrenmesi teknikleri kullanılmaktadır.

Bu çalışma kapsamında da üç ayrı yaş grubuna ait veri seti üzerinde Naive Bayes algoritmasının orijinal verideki başarı oranları ile veri ön işleme ve parametre optimizasyonu sonrasındaki başarı oranları karşılaştırılmış ve raporlanmıştır. Bu yöntemle eğitilen modellerin başarısı yüzde 2 ile yüzde 26 arasında artış göstermiştir. Çalışma, parametre optimizasyonu ve veri ön işleme tekniklerinin otizm tespit oranlarını artırıp artırmadığını görmek üzere yapılmıştır.

Literatürde farklı veri madenciliği algoritmaları ile Otizm teşhisi konusunda yapılan araştırmalar ve algoritmaların bazılarını şu şekilde özetlemek mümkündür:

Sedat Metlek'in "Otistik Spektrum Bozukluğunun Makine Öğrenme Algoritmaları ile Tespiti" adlı çalışmasında 12-36 ay arasındaki çocuklardaki OSB teşhisinde yardımcı olabilecek bir yazılım geliştirilmiştir. Nöro-gelişimsel bir gelişme bozukluğu olan Otizmin erken teşhisi ve teşhis doğruluk oranının yüksek olması bu çalışmanın öncelikli alanlarındandır. Bu çalışmada birbirinden farklı makine öğrenmesi algoritması ile otizm teşhisi yapılmıştır. Altı farklı makine öğrenmesi kullanılan bu çalışma ile yüksek doğruluklu erken OSB teşhisi amaçlanmıştır. Çalışmada gözetimli ve gözetimsiz modeller kıyaslanmaktadır. Bu test ve karşılaştırmaların sonucunda gözetimsiz algoritmalara oranla; gözetimli öğrenme algoritmalarının daha başarılı sonuçlar verdiği gözlemlenmiştir [1].

2019 yılında Cho ve arkadaşları tarafından yayınlanan "Automatic Detection of Autism Spectrum Disorder in Children Using Acoustic and Text Features from Brief Natural Conversations." Başlıklı çalışma da otizm teşhisi konusunda başarılı olmuş bir çalışmadır. Bu çalışmada öncelikle gerçek hayatta OSM saptaması yapılması amaçlanmıştır. Doktorlar ve okulların bulunduğu gerçek dünya çalışmasında doğal konuşmalarla teşhis edilebilen bir otomatik OSB saptama sistemi geliştirilmesi amaçlanmıştır. Projede Gradient tabanlı bir algoritma ile çalışılmış ve %76 teşhis başarısı gerçekleştirilmiştir [2].

Jaber Alwidian, Ammar Elhassan ve Ghnemat'ın çalışmasında ise bireyin otizmli olup olmadığı ilişkilendirme sınıflandırması tekniği kullanılarak yedi farklı algoritma ile test edilmiştir. Aralarındaki korelasyonlar incelenmiş ve en yüksek doğruluk oranı ile çalışan algoritma tespit edilmiştir. Bu çalışmanın sonucu olarak Birliktelik Kuralları Algoritmasına Dayalı Ağırlıklı Sınıflandırma (WCBA) algoritması en yüksek doğruluğu vermiştir [3].

Suman Raj ve Sarfaraz Masood tarafından 2019 yılında Uluslararası Sayısal Zekâ ve Veri Bilimi Konferansı'nda yayınlanan "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques" çalışması da bu gelişim bozukluğunu evre evre incelemiş ve buna göre üç gelişim evresinin analizini ortaya koymuştur. Bu çalışmada Naive Bayes, Support Vector Machine, Logistic Regression, KNN, Neural Network ve Convolutional Neural Network uygulamalarının bahsedilen evrelerdeki performansı analiz edilmiştir. Convolutional Neural Network uygulamasının doğruluk oranı daha yüksek çıkmıştır [4].

