ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİNİN TIP VE SAĞLIK HİZMETLERİNDE UYGULAMALARI

Didem ATİKTÜRK TAŞDELEN

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

ANKARA
2019

TEZ ONAYI

Didem ATİKTÜRK TAŞDELEN tarafından hazırlanan **"Veri Madenciliğinin Tıp ve Sağlık Hizmetlerinde Uygulamaları"** adlı tez çalışması 02/07/2019 tarihinde aşağıdaki jüri tarafından oy birliği ile Ankara Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman : Prof. Dr. Şahin EMRAH
Ankara Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı

Jüri Üyeleri :

Başkan: Doç. Dr. Süleyman TOSUN
Hacettepe Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı

Üye: Doç. Dr. Mehmet Serdar GÜZEL
Ankara Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı

Üye: Prof. Dr. Şahin EMRAH
Ankara Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı

Yukarıdaki sonucu onaylarım.

Prof. Dr. Özlem YILDIRIM
Enstitü Müdür Vekili

ETİK

Ankara Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez içindeki bütün bilgilerin doğru ve tam olduğunu, bilgilerin üretilmesi aşamasında bilimsel etiğe uygun davrandığımı, yararlandığım bütün kaynakları atıf yaparak belirttiğimi beyan ederim.

02/07/2019

Didem ATİKTÜRK TAŞDELEN

ÖZET

Yüksek Lisans Tezi

VERI MADENCILIĞININ TIP VE SAĞLIK HIZMETLERINDE UYGULAMALARI

Didem ATİKTÜRK TAŞDELEN

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Şahin EMRAH

Teknolojinin artmasıyla veri miktarının büyüklüğü ve işlenme sıklığı da günbegün artmaya devam etmektedir. Artan verilerden doğru veriye ulaşmak ve doğru bir şekilde analiz etmek yeni bir teknoloji gerektirmektedir. Bu büyük miktarlardaki veriler içerisindeki cevher veri, yönetilebilir olduğu ve yorumlandığı sürece değerlidir. Bu noktada veri madenciliği konseptiyle karşılaşılmaktadır.

Konu sağlık olduğunda ise doğru ve erken teşhis kritik öneme sahip olduğundan karar verme de çok önemli hale gelmektedir. Hasta olan kişiye erkenden tedaviye başlanabilmesi için hasta teşhisi konulması, hasta olmayan kişiye de gereksiz olduğu halde ilaç tedavisi uygulanmaması için doğru teşhisin en erken zamanda konulması toplum sağlığı açısından gereklidir. Burada makine öğrenmesi yoluyla karar verme konusunda makineler doktorlara yardımcı olmakta, böylece doğru tahminlerde bulunarak doktorların iş yükünü hafifletmektedir.

Bu çalışmada makine öğrenmesi metotları kullanılarak sınıflandırma işlemi yapılmıştır. Sağlık alanındaki verilerde perceptron öğrenme algoritması, K en yakın komşuluk, derin öğrenme metotları uygulanarak karşılaştırmalar yapılmış ve bir metot önerilmiştir. Uygulamalarda kullanılan veri kümesi ise UCI Makine Öğrenme Deposunda bulunan göğüs kanseri, pima yerlileri diyabet veri tabanı, bupa karaciğer hastalıkları, mamografik kitle verisi verileridir.

Temmuz 2019, 55 sayfa

Anahtar Kelimeler: Karar verme, derin öğrenme, perceptron öğrenme algoritması, akıllı sistemler, makine öğrenmesi, sağlık verisi, kanser verisi, diyabet, karaciğer hastalığı, veri madenciliği, yapay zeka

ABSTRACT

Master Thesis

APPLICATION OF DATA MINING IN MEDICAL SCIENCE AND HEALTH CARE

Didem ATİKTÜRK TAŞDELEN

Ankara University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Prof. Dr. Şahin EMRAH

The magnitude of the data and the frequency of processing data is continuing increase day by day with the advance of the technology. Analysing and extracting meaningful information from the big data requires a new approach. In this large amount of data, the ore data is valuable as long as it is manageable and interpreted. At this point it is encountered with the concept of data mining.

When the subject is health, decision making is critical since accurate and early diagnosis is critical. It is necessary in terms of public health that the patient should be diagnosed as soon as possible in order to start correct treatment early, and should prevent drug treatment if the patient is not unnecessary. Here, machines help doctors to decide on with machine decision making, reduce the workload of doctors with making accurate predictions.

In this study, the classification process is done by using machine learning methods. In the data in the health field, comparisons are made by applying the perceptron learning algorithm, K nearest neighbourhood, deep learning methods and a method is proposed. The data is used in applications are breast cancer, puma indians diabetes, bupa liver diseases, mammographic mass data in UCI Machine Learning Repository.

July 2019, 55 pages

Key Words: Decision making, deep learning, perceptron learning algorithm, intelligent systems, machine learning, health data, cancer data, diabetes, liver disease, data mining, artificial intelligence

TEŞEKKÜR

Bu çalışmayı yürütürken her koşulda bana fazlasıyla destek veren, bana farklı yaklaşım tarzları kazandırıp, benimle yeni bir algoritma geliştirme fikrini paylaşarak yeni bir algoritma geliştirmemize fırsat yaratan saygıdeğer hocam sayın Prof. Dr. Şahin EMRAH'a (Ankara Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı); bu süreçte bana destek olan bütün aile üyelerime ve iş arkadaşlarıma, en yapıcı eleştirileriyle kendimi geliştirme ateşini yakan hayat arkadaşım sevgili eşime ve son olarak bana çalışma azmi veren biricik kızım Bilge'me sevgi ve saygılarımı sunar teşekkürü bir borç bilirim.

Didem ATİKTÜRK TAŞDELEN
Ankara, Temmuz 2019

İÇİNDEKİLER

KISALTMALAR DİZİNİ

AUC	Area Under Curve
AI	Artificial Intelligence
ANN	Artificial Neural Network
BPN	Back Propagation Neural Network
CART	Classification and Regression Trees
CBC	Contraceptive Method Choice
CBFDT	Case-Based Fuzzy Decision Tree
CBR	Case-Based Reasoning
ECG	Electrocardiogram
EM	Expectation Maximization
FDT	Fuzzy Decision Tree
FM	Fuzzy Models
FR	Feature Ranking
FS	Feature Selection
FMM	Fuzzy Min – Max Sinir Ağı
GA	Genetic Algorithm
KNN	K-Nearest Neighbours
LS-SVM	Least Square Support Vector Machine
NB	Naive Bayes
NN	Neural Network
PCA	Principal Component Analysis
PCA-KNN	Principal Component Analysis - K-Nearest Neighbours
PCA-SVM	Principal Component Analysis - Support Vector Machine
PNN	Probabilistic Neural Network

RELU	Rectified Linear Unit
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operator Characteristic
RS	Rough Set
RS-SVM	Rough Set Support Vector Machine
SRA	Stepwise Regression Analysis
SVM	Support Vector Machine
UCI	UC Irvine Machine Learning Repository
WBDC	Wisconsin Breast Cancer Dataset
WDBC	Wisconsin Diagnostic Breast Cancer

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

1.1 Büyük Veri ve Veri Madenciliğinin Kullanıldığı Alanlar

Bilgisayarların, dijital ekranların, telefonların ve akıllı aletlerin kullanımının artmasıyla beraber üretilen veri miktarının büyüklüğü de her geçen gün artmaya devam etmektedir. Büyük veri tabanları, veri ambarları içinde saklanmış, ihtiyaç duyulan doğru veriye ulaşmak ve doğru kararları tahmin etmek yeni bir teknoloji gerektirmektedir. Veri içerisinden gizlenmiş verinin çıkarılması kabiliyeti her geçen gün daha da önemli hale gelmektedir.

Verileri yönetmek ve etkin şekilde analiz etmek karmaşık bir işlem olup yüksek performans gerektirir. Bu büyük miktarlardaki veriler içerisindeki cevher veri; yönetilebilir olduğu ve yorumlandığı sürece değerlidir. Bu noktada veri madenciliği konseptiyle karşılaşılmaktadır.

Konu sağlık olduğunda ise doğru ve erken teşhis kritik öneme sahip olduğundan karar verme de çok önemli hale gelmektedir. Bu noktada hasta olan kişiye erkenden tedaviye başlanabilmesi için hasta teşhisi konulması, hasta olmayan kişiye de gereksiz olduğu halde ilaç tedavisi uygulanmaması için doğru teşhisin en erken zamanda konulması toplum sağlığı açısından gereklidir. Bu noktada makina öğrenmesi yoluyla karar verme konusunda makineler doktorlara yardımcı olmakta, böylece doğru tahminlerde bulunarak doktorların iş yükünü hafifletmektedir ve onlara rehberlik ederek işlerini kolaylaştırarak hızlandırmaktadır.

Veri madenciliği ne aradığımızı ve nasıl bulacağımızı bilmediğimizde büyük veri içinden doğru veriyi bulma sanatıdır. Bu süreçte çoğu zaman hedefimiz bellidir, ancak hedefimize giden yolda sonuçları etkileyen faktörleri farkında olamayız ve bu noktada neyi aradığımızı ve nasıl bulacağımızı bilemezken bize en büyük yardımcı veri madenciliği yöntemleridir. Bu yöntemler arasında sınıflandırma, kümeleme ve birliktelik kuralları yer alır. Veri madenciliği hayatın her alanına bulunur ve etkin kullanıldığında yaşam kalitesini yükseltir.

1.1.1 Günlük hayattan örnekler

100 milyona yakın abonesiyle Netflix büyük veri teknolojisini etkin kullanır ve kullanıcılarının zevkine uyarlanmış bir hizmet sunar. Netflix algoritması, kullanıcıların davranış desenleriyle ilgili (ne izledikleri, ne atladıkları, ne aradıkları, nelerden hoşlandıkları gibi konularda) sürekli veri toplar. Bu verileri kişiselleştirilmiş film önerileri oluştururken kullanır ve hangi sınıflamada hangi filmlerin ilgili kullanıcıya hitap edeceğini belirler.

Bir uçağın uçması ile anlık olarak pilot ekranları, motor sistemleri, yakıt kullanımı, hava durumu bilgisi, olay raporları, kontrol pozisyonları, cihaz pozisyonları, uyarı modları gibi raporlamalar için milyonlarca veri üretilir. Ayrıca müşterilerle ilgili olarak üretilen veriler bulunmaktadır. Kişiye özel teklifleriyle müşteri memnuniyetini ve bağlılığını artırarak havayolu şirketleri rekabet avantajı sağlamaktadırlar.

Southwest Havayolları hem güvenli uçuşlar için hem de müşteri sadakati için veri madenciliği teknolojisi kullanmaktadır. Uçak verilerini analiz etmek üzere NASA ile ortaklık kurmuştur. NASA büyük veri ve veri madenciliği alanında sistem güvenliği ve müşteri memnuniyeti için çeşitli çalışmalar yapmıştır.

Akıllı çevre sistemleri, akıllı evler, akıllı ev aletleri de hayatımızda yaygınlaşmaya başlamıştır. Amazon'un Alexa, Apple Siri, Google Now, Microsoft – Cortana gibi kişisel asistanlar da yapay zeka teknolojisini kullanmaktadır

1.1.2 Sağlık alanındaki uygulamalar

Günlük hayatımızda sağlık alanıyla ilgili birçok akıllı uygulama mevcuttur. Giyilebilir teknolojiler koruyucu sağlık hizmetleri sunarak kullanıcıların yaşam kalitesini artırmaktadır. Google akıllı kontakt lensler anlık olarak gözyaşından glikoz ölçümü yapmakta ve gözlük kullananlar için gözün doğal odaklamasını ayarlar. Quell Relief akıllı dizlik vücuttaki ağrı sinyallerinin izleyerek kronik ağrıların azalmasına yardımcı olur. QardioCore giyilebilir ECG (Electrocardiogram - kalp grafiği) monitörü kalp

aktivitelerini izler. Sigarayı bırakmak amacıyla geliştirilen Smart Stop nikotin ölçen sensörler aracılığıyla gerektiğinde kullanıcıya ilaç verir. Akıllı saatler, akıllı bileklikler, akıllı kulaklıklar kullanımını yaygınlaşmıştır.

Ayrıca sağlık alanıyla ilgili çeşitli çalışmalar mevcuttur. Hastalık belirti (prognostic), teşhis (diagnostic) ve tedavisinde makinelerin tahmin etmesi için sınıflandırma, kümeleme, ilişkilendirmeler gibi bir takım veri madenciliği, makine öğrenmesi teknikleri kullanılır. Buna örnek verilecek olursa, literatürde kanser tahmini, mamografik görüntüler, skolyoz spinal hastalığı, kan üre konsantrasyonunun tahmini, hipertansiyon, diyabet, kardiyovasküler hastalıklar, akciğer grafileri, koroner arter hastalığı ile ilgili çalışmalar gösterilebilir.

1.2 Sağlık Uygulamalarında Önerilen Sınıflandırma Yöntemleri

Yapılan literatür taraması sırasında sınıflandırma probleminin çözümünde makineyi eğitmek için yapay sinir ağları, derin öğrenme modelleri, deri beslemeli sinir ağları, karar ağaçları, Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees - CART), bulanık modeller (Fuzzy Model - FM), nöro-bulanık teknikler, bulanık karar ağacı (Fuzzy Decision Tree - FDT), genetik algoritmalar (Genetic Algorithm - GA), destek vektör makineleri (Support Vector Machine - SVM), uzman sistemler, Rastgele Orman (Random Forest - RF), Saf Bayes (Naive Bayes - NB), k-ortalama (k-means), k-en yakın komşuluk (K-Nearest Neighbours - KNN) gibi birçok yöntem kullanıldığı görülmüştür. Ayrıca optimizasyonu sağlamak amacıyla özellik seçiminde durum tabanlı sebeplendirme, beklenti maksimizasyonu, temel bileşenler analizi (Principal Component Analysis- PCA), kaba kümeler (Rough Set - RS), kademeli regresyon analizi (Stepwise Regression Analysis - SRA), özyineli özellik seçimi (Recursive Feature Elimination - RFE) gibi yöntemlerin yaygın olarak kullanıldığı tespit edilmiştir.

