



**T.C. İSTANBUL T CARET
 NİVERSİTESİ**

FEN B LİMLERİ ENSTİT S 

**EEG S NYALLERİ İLE EPİLEPSİ KRİZİNİN TAHMİNLENMESİNDE
RASSAL ORMAN ALGORİTMASI İLE H PER PARAMETRE
OPTİMİZASYONUN UYGULANMASI**

Fatih Murathan YILMAZ

**Danışman
Dr.  ğr.  yesi Mustafa Cem KASAPBAŐI**

**Y KSEK LİSANS TEZİ
B LGİSAYAR M HENDİSLİĐİ ANAB LİM DALI
İSTANBUL - 2021**

KABUL VE ONAY SAYFASI

Fatih Murathan YILMAZ tarafından hazırlanan "**EEG SİNYALLERİ İLE EPİLEPSİ KRİZİNİN TAHMİNLENMESİNDE RASSAL ORMAN ALGORİTMASI İLE HİPER PARAMETRE OPTİMİZASYONUN UYGULANMASI**" adlı tez çalışması 01/03/2021 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman	Dr. Öğr. Üyesi Mustafa Cem KASAPBAŞI İstanbul Ticaret Üniversitesi
Jüri Üyesi	Prof. Dr. Serhat ÖZEKES Üsküdar Üniversitesi
Jüri Üyesi	Dr. Öğr. Üyesi Feyza Merve HAFIZOĞLU İstanbul Ticaret Üniversitesi

Onay Tarihi : 15.03.2021

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsünün 15.03.2021 tarih ve 2021/308 numaralı Yönetim Kurulu Kararının 1. maddesi gereğince, ders yüklerini ve tez yükümlülüğünü yerine getirdiği belirlenen "Fatih Murathan YILMAZ" (TC:12535488444) adlı öğrencinin mezun olmasına oy birliği ile karar verilmiştir.

Prof. Dr. Necip ŞİMŞEK
Enstitü Müdür

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmada,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

15.03.2021

Fatih Murathan YILMAZ

İÇİNDEKİLER

	Sayfa
İÇİNDEKİLER.....	i
ÖZET	iii
ABSTRACT	iv
TEŞEKKÜR.....	v
ŞEKİLLER DİZİNİ	vi
ÇİZELGELER DİZİNİ	viii
SİMGELER VE KISALTMALAR DİZİNİ	ix
1. GİRİŞ.....	1
2. LİTERATÜR ÖZETİ.....	2
3. MATERYAL VE YÖNTEM	4
3.1. Veri Seti	4
3.2. Ağaç Temelli Yaklaşım Ve Rassal Orman Algoritması	5
3.2.1. Sınıflandırma ve regresyon ağaçları	5
3.2.2. Sınıflandırma ve regresyon ağaçlarının kolektif metotlara evrimi.....	8
3.2.3. Rassal orman algoritması.....	9
3.3. Hiper Parametre Optimizasyonu Ve Temel Bileşenler Analizi	11
3.3.1. Hiper parametre optimizasyonu	11
3.3.2. Temel bileşenler analizi	13
3.4. Makine Öğrenmesi Tekniğinin Uygulanması	21
3.4.1. Python programlama dili ve özellikleri	22
3.4.2. Yardımcı yöntemler ve başarımlar metrikleri	22
3.4.2.1. K katlı çapraz doğrulama	23
3.4.2.2. Karmaşıklık matrisi	24
3.4.2.3. Doğruluk	25
3.4.2.4. Hassaslık	25
3.4.2.5. Özgüllük	26
3.4.3. Rassal orman algoritmasının uygulanması	26
4. ARAŞTIRMA BULGULARI VE TARTIŞMA.....	28
5. SONUÇ VE ÖNERİLER.....	36
KAYNAKLAR	37
ÖZGEÇMİŞ.....	43

ÖZET

Yüksek Lisans Tezi

EEG SİNYALLERİ İLE EPİLEPSİ KRİZİNİN TAHMİNLENMESİNDE RASSAL ORMAN ALGORİTMASI İLE HİPER PARAMETRE OPTİMİZASYONUN UYGULANMASI

Fatih Murathan YILMAZ

**İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı**

**Danışman: Dr. Öğr. Üyesi Mustafa Cem KASAPBAŞI
2021, 43 Sayfa**

Dünyadaki 50 milyondan fazla kişiden oluşan tüm nüfusun yaklaşık %1'i epilepsi ve epileptik nöbetlerden etkilenmektedir. Epileptik nöbetler, beynin elektriksel aktivitesindeki bir rahatsızlıktan kaynaklanır. Epilepsi nöbetinin saptanması genellikle elektroensefalografik (EEG) sinyal incelendikten sonra uzman görüşü tarafından gerçekleştirilir. Bu manuel bir süreçtir ve büyük ölçüde doktorun uzmanlığına dayanır. Bu nedenle, doktorların daha az hatayla teşhis koymasına yardımcı olmak için otomatik tanı veya yardım sistemleri gereklidir. Bu çalışmada, epileptik nöbetlerin varlığını sınıflandırmak için iyi bilinen bir veri kümesi kullanılmıştır. Veri setinin farklı konfigürasyonları literatürde bir kısmı Lojistik Regresyon, Dalgacık yöntemi, Karar Ağacı, Destek Vektör Makinesi, Yoğun Sinir Ağları, vb. birçok veri madenciliği ve makine öğrenme algoritması ile incelenmiştir. İyi tanı beklentisini karşılamak için Rassal Orman kullanılarak sınıflandırma modeli geliştirilmiştir ve sonuçlar aynı veri seti üzerinde incelenen farklı yöntemlerle karşılaştırılmıştır. Çalışılan deneylerin bazı vakalarında %99,78 oranında doğruluk, %99,95 özgüllük ve %99,61 hassasiyet elde edilmiştir ve sonuçlar modelinin başarılı şekilde sınıflandırdığını göstermektedir.

Anahtar Kelimeler: EEG, Hiper Parametre Optimizasyonu, Makine öğrenmesi

ABSTRACT

M.Sc. Thesis

IMPLEMENTATION OF HYPER PARAMETER OPTIMIZATION WITH RANDOM FOREST ALGORITHM FOR THE ESTIMATION OF THE EPILEPTIC SEIZURES WITH EEG SIGNALS

Fatih Murathan YILMAZ

**Istanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Computer Engineering**

**Supervisor: Asst. Prof. Dr. Mustafa Cem KASAPBAŞI
2021, 43 pages**

About %1 of the whole population of the world which constitutes more than 50 million people are affected by epilepsy and epileptic seizures. Epileptic seizures are caused by a disturbance in the electrical activity of the brain. Detecting epileptic seizure is generally carried out by the expert opinion after examining the electroencephalographic (EEG) signal. This is a manual process and heavily relies on the expertise of the physician. Therefore automated diagnosis or aiding systems are required to assist physicians to diagnose with fewer errors. In this study, a well known dataset is used for classifying the existence of epileptic seizures. Different configurations of the data set have been studied with many data mining and machine learning algorithms in the literature, some of which are Logistic Regression, Wavelet Method, Decision Tree, Support Vector Machine, Dense Neural networks, etc.. In this study, a classification model was developed by using Random Forest to meet the good diagnosis expectation, and results were compared with different methods studied on the same data set. In some cases of the studied experiments above 99,78 percent of accuracy, 99,95% specificity, and 99,61% sensitivity are obtained, indicating a good sign of classification model.

Keywords: EEG, Hyper Parameter Optimization, Machine Learning.

TEŞEKKÜR

Bu araştırma için beni yönlendiren, karşılaştığım zorlukları bilgi ve tecrübesi ile aşmamda yardımcı olan değerli Danışman Hocam Dr. Mustafa Cem KASAPBAŞI'na teşekkürlerimi sunarım. Tezimin her aşamasında beni yalnız bırakmayan babam Hikmet YILMAZ, annem Saime YILMAZ ve çekirdek ailemin diğer üyelerine sonsuz sevgi ve saygılarımı sunarım.

Fatih Murathan YILMAZ
İSTANBUL, 2021

ŞEKİLLER

	Sayfa
Şekil 3.1. Z ve O dizinindeki verilerin oluşturulmasında kullanılan beyin elektrot haritası.....	4
Şekil 3.2. 2 boyutlu 4 sınıfın şekilsel gösterimi.....	6
Şekil 3.3. Kurallı olarak ayrılmış sınıflar ve alt uzaylar	7
Şekil 3.4. S1 uzayını sınıflandıran ikili ağaç modeli	8
Şekil 3.5. Rassal orman algoritması karar verme mekanizması şematik gösterimi.....	10
Şekil 3.6. Grid arama algoritması akış diyagramı.....	12
Şekil 3.7. Temel bileşen analizi ile 2 boyutlu uzayda veri dağılımının gösterimi	18
Şekil 3.8. Üçüncü dereceden PCA uygulaması	19
Şekil 3.9. Beşinci dereceden PCA uygulaması.....	20
Şekil 3.10. Yedinci dereceden PCA uygulaması.....	20
Şekil 3.11. Dokuzuncu dereceden PCA uygulaması.....	21
Şekil 3.12. K katlı çapraz doğrulama yönteminin şekilsel gösterimi.....	23
Şekil 3.13. Karmaşıklık matrisi yerleşim gösterimi	24
Şekil 3.14. Oluşan ağaçlardan alınan örnek kesit.....	27

ÇİZELGELER

	Sayfa
Çizelge 3.1. Kuvvet uygulamaları sonrası sınıfların varyans değerleri	19
Çizelge 4.1. İdeal senaryoya ait 5 katlı çapraz doğrulama sonuçları.....	28
Çizelge 4.2. İdeal senaryo çapraz doğrulamalar sonucu oluşan kümülatif karmaşıklık matrisi.....	28
Çizelge 4.3. İdeal senaryonun diğer olası senaryolar ile başarımların karşılaştırılması.	29
Çizelge 4.4. Aynı veri setinin kullanıldığı önceki çalışmalar ve yaptığımız çalışmanın karşılaştırılması.	30

SİMGELER VE KISALTMALAR

ANN	Yapay Sinir Ağları
ApEn	Yaklaşık Entropi
CART	Sınıflandırma ve Regresyon Ağaçları
EEG	Elektroensefalografi
FN	Yanlış Negatif
FP	Yanlış Pozitif
ICA	Bağımsız Bileşenler Analizi
LDA	Lineer Diskriminant Analizi
MLP	Çok Katmanlı Sınıflayıcı
Numpy	Python programlama dilinde kullanılan Numpy kütüphanesi
PCA	Temel Bileşenler Analizi
Pd	Python programlama dilinde kullanılan Pandas kütüphanesi
RBF	Radyal Temeli Fonksiyon
Sklearn	Python programlama dilinde kullanılan Scikit-learn kütüphanesi
SVM	Destek Vektör Makineleri
TN	Doğru Negatif
TP	Doğru Pozitif

1. GİRİŞ

Epilepsi, dünya genelinde yaklaşık 50 milyon insanın sahip olduğu nörolojik bir hastalıktır (Kandel vd., 2000). Epilepsinin en yaygın özelliği hastaların geçirdiği krizlerdir. Bu krizlerin kaynağı beyinde bulunan nöron gruplarınca yapılan kontrolsüz elektrik boşaltımıdır (Acharya vd.,2018). Krizlere sebep olan bu elektrik boşalımının makineler yardımı ile gözlemlenmesi konusu 1970'li yıllarda başlamıştır (Viglione ve Walsh,1975). Fakat bu konu hakkında algoritmaların gelişmesi ve konunun fiziksel olarak çözümlenebilir bir probleme dönüşmesi yaklaşık olarak 30 sene sürmüştür (Litt ve Echauz,2002). Elektroensefalogram (EEG) teknolojisi beyin tarafından üretilen elektriksel potansiyeli kayıt etmeye yarar. EEG, beyindeki elektriksel anomalileri kaydetmek için yaygın olarak kullanılır. Nörolojik hastalıkların tanısında klinik olarak kullanılan en önemli cihazlardan biridir (Kannathal vd. 2005).

Epilepsi krizlerinin belirlenmesi için kullanılan EEG sinyalleri uzman kişilerin bu sinyallerin grafiklerinin gözle incelenmesi ve tecrübe ile karar verilmesi sürecine dayanmakta olup, uzman kişinin bilgi ve tecrübesi doğru kararlar verilmesinde önemli faktör olmaktadır. Bu sebeple alan uzmanlarına tahminleme sürecinde destek olabilecek sistemlere ihtiyaç duyulmakta, uzman hekimlere ulaşmayan hastalar için otonom karar verebilen gömülü sistemler için algoritmaların da geliştirilmesine ihtiyaç olduğu düşünülmektedir. Bu çalışmada bu ihtiyacı karşılamak için Rassal Orman kullanılarak bir sınıflandırma modeli geliştirilmiş ve aynı veri seti üzerinde çalışılan farklı yöntemlerle kıyaslanması gerçekleştirilmiştir.

2. LİTERATÜR ÖZETİ

Otomatik kriz tanısı konusunda 1995 yılında EEG sinyallerinden belirli özelliklerin dalgacık dönüşümü yardımıyla çıkartımı yapılmış ve elde edilen dönüştürülmüş veri yapay sinir ağları (ANN) yardımıyla sınıflandırılmış ve kayda değer sonuçlar elde edilmiştir (Kalayci ve Ozdamar,1995).

Nigam ve Graupe (2004), çok katmanlı eğrisel ön işleme filtresi kullanarak iki özellik çıkartımı olarak pik değerleri genliği ve pik frekansı çıkartımını yapmıştır ve bu çıkartımı yapay sinir ağlarında işlemişlerdir.

