TIP BİLİŞİMİNDE VERİ MADENCİLİĞİ YÖNTEMLERİ KULLANILARAK HASTALIKLARIN TAHMİN EDİLMESİ

PREDICTION OF DISEASES USING DATA MINING METHODS IN MEDICAL INFORMATICS

YASEMİN HANDE SITKI

DOÇ. DR. ESRA POLAT

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Egitim-Öğretim ve Sınav Yönetmeliğinin

İstatistik Anabilim Dalı İçin Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2020

Canım Ailem'e

ÖZET

TIP BİLİŞİMİNDE VERİ MADENCİLİĞİ YÖNTEMLERİ KULLANILARAK HASTALIKLARIN TAHMİN EDİLMESİ

Yasemin Hande SITKI

Yüksek Lisans, İSTATİSTİK Bölümü

Tez Danışmanı: Doç. Dr. Esra POLAT

Aralık 2020, 86 sayfa

Günümüzde veri madenciliği büyük boyutlu verilerden saklı kalan bilgiyi ortaya çıkararak, işletmelerin karar verme süreçlerine destek sağlamasından ötürü özellikle sağlık işletmelerinin idaresinde ve sağlıkla ilgili politikaların belirlenmesinde önemi gittikçe artan bir araçtır. Ayrıca son yıllarda sağlık alanında veri madenciliği algoritmaları kullanılarak hastalıkların teşhis edilmesine yönelik bilimsel yayınlar yapılmaktadır.

Bu tez çalışmasında, en sık kullanılan veri madenciliği sınıflandırma yöntemleri incelenmiş, dört farklı devlet hastanesinin üroloji branşına başvuran hastalardan toplanan veriler kullanılarak, üroloji branşında görülen 18 farklı hastalığın teşhisine yönelik bir çalışma yapılmıştır. Bu amaçla sınıflandırma algoritmalarından Rastgele orman (Random Forest), Rasgele ağaç (Random Tree), Çok katmanlı algılayıcılar (Multilayer Perception), K-en yakın komşu (IBk), Kstar, Lojistik regresyon, Naive Bayes ve ZeroR algoritması kullanılmış ve oluşturulan modellerin doğru sınıflandırma oranlarına yani hastalıkları ne kadar doğru teşhis ettiklerine bakılmıştır.

Bu algoritmalardan Rasgele orman, Lojistik regresyon ve Çok katmanlı algılayıcılar algoritmalarının teşhiste diğerlerine göre daha başarılı olduğu görülmüştür. İlerleyen

çalışmalarda üroloji branşına ilişkin hastalıklarının ya da başka branşlarda görülen hastalıkların teşhisine yönelik burada sözü geçen algoritmalar kullanılarak bir uygulama geliştirilebilir. Böylece sağlık çalışanlarına teşhiste fikir verilmesi ve iş yüklerinin azaltılması, erken teşhis ile hastalıkların önceden bulunarak tedavi sürelerinin daha kısa olması mümkün olabilir.

Anahtar Kelimeler: Üroloji, Veri Madenciliği, WEKA, Rastgele Orman Algoritması, Karar Ağacı Algoritması

ABSTRACT

PREDICTION OF DISEASES USING DATA MINING METHODS IN MEDICAL INFORMATICS

Yasemin Hande SITKI

Master of Science, Department of STATISTICS

Supervisor: Doç. Dr. Esra POLAT

December 2020, 86 pages

Nowadays, Data Mining is an increasingly important tool, especially in the administration of healthcare enterprises and in determining health-related policies, as it provides support for the decision-making processes of businesses by revealing information hidden from large dimensional data. Moreover, scientific publications have been made in the field of health in recent years to diagnose diseases using data mining algorithms.

In this thesis, most frequently used data mining classification methods were examined, and a study was conducted to diagnose 18 different diseaes in the urology branch using the data collected from the patients who applied to the urology branch of four different public hospitals. For this purpose, classification algorithms Random Forest, Random Tree, Multilayer Perception, IBk, Kstar one of the sample based algorithms, Simple Logistic and Naive Bayes from statistical algorithms and ZeroR from rule learning algorithms were used, and the correct classification rates of the created models, namely how correctly they diagnosed the diseases, were examined.