2. KURUMSAL TEMELLER VE KAYNAK ÖZETLERİ

Programcılığın başlangıcı Ada Byron'ın bir makina tarafından işlenebilen ilk algoritmayı yaratmasıyla olmuştur. Yapay zekanın başlangıcı ise, Alan Turing'in 2. Dünya Savaşı sırasında "Makineler düşünebilir mi?" sorusuna yanıt bulmaya çalışırken Enigma şifresini kırmasına dayanır. Yapay zeka ve makine öğrenmesi bu süre zarfında çok ilerlemiş ve içinde sağlık alanının da yaygın olarak kullanıldığı çok geniş bir yelpazeye yayılmıştır.

2.1 Sağlık Alanında Sınıflandırma Yöntemleri Üzerine Yapılan Çalışmalar

Bu bölümde şu ana kadar literatürde yapılmış olan diğer çalışmalarla ilgili olarak incelemelerde bulunulmuştur. Konuyla ilgili yapılan diğer çalışmaların genel bilgisi çizelge 2.1'de sunulmuştur. Çizelgede çalışmanın adı, çalışma ile ilgili özet bilgi, çalışmada kullanılan veri kümesi ve kullanılan yöntem yer almaktadır.

Çizelge 2.1 Yapılan diğer çalışmalar

Çalışmanın Adı	**Özet**	**Kullanılan Veri Kümesi**	**Kullanılan Yöntem**
Göğüs Kanseri Tanı Sistemi: Kaba Kümeler ve Olasılıksal Sinir Ağları Kullanarak Kombine Bir Yaklaşım (Revett vd. 2005)	Kaba kümeler (Rough Set – RS) kullanılarak boyut azaltılmış ve Olasılıksal sinir Ağı (Probabilistic Neural Network – PNN) uygulanmıştır.	Göğüs kanseri (WBCD)	RS ve PNN
Veri tabanı sınıflandırması için CBR (Case Based Reasoning – Durum tabanlı sebeplendirme) tabanlı bulanık karar ağacı yaklaşımı (Chang vd 2010)	CBFDT CBFDT'nin Sinir ağı (Neural Network - NN), SVM, KNN ile karşılaştırılması yapılmıştır.	Iris, Şarap, BUPA Karaciğer hastalıkları, Wisconsin Diagnostic Breast Cancer (WDBC), Gebeliği önleyici Yöntem Seçimi (Contraceptive Method Choice - CBC) (UCI)	Kademeli Regresyon Analizi (Stepwise Regression Analysis - SRA),FDT,GA kullanılması ile hibrit bir CBFDT modeli

Çizelge 2.1 Yapılan diğer çalışmalar (devam)

Tıbbi veri sınıflandırması için durum tabanlı sebeplendirme ve bulanık karar ağacını bir araya getiren hibrit bir model (Fan vd. 2011)	Veri setinin ön işleme tabi tutulması için durum tabanlı bir kümeleme yöntemi uygulanır, böylece her küme içinde daha homojen bir veri elde edilir. Daha sonra her kümedeki verilere FDT uygulanır ve belirlenen özelliklere ve hastalıklara dayanarak bir karar verme sistemi oluşturmak için GA uygulanır. Son olarak, her küme için bir dizi bulanık karar kuralları üretilir. CBFDT'nin SVM, KNN, NB, FDT ile karşılaştırılması yapılmıştır.	Göğüs Kanseri Wisconsin (WBCD), Bupa Karaciğer Hastalığı (UCI)	SRA, FDT, GA kullanılması ile hibrit bir CBFDT modeli
Tıbbi veri sınıflandırması için hibrit akıllı sistem (Seera vd. 2013)	Veri örneklerinden Bulanık Min-Max (Fuzzy Min Max - FMM) sinir ağı sayesinde aşamalı olarak öğrenebilir, CART sayesinde öngörülen çıktılarını açıklayabilir ve Rastgele Orman (RF) sayesinde yüksek sınıflandırma performansları elde edebilir.	Göğüs Kanseri Wisconsin (WBCD), Pima Yerlileri Diyabeti ve Bupa Karaciğer Hastalığı (UCI)	FMM, CART ve RF modelinden oluşan hibrit bir akıllı sistem
Tıbbi verilere uygulanan topluluk özellik sıralaması (Santos vd 2014)	Boyut azaltma işlemi için özellik seçme yöntemi olan özellik sıralaması SVM, BAG, RF, NB öğrenme algoritmaları ile kullanılmıştır.	Göğüs Kanseri (KDD Kupası 2008 web sitesi)	Özellik sıralaması, SVM, BAG, RF, NB
Makine Öğrenmesi Teknikleri Kullanılarak Meme Kanseri Teşhisinin Performans Değerlendirmesi (Bektaş ve Babur 2016)	Destek Vektör Makinesi (SVM), RF, K-Yıldız (K-star), Seçimli Algılayıcı Sinir Ağı	Kent Ridge 2	Destek Vektör Makinesi (SVM), RF, K-Yıldız (K-star), Seçimli Algılayıcı Sinir Ağı
Centroid Sınıflayıcılar Yardımıyla Meme Kanseri (Takcı 2016)	Takcı (2016) çalışmasında centroid sınıflayıcıları; C4.5, SVM, k-NN ve çok katmanlı algılayıcı (MLP) gibi yöntemlerle karşılaştırılmıştır.	Wisconsin (WBCD) (UCI)	Centroid Sınıflayıcılar

Çizelge 2.1 Yapılan diğer çalışmalar (devam)

Bulanık mantık yöntemi kullanılarak meme kanseri sınıflaması için bilgi tabanlı bir sistem (Nilashi vd. 2017)	Veri kümelemesinde Beklenti Maksimizasyonu (Expectation Maximization - EM), çoklu topluluk sorununu çözmede bir boyut azaltma tekniği olan Temel Bileşenler Analizi (Principal Component Analysis - PCA), bulanık kuralların üretilmesinde CART kullanılarak sınıflandırmada bulanık kural tabanlı bir sistem geliştirilmiştir.	Göğüs Kanseri Wisconsin (WDBC), Mamografik kitle veri kümesi (UCI)	EM, PCA, CART ve Bulanık Kural Tabanlı yöntemlerin bir kombinasyonu olan hibrit akıllı bir sistem

Literatür taraması sırasında incelenen konuyla ilgili yapılan diğer çalışmaların performans değerlendirmeleri çizelge 2.2'de sunulmuştur.

Çizelge 2.2 Yapılan diğer çalışmaların performans değerlendirmeleri

Çalışmanın Adı	**Önerilen Yöntemin Doğruluk Oranı**	
Göğüs Kanseri Tanı Sistemi: Kaba Kümeler ve Olasılıksal Sinir Ağları Kullanarak Kombine Bir Yaklaşım (Revett vd. 2005)	WBCD (9 özellik)	0.87
	WBCD (5 özellik)	0.86
	WBCD (3 özellik)	0.86
Veri tabanı sınıflandırması için CBR tabanlı bulanık karar ağacı yaklaşımı (Chang vd 2010)	Iris (NN)	0.9713
	Iris (KNN)	0.9406
	Iris (CBFDT)	0.989
	Şarap (NN)	0.9556
	Şarap (KNN)	0.9248
	Şarap (CBFDT)	0.9766
	Karaciğer Hast.(NN)	0.6182
	Karaciğer Hast.(KNN)	0.581
	Karaciğer Hast (CBFDT)	0.818

Çizelge 2.2 Yapılan diğer çalışmaların performans değerlendirmeleri (devam)

Veri tabanı sınıflandırması için CBR tabanlı bulanık karar ağacı yaklaşımı (Chang vd 2010)	WDBC (NN)	0.9726
	WDBC (KNN)	0.969
	WDBC (CBFDT)	0.984
	Gebeliği önleyici Yöntem Seçimi (NN)	0.6386
	Gebeliği önleyici Yöntem Seçimi (KNN)	0.4085
	Gebeliği önleyici Yöntem Seçimi (CBFDT)	0.762
Tıbbi veri sınıflandırması için durum tabanlı sebeplendirme ve bulanık karar ağacını bir araya getiren hibrit bir model (Fan vd. 2011)	Karaciğer Hastalığı (SVM)	0.776
	Karaciğer Hastalığı (KNN)	0.737
	Karaciğer Hastalığı (NB)	0.702
	Karaciğer Hastalığı (FDT)	0.683
	Karaciğer Hastalığı (CBFDT)	0.904
	WBCD (SVM)	0.981
	WBCD (KNN)	0.969
	WBCD (NB)	0.914
	WBCD (DT)	0.902
	WBCD (CBFDT)	0.9890
Tıbbi veri sınıflandırması için hibrit akıllı sistem (Seera vd. 2013)	WBCD	0.9884
	Pima Yerlileri Diyabeti	0.7839
	Karaciğer Hastalığı	0.9501

Çizelge 2.2 Yapılan diğer çalışmaların performans değerlendirmeleri (devam)

Tıbbi verilere uygulanan topluluk özellik sıralaması (Santos vd. 2014) Göğüs Kanseri (KDD Kupası 2008 web sitesi)	RF+FR (özellik sıralaması)	0.931
	RF+bütün özellikler	0.925
	BAG+FR (özellik sıralaması)	0.920
	BAG+bütün özellikler	0.908
	SVM+FR (özellik sıralaması)	0.931
	SVM+bütün özellikler	0.912
Makine Öğrenmesi Teknikleri Kullanılarak Meme Kanseri Teşhisinin Performans Değerlendirmesi (Bektaş ve Babur 2016)	Kent Ridge 2 (DVM)	0.8453
	Kent Ridge 2 (K-Yıldız)	0.8041
	Kent Ridge 2 (Rastgele Orman)	0.9072
	Kent Ridge 2 (Seçimli Algılayıcı Sinir Ağı)	0.8144
Centroid Sınıflayıcılar Yardımıyla Meme Kanseri (Takcı 2016)	WBCD (Euclidian tabanlı centroid sınıflayıcı)	0.9904
	WBCD (Manhattan tabanlı centroid sınıflayıcı)	0.9856
	WBCD (Cosine tabanlı centroid sınıflayıcı)	0.9135
Bulanık mantık yöntemi kullanılarak meme kanseri sınıflaması için bilgi tabanlı bir sistem (Nilashi vd. 2017)	WDBC	0.932
	Mamografik	0.941

Çizelge 2.2 Yapılan diğer çalışmaların performans değerlendirmeleri (devam)

Kronik hastalık tahmini için özellik seçimi ve sınıflandırma sistemleri (Jain ve Singh 2018)	Pima yerlileri diyabet (geleneksel sınıflandırıcı sistem - tahmin edici hibrit model)	0.9238
	Pima yerlileri diyabet (geleneksel sınıflandırıcı sistem - C4.5)	0.8127
	Pima yerlileri diyabet ve WBCD (adaptif sınıflandırıcı sistem - adaptif SVM)	1

Revett vd. (2005) çalışmalarında veri hacmini azaltmak için RS ve makine öğrenmesi kullanarak sınıflandırma yapmak için ise PNN kullanmıştır. Çalıştıkları veri kümesi WBCD'dır. 9 özellikli olan WBCD verisi RS ile 5 ve 3 özelliğe indirilir. Öğrenme algoritması PNN kullanılırken verinin %70'lik bölümü eğitim için %30'luk bölümü ise test için kullanılır. 9 özellikli veri kümesinden %87,5 özellikli veri kümesinden %86 ve 3 özellikli veri kümesinden %86 doğruluk oranları elde edilir. Buradan da görüldüğü gibi bu örnek için özellik azaltma performans değerlerini ciddi oran etkilememiştir.

Chang vd. (2010) çalışmalarında iris, şarap, karaciğer hastalığı, göğüs kanseri ve gebelik önleyici yöntem seçimi verilerinin sınıflandırılmasında FDT ve GA kullanarak durum tabanlı karar vermek sistemi oluşturmuşlardır. Çalışmalarında literatürde NN, SVM ve KNN ile yapılan diğer çalışmalarla karşılaştırmışlar ve durum tabanlı bulanık karar ağacı (Case-Based Fuzzy Decision Tree - CBFDT) çalışmalarının daha iyi sonuç verdiğini gözlemlemişlerdir. FDT girdi özellikleri arasından yorumlanabilir kurallar üretir. GA FDT'nin doğruluğunu artırmak için uygulanmıştır. GA'yı etkileyen faktörlerden popülasyon boyutu, neslin sayısı, çaprazlama, ve mutasyon oranı kullanılmıştır.

Fan vd. (2011) çalışmalarında göğüs kanseri ve karaciğer hastalığını sınıflandırmak için Chang vd.'nin çalışmalarına benzer yöntemler izlemiş ve FDT ve GA kullanarak durum

tabanlı karar veren hibrit bir model tasarlamışlardır. Veri setinin ön işleme tabi tutulması için durum tabanlı bir kümeleme yöntemi uygulanır, böylece her küme içinde daha homojen bir veri elde edilir. Daha sonra her kümedeki verilere FDT uygulanır ve belirlenen özelliklere ve hastalıklara dayanarak bir karar verme sistemi oluşturmak için GA uygulanır. Son olarak, her küme için bir dizi bulanık karar kuralları üretilir. Sinir ağları ve linear algoritmaların yüksek doğrulukla sınıflandırma yaptığını ancak bir kara kutu mantığıyla sınıflandırma sebeplerinin hangi özelliklerden geldiğini bulamadığını belirtmişlerdir. Kara kutu yaklaşımlarının neden olduğu bu problemi bulanık kurallar çözer.