Jahankhani vd. (2006) EEG sinyallerini dalgacık dönüşümü kullanarak farklı alt bantlara ayırmış ve buradan dalgacık katsayıları elde etmiş, elde ettiği veriyi radyal temelli fonksiyon (RBF) ve çok katmanlı sınıflayıcı (MLP) kullanarak sınıflandırmıştır.

Subasi (2005, 2006, 2007) EEG sinyallerini ayırık dalgacık dönüşümü (DWT) kullanarak zaman-frekans domeninde ayırtmıştır.

Tzallas vd. (2007) zaman-frekans analizi metodunu kullanarak EEG sinyallerinden özellik çıkarımını yapmışlar ve elde ettikleri özellikleri yapay sinir ağı yardımıyla sınıflandıracak modeli oluşturmuşlardır.

Srinivasan vd. (2007) yaklaşık entropi (ApEn) temelli yinelenen sinir ağlarının bir türü olan Elman sinir ağı modeli ile sınıflandırma yapan modeli kurgulamışlardır.

Guo vd. (2009) EEG sinyallerinde bulunan farklı frekans bantlarıyla ilişkili bağlı enerjiden, göreceli dalgacık enerjisi yöntemi ile özellik çıkartımı yaparak elde ettikleri özellikleri yapay sinir ağları yardımıyla sınıflandırma yapmak için kullanmışlardır.

Subasi ve Gursoy (2010) temel bileşenler analizi, bağımsız bileşenler analizi ve doğrusal diskriminant analizi yöntemlerini kullanarak sınıflandırma yapmışlardır.

Fernandez-Blanco vd. (2012) zaman-frekans analizi yöntemini daha önce çalışılmış metotlardan farklı olarak yıldız graf topoloji indisleri ile ele almışlardır.

Song ve Zhang (2013) permütasyon entropi, örnek entropi ve Hurst üssünü kullanarak farklı frekans bantlarından özellik çıkartımı yapmışlar ve genetik algoritma ile yaptıkları sınıflandırma yöntemiyle birlikte kullanmışlardır.

Kaya vd. (2014) tek boyutlu yerel ikili örüntü kullanarak EEG sinyallerinden özellik çıkartımı yaparak elde ettikleri özellikleri bayes ağı ile sınıflandırma yapmak için kullanmışlardır.

Fu vd. (2014) zaman-frekans analizi kapsamında Hilbert-Huang dönüşümüyle EEG sinyallerinden kalıplar elde etmişler ve bu yolla özellik çıkarımı yapmışlardır.

Chen (2014) ikili ağaç karmaşık dalgacık dönüşümü metodu ile sınıflandırma yapmıştır.

Zamir (2016) örnek tabanlı öğrenme algoritmaları ve C4.5 algoritması ile sınıflandırma yapmıştır.

Hassan vd. (2016) ayarlanabilir Q katsayılı dalgacık dönüşümü ile özellik çıkarımını yaparak elde ettikleri özellikleri torbalama yöntemi ile sınıflandırmada kullanmışlardır.

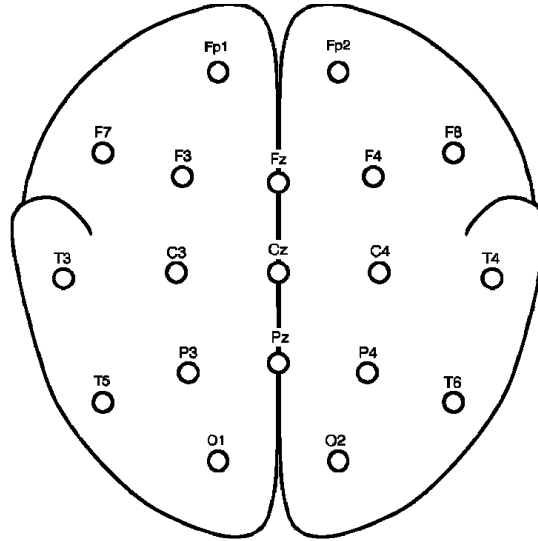
Hussein vd. (2018) EEG sinyalleri ile sınıflandırma yapmak için derin öğrenme yöntemini tercih etmişlerdir.

3. MATERYAL VE YÖNTEM

Bu bölümde araştırmanın konusu olan veri, veri ile çalışacak olan algoritma ve özellikleri, algoritmanın başarımını yükseltmek adına kullanılan yardımcı metotlar ve algoritmanın uygulanma yöntemi ele alınacaktır.

3.1. Veri Seti

Bu çalışmanın konusu olan EEG veri seti, Bonn Üniversitesinde hazırlanmıştır (Andrzejak vd., 2001). Bu veri seti 5 farklı dizine ayrılmış ve her bir dizinde 100 adet kayıt bulunmaktadır. Z ve O olarak etiketlenmiş dizinlerde sağlıklı ve uyanık halde bulunan deneklerden elde edilen kayıtlar mevcuttur. Z dizinindeki kayıtlarda deneğin gözleri açık konumunda iken O dizinindeki kayıtlarda deneğin gözleri kapalı konumdadır. Z ve O dizinlerindeki veriler, Şekil 3.1’de gösterilen haritadaki dışsal elektrotlar yardımı ile oluşturulmuştur. Elektrotların konumları ve isimleri bulundukları anatomik bölgeden türetilmiştir.



Şekil 3.1. Z ve O dizinindeki verilerin oluşturulmasında kullanılan beyin elektrot haritası (Andrzejak vd., 2001)

N, F ve S olarak etiketlenen dizinlerdeki kayıtlar epilepsi hastalarından elde edilmiştir. N ve F dizinleri bu epilepsi hastası deneklerin kriz geçirmediği anlardaki kayıtlardan oluşmaktadır. N dizininin kayıtları beyin epilepsi

krizlerine sebep olan elektriksel aktivitenin görülmediği bölgeden diğer bir deyiş ile epileptik olmayan bölgeden alınmıştır. F dizininde ise epileptik bölgeden elde edilmiş kayıtlar bulunmaktadır. S dizininde sadece epilepsi kriz esnasında elde edilen kayıtlar bulunmaktadır. Dizinlerdeki bütün kayıtlar 23.6 saniye boyunca 173.61 Hz örnekleme frekansı ile kaydedilen tek kanallı EEG sinyallerinden oluşmaktadır. Veriler zaman serilerine karşılık olarak 4097 veri noktasında örneklenebilir. Rochester Teknoloji Enstitüsü 4097 veri noktasını 23 adet yığına ayırmıştır. Yığınların her biri, 1 saniye için 178 veri noktasını veri noktasını göstermektedir (UCI Rep. 2020). Bu sayede makine öğrenmesi uygulamalarının eğitimi için mevcut veri sayısı 22 kat artırılmış olmaktadır. Son haliyle her bir dizin 2300 adet kayıt içermektedir ve toplam veri sayısı 11500 adete çıkmıştır.

3.2. Ağaç Temelli Yaklaşım ve Rassal Orman Algoritması

Morgan ve Sonquist (1963) ağaç temelli yaklaşımı ortaya çıkarmışlardır. Breiman vd. (1984) ağaç temelli yaklaşımı, kapsamını genişletip tahminleme mekanizmasını güçlendirerek sınıflandırma ve regresyon ağaçlarını (CART) geliştirmişlerdir.

3.2.1. Sınıflandırma ve regresyon ağaçları

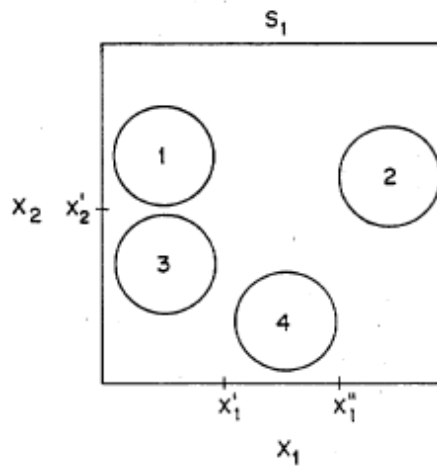
Sınıflandırma ve regresyon ağaçları, veriyi istatistiksel olarak en saf ve güvenilir şekilde ayırmayı hedefler. Bu ağaçlarda amaç genellemeler yaparak en büyük sınıf ayrımını yapabilmektir. Diğer bir deyiş ile veriyi sınıflandırabilmek için muhtemel en iyi bölünmeler hesaplanır. Bu bölünmeler belirli bazı metotlar ile bulunabilmektedir. En yaygın kullanılan bölünme metotları Gini ve Entropi yöntemleridir. Denklem 3.1'de Gini metodunun formülü ve Denklem 3.2'de Entropi metodunun formülü verilmiştir.

$$Gini = 1 - \sum_k^J (p_{ik})^2 \quad (3.1)$$

$$E(S) = - \sum_k (p_{ik}) \log_2 (p_{ik}) \quad (3.2)$$

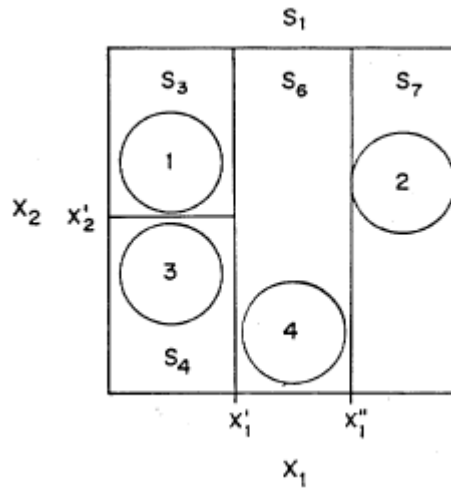
Denklemlerde verilen J değeri sınıf alabileceği özellikleri gösterirken, k değeri ise özellik iterasyon numarası olarak verilmiştir. P değeri sınıflandırıcı özelliğin o sınıf içindeki toplam özellik sayısına oranıdır. Formüller incelendiğinde görülecektir ki gini katsayısı 0 ile 0.5 arasında değer alan bir indistir. Gini katsayısı 0.5 değerine ne kadar yakınsa saflık o kadar az demektir. Bundan dolayı en iyi bölünme için düğüm kuralı oluşturmak için gini değerinin en düşük olanı tercih edilmektedir. Entropide de benzer durumlar söz konusudur. Entropi değeri ne kadar yüksek ise örnek uzay kümesinde düzensizlik o kadar fazladır çıkarımı yapılır. Gini indeksi veya entropi içinde ortak olan ve ağaç dallarının oluşmasını sağlayan ana kural bu değerlerinin en düşük seviyede olması yani saflığın fazla olmasıdır. Gini yöntemi entropi yöntemine göre daha yoğun tercih edilmektedir. Bunun sebebi entropi formülünde kullanılan logaritmadan dolayıdır. İşlem karmaşıklığını arttırdığından dolayı Entropi yöntemi, büyük sayılar içeren veri setleri için işlemsel yoğunluk olarak süre ve işlem miktarını arttıracığından dolayı Gini yöntemine göre daha az tercih edilmektedir.

Grajski vd. (1986) yaptıkları çalışma ile sınıflandırma ve regresyon ağaçlarının temelde nasıl çalıştığını 2 boyutlu ve 4 farklı sınıftan oluşan bir örnek üzerinde göstermişlerdir (Şekil 3.2).



Şekil 3.2. 2 boyutlu 4 sınıfın şekilsel gösterimi (Grajski vd., 1986)

Sınıflandırma ve regresyon ağaçlarında ana amaç her adımda en fazla sınıfı ayırmaktır. Şekil 3.2’de görülebileceği üzere x_1 ve x_2 doğrularınca sınırlandırılmış alanda 4 sınıftan oluşan S_1 uzayı bulunmaktadır. Sınıf ayrımını maksimum yapmak adına x_1 doğrusu üzerinde x' noktasından ilk ayırım yapılır. Böylece S_1 uzayından S_2 alt uzayı oluşturulmuş olunur ve bu yol ile 1 ve 3 sınıfları 2 ve 4 sınıflarından ayrı uzaya alınır. İkinci iterasyon olarak x_2 üzerinde bulunan x' noktasından uzay bölünmesi gerçekleştirilir. Bu şekilde S_2 uzayı S_3 ve S_4 alt uzaylarına ayrılmış olur. Diğer bir deyiş ile 1 ve 3 numaralı sınıflar birbirlerinden ayrılmış olur. Sınıfları kendi gruplarına ayırdığımız son kural x_1 üzerinde bulunan x'' noktasıdır. 2 ve 4 numaralı sınıfları içinde bulunduran S_5 uzayı x'' noktasından bölünerek S_6 ve S_7 alt uzayları elde edilmiştir. Tüm bu bölünmeler için yapılan her bir işlem ağacın bir düğümünü işaret etmektedir. Şeklin bölünmüş hali Şekil 3.3’de verilmiştir.

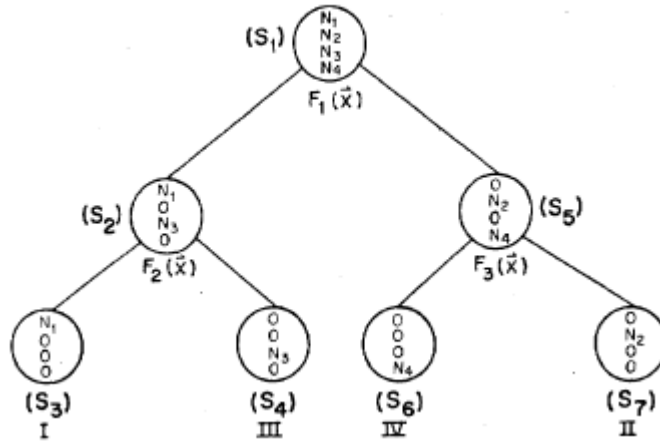


Şekil 3.3. Kurallı olarak ayrılmış sınıflar ve alt uzaylar (Grajski vd., 1986)

Şekil 3.3’de yapılan bu sınıflama probleminin çözümü Sınıflandırma ve Regresyon Ağaçlarının algoritmasını oluşturmaktadır. S_1 uzayında yüksek düzensizlik veya diğer bir ifade ile saf olmayan halde bulunan sınıflar düzensizliği azaltılarak yani saflaştırılarak alt uzaylara alınmışlardır.