Among these algorithms Random Forest, Simple Logistic and Multilayer Perception algorithms have been found to be more successful in diagnosis than others. In future studies, an application can be developed for the diagnosis of diseases related to urology branch or diseases seen other branches by using the algorithms mentioned here. Thus, it may be possible to give healthcare professionals an idea in the diagnosis and to reduce their workload, to find the diseases in advance with early diagnosis and to shorten the treatment period.

Keywords: Urology, Data Mining, WEKA, Random Forest Algorithm, Decision Tree Algorithm

TEŞEKKÜR

Bu tez çalışması süresince bana destek olarak değerli bilgi, birikim ve tecrübeleri ile yol gösteren ve bilgisini en iyi şekilde aktaran danışman hocam Doç. Dr. Esra POLAT başta olmak üzere, lisansüstü eğitimim boyunca bilgi ve tecrübeleri ile bana yol gösteren Hacettepe Üniversitesi İstatistik Bölümü'ndeki tüm hocalarıma, mesai arkadaşlarıma ve dostlarıma teşekkürlerimi sunarım.

Hayatım boyunca her koşulda bana destek olan, sabır gösteren, her zorlukta arkamda duran, bugünlere gelmemde en büyük maddi ve manevi katkıyı sunan canım aileme şimdi ve ömrüm boyunca teşekkürü bir borç bilirim.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

SİMGELER VE KISALTMALAR

Simgeler

Kısaltmalar

DVM	Destek Vektör Makinesi

1. GİRİŞ

Sağlık kuruluşlarının iş akışı, klinik uygulamaları, klinik araştırmaları, hasta bilgileri, kaynak yönetimi, politikalar ve araştırmaları çok fazla veri içerir. Bu veriler, veri madenciliği için benzersiz bir fırsat olmuştur. Geleneksel olarak, verilerden operasyonel bilgi çıkarmak için istatistiksel teknikler kullanılmıştır. Yeni bir yöntem olan veri madenciliği, ilişkilendirmeler, sınıflandırmalar ve tahminler açısından değerli sağlık bilgilerini keşfetmek ve onları kullanmak için bir fırsat sağlar. Sağlık hizmetlerine ilişkin bu tür kodlanmış bilgiler, sağlık hizmetlerinin sunumuna ilişkin stratejik öngörüler sağlayabilir.

Tıp alanı genişledikçe, her hastayı hastalıklardan, yan etkilerden ve tıbbi rahatsızlıklardan korumak her hekimin görevidir. Doktorlar, bilgi tabanlarını genişletmek için bir araç olarak taramalar, steteskoplar ve diğer cerrahi prosedürlerin yanı sıra veri madenciliğini de kullanabilir. Veri madenciliği, sağlık sektörünün birçok yönü için kullanılabilir.

Dolandırıcılığı tespit etmek, ilaç endüstrisinde ilaçların yan etkilerini değerlendirmek ve hatta genetiğe dayalı belirli hastalıkların teşhisi gibi alanlarda kullanılır [1]. Sağlık dünyasının veri madenciliği için yarattığı eşsiz fırsatlar göz önüne alındığında, veri madenciliği hastalıkların oluşumunu tahmin etmek için de kullanılabilir. Böylece doktorların iş yükünün rahatlaması, hata yapma ihtimallerinin düşmesi ve yanlış teşhisten doğabilecek maddi ve manevi zararların önüne geçilmesi gibi hem hasta, hem doktor hem de hastane yönetimi açısından birçok faydası vardır.

Veri madenciliği, sağlık sektörünün sağlık sistemleri için, verimsizlikler ile maliyeti azaltma ve iyileştirilmiş bakım sağlayan en iyi uygulamalarını belirlemek için sistematik olarak veri analizini sağlar. Bazı uzmanlar, bakımı iyileştirme ve maliyetleri düşürmenin toplam sağlık bakımı maliyetlerinin %30'unu oluşturabileceğine inanmaktadır [2]. Bununla birlikte, sağlık hizmetlerinin karmaşıklığı ve teknolojinin benimsenme hızının yavaş olması nedeniyle, veri madenciliği endüstrisi, bu verilerin ve analitik stratejilerin uygulanmasıyla birlikte

etkilidir. Bu konu özellikle Türkiye'de daha az ilgi görmüştür. Hastalıkların ortaya çıkmasını önlemek için veri madenciliğinin kullanılması, ülkenin sağlık sektörünün maliyetlerini düşürmede çok faydalı olabilir.