Fan vd. kümeleme algoritması için basit ve etkili olan k-means algoritması, kümelemede özellik seçiminde optimal parametrelerin ayarlanması için de Kademeli Regresyon Analizi (Stepwise Regression Analysis - SRA) kullanmıştır. FDT girdi değişkenleri arasındaki yorumlanabilir kuralları üretir. Bulanık kuralların üretilmesinde en yaygın kullanılan üç temel üyelik fonksiyonu (membership function) üçgen (triangle), trapezoidal (trapezoid) ve gauss (gauss)'tur. Fan vd. bunlar arasından üçgen üyelik fonksiyonunu (triangle membership) seçmiştir. Sonraki çalışmalarında ise hedefleri diğer üyelik fonksiyonlarını kullanmaktır. FDT'nin doğruluk oranını artırmak için ise GA kullanmıştır. GA'yı etkileyen 4 temel faktör popülasyon boyutu, neslin sayısı, çaprazlama, ve mutasyon oranı kullanılmıştır. Rastgele seçilen %75'lik veri, modeli eğitmek için kullanılırken, %25'lik veri ise modelin testi için kullanılmıştır. Fan vd. doğruluk oranlarını literatürde yapılan çalışmalardan SVM, KNN, NB, FDT ile elde edilmiş doğruluk oranları ile karşılaştırmış ve göğüs kanseri verisinde 0.9890, karaciğer hastalıkları verisinde 0.9040 ile durum tabanlı bulanık karar ağacının CBFDT'nin doğruluk oranının en iyi sonuç olduğunu tespit etmişlerdir.

Seera vd. (2013) tarafından yapılan çalışmada Bulanık Min - Max sinir ağı (FMM), CART ve RF modelinden oluşan hibrit bir akıllı sistem önerilmiş ve tıbbi veri sınıflandırma için bir karar destek aracı olarak etkinliği incelenmiştir. Seera vd. bu hibrit akıllı sistem ile, kurucu modellerin avantajlarından yararlanmak ve aynı zamanda sınırlamalarını hafifletmek amaçlanmaktadır. Model, veri örneklerinden FMM sinir ağı sayesinde aşamalı olarak öğrenebilir, CART sayesinde öngörülen çıktılarını

açıklayabilir ve RF sayesinde yüksek sınıflandırma performansları elde edebilir. Önerilen hibrit model, hata tespiti ve arıza teşhisi problemleri için çevrimdışı bir modele (yani FMM-CART) odaklanan Seera vd. (2012) önceki kendi çalışmalarının bir uzantısıdır.

Seera vd. tıbbi karar verme görevlerini desteklemek için çalışmasında bulanık sinir ağları, bulanık olasılıksal sinir ağları ve bulanık öğrenme vektör niceleme ağları gibi makine öğrenme modelleri kullanmıştır. Ancak bu modellerin en önemli kısıtlılığı, tahminlerini açıklayamama eksikliği olduğu çalışmasını göz önüne alarak, burada bir girdi vakasıyla uğraşırken nedeni ortaya koyan ve öngörüleri için gerekçe gösterebilen bir makine öğrenme tabanlı sistem geliştirmeye çalışmışlardır. Seera vd. göre CART, bir ağaç yapısı biçiminde kural çıkaran bir avantaja sahip olsa da, veri örneklerinden artan öğrenimde daha az esnektir. FMM, artımlı öğrenme özelliklerine sahip tek geçişli eğitim avantajına sahip olsa da, tahminlerini açıklamak için kurallar üretme yeteneğinden yoksundur. Öte yandan, RF, yüksek tahmin doğruluğu sağlamak için en iyi ağacın tanımlanabileceği bir CART topluluğu oluşturma yararına sahiptir. Bu nedenle hibrit model, veri örneklerinden Bulanık Min-Max sinir ağı sayesinde artan bir şekilde öğrenebilir, Sınıflandırma ve Regresyon Ağacı sayesinde öngörülen çıktılarını açıklayabilir ve Rastgele Orman sayesinde yüksek sınıflandırma performansları elde edebilir. Seera vd. iki temel amacı vardır. İlk amacı sistemin ürettiği öngörüye nasıl ulaştığını anlaması yani hastalığın sebebinin kuralını çıkarmaktır. İkinci amaç ise bir tarama sisteminin yüksek oranda hatalı negatif oranı, gerekli tıbbi müdahaleyi almaktan yoksun bırakarak hastaların riskini artırabilirken, yüksek düzeyde yanlış bir alarm oranı, hastalarda gereksiz endişe ve strese neden olacağından doğruluk oranını arttırmaktır. Bu iki amaca uygun olarak, Seera vd.'nin önerdikleri hibrit model, belirttikleri gibi sadece yüksek doğruluk, hassasiyet ve özgüllük oranlarını elde etmekle kalmayıp aynı zamanda karar ağacı şeklinde öngörüleri için açıklama da sağlayabilir. Çalışmalarında UCI Makine Öğrenimi Deposundan Göğüs Kanseri (WBCD), Pima Yerlileri Diyabeti Karaciğer Hastalığı verilerini kullanmışlardır. 2, 5 ve 10 kat çapraz doğrulamaya denenmiştir. 10-kat çapraz doğrulamayla; WBCD'de FMM 0.9526, FMM-CART 0.9571, FMM-CART-RF 0.9884; Pima Yerlileri'nde FMM 0.6928, FMM-CART 0.7135, FMM-CART-RF 0.7839; Diyabeti Karaciğer Hastalığında FMM 0.6725,

FMM-CART 0.9261, FMM-CART-RF 0.9501 doğruluk oranları elde edilmiştir.

Bektaş ve Babur (2016) yaptıkları çalışmada Destek Vektör Makinesi (SVM), RF, K-Yıldız (K-star), Seçimli algılayıcı sinir ağı algoritmalarını kullanarak Kent Ridge 2 mikrodizi veri seti kullanarak sınıflandırma yapmışlar ve sonuçları karşılaştırmışlardır. Performans ölçütü olarak doğruluk, duyarlılık, kesinlik, Alıcı Çalışma Karakteristiği (Receiver Operator Characteristic - ROC) eğrisi ve eğri altındaki alan (Area Under Curve - AUC) değerlerini kullanmışlardır. Bektaş ve Babur (2016) iki sınıflı verilerin tahmininde geleneksel SVM'ye göre daha gelişmiş olan LibSVM kullanmışlar ayrıca K-star algoritmasında iki özelliği birbirine bağlayan en kısa uzaklık olarak Kolmogorov mesafesini dikkate almışlardır. 10 kat çapraz doğrulama kullanmışlardır. Performans ölçütleri çizelge 2.2'de gösterildiği gibidir.

Takcı (2016) çalışmasında centroid sınıflayıcıları; C4.5, SVM, k-NN ve çok katmanlı algılayıcı (MLP) gibi yöntemlerle karşılaştırılmıştır. Euclidian tabanlı centroid sınıflayıcı %99,04 değeriyle orijinal Wisconsin veri setinde diğer sınıflayıcıları geçerek en iyi sonucu vermiştir (Takcı, 2016). Performans ölçümü için hem doğruluk ölçümüyle hem de ROC analizi yöntemi kullanılmıştır. Sınıflayıcıları işlem hızı açısından da değerlendirerek centroid tabanlı sınıflayıcıların diğerlerinden belirgin derecede hızlı olduğu değerlendirerek düşük işlem maliyetine dikkat çekmiştir. Düşük işlem maliyeti ve yüksek tanıma doğruluklarına sahip centroid sınıflayıcılar diğer sınıflayıcılar gibi meme kanseri teşhisinde kullanılabilir sınıflayıcılardır (Takcı, 2016).

Santos vd. (2014) çalışmalarında ilişkilendirmelerin tanımlamalarına bağlı olduğu için özelliklerin azaltılmasının sınıflandırmada önemli hassas ve önemli olduğunu belirtmiş, göğüs kanseri verilerinde kullandıkları etkili özellik sıralaması algoritmalarından bahsetmiştir. Literatürde karar vermek için en uygun hangi özelliklerin kullanıldığını öğrenmek için birçok makine öğrenme algoritması geliştirilmiştir. Zayıf, tekrarlı ve gereksiz özelliklerin atılarak boyut azaltma işlemi çokça istenen bir durumdur. Santos vd. çalışmalarında boyut azaltmak için bir özellik seçme (Feature Selection - FS) yöntemi olan topluluk özellik sıralaması (Feature Ranking - FR) algoritması kullanmışlardır. Diğer boyut azaltma yöntemlerin aksine Santos vd.nin bu

çalışmalarında tekrarlanan veri silinmez, kullanılacak veri ana kümenin alt kümesini kullanılır böylece yorumlamada daha iyi sonuçlar alırlar. Deneysel karşılaştırma yapmak için SVM, BAG, RF, NB öğrenme algoritmalarını Göğüs Kanseri (KDD Kupası 2008 web sitesi) verisinde uygulamıştır. Sınıflandırıcıların performans değerlendirmeleri AUC, hassasiyet ve yanlış pozitif oranı kullanılarak yapılmıştır.

Nilashi vd. (2017) tarafından yapılan çalışmanın amacı, meme kanseri hastalığı verilerinde hastalıkla ilgili tahmin yürütmektir. Bunu yaparken, bulanık kural gerekçelendirme yöntemini kullanmış, bulanık kuralları keşfederek tahmin modelleri oluşturmuşlardır. Yazarlar öncelikle verilerde önişlemler yapmışlardır. Birçok alanda kümeleme yöntemi performans artırıcı bir işlem olduğundan, tıbbi alanda hastalık tanı sistemlerinin doğruluğunu arttırmada da bu yöntem önemli bir rol oynamıştır. Daha sonra kümeleme tekniği olarak Beklenti Maksimizasyonu (Expectation Maximization - EM) kullanmışlardır. Bu kümeleme işlemi sınıflandırıcının verilerden tahmin modellerini daha iyi öğrenebilmesini sağlamıştır. Bu işlemden sonra boyut azaltma tekniği olan Temel Bileşenler Analizi (Principal Component Analysis - PCA) kullanmıştır. Böylece potansiyel ses elenmiştir. Bulanık kuralların üretilmesinde CART kullanmıştır. Kümeleme işlemiyle oluşmuş her kümeye tahmin modelleri inşa etmiştir. Böylece EM, PCA, CART ve Bulanık Kural Tabanlı yöntemlerinin kullanıldığı akıllı bir hibrit sistem geliştirmiştir.

Nilashi vd. bu çalışma için Wisconsin Teşhis Meme Kanseri ve Mamografik kitle veri kümeleri kullanılmıştır. Nilashi vd. göre bu veri kümelerindeki özellikler, göğüs kanseriyle ilgili önceki çalışmalarda kullanılan en fazla risk faktörü olan özelliklerdir, dolayısıyla bu veriler literatürde kullanılan yöntemlerin karşılaştırılması için en yaygın referans gösterilen veri kümeleridir. Nilashi vd. çalışmasında tüm kümelerde regresyon ağaçları oluşturmuştur. İnşaat iki ayrı aşamada, büyüme ve budama aşamasından oluşur. Büyüme aşamasında yukarıdan aşağıya, ağaç, düğümleri özyinelemeli bir şekilde bölerek yapar. Budama aşamasında aşağıdan yukarıya, müteakip düğümleri çıkararak alakasız dalları ortadan kaldıran ağaç budanır, bir düğüm bir yaprağa dönüştürür. Büyüme ve budama sırasında regresyon ağacının amacı budama aşamasındaki optimal ağacı aramak ve bulmaktır. Nilashi vd. önerdikleri yöntem için 10 kat çapraz doğrulama

uygulamıştır ve sınıflandırma doğruluğu için ROC eğrisinin altındaki alanı (AUC) kullanmıştır. Sonuçlarda WBCD için Temel Bileşenler Analizi Destek Vektör Makinesi (Principal Component Analysis Support Vector Machine - PCA-SVM), Temel Bileşenler Analizi K-En yakın komşuluk (Principal Component Analysis – K Nearest Neighbours - PCA-KNN) ve karar ağacı algoritmaları ile karşılaştırıldığında 0.932 doğruluk oranıyla en yüksek performansı aldığı görülmektedir. Mamografik kütle verisi için ise 0.941 doğruluk oranı olduğu görülmektedir.

Jain ve Singh (2018) uygun özellik seçiminin sınıflandırmanın doğruluğunu önemli ölçüde etkilediğini belirtmiş ve çeşitli özellik seçimi yöntemlerini inceleyerek avantaj ve dezavantajlarını incelemişlerdir. Ayrıca kronik hastalık tahmini için geleneksel sınıflandırma sistemleri, adaptif sınıflandırma sistemleri ve paralel sınıflandırma sistemlerini incelemişlerdir. Diyabet, kardiyovasküler hastalıklar, artrit, kanser, hepatit C, hipertansiyon, talasemi gibi kronik hastalıkların erken tespitinin ve etkili tedavinin her zaman hastalar için yararlı olduğu tespit edilmiştir. Jain ve Singh'e göre eğer veri doğru, eksiksiz, tekrarsız ve sesten arındırılmış ise, veri madenciliği ile tahmin etmek hızlı ve kolaydır. Jain ve Singh özellik seçimi yaklaşımlarını üç kategoride incelemiştir ve karşılaştırmalar yapmışlardır. Bunlar; filtreleme (Filter) yöntemleri, sarmal (wrapper) yöntemler, gömülü (embedded) yöntemlerdir. Ayrıca hibrit yöntemler de son zamanlarda öne çıkan ve kullanılan yaklaşımlardır. Filtreleme yöntemleri öğrenme algoritması kullanılmadan önce yapılır, ancak sarmal ve gömülü yöntemler, kullanılacak öğrenme algoritmasına bağlı kalarak özellik seçimi yapar. Filtreleme yöntemleri hacimli veri tabanları için sarma yöntemlerine göre daha fazla tercih edilirler. Ancak filtreleme yöntemlerinin kısıtı bir özelliğin diğerine bağımlılığını göz ardı etmeleri ve en kullanışlı özellikleri seçememeleridir. Sarmal yöntemlerde ise, eğer başka bir öğrenme algoritmasının kullanılması gerekiyorsa, bu yöntemin tekrar uygulanması gerekir çünkü hep öğrenme algoritması için optimal değerler farklıdır. Ayrıca, sarmal yöntemler çok karmaşık ve küçük eğitim veri kümelerinde ezber yapmaya meyillidir. Gömülü yöntemler ise eğitim verilerinin eğitim kümesine ve doğrulama kümesine ayrılmasını gerektirmediğinden daha hızlı bir çözüm sunar ve sarmal tekniklere göre daha az ezber (over-fitting) yapar. Ayrıca, gömülü yöntemlerin hesaplama karmaşıklığı, sarmal yöntemlerinden daha iyidir. Gömülü yöntemlerle ilgili

en büyük sınırlama, sınıflandırıcıya bağlı olarak kararlar almasıdır. Hibrit yöntemler ise özelliklerin sayısını önemli ölçüde azaltmış ve diğer özellik seçim algoritmalarına kıyasla sınıflandırma doğruluğunu arttırmıştır. Ayrıca, hibrit yöntemler ile hesaplanan maliyet ve zaman da azalmıştır. Jain ve Singh, SVM, K-en yakın komşuluk, NB, Sinir Ağları, Bayes Ağları, C4.5 sınıflayıcısı gibi sınıflandırıcıların adaptif sistemlerde kullanıldığını ayrıca sağlık endüstrisinde kullanılan paralel sınıflama sistemlerinde ağırlıklı olarak hadoop, STORM, Map Reduce programlama teknolojilerinin bulunduğunu belirtmiştir.

