T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ KULLANILARAK MİKRODALGA VERİLERİ İLE BEYİNDEKİ KİTLESEL LEZYONLARIN TEŞHİSİ

TUĞBA VURAL

YÜKSEK LİSANS TEZİ
MATEMATİK MÜHENDİSLİĞİ ANABİLİM DALI
MATEMATİK MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
YRD. DOÇ. DR. BİROL ASLANYÜREK

İSTANBUL, 2017

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ KULLANILARAK MİKRODALGA VERİLERİ İLE BEYİNDEKİ KİTLESEL LEZYONLARIN TEŞHİSİ

Tuğba VURAL tarafından hazırlanan tez çalışması 16.06.2017 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Matematik Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Yrd. Doç. Dr. Birol ASLANYÜREK

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Yrd. Doç. Dr. Birol ASLANYÜREK

Yıldız Teknik Üniversitesi

Prof. Dr. Hülya ŞAHİNTÜRK

Yıldız Teknik Üniversitesi

Yrd. Doç. Dr. Tolga Ulaş GÜRBÜZ

Gaziantep Üniversitesi

ÖNSÖZ

Yüksek Lisans çalışmalarım boyunca, bana yol gösteren bilgi ve desteğini hiç bir zaman esirgemeyen değerli hocam ve tez danışmanım Yrd. Doç. Dr. Birol ASLANYÜREK'e,

Ayrıca bana desteklerini hiç bir zaman esirgemeyen ve tüm öğretim hayatım boyunca her zaman yanımda olan çok sevdiğim aileme,

Sonsuz teşekkürler...

Mayıs, 2017

Tuğba VURAL

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

p	Test sınıfının eleman sayısı
x_i	Eğitim sınıfı elemanlarının konumu
x_j	Test sınıfı elemanlarının konumu
TP	True Pozitif
TN	True Negatif
FP	False Pozitif
FN	False Negatif
K	Kappa katsayısı
P_0	Kabul edilen oran
P_c	Kabul edilmesi beklenen oran
mm	Milimetre
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error

KISALTMA LİSTESİ

BT	Bilgisayarlı Tomografi
kNN	k Nearest Neighbour
KDD	Knowledge Discovery Databases
MRG	Manyetik Rezonans Görüntüleme
MWT	Microwavw Tomography
SAS	Statistical Analysis Software
SVM	Support Vector Machine
VTYS	Veri Tabanı Yönetim Sistemleri
YSA	Yapay Sinir Ağları

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

VERİ MADENCİLİĞİ KULLANILARAK MİKRODALGA VERİLERİ İLE BEYİNDEKİ KİTLESEL LEZYONLARIN TEŞHİSİ

Tuğba VURAL

Matematik Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Yrd. Doç. Dr. Birol Aslanyürek

Mikrodalga tomografisi metalik olmayan cisimlerin elektromagnetik malzeme özelliklerini görselleştirmek için kullanılan etkili tanısal yöntemlerden biridir. Son yıllarda biyotıpta mikrodalga görüntüleme yöntemleri kullanarak kitle algılama uygulamaları üzerinde yapılan çalışmalara artan bir ilgi olduğu görülmektedir.

Ayrıca, gelişen teknolojiyle beraber depolanan veri miktarları da hızla çoğalmaktadır. Bu durum, oluşan veri yığınlarından ihtiyaç duyulan bilgiyi elde etme problemini de beraberinde getirmiştir. Veri madenciliği, büyük hacimli, ticari veya bilimsel veri setlerinden işe yarar bilgiler elde etme ile ilgili sorunları ele alan yeni ve hızla gelişen bir alandır. Bu bağlamda sağlık sektörü de veri madenciliği disiplinin etkin kullanılabileceği alanlardandır.

Bu tezde tıp alanında sıkça kullanılan mikrodalga saçılımı ve veri madenciliği yöntemleri biraraya getirilerek, beyindeki kitlesel lezyonların tespitine dair bir çalışma sunulmuştur. Çalışmada öncelikle gerçekçi beyin modellerinden saçılan elektromagnetik alan Moment Yöntemi ile sentetik olarak hesaplanmıştır. Üretilen sentetik veriler kullanılarak çeşitli veri madenciliği teknikleriyle beyin lezyonlarının var olup olmadığına dair sonuçlar elde edilmiş ve yöntemlerin başarı oranları karşılaştırılmıştır.

Anahtar Kelimeler: Kitlesel lezyon, veri madenciliği, biyomedikal görüntüleme

ABSTRACT

MICROWAVE BRAIN MASS LESION DIAGNOSIS BY USING DATA MINING

Tuğba VURAL

Department of Mathematical Engineering

MSc. Thesis

Adviser: Yrd. Doç. Dr. Birol Aslanyürek

Microwave tomography is one of the effective diagnostic methods used to visualize the electromagnetic material properties of non-metallic objects. In recent years it appears to be an increasing interest in studies on mass sensing applications by using microwave imaging methods in biomedicine.

In addition, the amount of data stored with the developing technology is increasing rapidly. This condition has brought about the problem of obtaining the needed information from the data chunks. Data mining is a new and rapidly evolving field that deals with the issues of obtaining useful information from large volumes of commercial or scientific data sets. In this context, health sector is also the field where data mining discipline can be used effectively.

In this thesis, through bringing together of microwave scattering and data mining methods which are frequently used in medical field, a study on the detection of mass lesions in the brain is presented. In the study, firstly the electromagnetic field scattered from realistic brain models is synthetically calculated by method of Moment. Using the generated synthetic data, with various data mining techniques, the results whether the brain lesions exist or not are gained and the methods which become successful have been compared.

Key words: Mass lesion, Data mining, Biomedicine imaging

BÖLÜM 1

GİRİŞ

1.1 Literatür Özeti

Beyindeki anormal doku kısımları beyin lezyonları olarak adlandırılır. Testler sonucu teşhis konulana kadar birçok farklı anormalliğe beyin lezyonu denilebilir. İnme, tümörler ve anevrizmalar beyinde en yaygın görülen lezyonlardır [1].

Beyin tümörü, beyin hücrelerinin anormal bir şekilde veya kontrolsüz olarak büyümesi şeklinde nitelendirilir. Beyin kafatası ile çevrili olduğundan tümörlü hücrelerin büyümesi normal beyin dokularına basınç uygulamaya başlaması anlamına gelir ve bu durum iltihaba ve beyin şişmesine sebep olabilir [2].

Anevrizmalar çoğunlukla damarların çatallanma bölgelerinde görülür. Beyindeki atardamar duvarının zayıflaması sonucu ortaya çıkan balonlaşmaya anevrizma denir. Anevrizmalar damar duvarlarında incelmeye ve zayıflamaya sebep olur. Damarın zayıfladığı bu yerden yırtılması sonucu oluşan kanamalar inmeye, komaya hatta ölüme neden olabilir [3]. Beyindeki bu tür lezyonların erken saptanması hastada kalıcı hasar olmaması veya hastanın hayatta kalması için ciddi önem arz etmektedir.

Beyin lezyonlarının görüntülenmesi veya tespiti için çoğunlukla BT (Bilgisayarlı Tomografi) ve MRG (Manyetik Rezonans Görüntüleme) yöntemleri kullanılır [4].

1972 yılında Sir Godfrey Hounsfield'in ilk bilgisayarlı tomografi cihazını geliştirmesiyle kullanılmaya başlanan BT, vücudun incelenen bölgesinin kesitsel görüntüsünü oluşturmak için x- ışını kullanılarak uygulanan bir radyolojik teşhis yöntemidir. Kesit düzlemindeki her noktanın x-ışınını zayıflatma değeri bulunur. 2000'lere gelindiğinde

ise çoklu kesit BT görüntüsü ile üç boyutlu görüntüler oluşturularak daha fazla bilgi elde edilmektedir. Ancak bunun yanı sıra çeşitli dezavantajları da mevcuttur. BT esnasında bir röntgen filmine göre 10 kat daha fazla radyasyona maruz kalınır. Kullanılan radyasyon hem vücuttaki organların hem de çekim yapılan ortamdaki havanın yapısında iyonlaşmaya sebep olmaktadır. İyonlaşma ise DNA'ya zarar verebilir. Bunun sonucunda ise hücre ölümleri veya kalıtsal bozukluklar meydana gelebilir [4].

1973 yılında Lauterbur ve Damadian'ın Manyetik Rezonansı geliştirmesiyle başlayan süreçle birlikte 1990 yılına gelindiğinde Ultrafast MR görüntüleme tekniği ortaya kondu ve beynin fonksiyonel MR incelemesi geliştirildi. MR tekniğinde, büyük mıknatıslarla güçlü bir manyetik alan oluşturulur ve bu alan içinde görüntülenmek istenen bölgeye radyo dalgaları gönderilir. Gönderilen radyo dalgaları bölgedeki dokulara çarpıp geri döner ve verileri bilgisayar ortamına görüntü olarak yansıtır. Böylece sağlıklı ve hasta dokular arasındaki farklılıklar saptanır. BT 'ye oranla daha zararsız olsa da kalp pili taşıyan hastalar ve hamileliğin erken dönemlerindeki kişiler için riskli bir yöntemdir. Ayrıca MRG yöntemi kapalı yerde bulunma korkusu (klostrofobisi) olanlar için rahatsız edici olabilmektedir [5].

BT ve MRG cihazlarının yüksek maliyetli oluşları ve kolay taşınabilir olmayışları da uzmanları alternatif aramaya yönlendiren etmenlerdendir. Mikrodalga frekanslarında beyindeki sağlıklı dokular ve lezyonların elektromagnetik malzeme özellikleri arasında kayda değer bir fark olması, beyindeki lezyonların tespitinde mikrodalgaların kullanılmasına olanak verir. Bu nedenle beyindeki kitlesel yapıdaki lezyonların tespitinde son yıllarda mikrodalga teknolojileri ön plana çıkmıştır. Mikrodalga ışınlarının kullanılmasının yukarıda bahsedilen yöntemlere nispeten bazı üstünlükleri bulunmaktadır. Diğer yöntemlerin aksine mikrodalgalar iyonize olmayan ışıma yaparlar bu yüzden uzun süreli ve güvenli görüntüleme olanağı doğar. Ayrıca mikrodalga teknolojisine dayanan cihazlar göreli olarak daha kolay taşınabilirdir ve düşük maliyetlidir [3].

Söz konusu avantajları nedeniyle beyindeki lezyonların teşhisi için son yıllarda çeşitli mikrodalga teknikleri geliştirilmiştir. Bu teknikler kabaca ultra geniş bantlı radar tabanlı yöntemler [6], [7] ve mikrodalga tomografisi [8-12] olarak ikiye ayrılır. Ultra geniş bantlı

radar tabanlı yöntemlerde öncelikle kafatası mikrodalga sinyalleri ile aydınlatılır ve geri saçılmış enerji (backscattered energy) dağılımı kullanılarak lezyonlar tespit edilmeye çalışılır. Mikrodalga tomografisinde ise öncelikle, beyin, etrafına yerleştirilmiş verici mikrodalga antenleriyle aydınlatılır ve yine beynin etrafına yerleştirilmiş mikrodalga alıcı antenlerle beyinden saçılan elektromagnetik alan ölçülür. Beyindeki sağlıklı dokuların ve varsa lezyonların bilgisini içeren saçılan alan için ters saçılma probleminin çözümü Newton Yöntemi [8], Doğrusal Örnekleme Yöntemi (Linear Sampling Method) [9], İteratif Born Yöntemi [10], Kontrast-Kaynak (Contrast Source) Yöntemi [11] ve İteratif Gauss-Newton Yöntemi [12] gibi sayısal yöntemlerle aranır.

Yukarıda sözü edilen mikrodalga tekniklerinde kullanılan ters saçılma problemleri kötü kurulmuş (ill-posed) olduğu için çözümün varlığı veriye sürekli bir şekilde bağlıdır [13]. Bundan dolayı; buna benzer karmaşık matematiksel yapılar içeren problemlerin çözümü özellikle beyin gibi komplex yapılar için çeşitli zorluklar içerir ve en küçük ölçüm hataları bile birçok durumda hatalı sonuçlara neden olur. Bu zorluklarla uğraşmamak için sınıflandırmaya dayalı bazı veri madenciliği yöntemleri kullanılabilir. Oysaki mikrodalgalar aracılığıyla beyin lezyonlarının tespitine dair yapılan çalışmaların sayısı çok fazla değildir [14-16]. Sundström 2014 yılında [14]'da en küçük kareler yöntemine dayanan iç çarpım alt uzayı sınıflandırıcı (Inner-product subspace classifier), Destek Vektör Makinesi (Support Vector Machine, SVM) ve bir tür k en yakın komşuluk (k-nearest neighbour, KNN) algoritması kullanarak mikrodalga beyin lezyonu tespiti yapan bir çalışmayı yayınlamıştır. Wu ise 2016 yılında aynı problem için yine SVM[15] ve alt uzay sınıflandırmasına [16] dayanan birer yöntem sunmuştur.

1.2 Tezin Amacı

Bu tezin amacı, mikrodalga ölçümlerini kullanarak çeşitli veri madenciliği modelleri aracılığıyla beyindeki kitlesel lezyonları tespit etmektir. Öncelikle kafatası etrafına yerleştirilen verici antenlerle beynin aydınlatılıp, yine kafatası etrafına yerleştirilen alıcı antenlerle beyinden saçılan alanın ölçüldüğü varsayılmıştır. Bu kapsamda, elektomagnetik malzeme özellikleri açısından gerçekçi sayısal bir beyin modelinden saçılan alan Moment Metodu (MoM) kullanılarak sentetik olarak üretilmiştir. Üretilen

sentetik veriler veri madenciliği yöntemleriyle modellenip, başarı oranları karşılaştırılacak ve en başarılı model test edilecektir.

Bu tezde, ikinci ve üçüncü bölümler kısaca mikrodalga tomografisinin biyomedikal uygulamaları ve genel özellikleriyle veri madenciliğine ayrılmıştır. Tezin dördüncü bölümünde beyindeki lezyonların tespitine dair uygulama ve sonuçları ayrıntılarıyla ele alınmıştır. Son bölümde ise sonuçlar ve öneriler sunulmuştur.

1.3 Orijinal Katkı

Literatür özetinde belirtildiği üzere; mikrodalga görüntüleme problemleri karmaşık bir matematiksel yapıya sahip, kötü kurulmuş ve doğrusal olmayan denklemler halinde ifade edilirler. Bu nedenle bu problemi çözebilecek sayısal yöntemler kullanıldığında beyin gibi karmaşık yapılar için başarı oranı düşmektedir. Bu nedenlerle, veri madenciliği yöntemleri klasik sayısal yöntemlere ciddi bir alternatif oluşturmaktadır. Oysaki veri madenciliği kullanılarak mikrodalga beyin lezyonu tespiti için az sayıda çalışma yapılmıştır ve bunlarda az sayı da yöntem denenmiştir. Bu tezde ise, beyin lezyonu tespiti için çok sayıda veri madenciliği yöntemi uygulanarak, sonuçları karşılaştırılmış ve en başarılı yöntemler tespit edilmiştir.