Ağaçlarda kök düğümden uzaklaştıkça düzensizliğin azalması beklenir. Diğer bir deyiş ile daha alt seviyelerde bulunan düğümlerdeki düzensizlik veya saf olmama durumu daha azdır.

Şekil 3.4’de S_1 uzayını en doğru şekilde sınıflandıran ağacın şeması verilmiştir. Şekilden de görülebileceği üzere kök düğümde tüm sınıflar bir aradadır ve bu düğüm S_1 uzayını temsil etmektedir. İlk ayrımı yapan kural uygulandığında S_1 uzayı, S_2 ve S_5 alt uzaylarına ayrılmıştır. S_1 uzayı, İkinci ve üçüncü bölünmeler ile 1 numaralı sınıfı temsil etmekte olan S_3 alt uzayı, 2 numaralı sınıfı temsil etmekte olan S_7 alt uzayı, 3 numaralı sınıfı temsil etmekte olan S_4 alt uzayı, 4 numaralı sınıfı temsil etmekte olan S_6 alt uzayına ayrılmıştır. Burada dikkat edilmesi gereken en temel nokta en düşük seviyedeki alt uzaylarda entropi ve gini değerleri 0 olarak bulunmaktadır ve daha üst seviyelere gittikçe düzensizlik ve gini değeri artış gösterir.



Şekil 3.4. S_1 uzayını sınıflandıran ikili ağaç modeli (Grajski vd., 1986)

3.2.2. Sınıflandırma ve regresyon ağaçlarının kolektif metotlara evrimi

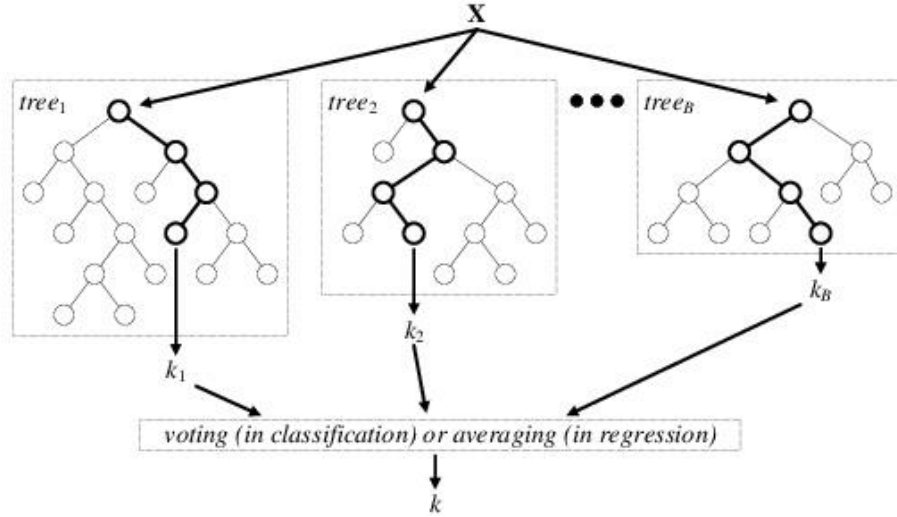
Sınıflandırma ve regresyon problemlerinde kullanılan karar ağacı yaklaşımını geliştiren Breiman (1996) torbalama adını verdiği yöntem ile birden fazla ağacın birlikte daha isabetli kararlar verebileceğini ispatlamıştır. Torbalama yöntemi, veri setinden rastgele olarak seçilen sınıflandırıcılardan bir ağaç kümesi kurgulamak üzerine kuruludur. Ho (1998) rassal alt uzaylar metodu ile rastgele

seçilmiş alt veri kümeleri ile her bir ağacı dahil olduğu uzayın özellikleri ile geliştirmiştir. Rassal alt uzay metodu, torbalama yönteminden farklı olarak her bir ağaç için farklı sayıda sınıflandırıcı özelliği ağaçları yaratırken kullanır.

Torbalama metodu ile yapılan deneysel çalışmalarda doğruluk oranının diğer metotlara göre daha yüksek olduğu görülmüş ve ağaç topluluklarının birlikte verdiği kararların tek bir karar ağacına göre oylama yöntemi sayesinde korelasyonu daha yüksek ağaçların karar vermesinden dolayı daha başarılı sınıflandırma yaptığı ortaya konulmuştur. Ağaç gruplarının sağladığı bu başarı ve araştırmacıların popüler olan kolektif ağaç yapısının sınıflandırma problemleri üzerine çalışmaları sonraki araştırmalarda Rassal Orman algoritmasının ortaya çıkmasına ortam hazırlamıştır.

3.2.3. Rassal orman algoritması

Breiman (2001) Rassal Orman algoritmasını tanımladığı çalışmasında, standart ağaç düğümü için en iyi bölünme belirlenirken tüm veri seti içerisindeki en iyi tahminleyici seçilmesi kuralını rassallık kuramıyla uyarlayarak her bir ağaç düğümü için en iyi bölünme ön yüklenme için hazırlanmış alt veri setinin içinden seçilecek rastgele bir özellik olması olarak geliştirmiştir. Çok sayıda ağacın oluşmasının ardından orman sınıflamaya uygun hale gelmiştir ve tekil olarak her biri kendi kurallarına uygun olan sınıfa oy verir. Rassal Orman algoritması sınıflandırma problemlerinde çoğunluğun haklılığı prensibine göre sınıflandırma yapar. Yani en çok oyu alan sınıf Rassal Ormanın tahminini işaret etmektedir. Regresyon problemlerinde ise Rassal Ormanın tahmini ormanda bulunan bütün ağaçların ortalaması alınarak bulunur. Şekil 3.5’de görülen diyagramda Rassal Orman algoritmasında ağaçların yapısı ve karar verme mekanizması gösterilmektedir.



Şekil 3.5. Rassal orman algoritması karar verme mekanizması şematik gösterimi (Suparwito, 2019)

Olasılık kuramının büyük sayılar teoremine göre aynı deney defalarca tekrarlandığında elde edilen sonuç beklenen değere yakınsar (Dekking vd.,2005). Bu teoremden yola çıkarak Rassal Orman algoritması için büyük sayılar asla aşırı öğrenmiş olmaz yorumu yapılabilmektedir. Rassal Orman algoritmasının temel diğer özelliklerinden biri de bir karar ağacı koleksiyonu olmasına rağmen tekil olarak hiçbir ağaç budanmaz. Tamamen rastgele seçilen sınıflandırıcı özelliklerden ve alt veri kümelerinden kurulan bu tümleşik ağaç sistemi diğer algoritmalarda ihtiyaç duyulduğu gibi özellik çıkarımına da ihtiyaç duymamaktadır. Sınıflandırma ağaçlarının birlikte karar vermesi prensibine dayanan Rassal Orman algoritması, deneylerde ortaya koyulduğu üzere tekil olarak ağaçların sağladığı doğruluk oranı ortalamasının üzerinde doğruluk oranlarına ulaşabilmektedir.

Bu algoritmanın hiper parametre optimizasyonu yapılabilecek noktaları karar ağaçlarının sayısı, ağaçların dallarının minimum ve maksimumları, düğümlerdeki bölünmeleri belirleyecek ayırma kriteri, kullanılacak sınıflandırıcı özellik adedinin maksimum ve minimumları ve eğer entropi temelli bir ağaç yapısı kurgulanacak ise bölünme ve dalların oluşmasında kullanılacak saflık değer sabitleri olarak ele alınabilir.

3.3. Hiper Parametre Optimizasyonu Ve Temel Bileşenler Analizi

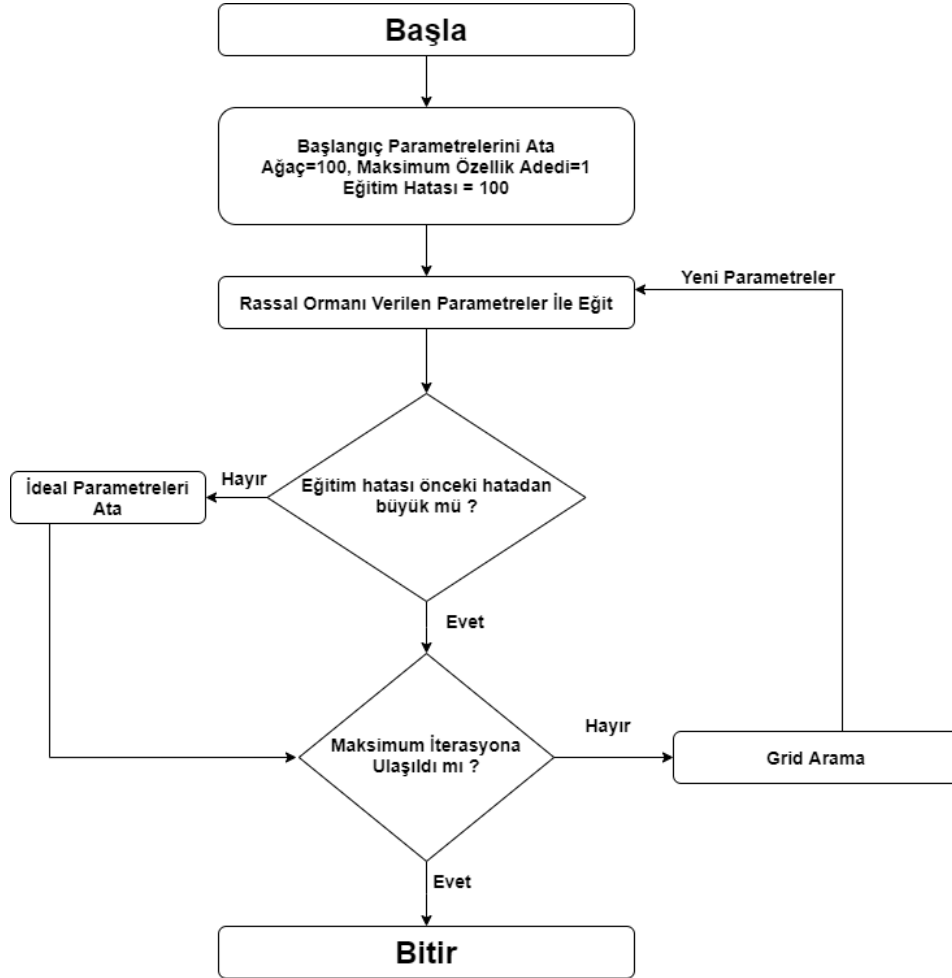
Sınıflandırma problemlerinde algoritma seçiminin problemin çözümü için yeterli olmadığı durumlarda ön işleme ve hiper parametre optimizasyon teknikleri uygulanmaktadır.

3.3.1 Hiper parametre optimizasyonu

Hiper parametre optimizasyonu, makine öğrenmesi problemlerinde yaygın olarak kullanılan popüler araştırma konularından biridir (Bergstra vd.,2011). Temelde tahmin ve sınıflandırmaların daha isabetli, daha hassas veya özellik seçimlerinin daha başarılı yapılabilmesi için algoritmalara ait parametrelerin probleme uygun hale getirilmesini esas almaktadır. Hiper parametre optimizasyonunu ile başarılı sonuçlar elde etmek için öncelikle probleme en uygun algoritmanın seçilmesi önemli bir yer tutar (Bergstra ve Bengio,2011). Daha önce bu konuda yapılmış çalışmalardan da görülebileceği üzere Rassal Orman gibi algoritmalarda ağaç sayısı tahminleyici başarısını büyük ölçüde etkilemektedir (Tantithamthavorn vd.,2016). Jiang vd. (2008) ile Tosun ve Bener (2009), yaptıkları çalışmalarda Rassal Orman algoritmasının varsayılan değerler ile uygulandığında sonuçların ideal seviyenin altında kaldıklarını görmüşlerdir. Tüm bunlardan çıkarımla bu çalışmada grid arama algoritması uygulanmıştır. Grid arama algoritması temelde hiper parametre uzayından elde edilen alt kümelerde iteratif olarak çalışan kapsamlı bir arama algoritmasıdır (Syarif vd.,2016). Grid aramada metot parametreler için başlangıç ve bitiş değerlerinin verilmesi ve başlangıç bitiş değeri arası kaç adımda tamamlanacak tanımlanması gereklidir. Üç farklı ölçek kullanılabilir. Bunlar lineer ölçek, kareli ölçek ve logaritmik ölçektir. Bu çalışmada ağaç sayısı ve sınıflandırıcı özellikler için grid arama ile hiper parametre optimizasyonu lineer ölçek ile uygulanmıştır. Kim (1997), grid arama algoritmasının uygulanması için ideal test noktalarının belirlenmesinde toplam işlem sayısının önemli bir kriter olduğunu ortaya koymuştur. Denklem 3.3'te toplam işlem sayısının formülü verilmiştir.

$$N(n, r) = r * n^m \quad (3.3)$$

Denklemden N toplam işlem adedini temsil ederken, r iterasyon sayısını, n özellik adedini ve m toplam parametre değer uzayını temsil eder. Bu çalışma için 178 özellik olduğundan dolayı r değeri 178 olarak alınmıştır. Toplam işlem adedinin büyük sayılara çıkması istenmediğinden dolayı iterasyon ve özellik sayısı parametre uzayının genişliği bu kısıt çerçevesinde belirlenmiştir. Ağaç sayısı 100 adetden başlamak üzere 8000 adet ağaca kadar ve sınıflandırıcı özellik sayısı 1'den başlamak üzere 178 adete kadar iteratif olarak denenmiştir. Şekil 3.6'da grid arama algoritmasının akış diyagramı verilmiştir. Tüm bu iteratif parametre denemelerinin sonucu olarak en ideal hiper parametrelerin 2000 ağaç ve 15 adet özellik olduğu görülmüştür.