3. MATERYAL VE YÖNTEM

Bu çalışmada sınıflandırma yöntemlerinden olan yapay sinir ağları (perceptron öğrenme algoritması, derin öğrenme metotları) ve K en yakın komşuluk yöntemi kullanılmıştır.

3.1 Yapay Sinir Ağları

Yapay sinir ağları, biyolojik sinir ağlarının çalışma mantığıyla çalışır. Yapay sinir ağlarına öncelikle veri verilerek/yüklenerek yapay sinir ağının bilgi toplaması sağlanmakta ve bu yöntemle eğitilmektedir. Her bir birimin kendi belleği vardır ve bu birimler çeşitli işlemlerden geçerek yapay sinir ağının topladığı bilgi ile karşılaşmadığı veriler için tahminde bulunması sağlanır. Daha sonra gerçek değerler ile tahmini değerlerin karşılaştırılması yapılarak yapay sinir ağının hangi oranda doğru eğitildiği hesaplanarak performansı elde edilir.

Yapay sinir ağları yapay sinir hücrelerinin birbirine bağlanmasıyla oluşur ve bu yapay sinir hücrelerine işlem elemanı denir. İşlem elemanlarının 5 temel özelliği vardır. (Öztemel, 2006)

Girdiler: Yapay sinir ağına girdi verileridir. Girdi verileri yapay sinir ağlarına verilerek çeşitli işlemlerden geçerek eğitilir.

Girdilerin ağırlıkları: Her girdi verisinin çıktıyı etkileyen bir ağırlığı vardır.

Toplam Fonksiyonu: İşlem elemanına gelen net girdiyi hesaplar. Toplam fonksiyonu için; toplam, çarpım, maksimum, minimum, kümülatif toplam gibi fonksiyonlar kullanılabilir. Aşağıda toplam fonksiyonu için çarpım işlemi kullanılmıştır. Aşağıda ağırlık ve girdi vektörünün skaler çarpımının toplamı net değerini vermektedir.

$$net = < \omega, x > = \omega^T x = \sum_{j=1}^{n} \omega_j \, x_j = \omega_1 x_1 + \cdots + \omega_n x_n$$

Aktivasyon fonksiyonu: İşlem elemanına gelen net girdi işlenerek bir çıktı üretilir. Aktivasyon fonksiyonu için; lineer fonksiyon, unipolar/bipolar basamak (step) fonksiyonu, parçalı doğrusal fonksiyon, unipolar/bipolar sigmoid fonksiyonu, radyal temelli (Gaussian) fonksiyon gibi fonksiyonlar kullanılır. Aşağıda belirtilen $f(net)$ aktivasyon fonksiyonu, eşik değeri olan step fonksiyonudur ve o çıktı değerini belirler. θ değeri eşik değeridir.

$o = f(net) = f(< \omega, x >) = f(\omega^T x)$

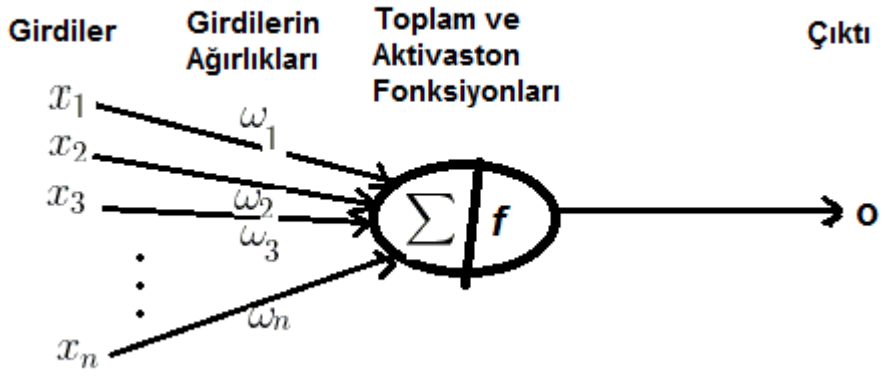
$o = f(net) = f\left(\sum_{j=1}^{n} \omega_j \; x_j\right)$

$$o = f(net) = \begin{cases} 1, & \omega^T x \geq \theta \\ 0 & \end{cases}$$

Sonuç olarak θ değerinden büyük olan net'ler için çıktı değeri 1, küçük olan net'ler için çıktı değeri 0'dır.

Çıktı: Girdi verilerinin işlem elemanında işlemden geçerek bir çıktı üretmesidir.

Şekil 3.1'de n girdili, tek çıktılı işlem elemanı gösterilmiştir.

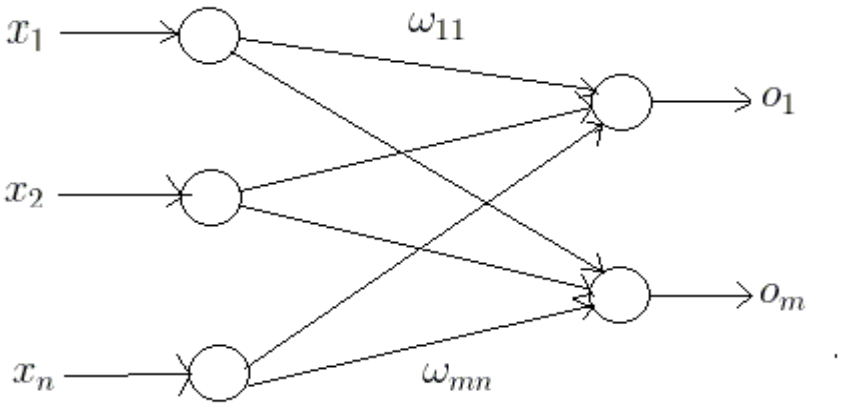


Şekil 3.1 n girdili, tek çıktılı işlem elemanı

3.1.1 Perceptron öğrenme

Perceptron öğrenme algoritması, en basit yapay sinir ağı yöntemlerindendir. Yapısı tek katmanlı bir ileri beslemeli ağ şeklindedir. Hata düzeltmeleri yaparak öğrenme sağlanır. Doğrusal ayrılabilir problemlerin çözümü için kullanılır. Şekil 3.2 n özellikli K tane

girdi ve m özellikli K tane çıktı olan bir eğitim kümesini gösterir. Çizelge 3.1 ise parametrelerin tanımlarını gösterir.



Şekil 3.2 Tek katmanlı ileri beslemeli ağ

No	Girdi Değerleri	Gerçek Çıktı Değerleri
1.	$x^1 = (x_1^1, ... , x_n^1)$	$y^1 = (y_1^1, ... , y_m^1)$
.	.	.
.	.	.
.	.	.
K.	$x^K = (x_1^K, ... , x_n^K)$	$y^K = (y_1^K, ... , y_m^K)$

Çizelge 3.1 Parametrelerin tanımları

$\omega_{i\,j}$	j'inci girdinin i'inci çıktıya olan ağırlığı
ω_i	i'inci çıktının ağırlık vektörü
x_1^1	Birinci girdinin birinci özelliği
x^1	Birinci girdi vektörü
x_1^K	K'ıncı girdinin birinci özelliği
x^K	K'ıncı girdi vektörü
x_1^k	k'ıncı girdinin birinci özelliği
x^k	k'ıncı girdi vektörü
y_i	i'inci istenen çıktı
o_i	i'inci hesaplanan çıktı
ŋ	Öğrenme oranı. Çok küçük bir değer seçilir.
ε	Hata değeri
n	Girdinin özellik sayısı
m	Çıktının özellik sayısı
θ	Eşik değeri. Burada 0'a eşitlenir.

Her bir k değeri için toplam fonksiyonu, aktivasyon fonksiyonuyla çıktı değeri ve ağırlık vektörü hesaplanır.

Öncelikle toplam fonksiyonu net hesaplanır.

$$\text{net} = \sum_{j=1}^{n} \omega_j \; x_j$$

Çıktı değerini hesaplayan aktivasyon fonksiyonu aşağıda belirtilen eşik değeri olan step fonksiyondur. Eşik değeri $\theta = 0$ 'dır.

$$o_i(x) = işaret(< \omega_i\ ,\quad x >) = \begin{cases} 1, & < \omega_i\ ,\quad x > \geq 0 \\ 0, & < \omega_i\ ,\quad x > < 0 \end{cases}$$

k'ıncı girdi verisinin çıktısı aşağıdaki gibi hesaplanır.

$o_i(x^k) = işaret(< \omega_i\ ,\ x^k >) = y_i^k$, i = 1,…, m

Amaç bütün k'lar için ω_i ağırlık vektörünü bulmaktır. Ağırlık vektörünü hesaplanırken, hesaplanan çıktı değeri ile gerçek çıktı değerlerinin karşılaştırması yapılır. Böylece; istenen çıktı ile hesaplanan çıktı farklı çıkarsa ağırlık vektöründe hata düzeltme yapılması gerektiği anlamına gelir. Sonuçlar aynı çıkarsa hata düzeltme yapılmaz. Perceptron öğrenme yönteminde ağırlık düzeltmeleri aşağıdaki gibi olur.

$\omega_i := \omega_i + \eta(y_i - o_i)\, x\ ,\quad \mathrm{i} = 1,\ldots,\mathrm{m}$

$$\omega_{ij} := \omega_{ij} + \eta(y_i - o_i)\, x_j\ ,\quad \mathrm{j} = 1,\ldots,\mathrm{n}$$

Burada $\eta > 0$ öğrenme oranıdır ve çok küçük bir değer olarak seçilir.

Hata değeri ise hataların kümülatif toplamıdır.

$$\varepsilon := \varepsilon + \frac{1}{2}\,\|y - o\|^2$$

Öğrenme döngüsü boyunca hata düzeltmeleri yapıldığından perceptron öğrenme algoritması bir hata düzeltme algoritmasıdır. Bütün ağırlık vektörleri değişmeden kaldığında ise öğrenme durmuş demektir.

Burada perceptron öğrenme algoritmasının sözde kodu yer almaktadır; (Fuller 1995)

K tane olan bir eğitim kümesinde

$$(x^1,\ y^1), \ldots, (x^K,\ y^K)$$

$x^k = (x_1^k, \ldots, x_n^k)$, $y^k = (y_1^k, \ldots, y_m^k)$, k = 1, … , K

Aşama 1: $\eta > 0$ öğrenme oranı olarak seçilir.

Aşama 2: Ağırlıklara ω_i rastgele küçük değerler atanır. Hata değeri $\varepsilon = 0'$a ve $k \coloneqq 1$ eşitlenir,

Aşama 3: Öğrenme burada başlar. x^k sunulur, $x \coloneqq x^k$, $y \coloneqq y^k$ ve çıktı $o = o(x)$ hesaplanır.

$$o_i(x) = işaret(< \omega_i , \quad x >) = \begin{cases} 1, & < \omega_i , \quad x > \geq 0 \\ 0, & < \omega_i , \quad x > < 0 \end{cases}$$

Aşama 4: Ağırlıklar aşağıdaki şekilde güncellenir.

$\omega_i \coloneqq \omega_i + \eta(y_i - o_i)\, x , \quad i = 1, \ldots, m$

Aşama 5: ε ye mevcut hatanın eklenmesiyle kümülatif hata döngüsü hesaplanır.

$\varepsilon \coloneqq \varepsilon + \frac{1}{2} \|y - o\|^2$

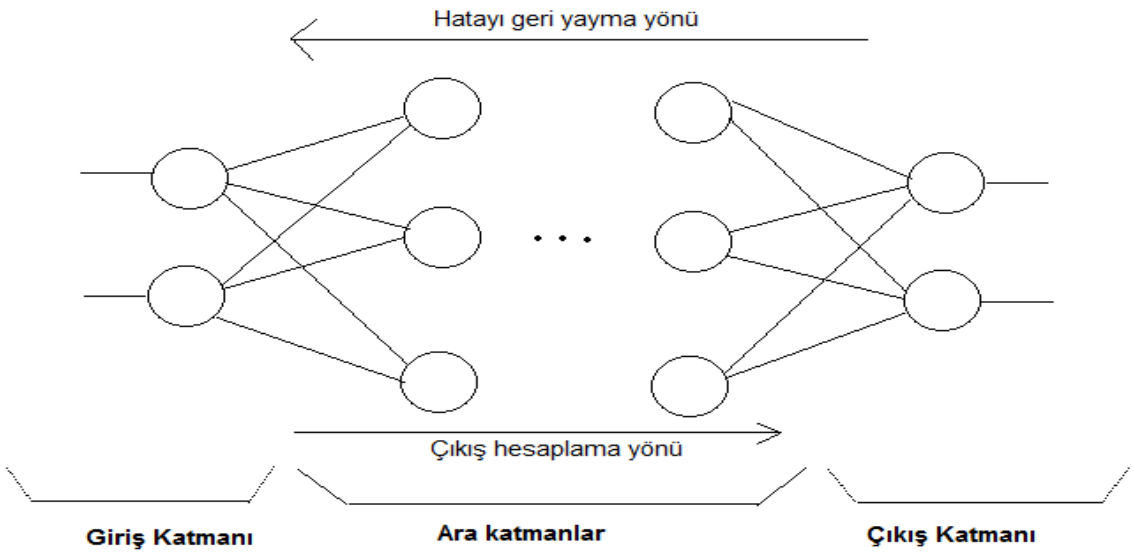
Aşama 6: Eğer k<K ise $k \coloneqq k + 1$ 'dir ve **Aşama 3**'e geri dönerek öğrenme devam ettirilir, ya da **Aşama 7**'e gidilir.