M. S. Mythili ve A. R. Mohamed Shanavas "Sınıflandırma Teknikleri Kullanılarak Otizm Spektrum Bozuklukları Üzerine Bir Çalışma" isimli bir makale yayınlamışlardır. Bu makalede otizm gelişim bozukluğu düzeyleri veri madenciliği ve sınıflandırma algoritmalarıyla tespit edilmeye çalışılarak, çocukların eğitim ve algılama performansının artırılması üzerine çalışılmıştır. Bulanık bilişsel harita optimizasyonunun doğruluğundan bahsedilmiştir [5].

Hailong Li ve arkadaşları tarafından yayınlanan "Optimize Edilmiş Makine Öğrenimi Modelleri ve Kişisel Karakteristik Verilerle Otizm Teşhisini Geliştirmek" adlı çalışmada kişisel veriler de incelenmiş ve benimsenen klinik yöntemlerden farklı bir yol izlenmiştir. Veriler üzerinden yaptıkları çalışmada bireyin yaş, cinsiyet, IQ seviyesi gibi kişisel verileri de kullanılmış ve denetimli makine öğrenimi deneyleri yapılmıştır. Karakteristik özelliklerin etkisinin gözlemlendiği bu çalışmada 9 ayrı denetimli makine öğrenim modeli test edilmiştir. Yapay sinir ağı modeli ve k-en yakın komşu algoritması bu çalışmada en yüksek verimliliği gösteren algoritmalar olmuştur [6].

Devika Varshini G ve Chinnaiyan R, 2020 yılında "Otizm Spektrum Bozukluğunun Tahmini için Optimize Edilmiş Makine Öğrenimi Sınıflandırma Yaklaşımları" adlı çalışmalarını yayınlamışlardır. Bu çalışmada farklı gelişim evrelerindeki bireylerin barındırdığı otizm özellikleri incelenmiş ve sınıflandırılmıştır. Bu projede psikologların kullandığı otizm belirtilerinin incelendiği puanlama

yönteminden yararlanılmıştır. Lojistik Regresyon ve Random Forest gibi algoritmalar yardımıyla otizm saptamasının analizi yapılmıştır. KNN algoritmasında en yüksek doğruluk oranına ulaşılmıştır [7].

Fatiha Nur Büyükoflaz ve Ali Öztürk'ün "Makine Öğrenmesi Algoritmaları ile Çocuklarda Erken Otizm Teşhisi" çalışmasında; Naive Bayes, K-En Yakın Komşu ve Random Forest gibi algoritmalar kullanılmıştır. Kendi aralarında performanslarının değerlendirilmesi yapılmış ve bunun sonucunda en yüksek performansı Random Forest algoritması vermiştir. Naive Bayes uygulaması %96,55 başarı oranı vermiştir [8].

Azian Azamimi Abdullah, Saroja Rijal ve Satya Ranjan Dash'ın yayınladıkları "Otizm Spektrum Bozukluğunun Sınıflandırılmasına Yönelik Makine Öğrenimi Algoritmalarının Değerlendirilmesi" isimli çalışmalarında veri setleri merkezde tutulmuş ve böylece otizm sınıflandırılması makine öğrenimi algoritmaları vasıtasıyla yapılmaya çalışılmıştır. Veri setlerinin çeşitlendirildiği bu çalışmada modellerin doğruluk oranları ve performansına odaklanılmamış, veri çeşitliliği merkezde tutulmuştur. Denetimli makine öğrenimi algoritmalarının ve çeşitli test yöntemlerinin denendiği bu çalışmada Lojistik Regresyon algoritmasının en yüksek doğruluğu verdiği açıklanmıştır [9].

Ayşe Demirhan'ın 2018 yılında yayınlamış olduğu "Otizm Spektrum Bozukluk Vakalarını Belirlemede Makine Öğrenme Yöntemlerinin Performansı" adlı çalışmasında Destek Vektör Makineleri, K-En Yakın Komşu ve Random Forest algoritmaları kullanılmıştır. Otizm bozukluğunun en hızlı ve doğru saptanmasının amaçlandığı bu çalışmada algoritmaların kendi içerisinde performansı değerlendirilmiştir. Random Forest algoritması ile en yüksek performans elde edilmiştir. Bu çalışmaya göre otizm bozukluğunun en doğru şekilde tespit edilmesinde Random Forest algoritmasının tam başarı sağladığı belirtilmiştir [10].