Şekil 3.6. Grid arama algoritması akış diyagramı

3.3.2 Temel bileşenler analizi

Veri setinde bulunan tüm sınıflandırıcı özellikler veri setinin sınıflandırılmasında aynı oranda önem teşkil etmez. Varyansı yüksek olan ve diğer özellikler ile korelasyonu düşük olan özellikler veriyi sınıflandırmak için daha başarılı sonuçlar verirler (Hotelling, 1933). Buradan çıkarımla veriyi sınıflandırmaya uygun hale getirmek için Temel Bileşenler Analizi (PCA) yöntemi uygulanır. PCA algoritması özellik sayısı fazla olan veri setlerinde veri boyutunu azaltmak için kullanılan bir yöntemdir (Jolliffe, 2003). PCA gerek sınıflandırmak gerek ise özellik çıkarmak için kullanılabilen bir yöntemdir. Yöntemi gereği korelasyonu yüksek özellikleri elimine ederek varyans değeri yüksek ve sınıflandırmaya uygun özelliklerden bir veri izdüşüm alt uzayı oluşturmaktır. Bu yolla boyutu azaltılmış olan veri grubu ana veri setini tam olarak temsil edebilmektedir.

Algoritma, öncelikli olarak veri setinin varyansını hesaplar. Varyans hesaplaması için ilk adım tüm özelliklerin içinde bulunduğu kovaryans matrisini oluşturmaktır. Kovaryans matrisi oluşturulduktan sonra bu matrise ait özdeğer ve özvektörler bulunur. Hesaplanan özdeğerlerden büyük olanlar, küçük olanlara göre veri setini daha iyi temsil edebilecek grubu ifade eder. Özdeğeri en küçük olan özellik diğer özelliklere doğrusal olarak bağımlı bir özellik olduğu anlamını taşır. Varyansı veren formül Denklem 3.4'te verilmiştir.

$$var(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{n-1} \quad (3.4)$$

Denklemden yer alan n değeri özellik sayısını gösterirken $\bar{X}$ ifadesi ortalama vektörünü gösterir. Önceki kısımda açıklanan PCA'nın genel çalışma prensibi içerisinde yer alan kovaryans matrisini oluşturabilmek için özelliklerin kovaryansının bulunması gerekir. Denklem 3.5'de kovaryans formülü verilmiştir.

$$cov(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \quad (3.5)$$

Kovaryans ifadesi iki özellik arasındaki korelasyonu gösteren ifadedir. Kovaryans pozitif ise iki özelliğin birlikte arttığını ve azaldığını gösterir. Aynı şekilde negatif kovaryans aralarındaki ters orantıyı yani bir özellik artış eğiliminde iken diğer özelliğin azalış eğiliminde olduğunu gösterir. Kovaryans değeri 0 olan özellikler birbirinden bağımsız ve korelasyonsuz özellikler olarak değerlendirilir.

Temel Bileşenler Analizi için kullanılan kovaryans matrisi kare bir matristir. Özelliklerin birbiri ile olan ilişkilerini ortaya koyan bu matris toplam özellik sayısı boyutlarındadır. Denklem 3.6'da k adet özelliğe sahip bir veri seti için kovaryans matrisi örneği verilmektedir.

$$C_{k \times k} = \begin{bmatrix} cov(x_1, x_1) & \cdots & cov(x_1, x_k) \\ \vdots & \ddots & \vdots \\ cov(x_k, x_1) & \cdots & cov(x_k, x_k) \end{bmatrix} \quad (3.6)$$

Elde edilen kovaryans matrisi, Denklem 3.7'de ifade edilen kovaryansın kendi arasında değişme özelliğinden dolayı simetrik bir matristir.

$$cov(x_n, x_m) = cov(x_m, x_n) \quad (3.7)$$

Kovaryans matrisinin asal köşegeni üzerinde bulunan elemanlar özelliklerin kendi ile kovaryansları olduğu için Denklem 3.8'de verilen eşitliğe göre köşegen üzerindeki elemanlar özelliklerin varyanslarına eşittir.

$$cov(x_m, x_m) = var(x_m) \quad (3.8)$$

Elde edilen kovaryans matrisinden çıkarılan özvektörler PCA algoritmasının model oluştururken kullandığı vektörlerdir. Bir vektör matris ile çarpıldığında eğer yönü değişmiyor sadece büyüklüğünde değişim yaşıyor ise o vektöre

matrisin özvektörü denmektedir. Ve özvektör için değişen büyüklük miktarı bir değer ile ifade edilebilir. Bu değere özdeğer adı verilir. Denklem 3.9'da özdeğer, matris ve özvektör eşitliği ile gösterilmektedir.

$$C_{kxk} \bar{v} = \theta \bar{v} \quad (3.9)$$

Denklemde ifade edilen θ değeri özdeğeri gösterirken $\bar{v}$ değeri özvektörü simgelemektedir. Özvektörler kare matrisler için bulunabilmektedir fakat her matris için bir özvektör olma zorunluluğu yoktur. Simetrik matrislerde bulunan bütün özvektörler hiçbir özvektör birbiri ile paralel özellik göstermez (Shlens,2014).

Bu çalışmanın konusu olan veri seti için 11500x178 boyutlarında bir matris olacaktır. Matrisin örnek gösterimi Denklem 3.10'da verilmiştir.

$$X = \begin{bmatrix} x_1^1 & \cdots & x_1^{178} \\ \vdots & \ddots & \vdots \\ x_{11500}^1 & \cdots & x_{11500}^{178} \end{bmatrix} \quad (3.10)$$

Veri setinin temsilinde kullanılacak özellik kümesinin belirlenmesi için bütün ağırlıkları dengelemek adına veri setine normalizasyon uygulanır. Uygulanan normalizasyon sonrası özelliklerin ortalamasının 0 olması sağlanır. Özelliklere uygulanan normalizasyon Denklem 3.11'de gösterilmektedir.

$$m_j = \frac{1}{n} \sum_{i=1}^n (x_j^i) \quad (3.11)$$

Denklemde verilen m değeri X matrisi için 11500 adet olan satır değerini temsil eder. j değeri 1'den başlamak üzere 11500'e kadar devam eden satırın sırasını göstermektedir. n değeri toplam özellik miktarının değeridir ve üzerinde çalıştığımız veri seti için bu parametrenin değeri 178'dir. Her bir özellik için

$m \times 1$ büyüklüğünde bir vektör tanımlanır ve bu vektör her bir satır değerinden çıkarılarak 0 ortalamalı yeni matris elde edilir. Normalizasyona uygun tanımlanan matris Denklem 3.12’de verilmiştir.

$$X_{NM} = \begin{bmatrix} x_1^1 - m_1 & \cdots & x_1^n - m_1 \\ \vdots & \ddots & \vdots \\ x_m^1 - m_m & \cdots & x_m^n - m_m \end{bmatrix} \quad (3.12)$$

Elde edilen normalize edilmiş veri setinden Denklem 3.13’de görülen matris çarpımı ile kovaryans matrisi elde edilir.

$$C = \frac{1}{n} \begin{bmatrix} x_1^1 - m_1 & \cdots & x_1^n - m_1 \\ \vdots & \ddots & \vdots \\ x_m^1 - m_m & \cdots & x_m^n - m_m \end{bmatrix}_{m \times n} * \begin{bmatrix} x_1^1 - m_1 & \cdots & x_1^n - m_1 \\ \vdots & \ddots & \vdots \\ x_m^1 - m_m & \cdots & x_m^n - m_m \end{bmatrix}_{n \times m}^T \quad (3.13)$$

Denlemde görüldüğü üzere $m \times n$ boyutlarındaki normalizasyon uygulanmış veri seti transpoze edilmiş kendi ile çarpılmış ve oluşan matrisin her bir elemanı toplam özellik sayısına bölünmüştür. Bu şekilde elde edilmek istenen özdeğer ve özvektörler için kovaryans matrisi oluşturulmuş olmaktadır. Denklem 3.9’da verilmiş olan özdeğer fonksiyonuna elde ettiğimiz matris uygulandığında Denklem 3.14’de verilen eşitlik oluşmaktadır.

$$C_{m \times m} \bar{v} = \theta \bar{v} \quad (3.14)$$

Eşitliğin her iki tarafında da bulunan $\bar{v}$ vektörü birbirine bölündüğünde ortaya birim matris çıkmaktadır ve denklemde I olarak gösterilmiştir. Ve kovaryans matrisinin θ değeriyle farkının determinanı 0 olmalıdır. Denklem 3.15’de ilgili eşitlik verilmektedir.

$$\det(C - \theta I) = 0 \quad (3.15)$$

Denklem 3.14’de verilen eşitlik yardımı ile kovaryans matrisinde her bir özellik için özdeğerler bulunmuş olur. İşlem sonucunda elde edilen özdeğerlerden en büyüğü veri setini en iyi temsil eden özellik yani temel bileşen olmaktadır. Her bir özdeğer kendilerine bağlı özvektörü gösterdiği için bunlar büyükten küçüğe doğru sıralandığında veri setini en iyi temsil eden özellikler elde edilmiş olunur.

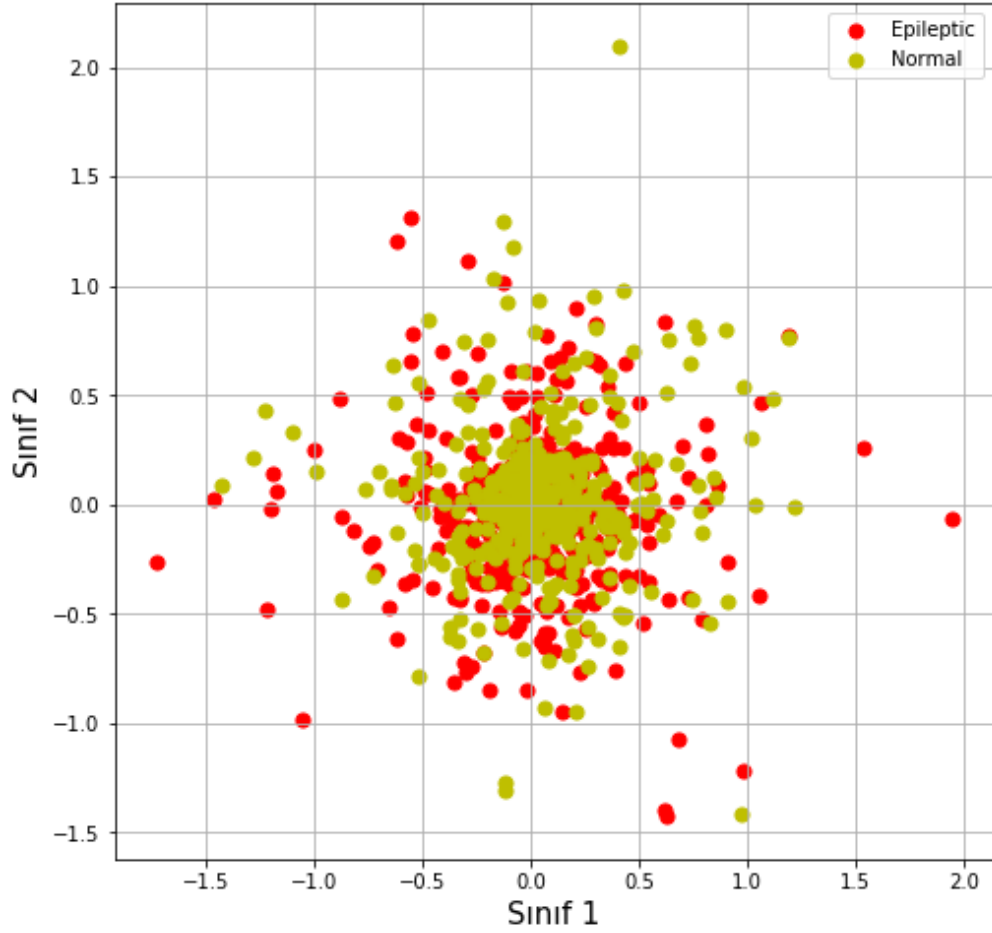
Veri setini temsil edecek özellikler belirlenirken özdeğeri en yüksek olanları seçip en küçük olanları elimine etmek belirli bir miktarda veri kaybına neden olacaktır. Bu veri kaybına ne kadar tolerans gösterileceği problemin tipine, veri集中的 özelliklerin birbiri arasındaki korelasyona bağlıdır. Kaç adet özelliğin elimine edileceğini veren denklem, Denklem 3.16’da verilmiştir.

$$\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^m x_i} < \mu \quad (3.16)$$

Denklemden verilen μ değeri tolerans sabitinin ifadesidir. m toplam özvektör sayısını ifade ederken n seçilecek olan özvektör sayısını ifade eder. Özellikler arasındaki korelasyonun yüksek olduğu veri setlerinde seçilecek olan özellik sayısı yeterince az olduğunda bile yüksek bir başarımla elde edilebilir.