Aşama 7: Öğrenme döngüsü tamamlanır. $\varepsilon = 0$ değeri için öğrenme aşaması durdurulur. Eğer $\varepsilon >$ ise $\varepsilon = 0$ eşitlenir, $k \coloneqq 1$ ve **Aşama 3**'e giderek yeni bir öğrenme döngüsü başlatılır.

Perceptron öğrenme algoritmasının en büyük sorunu doğrusal olmayan problemlerin çözümünde kullanılamamasıdır.

3.1.2 Çok katmanlı yapay sinir ağları ve derin öğrenme

Şekil 3.3'te çok katmanlı bir yapay sinir ağı yapısı sunulmuştur. Giriş değerleri giriş katmanı, çıkış değerleri çıkış katmanı ve giriş ile çıkış arasında kalan katmanlara ara katman veya gizli katman denir.



Şekil 3.3 Çok katmanlı yapay sinir ağı

Derin öğrenmedeki derin ifadesi ara katmandaki sinir elemanlarının derinliğini yani gizli katmanın kaç katmandan oluştuğunu ifade eder (Anonim 2019).

Problemin cinsine ve veriye göre yapay sinir ağının katman sayısı belirlenmeli, buna göre model oluşturulmalıdır. Oluşturulan model üzerinde çıktı fonksiyonuna gelen değerler geri besleme ile geri yayılım yaparak ağırlık değerleri güncellenmekte ve en son aşamada ağırlık değeri bulunmaktadır. Yapay sinir ağının eğitilmesi aslında optimal ağırlık değerlerinin bulunması anlamına gelir ve bulunan bu optimal ağırlıklar üzerinden tahminde bulunur.

Geri beslemeli olabilmesi için aktivasyon fonksiyonunun türevinin alınabilir bir fonksiyon olması önemlidir.

3.1.3 Önerilen algoritma

Ayrıntısının aşağıdaki gibi olduğu K tane n özellikli girdi ve K tane m çıktımız vardır;

	Girdiler	Çıktılar
1.	$x^1 = (x_1^1, \ldots, x_n^1)$	$y^1 = (y_1^1, \ldots, y_m^1)$
.	.	.
.	.	.
.	.	.
K.	$x^K = (x_1^K, \ldots, x_n^K)$	$y^K = (y_1^K, \ldots, y_m^K)$

K: Veri sayısı

x: Girdi verisi

y: Çıktı verisi

n: Girdilerdeki özellik sayısı

m: Çıktılardaki özellik sayısı

Problemi basite indirgemek için başlangıçta girdiler 2 özellikli ve çıktılar 1 özellikli olarak düşünülür. Bu problem daha sonra n özellikli girdi ve 1 özellikli çıktı olarak adapte edilecektir. Ayrıca veri tabanımızda K tane kayıtımız vardır. Dolayısıyla:

	Girdiler	Çıktılar
1.	$x^1 = (x_1^1, x_2^1)$	$y^1 = (y^1)$
.	.	.
.	.	.
.	.	.
K.	$x^K = (x_1^K, x_2^K)$	$y^K = (y^K)$

değerlerini almaktadır.

Ayrıca y çıktısının alabildiği değerler C_1 ve C_2 olarak iki sınıfa ayrılmıştır.

$$y = \{C_1, C_2\}$$

Girdiler 2 özellikli olduğu için sistem 2 boyutlu uzayda düşünülmelidir. 2 boyutlu sistemde her girdi bir "nokta" ve her çıktı doktorların aldığı gerçek "kararlar" olarak düşünülür. 2 noktadan geçen bir doğru denklemi bulmak için 2 nokta alınır.

2 noktadan geçen 2 boyutta çizgi denklemi;

$ax_1 + b\,x_2 + c = 0$ 'dir. a,b,c katsayılardır ve girdiler;

x_1^k, x_2^k
x_1^{k+1}, x_2^{k+1} ‘dir.

Herhangi 2 noktadan geçen olası bütün çizgi denklemlerini bulabilmek için K tane kayıtta noktaların tüm 2'li bütün kombinasyonları bulunur. Her bir çizgi için çizgi denklemleri bulunduktan sonra, veri tabanında olan her nokta doğru denkleminde yerine konulur ve sonucun 0'dan büyük olup olmadığına bakılır. Ve aşağıda açıklanan bazı koşullara göre, veri tabanındaki kişilerin kayıtları olan noktaların toplamı hesaplanacaktır.

4 durum vardır;
y $= C_1$ iken
ax_1 + bx_2 + c $\geq$ 0 iken m’yi hesapla
ax_1 + bx_2 + c < 0 iken k’yı hesapla

y $= C_2$ iken; noktalar:
ax_1 + bx_2 + c $\geq$ 0 iken t’yi hesapla
ax_1 + bx_2 + c < 0 iken j’yi hesapla
Bu sayede 2 boyutlu uzay 2 parçaya bölünmüş olur.

Kriter1 ve Kriter2 değerleri bulunur;

$$Kriter1 = \frac{k + t}{K}$$

$$Kriter2 = \frac{m + j}{K}$$

Kriter değeri bulunur;

$$Kriter = enk\text{üçük}\ (Kriter1, Kriter2)$$

Yukarıda açıklanan her döngü (2’li kombinasyon) için kriter değeri kontrol edilir. Eğer yeni hesaplanan kriter eski kriter değerinden küçük ise kriter değeri değiştirilir. Burada amaç en küçük kriter değerine yaklaşmaktır.

$$GenelKriter = enk\text{üçük}\ (Kriterler)$$

Genel kriter algoritmada hata oranını belirler. Genellikle bu hata oranı çok küçüktür çünkü küçük değerlere yakınsayarak daima en küçük oranları bulmaktadır.

Problem, Wisconsin Üniversitesi verilerinden Göğüs kanseri verilerine yani n (ilk özellik id ve son özellik sınıf hariç 9) özellikli girdi ve 1 özellikli çıktının olduğu örneğe adapte edilir.

Veri tabanımızda K tane (699) kayıt vardır.

	Girdiler	Çıktılar
1.	$x^1 = (x_1^1, ... , x_n^1)$	$y^1 = (y^1)$
.	.	.
.	.	.
.	.	.
K.	$x^K = (x_1^K, ... , x_n^K)$	$y^K = (y^K)$

Ayrıca karar olan çıktılar 2 (iyi huylu) ve 4 (kötü huylu) değerlerini almaktadır.

y = {2,4}

değerlerini almaktadır.

y = $\{C_1 , C_2 \}$

Girdiler n özellikli olduğu için uzay N boyutlu olarak düşünülecektir.

2 boyutlu uzayda olduğu gibi girdiler (her bir hasta kaydı) N boyutlu denklemde "nokta" olarak ve çıktılar doktorların verdiği gerçek "kararlar" olarak düşünülürse, n noktadan geçen düzlemi bulmak için n tane nokta kullanılır. N boyutta n tane noktadan geçen düzlem denklemi;

ax_1 + bx_2 $+ \cdots + tx_n + u$ = 0'dır. a, b, ...t, u katsayılardır.

x_1 , x_2 ..... , x_n ler bir (girdinin) hastanın hastalık özellikleridir.

Dolayısıyla k'ıncı girdide (hastada)

x_1^k, x_2^k ..... , x_n^k k'ıncı girdide n tane özellik vardır.

Burada 2-boyut n-boyut hale gelmektedir. Bütün olası noktalar için, olası bütün düzlemler ve/veya hiperdüzlem denklemleri (burada n-boyut olduğundan doğru düzleme/hiperdüzleme dönüşmektedir) bulabilmek için K tane kayıt içinde olası bütün n-kombinasyonları bulunacaktır.

Yukarıda açıklanan 4 koşula göre, kayıtlardaki m, k, t, j değerleri bulunur.

y = C_1 iken

ax_1 + bx_2 + $\cdots$ + tx_n + $u \geq 0$ iken m'yi hesapla

ax_1 + bx_2 + $\cdots$ + tx_n + $u < 0$ iken k'yı hesapla

y = C_2 iken; noktalar:

ax_1 + bx_2 + $\cdots$ + tx_n + $u \geq 0$ iken t'yi hesapla

ax_1 + bx_2 + $\cdots$ + tx_n + $u < 0$ iken j'yi hesapla

Kriter1 ve Kriter2 değerleri bulunur;

$$Kriter1 = \frac{k + t}{K}$$

$$Kriter2 = \frac{m + j}{K}$$

Kriter değeri bulunur.

$$Kriter = enk\text{üçük}\ (Kriter1, Kriter2)$$

Test edildiğinde;

$K = m + k + t + j$ olduğu görülür.

Burada en küçük kriter değeri ve düzlem denklemi bulunduğunda öğrenme tamamlanır. Test aşamasında, en küçük kriter değerinin bulunduğu düzlem denklemi, doktor kararlarını bilmediğimiz kayıtların sonuçları hakkında tahminler yürütmede bize yardımcı olacaktır.

Bütün kombinasyonlara bakılırken, kayıtları hatasız olarak 2 parçaya ayıran "kriter=0" durumu bulunduğunda döngü durur.

Bütün kombinasyonlara bakıldığında veya kriter değeri 0'a eşitlendiğinde öğrenme döngüsü biter. Algoritmanın çalışması bittiğinde elde bir düzlem denklemi ve genel bir kriter değeri vardır. "Genel kriter" değerine yaklaşmak için bu denklem ve hata oranı ile yeni kayıtlarda testler yapacağız. Böylece bu kriter sayesinde belli bir hata payı ile doktor kararlarını bilmediğimiz hastaların kayıtları için tahminlerde bulunabileceğiz.

Burada n-boyutlu uzay için önerilen algoritmanın sözde kodu vardır.

Aşama 1: K tane kayıt içinde olası bütün n-kombinasyonları bulunur.

Aşama 2: Bütün n-kombinasyonları için $ax_1 + bx_2 + \cdots + tx_n + u = 0$ olan bir düzlem denklemi bulunur. Burada a, b, ...t, u katsayılardır ve girdiler

$x^1 = (x_1^1, \ldots, x_n^1)$

.

.

.

$x^K = (x_1^K, \ldots, x_n^K)$ şeklindedir.

Aşama 3: Veri tabanında olan her nokta denklemde yerine koyulur ve sonucun 0'dan büyük olup olmadığı öğrenilir. 4 koşula göre, kayıtlar olan noktalar sayılır.

4 koşul vardır:

$y = C_1$ iken

$ax_1 + bx_2 + \cdots + tx_n + u \geq 0$ iken m'yi hesapla

$ax_1 + bx_2 + \cdots + tx_n + u < 0$ iken k'yı hesapla

$y = C_2$ iken; noktalar:

$ax_1 + bx_2 + \cdots + tx_n + u \geq 0$ iken t'yi hesapla

$ax_1 + bx_2 + \cdots + tx_n + u < 0$ iken j'yi hesapla

Aşama 4: Kriter1, Kriter2, Kriter ve Genel Kriteri bulunur.

$$Kriter1 = \frac{k + t}{K}$$

$$Kriter2 = \frac{m + j}{K}$$

$$Kriter = enk\text{üçük}\ (Kriter1, Kriter2)$$

$$GenelKriter = enk\text{üçük}\ (Kriterler)$$

Aşama 5: Kriter=0 ise dur.

Aşama 6: Öğrenme döngüsü burada tamamlanır. Belli bir hata payı ile yeni kayıtlar düzlem denkleminde test edilir.

3.2 K-En Yakın Komşuluk Yöntemi

K-en yakın komşuluk algoritması temel mantığı basittir ve algoritmadaki k, bir noktaya en yakın olan k tane komşu sayısını belirtir. Bir veri kümesi içinde bir nokta alınır ve bu nokta, kendisine en yakın k tane komşudan sınıflaması fazla olan komşunun sınıfında sınıflandırılır. Veri kümesine yeni bir nokta eklendiğinde ise bu noktanın her bir noktaya uzaklığı hesaplanır ve k tane kendine en yakın komşusu içinden sayısı fazla olanın sınıfında sınıflandırılır.

K-en yakın komşuluk algoritması için uzaklık fonksiyonları için genelde Euclidean, Manhatton, Minkowski fonksiyonları kullanılır. k değeri genelde tek sayı olarak seçilir, bunun sebebi ise bir noktanın sınıfı kendisine en yakın komşularının sınıflarına göre belirlenirken sınıflardan birinin seçimi için kesinlik oluşturmasıdır.

4. ARAŞTIRMA BULGULARI

4.1 Veri Kümesi Bilgisi

Yapılan çalışmalar esnasında UCI Makine Öğrenme Deposunda bulunan Wisconsin göğüs kanseri, Pima Yerlileri diyabet veri tabanı, Bupa karaciğer hastalıkları mamografik kitle verileri kullanılmıştır. Veri kümeleri öğrenme ve test kümesi olmak üzere ikiye parçaya ayrılmıştır. Aksi belirtilmedikçe öğrenme modellerinde verilerin %70'i üzerinde öğrenme modeliyle öğrenme yapılmış, %30'u üzerinde ise testler yapılmıştır. Derin öğrenme için performans değerlendirmeleri doğruluk (accuracy), hassasiyet (precision), kesinlik (recall) ve f-ölçüsü (f-score) değerleri hesaplanarak yapılmıştır.

Wisconsin göğüs kanseri veri kümesi için özellik bilgileri çizelge 4.1'de sunulmuştur. Burada sınıflandırma özelliği son sütun ile belirlenmiştir. 15 Temmuz 1992'de Wisconsin Üniversitesi tarafından üretilen veri tabanında 699 hastanın verisi bulunmaktadır.