Veri madenciliği aynı zamanda bilgi keşfi olarak da bilinir. Veri madenciliği, işletmelerin problemlerini çözmesine, yeni fırsatları elde etmesine ve riskleri düşürmesine destekçi iş bilgilerini elde etmek için, büyük boyutlu verilerin analiz edilmesidir. Bu veri bilimi alanı, ismini büyük bir veritabanından önemli olan yani işe yarayan bilgileri aramakla değerli taşları çıkarmak için dağları kazmak arasındaki benzerlikten alır. Bu iki süreç de saklı kalmış değeri bulmak için büyük miktarlarda malzeme üzerinden analiz yapılmasını gerektirmektedir.

Bu tez çalışmasında, ikinci bölümde veri madenciliği kavramından, özellikle sağlık alanında olmak üzere farklı uygulama alanlarından ve popüler veri madenciliği algoritmalarından bahsedilmiştir. Üçüncü bölümde, bu tezde uygulanan sınıflandırma yönteminden, sınıflandırmada algoritmaları karşılaştırırken kullanılan ölçütlerden ve kısaca WEKA programından bahsedilmiştir. Dördüncü bölümde üroloji branşına başvuran hastalardan elde edilen büyük veriden yararlanarak, yaygın kullanılan veri madenciliği algoritmalarından hangisinin ürolojik hastalıkların teşhisinde daha başarılı olduğu WEKA programında yapılan uygulama ile ortaya konmuştur. Karşılaştırma sonucunda hastaların hastalıklarını doğru teşhis etmede öne çıkan sınıflandırma algoritmaları belirlenmiş ve sonuçlar yorumlanmıştır.

2. GENEL BİLGİLER

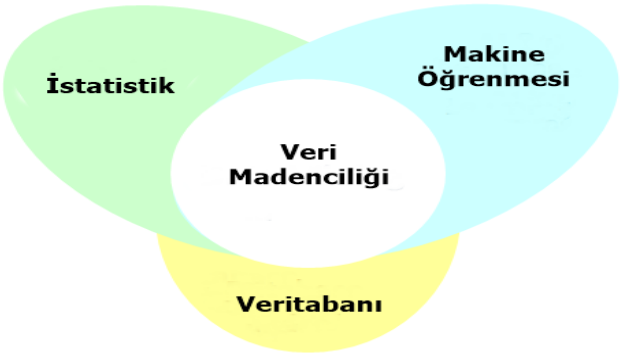
2.1. Veri Madenciliği

Bu bölümde, veri madenciliği kavramından kısaca bahsedilecektir. Bu tez çalışmasında sağlık alanında bir uygulamaya yer verildiğinden dolayı bu sektördeki kullanımı detaylandırılacaktır. Veri madenciliğinde kullanılan algoritmalardan bahsedilecektir. En son olarak ise sağlık sektörü gibi başka sektörlerdeki farklı uygulamalarından da bahsedilecektir.

2.1.1. Veri Madenciliği Tanımı

Veri madenciliği, ham verilerden bilgiyi belirlemeyi ve keşfetmeyi amaçlayan bir araştırma alanıdır. Bu bilgi tanıma ne kadar doğru olursa, yaklaşım o kadar başarılı olur [1]. Günümüzde veri, araştırmacılar, endüstri ve şirketler için çok önemli konulardan biridir. Modern çağda veri üretimi çok daha gelişmiş olduğu için, veri madenciliği de popüler bir bilim haline gelmiştir.

Ham verilerden bilgiyi keşfetmek, veri madenciliğinin ana görevidir. Veri madenciliği adı verilen ham verilerden bilgi keşfetmenin adımları vardır. Bunun için önce veriler hazırlanıp standartlaştırılır, ardından özellik seçim yöntemlerine göre modeli oluşturmak için gerekli olan özellik çıkarılır. Bu adımlardan sonra, sınıflandırma modelini oluşturmak için makine öğrenimi algoritmaları kullanılır. Sonunda en başarılı algoritma belirlenir.



Şekil 2. 1 Veri madenciliğinin başka bilimler ile ilişkisi [2].