Tüm bu matematiksel arkaplan dahilinde PCA algoritması özellik çıkarımı ve veri görselleştirmeleri için kullanılan önemli bir algoritma olduğu ve tüm bu çıkarımların kolayca yapılabileceği görülmektedir (Jimenez ve Landgrebe, 1998).

Epileptik ve normal olarak iki sınıfa ayırmaya çalıştığımız veri seti 2 boyutlu düzlemde dağılımları gözlenmiş ve açıklanabilir varyans oranı hesaplanmıştır. Şekil 3.7’de verilerin herhangi bir işleme tabi tutulmadan saf halleri görülmektedir.



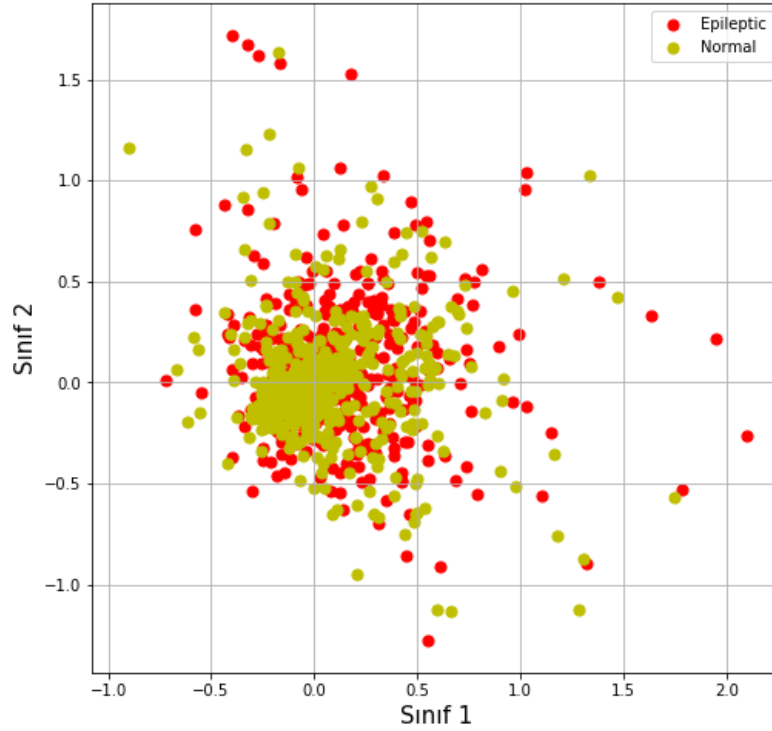
Şekil 3.7. Temel bileşen analizi ile 2 boyutlu uzayda veri dağılımının gösterimi.

Şekilde görüldüğü üzere sınıflamaya çalıştığımız bu verinin sınıf merkezleri birbirlerine çok yakın bir durumdadır. 2 boyutlu uzayda sınıflamaya çalıştığımız bu veri seti birçok veri noktası için üst üste gözükmemektedir. Burada sınıflama başarımını artırmak için veri üzerindeki sınıflandırıcı özelliklerin standart normalizasyon uygulandıktan sonra kuvvetlerini alarak veriyi bu şekilde manipüle etme yolu tercih edildi. Sınıflandırıcı özelliklerin kuvvetlerin alınmasındaki temel amaç, varyans farkını yani toplam varyans kazancını artırabilmektir. Her bir kuvvet uygulama iterasyonundan sonra sınıfların kendi bazında varyans kazançları kaydedilmiştir ve bu varyans kazançlarının farkının mutlak değeri alınarak toplam varyans kazancı elde edilmiştir. Sınıf özelliklerinin varyans değerleri, sınıfların varyans kazançları ve toplam kazanç değerleri

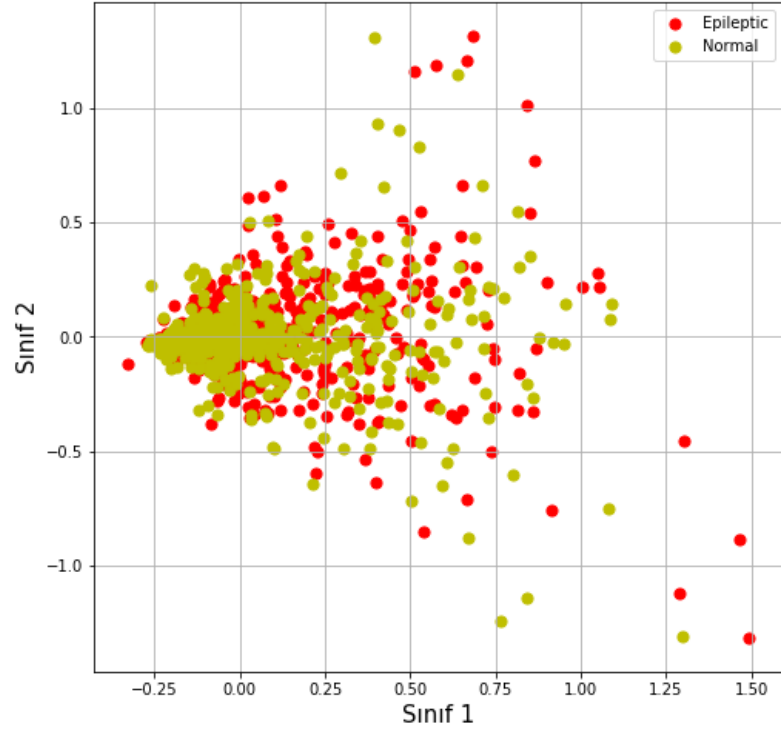
Çizelge 3.1’de görülebileceği şekilde kayıt edildi. Kuvvetler uygulandıktan sonra sırası ile Şekil 3.8, Şekil 3.9, Şekil 3.10, Şekil 3.11’deki gibi dağılımlar oluşmuştur.

Çizelge 3.1. Kuvvet uygulamaları sonrası sınıfların varyans değerleri.

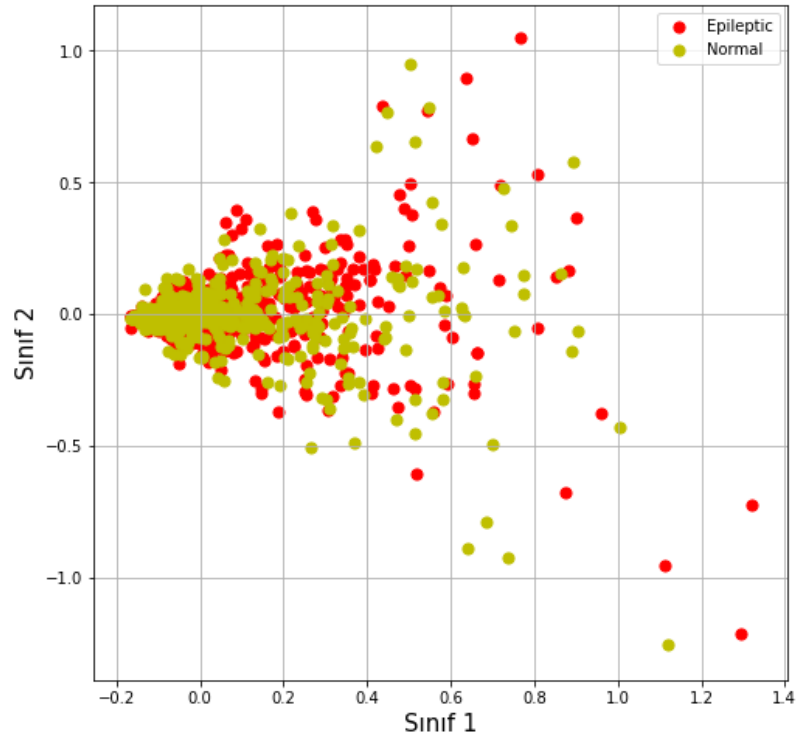
PCA Uygulanan Kuvvet	Normal	Epileptik	Normal Kazanç	Epileptik Kazanç	Toplam Kazanç
1	0.06245396	0.0582515			
3	0.07090286	0.0636541	+0.00844	+0.00540	0.00304
5	0.10917716	0.0594337	+0.03827	-0.00422	0.04249
7	0.12033857	0.0537303	+0.01116	-0.00570	0.01686
9	0.11001304	0.0493376	-0.01032	-0.00439	0.00593



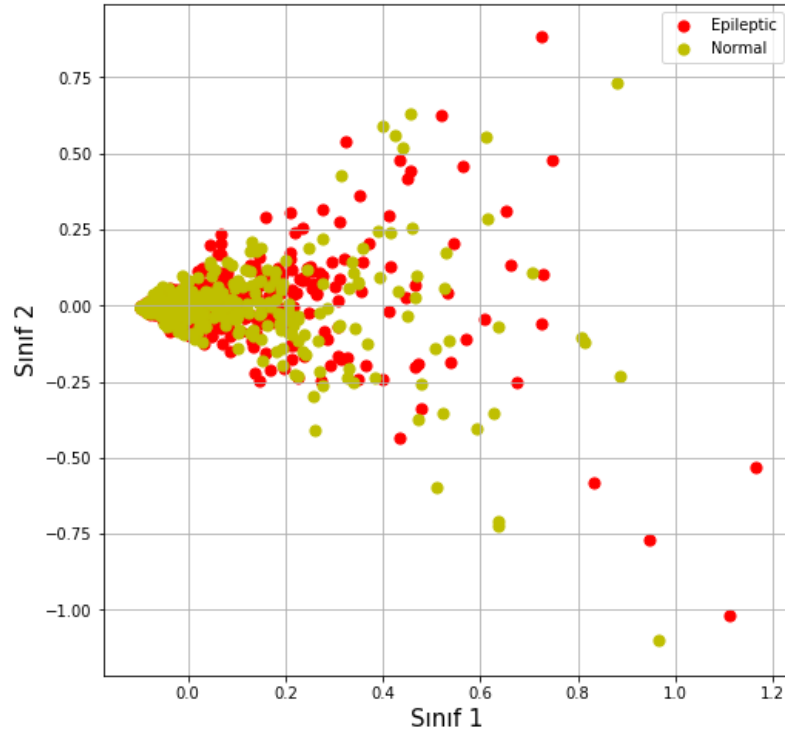
Şekil 3.8. Üçüncü dereceden PCA uygulaması



Şekil 3.9. Beşinci dereceden PCA uygulaması



Şekil 3.10. Yedinci dereceden PCA uygulaması



Şekil 3.11. Dokuzuncu dereceden PCA uygulaması

Tablodan görüleceği üzere sınıflandırıcı özelliklerin beşinci dereceden kuvveti alındığında normal sınıfının varyansı kayda değer bir biçimde artmış ve epileptik sınıfının varyansı yeterli seviyede azalmıştır ve tüm bunların sonucu olarak toplam varyans kazancı maksimum olarak hesaplanmıştır. Bu varyans kazancı, verilerin birbirinden kayda değer ölçülerde ayrılabilmesini ve sınıflandırma başarımının artacağını göstermektedir.

3.4. Makine Öğrenmesi Tekniğinin Uygulanması

Bu bölümde, araştırmanın konusu olan rassal orman algoritmasının makine öğreniminin yapılacağı programlama dili ve özellikleri, uygulama için kullanılacak olan yardımcı metot ve metrikler ile algoritmanın gerçekleştirilmesi için izlenen adımlar açıklanmaktadır.

3.4.1. Python programlama dili ve özellikleri

Python yüksek seviyeli, yordamsal, nesne yönelimli, birçok alanda kullanılan bir programlama dilidir. 1991 yılında nesne yönelimli bir programlama dili etiketiyle Van Rossum tarafından bilim literatürüne kazandırılmıştır (Van Rossum,1991). Ölçek farketmeksizin projelere uygunluğu, temiz ve mantıksal kod yazımı ve yapısal öğelerin dile uygunluğu ile yazılımcıların sade kod yazması hedeflenmiştir (Kuhlman,2012)

Python, dinamik veri tipini yaklaşımına uygun olarak tasarlanmıştır. Nesne yönelimli programlama yaklaşımının önemli öğelerinden biri olan çöp toplayıcısı özelliğininde içinde barındırır. Aynı zamanda farklı programlama modellerine uygunluğu ile esnek bir yapı sunmaktadır. Fonksiyonel, nesne yönelimli veya prosedürel tüm yazılımlar Python dili ile gerçekleştirilebilmektedir.

Python, Perl dilinin bir işi yapmanın birden çok yolu vardır prensibine karşıt olarak bir işi yapmanın yalnız ve yalnız bir yolu vardır prensibini ortaya koymuştur (Peters,2004).

Python programlama dilini kullanan geniş bir yazılım geliştiricileri ağı mevcuttur. Bu dil ile uygulama geliştiren veya hesaplama yapan yazılım geliştiricileri Python çevresine farklı görevleri yerine getiren kütüphaneler kazandırmışlardır. Pandas, Numpy ve Scikit-learn gibi kütüphaneler makine öğrenmesi ve veri madenciliği gibi konularda çalışma yapanların kullandığı temel kütüphaneler haline gelmiştir.

3.4.2. Yardımcı yöntemler ve başarımlar metrikleri

Yardımcı yöntemler ve başarımlar metrikleri, sınıflandırma problemlerinin çözümünü kolaylaştıran ve deney çıktılarını nicel bir şekilde gözlemlemeye imkan sağlayan yol ve yöntemler bütünüdür. Bu konuda yapılmış çalışmalarla, bu çalışmayı kıyaslamayı mümkün kılan metrikler bu bölümde ele alınacaktır.