BÖLÜM 2

MİKRODALGA TOMOGRAFİ TEKNOLOJİSİ

2.1 Biyomedikal Niceliklerin Görüntülenmesi İçin Mikrodalga Tomografi Teknolojisi

Mikrodalga tomografisi (Microwave Tomography, MWT), bilinmeyen cisimlerin tespit edilmesi ve cismi kuşatan bir dizi alıcıyla saçılan alanın ölçülerek niceliksel olarak kategorize edilmesi mantığına göre hareket eder. Bu saçılan alanlar bir veya bir kaç verici antenle gönderilen ışınlarla elde edilmektedir. Geçtiğimiz yirmi yılda MWT, meme kanseri görüntüleme, beyin görüntüleme ve akciğer kanseri görüntüleme gibi biyomedikal uygulamalar için nicel bir görüntüleme yöntemi olarak oldukça dikkat çekmiştir. Bu uygulamalar çoğunlukla diğer görüntüleme tekniklerinin eksikliklerine bağlı olarak ortaya çıkmıştır. Mikrodalgalar kullanılarak beyin görüntülemenin fizibilitesi (yapılabilirliği) uzun zamandır incelenmektedir. Bu çalışmaların çoğu inme ve kitle görüntüleme üzerine yoğunlaşmıştır. Beynin bir bölümüne kan girişi aniden kesildiğinde veya beyindeki kan damarları patladığında beyin hücrelerini çevreleyen boşluklara kan akar ve inme meydana gelir. İki çeşit inme vardır: iskemik (tıkanıklığa bağlı olarak yeterli kan akışı içinde) ve hemorajik (beynin içine veya çevresinde kanama). Şu anda inmeyi saptamak için kullanılan teknikler BT ve MRG' dir. Bunlar oldukça etkili araçlardır ancak pahalıdırlar ve her hastane de uygun donanım bulunmamaktadır. Buna ek olarak taşınabilir değillerdir ve sıklıkla görüntülemeye uygun değildir.

2.2 Biyomedikal Mikrodalga Tomografi: Avantajları ve Zorluklar

Mikrodalga görüntüleme, kanser ve inme izleme gibi biyomedikal uygulamalar için daha az maliyetli bir görüntüleme yöntemi olma potansiyeline sahiptir. Ayrıca

mikrodalga görüntüleme, noninvaziv olma (kansız olduğu için hastaya fiziksel bir zarar verme ihtimali olmayan), gerçekleştirilmesi kolay ve düşük maliyetli olma ve daha az rahatsızlık verme gibi birçok çekici özellik sunar. Mikrodalga görüntüler, cisimlerin elektriksel özellik dağılımlarının haritaları olarak tanımlanabilir. Malzemeler elektrik özellikler bakımından dielektrik katsayısı ve iletkenlik katsayısı olmak üzere iki parametre ile ifade edilirler[1]. Bu özellikler, dokudaki sıcaklık yoğunluğu ve su içeriği gibi fiziksel özelliklere bağlı olarak değişir. Cilt ve yağ gibi düşük oranda su içeren dokular, kan ve kanserli dokular gibi yüksek su içerikli dokulara göre daha düşük dielektrik katsayısı değerlerine sahiptir.

Biyomedikal nicelik görüntüleme pratik bir görüntüleme yöntemi haline gelmeden önce dikkat çekici birçok avantajının yanında bazı zorluklarla da karşı karşıyadır.

- Dokular genel olarak çok kayıplıdır (iletkenlik katsayıları yüksektir). Bu, özellikle yüksek frekanslarda mikrodalgaların penetrasyon derinliğini sınırlar.
- Mikrodalga görüntüleme, dielektrik özelliklerde kontrasta dayalıdır. Bazı dokular dielektrik özelliklerinde büyük bir kontrast gösterirken (yağ ve kanser dokuları gibi) bazılarının büyük bir kontrastı yoktur.
- Alan ölçümleri, özellikle faz ölçüm ve kalibrasyonunda çok hassas yapılmalıdır.
- Modelleme ve kalibrasyon hataları konusunda da üstesinden gelinmesi gereken zorluklar vardır [17].

2.3 Mikrodalga Tomografisinin Klinik Kullanımı

Mikrodalga donanım tasarımı ve hesaplama gücündeki son gelişmelerde MWT'nin uygun maliyetli bir tarama aracı olarak görülmesinin de katkısı büyüktür. Bu nedenle, araştırmacılar son on yılda MWT'nin klinik kullanım alanını genişletmek için birçok başarıya imza atmıştır. Bu başarılar iki ana gruba ayrılabilir: Ters saçılma problemlerini çözen algoritmalar kullanılarak hedeflenen görüntülemenin yapılması ve alıcı-verici antenleri de içeren etkili donanım düzeninin geliştirilmesi.

[1] Magnetik özellik gösteren yapılarda ayrıca magnetik geçirgenlik katsayısı olarak adlandırılan üçüncü bir parametre daha kullanılır. Beyin dokularının da dâhil olduğu magnetik olmayan yapılarda bu katsayı boş uzayın magnetik geçirgenlik katsayısıyla aynıdır.

MWT'de ele alınan yapının görüntülenmesi, ters saçılma probleminin çözülmesine dayanır. Bu problem genellikle ölçülen ayrık elektrik alanlar ile ileri sayısal çözüm teknikleriyle (Kontrast Kaynak Yöntemi gibi) hesaplanan alanlar arasındaki hatanın asgariye indirildiği bir optimizasyon problemi olarak ele alınır. Ortaya çıkan görüntünün çözünürlüğü, ileri sayısal çözüm tekniğinin karmaşıklığını belirler. Optimizasyon arama alanı olası dielektrik özelliklerine ve doku bileşimlerine bağlıdır. Dolayısıyla doğru dielektrik özelliklere sahip yüksek çözünürlüklü bir görüntü yoğun bir hesaplama problemi demektir. Buna ek olarak problem hatalı olabilir, çözüm benzersiz olabilir veya hiç mevcut olmayabilir. Bu durum, sınırlı örnekleme noktalarından kaynaklanmaktadır. Ek olarak çözüm istikrarsız olabilir yani olay alanındaki küçük bir değişiklik çözümde büyük bir değişiklik meydana getirebilir. Bu zorluklar matematiksel açıdan problemin kötü-kurulmuş olması anlamına gelir. Bu durumun üstesinden gelmek için birçok etkin ve düzenleyici yöntem önerilmiştir. Tikhonov [18], Krylov [19] ve çarpıcı regülerizasyonlar; [20] gauss-newton inversion [21] gibi yerel ve deterministik optimizasyon yöntemlerinde kullanılan değişkenlerden bazılarıdır. Görüntülenmek istenen yapının karmaşıklığı artıkça ters saçılma problemlerinin çözümünde kullanılan sayısal yöntemlerin başarısı azalmaktadır.

BÖLÜM 3

VERİ MADENCİLİĞİ

3.1 Veri Madenciliğinin Tarihçesi

1950'li yılların başlarında sayımlar için ilk bilgisayarların kullanılmasıyla veri madenciliğinin temelleri atılmıştır. 1960'larda verilerin artmasıyla veri kolleksiyonları oluşmuş ve bu da veri tabanlarını (hiyerarşik ve ağ modelleri) ortaya çıkarmıştır. 1970'lere gelindiğinde ilişkisel veri modeli kurulmuş ve ilişkisel veri tabanı yönetim sistemleri uygulamaları başlamıştır. 1980'lerde ilişkisel Veri Tabanı Yönetim Sistemlerinin (VTYS) yaygınlaşmasıyla birlikte mekansal, bilimsel, mühendislik vs. gibi uygulamaya yönelik VTYS oluşturulmuştur. 1990'larda ise rutin işlemlerden derlenen büyük miktarda verilerin optimum şekilde nasıl kullanılabileceği araştırılmaya başlanmıştır. Bu aşamada gerçekleşen en önemli olaylardan bir kaçı ise şunlardır: 1989'da Knowledge Discovery Databases (KDD)-89 Veri tabanlarında bilgi keşfi çalışma grubu toplantısı yapılmıştır. 1991'de (KDD)-89'un sonuç bildirgesi olarak gösterilebilecek 'Knowledge Discovery in Real Databases: A Report on the KDD-89 Workshop ' makalesi yayınlanmıştır. 1992'de veri madenciliği konusunda ilk yazılım geliştirilmiştir. 1995'de 1. Uluslararası Bigi Keşfi ve Veri Madenciliği Konferansı'nın (KDD-95) açılışı yapılmıştır. 2000'lere gelindiğinde ise geliştirilen veri ambarlarını ileri düzeyde algoritmalar, çok işlemcili bilgisayarlar ve büyük veri tabanları takip etmektedir [22].

3.2 Veri Madenciliğinin Önemi

Verileri dönüştürmenin geleneksel metodu, el ile yapılan analizlere ve yorumlara dayanır [23]. Örneğin sağlık sektöründe uzmanlar periyodik olarak üç ayda bir mevcut eğilimleri ve değişiklikleri analiz ederler. Daha sonra sponsor sağlık örgütlerine ayrıntılı bir rapor hazırlarlar. Bu rapor sağlık yönetimi için gelecek adına karar verme ve planlama adına temel oluşturur. Bilim, pazarlama, finans, sağlık, perakende ya da bunlara benzer başka bir alanda klasik veri analizi,bir veya daha fazla analistin veri ile haşır neşir olmasına ve veriler, kullanıcılar ve ürünler arasında bir arayüz olarak hizmet vermesine dayanır. Bunlar ve diğer bir çok uygulama için her geçen gün hacmi artan veri setlerini bu formda ayıklama ve kullanılır hale getirme yavaş, pahalı ve son derece özneldir. Dolayısıyla veri madenciliği, hepimizin yaşamının bir parçası haline gelmiş bu sorunu çözmek için ortaya konulan bir girişimdir [24].

3.3 Veri Madenciliği Uygulama Alanları

Veri madenciliği verilerin sürekli ve yoğun olarak üretildiği her alanda kullanılabilir. En yaygın kullanım alanları şöyle özetlenebilir [25], [26]:

Bilişim ve Mühendislik: Günümüzde bilgisayar, internet ve medya teknolojilerindeki olağanüstü gelişmeler karar vericileri bir veri okyanusuyla karşı karşıya bırakmıştır. Bu yüzden veri madenciliğinin en yaygın kullanıldığı sektörlerdendir. İnternet dolandırıcılığının tespit edilmesinde, bilgisayar sistemlerine ve ağlarına izinsiz erişimin saptanmasında, biyolojik kimlik tespitinde ve yapay zeka uygulamalarında kullanılmaktadır.

Bankacılık: Farklı finansal durumlar arasındaki gizli ilişkilerin bulunmasında, kredi kartı dolandırıcılıkların tespitinde, risk analizleri ve risk yönetiminde, müşteri eğilim analizi ve kar analizi gibi alanlarda kullanılır.

Sigortacılık: Yeni poliçe talep edecek müşterilerin tahmin edilmesinde, riskli müşteri formatının belirlenmesinde, dolandırıcılıkların tespitinde kullanılır.

Pazarlama: Sepet analizi yaklaşımıyla alınan ürünler arasındaki ilişkileri belirleyip işletmenin karının artırılmasında, müşterilerin rutin alışkanlıkları arasındaki ilişkinin

saptanmasında ve uygun modellerin kurulmasında, satın alma eğilimlerinin ve pazarlama kampanyalarının ve stratejilerinin belirlenmesinde kullanılır.

Sağlık, Biyoloji ve Farmakoloji: Tıp bilimi, tıp ve bilgi teknolojilerin sürekli kesiştiği, tıbbi durumlara çözümler sunan, biyomedikal verileri saklayan ve bu verilere hızlı erişim ve optimum kullanımlarını sağlayan çok disiplinli bir bilim dalıdır. Ayrıca, tıp bilimi tıbbi bilginin elde edilmesi, derlenmesi, paylaşılması, kullanılması, hastalar için tedavi ve bakım yöntemlerinin belirlenmesi ve bu yöntemlerin geliştirilmesi ile ilgilenir. Tıptaki bilgilerin kullanımındaki değişim sağlık hizmeti sunan kişileri fazlasıyla etkilemiştir. Bilgisayarların kullanımı, verilerin paylaşımı, ekip çalışması ve bilgi kavramlarına dayalı uygulamalar daha yaygın hale gelmiştir. Bilgisayarlar, hasta bakım hizmetlerine destek sağlama, sağlık hizmetlerinin kalitesinin kontrol edilmesi ve değerlendirilmesi gibi doğrudan tıbbi hizmetleri sağlamanın yanı sıra; karar verme, yönetim, planlama ve tıbbi araştırma gibi yönetsel ve akademik işlevlerde daha yaygın olarak tercih edilmeye başlamıştır. Veri madenciliği, büyük veri tabanlarından gizli kalıpları keşfetme sürecidir. Geleneksel metotları kullanarak çözmenin uzun zaman alacağı işlemler, veri madenciliği sürecini kullanarak daha hızlı ve kolay bir şekilde çözülebilir. Bu yüzden, veri madenciliğinin tıp biliminde popülaritesi gitgide artmaktadır. Tarama testlerinden elde edilen verileri kullanarak çeşitli kanserlerin ön tanısında, acil servislerde hasta semptomlarına göre risk ve önceliklerin tespitinde, İlaç geliştirmede, hastalıkların tespiti ve tedavi sürecinin belirlenmesinde, DNA sıra analizi ile hastalıklara neden olan gen sıralamasının belirlenmesinde kullanılır [27].

Eğitim Sektörü: Öğrenci işlerinde veriler analiz edilerek öğrencilerin başarı ve başarısızlık nedenleri, başarının artırılması için tahminler yürütülmesi, okul başarısı ile üniversite giriş puanları arasında bir ilişkinin var olup olmadığı gibi alanlarda kullanılıp okulun başarı performansı artırılabilir.

Meteoroloji ve Atmosfer Bilimleri: Bölgesel iklim ve yağış haritaları oluşturma, hava tahminleri yapma gibi alanlarda kullanılır.