Çizelge 4.1 Wisconsin göğüs kanseri veri kümesi için özellik bilgileri

No	Özellik	Aralık
1.	Örnek Kod Numarası	ID Numarası
2.	Küme Kalınlığı	1 - 10
3.	Hücre Boyutunun Benzerliği	1 - 10
4.	Hücre Biçiminin Benzerliği	1 - 10
5.	Marjinal Bağlılık	1 - 10
6.	Tek Epitelyal Hücre Boyutu	1 - 10
7.	Yalın Çekirdekler	1 - 10
8.	Mülayim Kromatin	1 - 10
9.	Normal Nükleoli	1 - 10

Çizelge 4.1 Wisconsin göğüs kanseri veri kümesi için özellik bilgileri (devam)

10.	Mitoz	1 - 10
11.	Sınıf	2: İyi huylu için, 4: Kötü huylu için.

Eksik Özellik Değerleri olup bazı özellikleri '?' ile gösterilen 16 hasta mevcuttur. 699 hasta içinden 458 hastanın (65.5%) hastalık sınıflaması 'İyi huylu', 241 hastanın (34.5%) hastalık sınıflaması ise 'Kötü huylu' olarak yapılmıştır. Örnek olarak;

- 1017023,4,1,1,3,2,1,3,1,1,2
- 1017122,8,10,10,8,7,10,9,7,1,4

1017023 id numaralı hasta, doktorun '2' kararı ile 'İyi huylu' kanserdir.
1017122 id numaralı hasta, doktorun '4' kararı ile 'Kötü huylu' kanserdir.

Bupa karaciğer hastalığı verisi özellik bilgileri , Çizelge 4.2'de sunulmuştur. 15 Mayıs 1990'da BUPA Tıbbi Araştırma Şirketi tarafından üretilen veri kümesi 345 bekar erkek bireyin kaydını tutar. Kayıtlarda eksik veri yoktur. İlk 5 değişken karaciğer rahatsızlıklarında aşırı alkol tüketiminden etkilenebileceği düşünülen kan testleridir. 5'ten fazla içilen miktar verilerde seçici hale gelmektedir.

Çizelge 4.2 Bupa karaciğer hastalığı veri kümesi için özellik bilgileri

No	Özellik
1.	Ortalama corpuscular hacmi
2.	alkalin fosfotaz (alkphos)
3.	alamin aminotransferaz (sgpt)
4.	aspartat aminotransferaz (sgot)
5.	gama-glutamil transpeptidaz (gammagt)
6.	Günlük tüketilen yarım bardak alkollü içeceğe eşdeğer içecek miktarı
7.	Verileri ikiye ayırmak için kullanılan veri

Pima Yerlileri diyabet veri kümesi özellik bilgileri çizelge 4.3'te sunulmuştur. 9 Mayıs 1990'da Ulusal Sindirim ve Böbrek hastalıkları ve Diyabet Enstitüsü tarafından üretilen veri kümesinde 768 kayıt vardır. Sınıflandırmanın 1 olması diyabet hastalığının mevcut olması anlamına gelir. 500 hastada diyabet mevcut değilken, 268 hastada diyabet mevcuttur.

Çizelge 4.3 Pima yerlileri diyabet veri kümesi için özellik bilgileri

	Özellik	
1.	Hamilelik sayısı	
2.	2 saatlik oral glukoz tolerans testinde plazma glikoz konsantrasyonu	
3.	Diyastolik kan basıncı (mm Hg)	
4.	Triceps cilt kıvrım kalınlığını (mm)	
5.	2 saatlik serum insülini (mu U / ml)	
6.	Vücut kitle indeksi (kg cinsinden ağırlık / (m cinsinden yükseklik) ^ 2)	
7.	Diyabet soyağacı fonksiyonu	
8.	Yaş (yıl)	
9.	Sınıf değişkeni (0 veya 1)	1: Diyabet 0: Diyabet değil.

Mamografig kitle verisi özellik bilgileri Çizelge 4.4'te sunulmuştur. Ekim 2007'de Erlangen-Nuremberg Üniversitesi Radyoloji Enstitüsü tarafından üretilen veri kümesinde 961 kayıt vardır. Önem derecesinin 0 olması iyi huylu kitlenin varlığı, 1 olması ise kötü huylu kitlenin varlığı anlamına gelir. 445 kişide iyi huylu kitle varken, 516 kişide kötü huylu kitle mevcuttur. Eksik Özellik Değerleri olup bazı özellikleri '?' ile gösterilen 162 hasta mevcuttur.

Çizelge 4.4 Mamografig kitle veri kümesi için özellik bilgileri

	Özellik	
1.	BI-RADS değerlendirmesi	1 - 5 arası değerler
2.	Yaş	
3.	Kitle şekli	1: yuvarlak 2: lobüler 3: düzensiz
4.	Kitle marjı	1: sınırlandırılmış 2: mikro kütleli 3: gizlenmiş 4: gizlenmiş 5: eklenmiş
5.	Kitle yoğunluğu	1: yüksek 2: izo 3: düşük 4: yağ içeren
6.	Önem derecesi	0: iyi huylu 1: kötü huylu

4.2 Uygulama Altyapısı

Bu çalışmada makine öğrenmesi algoritmalarının uygulanması için geliştirme ortamı açık kaynak kodlu bir platform olan Java programlama dili kullanılmıştır. Daha önce Java program dili kullanılarak uygulamalar geliştirildiğinden dolayı daha hızlı adapte olunduğu için bu programlama dili kullanılmak üzere seçilmiştir. Perceptron öğrenme ve KNN algoritmaları için JAVA programlama dili kullanılarak model geliştirilmiştir. Derin öğrenme için ise Javanın derin öğrenme algoritmaları için tasarladığı deeplearning4java kütüphanesi kullanılmıştır.

Geliştirme yapılırken kişisel bilgisayar olan MacBook Pro kullanılmıştır. Intel Core i7 2,3 GHz işlemcili, 16 GB bellekli, dizüstü bir bilgisayardır. Derin öğrenmenin yüksek performans gerektirdiği aşikardır. Özellikleri daha iyi bilgisayarlar için daha iyi sonuçlar vermektedir. Kontrol için özellikleri daha iyi olan başka bir bilgisayarda

deeplearning4java kütüphanesi ile derin öğrenme modeli göğüs kanseri (WBCD) için denenmiş, sürenin üçte bir düştüğü görülmüştür.

4.3 Performans Değerlendirme

Göğüs Kanseri Verisi için Hata Matrisi çizelge 4.5'de gösterilmiştir.

Çizelge 4.5 Göğüs kanseri verisi için hata matrisi

	Negatif **İyi huylu (Tahmin)**	**Pozitif** **Kötü Huylu (Tahmin)**
Negatif **İyi huylu (Gerçek)**	DN (TN)	YP (FP)
Pozitif **Kötü Huylu (Gerçek)**	YN (FN)	DP (TP)

Verilerde kişiler İyi huylu (Hasta olmayan) ve Kötü huylu (Hasta) olarak sınıflandırılmıştır.

· **DN (Doğru Negatif – True Negative) :** Modelin doğru bir şekilde "Hasta olmayan" kişiyi "Hasta olmayan" olarak sınıflandırmasıdır. Doğru uygulamayla hasta olmayan kişiye ilaç tedavisine başlanmaz.

· **DP (Doğru Pozitif – True Positive):** Modelin doğru bir şekilde "Hasta" kişiyi "Hasta" olarak sınıflandırmasıdır. Doğru uygulamayla hasta olan kişiye ilaç tedavisine başlanır.

· **YN (Yanlış Negatif – False Negative):** Modelin yanlış bir şekilde "Hasta" kişiyi "Hasta olmayan" olarak sınıflandırmasıdır. Yanlış uygulamayla hasta olan kişi için ilaç tedavisine başlanmayacağı için kişi tedaviden mahrum kalarak ölüm riski ile karşılaşır.

· **YP (Yanlış Pozitif – False Positive):** Modelin yanlış bir şekilde "Hasta olmayan" kişiyi "Hasta" olarak sınıflandırmasıdır. Yanlış uygulamayla kişiye gereksiz olduğu halde ilaç tedavisi uygulanır.

Bu tablo "Hasta olmayan" ve "Hasta" olarak genel olarak hazırlanmış olup, hastalığın kanser, diyabet veya karaciğer hastalığı olması durumuna göre değişiklik göstermekte olacağı düşünüldüğünde; hastalığın ölümcül olması veya göz ardı edilebilir bir hastalık olması durumuna göre YN ve YP değerleri dikkate alınmaktadır. Hata matrisinden elde edilen ve performans değerlendirmeleri için bu çalışmada Deeplearning4java kütüphanesi ile kullanılan oranlar doğruluk (accuracy), hassasiyet (precision), kesinlik (recall) ve f-ölçüsü (f-score) oranlarıdır.

· **Doğruluk (Accuracy) :** Doğru sınıflandırılan kişi (hasta veya hasta olmayan) sayısının hasta ve hasta olmayan bütün kişilerin sayısına oranıdır.

$$Doğruluk(Accuracy) = \frac{DP + DN}{P + N} = \frac{DP + DN}{DP + YP + DN + YN}$$

· **Hassasiyet (Precision) :** Doğru sınıflandırılan hasta sayısının, hasta olan bütün kişilerin sayısına oranıdır. Aynı zamanda doğru pozitif oranı olarak adlandırılır.

$$Hassasiyet = \frac{DP}{DP + YN}$$

$$DoğruPozitifOranı = \frac{DP}{DP + YN}$$

· **Kesinlik (Recall) :** Doğru sınıflandırılan hasta sayısının, hasta olarak sınıflandırılan bütün kişilerin sayısına oranıdır.

$$Kesinlik = \frac{DP}{DP + YP}$$

· **F-Ölçüsü (F-Score) :** Kesinlik ve hassasiyet değerinin harmonik ortalamasıdır.

$$F - \text{Ölçüsü} = \frac{(\beta^2 + 1) * Kesinlik * Hassasiyet}{\beta * Kesinlik + Hassasiyet}$$

$\beta = 1 \quad iken$

$$F1 - \text{Ölçüsü} = 2\frac{Kesinlik * Hassasiyet}{Kesinlik + Hassasiyet}$$

$$F1 - \text{Ölçüsü} = \frac{2 * DP}{2DP + YP + YN}$$

Göğüs kanseri (WBCD) verisi için derin öğrenme modeli için çalıştırılması planlanan değişkenler çizelge 4.6 da gösterilmiştir.

Çizelge 4.6 Göğüs kanseri verisi için testlerde kullanılan değişkenler

Aktivasyon Fonksiyonu	TANH
Öğrenme Oranı	0.01 , 0.03, 0.05, 0.1, 0.3
Ağırlık	XAVIER
Kayıp Fonksiyonu	COSINE_PROXIMITY
Çıkış Aktivasyon Fonksiyonu	TANH

Göğüs kanseri (WBCD) verisi için çizelge 4.6'da belirtilen değişkenler ve iterasyon sayısı 1000 olacak şekilde derin öğrenme modeli oluşturulmuştur (Çizelge 4.7). Belirtilen sonuçlar yukarıdan aşağıya doğruluk (accuracy), hassasiyet (precision), kesinlik (recall) ve f-ölçüsü (f-score) değerlerini göstermektedir.

Çizelge 4.7 Göğüs kanseri verisi için iterasyon sayısı 1000 iken değişik öğrenme oranları ve katmanlara göre derin öğrenme test sonuçları

	Öğrenme Oranları				
Doğruluk **Hassasiyet** **Kesinlik** **F1_Ölçüsü**	0.01	0.03	0.05	0.1	0.3
2 katman(3,2)	0.972 0.962 0.967 0.949	0.975 0.966 0.972 0.955	0.978 0.97 0.974 0.959	0.977 0.97 0.972 0.958	0.975 0.968 0.969 0.954
2 katman (8,4)	0.974 0.965 0.97 0.953	0.975 0.968 0.969 0.954	0.978 0.971 0.973 0.96	0.977 0.971 0.972 0.958	0.975 0.97 0.965 0.952
3 katman(8,4,4)	0.975 0.967 0.97 0.954	0.976 0.969 0.971 0.957	0.977 0.969 0.972 0.957	0.978 0.972 0.973 0.959	0.975 0.969 0.966 0.953
4 katman(8,4,5,4)			0.977 0.97 0.971 0.957	0.976 0.971 0.97 0.957	0.974 0.967 0.966 0.951
6 katman(8,4,5,5,5,4)					0.97 0.965 0.959 0.944

Çizelgede belirtilen sonuçlar yukarıdan aşağıya doğruluk (accuracy), hassasiyet (precision), kesinlik (recall) ve f-ölçüsü (f-score) değerlerini göstermektedir.

Çizelge 4.8 Göğüs kanseri verisi için iterasyon sayısı 100 iken değişik öğrenme oranları ve katmanlara göre derin öğrenme test sonuçları

	Öğrenme Oranları				
Doğruluk **Hassasiyet** **Kesinlik** **F1_Ölçüsü**	0.01	0.03	0.35	0.1	0.3
2 katman (3,2)	0.927 0.923 0.888 0.848	0.963 0.957 0.952 0.932	0.963 0.951 0.956 0.933	0.974 0.967 0.968 0.952	0.977 0.968 0.973 0.958
2 katman(8,4)	0.948 0.935 0.933 0.901	0.968 0.956 0.964 0.942	0.975 0.967 0.970 0.954	0.973 0.966 0.966 0.950	0.975 0.968 0.970 0.955
3 katman (8,4,4)	0.950 0.942 0.932 0.906	0.969 0.960 0.963 0.944	0.967 0.957 0.960 0.940	0.975 0.966 0.972 0.955	0.976 0.969 0.971 0.956
4 katman (8,4,5,4)	0.965 0.953 0.957 0.933	0.971 0.960 0.966 0.946	0.974 0.964 0.970 0.952	0.973 0.963 0.968 0.950	0.977 0.971 0.972 0.959
6 katman (8,4,5,5,5,4)					0.977 0.970 0.972 0.958

Ayrıca kütüphanede mevcut olan fonksiyonlar arasından aşağıda çizelge 4.9 ile belirtilen aktivasyon fonksiyonları, öğrenme oranları, ağırlıklar, kayıp fonksiyonları ve çıkış aktivasyon fonksiyonlarının tüm kombinasyonu için çalışmalar yapılmıştır. Çıkan sonuçlar arasından en yüksek doğruluğu (accuracy) olan test sonuçları belirlenmiş ve bunlar üzerinde değerlendirmeler yapılmıştır. Testler yapılırken eksik verisi olan veri kümeleri için verileri yok saymak ve eksik veriler için rastgele değerler atamak şeklinde 2 farklı işlem yapılmıştır. Bunlardan Göğüs kanseri verisinde her iki yöntem kullanılarak, Mamografig kitle verileri yok sayarak işlemler yapılmıştır.