“Günümüzde, bilgi sistemleri, büyük miktarlarda veriyi saklayarak, işleyerek ve analiz ederek karar verme süreçlerinde yararlı bilgiler sağladığından birçok kuruluş için vazgeçilmezdir.” [3].

Şekil 2.1’de veri madenciliğinin başka bilimler ile ilişkisi görülmektedir. Veri madenciliğinin birçok bilim ile ilişkisi vardır, bu bilimler veri tabanı, istatistik ve makine öğrenmesi bilimleri olarak bilinir.

Veri madenciliği, dijital çağın gelişiyle öne çıkan yeni bir konu değil, bir asırdan fazla bir süredir araştırmacılar arasında bir kavram haline gelmiştir ve 1930'lardan beri veri madenciliği bir araştırma konusu olarak ilgi çekmiştir. Veri madenciliğine ilişkin ilk örneklerden biri, 1936'da Alan Turing'in modern bilgisayarlara benzeyen hesaplamalar yapabilen evrensel bir makine fikrini tanıtmasıyla ortaya çıkmıştır. 1930'lardan bugüne, veri madenciliği önemli bir uygulama alanı oldu ve bu bilim gittikçe büyüyerek daha fazla ilgi görmüştür[1].

2.1.2. Veri Madenciliği Süreci

Veri madenciliği 4 önemli aşamadan oluşmaktadır: 1- Veri Toplama 2- Veri Temizleme 3- Model Oluşturma 4- Model Kullanma. Her veri madenciliği uygulaması bu aşamalardan oluşmaktadır.

Şekil 2. 2 Veri Madenciliği Aşamaları

1-Veri Toplama

Veri toplama, veri madenciliğinde en önemli adımdır. Adından da belli olmak üzere veri madenciliğinde her şey veriden başlar. Her alanda veri temin etmenin kendi zorlukları vardır. Mesela sağlık alanında veri temin etmek en zor aşamadır; çünkü hastaların verilerinin gizliliği de söz konusudur. Verilerin kullanabilmesi için verilerin tek bir veri tabanında veya veri ambarında birleştirilmesi gerekir. Elde edilen veriler bir sonraki adımda temizlenecek ve yararlı olabilecek veriler çıkartılacaktır [5].

2-Veri Temizleme

Veri temizleme veya veri kontrolü, bozuk veya yanlış verinin temizlenmesi ve çıkartılması aşamasıdır. Bu etapta verilerin üzerine bir ön işleme yapılır. Veri temizleme işlemi ne kadar düzgün yapılırsa kurulacak modelin başarısı da o kadar artar[1]. Bu aşamada, veriler model oluşturmak için uygun hale getirilir.

3-Model Oluşturma

Veri madenciliğinde modeli oluşturmamız için farklı yöntemler hakkında düşünme ve veri için en ideal olanı seçmek gerekmektedir. Bu aşama tek bir işlem gibi görünmesine karşın, detaylı işlemler içerebilir. Bu amaca varmak için çeşitli algoritmalar geliştirilmiştir. Bu algoritmaların geneli, aynı veri setiyle değişik modelleri modellemeye ve en iyisini seçmek amacıyla en iyisini yapmak için uğraşan bir metoda dayanmaktadır. Veri madenciliği algoritmaları uygulandıktan sonra elde edilen sonuçlar yorumlanır ve çalışmanın ne kadar başarılı olup olmadığı

değerlendirilir. Bu adımda farklı algoritmalar uygulanır ve farklı algoritmalardan elde edilen sonuçlar karşılaştırılır. En iyi sonuç ise gelecek adımda kullanmak için seçilir. Eldeki sonuçlar başka çalışmalardaki sonuçlar ile karşılaştırılır. Genellikle, algoritmanın kolay uygulanabilirliği ve doğruluğu, çalışmada kullanılacak yöntemi belirlemede temeli oluşturur.

Bu çalışmada da üroloji hastalıklarını tahmin etmek için bu aşamalar uygulanacak ve farklı veri madenciliği algoritmaları kullanılarak en başarılı sonucu veren algoritma belirlenecektir.