Yukarıda bahsedilen çalışmalar ve benzerleri incelendiğinde kimi çalışmalarda tek bir makine öğrenme algoritması kiminde birden çok makine öğrenmesi algoritması kullanıldığı saptanmıştır. Otizm teşhisi için denetimli makine öğrenme algoritmaları kullanılmaktadır. Çalışmalar farklı veri setleri kullansalar da hepsinin ortak amacı teşhis işlemi ve teşhisin iyileştirilmesidir. Çalışmamız da bu kapsamda ele alınmış ve bir makine öğrenmesi algoritması üzerinde parametre iyileştirme çalışmaları ile sonuç iyileştirme yapılmıştır.

II. MATERYAL VE YÖNTEM

A. *Veri Setleri*

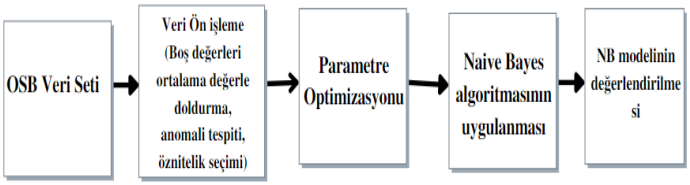
Bu çalışmada, yetişkinlik, ergenlik ve bebeklik dönemlerinin veri setleri kullanılmıştır. Veri setleri, Otizm teşhisinde etkili olan parametrelere ait verileri içermektedir. Veri setlerinde, otizm teşhisinde etkili olduğu belirlenmiş davranışsal ve bireysel 10 özellik bulunmaktadır. Ergenlik dönemi veri seti 104, yetişkinlik dönemi veri seti 704 ve çocukluk dönemi veri seti 292 adet örnek içermektedir [11], [12] - [13]. Tablo 1'de veri setlerinin özellikleri gösterilmiştir.

Tablo 1. Veri Setlerinin Özellikleri

Sütun	Açıklaması
1	Hasta yaşı (Yıl olarak)
2	Hasta cinsiyeti
3	Etnik köken
4	Kişinin sarılıkla doğup doğmadığı durumu
5	Herhangi bir yakın aile üyesinde osb var mı?
6	Ebeveyn, kendi, bakıcı, sağlık personeli, klinisyen vb.
7	Yaşanılan ülke
8	Kullanıcının bir tarama uygulaması kullanıp kullanmadığı
9	Yaş kategorisine göre seçilen tarama yöntemlerinin türü (0=bebek, 1=çocuk, 2=ergen, 3=yetişkin)
10-19	Davranışsal özelliklerle ilgili 10 soru
20	Kullanılan tarama yönteminde elde edilen nihai puan

B. *Çalışma Metodolojisi*

Şekil 1'de çalışmada yapılan iş akışı gösterilmektedir.



Şekil 1. Çalışma iş akış diyagramı

Öncelikle veri setleri üzerinde aykırı verilerin temizlenmesi ve boş verilerin sütun ortalamasındaki değer alınarak doldurulması adımları uygulanarak

veri ön işleme adımı tamamlanmıştır. Bu adım ile veri setleri maksimum verim alınacak duruma getirilmiştir. Daha sonra parametre optimizasyonu adımına geçilmiştir. Bu adımda modelleri eğitirken k-nn algoritması için farklı k değerlerinin denenmesi, Naive Bayes için laplace yumuşatması kullanılıp kullanılmama durumu, Lojistik regresyon için doğrusal sütunların kaldırılıp kaldırılmaması, p-değerlerinin hesaplanıp hesaplanmaması ve Rastgele Orman algoritması içinde ön budama yapılıp yapılmaması, karışık bölme ve farklı maksimum derinlik değerlerinin kullanılması ile modeller eğitilmiştir. Modeller test edilirken ise çapraz doğrulama katlama sayısı, tek çıkışlı çapraz doğrulama, ve ağırlık seçimi gibi parametrelerin farklı farklı değerleri ile modeller test edilmiştir. Parametre optimizasyonu adımı belirlenen parametlerin farklı farklı kombinasyonları ile birçok kez tekrarlanmıştır.