Deeplearning4java kütüphanesi karar kümesini 0,1,2 gibi değerleri desteklediği için sınıflama özelliği olan değerler karar sayısına göre 0,1,2... şeklinde değerlere eşitlenerek çalışmalar yapılmıştır. Göğüs kanseri için 2 (iyi huylu), 4 (kötü huylu) değerleri sırasıyla 0 (iyi huylu), 1 (kötü huylu) şeklinde güncellenerek veri kümesi kullanılmıştır. Benzer şekilde bupa karaciğer 1,2 kararları sırasıyla 0,1 şeklinde güncellenmiştir. Diyabet ve Mamografig kitle verisindeki kararlar zaten 0,1,2 sırasında olduğundan o veri kümeleri için bir işlem yapılmamıştır.

Çizelge 4.9 İterasyon sayısı 1000 iken tüm veri kümelerine uygulanan derin öğrenmede kullanılan değişkenler

Aktivasyon Fonksiyonu	TANH, RELU, SIGMOID
Öğrenme Oranı	0.01, 0.03, 0.05, 0.1, 0.3
Ağırlık	XAVIER, RELU
Kayıp Fonksiyonu	COSINE_PROXIMITY, KL_DIVERGENCE, MSE, NEGATIVELOGLIKELIHOOD
Çıkış Aktivasyon Fonksiyonu	TANH, RELU, SIGMOID

Göğüs kanseri verisi için Çizelge 4.9'da belirtilen değişkenlerle testler yapılmıştır.

3 katmanlı model için toplam çalışma süresi yaklaşık 13 saat 30 dakika (1000 iterasyon), 1 saat 30 dakika (100 iterasyon), 4 katmanlı modelde ise 15 saat 20 dakikadır. Testlerde KL_DIVERGENCE ile yapılan çalışmalarda başarılı sonuçlar alınmadığı (Doğruluk >= 0.98) için çizelge 4.10'da yer almamaktadır.

Göğüs kanseri verisinde 3 katmanlı (8,4,4 düğümü olan) bir model uygulandığında Doğruluk >= 0.98 olan sonuçların kullanılan değişkenlere göre dağılımları çizelge 4.10'da gösterilmiştir.

Çizelge 4.10 Göğüs kanseri verisinde 3 katmanlı (8,4,4 düğümü olan) bir model uygulandığında Doğruluk >= 0.98 olan sonuçların kullanılan değişkenlere göre dağılımları

		iteration:1000 (13 saat 30 dak)		**itaration:100 (1 saat 30 dak)**	
Aktivasyon Fonksiyonu	TANH	45	72	26	28
	SIGMOID	3		0	
	RELU	24		2	
Ağırlık	XAVIER	35	72	13	28
	RELU	37		15	
Öğrenme Oranı	0.01	8	72	0	28
	0.03	14		2	
	0.05	14		5	
	0.1	19		12	
	0.3	17		10	
Kayıp Fonksiyonu	NEGATIVE LOGLIKELIHOOD	1	72	1	28
	COSINE_PROXIMITY	32		14	
	MSE	39		15	
Çıkış Aktivasyon Fonksiyonu	TANH	34	72	14	28
	SIGMOID	34		12	
	RELU	4		2	

Bu sonuçlar (Çizelge 4.10) arasından en yüksek değeri alan değişkenler Doğruluk=0.989 değeri ile çizelge 4.11'de gösterilmiştir.

Çizelge 4.11 En yüksek Doğruluk = 0.989 değeri alan modelde kullanılan değişkenler

Aktivasyon Fonksiyonu	TANH
Öğrenme Oranı	0.1
Ağırlık	XAVIER
Kayıp Fonksiyonu	MSE
Çıkış Aktivasyon Fonksiyonu	RELU

Göğüs kanseri verisinde 4 katmanlı (8,4,5,4 düğümü olan) bir model uygulandığında Doğruluk >= 0.98 değerleri alma dağılımları çizelge 4.12'de gösterilmiştir.

Çizelge 4.12 Göğüs kanseri verisinde 4 katmanlı (8,4,5,4 düğümü olan) bir model uygulandığında Doğruluk >= 0.98 olan sonuçların kullanılan değişkenlere göre dağılımları

		iteration:1000 15 saat 20 dak.	
Aktivasyon Fonksiyonu	TANH	45	62
	SIGMOID	1	
	RELU	16	
Ağırlık	XAVIER	28	62
	RELU	34	
	SIGMOID	31	
	RELU	5	

Çizelge 4.12 Göğüs kanseri verisinde 4 katmanlı (8,4,5,4 düğümü olan) bir model uygulandığında Doğruluk >= 0.98 olan sonuçların kullanılan değişkenlere göre dağılımları (devam)

Öğrenme Oranı	0.01	8	62
	0.03	13	
	0.05	11	
	0.1	14	
	0.3	16	
Kayıp Fonksiyonu	NEGATIVELOG LIKELIHOOD	1	62
	COSINE_PROXI MITY	27	
	MSE	35	
Çıkış Aktivasyon Fonksiyonu	TANH	26	62

Pima Yerlileri Diabet hastalığı verisinde 3 katmanlı (8,4,4 düğümü olan) bir model uygulandığında Accuracy>=0.75 değerleri olan sonuçların kullanılan değişkenlere göre dağılımları Çizelge 4.13'te gösterilmiştir. Model 3*5*2*4*3=360 farklı değişken ile çalıştırılmıştır. Burada 3 katmanlı olan modelde 360 farklı değişkenden 1000 iterasyon için 74'sinde, 100 iterasyon için 19'unda doğruluk 0.75 üzeri sonuç alınmıştır. Test sonuçlarının başarısının Göğüs kanseri verisinde olduğu gibi 1000 iterasyon ile çalışan modelin 100 iterasyon ile çalışan modele göre daha iyi olduğu gözlemlenmiştir. Modeller tutarlılık sağlanması açısından 20 defa çalıştırılmış ve ortalaması alınarak sonuçlar yazılmıştır.

3 katmanlı model için toplam çalışma süresi yaklaşık 14 saat 15 dakikadır. (1000 iterasyon), 1 saat 33 dakika (100 iterasyon) dır.

Çizelge 4.13'te, 1000 iterasyonlu test sonuçları üzerinde daha ayrıntılı yorum yapabilmek için, çıkış aktivasyon fonksiyonuna (TAN+RELU+SIGMOID) göre sonuçlar parçalanmıştır. Burada çıkış aktivasyon fonksiyonu TAN en fazla yüksek sonuç almış olup bunlar arasında da aktivasyon fonksiyonu, ağırlık, öğrenme oranı, kayıp fonksiyonları arasında değerlendirmeler yapılabilmesi için sonuçlar TAN+RELU+SIGMOID (örneğin 20+15+16) olacak şekilde yazılmıştır.

Örneğin anlamı; 1000 iterasyonlu modelde, 320 sonuç alınmış, bu sonuçlardan 74'ü doğruluk 0.75 üzeri çıkmıştır. Aktivasyon fonsiyonu TANH olan sonuçlardan 20'sinin çıkış fonksiyonu TANH, 15'inin çıkış fonksiyonu RELU, 16'sının çıkış fonksiyonu ise SIGMOID'tir.

Göğüs kanseri verisinde olduğu gibi bundan sonra yapılacak testlerde (Diyabet, Bupa Karaciğer hastalığı, Mamografig verilerinde) KL_DIVERGENCE ile yapılan çalışmalarda başarılı sonuçlar alınmadığı (Doğruluk >= istenen değer) için çizelge 4.11, 4.12, 4.14, 4.16, 4.18'de yer almamaktadır. Ayrıca her bir çalışmada testler tutarlılık sağlanması için 20 defa çalıştırılarak ortalamaları alınmıştır. Buradan görüldüğü üzere testlerin hepsi için 1000 iterasyon 100 iterasyona göre iyi sonuçlar vermiştir. Kayıp fonksiyonları arasında en kötü sonuç, belirlenen değerin altında kalıp çizelgeye yazmaya gerek görülmeyecek kadar kötü sonuç alan KL_DIVERGENCE fonksiyonuna, sonrasında çok düşük değerler alan NEGATIVELOGLIKELIHOOD fonksiyonuna aittir.

Çizelge 4.13 Pima yerlileri diyabet hastalığı verisinde (8,4,4 düğümü olan) bir model uygulandığında Doğruluk >= 0.75 olan sonuçların kullanılan değişkenlere göre dağılımları

		iteration:1000 (14 saat 15 dak.)		**itaration:100 (1 saat 33 dak.)**	
		(çıkış aktivasyon fonksiyonuna göre sonuçlar parçalanmıştır) TAN+RELU+ SIGMOID	74		
Aktivasyon Fonksiyonu	TANH	20+15+16		18	19
	SIGMOID	2+0+0		0	
	RELU	11+0+10		1	
Ağırlık	XAVIER	16+9+13	74	12	19
	RELU	17+6+13		7	
Öğrenme Oranı	0.01	4+2+0	74	0	19
	0.03	5+4+4		0	
	0.05	7+3+6		1	
	0.1	8+3+8		6	
	0.3	9+3+8		12	

Çizelge 4.13 Pima yerlileri diyabet hastalığı verisinde (8,4,4 düğümü olan) bir model uygulandığında Doğruluk >= 0.75 olan sonuçların kullanılan değişkenlere göre dağılımları (devam)

Kayıp Fonksiyonu	NEGATIVE LOGLIKELIHOOD	0+0	74	0	19
	COSINE_PROXIMIT	16+5+14		8	
	MSE	17+10+12		11	
Çıkış Aktivasyon Fonksiyonu	TANH	33	74	9	19
	SIGMOID	26		4	
	RELU	15		6	

Çizelge 4.10 ve çizelge 4.13'teki sonuçlardan da anlaşılacağı üzere 3 katmanlı (8,4,4 düğümü olan) ağ modelinin 4 katmanlı ağ modeline göre daha iyi sonuçlar verdiği tespit edilmiştir. 3 ve 4 katmanlı model Çizelge 4.9'da belirtilen değişkenlerin olduğu 3*5*2*4*3=360 farklı kombinasyon ile çalıştırılmıştır. Doğruluk değeri 0.98'den büyük olan sonuçlar belirlenmiş ve bunların hangi değişkene göre olduğu çizelge 4.10 ve çizelge 4.13'te belirtilmiştir. Burada 1000 iterasyon için; 3 katmanlı olan modelde 360 farklı değişkenden 72'sinde doğruluk 0.98 üzeri sonuç alınırken, 4 katmanlı olan modelde 62'sinde doğruluk 0.98 üzeri sonuç alınmıştır. Ayrıca Çizelge 4.10 incelendiğinde 360 farklı sonuçtan doğruluk 0.98'den büyük olan sonuçların 1000 iterasyon için 72 iken, 100 iterasyon için 28 olduğu görülmektedir. Sonuç olarak test sonuçlarının başarısının 1000 iterasyon ile çalışan modelin 100 iterasyon ile çalışan modele göre, 3 katmanlı ağ modelinin 4 katmanlı ağ modeline göre daha daha başarılı olduğu gözlemlenmiştir. Modeller tutarlılık sağlanması açısından 20 defa çalıştırılmış ve ortalaması alınarak sonuçlar yazılmıştır.

Çizelge 4.13'teki sonuçlar arasından en yüksek değeri alan dört değişken çizelge 4.14'te sunulmuştur.

Çizelge 4.14 Pima yerlileri diyabet kitle verisinde en yüksek dört doğruluk değeri alan modelde kullanılan değişkenler

Aktivasyon Fonksiyonu	TANH	TANH	TANH	TANH
Öğrenme Oranı	0.05	0.05	0.03	0.1
Ağırlık	XAVIER	RELU	XAVIER	XAVIER
Kayıp Fonksiyonu	MSE	MSE	COSINE_PROXIMITY	MSE
Çıkış Aktivasyon Fonksiyonu	TANH	TANH	TANH	RELU
Doğruluk **Hassasiyet** **Kesinlik** **F1_Ölçüsü**	0.784 0.768 0.735 0.647	0.785 0.770 0.734 0.645	0.792 0.783 0.740 0.653	0.796 0.785 0.745 0.661

Bupa karaciğer hastalığı verisinde 3 katmanlı (8,4,4 düğümü olan) bir modelde uygulandığında Doğruluk >= 0.7 değerleri alma dağılımları çizelge 4.15'da gösterilmiştir.

Çizelge 4.15 Bupa karaciğer hastalığı verisinde (8,4,4 düğümü olan) bir model uygulandığında Doğruluk >= 0.7 olan sonuçların kullanılan değişkenlere göre dağılımları

		iteration:1000 (13 saat 20 dak)		**itaration:100** (1 saat 20 dak)	
Aktivasyon Fonksiyonu	TANH	34	40	3	3
	SIGMOID	0		0	
	RELU	6		0	
Ağırlık	XAVIER	22	40	3	3
	RELU	18		0	
Öğrenme Oranı	0.01	0	40	0	3
	0.03	5		0	
	0.05	9		0	
	0.1	12		0	
	0.3	14		3	
Kayıp Fonksiyonu	NEGATIVE LOGLIKELIHOOD	0	40	0	3
	COSINE_PROXIMITY	17		2	
	MSE	23		1	
Çıkış Aktivasyon Fonksiyonu	TANH	18	40	1	3
	SIGMOID	13		0	
	RELU	9		2	

Çizelge 4.15'teki sonuçlar arasından en yüksek değeri alan diğer değişkenler performans değerleri ile Çizelge 4.16'de gösterilmiştir.