3.4 Veri Madenciliği Eksikleri ve Uygulamada Karşılaşılan Zorluklar

Etkin bir veri madenciliği süreci yönetmek için duyulan ihtiyaçlardan ilki, bilgi keşfi sisteminden ne tür özniteliklerin bekleneceğinin ve veri madenciliği tekniklerinin gelişiminde ortaya çıkabilecek güçlüklerin tanımlanmasıdır.

a) Farklı Tip Verilerin Yönetimi

Bilgi keşfi sistemlerinde, etkili bir veri madenciliği yapılabilmesine olanak tanıyan, farklı uygulamalarda kullanılan çok fazla veri ve veri tabanı tipi vardır. Mevcut veri tabanlarının çoğu ilişkiseldir bu yüzden veri madenciliği sistemi ilişkisel veri tabanları üzerinde etkili bir bilgi keşfi gerçekleştirir. Geçerli birçok veri tabanı türü, multimedya verileri, uzaysal ve zamansal veriler gibi karmaşık veri türleri içerir. Bu tip karmaşık veri tiplerinde etkin veri madenciliği için güçlü bir sistem kullanılmalıdır. Ne var ki veri tiplerinin ve madencilik amaçlarının çeşitliliği tüm veri tipleri için tek bir sistem kullanılmasını imkansız kılar. İşlem verileri, uzaysal veriler, multimedya verileri gibi belirli veri türleri üzerine bilgi madenciliği için özel veri madenciliği sistemleri oluşturulmalıdır.

b) Veri Madenciliği Algoritmalarının Verimliliği ve Ölçeklenebilirliği

Veri tabanlarında büyük miktardaki verilerden etkili bir şekilde bilgi ayıklamak için, bilgi keşfi algoritmaları etkin ve ölçeklenebilir olmalıdır. Yani, bir veri madenciliği algoritmasının çalışma süresi, büyük veri tabanlarında öngörülebilir ve kabul edilebilir olmalıdır. Üstel veya orta dereceli polinom karmaşıklığına sahip algoritmalar kullanımda pratik olmayacaktır.

c) Veri Madenciliği Sonuçlarının Yararlılığı, Kesinliği ve İfade Gücü

Bulunan bilgi, veri tabanının içeriğini doğru şekilde tasvir etmeli ve uygulamalar için faydalı olmalıdır. Hatalar, yaklaşık kurallar ve niceliksel kurallar formunda belirsizlik ölçüleri ile ifade edilebilmelidir. Gürültülü ve standart dışı veriler daha hassas ele alınmalıdır. Bu aynı zamanda, istatistiksel, analitik ve simülatif modeller inşa ederek keşfedilen bilginin kalitesini ölçmek için sistematik bir çalışmayı da destekler.

d) Çeşitli Veri Madenciliği Sonuçlarının İfadesi

Büyük miktarda veri setlerinden farklı bilgi türleri keşfedilebilir ve bu bilgi farklı bakış açılarından incelenmek ve farklı şekillerde sunulmak istenilebilir. Bu hem veri madenciliği taleplerini hem de keşfedilen bilginin üst düzey grafik kullanıcı ara yüzlerinde ifade edilmesini gerektirir. Böylece keşfedilen bilgiler, veri madenciliği uzmanı olmayan kişilerce anlaşılabilir ve doğrudan kullanılabilir. Bu aynı zamanda keşif sisteminin, anlamlı bilgi gösterim tekniklerini benimsemesini gerektirir.

e) Çoklu Soyutlama Seviyelerinde İnteraktif Madencilik Bilgisi

Bir veri tabanından tam olarak neyin keşfedileceğini tahmin etmek zordur. Üst düzey bir veri madenciliği sorgusu, daha ileri araştırmalar için bazı ilginç izler gösterebilecek bir araştırma olarak ele alınmalıdır. İnteraktif keşif teşvik edilmelidir. Bu da, etkileşimli olarak veri madenciliği talebini yeniden yapılandırır, veriye odaklanmayı dinamik olarak değiştirir, süreci kademeli olarak derinleştirir ve sonuçları birden fazla soyutlama seviyesinde ve farklı açılardan görüntülemeyi sağlar.

f) Farklı Veri Kaynaklarından Bilgi Madenciliği

İnternet de dahil olmak üzere yaygın olarak bulunan yerel ve geniş alanlı bilgisayar ağları, bir çok veri kaynağı ile bağlantı kurmakta ve büyük, dağınık, heterojen veri tabanları oluşturmaktadır. Bu durum bilgi keşfi için yeni zorluklar ortaya koymaktadır. Veri madenciliği heterojen veri tabanlarında basit sorgu sistemleri tarafından pek keşfedilmemiş üst düzey veri ilişkilerini açıklamaya yardımcı olabilir.

g) Gizlilik ve Veri Güvenliğini Koruma

Verilerin birçok açıdan ve farklı soyutlama düzeylerinde görülebilmesi gizlilik ihlali tehdidi oluşturur ve bu durum veri güvenliğini korumayı gerektirir. Veri madenciliği sırasında hassas bilgilerin ifşa edilmesinin önlenmesi için hangi güvenlik önlemlerinin geliştirilebileceğini incelemek önemlidir [28].

3.5 Veri Madenciliğinin Yararlandığı Disiplinler ve Teknolojiler

Uygulamaya dayalı bir alan olarak veri madenciliği, istatistik, makine öğrenimi, veri tabanı ve veri ambarları sistemleri, görselleştirme, algoritmalar gibi bir çok tekniği sistemine dahil etmiştir. İstatistik ve Makine Öğrenimi aşağıda kısaca özetlenmiştir.

3.5.1 İstatistik

İstatistik, verilerin toplanması, analizi, yorumlanması ya da açıklanması ve sunumu üzerine çalışmalar yapar. Bu bağlamda veri madenciliği istatistikle iç içe bir ilişki halindedir. İstatistiksel bir model, rastgele değişkenler ve bunların ilişkili olasılık dağılımları açısından hedef sınıftaki nesnelerin davranışlarını tanımlayan bir dizi matematiksel işlevdir. İstatistiksel teknikler veri ve veri sınıflarını modellemek için yaygın olarak kullanılmaktadır. Örneğin veri karakterizasyonu ve sınıflandırma gibi veri madenciliği görevlerinde hedef sınıfların istatistiksel modelleri oluşturulabilir. Alternatif olarak, veri madenciliği görevleri istatistiksel modellerin üzerine kurulabilir. Örneğin, gürültüyü ve eksik veri değerlerini modellemek için istatistik kullanılabilir. Sonrasında veri madenciliği veriyi işleme sürecinde bu modeli kullanır. Geniş bir veri kümesi üzerindeki bilgi keşfi sürecinde tamamen istatistiksel yöntemlerin ölçeklendirilmesi çoğu zaman ciddi bir zorluk oluşturmaktadır. Birçok istatistiksel yöntem hesaplamada ciddi karmaşıklığa sahiptir ve hesaplama maliyetini azaltmak için algoritmalar dikkatle tasarlanmalıdır.

3.5.2 Makine Öğrenimi

Makine öğrenimi, bilgisayarların verilere dayalı olarak nasıl öğrenebileceğini (veya performanslarını nasıl geliştirdiğini) araştırır. Temel araştırma alanı, karmaşık kalıpları tanıyan ve verilere dayalı akıllı kararlar veren bilgisayar programlarıdır. Örneğin, bir dizi örnekten öğrendikten sonra mail üzerindeki el yazısı posta kodlarını otomatik tanıyan bir bilgisayar programı, makine öğrenimi problemidir. Makine öğrenimi oldukça hızlı gelişen bir disiplindir. Aşağıda, veri madenciliği ile yakından ilişkili olan temel makine öğrenme problemlerinden bahsedilmiştir [29].

3.5.2.1 Denetimli Öğrenme

Denetimli öğrenme temel olarak sınıflandırma ile eş anlamlıdır. Öğrenmedeki denetim, eğitim veri setindeki etiketli örneklerden gelir. İlk etapta, eğitim için bir dizi veri ve o verilerden çıkan sonuçlar verilerek girdi ve çıktı değerleri için bir fonksiyon oluşturulması ve bu şekilde modelin öğretilmesi yöntemiyle çalışır.

3.5.2.2 Denetimsiz Öğrenme

Denetimsiz öğrenme esasen kümeleme ile eş anlamlıdır. Başlangıçta örnek sınıf etiketi olmadığından öğrenme süreci denetlenmez. Tipik olarak veri seti içindeki sınıfları keşfetmek için kümeleme kullanılır. Büyük veri kümelerini etiketlemenin maliyetli olduğu durumlarda ya da sınıf etiketlerinin bilinmediği durumlarda denetimsiz öğrenme kullanılmaktadır.

3.5.2.3 Yarı Denetimli Öğrenme

Yarı denetimli öğrenme, bir modeli öğretirken hem etiketli hem de etiketlenmemiş örneklerden faydalanır. Bu yaklaşımda sınıf modellerini öğretmek için etiketli örnekler kullanılır, sınıflar arasındaki sınırları ayrıştırmak için etiketsiz örneklerden faydalanılır.

3.5.2.4 Aktif Öğrenme

Aktif öğrenme, kullanıcıların öğrenme sürecinde aktif bir rol oynamasına olanak tanıyan bir makine öğrenme yaklaşımıdır. Aktif öğrenme yaklaşımı, bir kullanıcıya (örn.alan uzmanı) etiket vermesini isteyebilir. Amaç, kişilerin bilgilerini aktif olarak elde ederek modelin kalitesini optimize etmektir.

3.6 Veri Madenciliğinde Kullanılan Yazılımlar

Veri madenciliğinde kullanılan yazılımlar ticari ve açık kaynak kodlu olmak üzere iki gruba ayrılmaktadır. Ticari yazılımlara SPSS Clementine, Angoss, SAS, KXEN, SQL Server, MATLAB ve ORACLE ın bunun için geliştirilen modülleri örnek verilebilir. Açık kaynak kodlu yazılımlardan bazıları ise Weka, RapidMiner, R, Keel, Knime örnek olarak verilebilir [30].

3.6.1 Weka

Waikato Environment for Knowledge Analysis kelimelerinin kısaltmasıdır. Java tabanında geliştirilmiştir, makine öğrenmesi algoritmaları içeren ve SQL veri tabanlarına erişim sağlayabilen bir veri madenciliği yazılımıdır. Weka ya özel olarak tasarlanmış .arff dosya formatı ile çalışmaktadır. Veriler üzerinde sınıflandırma, kümeleme, özellik seçimi ve görselleştirme yapabilmektedir. Bu çalışmada veri madenciliği yazılımı olarak Weka kullanılmıştır.

3.6.2 RapidMiner

RapidMiner, Ralf Klinkenberg, Ingo Mıerswa ve Simon Fischer tarafından Dortmund Teknoloji Üniversitesi Yapay Zeka biriminde geliştirilmiştir. Diğer veri madenciliği yazılımlarından farklı olarak 22 adet dosya formatındaki veriyi işleyebilmektedir. Ön işleme, veri analizi, sınıflama, birliktelik kuralları çıkarımı gibi işlemleri içermektedir. Excell dosyalarıyla bağlantı kurabilir ve içerisinde script yazılabilir. Çok çeşitli veri tabanlarını desteklediği için en kapsamlı yazılımlardan biri sayılır.

3.6.3 R

İstatistiksel hesaplamalar, veri analizleri ve grafikler için Auckland Üniversitesi bilim adamları tarafından geliştirilmiştir. Bilgi keşfi sürecinde çok fazla tercih edilmemektedir.

3.6.4 Keel

Weka gibi Java tabanlı bir yazılımdır. Granda Üniversitesi tarafından geliştirilmiştir. Yapay zekaya dayanan sınıflama ve Kural tabanlı kümeleme algoritmalarının çoğunu içerir ancak veri görselleştirme açısından oldukça zayıftır.

3.6.5 Knime

Konstanz Üniversitesi görsel veri madenciliği araştırma grubu tarafından Eclipse Rich Client Platformu üzerinde geliştirilen, genişletilebilme özellikleri ile ön plana çıkan bir yazılımdır. Knime, .txt, .arff, .table formatında dosyalardan veri alabilmektedir.

3.7 Veri Madenciliği Metodolojisi

Veri madenciliği, bilgi keşfi kavramının ortaya çıkmasında sürecin bir parçası olsa da zamanla kısmen de olsa bilgi keşfi kavramıyla benzer olarak kullanılmaya başlanmıştır. Dolayısıyla veri madenciliğinin gelişimi ile birlikte bir metodoloji ortaya koyma ihtiyacı doğmuştur. Veri madenciliği için kullanılan en yaygın iki metodoloji, SEMMA VE CRISP-DM' dir.

3.7.1 SEMMA

SEMMA metodu Statistical Analysis Software (SAS) Enstitüsü tarafından hazırlanmış olup veri madenciliği aşamalarını organize etmek için baz alınan genel bir yapıdır. SAS Enstitüsü SEMMA nın bir veri madenciliği metodolojisi olmadığını, temel görevler için veri madenciliği işlev aracı setinin yerine getirmesi gereken mantıksal bir organizasyon olduğunu iddia etmektedir. Bununla birlikte SEMMA, uygulamada yaygın olarak veri madenciliği metodolojisi olarak kullanılmaktadır [31]. SEMMA bir akronimdir ve Sample (Örnekle), Explore (Keşfetme), Modify (Değiştirme), Model (Modelleme), Assess (Değerlendir) anlamlarına gelmektedir [32]:

a) Örnekleme: SAS enstitüsü, veri madenciliği işinde veri setini güvenilir bir şekilde temsil edebilecek, kolaylıkla işlem yapabilecek kadar küçük aynı zamanda büyük veri setindeki en önemli bilgileri aktaracak kadar da büyük bir veri setini örnekleme stratejisini savunmaktadır. Bütün veri seti yerine en önemli bilgileri aktarabilecek temsilci örneklemi madenlemek işlem süresini oldukça düşürecektir. Eğer bir örüntü temsilci örneklemde gözlemlenmiyorsa bütün veri setini etkileyecek kadar önemli değildir denilebilir.

b) Keşfetme: Keşfetme, önceden bilinmeyen, beklenmeyen ilişkilerin ve anomalilerin araştırılması anlamına gelir böylece araştırma süreci hızlandırılmış olur.

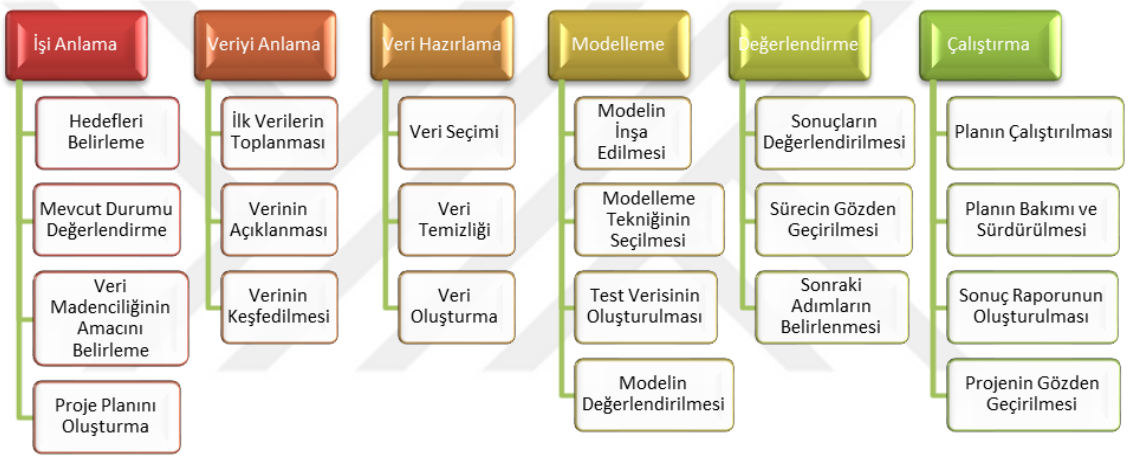
c) Değiştirme: Model seçme aşamasına odaklanır ve değişkenler oluşturulup seçerek veya dönüştürerek veri değiştirilebilir.

d) Modelleme: İstenilen çıktıyı en iyi oranla tahmin edebilecek veri kombinasyonu için bir model oluşturulur. Ağaç tabanlı modeller, sinir ağları, lojistik modeller, temel bileşenler analizi, zaman serisi analizi gibi modeller veri madenciliğinde kullanılan modelleme teknikleridir.

e)Değerlendir: Süreçte keşfedilmiş bilginin güvenilirliği, yararının değeri ve performansının ne kadar iyi olduğu değerlendirilir. Değerlendirmede yaygın olarak kullanılan bir yöntem de modeli, örnekleme aşamasında kullanılmayan verilerin bir kısmına uygulamaktır [33].