Tablo 2. Rapidminer parametre optimizasyonu iterasyon değerleri

İterasyon	Naive Bayes Laplace Yumuşatması	Ağırlık Seçimi	Çapraz Doğrulama Katlama Sayısı	Doğruluk
122	Hayır	0.050	14	0.936
123	Evet	0.060	14	0.936
243	Evet	0	31	0.951
124	Hayır	0.060	14	0.936
424	Hayır	0.020	54	0.948
125	Evet	0.070	14	0.940
126	Hayır	0.070	14	0.933
127	Evet	0.080	14	0.936
128	Hayır	0.080	14	0.933
425	Evet	0.030	54	0.946
244	Hayır	0	31	0.936

Tablo 2' de Naïve Bayes algoritması için farklı parametrelerle yapılan iterasyonlar ve sonuçları verilmiştir. Tüm algoritmalar için her bir iterasyon sonucunda modeller değerlendirilerek maksimum doğruluk oranına sahip modeller elde edilmeye çalışılmıştır. Tablo 2'deki değerler rapidminer sonuç ekranından alınmıştır.

C. Veri Ön İşleme

Veri ön işlemede, eksik ve tutarsız veriler de düzeltme yaparak veri setlerinden maksimum verim elde edilmeye çalışılmaktadır. Veri ön işleme sonrasında genellikle algoritmaların başarı oranlarında artış görülmektedir [14]. Bu çalışmada aykırı verilerin tespiti için Öklid mesafesi kullanılıp aykırı veriler veri setlerinden temizlenmiştir. Sütunun ortalama değeri alınarak da o sütundaki eksik değerler doldurulmuştur.

D. Parametre Optimizasyonu

Çalışmanın parametre optimizasyonu aşamasında modeller eğitilip test edilirken farklı parametre kombinasyonları ile modeller eğitilip test edilmiştir. Aşağıdaki parametreler ile farklı kombinasyonlar oluşturulup eğitim ve test aşamaları tekrarlanmıştır.

- K-NN algoritması için farklı k değerlerinin denenmesi
- Naive Bayes için laplace yumuşatması kullanılıp kullanılmama durumu
- Lojistik regresyon için doğrusal sütunların kaldırılıp kaldırılmaması, p-değerlerinin hesaplanıp hesaplanmaması
- Rastgele Orman algoritması içinde ön budama yapılıp yapılmaması, karışık bölme ve farklı maksimum derinlik değerlerinin kullanılması
- Minimum 2 maksimum 60 katlama değerlerinin çapraz doğrulamada kullanılması
- Öznitelik ağırlıklarının tespitinde Gini indeksi ve korelasyon yöntemlerinin kullanılması [15] – [16]
- Minimum 0, maksimumu 0.155 eşik değerinin öznitelik eşik değeri için kullanılması

Bu parametrelerin farklı kombinasyonları kullanılarak algoritmaların başarı durumları kayıt edilmiştir.

E. Naive Bayes (NB)

Naive Bayes denetimli bir sınıflandırma algoritmasıdır. Naive Bayes modellerinin eğitilmesinde etiketli veriler kullanılmaktadır. Etiketli veriler kullanılarak olasılık işlemleri yapılır ve olasılık durumları ile model oluşturulur. Etiketsiz veriler modele verilerek eğitim aşamasındaki olasılık durumları kullanılarak etiketsiz verilerin etiketlenmesi sağlanmış olur. Kategorisi belirli yani etiketli veri sayısının fazlalığı modelin başarısına olumlu olarak yansımaktadır [17].