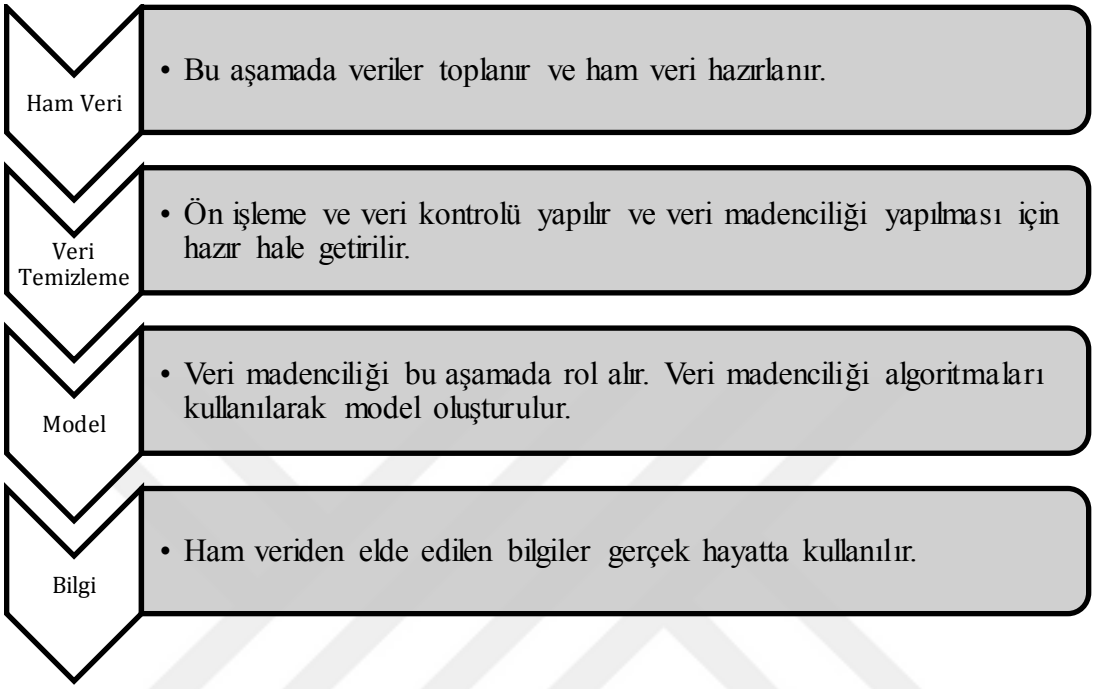
4-Model Kullanma

Veri madenciliği çalışmalarının nihai hedefi kullanılabilir bir model oluşturmaktır. Veriler toplanır, temizlenir ve daha sonra farklı algoritmalar kullanılarak en başarılı algoritma belirlenir. Bunlar yapıldıktan sonra yeni veriler üzerine veri madenciliği yapmak için elde edilen en iyi model kullanılır [1].

2.2. Sağlıkta Veri Madenciliği

Tıbbi veri; hasta kayıtları gibi sağlık hizmeti bilgilerini depolayan veri tabanları anlamına gelmektedir. Bilişim Teknolojisi'nin gelişmesi ile paralel olarak, bu tür tıbbi verilerin birçoğu elektronik formlarda saklanmaya başlamıştır. Söz konusu veri tabanları büyük miktarda veri içerir. Tıbbi veriler röntgen, manyetik rezonans görüntüleme (MRI), bilgisayarlı tomografi (CT), ultrason gibi farklı kaynaklardan elde edilebilir. Böylelikle veri hacmindeki büyüme ve sayısallaştırılmış verinin depolanması için gerekli veri tabanlarındaki artışı katlanarak devam etmiştir [1]. Öte yandan, ham tıbbi veriler; görüntülemeler, hasta ile görüşmeler, laboratuvar verileri ve hekimin gözlemleri ve değerlendirmeleri gibi farklı kaynaklardan toplanabilir [2]. Tıbbi verinin çeşitli türleri mevcuttur. Bunlar; görüntüler, veri kümeleri, sinyaller, dalga boyları vb. olabilir. Mevcut senaryoda, bilgi toplama araçları alanındaki araştırmaların ve geliştirmelerin bir sonucu olarak, elektronik formatta erişilebilir olan büyük miktarda bilgi veya veriye tanıklık edebiliriz. Bu derece büyük miktarlarda veri ve bilginin depolanabilmesi için veri tabanı boyutlarının da önemli ölçüde arttığı açıktır [3].