3.7.2 CRISP-DM

CRISP-DM (Cross-Industry-Standart Process for Data Mining) 1996 nın sonlarına doğru tasarlanan veri madenciliği için standart bir süreç modelidir. Uygulamaya dönük, aşamaları katı olmayan yani her aşamada ileri geri harekete izin veren dinamik bir sürece sahip olan 6 adımlık bir döngüden oluşmaktadır [34].



Şekil 3.1 Crispi-DM süreci

1.İşi anlama: İlk aşama projenin gereksinimlerini anlamaya ve bu bilgiyi veri madenciliği problemi tanımına dönüştürmeye yönelik hedeflere ulaşmak için tasarlanmış bir ön plana odaklanmaktadır.

2.Veriyi Anlama: Veriyi anlama aşaması ilk veri toplama işlemiyle başlayan ve verilere aşina olma, veri kalitesi sorunlarını tanımlama, verilere ilişkin ilk bakış açısını keşfetme, gizli bilgiler için ilginç alt grupları tespit etme aktivitelerine kadar devam eden süreçtir.

3.Veri Hazırlama: Veri hazırlama süreci, başlangıçtaki ham verilerden son veri kümesini oluşturmak için yapılan tüm faaliyetleri kapsamaktadır.

4.Modelleme: Bu kısımda, çeşitli modelleme metotları seçilir ve uygulanır ve modellerin parametreleri optimum değerlere göre ayarlanır.

5.Değerlendirme: Bu aşamada veri analizi perspektifinden, yüksek kaliteye sahip olduğu sonucuna varılan bir model oluşturulmuş olur. Modelin son çalıştırılma kısmına geçmeden önce, modelin daha kapsamlı bir şekilde değerlendirilmesi ve uygulanan adımların gözden geçirilmesi ve hedeflerin doğru bir şekilde yerine getirildiğinden emin olmak önemlidir. Bu aşamanın sonunda veri madenciliği sonuçlarının kullanımı ile ilgili bir karara varılmalıdır.

6.Çalıştırma: Modelin oluşturulması genellikle projenin sonu değildir. İhtiyaçlara bağlı olarak, çalıştırma aşaması rapor oluşturmak kadar basit olabilir veya tekrarlanabilir bir veri madenciliği süreci uygulayacak kadar karmaşık olabilir.

3.8 Veri Madenciliği Metotları

Veri madenciliği metotları, dayandıkları fonksiyonlar baz alınarak; sınıflandırma ve regresyon, kümeleme ve birliktelik kuralları olmak üzere 3 temel gruba ayrılabilir.

3.8.1 Sınıflandırma ve Regresyon

Sınıflandırma ve regresyon, önemli veri sınıfları üretebilen veya gelecekteki eğilimi öngörebilecek modeller oluşturabilen veri analiz yöntemleridir. Sınıflandırma kategorik değişkenleri öngörürken, sürekli değişkenleri tahmin etmek için regresyon kullanılır. Sınıflandırma ve regresyonda kullanılan temel teknikler şunlardır:

- Naive Bayes
- Karar Ağaçları
- k-En yakın komşu algoritması
- Yapay sinir ağları
- Genetik Algoritmalar
- Destek Vektör Makinesi (Support Vector Machine,SVM)

3.8.1.1 Naive Bayes

Naive Bayes, temelini Bayes teoreminden alan bir sınıflandırma metodudur. Naive Bayes sınıflandırması, olasılık kurallarına göre tanımlanmış hesaplamalar ile girilen verilerin kategorisini tespit etmeye çalışır.

Algoritma birden fazla sınıftan oluşan veri setlerinde kullanılır ve her sınıfın alt örneğinden oluşan durumların gerçekleşme olasılıklarının çarpımlarının tüm sınıfın ihtimali ile çarpılıp normalize edilmesi mantığıyla çalışır. Sisteme belirli oranda öğretilmiş veri seti girilir ve bu verilerin sınıfları mutlaka belirlenmiş olmalıdır. Öğretilmiş veri seti üzerinde yapılan işlemler sonucu elde edilen olasılık değerleri ile yeni test verileri işletilir ve test verilerinin sınıfı tespit edilmeye çalışılır. Öğretilmiş veri sayısı ne kadar fazla ise test verisinin sınıfını doğru tahmin etme oranı da o kadar fazladır. Algoritmanın uygulanmasında niteliklerin birbirinden bağımsız olması gibi bazı kabuller yapılır. Çünkü eğer nitelikler birbirini etkiliyorsa bu olasılıkları hesaplamak oldukça zordur. Niteliklerin hepsinin eşit derecede önemli olduğu varsayılır. Bu durum ise algoritmanın yetersizliğidir. Örneğin, şeker hastalığı teşhis probleminde, hastanın tansiyon değeri ile kilosunun aynı değerde birer özellik olarak alınması mantıklı değildir.

Bu metot, makine öğreniminde öğreticili öğrenme kategorisindedir. Sınıflandırılması istenen kümelerin ve örneklerin ait oldukları sınıflar bilinmektedir. Yüksek hacimli, henüz tamamlanmamış ve karmaşık veri setlerinin sınıflandırılması için idealdir. Metin dokümanlarının sınıflandırılmasında yaygın olarak kullanılır. Uygulanabilirliği ve performansı sebebiyle tercih edilen bir algoritmadır. Performansı kategorik verilerde sayısal verilere göre daha iyidir. Naive Bayes algoritmasının uygulanmasındaki bir diğer zorluk ise olasılıkların başlangıç değerlerine ihtiyaç duyulmasıdır. Bu olasılıkların bilinmemesi durumunda ise çoğunlukla eldeki veriler hakkında temel bilgilere ve veri dağılımlarına dayanarak kestirilebilir.

3.8.1.2 Karar Ağaçları

Karar ağaçları sınıflandırma teknikleri arasında en yaygın kullanılan metotlardandır. Karar ağaçları kök adı verilen ve girdi almayan bir düğümle başlar ve her biri birer girdiyi ifade eden iç düğümlerden oluşur. Test düğümlerinin çıktıları başka bir düğüm için girdi olarak kullanılırken, yaprak düğümlerinin çıktıları başka bir düğüm için girdi olarak kullanılmaz. İç karar düğümleri, verilerin test edildiği, soruların sorulduğu ve yönelecekleri yönün belirlendiği düğümlerdir. Dallar, soruların cevaplarını gösterir. Uç yapraklar ise kategorinin bulunduğu sınıf etiketlerini içerir. En genel şekilde ifade

edilecek olursa; eğitilecek veri setinden ağaç yapısı oluşturulur, bu yapı test seti ile değerlendirilir ve sınıfı bilinmeyen verinin sınıfı en yüksek başarı oranıyla tahmin edilmeye çalışılır.

- **Hoeffding Tree**

Farklı alanlarda başarıyla uygulanan Random Forest, C4.5, LMT (Logistik Model Trees) gibi çeşitli karar ağacı algoritmaları bulunmaktadır. Hoeffding Tree de yaygın olarak tercih edilenlerden bir tanesidir. Hoeffding Tree, dağılımı üreten örneklerin zamanla değişmez olduğunu varsayan, büyük veri setlerinden öğrenme yeteneğine sahip artımlı bir karar ağacı algoritmasıdır. Hoeffding ağaçlar küçük bir örneğin, optimal bölen özelliği seçmek için yeterli olabileceği gerçeğinden faydalanır. Her bir düğümün nasıl parçalanacağı ile ilgili karar vermek için Hoeffding sınırı adı verilen eşitsizliği kullanır. Hoeffding eşitsizliği, beklenen değerinden ayrılan bağımsız rastgele değişkenlerin toplam olasılığı üzerinde bir üst sınır sağlar. Bu algoritma diğer karar ağacı algoritmalarına nazaran depolama sorunlarını minimize ederek problemin uygun bir hesaplama maliyeti ile sonuçlandırılmasını sağlar.

- **J-48**

J-48, en yaygın örüntü tanımlama tekniklerinden biri olan C4.5 algoritmasının yeniden yazılması ile oluşturulan bir karar ağacı algoritmasıdır. Algoritma yukarıdan aşağıya doğru bir yaklaşım izlemektedir ve buradaki en önemli işlem belirli bir düğümdeki özelliğin tespit edilmesidir. Bir veri setini parçalara bölerken önce en üst seviyeyi (kök) sonra daha alt seviyeleri (dallar ve yapraklar) inceler. Karar ağacının en üstüne yerleştirilecek olan niteliği bilgi kazanımı ya da diğer bir ifadeyle entropisi (bilgi) hesaplanarak belirlenir. Bilgi Kazanımı Teorisine (Information Gain Theory) göre ağaç yapısında tepedeki sınıfı tespit etmek için tüm alt sınıflar için bilgi kazanımı değerleri (entropi) hesaplanır ve en yüksek değeri veren özellik karar ağacına karar olarak yerleştirilir. Diğer bir ifadeyle; bu şekilde, zayıf dalları ortadan kaldırmak için bir budama işlemi yapılmış olur [35].

3.8.1.3 k- En Yakın Komşu Algoritması (kNN)

kNN makine öğrenimi algoritmalarının en basit ve anlaşılır olanlarındandır. Sınıfı bilinen eğitim setindeki gözlem değerlerinin (attribute) her birinin, sınıfı tespit edilmek istenen test gözlem değerine olan mesafeleri hesaplanır ve bu uzaklıklardan en küçük olan (en yakın olan k tane komşu) k tanesi seçilir. Seçilen bu özelliklerin sınıfları belirlenir ve test verisinin de bu sınıfa ait olması beklenir. K en yakın komşu değerini belirlemek sınıflandırmanın başarısı için önemlidir. K=1 seçildiğinde; seçilen komşu gürültülü veri ise model çok etkilenir. K çok büyük seçildiğinde ise algoritma tek sınıf gibi algılar ve sınıftaki verilerin modu seçilmiş olur. Uzaklıkların bulunmasında Öklid uzaklık formülü kullanılabilir. *p* test sınıfının eleman sayısı, x_i eğitim sınıfı elemanlarının konumu ve x_j ise test sınıfı elemanlarının konumu olmak üzere Öklid uzaklık formülü;

$$d(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \qquad (3.1)$$

şeklinde tanımlanır.

En basit sınıflandırma tekniklerinden birisi olduğu için çokça tercih edilir ve sayısal verilerde uygulanması kategorik verilere göre daha basittir. Ayrıca gürültülü verilere karşı dayanıklı olması da sınıflandırma problemlerinde özellikle tercih edilmesinin sebeplerindendir. Ancak milyonlarca verinin olduğu veri setlerinde hesaplanacak uzaklık sayısı da çok fazla olacağından işlem süresinin uzun olması ve maliyetin yüksek olması gibi dezavantajları vardır ama bazı indeksleme yöntemleriyle bu maliyet düşürülebilir.

3.8.1.4 Yapay Sinir Ağları

Yapay sinir ağları (YSA), insan beyninin en temel işlevlerinden olan öğrenme özelliğini gerçekleştirebilen bilgisayar sistemleridir. Öğrenmeyi tecrübeler yani girdi olarak verilen örnekler yardımıyla gerçekleştirirler. YSA, öğrenme yoluyla keşfedebilme, yeni bilgiler oluşturabilme gibi faaliyetleri otomatik olarak gerçekleştirebilmek amacı ile geliştirilmişlerdir. Öğrenmeye ek olarak bilgiler arasında ilişki kurma özelliğine de sahiptir. Sınıflandırma, kümeleme, tahminleme, kontrol, veri birleştirme ve kavramsallaştırma YSA'nın temel işlevlerindendir. Finansal ve endüstriyel çalışmalar,

askeri ve savunma uygulamaları, robotbilim, tıp ve sağlık sektörü, görüntü işleme ve örüntü tanıma gibi alanlarda kullanılmaktadır.

Yapay sinir ağları birbirine bağlı yapay sinir hücrelerinden ve katmanlardan oluşur. Aradaki tüm bağlantılar bir ağırlık değerine sahiptir. Bilgiler bu ağırlık değerlerinde muhafaza edilip tüm ağa yayılır. YSA 3 katmandan oluşur.

Girdi katmanı: Kullanıcıdan bilgileri alır. Bu katmanda bilgiler işleme tabi tutulmaz.

Ara katmanlar: Gelen bilgileri işlerler. Bir adet ara katman birçok probleme çözüm getirebilir. Eğer ağa öğretilmek istenen problemin karmaşıklığı artarsa ya da girdi ve çıktı arasındaki ilişki doğrusal değilse birden fazla sayıda ara katman kullanılabilir.

Çıktı katmanı: Ara katmandan gelen bilgileri işler ve ağın ürettiği çıktıyı bulur, bu çıktı kullanıcıya iletilir [36].

Bir problemde YSA kullanılabilmesi için ilk olarak girdi ve çıktı değişkenleri belirlenir. Daha sonra veriler toplanır. Veri kümesindeki her durum girdi ve çıktı değişkenlerinin değerlerinden oluşur ve eğitim ve test aşamalarında kullanılır. Veri düzenlenip değişkenler belirlendikten sonra kullanılacak ağ modeli ve ağ konfigürasyonu yani katmanlar ve katmanlardaki nöron sayıları belirlenmelidir. Her katmanda belirli sayıda nöronun bir sonraki katmanlara bağlanmasıyla ağ topolojisi oluşturulur. Probleme göre bir sinir ağı belirlemek daha yüksek performans elde etmek adına önemlidir. Ağın yapısını oluşturmak basit bir işlem değildir, deneme ve yanılmalar gerektirir. Ancak son zamanlarda ilgili yazılımların artmasıyla karmaşık ve uzun süren matematiksel modelleri çözmeye gerek kalmadan çok sayıda deneme yanılma işlemi kısa sürede yapılabilmektedir ve bu durum ağ topolojisi oluşturmayı kolaylaştırmıştır [37]. Son olarak ise ağ, geçmişe ait çeşitli girdi ve çıktı değerlerine sahip veri setini kullanıp eğitilerek aradaki ilişki öğretilir. Bu esnada bir eğitim algoritması kullanılarak hatanın en aza indirgenmesi amaçlanır [38].