F. Lojistik Regresyon (LR)

Lojistik regresyon istatiksel tabanlı sınıflandırma algoritmasıdır. Birçok sınıf arasındaki örnek veya veri noktasının hangi sınıfta olduğunu tahmin etmek

için kullanılmaktadır. Özellikle ikili sınıflandırmada yaygın olarak değerlendirilmektedir. Sürekli veri seti için kullanılmaktadır [18].

G. K-NN

KNN denetimli bir öğrenme yaklaşımıdır ve en basitidir. Regresyon problemlerinin yanı sıra sınıflandırma için de kullanılır. Yakınlarda benzer verilerin mevcut olduğunu varsayar. 'K' kısmı seçilecek başlangıç noktasının sayısını gösterir. Hatayı azaltmak için dikkatli seçilmelidir. Dolayısıyla mesafe, yakınlık veya yakınlık açısından olabilecek benzerlik fikrine dayanmaktadır. En yaygın uzaklık ölçüsü Öklid uzaklığıdır [19].

H. Rastgele Orman (RO)

Rastgele Orman sınıflandırma problemlerinde yaygın olarak kullanılan denetimli bir sınıflandırma algoritmasıdır. Birçok karar ağacının bir araya gelip orman oluşturduğu öğrenme yöntemidir. Bu orman etiketsiz verilerin doğru bir şekilde sınıflandırılması için oluşturulur. Temel olarak karar ağaçlarından faydanılmaktadır. Ağaç sayısının ve rastgeliliğin arttırılmasıyla daha doğru sonuçlar elde edilebilir [20].

İ. Modellerin Değerlendirilmesi

Performans değerlendirilmesinde karmaşıklık matrisi kullanılmıştır [21]. Değerlendirme aşamasında 1. ve 4. denklemlerinde verilen değerler ve ROC eğrisi altındaki alan kullanılmıştır [22].

Tablo 3. Karmaşıklık Matrisi

Gerçek		Tahminlenen: **Doğru**	Tahminlenen: **Yanlış**
	Doğru	DP	YN
	Yanlış	YP	DN

DP: Hasta olup, doğru tahmin edilen
YN: Hasta olup, yanlış tahmin edilen
YP: Hasta olmayıp, yanlış tahmin edilen
DN: Hasta olmayıp, doğru olarak tahmin edilen

$$\text{Doğruluk Oranı} : \frac{(DP+DN)}{DP+YN+DN+YP} \quad (1)$$

$$\text{Hassasiyet} : \frac{DP}{DP+YP} \quad (2)$$

$$\text{Geri Çağırma} : \frac{DP}{DP+YN} \quad (3)$$

$$\text{F1 Skoru: } 2x\left(\frac{HassasiyetxGeriÇağırma}{Hassasiyet+GeriÇağırma}\right) \quad (4)$$

III. BULGULAR

Tablo 4' de çocukluk veri seti üzerinde herhangi bir veri ön işleme yapılmadan ve parametre optimizasyonu yapılmadan elde edilen modellerin başarı oranları görülmektedir. En yüksek doğruluk değeri %98.28 ile K-NN, en düşük doğruluk değeri ise %85.71 ile Rastgele Orman algoritmasından elde edilmiştir.

Tablo 4. Çocuk ham veri seti ile algoritmaların performansları

	NB	**LR**	**K-NN**	**RO**
Doğruluk Oranı %	93.80	89.66	98.28	85.71
AUC	0.982	0.917	0.99	0.96
Hassasiyet %	89.99	91.43	97.22	90.91
F1 Skoru %	93.26	91.86	98.80	86.96

Ham veri ile yapılan değerlendirmeden sonra veri ön işleme ve parametre optimizasyonu ile modeller oluşturulup sonuçları değerlendirilmiştir.