3.4.2.1. K katlı çapraz doğrulama

K katlı çapraz doğrulama tekniği ile veri seti eğitim ve test verisi olarak ayırmak yerine tüm veri seti dönüşümsel olarak hem eğitim hem de test için kullanılmış olunur. Bu tekniğin ortaya çıkış amacı makine öğrenmesi modellerinin sadece eğitim için ayrılmış olan veriler ile bağımlılığını ortadan kaldırmak ve makine öğrenmesi tekniklerinin en önemli sorunlarından biri olan aşırı öğrenmenin önlenmesidir. Tüm verinin hem eğitim hem test için kullanılması birçok açıdan faydalıdır. Bu sayede eğitim kümesindeki örnek sayısı artırılarak daha geniş bir eğitim seti ile model eğitilmiş olunur ve aynı şekilde tüm veri setinin ayrı parçalarından alınmış örnekler ile eğitilen modelin test edilmesi sonuçların daha güvenilir olmasına olanak sağlamaktadır.

Tekniğe adını da vermiş olan k sayısı verinin toplam kaç parçaya ayrılacağını ifade eden 1'den büyük bir doğal sayıdır. Bu sayı adedince bölümlere ayrılan verinin ilk parçası test verisi olarak ayrılır ve geri kalanı ile model eğitimi yapılır. Yöntem gereği iteratif olarak ve bölüm sıra numarası ile test için ayrılır ve geri kalan veri seti eğitim verisi olarak kullanılır. Yöntem, Şekil 3.12'de gösterilen şekilde bir işleyişe sahiptir.

5-fold CV	DATASET				
Estimation 1	Test	Train	Train	Train	Train
Estimation 2	Train	Test	Train	Train	Train
Estimation 3	Train	Train	Test	Train	Train
Estimation 4	Train	Train	Train	Test	Train
Estimation 5	Train	Train	Train	Train	Test

Şekil 3.12. K katlı çapraz doğrulama yönteminin şekilsel gösterimi (Packtpub,2020).

Şekilden de görülebileceği üzere veri seti 5 parçaya ayrılmış ve her iterasyonda test için kullanılmamış bir sonraki parça test için ayrılmış ve geri kalan veriler ile eğitim seti oluşturulmuştur.

3.4.2.2. Karmaşıklık matrisi

Karmaşıklık Matrisi sınıflandırma problemlerinde kullanılan sonuçların gösterim aracıdır. Makine öğrenmesi modellerinin test girdileriyle üretilen çıktıların bu matris yardımı ile başarı ölçümlerinin hesaplanması gerçekleştirilir. Şekil 3.13’de örnek yerleşim gösterilmiştir.

		Öngörülen sınıf (Predicted Class)	
Gerçek Sınıf (Actual Class)		C ₁ (Positive)	C ₂ (Negative)
	C ₁ (Positive)	True positive TP	False negative FN
	C ₂ (Negative)	False positive FP	True negative TN

Şekil 3.13. Karmaşıklık matrisi yerleşim gösterimi.

Şekilde görülebileceği gibi verinin gerçek sınıf ve makine öğrenmesi modelinin tahminlemiş olduğu sınıf olarak iki farklı uzay bulunmaktadır. Test sonuçları bu iki uzayın kesiştiği matrise aktarılır. Doğru Pozitif (True Positive) değeri gerçek sınıfı pozitif olan verilerin tahminleyici model tarafından da doğru olarak tahmin edilenlerin toplamının ifadesidir. Yanlış Negatif (False Negative) değeri gerçekte negatif olan verilerin tahminleyici model tarafından doğru sınıfına dahil edildiği verilerin toplam ifadesidir. Yanlış Pozitif (False Positive) gerçekte doğru olan verilerin tahminleyici tarafından yanlış sınıflandırma sonucu negatif olarak işaretlenmiş verilerin toplam ifadesidir. Doğru Negatif (True Negative) gerçek değeri negatif olan verilerin tahminleyici tarafından doğru bir şekilde sınıflandırılarak negatif olan tahminlenen verilerin toplam ifadesidir.

Karmaşıklık Matrisi makine öğrenmesi metriklerinin uygulanmasında çok önemli bir yere sahiptir. Bu matristen yararlanılarak doğruluk, hassaslık ve özgüllük gibi başlıca temel metrikler uygulanarak sınıflandırma probleminin hangi ölçülerde çözülebildiği rahatça görülebilir.

3.4.2.3. Doğruluk

Doğruluk, sınıflandırma problemlerinde tahminleyici modelin veri setini ne kadar doğru sınıflandırdığını gösteren bir metriktir. Karmaşıklık Matrisi'ndeki Doğru Pozitif ve Doğru Negatif değerlerin yapılan tüm tahminlere oranı doğruluğu vermektedir. Denklem 3.17'de verilen formül ile bulunmaktadır. Doğruluk değeri 1 ve 0 arasında değerler alır ve 100 ile çarpılarak doğruluk yüzdesi elde edilir. Doğruluk değerinin 1'e yakınsaması modelin başarılı sınıflama yaptığını gösterirken benzer şekilde 0'a yakınsayan modellerin başarısız sınıflandırma yaptığı yorumu yapılır.

$$\text{Doğruluk} = \frac{\text{Doğru Pozitif} + \text{Doğru Negatif}}{\text{Tüm Durumlar Toplamı}} \quad (3.17)$$

3.4.2.4. Hassaslık

Hassaslık, Karmaşıklık Matrisi üzerinde Doğru Pozitif olarak bulunmuş toplam değerinin Doğru Pozitif ve Yanlış Negatif'lerin toplamına oranı olarak tanımlanır. Daha genel anlamda tahminleyicinin doğru tahmin ettiği pozitif değerlerin gerçekte pozitif olan değerlere oranıdır. Yanlış Negatif değerler gerçekte pozitif olan değerler olduğu için bu küme doğru olarak tahmin edilmiş değerler ile toplanarak örnek uzayını diğer bir deyiş ile oluşan oranın paydasını oluşturur. Denklem 3.18'de formülü verilmiştir.

$$\text{Hassaslık} = \frac{\text{Doğru Pozitif}}{\text{Doğru Pozitif} + \text{Yanlış Negatif}} \quad (3.18)$$

3.4.2.4. Özgüllük

Özgüllük, Karmaşıklık Matrisi üzerinde Doğru Negatif olarak bulunmuş toplam değerinin Doğru Negatif ve Yanlış Pozitiflerin toplamına oranı olarak tanımlanır. Daha genel anlamda tahminleyicinin doğru tahmin ettiği negatif değerlerin gerçekte negatif olan değerlere oranıdır. Yanlış Pozitif değerler gerçekte negatif olan değerler olduğu için bu küme doğru olarak tahmin edilmiş değerler ile toplanarak örnek uzayını diğer bir deyiş ile oluşan oranın paydasını oluşturur. Denklem 3.19’da formülü verilmiştir.

$$\text{Özgüllük} = \frac{\text{Doğru Negatif}}{\text{Doğru Negatif} + \text{Yanlış Pozitif}} \quad (3.19)$$

3.4.3. Rassal orman algoritmasının uygulanması

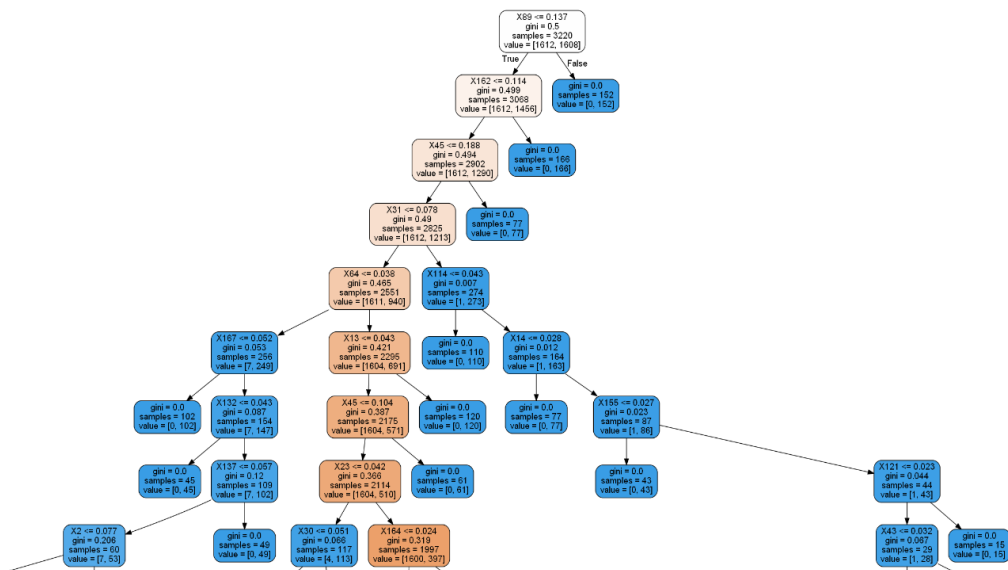
Sınıflandırma uygulaması Python 3.6 versiyonu üzerinde gerçekleştirilmiştir. Pandas ve scikit-learn kütüphaneleri bu çalışmanın bağımlı olduğu ana kütüphanelerdir. Pandas (Pd) kütüphanesi veri setimizin veri çerçevesine dönüştürülmesi, kullanmayacağımız verilerin elimine edilmesi işlerinde kullanılmıştır. Scikit-learn (Sklearn) kütüphanesi verinin normalize edilmesinde, algoritmanın yüklenmesinde, çapraz doğrulamanın gerçekleştirilmesinde, eğitim sonucunda oluşan modelin test verileriyle test edilmesinde ve tüm bu deney sonuçlarının karmaşıklık matrisi ve doğruluk raporlarının hesaplamasında kullanılmıştır. Uygulamanın işleyiş sırası aşağıda verilmiştir.

- 1- Veri bulunduğu dosya konumundan Pd *read_csv* fonksiyonu ile okunmuştur.
- 2- Veri setinde bulunan normal ve epileptik sınıfları dışında kalan sınıflar, Pd *map* fonksiyonu yardımıyla aynı sınıf etiketiyle işaretlenmişlerdir. Bu etiketlenen veri grubu Pd *drop* fonksiyonu ile veri setinden elimine edilmiştir.
- 3- Sınıflandırıcı özellikler ve sınıf etiketi ayrı birer veri çerçevesine

atanmışlardır.

- 4- Sklearn altında bulunan *MinMaxScaler* sınıfıyla veri çerçevesi normalize edilmiştir.
- 5- Normalleştirilmiş veri beş kez kendi ile çarpılarak beşinci kuvveti alınmış ve yeni sınıflandırıcı özellikler olarak atanmıştır.
- 6- Sklearn *train_test_split* fonksiyonu yardımıyla tüm datanın %30'u test için ayrılmış, geri kalanı eğitim seti olarak atanmıştır.
- 7- Sklearn altında bulunan algoritmanın kurucu fonksiyonuna hiper parametre ayarları olan ağaç sayısı 2000 ve maksimum sınıflandırıcı özellik 15 olarak verilerek algoritma çağırımı yapılmıştır.
- 8- Algoritma eğitim seti ile eğitilmiş ve 5 iterasyon çapraz doğrulama uygulanmıştır.
- 9- Karmaşıklık matrisi, doğruluk, hassaslık ve özgüllük değerleri hesaplanmıştır.

Uygulamanın eğitim sonrasındaki adımlarında eğitilen modeldeki ağaçlardan rastgele seçim yaparak modelin oluşmasındaki aşamalar kontrol edilmiştir. Şekil 3.14’de oluşan ağaçlardan birinin örnek kesiti verilmiştir.



Şekil 3.14. Oluşan ağaçlardan alınan örnek kesit

4. ARAŞTIRMA BULGULARI VE TARTIŞMA

Çalışmanın çıktıları bu bölümde iki farklı şekilde ele alınacaktır. İlk kısımda algoritma üzerinde hiper parametre optimizasyonu yapılmadığı ve verinin kuvvetinin değiştirilmediği senaryolar ile bütün hepsinin uygulandığı ideal senaryo karşılaştırılacaktır. İkinci olarak bu çalışmadan önce bu veri seti ile çalışmış araştırmacıların derlenen sonuçları ile karşılaştırma yapılacaktır. Başarım ölçeğimiz doğruluk, hassasiyet ve özgüllük değerleri üzerinden karşılaştırılacaktır.

Çizelge 4.1’de ideal senaryo için 5 kat çapraz doğrulama sonuçları verilmektedir. Çizelge 4.2’te ideal senaryoya ait karmaşıklık matrisi verilmektedir. Çizelge 4.3’de ideal senaryo ile diğer durumların karşılaştırılması verilmiştir.

Çizelge 4.1. İdeal senaryoya ait 5 katlı çapraz doğrulama sonuçları

Çapraz Doğrulama 1	Çapraz Doğrulama 2	Çapraz Doğrulama 3	Çapraz Doğrulama 4	Çapraz Doğrulama 5	Ortalama Doğruluk	Standart Sapma
99.78	99.67	100	99.67	99.78	99.78	(+/- 0.0011)

Çizelge 4.2 İdeal senaryo çapraz doğrulamalar sonucu oluşan kümülatif karmaşıklık matrisi.

Gerçek\Tahmin	Normal	Epileptik	Toplam
Normal	2299	1	2300
Epileptik	9	2291	2300

Çizelge 4.3 İdeal senaryonun diğer olası senaryolar ile başarımların karşılaştırması

Senaryo	Doğruluk	Hassasiyet	Özgüllük
Varsayılan Parametreler	99.08	98.57	98.90
Hiper Parametreler	99.71	99.48	99.95
Varsayılan Parametreler ve Kuvvet uygulaması	98.97	98.62	99.43
İdeal Senaryo	99.78	99.61	99.95

Tablolardan da görülebileceği üzere varsayılan parametreler ile ideal senaryo arasında kayda değer bir farklılık söz konusudur. Hiper parametre optimizasyonu ve ideal senaryo arasındaki farklılık, varsayılan değer ile olan kadar önemli gözükmemektedir fakat parametre optimizasyonundan sonra doğruluk oranını artırdığı için varyans analizinden yaptığımız çıkarımında başarımda katkısının olduğu açıkça görülebilmektedir.