3.8.1.5 Genetik Algoritmalar

Genetik algoritmalar, mantığı doğal seçim ilkelerine dayanan bir optimizasyon yöntemidir. Olasılık kurallarına göre çalışırlar, sadece amaç fonksiyonuna ihtiyaç

duyarlar ve parametre kümesini değil kodlanmış biçimlerini kullanırlar. Çözüm uzayının tamamını değil bir kısmını taradıkları için de daha kısa sürede çözüme ulaşırlar.

Çalışma mantığı ise şöyledir:

Olası çözümlerin kodlandığı bir çözüm grubu oluşturulur. Bu çözüm grubu popülasyon çözümlerin kodları da kromozom olarak adlandırılır. Popülasyondaki birey sayısı rastgele olarak belirlendikten sonra uygunluk fonksiyonu ile her kromozomun ne kadar iyi olduğu hesaplanır (evalution). Bu fonksiyon genetik algoritmanın beyni görevindedir. Algoritmanın başarısı bu fonksiyonun verimlilik ve hassaslık derecesi ile doğru orantılıdır. Daha sonra bu kromozomlar eşlenerek yeniden kopyalama ve değiştirme işlemleri uygulanır. Böylece yeni bir popülasyon oluşturulur. Kromozomların uygunluk değerlerine göre kromozomların eşlenmesi yapılır. Eski kromozomlar yok edilerek sabit büyüklükte bir popülasyon meydana getirilir ve tüm kromozomların uygunlukları yeniden hesaplanır ve popülasyonun başarısı bulunur. Algoritma sürekli çalıştırılarak çok sayıda popülasyon oluşturulup hesaplanır ve o ana kadar bulunan en iyi kromozom aranılan sonuçtur çünkü hesaplamalar yapılırken hep en iyi bireyler saklanmaktadır. [39]

3.8.1.6 Destek Vektör Makinesi (Support Vector Machine, SVM)

SVM, yaklaşık doğruluk açısından günümüzü en güçlü sınıflandırma algoritmalarındandır. Güçlü matematiksel temellere ve istatistiksel öğrenme teorisine dayanır. Farklı sınıflara ait örnekleri birbirinde ayıran pek çok doğrusal düzlem bulmak mümkündür ancak SVM farklı sınıflara ait destek vektörleri arasındaki uzaklığı maksimize eden bir ayırma hiper düzleminin tespitini hedefler. Yöntemin temelinde, farklı sonuçlara sahip örnekleri birbirinden ayıran bir hiper düzlem (hyperplane) mantığı vardır. Öncelikli olarak iki sınıflı problemler için tasarlanmış olan SVM, her iki sınıfın en yakın noktasına maksimum mesafede bulunan optimum hiper düzlemi bulur. Bulunan optimum hiper düzleme en yakın olan örnek setine destek vektör denir. Optimum hiper düzlemi bulmak doğrusal sınıflandırmayı sağlar. Çoğu zaman gerçek veri setleri için farklı sınıfların örneklerini açıkça ayıran bir hiper düzlem mevcut değildir. Bu sorunu çözmek için veri kümesini mümkün olduğunca temiz, yani bir sınıfın

mümkün olan en üst düzeyde geçerli olduğu iki kümeye bölen gevşek sınır (soft margin) yaklaşımı önerilir [40].

3.8.2 Kümeleme

Kümeleme, verilerin ortak özelliklerine göre gruplandırılması işlemidir. Aynı gruptaki nesneler diğer kümelerdekilere oranla birbirlerine daha çok benzemektedirler. Sınıflandırmanın aksine kümelemede hiçbir veri sınıfı yoktur. Bazı veri setlerinde kümeleme, sınıflandırmanın ön işlemi olabilir. Kullanılacak kümeleme algoritmasının seçimi veri türüne ve amaca bağlıdır. Belli başlı kümeleme yöntemleri şunlardır:

- Bölme Yöntemleri
- Hiyerarşik Yöntemler
- Yoğunluk Tabanlı Yöntemler
- Model Tabanlı Yöntemler
- Izgara Tabanlı Yöntemler

3.8.3 Birliktelik Kuralları

Birliktelik kuralları incelemesi, büyük veri grupları arasındaki birliktelik ilişkilerini bulmak için kullanılır. Toplanan ve depolanan veri miktarları her geçen gün arttığı için, sağlık kurumları veri tabanlarındaki birliktelik kurallarını bulmak istemektedir. Tıbbi işlemlerin büyük veri setlerindeki enteresan ilişkileri bulmak sağlık kurumlarının karar verme mekanizmalarını daha etkili hale getirmektedir. En yaygın kullanılan birliktelik kuralı Pazar sepet analizidir. Bu yöntem, sağlık kurumuna başvuran bireylerin başvuruları arasındaki ilişkileri belirleyerek, hastalara uygulanan tedavi ve bakım yöntemlerinin belirlenmesini ve iyileştirilmesini analiz etmektedir [41].

3.9 Sınıflandırma Algoritmalarının Kıyaslanmasında Kullanılan Kriterler

Veri madenciliği uygulamasında ortaya çıkacak modelin başarısı; veri ön işleme aşaması, parametre seçimi ve test kümesi seçimi ile yakından ilgilidir. Testler sonucunda elde edilen sonuçlar karışıklık matrisi ile gösterilir. Karışıklık matrisinde

satırlar kullanılan test kümesine ait gerçek değerleri, sütunlar ise modelin tahmin ettiği değerleri gösterir.

Çizelge 3.1 Karışıklık matrisi

		Tahmin Edilen Sınıf	
		Sınıf=1	Sınıf=0
Doğru Sınıf	Sınıf=1	a	b
	Sınıf=0	**c**	d

Burada; “a”, True Pozitif (TP) anlamına gelmektedir yani algoritma, kanama var sınıfındaki “a” tane örneği kanama var şeklinde sınıflandırılmıştır. “b”, False Pozitif (FP) anlamına gelmektedir yani algoritma, kanama var sınıfındaki “b” tane örneği kanama yok şeklinde sınıflandırmıştır. “c”, False Negatif (FN) anlamına gelmektedir yani algoritma, kanama yok sınıfındaki “c” tane örneği kanama var şeklinde sınıflandırmıştır. “d”, True Negatif (TN) anlamına gelmektedir yani algoritma, kanama yok sınıfındaki “d” tane örneği kanama yok şeklinde sınıflandırmıştır.

Kullanılacak modelin başarısı çeşitli kriterlere göre belirlenir. Bu kavramlar; Doğruluk-hata oranı, duyarlılık, kesinlik, F- ölçütü ve Kappa istatistiğidir.

3.9.1 Doğruluk- Hata Oranı (Accuracy- Error Rate)

Model başarımının ölçülmesinde kullanılan en popüler ve en kolay yöntemdir. Doğru sınıflandırılmış örneklerin sayısının (TP+TN), toplam örnek sayısına (TP+TN+FP+FN) oranına doğruluk oranı denir. Hata oranı ise bu değeri 1 e tamamlayan eşlenik değeridir. Yani, yanlış sınıflandırılmış örneklerin miktarının (FP+FN), toplam örnek miktarına (TP+TN+FP+FN) oranıdır.

$$\text{Doğruluk} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{3.2}$$

$$\text{Hata Oranı} = \frac{(\text{FP} + \text{FN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{3.3}$$

3.9.2 Kesinlik (Precision)

Sınıfı 1 olarak tahmin edilmiş pozitif örnek sayısının (TP), toplam pozitif örnek sayısına (TP+FN) oranıdır.

$$\text{Duyarlılık} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{3.4}$$

3.9.3 F-Ölçütü

Kesinlik ve duyarlılık ölçütleri tek başlarına anlamlı bir karşılaştırma sonucu çıkarmaya yeterli değildir. Bu yüzden her iki ölçütü beraber değerlendirmek gerekir. Bunun için kesinlik ve duyarlılığın harmonik ortalamasından oluşan F-ölçütü tanımlanmıştır. [42]

$$\text{F\_Ölçütü} = \frac{2\text{x Duyarlılık x Kesinlik}}{(\text{Duyarlılık} + \text{Kesinlik})} \tag{3.5}$$

3.9.4 Kappa İstatistiği

İki ya da daha fazla gözlemci aynı şeyi değerlendiriyorsa aralarındaki varyasyon her durumda ölçülebilir. K (Kappa katsayısı), tam uyum varsa 1 olan diğer durumlarda -1 ile +1 arasında değişen bir değerdir. Gözlenen uyum ile şansa bağlı uyumun eşit olduğu ya da gözlenen uyumun şansa bağlı uyumdan büyük olduğu durumlarda K≥0 iken, gözlenen uyumun şansa bağlı uyumdan küçük olduğu durumlarda K<0 olur. (K<0) değerlerinin güvenilirlik açısından bir anlamı olmadığından yorumlanabilir aralığı 0 ile +1 arasındadır. P_0 kabul edilen oran, P_c kabul edilmesi beklenen oran olmak üzere; Kappa değeri (3.7)' de verilen denklemle hesaplanır [43].

$$K = \frac{(P_0 - P_c)}{(1 - P_c)}. \tag{3.7}$$

BÖLÜM 4

VERİ MADENCİLİĞİ YÖNTEMLERİYLE MİKRODALGA BEYİN LEZYONU TESPİTİ

Bu çalışmada veri madenciliği süreci olarak CRISP-DM metodolojisi takip edilmiştir. CRISP-DM süreci işin anlaşılmasıyla başlar, hedefin ne olduğu, mevcut durum ve projenin planı bu aşamada belirlenir. Daha sonra, ilk verilerin toplandığı ve veri setinin niteliklerinin kavrandığı veriyi anlama aşamasına geçilir. Bu iki aşama birbiriyle geri bildirimli olarak işler. Daha sonra, veri hazırlama aşamasına geçilir ki; bu aşamada gerekiyorsa veri temizliği yapılır ve ardından kullanılacak veriler seçilir ve uygulama için veri setleri oluşturulur. Sonrasında hazırlanan veri setleri modellemede kullanılır, probleme uygun modeller bir ya da birden fazla deneme sonucu belirlenir ve test verileri oluşturularak modeller test edilir. Bu kısım da tamamlandıktan sonra değerlendirme kısmına geçilir, yapılan modelin istatistiksel açıdan geçerliliği, üretilen sonuçların doğru olup olmadığı, planlanan iş ihtiyacını karşılayıp karşılamadığı değerlendirilir. Son olarak, geçerliliğine karar verilen modeller çalıştırma aşamasına aktarılır ve analizler için kullanıma sunulur.

4.1 İşi Anlama

İşin Hedeflerini Belirleme: Beyindeki bir kan damarı patladığında kitlesel lezyonlar oluşur. Beyin dokusu etrafında biriken kan zamanında tespit edilemezse ciddi hastalıklara hatta ölümlere dahi neden olabilir. Bu sebeple çalışmanın amacı, zaman kısıtıyla beraber ciddi bir hayati risk taşıyan bu probleme çözüm getirmek adına alanın uzmanlarına farklı bir bakış açısı kazandırmak ve bununla birlikte bir çözüm önerisi sunmaktır.

Mevcut Durumu Değerlendirme: Beyinde oluşan kitlesel lezyonların görüntülenmesi ya da tespit edilmesi için farklı alanlarda çok sayıda çalışma sürdürülmektedir. Beyindeki bu tür lezyonların taranması için çoğunlukla BT ve MRG yöntemleri kullanılır. Ancak bu yöntemlerin maliyetlerinin çok yüksek olması, maruz kalınan radyasyon oranının fazlalığı gibi çeşitli dezavantajlarının olması uzmanları, iyonize olmayan ışıma yapmalarından dolayı uzun süreli ve güvenli görüntüleme imkanı sunan mikrodalga teknolojisine yönlendirmiştir. Bu yüzden bu çalışmanın temeli de mikrodalga teknolojisine dayanmaktadır.

Veri Madenciliğinin Amacını Belirleme: Teknolojinin gelişmesiyle birlikte veriler dijital ortamlarda saklanmaya başlamış ve veri tabanlarının sayısı her geçen gün daha da yüksek boyutlara ulaşmıştır. Bu durum, mevcut veri yığınlarından anlamlı ve faydalı bilgilerin elde edilmesi için veri madenciliğini zorunlu kılmaktadır. Tıbbi verilerin her geçen gün artan miktarları ve hayati önem taşımaları sebebiyle veri madenciliği teknikleri tıp alanında da popülaritesini günden güne artırmaktadır. Bu yüzden çalışmada elde edilen veriler arasındaki gizli ilişkiler veri madenciliği yöntemleri kullanılarak analiz edilmiştir.

Proje Planını Oluşturma: Bu çalışmada veriler MoM kullanılarak MATLAB uygulamasında sentetik olarak üretilmiştir. MoM, elektromagnetik saçılma problemlerinde kullanılan ele alınan cismin geometrisini ayrık hücrelere ayıran ve neticede ise integral denklemleri matris denklemler haline getiren nümerik bir tekniktir. Bu çalışmada lezyon, kanama olarak belirlenmiştir. Kanama içeren ve içermeyen beyin modelleri için MoM ile elde edilen sentetik veriler, Naive Bayes, k-en yakın komşu algoritması ve karar ağaçları gibi makine öğrenmesi tekniklerinde kullanılarak kanama olup olmadığına dair sonuçlar elde edilmiştir.

4.2 Veriyi Anlama

İlk Verilerin Toplanması: Bu çalışmada gerçekçi elektromagnetik malzeme özellikleri içeren sayısal bir kafa modeli kullanılmıştır [44]. Veriler bu kafa modeli üzerinden moment metodu (MoM) kullanılarak [45], MATLAB uygulamasında sentetik olarak üretilmiştir. Beyin, etrafına yerleştirilmiş bir mikrodalga kaynağıyla aydınlatılmış ve yine etrafına yerleştirilmiş mikrodalga alıcı antenlerle saçılan elektromagnetik alan ölçülmüştür. Saçılan elektromagnetik alan verisi beynin şekli ve malzeme özellikleriyle

birlikte, varsa beyindeki kanamalı dokunun bilgisini de içerir. Böylece saçılan alanın ölçülmesiyle elde edilen veriler yardımıyla kanamalı dokuların tespiti mümkün hale gelir.

Verinin Açıklanması: Simülasyon sonucu üretilen veriler hem bir metin dosyasına hem de bir excell dosyasına yazdırılmıştır. Metin dosyasının içeriğinde verinin hangi kaynaktan üretildiği, sağlıklı olup olmadığı ve ölçülen sayısal değerler yer almaktadır. Metin dosyaları, ölçüm yapabilmek için arff formatında bir veri setinde bir araya getirilmiştir. Excell dosyası ise, ölçüm yarıçapı, ölçüm merkezlerinin koordinatı, kaynak sayısı ve yarıçapı, kaynak merkezlerinin koordinatı, frekans, özellik sayısı ve alan değerlerini içerir.