Tablo 5. Veri ön işleme ve parametre optimizasyonu sonrası çocukluk veri setinde algoritmaların performansları

	NB	**LR**	**K-NN**	**RO**
Doğruluk Oranı %	97.58	100	99.2	96.73
AUC	0.994	1	0.99	0.98
Hassasiyet %	95.03	100	98.55	98.30
F1 Skoru %	97.03	100	99.27	98.25

Tablo 5' de görüldüğü üzere bu çalışmadaki yöntem kullanılarak algoritmaların başarı oranlarında artış sağlanmıştır. Ayrıca çocukluk döneminde erken teşhis için algoritmalar arasında daha detaylı ve kapsamlı eğitim ve test süreçleri gerçekleştirilmiş ve başarılı sonuçlar elde edilmiştir. Bu adım sonrasında en başarılı algoritma, ham veri ile herhangi bir optimizasyon olmadan yapılan çalışmada en düşük sonuçlardan birini veren Lojistik Regresyon algoritması olmuştur. Lojistik regresyon tüm verilerin durumlarını doğru olarak tahmin etmiştir. Bu adımlar sonrasında en düşük doğruluk oranı %96.73 ile Rastgele Orman algoritmasından elde edilmiştir.

Daha sonra ham ergen veri seti ile çalışmaya devam edilmiştir. Tablo 6'da görüldüğü üzere en düşük doğruluk oranı %70.97 ile Lojistik Regresyon algoritmasından, en yüksek doğruluk oranı ise %90.48 ile K-NN algoritmasından elde edilmiştir.

Tablo 6. Ergen ham veri seti ile algoritmaların performansları

	NB	LR	K-NN	RO
Doğruluk Oranı %	88.73	70.97	90.48	83.87
AUC	0.963	0.735	0.97	0.98
Hassasiyet %	86.43	73.68	85.62	78.26
F1 Skoru %	90.26	75.66	92.06	87.21

Ham veri ile yapılan değerlendirmeden sonra veri ön işleme ve parametre optimizasyonu ile modeller oluşturulup sonuçları değerlendirilmiştir.

Tablo 7. Veri ön işleme ve parametre optimizasyonu sonrası ergen veri setinde algoritmaların performansları

	NB	LR	K-NN	RO
Doğruluk Oranı %	95.74	96.94	97.73	100
AUC	0.985	0.99	0.98	1
Hassasiyet %	96.67	98.81	96.88	100
F1 Skoru %	96.69	98.1	98.63	100

Tablo 7' de görüldüğü üzere bu çalışmadaki yöntem kullanılarak ergenlik veri setinde tüm algoritmaların başarı oranlarında artış gerçekleşmiştir. Ayrıca ergenlik dönemi içinde algoritmalar arasında daha detaylı ve kapsamlı eğitim ve test süreçleri gerçekleştirilmiş ve başarılı sonuçlar elde edilmiştir. Bu adım sonrasında en başarılı algoritma, tüm verileri doğru şekilde etiketleyen Rastgele Orman algoritması olmuştur. Tüm algoritmalar %95 üzerinde başarı göstermiştir. En yüksek artış ise %25.97 ile Lojistik Regresyon algoritmasında gerçekleşmiştir.

Son olarak yetişkinlik dönemi veri seti ile devam edilmiştir. Tablo 8' de görüldüğü üzere en düşük başarı oranı 89.57 ile Lojistik Regresyon algoritmasından, en yüksek başarı oranı ise 98.58 ile K-NN algoritmasından elde edilmiştir.

Tablo 8. Yetişkin ham veri seti ile algoritmaların performansları

	NB	LR	K-NN	RO
Doğruluk Oranı %	96.44	89.57	98.58	94.31
AUC	0.994	0.961	0.99	0.99
Hassasiyet %	92.98	84.06	98.21	98.31
F1 Skoru %	93.20	86.68	97.33	91.14

Yetişkin ham veri ile yapılan değerlendirmeden sonra veri ön işleme ve parametre optimizasyonu adımları uygulandıktan sonra modeller oluşturulup sonuçları değerlendirilmiştir.