Tartışma bölümünün bundan sonraki safhasında çalışmamızın aynı veri seti üzerinde yapılmış önceki çalışmalar ile karşılaştırılması ele alınacaktır. Çizelge 4.4'te derlenen öncü çalışmaların bu çalışma ile karşılaştırılması verilmektedir. Bu çalışma, diğer araştırmacıların Z ve S olarak etiketlediği iki sınıflı veri seti üzerinde yapılmış olan çalışmalar ile kıyaslanmıştır. Beyindeki doğrusal olmayan elektrik dalgalarının gözlenmesi üzerine oluşturulan bu veri setine dair normal ve epileptik olarak sınıflandırma çalışmalarının ilk örneğinin yapıldığı 2004 yılından günümüze kadar birçok farklı yöntem ile sınıflandırmalar gerçekleştirilmiştir. Bu tablo detaylı olarak incelenirse görülecektir ki yapay sinir ağları ve türevlerini oluşturan sınıflandırma yöntemleri ağırlıklı olarak kullanılmıştır. Ve yine tüm çalışmalar incelendiğinde en yaygın tercih edilen ön işleme modeli ayrık dalgacık dönüşümü ve bunu takip eden dalgacık dönüşüm modellerinin türevlerinden oluşmaktadır. Veri setinde kategorik bir veri olmaması bu veri seti için yapay sinir ağı modelinin başarılı olması için ön görülebilirdir. Bu veri seti ile çalışan araştırmacılar bu yöntem ile gayet yeterli ve

başarılı sonuçlar elde etmişlerdir. Yapay sinir ağı modelinin dışında en çok tercih edilen diğer algoritmalar destek vektör makineleri ve k-NN algoritmalarıdır. Dalgacık dönüşümleri ve radyal temelli fonksiyonlar çerçevesinde bu algoritmalarda da gayet yeterli ve başarılı sayılabilecek sonuçlar elde edilmiştir. Hassan vd. (2016), bu çalışmasında konusu olan bir grup sınıflandırıcının bir araya getirilerek sınıflandırma yapılması yaklaşımını başarılı bir şekilde gerçekleştirmişlerdir. Bu çalışma ile EEG sinyallerinin parçalara ayrılarak çoğaltılmış halinden %99,78 gibi oldukça başarılı, aşırı öğrenme sorunuyla asla karşı karşıya gelmeyecek ve paralel olarak çalışabilen bir yöntemin başarımını deneysel olarak ispatlamış oluyoruz.

Çizelge 4.4 Aynı veri setinin kullanıldığı önceki çalışmalar ve yaptığımız çalışmanın karşılaştırılması.

Yazar(lar)	Çalışma Başlığı	Metot	Doğruluk (%)
Nigam ve Graupe (2004)	A neural-network-based detection of epilepsy	Doğrusal olmayan ön işleme filtresi ile yapay sinir ağı	97.2
Srinivasan vd. (2005)	Artificial neural network based epileptic detection using time-domain and frequency-domain features	Zaman ve frekans alan özellikli yinelenen sinir ağı	99.6
Kannathal vd. (2005)	Entropies for detection of epilepsy in EEG	Entropi ölçümlü adaptif neuro-fuzzy çıkarım sistemi	92.22
Sadati vd. (2006)	Epileptic seizure detection using neural fuzzy networks	Ayrık dalgacık dönüşümlü adaptif neuro-fuzzy ağı	85.9
Polat and Güneş (2007)	Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform	Hızlı Fourier dönüşümü ve karar ağaçları	98.72
Subasi (2007)	EEG signal classification using wavelet feature extraction and amixture of expert model	Ayrık dalgacık dönüşümlü uzman model karışımı	95
Tzallas vd. (2007)	Automatic seizure detection based on	Zaman-frekans analizi ve yapay sinir ağı	100

	time–frequency analysis and artificial neural networks		
Srinivasan vd. (2007)	Approximate entropy-based epileptic EEG detection using artificial neural networks	Yaklaşık entropi-Elman ağı (yinelenen sinir ağı)	100
Guo vd. (2009)	Classification of EEG signals using relative wavelet energy and artificial neural networks	Ayrık dalgacık dönüşümlü göreceli dalgacık enerjisi ve yapay sinir ağı	95.2
Ocak (2009)	Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy	Ayrık dalgacık dönüşümü ve yaklaşık entropi ile yapay sinir ağı	96
Guo vd. (2010)	Epileptic seizure detection using multiwavelet transform based approximate entropy and artificial neural networks	Çoklu dalgacık dönüşümü entropi yaklaşımı ve çok katmanlı yapay sinir ağı	100
Subasi ve Gursoy (2010)	EEG Signal classification using PCA, ICA, LDA and support vector machine	Temel bileşenler analizi (PCA), bağımsız bileşenler analizi (ICA) ve doğrusal diskriminant analizi (LDA) ile destek vektör makinesi	98.75 (PCA), 99.5 (ICA), 100 (LDA)
Kumar vd. (2012)	Epileptic seizures detection in EEG using DWT-based ApEn and artificial neural network	Ayrık dalgacık dönüşümü, yaklaşık entropi ile yapay sinir ağları	100
Fernandez-Blanco vd. (2012)	Automatic seizure detection based on star graph topological indices	Genelleştirilmiş diskriminant analizi	99
Dehuri vd. (2013)	Epileptic seizure identification from electroencephalography signal using DE-RBFNs ensemble	Radyal temelli fonksiyon ile yapay sinir ağları	100

Xie ve Krishnan (2013)	Wavelet-based sparse functional linear model with applications to EEGs seizure detection and epilepsy diagnosis	Dalgacık temelli ayrık doğrusal model ile k-NN	100
Song ve Zhang (2013)	Automatic recognition of epileptic EEG patterns via Extreme Learning Machine and multiresolution feature extraction	Aşırı Öğrenme Makineleri	94.20
Lee vd. (2014)	Classification of normal and epileptic seizure EEG signals using wavelet transform, phase-space reconstruction, and Euclidean distance	Dalgacık dönüşümü ile ağırlıklı fuzzy üyeli fonksiyon	98.17
Kaya vd. (2014)	1D-local binary pattern based feature extraction for classification of epileptic EEG signals	Tek boyutlu ikili desen temelli özellik çıkartımı ve bayes ağı	99.50
Zhu vd. (2014)	Epileptic seizure detection in EEGs signals using a fast weighted horizontal visibility algorithm	Hızlı ağırlıklı yatay görünür algoritma ve k-NN	100
Kumar vd. (2014)	Epileptic seizure detection using DWT based fuzzy approximate entropy and support vector machine	Ayrık dalgacık dönüşümü temelli fuzzy entropi yaklaşımı ile destek vektör makineleri	100
Fu vd. (2014)	Classification of seizure based on the time-frequency image of EEG signals using HHT and SVM	Hilbert-Hung dönüşümü temelli zaman-frekans kalıbı ile destek vektör makineleri	99.12
Chen (2014)	Automatic EEG seizure detection using dual-tree complex wavelet-Fourier features	Çift ağaç karmaşık dalgacık fourier özellikleri ile k-NN	100
Dhiman vd. (2014)	Genetic algorithms tuned expert model for detection of	Dalgacık paketi ve genetik algoritma ile destek vektör makineleri	100

	epileptic seizures from EEG signatures		
Yuan vd. (2014)	Epileptic EEG classification based on kernel sparse representation	Radyal Temelli Fonksiyon ile destek vektör makineleri	98.63
Zhu vd. (2014)	Epileptic seizure detection in EEGs signals using a fastweighted horizontal visibility algorithm	Hızlı fourier dönüşümü ile k-NN	99
Kang vd. (2015)	An efficient detection of epileptic seizure by differentiation and spectral analysis of electroencephalogram s	İkinci dereceden diskriminant analizi ve çok katmanlı yapay sinir ağı	99.78
Tawfik vd. (2015)	A hybrid automated detection of epileptic seizures in EEG records	Dalgacık dönüşümü ve ağırlıklı permütasyon entropi algoritması ile destek vektör makineleri ve yapay sinir ağları	99.5
Peker vd. (2016)	A novel method for automated diagnosis of epilepsy using complex-valued classifiers	Sürekli dalgacık dönüşümü ve evrimsel sinir ağları	99.5
Zamir (2016)	Detection of epileptic seizure in EEG signals using linear least squares preprocessing	Lojistik, LazyIB1, LazyIB5 ve J48 algoritmaları	100
Hassan vd. (2016)	Epileptic seizure detection in EEG signals using tunable-Q factor wavelet transform and bootstrap aggregating	Ayarlanabilir Q faktör dalgacık dönüşümü ile torbalama yöntemi	100
Bhardwaj vd. (2016)	A novel genetic programming approach for epileptic seizure detection	Ampirik mod ayrıştırması ile Kartezyen genetik programlama algoritmaları	98.64
Zhang ve Chen (2016)	LMD based features for the automatic seizure detection of EEG signals using SVM	Yerel ortalamalar ayrıştırması ile genetik algoritma ve destek vektör makineleri	100

Swami vd. (2016)	A novel robust diagnostic model to detect seizures in electroencephalography	Sürekli dalgacık dönüşümü ile genelleştirilmiş regresyon sinir ağı	100
Martinez-del-Rincon vd. (2017)	Non-linear classifiers applied to EEG analysis for epilepsy seizure detection	Ayrık dalgacık dönüşümü ve radyal temelli fonksiyon ile destek vektör makineleri	99.85
Chen vd. (2017)	Automatic epileptic seizure detection in EEG using non subsampled wavelet-Fourier features	Ayrık dalgacık dönüşümü ile yapay sinir ağları ve destek vektör makineleri	100 (ANN), 78 (SVM)
Bhattacharyya vd. (2017)	Tunable-Q wavelet transform based multiscale entropy measure for automated classification of epileptic EEG signals	Ayarlanabilir Q faktör dalgacık dönüşümü ile destek vektör makineleri	100
Sharma vd. (2017)	A new approach to characterize epileptic seizures using analytic time-frequency flexible wavelet transform and fractal dimension	Analitik zaman frekanslı esnek dalgacık dönüşümü ve destek vektör makineleri	100
Zhou vd. (2018)	Epileptic seizure detection based on EEG signals and CNN	Evrimsel sinir ağı	100
Hussein vd. (2018)	Epileptic Seizure Detection: A Deep Learning Approach	Uzun kısa bellek hücre temelli tekrarlayan sinir ağı	100
Yavuz vd. (2018)	An epileptic seizure detection system based on cepstral analysis and generalized regression neural network	Logaritmik spektrumun ters fourier dönüşümüne dayalı sepstral analiz ile genelleştirilmiş regresyon yapay sinir ağı	99
Amin vd. (2020)	A novel approach based on wavelet analysis and arithmetic coding for automated detection and diagnosis of epileptic seizure in	Ayrık dalgacık ve aritmetik kodlama ile destek vektör makineleri ve yapay sinir ağları	100

	EEG signals using machine learning techniques		
Sharma vd. (2020)	Seizures classification based on higher order statistics and deepneural network	Derin sinir ağı	100
Bu Çalışma	EEG sinyalleri ile epilepsi krizinin tahminlenmesinde Rassal Orman algoritması ile hiper parametre optimizasyonun uygulanması	Hiper parametre optimizasyonu ve Rassal Orman algoritması	99.78

Çalışılan deneylerin vakalarında %99,78 oranında doğruluk, %99,95 özgüllük ve %99,61 hassasiyet elde edilmiştir ve sonuçlar modelin başarılı şekilde sınıflandırdığını göstermektedir. Özgüllük ve hassasiyet değerleri kıyaslanan tüm çalışmalarda verilmediği için tabloda yer almamıştır.

5. SONUÇ VE ÖNERİLER

Bu çalışmada epilepsi rahatsızlığı olan kişilerin belirlenmesi için, Rassal Orman algoritması ile hiper parametre optimizasyonu yaparak EEG sinyallerinden epilepsi krizi ve normal beyin durumlarını %99,78 doğruluk oranı ile tahmin edebilecek sistemin genel tasarımı ele alınmıştır ve bu veri setinde daha önce yapılmış çalışmalar incelendiğinde yapay sinir ağları dahil olmak üzere birçok farklı yöntem ile aynı seviyede sonuçlar elde edildiği görülmektedir. Somut rakamlar vermek gerekirse; çalışılan deneylerin vakalarında %99,78 oranında doğruluk yanında, %99,95 özgüllük ve %99,61 hassasiyet de elde edilmiştir. Rassal Ormanlar paralel olarak kurulabilen, eğitimde zaman maliyeti düşük ve doğruluk oranı yüksek, özellik çıkartımı ve veri ön işleme gibi ekstra işlemlere ihtiyaç duymayan bir algoritma olduğu görülmektedir. EEG verileri gibi sürekli verilerin işlenmesinde bir alternatif olarak değerlendirilebileceği çalışmamızın sonuçlarından biridir. Sonuçlar ve değerlendirmeler ışığında modelinin başarılı bir şekilde sınıflandırma yaptığı görülmektedir.