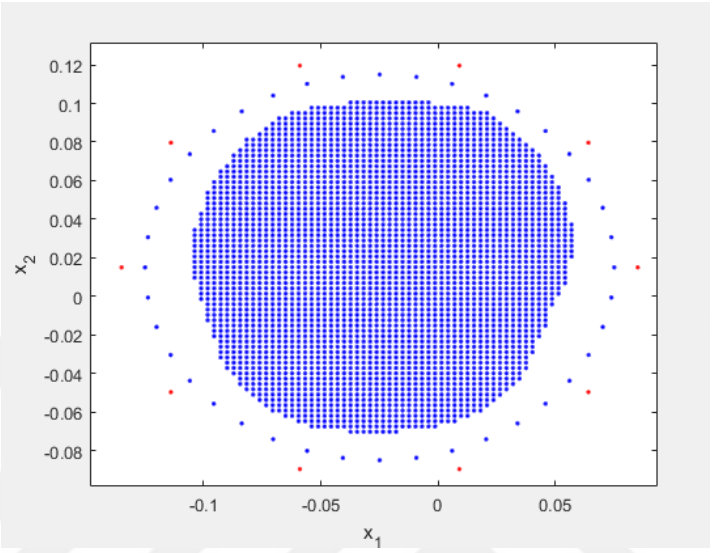
4.3 Veri Hazırlama

Veri Seçimi: Hangi veri setlerinin kullanılacağına karar verme ve ihtiyaç duyulan ek verilerin toplanması aşamasıdır. Çalışmada kullanılan kafa modelinin enine (transverse) düzlemine göre her bir farklı kesiti farklı bir kafa modeli olarak kurgulanmış ve farklı kesitlerden alınan veriler modellemede kullanılmak üzere bir veri setinde birleştirilmiştir. Kafa modelindeki sıra numarası 135, 136, 139, 140, 143, 144 olan enine düzlemler kullanılarak elde edilen veriler eğitim seti için bir araya getirilmiştir. Kafa modelindeki sıra numarası 137, 138, 141, 142, 145, 146 olan enine düzlemler ise test seti için bir araya getirilmiştir.

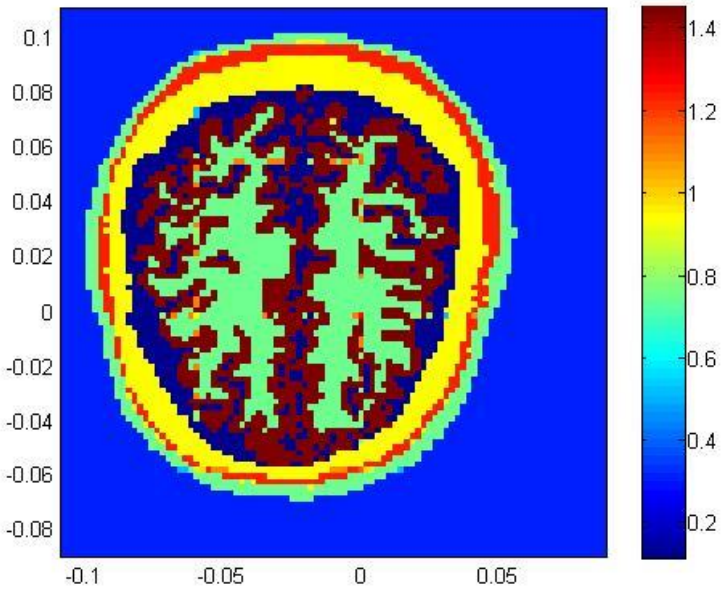
Veri Temizliği: Birçok uygulamada veriler tutarsız veya eksik olabilir. Bu nedenle veri kümesinde aykırılıkların bulunması ve tutarsızlıkların giderilmesi için veri temizleme işlemi yapılır. Ancak yapılan çalışmada veriler sentetik olarak üretildiği için bu aşamaya gerek kalmamıştır.

Veri Oluşturma: Beyin modeli 10 farklı konumda bulunan noktasal elektromagnetik kaynaklarla aydınlatılmış ve 40 farklı noktada bulunan alıcı mikrodalga antenlerle saçılan elektromagnetik alan değerleri ölçülmüştür. Kanamalı hücre içeren verilerde lezyon büyüklüğü 10 mm ile 15 mm arasında random olarak, lezyonun konumu da yine random olarak belirlenmiştir. MATLAB uygulamasından alınan ölçüm bölgesi, sağlıklı beyin ve lezyon içeren beyin şekilleri aşağıda verilmiştir. Şekil 4.1'de bir çember üzerinde gösterilen kırmızı noktalar verici antenlerin (noktasal kaynakların) konumunu, farklı bir çember üzerinde gösterilen mavi noktalar ise alıcı antenlerin konumunu

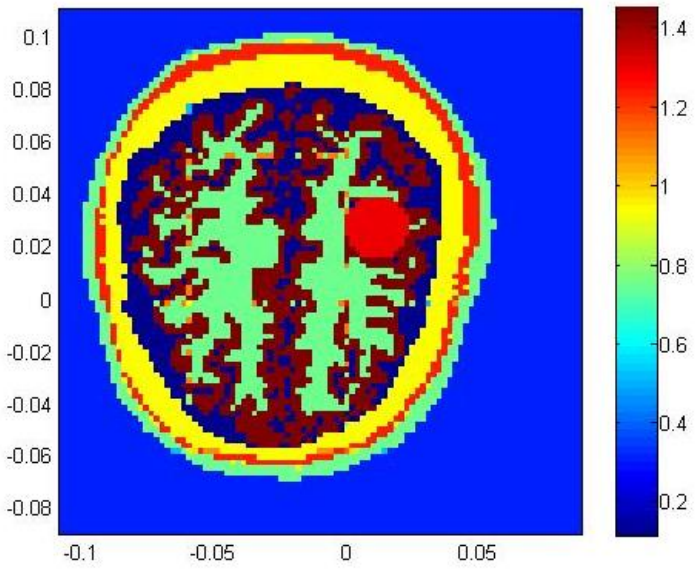
göstermektedir. Ayrıca merkezde bulunan mavi noktalar kümesi ise MoM kullanılırken yapılan hücrelendirmelerin merkezini temsil etmektedir. Şekil 4.2 ve 4.3 ise sırasıyla sağlıklı ve lezyon içeren örnek beyin modellerinin iletkenlik katsayısının dağılımı gösterilmektedir.



Şekil 4.1 Noktasal Kaynaklar ve Ölçüm Noktaları



Şekil 4.2 Sağlıklı Beyin



Şekil 4.3 Lezyonlu doku içeren beyin

Simülasyon Model: Bu çalışmanın temeli gerçekçi elektromagnetik malzeme özellikleri içeren sayısal bir kafa modeli kullanmaya dayanmaktadır [44]. Bu model, kabaca, bir insan kafatasının MR görüntüsündeki deri, yağ, kan, gri madde gibi dokuların tespit edilmesi ve bu dokuların elektromagnetik özelliklerinin (dielektrik ve iletkenlik katsayıları) kullanılmasıyla elde edilmiştir. Model 2 mm x 2 mm x 2 mm boyutlarında 181*169*191 kübik elemandan oluşur. Bu çalışmada iki boyutlu bir görüntüleme yaklaşımı göz önünde bulundurulmuştur. Kanın göreli dielektrik katsayısı 80, iletkenlik katsayısı ise 1.3 S/m olarak alınmıştır [46].

Veriyi Formatla: Ölçülen veriler kompleks değerlerden oluşmaktadır. Reel kısım ve imajiner kısım farklı anlamlara gelmektedir ve ayrı birer özellik olarak tanımlanmışlardır. Oluşturulan veri dosyaları modellemede kullanılmak üzere bir veri setinde birleştirilmiştir. Veri setine ait bir kısım aşağıda gösterilmiştir.

```
%  ========== AÇIKLAMALAR ==================
%  Beyin Modeli : BRAIN_MATRICE_duzlem_1_slice_135_
%  Kanama Durumu : 0
%  min_R_kanama : 0.0000   max_R_kanama: 0.0000
%  Gürültü Oranı : 0.010
%  Ölçüm Noktası Sayısı : 40
%  Ölçüm Yarıçapı : 0.100
%  Ölçüm Çemberinin merkezi:(-0.0250 , 0.0150)
%  Kaynak Sayısı : 10
%  Kaynak Yarıçapı : 0.110
%  Kaynak Çemberinin merkezi:(-0.0250 , 0.0150)
%  Frekans : 500000000
%  ======================================
%  Attribute Sayısı: 801
%  Örnek Sayısı: 40
%  ======================================
@relation breast

@attribute T1M1Re numeric
@attribute T1M2Re numeric
@attribute T1M3Re numeric
@attribute T1M4Re numeric
@attribute T1M5Re numeric
@attribute T1M6Re numeric
@attribute T1M7Re numeric
@attribute T1M8Re numeric
@attribute T1M9Re numeric
@attribute T1M10Re numeric
@attribute T1M11Re numeric
@attribute T1M12Re numeric
```

Şekil 4.4 Veri setine ait bir kısım

4.4 Modelleme

Modelin İnşa Edilmesi: Çalışmanın bu aşamasında en uygun modelleri belirleyebilmek için WEKA da mevcut bulunan ve nümerik girdiyi destekleyen algoritmalar sınıflandırma doğruluk oranları baz alınarak kıyaslanmıştır. Gelen dalganın frekansı 1 GHz seçilmiş ve kafatası modelinin 135. enine kesiti alınarak 240 sağlıklı 240 hasta toplam 480 adet veri ile veri seti elde edilmiştir. Bu veriler çeşitli veri madenciliğiyle eğitilmiştir. Ayrıca sırasıyla %1, %3, %5 ve %10 gürültü eklenerek bu işlem tekrarlanmıştır. Çizelge 4.2'de her bir sınıflandırıcı yöntem için başarı oranları verilmiştir.

Modelleme Tekniğinin Seçilmesi: Eğitim sonuçları kıyaslandığında en başarılı algoritmalar Naive Bayes, k-NN algoritması ve Hoeffding Tree olarak belirlenmiş ve modeller bu algoritmalar ile oluşturulmuştur**.**

Çizelge 4.1 Model başarım oranları

Sınıflandırma Fonk.Adı	Noise 0.00	Noise 0.01	Noise 0.03	Noise 0.05	Noise 0.10
Naive Bayes	100%	100%	100%	100%	100%
Naive Bayes Updateable	100%	100%	100%	100%	100%
Logistic	100%	95%	100%	80%	75%
Multilayer Perceptron	100%	95%	100%	100%	90%
SMO	100%	95%	100%	100%	90%
Ibk	100%	100%	100%	100%	95%
Kstar	35%	85%	70%	95%	80%
LWL	95%	100%	100%	85%	65%
Iterative Classifier Optimizer	100%	100%	95%	90%	65%
AdaBoostM1	100%	100%	95%	95%	75%
Attibute Selected Classifier	100%	100%	85%	80%	75%
Bagging	100%	100%	95%	60%	55%
Classification Via Regression	100%	100%	85%	80%	75%
Filtered Classifier	100%	100%	85%	80%	75%
Logit Boost	100%	100%	95%	90%	60%
Random Committee	100%	100%	100%	100%	80%
Multi Class Classifier Updateable	100%	100%	100%	100%	90%
Randomizable Filtered Classifier	100%	100%	100%	95%	85%
Random Subspace	95%	100%	50%	50%	50%
Input Mapped Classifier	50%	100%	50%	50%	50%
Decision Table	100%	75%	65%	75%	75%
Jrip	100%	95%	80%	80%	45%
PART	100%	100%	85%	80%	75%
Decision Stump	100%	100%	85%	85%	65%
Hoeffding Tree	100%	100%	100%	100%	95%
J48	100%	85%	85%	80%	75%
LMT	100%	95%	95%	100%	90%
Random Forest	100%	100%	100%	100%	55%
Random Tree	100%	80%	80%	65%	85%
Rep Tree	100%	85%	85%	85%	50%

Test Verisinin Oluşturulması: Modelleme tekniklerinin seçilmesinde kafatasının tek bir enine kesiti için yukarıda verilen sınıflandırıcılar seçilmişti. Şimdi ise; seçilen sınıflandırıcılar kafatası modelinin 135, 136, 139, 140, 143, 144 numaralı kesitlerinden üretilen 480 adet (240 sağlıklı, 240 lezyonlu) veriyle eğitilmiş ve 137, 138, 141, 142, 145, 146 numaralı kesitlerden üretilen 960 adet (480 sağlıklı, 480 lezyonlu) veriyle test

edilmiştir. Burada frekans yine 1 GHz seçilirken, gürültü oranı %1 ve %3 olarak alınmıştır.

Modellerin Değerlendirilmesi: Öncelikle %1 gürültü için Naive Bayes, kNN ve Hoeffding Tree modelleri için ayrı ayrı test sonuçları elde edilmiştir. Naive Bayes algoritmasının karışıklık matrisi ve sınıflandırma sonuçları sırasıyla Çizelge 4.3 ve Çizelge 4.4'te verilmiştir. Algoritma, true sınıfında (kanama var) olması gereken 5 örneği false (kanama yok) sınıfında; false sınıfında olması gereken 44 örneği true sınıfında tahmin ederek % 94.89 başarı oranı sağlamıştır.

Çizelge 4.2 Naive Bayes Modeli Karışıklık Matrisi Sonuçları

Karışıklık Matrisi		Tahmin Edilen Sınıf	
		Sınıf=1	Sınıf=0
Doğru Sınıf	Sınıf=1	475	5
	Sınıf=0	44	436

Çizelge 4.3 Naive Bayes Modeli Sınıflandırma Sonuçları

Toplam Örnek Sayısı	960	Özellik Sayısı	801
Doğru Sınıflandırılan Örnek Sayısı	911	Doğru Sınıflandırma Oranı (%)	%94.8958
Yanlış Sınıflandırılan Örnek Sayısı	49	Yanlış Sınıflandırma Oranı (%)	%5.1042

kNN algoritmasının karışıklık matrisi ve sınıflandırma sonuçları ise sırasıyla Çizelge 4.5 ve Çizelge 4.6'da verilmiştir. Bu algoritma, true sınıfında (kanama var) olması gereken 8 örneği false (kanama yok) sınıfında; false sınıfında olması gereken 480 örneği tamamını false sınıfında tahmin ederek % 99.16 başarı oranı sağlamıştır.

Çizelge 4.4 kNN modeli karışıklık matrisi sonuçları

Karışıklık Matrisi		Tahmin Edilen Sınıf	
		Sınıf=1	Sınıf=0
Doğru Sınıf	Sınıf=1	472	8
	Sınıf=0	0	480

Çizelge 4.5 kNN modeli sınıflandırma sonuçları

Toplam Örnek Sayısı	960	Özellik Sayısı	801
Doğru Sınıflandırılan Örnek Sayısı	952	Doğru Sınıflandırma Oranı (%)	%99.1667
Yanlış Sınıflandırılan Örnek Sayısı	8	Yanlış Sınıflandırma Oranı (%)	%0.8333

Benzer şekilde Hoeffding tree algoritmasının karışıklık matrisi ve sınıflandırma sonuçları elde edilmiştir(Bkz. Çizelge 4.7 ve Çizelge 4.8). Hoeffding tree algoritması, true sınıfında (kanama var) olması gereken 131 örneği false (kanama yok) sınıfında; false sınıfında olması gereken 274 true (kanama var) sınıfında tahmin ederek % 57.81 başarı oranı sağlamıştır.