Tablo 9. Veri ön işleme ve parametre optimizasyonu sonrası yetişkin veri setinde algoritmaların performansları

	NB	LR	K-NN	RO
Doğruluk Oranı %	98.71	100	99.71	100
AUC	0.99	1	0.919	1
Hassasiyet %	97.33	100	99.08	100
F1 Skoru %	97.57	100	99.53	100

Tablo 9' da görüldüğü üzere bu çalışmadaki yöntem kullanılarak yetişkinlik veri setinde tüm algoritmaların başarı oranlarında önemli derecede artış gerçekleşmiştir. Ayrıca yetişkinlik dönemi içinde algoritmalar arasında daha detaylı ve kapsamlı eğitim ve test süreçleri gerçekleştirilmiş ve başarılı sonuçlar elde edilmiştir. Bu adım sonrasında en başarılı algoritmalar, tüm verileri doğru şekilde etiketleyen Rastgele Orman ve Lojistik Regresyon algoritmaları olmuştur. Tüm algoritmalar %98 üzerinde başarı göstermiştir. Tüm algoritmaların ham veri ile optimizasyon yapılmadan elde edilen sonuçları ile çalışmadaki yöntem kullanılarak elde edilen sonuçları karşılaştırılmıştır. Böylece üç veri seti içinde çalışma tamamlanmıştır.

IV. TARTIŞMA

Literatür incelendiğinde çalışmaların büyük bir bölümünde algoritmaların otizm teşhisindeki başarıları ham veri üzerinden karşılaştırılmıştır. Çalışmadaki yöntem ve bulgularda elde edilen sonuçlar incelendiğinde ise üç veri seti içinde algoritmaların başarı oranlarında artış gerçekleştiği görülmektedir. Ayrıca her algoritmanın artış oranlarındaki farklılıktan dolayı, ham veri ile optimizasyon yapılmadan elde edilen algoritmaların başarı sırası ile çalışmadaki adımlar uygulandıktan sonra algoritmaların başarı sıralamalarının değiştiği görülmektedir. Bu çalışmadaki yöntem ile makine öğrenimi algoritmalarının sınıflandırma kabiliyetlerinin daha detaylı incelenmesi gerektiğini göstermektedir.

Ayrıca çalışmadaki yöntemin, diğer çalışmalarda da incelenen hastalıkların tedavisine ve teşhis de kullanılacak algoritmanın seçilmesinde daha doğru kararlar alınmasında yardımcı olabileceği gözlemlenmiştir.

V. SONUÇLAR

Bu çalışmada çocukluk, ergenlik ve yetişkinlik dönemi veri setleri kullanılarak Otizm erken teşhisi için makine öğrenimi algoritmaları ile sınıflandırma yapılmıştır. Literatürdeki çalışmalara bakıldığında çalışmaların büyük bir kısmında algoritmaların otizm teşhisindeki başarı oranları karşılaştırılmıştır. Çalışmaların genelinde kullanılan algoritmalar bu çalışma kapmasında seçilmiştir. Tüm algoritmaların performansları otizm erken teşhisi için değerlendirilmiştir. Algoritmaların ham veri ile elde sonuçları ile çalışmadaki yöntem kullanıldıktan sonraki başarı oranları karşılaştırılmıştır.

Çocuk veri setinde %4 ile %11 değerleri arasında başarı artışı görülmüştür. En yüksek başarı değerini optimizasyon sonrasında Lojistik Regresyon algoritması vermiştir. Ergen veri setinde %7 ile % 25 arasında başarı artışı görülmektedir. En yüksek başarı oranı Random Forest algoritmasından elde edilmiştir. Yetişkin veri setinde %2 ile %11 arasında başarı artışı elde edilmiştir. En yüksek başarı oranları Random Forest ve Lojistik Regresyon algoritmalarından elde edilmiştir. Sonuç olarak, çalışmadaki yöntem kullanılarak algoritmaların üç veri seti üzerinde de başarılarının arttığı görülmüştür.

TEŞEKKÜR

Bilgi ve deneyimlerinden sürekli yararlandığım, çalışmanın her aşamasında yardımlarını esirgemeyen Doç. Dr. Hidayet TAKCI' ya çok teşekkür ederim.