Çizelge 4.16 Bupa karaciğer verisinde en yüksek dört doğruluk değeri alan modelde kullanılan değişkenler

Aktivasyon Fonksiyonu	TANH	TANH	TANH	TANH
Öğrenme Oranı	0.05	0.05	0.05	0.1
Ağırlık	XAVIER	RELU	XAVIER	XAVIER
Kayıp Fonksiyonu	MSE	MSE	MSE	MSE
Çıkış Aktivasyon Fonksiyonu	TANH	TANH	RELU	TANH
Doğruluk **Hassasiyet** **Kesinlik** **F1_Ölçüsü**	0.726 0.712 0.695 0.789	0.726 0.712 0.697 0.787	0.728 0.713 0.705 0.785	0.728 0.714 0.701 0.788

Mamografig kitle hastalığı verisinde 3 katmanlı (8,4,4 düğümü olan) bir modelde uygulandığında Doğruluk >= 0.75 değerleri alma dağılımları çizelge 4.17'de gösterilmiştir.

Çizelge 4.17 Mamografig kitle hastalığı verisinde (8,4,4 düğümü olan) bir model uygulandığında Doğruluk >= 0.75 olan sonuçların kullanılan değişkenlere göre dağılımları

		iteration:1000 (10 saat)		**itaration:100 (1 saat 30 dak)**	
Aktivasyon Fonksiyonu	TANH	20	86	32	39
	SIGMOID	2		0	
	RELU	17		7	
Ağırlık	XAVIER	19+17+6	86	19	39
	RELU	20+18+6		20	
Öğrenme Oranı	0.01	7+4+1	86	0	39
	0.03	8+7+3		5	
	0.05	8+8+3		6	
	0.1	7+8+2		11	
	0.3	9+8+3		17	
Kayıp Fonksiyonu	NEGATIVE LOGLIKELIHOOD	0+0+0	86	0	39
	COSINE_PROXIMITY	18+18+2		19	
	MSE	21+17+10		20	
Çıkış Aktivasyon Fonksiyonu	TANH	39	86	19	39
	SIGMOID	35		15	
	RELU	12		5	

Çizelge 4.17'deki sonuçlar arasından en yüksek değeri alan diğer değişkenler performans değerleri ile Çizelge 4.18'de gösterilmiştir.

Çizelge 4.18 Mamografig kitle verisinde en yüksek dört doğruluk değeri alan modelde kullanılan değişkenler

Aktivasyon Fonksiyonu	TANH	TANH	TANH	TANH
Öğrenme Oranı	0.3	0.3	0.3	0.3
Ağırlık	XAVIER	XAVIER	XAVIER	XAVIER
Kayıp Fonksiyonu	MSE	MSE	COSINE_PROXIMITY	COSINE_PROXIMITY
Çıkış Aktivasyon Fonksiyonu	RELU	SIGMOID	SIGMOID	TANH
Doğruluk **Hassasiyet** **Kesinlik** **F1_Score**	0.805 0.815 0.801 0.778	0.806 0.818 0.803 0.777	0.807 0.817 0.803 0.78	0.809 0.818 0.806 0.783

Çizelge 4.19 KNN algoritmasının veri kümelerinde karşılaştırılması

	k=1	**k=3**	**k=5**	**k=10**
Göğüs kanseri	0.970	0.975	0.970	0.955
Pima Yerlileri diyabet	0.647	0.656	0.660	0.673
Bupa karaciğer hastalıkları	0.466	0.533	0.563	0.553
Mamografik kitle	0.420	0.375	0.375	0.541

5. SONUÇ

Literatürde şu ana kadar yapılan çalışmalar ilgili incelemeler yapılmış, perceptron öğrenme algoritması ve KNN kullanılarak sağlık verilerinde sınıflandırma çalışması yapılmıştır. Ayrıca derin öğrenme üzerine bir model kurulmuş, bu model çeşitli değişkenlerle çalıştırılarak sınıflama yapılmıştır. Kullanılan veri kümesi UCI Makine Öğrenme Deposunda bulunan göğüs kanseri, pima yerlileri diyabet veri tabanı, bupa karaciğer hastalıkları, mamografik kitle verileridir. Oluşturulan derin öğrenme modelinde performans değerleri en iyi çıkan sonuçlar ele alınarak, her veri kümesi için en iyi sonuçları veren değişkenlerin değerlendirilmesi yapılmıştır. Çizelge 5.1'de kullanılan değişkenler ve sonuçları gösterilmiştir. Performans değerlendirmesi doğruluk, hassasiyet, kesinlik ve f-ölçüsü kullanılarak yapılmıştır. Sonuçların tutarlı olması açısından her veri kümesi için model 20 kere çalıştırılarak sonuçların ortalaması alınmıştır.

Çizelge 5.1 Oluşturulan derin öğrenme modelinde en yüksek sonuçları alan parametreler ve sonuçları

	Aktivasyon Fonksiyonu **Öğrenme Oranı Ağırlık** **Kayıp Fonksiyonu** **Çıkış Aktivasyon Fonk.**	**Doğruluk** **Hassasiyet** **Kesinlik** **F1_Score**
Göğüs kanseri	TANH 0.1 XAVIER MSE RELU	0.989 0.982 0.985 0.975
Pima Yerlileri diyabet	TANH 0.1 XAVIER MSE RELU	0.796 0.785 0.745 0.661
Bupa karaciğer hastalıkları	TANH 0.1 XAVIER MSE TANH	0.728 0.714 0.701 0.788
Mamografik kitle	TANH 0.3 XAVIER COSINE_PROXIMITY TANH	0.809 0.818 0.806 0.783

KNN algoritmasında k değeri 1,3,5 ve 10 değerleri kullanılarak performans değerlendirmesi yapılmış ve bu sonuçlar; oluşturulan derin öğrenme modelindeki en iyi sonuçlar (Çizelge 5.1) ile karşılaştırılarak çizelge 5.2'de gösterilmiştir. Çizelgede görüldüğü gibi, oluşturulan derin öğrenme modeli ile elde edilen sonuçlar KNN'ye göre daha iyidir.

Çizelge 5.2 KNN ve oluşturulan derin öğrenme modelinin doğruluklarının karşılaştırılması

Doğruluk	**KNN k=1**	**KNN k=3**	**KNN k=5**	**KNN k=10**	**Oluşturulan Derin Öğrenme Modeli**
Göğüs kanseri	0.970	0.975	0.970	0.955	0.989
Pima Yerlileri diyabet	0.647	0.656	0.660	0.673	0.796
Bupa karaciğer hastalıkları	0.466	0.533	0.563	0.553	0.728
Mamografik kitle	0.420	0.375	0.375	0.541	0.809

Önerilen algoritmanın 2 ve 3 özellikli versiyonu belirtilen dört veri kümesinde çalıştırılarak doğruluk oranları bulunmuştur. Bu sonuçlar; çizelge 5.2'te gösterilen KNN ile elde edilen sonuçlardan en iyi olan k=5 ile elde edilen sonuçlar ve çizelge 5.1'de gösterilen oluşturulan derin öğrenme modeli sonuçları ile karşılaştırılarak çizelge 5.3'te gösterilmiştir. Çizelge 5.3'te görüldüğü gibi önerilen algoritma 3 özellik ile çalıştırıldığında 2 özellikle çalıştırdığında elde edilen sonuçlara göre daha iyi doğruluk değerleri elde edilmiş ancak çalışma süreleri dikkate alınacak ölçüde uzamıştır. Doğruluk değerlerinin tutarlı olması için 2 özellikli versiyonu için 20 kere çalıştırılarak ortalamaları alınmıştır. 3 özellikli versiyonu için çalışma sürelerinin uzunluğu dikkate alınarak, tutarlılığın sağlanması için pima yerlileri diyabet verisinde ancak 10 kere,

göğüs kanseri verisinde ancak 1 kere, mamografik kitle verisinde ancak 1 kere, bupa karaciğer hastalıkları verisinde ise 20 kere çalıştırma yapılabilmiştir. Bu sonuçların ortalama değerleri bulunmuş ve doğruluk değerleri belirlenmiştir. Doğruluklar karşılaştırıldığında; göğüs kanseri, pima yerlileri diyabet ve bupa karaciğer hastalıkları verisinde önerilen algoritmanın 3 özellik ile çalıştırılması ile elde edilen sonuçlar oluşturulan derin öğrenme modelindeki doğruluklara yaklaşmıştır. Mamografig kitle verisinde ise önerilen algoritmanın 3 özellikle ile çalıştırılmasından elde edilen doğruluk, oluşturulan derin öğrenme modelindeki doğruluktan daha iyi çıkmıştır. Çalışma süreleri olarak değerlendirildiğinde önerilen algoritmanın 2 özellikle çalıştırılmasındaki süre oluşturulan derin öğrenme modelindekinden az olurken, önerilen algoritmanın 3 özellikle çalıştırılmasından çok olmuştur. Önerilen algoritmanın 3 özellik ile çalışması veri kümeleri arasında değerlendirildiğinde ise en uzun süreli çalışan veri mamografik kitle verisi olmuştur.

Çizelge 5.3 KNN, oluşturulan derin öğrenme modeli ve önerilen algoritmanın doğruluklarının karşılaştırılması

Doğruluk	**KNN k=5**	**Derin Öğrenme**	**Önerilen Algoritma (2 Özellik)**	**Önerilen Algoritma (3 Özellik)**
Göğüs kanseri	0.970	0.989 (13 saat 30 dak)	0.943 (20 kere çalıştırılmıştır) (40 dak)	0.962 (1 kere çalıştırılmıştır) (25 saat 40 dak)
Pima Yerlileri diyabet	0.660	0.796 (14 saat 15 dak.)	0.715 (20 kere çalıştırılmıştır) (50 dak)	0.743 (10 kere çalıştırılmıştır) (81 saat)
Bupa karaciğer hastalıkları	0.563	0.728 (13 saat 20 dak)	0.663 (20 kere çalıştırılmıştır) (3 dak)	0.701 (20 kere çalıştırılmıştır) (3 saat)
Mamografik kitle	0.375	0.809 (10 saat)	0.791 (20 kere çalıştırılmıştır) (30 dak)	0.825 (1 kere çalıştırılmıştır) (108 saat 40 dak)

Uzun süre çalışması göz önünde bulundurulduğundan ve mevcuttaki 2 ve 3 özellikle ile çalıştırılan test sonuçlarının doğruluklarının yüksek olmasından dolayı 4 ve daha fazla özellik için testler yapılmamıştır.

KAYNAKLAR

Abdel-Aal, R.E. 2005. GMDH-based feature ranking and selection for improved classification of medical data. Journal of Biomedical Informatics, 38 (2005) 456–4

Anonim, Web Sitesi: https://www.udemy.com/deep-learning-ve-python-adan-zye-derin-ogrenme-5 Erişim Tarihi: 20.02.2019

Bektaş, B., Babur, S. 2016. Makine Öğrenmesi Teknikleri Kullanlarak Meme Kanseri Teşhisinin Performans Değerlendirmesi, TıpTekno'16 Tıp Teknolojileri Kongresi, 27-29 Ekim, Antalya

Chang, P.C., Fan, C.Y., Dzan, W.Y. 2009. A CBR-based fuzzy decision tree approach for database classification. Expert Systems with Applications 2009, 37(2010) 214-225

Fan, C.Y., Chang, P.C., Linb, J.J., Hsiehb, J.C. 2011. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Applied Soft Computing, 11(2011) 632-644

Fuller, R. 1995. Neural Fuzzy Systems. 157-169

Jain, D., Singh, V. 2018. Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal, 19 (2018) 179–189

Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L. 2017. A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics and Informatics, 34(2017) 133-144

Öztemel, E. 2006. Yapay Sinir Ağları. Papatya Yayıncılık, 45-56, İstanbul, Ankara, İzmir, Adana

Revett, K., Gorunescu, F., Gorunescu, M., El-Darzi, E., Ene, E. 2005. A Breast Cancer Diagnosis System: A Combined ApproachUsing Rough Sets and ProbabilisticNeural Networks. Eurocorn 2005, 22-24 November, Belgrade

Santos, V., Datia, N., Pato, M.P.M. 2014. Ensemble feature ranking applied to medical data, Procedia Technology 17 (2014) 223 – 230

Seera, M., Lim, C. 2013. A hybrid intelligent system for medical data classification. Expert Systems with Applications, 41(2014) 2239-2249

Takcı, H. 2016. Centroid Entroid Sınıflayıcılar Yardımıyla Meme Kanseri. Journal of the Faculty of Engineering and Architecture of Gazi University 31 (2) 323-330

Taşdelen, D.A., Amrahov, Ş.E. 2017. Classification Problem of Breast Cancer Patients. International Conference on Theoretical and Applied Computer Science and Engineering (ICTACSE, 2017) Ankara, Turkey, November 10, 2107

ÖZGEÇMİŞ

Adı Soyadı : Didem ATİKTÜRK TAŞDELEN

Doğum Yeri : Ankara

Doğum Tarihi : 20.01.1986

Medeni Hali : Evli

Yabancı Dili : İngilizce

Eğitim Durumu

Lise : Kalaba Anadolu Lisesi (2004)

Lisans : Ankara Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği (2009)

Yüksek Lisans : Ankara Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Bölümü (Eylül 2009–Temmuz 2019)

Çalıştığı Kurumlar

Sistem Geliştirme ve İşletim Kontrol Uzmanı, Milli Savunma Bakanlığı, 2012-Devam ediyor

Veri tabanı Uzmanı, Cybersoft, 2009-2012

Uluslararası Kongre Sunum

Taşdelen, D.A., Amrahov, Ş.E. 2017. Classification Problem of Breast Cancer Patients. International Conference on Theoretical and Applied Computer Science and Engineering (ICTACSE, 2017) Ankara, Turkey, November 10, 2017