KAYNAKLAR

- Acharya U.R., Oh S.L., Hagiwara Y., Tan J.H., Adeli H., 2018. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals, *Computers in Biology and Medicine*, 100 270–278.
- Amin H.U., Yusoff M.Z., Ahmad R.F., 2020. A novel approach based on wavelet analysis and arithmetic coding for automated detection and diagnosis of epileptic seizure in EEG signals using machine learning techniques. *Biomedical Signal Processing and Control* 56, 101707.
- Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P, Elger CE., 2001. Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, *Phys. Rev. E*, 64, 061907.
- Bergstra J., Bardenet R., Bengio Y., Kegl B., 2011. Algorithms for hyper-parameter optimization. In *NIPS*.
- Bergstra J., Bengio Y., 2012. Random search for hyper-parameter optimization.
- Bhardwaj A., Tiwari A., Krishna R., Varma V., 2016. A novel genetic programming approach for epileptic seizure detection, *Comput. Methods Programs Biomed.* 124, 2–18.
- Bhattacharyya A., Pachori R.B., Upadhyay A., Acharya U.R., 2017. Tunable-Q wavelet transform based multiscale entropy measure for automated classification of epileptic EEG signals, *Appl. Sci.* 7 (4), 385.
- Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984. *Classification and Regression Trees*. Belmont, CA: Wadsworth International.
- Breiman, L., 1996. Bagging predictors. *Machine Learning* 26(2), 123–140.
- Breiman, L., 2001. Random Forests. *Machine Learning* 45 (1), 5–32.
- Chen G., Xie W., Bui T.D., Krzyzak A., 2017. Automatic epileptic seizure detection in EEG using non subsampled wavelet-Fourier features, *J. Med. Biol. Eng.* 37 (1), 123–131.
- Chen G., 2014. Automatic EEG seizure detection using dual-tree complex wavelet-Fourier features, *Expert Syst. Appl.* 41, 2391–2394.
- Dehuri S., Jagadev A.K., Cho S.-B., 2013. Epileptic seizure identification from electroencephalography signal using DE-RBFNs ensemble, *Procedia Comput.Sci.* 23, 84–95.

- Dekking F.M., Kraaikamp C., Lopuhaa H.P., Meester L.E., 2005. A Modern Introduction to Probability and Statistics.
- Dhiman R., Saini J.S., Priyanka, 2014, Genetic algorithms tuned expert model for detection of epileptic seizures from EEG signatures, *Appl. Soft Comput.* 19, 8–17.
- Fernandez-Blanco E., Rivero D., Rabuñal J., Dorado J., Pazos A., Munteanu C.R., 2012. Automatic seizure detection based on star graph topological indices, *J.Neurosci. Methods* 209, 410–419.
- Fu K., Qu J., Chai Y., Dong Y., 2014. Classification of seizure based on the time-frequency image of EEG signals using HHT and SVM, *Biomed. SignalProcess. Control* 13, 15–22.
- Grajski, K. A., Breiman, L., Di Prisco, G. V., & Freeman, W. J., 1986. Classification of EEG spatial patterns with a tree-structured methodology: CART. *IEEE transactions on biomedical engineering*, (12), 1076-1086.
- Guo L, Rivero D, Seoane J, Pazos A., 2009. Classification of EEG signals using relative wavelet energy and artificial neural networks. In: *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, 177–84.
- Guo L, Rivero D, Pazos A., 2010. Epileptic seizure detection using multiwavelet transform based approximate entropy and artificial neural networks. *Journal of Neuroscience Methods* 193, 156–163.
- Hassan A.R., Siuly S., Zhang Y., 2016. Epileptic seizure detection in EEG signals using tunable-Q factor wavelet transform and bootstrap aggregating, *Comput.Methods Programs Biomed.* 137, 247–259.
- Ho T.K., 1998. The random subspace method for constructing decision forests. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(8), 832–844.
- Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6), 417.
- Hussein R., Palangi H., Ward R., Wang Z.J., 2018. Epileptic Seizure Detection: A Deep Learning Approach
- Jahankhani P, Kodogiannis V, Revett K., 2006. EEGsignal classification using wavelet feature extraction and neural networks. In: *IEEE John Vincent Atanasoff 2006 International Symposium on Modern Computing (JVA'06)*, 52–7.
- Jiang Y., Cukic B., Menzies T., 2008. Can Data Transformation Help in the Detection of Fault-prone Modules? In *Proceedings of the workshop on Defects in Large Software Systems*, 16–20

- Jimenez, L. O., & Landgrebe, D. A., 1998. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 28(1), 39-54.
- Jolliffe, I. T., 2003. Principal component analysis. *Technometrics*, 45(3), 276.
- Kalayci T, Ozdamar O., 1995. Wavelet preprocessing for automated neural network detection of EEG spikes. *IEEE Engineering in Medicine and Biology Magazine*, 14(2), 160–6.
- Kandel E, Schwartz J, Jessell T., 2000. *Principles of Neural Science*. New York: McGraw-Hill, Health Professions Division.
- Kang J.-H., Chung Y.G., Kim S.-P., 2015. An efficient detection of epileptic seizure by differentiation and spectral analysis of electroencephalograms, *Comput. Biol.Med.* 66, 352–356.
- Kannathal N., Choo M.L Acharya U.R., Sadasivana P.K., 2005. Entropies for detection of epilepsy in EEG, *Comput. Methods Programs Biomed.* 80, 187–194.
- Kannathal N, Choo M, Acharya U, Sadasivan P., 2005. Entropies for detection of epilepsy in EEG. *Computer Methods and Programs in Biomedicine* ,80(3),187–94.
- Kaya Y., Uyar M., Tekin R., Yıldırım S., 2014. 1D-local binary pattern based feature extraction for classification of epileptic EEG signals, *Appl. Math. Comput.* 243, 209–219.
- Kim, J., 1997. Iterated grid search algorithm on unimodal criteria (Doctoral dissertation, Virginia Tech).
- Kuhlman, D., 2012. *A Python Book: Beginning Python. Advanced Python, and Python Exercises*.
- Kumar Y., Dewal M., Anand R., 2012. Epileptic seizures detection in EEG using DWT-based ApEn and artificial neural network, *Signal Image Video Process*, 1–12.
- Kumar Y., Dewal M., Anand R., 2014. Epileptic seizure detection using DWT based fuzzy approximate entropy and support vector machine, *Neurocomputing* 133, 271–279.
- Lee S.-H., Lim J.S., Kim J.-K., Yang J., Lee Y., 2014. Classification of normal and epileptic seizure EEG signals using wavelet transform, phase-space reconstruction, and Euclidean distance, *Comput. Methods Programs Biomed.* 116, 10–25.

- Litt B, Echauz J., 2002. Prediction of epileptic seizures, *The Lancet Neurology*, V1 (1), 22-30.
- Martinez-del-Rincon J., Santofimia M.J., del Toro X., Barba J., Romero F., Navas P., Lopez J.C., 2017. Non-linear classifiers applied to EEG analysis for epilepsyseizure detection, *Expert Syst. Appl.* 86, 99–112
- Morgan J. N. ve Sonquist J. A., 1963. Problems in the analysis of survey data, and a proposal, *J. Amer. Statist. Ass.*, 58, 415-434.
- Nigam V, Graupe D., 2004. A neural-network-based detection of epilepsy. *Neurological Research*, 26(1), 55–60.
- Ocak H., 2009. Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy, *Expert Systems with Applications* 36 (2), 2027–2036.
- Packtpub, 2020, https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789617740/2/ch02lvl1sec14/k-fold-cross-validation. Son erişim tarihi 1 Haziran 2020.
- Peker M., Sen B., Delen D., 2016. A novel method for automated diagnosis of epilepsy using complex-valued classifiers, *IEEE J. Biomed. Health Inform.* 20(1), 108–118.
- Peters, Tim, 2004. PEP 20 – The Zen of Python. Python Enhancement Proposals. Python Software Foundation.
- Polat K., Günes, S., 2007. Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. *Applied Mathematics and Computation*, 187(2), 1017–26.
- Sadati N, Mohseni H.R., Magshoudi A., 2006. Epileptic seizure detection using neural fuzzy networks, in: *Proceedings of the IEEE International Conference on Fuzzy Systems*, Canada, 596–600.
- Sharma M., Pachori R.B., Acharya U.R., 2017. A new approach to characterize epileptic seizures using analytic time–frequency flexible wavelet transform and fractal dimension, *Pattern Recogn. Lett.* 94, 172–179.
- Sharma M., Pachori R.B., Sircar P., 2020. Seizures classification based on higher order statistics and deepneural network. *Biomedical Signal Processing and Control* 59, 101921.
- Shlens, J., 2014. A Tutorial on Principal Component Analysis.

- Song Y., Zhang J., 2013. Automatic recognition of epileptic EEG patterns via Extreme Learning Machine and multiresolution feature extraction, *Expert Syst. Appl.* 40, 5477–5489.
- Srinivasan V, Eswaran C, Sriraam N., 2005. Artificial neural network based epileptic detection using time-domain and frequency-domain features. *Journal of Medical Systems*, 29(6), 647–60.
- Srinivasan V., Eswaran C., Sriraam N., 2007. Approximate entropy-based epileptic EEG detection using artificial neural networks, *IEEE Transaction on Information Technology in Biomedicine* 11 (3), 288–295.
- Subasi A., 2005. Epileptic seizure detection using dynamic wavelet network. *Expert Systems with Applications*, 29(2), 343–55.
- Subasi A., 2006. Automatic detection of epileptic seizure using dynamic fuzzy neural networks. *Expert Systems with Applications*, 31(2), 320–8.
- Subasi A., 2007. EEG signal classification using wavelet feature extraction and amixture of expert model. *Expert Systems with Applications*, 32(4), 1084–93.
- Subasi A., Gursoy M.I., 2010. EEG Signal classification using PCA, ICA, LDA and support vector machine, *Expert Systems with Applications* 37, 8659–8666.
- Suparwito, H., 2019. Factors Influencing the Difficulty Level of the Subject: Machine Learning Technique Approaches. *International Journal of Applied Sciences and Smart Technologies*, 1(1), 65-82.
- Syarif, I., Prugel-Bennett, A., Wills G., 2016. SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance. *Telkomnika*, 14, 1502–1509.
- Swami P., Gandhi T.K., Panigrahi B.K., Tripathi M., Anand S., 2016. A novel robust diagnostic model to detect seizures in electroencephalography, *Expert Syst. Appl.* 56, 116–130.
- Tantithamthavorn C., McIntosh S., Hassan A. E., and Matsumoto K., 2016. Automated parameter optimization of classification techniques for defect prediction models. *IEEE/ACM 38th IEEE International Conference on Software Engineering*.
- Tawfik N.S., Youssef S.M., Kholief M., 2015. A hybrid automated detection of epileptic seizures in EEG records, *Comput. Electr. Eng.*
- Tosun A., Bener A., 2009. Reducing false alarms in software defect prediction by decision threshold optimization. In *Proceedings of the International*

Symposium on Empirical Software Engineering and Measurement, 477–480.

Tzallas A, Tsipouras M, Fotiadis D., 2007. Automatic seizure detection based on time–frequency analysis and artificial neural networks. Computational Intelligence and Neuroscience, 13.

UCI Machine Learning Repository, 2020
<https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition>.
Son erişim tarihi 17 Mayıs 2020.

Van Rossum, G., 1991. Python: An Object Oriented Programming Language.

Viglione SS, Walsh GO, 1975. Proceedings: Epileptic seizure prediction. Electroencephalography and clinical neurophysiology, 39 (4), 435–436.

Yuan Q., Zhou W., Yuan S., Li X., Wang J., Jia G., 2014. Epileptic EEG classification based on kernel sparse representation, Int. J. Neural Syst. 24, 1450015.

Xie S., Krishnan S., 2013. Wavelet-based sparse functional linear model with applications to EEGs seizure detection and epilepsy diagnosis, Med. Biol. Eng.Comput. 51, 49–60.

Zamir Z.R., 2016. Detection of epileptic seizure in EEG signals using linear least squares preprocessing, Comput. Methods Programs Biomed. 133, 95–109.

Zhang T., Chen W., 2016. LMD based features for the automatic seizure detection of EEG signals using SVM, IEEE Trans. Neural Syst. Rehabil. Eng. 25 (8),1100–1108.

Zhou M., Tian C., Cao R., Wang B., Niu Y., Hu T., Guo H., Xiang J., 2018. Epileptic seizure detection based on EEG signals and CNN, Frontiers Neuroinform. 12, 95.

Zhu G., Li Y., Wen P.P., 2014. Epileptic seizure detection in EEGs signals using a fastweighted horizontal visibility algorithm, Comput. Methods Programs Biomed.115, 64–75.

ÖZGEÇMİŞ

Adı Soyadı : Fatih Murathan YILMAZ
Doğum Yeri ve Yılı : Afyonkarahisar, 24/09/1992
Medeni Hali : Bekar
Yabancı Dili : İngilizce
E-posta : fmurathanyilmaz@gmail.com



Eğitim Durumu

Lise : Servergazi Lisesi, 2010
Lisans : Sakarya Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 2015.

Mesleki Deneyim

Vizyoneks Bilgi Teknolojileri	2015-2016
Bridge Yazılım Danışmanlık	2018-2019
Turkcell Teknoloji Araştırma ve Geliştirme	2019-...(devam ediyor)

Yayın

Yılmaz, F., Kasapbaşı, M., 2021. EEG SİNYALLERİ İLE EPİLEPSİ KRİZİNİN TAHMİNLENMESİNDE RASSAL ORMAN ALGORİTMASI İLE HİPER PARAMETRE OPTİMİZASYONUN UYGULANMASI. İstanbul Ticaret Üniversitesi Teknoloji ve Uygulamalı Bilimler Dergisi, 3 (2), 189-203.