KAYNAKLAR

[1] Metlek, S., & Kayaalp, K. (2020). Otistik Spektrum Bozukluğunun Makine Öğrenme Algoritmaları ile Tespiti. Zeki Sistemler Teori ve Uygulamaları Dergisi, 3(2), 60-68.

[2] Cho S, Liberman M, Ryant N, Cola M, Schultz RT, Parish-Morris J. (2019). Automatic Detection of Autism Spectrum Disorder in Children Using Acoustic and Text Features from Brief Natural Conversations. Proceedings of the Annual Conference of INTERSPEECH; (pp. 2513-2517), Graz, Austria.

[3] Alwidian, J., Elhassan, A., Ghnemat, R. (2020). Predicting autism spectrum disorder using machine learning technique. International Journal of Recent Technology and Engineering, 8(5), 4139-4143.

[4] Raj, S., & Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. Procedia Computer Science, 167, 994-1004.

[5] Mythili, M. S., & Shanavas, A. M. (2014). A study on Autism spectrum disorders using classification techniques. International Journal of Soft Computing and Engineering, 4(5), 88-91.

[6] Parikh, M. N., Li, H., He, L. (2019). Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data. Frontiers in computational neuroscience, 13, 9.

[7] Devika Varshini, G., Chinnaiyan, R. (2020). Optimized Machine Learning Classification Approaches for Prediction of Autism Spectrum Disorder. Ann Autism Dev Disord, 1(1), 1001

[8] Büyükoflaz, F. N. ve Öztürk, A. (2018). "Early autism diagnosis of children with machine learning algorithms,". Signal Processing and Communications Applications Conference (SIU), pp. 1-4, doi: 10.1109/SIU.2018.8404223.

[9] Abdullah, A. A., Rijal, S., Dash, S. R. (2019). Evaluation on Machine Learning Algorithms for Classification of Autism Spectrum Disorder (ASD). In Journal of Physics: Conference Series 1372 1 012052.

[10] Demirhan, A. (2018). Performance of machine learning methods in determining the autism spectrum disorder cases. Mugla Journal of Science and Technology, 4(1), 79-84.

[11] Anonim, "Autistic Spectrum Disorder Screening Data for Adolescent", https://archive.ics.uci.edu/ml/machine-learning-databases/00420/. (Erişim Tarihi: 12.02.2023).

[12] Anonim, "Autistic Spectrum Disorder Screening Data for Adult", https://archive.ics.uci.edu/ml/machine-learning-databases/00426/. (Erişim Tarihi: 12.02.2023).

[13] Anonim, "Autistic Spectrum Disorder Screening Data for children", https://archive.ics.uci.edu/ml/machine-learning-databases/00419/. (Erişim Tarihi: 12.02.2023).

[14] Kotsiantis, S. B., Kanellopoulos, D., Pintelas, P. E. (2006). Data preprocessing for supervised leaning. International journal of computer science, 1(2), 111-117.

[15] Jiang, L., Zhang, L., Li, C., Wu, J. (2018). A correlation-based feature weighting filter for naive Bayes. IEEE transactions on knowledge and data engineering, 31(2), 201-213.

[16] Manek, A. S., Shenoy, P. D., Mohan, M. C. (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World wide web, 20(2), 135-154.

[17] Atan, S. (2020). KNN, Naive Bayes ve Karar Ağacı Makine Öğrenme Algoritmaları, Bu Algoritmaların Sosyal Bilimlerde Kullanım İmkânları. SocArXiv. doi:10.31235/osf.io/8r5pu

[18] Raj, S., & Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. Procedia Computer Science, 167, 994-1004.

[19] Mezquita, Y., Alonso, R. S., Casado-Vara, R., Prieto, J., & Corchado, J. M. (2021). A review of k-nn algorithm based on classical and quantum machine learning. In Distributed Computing and Artificial Intelligence, Special Sessions, 17th International Conference (pp. 189-198). Springer International Publishing.

[20] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25, 197-227.

[21] Tan, P., Steinbach, M., Kumar, V. (2014). "Performance Measure" in Introduction to Data Mining, Pearson Education Limited (UK).

[22] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.