Çizelge 4.6 Hoeffding tree modeli karışıklık matrisi sonuçları

Karışıklık Matrisi		Tahmin Edilen Sınıf	
		Sınıf=1	Sınıf=0
Doğru Sınıf	Sınıf=1	349	131
	Sınıf=0	274	206

Çizelge 4.7 Hoeffding tree modeli sınıflandırma sonuçları

Toplam Örnek Sayısı	960	Özellik Sayısı	801
Doğru Sınıflandırılan Örnek Sayısı	555	Doğru Sınıflandırma Oranı (%)	%57.8125
Yanlış Sınıflandırılan Örnek Sayısı	405	Yanlış Sınıflandırma Oranı (%)	%42.1875

Yukarıda %1 gürültü oranı için gerçekleştirilen işlemler, gürültü oranı %3'e çıkarılarak tekrarlanmıştır. Naive Bayes algoritmasının başarı oranının %94.89'dan % 50 ye, kNN algoritmasının başarı oranının %99.16'dan %82.5'e ve Hoeffding Tree algoritmasının başarı oranının ise %57.81'den %50'ye düştüğü gözlemlenmiştir. Bu durumda kNN algoritmasının diğerlerine göre gürültüden en az etkilenen algoritma olduğu söylenebilir.

%1 ve %3 gürültü oranları için her bir algoritmanın doğruluk hata oranı, ortalama mutlak hata (MAE), ortalama hata kareleri karekökü (RMSE) ve Kappa istatistiği Çizelge 4.9'da verilmiştir.

MAE; mutlak hata ortalamasını verir. Mutlak hata ölçülen değerle gerçek değer arasındaki farkı ifade eder. Mutlak hata sıfıra ne kadar yakınsa modelin tahmin yeteneği de o kadar iyidir. Buna göre; kNN algoritmasının her iki gürültü oranı için tahmin yeteneğinin daha iyi olduğu görülmektedir.

Çizelge 4.8 Test istatistikleri

	gürültü=0.01				gürültü =0.03			
	Doğruluk	MAE	RMSE	Kappa	Doğruluk	MAE	RMSE	Kappa
Naive Bayes	%94.89	0.05	0.22	0.89	%50	0.50	0.70	0
kNN	%99.16	0.01	0.09	0.98	% 82.5	0.18	0.41	0.65
Hoeffding Tree	%57.81	0.40	0.55	0.15	%50	0.50	0.70	0

RMSE; gerçek sınıflandırma değerleri ile modelin tahmini arasındaki hata oranını saptamak için kullanılır. RMSE değerinin 0 olması kullanılan modelin mükemmel olması demektir.

Kappa değeri ise 1'e yakınlığı tahmin edilen sınıf ile gerçek sınıfın uyumunu gösteren bir istatiksel ölçüdür.

Tüm sonuçlar göz önünde bulundurulduğunda; mutlak hata ortalaması en düşük, gürültüden en az etkilenen ve doğruluk oranı % 99,16 ile en yüksek algoritma kNN algoritması olmuştur.

4.5 Çalıştırma

Bu tezde beynin farklı enine kesitleri kullanılarak iki boyutlu saçılma problemi için beyin lezyonun tespiti ele alınmıştır. Elde edilen başarılı sonuçlar göstermektedir ki; bu yaklaşım üç boyutlu problemler için genişletilebilir ve dolayısıyla klinik ortamda gerçek ölçüm verileriyle test edilebilir.

BÖLÜM 5

SONUÇLAR ve ÖNERİLER

Beyindeki lezyonların mikrodalga tomografisi ile tespitinde genelde yüksek oranda matematiksel karmaşıklığa sahip ters saçılma problemlerini çözen sayısal yöntemler kullanılmaktadır. Lezyonların tespitinde veri madenciliği yöntemlerini kullanmak ise bu zorlukları azaltmaktadır. Ayrıca klinik ortamda ölçüm düzeneğinden kaynaklanan gürültülü veri ile sınıflandırma algoritmaları eğitildiğinden dolayı, bu yöntemler ölçüm düzeneğinden daha az etkilenir. Bir başka ifadeyle, ölçümün yapıldığı ortamdaki düzeneğin tüm etkileri eğitim aşamasında dikkate alınır ve bu durum pratik uygulamalarda avantaj sağlar.

Bu çalışmada, öncelikle gerçekçi özelliklere sahip bir beyin modelinin enine kesitleri kullanılarak moment yöntemiyle sentetik veriler üretilmiştir ve çeşitli makine öğrenmesi algoritmaları bu verilerle modellenmiştir. Bu algoritmalar temel olarak gürültüye dirençlerine göre kıyaslanmış ve başarı durumları değerlendirilmiştir. Buna göre, tahmin edici modeller arasından kullanılan eğitim seti ile beyin lezyonun tespitine dair en iyi sonuçları üreten Naive Bayes, kNN ve Hoeffding Tree algoritmaları seçilmiştir.

Naive Bayes algoritması değişkenlerin birbirinden bağımsız olması temeline dayanır. Uygulamanın başlangıcında bu yönde bir kabul yapılır ve her bir özelliğe eşit değer verilir. Bu durum, değişkenleri gerçekte birbirinden bağımsız olmayan, değişkenleri arasındaki ilişkinin bilinmediği veya her bir özelliğin sınıflandırma açısından eşit bir öneme sahip olmadığı veriler için dezavantaj oluşturmaktadır. Bu çalışmada gerçekleştirilen simülasyonlara göre, başlangıçta Naive Bayes algoritmasının başarı

oranı oldukça yüksek olsa da gürültü oranı %1'den %3'e çıkarıldığında başarı oranı ciddi oranda düşmüştür. Bu da, algoritmanın klinik ortamlarda gürültü miktarına karşı oldukça hassas olduğunu göstermektedir.

Hoeffding Tree algoritması, ağacın dağılımını meydana getiren örneklerin zamanla değişmez olduğunu varsayarak çalışmaktadır. Modelleme aşamasında beyin modelinin tek bir enine kesitinden elde edilen verilerle %95 başarı oranı sağlanmıştır. Ancak beyin modelinin farklı enine kesitlerinden elde edilen örneklerle çalıştırıldığında başarı oranı yaklaşık %58'e düşmüştür. Bu durum, belli bir beyin modeli için oluşan ağaç yapısının farklı beyin modelleri için yetersiz kaldığını göstermektedir.

Yeni bir veri geldiğinde eğitim verilerini baz alarak en yakın komşularına göre sınıflandırma yapan kNN algoritması, gürültü miktarının artmasından en az etkilenen ve dolayısıyla başarı oranlarının en az düştüğü algoritma olmuştur. Bu da algoritmanın gürültüye karşı diğerlerine nazaran daha dayanıklı olduğunu göstermektedir.

Bu çalışmanın devamında beyin lezyonlarının varlığıyla birlikte konumu ve büyüklüğü de tespit edilebilir. Ayrıca problem kolayca üç boyutlu olarak ele alınıp, söz konusu yöntemler klinik ortamda test edilebilir.

KAYNAKLAR

[1] Karatağ, O., (2005), İntrakranyal Yer Kaplayıcı Lezyonların Ayırıcı Tanısında MR Spektroskopinin Yeri, Radyoloji Uzmanlık Tezi, Şişli Etfal Eğitim ve Araştırma Hastanesi, Radyoloji Kliniği, İstanbul.

[2] Hemen Sağlık, Beyin Lezyonları, www.hemensaglik.com/makale/beyin-lezyonlari, 25 Mayıs 2016.

[3] Polat, B., (2015), "Beyin Kanamalarının Tespitinde Mikrodalga Görüntüleme Teknikleri Üzerine", Beykent Üniversitesi Fen ve Mühendislik Bilimleri Dergisi, 8(2): 191-210.

[4] Gunnarsson, T., (2007), Microwave Imaging of Biological Tissues: Applied Toward Breast Tumor Detection, Licentiate Thesis, Mälardalen University, Sweden.

[5] Arıöz, U., Beyin Görüntüleme Teknikleri ve Algoritmaları, Gazi Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği, ceng.gazi.edu.tr/~uarioz/beyin/BEYIN GORUNTULEME 1 HAFTA GIRIS.pdf.

[6] Ireland, D., (2011). "Microwave Head Imaging for Stroke Detection", Progress In Electromagnetics Research M, 21:163-175.

[7] Mohammed, B.J., Abbosh, A.M., Mustafa, S. ve Ireland, D., (2014). "Microwave System for Head Imaging", IEEE Transaction on Instrumentation and Measurement, 63(1):117-123.

[8] Semenov, S.Y. ve Corfield, D.R., (2008). "Microwave Tomography for Brain Imaging: Feasibility Assessment for Stroke Detection", International Journal of Antennas and Propagation, Article ID:254830, 1-8.

[9] Scapaticci, R., Donato, l.D., Catapano, I. ve Crocco, L., (2012). "A Feasibility Study on Microwave Imaging for Brain Stroke Monitoring", Progress In Electromagnetics Research B, 40:305-324.

[10] Ireland, D., Bialkowski, K. ve Abbosh, A., (2013). "Microwave Imaging for Brain Stroke Detection Using Born Iterative Method", IET Microwaves, Antennas & Propagation, 7(11):909-915.

[11] Munawar, A., Mustansar, Z. ve Maqsood, A., (2016). "Analysis of Microwave Scattering from a Realistic Human Head Model for Brain Stroke Detection

Using Electromagnetic Impedance Tomography", Progress In Electromagnetics Research M, 52:45-56.

[12] Bisio, I., Fedeli, A., Lavagetto, F., Luzzati, G., Pastorino, M., Randazzo, A., ve Tavanti, E., (2016). "Brain stroke detection by microwave imaging systems: Preliminary two-dimensional numerical simulations", 2016 IEEE International Conference on Imaging Systems and Techniques (IST), 4-6 Ekim 2016, Yunanistan, 330-334.

[13] Hadamard, J., (1923). Lectures on Cauchy's Problem in Linear Partial Differential Equations, Yale University Press, New Haven.

[14] Sundström, C., (2014). Machine Learning Algorithms for Stroke Diagnostics, Yüksek Lisans Tezi, Chalmers Unıversıty of Technology, Department of Signals and Systems, Göteborg.

[15] Wu, Y., Zhu, M., Li, D., Zhang, Y. ve Wang, Y., (2016). "Brain stroke localization by using microwave-based signal classification", 2016 International Conference on Electromagnetics in Advanced Applications (ICEAA), 19-23 Eylül 2016, Avustralya, 828-831.

[16] Wu, Y., Xu, X. ve Zhu, M., (2016). "Subspace Pattern Recognition Method for Brain Stroke Detection", Progress In Electromagnetics Research Letters, 46:119-126.

[17] Noghaninan, S., (2012), Microwave Tomography for Biomedical Quantitative Imaging, Electrical and Electronics, University of North Dakota, USA, 1:3.

[18] Tikhonov, A.N., Arsenin, V.Y., (1977), Solutions off ill-Posed Problems, John Wiley & Sons, New York.

[19] Hanke, M., (1995), Conjugate and Gradient Type Methods for Ill-Posed Problems, Chapman and Hall/ CRC.

[20] Mojabi, P., LoVetri, J., (2009), Overview and Classification of Some Regularization Techniques for the Gauss-Newton Inversion Method Applied to Inverse Scattering Problems. IEEE Trans Antennas Propag 57: 2658-2665.

[21] Joachimowicz, N., Pichot, C., Hugonin, J.P., (1991), Inverse Scattering: an Iterative Numerical Method for Electromagnetic Imaging, IEEE Trans Antennas Propag 39: 1742- 1753.

[22] Mitchell, T., (1999), Machine Learning and Data Mining, Center for Automated Learning and Discovery Scholl of Computer Science, Carnegie Mellon University.

[23] Akgöbek, Ö., Çakır, F., (2009), "Veri Madenciliğinde Bir Uzman Sistem Tasarımı", Akademik Bilişim'09-XI.Akademik Bilişim Konferansı Bildirileri, Harran Üniversitesi, Şanlıurfa.

[24] VeriMadenciliği,BaşkentÜniversitesi,http://mail.baskent.edu.tr/~20410964/DM-1.pdf. 01 Eylül 2016.

[25] Fayyad, U., Piatetsky, G., Shapiro, G., and Smyth, P., (1997), From Data Mining to Knowledge Discovery in Databases, Al Magazine Volume 17, No: 3.

[26] Kumdereli, Ü., (2012), Tıp Bilişimi veVeri Madenciliği Uygulamaları, EEG Sinyallerindeki Epileptiform Aktiviteye Veri Madenciliği Yöntemlerinin Uygulanması, Yüksek Lisans Tezi, Trakya Üniversitesi, Edirne.

[27] Dalkılıç, G., Türkmen, F., (2002), "Karınca Kolonisi Optimizasyonu", Yüksek Performanslı Bilişim Sempozyumu, Kocaeli.

[28] Chen, M. and S., Han, J., Yu, P., Data Mining an Overview from Database Perspective, U.S.A.

[29] Jiawei, H., Kamber, M., (2006), "Data Mining Concepst and Techniques", Second Edition, Morgan Kaufmann Publishers, San Francisco.

[30] Kaya, M., Özel, A.S., (2014), Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması, Akademik Bilişim Konfereansları.

[31] Kızılkan, Ö., Küçüksılle, E.U., Kabul, A., (2015), Data Mining Approach for Analysis of Variable Speed Refrigeration System, Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü Dergisi, 19(1):19-26

[32] Koyuncugil, A.S., (2006), Bulanık Veri Madenciliği ve Sermaye Piyasalarına Uygulanması, Doktora Tezi, Ankara Üniversitesi, İstatistik Anabilim Dalı, Ankara.

[33] Tokmak, D., (2011), Knowledge Discovery and a Comparison of Data Mining Tools on Iris Dataset, Thesis of Master, Departmant of Computer Engineering, Başkent University, Ankara.

[34] Chapman, P., "The Crısp-DM User Guide", Brussels SIG Meeting, NCR Systems Engineering Copenhagen, http://lyle.smu.edu/~mhd/8331f03/crisp.pdf, 10 Mayıs 2016.

[35] Farboudi, S., (2009), Tıp Bilişiminde İstatistiksel Veri Madenciliği, Yüksek Lisans Tezi, Hacettepe Üniversitesi, İstatistik Anabilim Dalı, Ankara.

[36] Öztemel, E., (2012), Yapay Sinir Ağları, 3.Basım, Papatya Yayıncılık, İstanbul

[37] Uğur, A., Kınacı, A.C., (2006), Yapay Zeka Teknikleri ve Yapay Sinir Ağları Kullanılarak Web Sayfalarının Sınıflandırılması, Türkiye'de İnternet Konferansı, 21-23 Aralık 2006, Ankara.

[38] Tektaş, A., Karataş, A., (2004), "Yapay Sinir Ağları ve Finans Alanına Uygulanması: Hisse Senedi Fiyat Tahminlemesi", İktisadi İdari Bilimler Dergisi, 18:3-4.

[39] Emel, G., Taşkın, Ç., (2002), "Genetik Algoritmalar ve Uygulama Alanları", İktisadi İdari Bilimler Fakültesi Dergisi, Uludağ Üniversitesi, Cilt XXI, 1:129-152.

[40] Bellazzi, R., Zupan, B., (2008) Predictive Data Mining in Clinical Medicine: Current Issues and Guidlines, International Journal of Medical Informatics 77, 81-97.

[41] Doğan, Y., (2015), Data Mining and Knowledge Discovery in Medical Information Systems, Dokuz Eylül University Graduate School of Natural and Applied Sciences, İzmir.

[42] Nizam, H., Akın, S.S., (2014), "Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması", Türkiye'de İnternet Konferansı, İzmir.

[43] Viera, A.J., Garrett, J.M., (2005), "Understanding Interobserver Agreement: the Kappa Statisc", Family Medicine, 37(5): 360-3.

[44] Yelkenci, T. ve Magerl, G., (2000), 'Die Berechnung der von Mobiltelefonen im menschlichen Kopf hervorgerufenen spezifischen Absorptionsraten' , Elektrotech. Inftech., 117 (11): 744–749.

[45] Richmond, H.J. , (1965), "Scattering by Dielectric Cylinder of Arbitrary Cross Section Shape", IEEE Transactions on Antennas and Propagation, 1792-1799.

[46] John, G.H., Langley, P., (1995), Estimating Continuous Distributions in Bayesian Classifiers, In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338-345.

[47] Landwehr, N., Hall, M., and Frank, E., (2005), Logistic Model Trees.

[48] Hastie T., Tibshirani, R., (1998), Classification by Pairwise Coupling, Advances in Neural Information Processing Systems.

[49] Aha, D., Kibler, D., (1991), Instance-based Learning Algorithms, Machine Learning, 6:37-66.

[50] Cleary, J.G., Trigg, L.E., (1995), K:An Instance-based Learner Using an Entropic Distance Measure, 12th International Conference on Machine Learning, 108-114.

[51] Atkeson, C., Moore, A., and Schaal, S., (1996), Locally Weighted Learning, AI Review.

[52] Freund, Y., Schapire, R.E., (1996), Experiments with a New Boosting Algorithm, Thirteenth International Conference on Machine Learning, San Fransisco, 148-156.

[53] Breiman, L., (1996), Bagging Predictors, Machine Learning, 24 (2):123-140.

[54] Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I.H., (1998), Using Model Trees for Classification , Machine Learning, 32 (1):63-76.

[55] Kohavi, R., (1995), Wrappers for Performance Enhancement and Oblivious Decision Graphs, Departmant of Computer Science, Stanford University.

[56] Kuncheva, L.I., (2004), Combining Pattern Classifiers: Methods and Algorithms, John Wiley and Sons, Inc..

[57] Holte, R.C., (1993), Very Simple Classification Rules Perform Well on Most Commonly Used Datasets, Machine Learning, 11:63-91.

[58] Breiman, L., (2001), Random Forests, Machine Learning, 45 (1):5-32

[59] Quinlan R., (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.

EK-A

WEKA PLATFORMU SINIFLANDIRICILARI

A-1 Bayes Sınıfı

BayesNet: Bayes Network; çeşitli arama algoritmalarını ve bir Bayes Network sınıflandırıcısı için taban ölçümü kalite ölçütlerini kullanarak öğrenir. Kullanıcıya, veri yapıları özelliklerini (ağ yapısı, koşullu olasılık dağılımları, vb.) ve K2 ve B gibi öğrenme algoritmalarında da mevcut olan imkanları sağlar.

Naive Bayes: Tahmin edici sınıfları kullanan Naive Bayes sınıflandırıcısı için tahmin edici hassas değerler, eğitim verilerinin analizine dayanarak seçilir. Bu nedenle sınıflandırıcı, güncellenebilir bir sınıflandırıcı değildir. Güncellenebilir sınıflandırıcı işlevselliğine ihtiyaç duyulduğunda NaiveBayesUpdateable sınıflandırıcı kullanılabilir. NaiveBayesUpdateable sınıflandırıcısı sayısal öznitelikler için varsayılan olarak hassasiyet değerini 0,1 olarak kullanır.

Naive Bayes Multinominal: Bu sınıflandırıcı, çok terimli bir NaiveBayes sınıflandırıcısı oluşturmak ve kullanmak için oluşturulmuştur. Metin sınıflandırmak için çok terimli Naive Bayes modeli öne sürülmüştür ve çok değişkenli Bernoulli modeline göre frekans bilgisi desteği sağladığından ötürü üstün bir performans göstermektedir.

Naive Bayes Multinominal Text: Multinominal Naive Bayes Text sınıflandırıcısı doğrudan ve sadece string değerleri üzerinde çalışır. Diğer veri tipleri de girdi olarak kabul edilir ancak eğitim ve sınıflandırma sırasında yok sayılır [47].

A-2 Functions Sınıfı

Logistic: Sınıflandırıcıda eksik değerler, eksik değer değiştirici filtresi (ReplaceMissingValuesFilter) kullanılarak değiştirilir ve nominal nitelikler NominaltoBinaryFilter kullanılarak sayısal özniteliklere dönüştürülür [48].

Multilayer Perceptron: Örnekleri sınıflandırmak için geri yayılım (backpropagation) kullanan bir sınıflandırıcıdır. Manuel olarak ya da algoritma kullanılarak veya her ikisi tarafından da oluşturulabilir. Ağ, eğitim zamanı boyunca izlenebilir ve değiştirilebilir.

SGD: Stochastinc Gradient Descent (SGD), küresel olarak tüm eksik değerleri değiştirir ve nominal nitelikleri ikili değerlere dönüştürür. Çıktıdaki katsayıların normalleştirilmiş değerlere dayandırılması için tüm nitelikleri normalleştirir.

SMO: Sequential Minimal Optimization (SMO), bir destek vektör sınıflandırıcısını eğitmek için sıralı minimum optimizasyon algoritması kullanır. Bu uygulama tüm eksik değerleri yeniden konumlandırır ve nominal nitelikleri ikili değerlere çevirir. Varsayılan olarak tüm öznitelikleri normalleştirir. Bu durumda çıktıdaki katsayılar orijinal verilere değil normalleştirilmiş verilere dayanır [49].

VotedPerceptron: Bu uygulama tüm eksik değerleri yeniden konumlandırır ve nominal nitelikleri ikili değerlere çevirir.

A-3 Lazy Sınıfı

kNN: k en yakın komşu sınıflandırıcısıdır. Çapraz geçerlilik temelinde k'nın uygun değerini seçebilir. Ayrıca mesafeye göre ağırlık belirlenebilir [50].

kstar: Örnek tabanlı bir sınıflandırıcıdır. Başka bir deyişle bir test örneğinin sınıfı aynı benzerlik fonksiyonlarıyla belirlenen bir eğitim örneğinin sınıfına göre belirlenebilir. Entropi tabanlı bir uzaklık fonksiyonu kullanması açısından diğer örnek tabanlı sınıflandırıcılardan farklıdır [51].

LWL: Locally Weighted Learning (LWL), yerel olarak ağırlıklandırılmış bir makine öğrenme algoritmasidir. Ağırlıklı örnekleri atamak için örnek tabanlı bir algoritma kullanır. Sınıflandırma ve doğrusal regresyon yapılabilir [52].

A-4 Meta Sınıfı

adaboostm1: adaboost m1 yöntemini kullanarak nominal sınıf sınıflandırıcısını yükseltmek için kullanılan bir sınıftır. Sadece nominal sınıf problemlerini çözebilir.

Çoğunlukla beklenenin üzerinde performansı artırır ama bazen sisteme aşırı yüklenir [53].

Attribute selected classifier: Eğitim ve test verilerinin boyutsallığı sınıflandırıcıya geçmeden önce özellik seçimi ile azaltılır.

Bagging: Varyansı azaltmak için eğitim verisinden birden fazla örnek oluşturarak elde edilen daha zayıf eğitim verilerinin temel öğreniciler tarafından öğrenilmesi sonucu ortaya çıkan modellerin karşılaştırılması mantığına dayanır. Bu metod çapraz sorgulama metoduna benzemektedir [54].

Classificationviaregression: Regresyon yöntemlerini kullanarak sınıflandırma yapmakta kullanılır. Sınıf ikili hale getirilir ve her sınıf değeri için bir regresyon modeli oluşturulur [55].

Cost sensitive classifier: Temelinde maliyet duyarlı olan bir meta sınıflandırıcıdır. Maliyet duyarlılığını ortaya koymak için iki yöntem kullanılabilir. Her sınıfa göre belirlenen toplam maliyete göre eğitim örneklerini yeniden ağırlıklandırma veya sınıfı, asgari olarak beklenen en yanlış sınıflandırma maliyetiyle öngörmek. Performans, temel sınıflandırıcının olasılık tahminlerini iyileştirmek için bagging sınıflandırıcısını kullanılarak artırılabilir.

Cv parameter selection: Herhangi bir sınıflandırıcı için çapraz doğrulama ile parametre seçimi gerçekleştiren sınıftır [56].

Filtered classifier: Rastgele bir filtreden geçirilen veriler üzerinde keyfi bir sınıflandırıcı çalıştırmak için kullanılan sınıftır. Sınıflandırıcı gibi filtrenin de yapısı yalnızca eğitim verilerine dayanır ve test örneklerinin yapıları değiştirlmeden filtre tarafından işlenir.

Iterative classifier optimizer: Bu sınıflandırıcı, çapraz doğrulama kullanan logitboost gibi bir iterative sınıflandırıcı için en iyi yineleme sayısını seçer. Çapraz doğrulama kullanarak verilen yinelemeli sınıflandırıcının yinelemelerinin sayısını optimize eder.

Logit boost: Bu sınıf, temel öğrenici olarak bir regresyon şeması kullanarak sınıflandırma yapar ve çok sınıflı problemleri işleyebilir.

Multi class classifier: 2 sınıflı sınıflandırıcılarla çok sınıflı veri kümelerini işlemek için kullanılır. Bu sınıflandırıcı aynı zamanda artan doğruluğa yönelik çıkış kodlarını düzeltme uygulamasına izin verebilir.

Multi scheme: Sınıflandırıcıları, eğitim verileri üzerinde çapraz doğrulama veya performans kullananlar arasından seçmeye yarayan bir sınıftır. Performans sınflandırma oranı yüzdesi veya ortalama karesel hata temel alınarak ölçülür.

Random committee: Rastgele hale getirilebilir taban sınıflandırıcıları topluluğu oluşturmak için kullanılan bir sınıftır. Her taban sınıflandırıcı farklı bir rastgele sayı köküyle oluşturulmuştur. Son tahmin ise kişisel taban sınıflandırıcıları tarafından üretilen tahminlerin düz ortalamasıdır.

Random subspace: Bu yöntem, eğitim verisi üzerinde en yüksek doğruluk oranını korur ve karmaşıklık arttıkça genelleme doğruluğunu iyileştiren karar ağacı tabanlı bir sınıflandırıcı oluşturur. Sınıflandırıcı, rasgele seçilen alt uzaylarda oluşturulan ağaçların özellik vektörünün bileşenlerinin alt gruplarını sözde rastgele seçerek sistematik olarak yapılandırılmış birden fazla ağaçtan oluşur.

Stacking: Yığın yapısını kullanarak birçok sınıflandırıcıyı birleştirir. Sınıflandırabilir veya kümeleme yapabilir.

Vote: sınıflandırıcıları birleştiren bir sınıftır. Sınıflandırma için olasılık tahminlerinin farklı kombinasyonları mevcuttur [57].

A-5 Misc Sınıfı

Input mapped classifier: uyumsuz eğitim ve test verilerini, sınıflandırıcıların oluşturulduğu eğitim verileri ile gelen test örneklerinin yapısı arasında bir eşleme oluşturarak adresleyen wrapper sınıflandırıcısıdır. Gelen örneklerde bulunmayan model öznitelikler eksik değer olarak alınır böylece sınıflandırıcının daha önce görmediği gelen nominal öznitelik değerler kullanılır. Yeni bir sınıflandırıcı eğitilebilir veya varolan bir sınıflayıcı dosyadan yüklenebilir.

Serialized classifier: Bu sınıflandırıcı seri hale getirilmiş modeller yükler ve tahminler yapmak için onu kullanılır.

A-6 Rules Sınıfı

Oner: 1-r sınıflandırıcıyı oluşturmak ve kullanmak için kullanılan sınıftır. Bir başka deyişle sayısal öznitelikleri ayıran tahminler için minimum hata veren özniteliği kullanır [58].

Part: Part, karar listesi oluşturmak için kullanılan bir sınıftır. Her yinelemede kısmi bir c4.5 karar ağacı oluşturur ve "en iyi" yaprağı bir kural haline getirir.

Zero r: 0-r sınıflandırıcısı oluşturmak ve kullanmak için oluşturulmuştur. Ortalama değeri (sayısal bir değer için) veya modu (nominal bir sınıf için) tahmin eder.

A-7 Trees Sınıfı

Decision stump: Bir karar kökü oluşturmak ve kullanmak için oluşturulmuştur. Genellikle yükseltme algoritması ile birlikte kullanılır. Eksik veriler ayrı bir değer olarak değerlendirilir.

Lmt: Yapraklarında lojistik kümeleme fonksiyonlarına göre sınıflandırma yapan lojistik model ağaçları inşa etmek için oluşturulmuş bir sınıflandırıcıdır. Algoritma sayısal ve nümerik özniteklikler ile eksik değerler olmak üzere ikili ve çok sınıflı hedef değişkenleri ele alabilir.

Random forest: Rastgele ağaçlardan bir yapı oluşturmak için üretilmiştir [59].

Random tree: Her düğümde k adet rastgele seçilmiş niteliği göz önüne alan bir ağaç oluşturmak için üretilmiştir. Budama yapmaya izin vermez. Ayrıca sınıflandırma olasılıklarının (veya regresyon durumundaki hedef ortalamanın) tahmin edilmesine, destekleme kümesine (yedekleme) dayalı olarak izin verme seçeneğine sahiptir.

Rep tree: Hızlı karar ağacı öğrenicisidir. Bilgi kazanımı (varyansı) kullanarak bir karar/regresyon ağacı kurar ve azaltılmış hata ayıklama kullanarak onu siler. Yalnızca sayısal öznitelikler değerlerini bir kez alır. Eksik değerler, ilgili örnekleri parçalara ayırarak ele alınır [60].

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Tuğba VURAL

Doğum Tarihi ve Yeri : 06.08.1990/ÇİVRİL

Yabancı Dili : İngilizce

E-posta : tugbav34@gmail.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Y. Lisans	Matematik Mühendisliği	Yıldız Teknik Üniversitesi	2017
Lisans	Matemetik Mühendisliği	Yıldız Teknik Üniversitesi	2013
Lise	Sayısal	Emine Özcan AnadoluLisesi	2013

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2016-	Austria Card Turkey Kart Operasyonları A.Ş.	Bilgi Teknolojileri Yetkili Yardımcısı