Aslında veri madenciliği, bilgi keşfi olarak bilinen daha geniş bir olasılığın temel bir sürecidir. Veri madenciliği ve bilgi keşfi arasındaki karşılıklı ilişki, Şekil 2.3'de gösterilmektedir.



Şekil 2. 3 Bilgi keşfi sürecinde veri madenciliğinin rolü

Tıbbi veriler, yalnızca yeni veri tabanı teknolojileri ve internet yoluyla var olabilen yüzlerce kamusal ve özel veri tabanında erişilebilir vaziyettedir. Sağlık hizmetleri sektörünün her yıl terabaytlarca veri üretebileceği tahmin edilmektedir. Aslında, kaliteli sağlık hizmeti için faydalı bilgiler elde etmek zor ve önemli bir iş olup, bu amaçla günümüz veri tabanlarında çok büyük miktarlarda veri mevcuttur. Bununla beraber, buradan ayıklanan/çıkarılan bilgi, neredeyse yok denecek kadar azdır. Dolayısıyla, somut bilgi elde etmenin mümkün olabilmesi için, verilerin etkin organizasyonu, analizi ve yorumlanması son derece büyük öneme sahiptir. Gerçekte, bu büyük tıbbi veritabanlarının yönetimi için elverişli örüntülerin ve bu örüntülerden gizli bilgilerin keşfedilmesi adına, farklı hesaplama tekniklerine ihtiyaç duyulmaktadır[4]. Veri madenciliği sürecinde, genellikle, muazzam büyüklükte gözlemsel veri kümeleri analiz edilir ve sonrasında veri sınıflandırması amacıyla işe yarar gizli örüntüler ayıklanır. Günümüzde veri madenciliği tıbbi veriler ile de buluşmuştur. Bunun nedeni; hastalar, hastaların tıbbi durumları ve tedavileri

arasındaki karmaşık ilişkinin anlaşılır biçimde analiz edilmesi için [7], tıbbi verilerden bilinmeyen, değerli ve gizli bilgilerin tespitine yönelik etkili tekniklere ihtiyaç duyulmasıdır[6].

2.2.1.Veri Madenciliği Görevleri ve Sağlık Hizmetleri Alanında Kullanımı

Bir uygulama alanından diğerine değişiklik gösteren çeşitli veri madenciliği modelleri mevcuttur. Ancak genel anlamda, öngörücü model ve betimleyici model olmak üzere iki grupta kategorize edilmektedir. Şekil 2.4'te, tıp ve sağlık hizmetleri alanına ilişkin bir takım önemli veri madenciliği görevlerine yer verilmektedir.

Şekil 2. 4 Veri Madenciliği Görevleri

(i) Özetleme: Özetlemede, veri seti soyutlanır ve bu da, verilerin genel bir incelemesini sunan daha küçük bir veri setiyle sonuçlanır. Dolayısıyla özetleme, verilerin soyutlanması veya genelleştirilmesi olarak da ifade edilebilir. Özetleme, birçok soyutlama düzeyine kadar gerçekleştirilebilir ve farklı bakış açılarından ele alınabilir. Örneğin; çağrının detaylarına bakmak yerine arama süresi, arama sayısı ve arama sırasında ortaya çıkan tutar özetlenebilir. Aynı şekilde, aramalar da ulusal aramalar veya uluslararası aramalar temelinde özetlenebilir. Farklı soyutlama

düzeylerinin bu kombinasyonları, verilerde mevcut olan çeşitli örüntüler ve düzenlilikler hakkında bilgi vermektedir [8].

(ii) İlişkilendirme: İlişkilendirme, büyük veritabanlarında nesnelerin bir aradalığını veya bağlantısını arar. Bu tür bir bağlantı, ilişkilendirme kuralı olarak bilinir. Bir ilişkilendirme, nesneler arasında mevcut olan ilişkileri açığa çıkarır. İlişkilendirmenin temel amacı, nesneler arasında var olan ilişkileri, yani, başka bir nesnede bir nesne kümesinin varlığını bulmaktır [9]. İlişkilendirme kuralları genellikle pazarlama, emtia yönetimi, reklamcılık gibi alanlarda kullanılır. Bu ilişkilendirme kurallarından, çeşitli nitelikler arasında var olan ilişkiler ve örüntüler çıkarılır. Birlikteliğe dayalı veri madenciliği, nitelikler arasındaki ilişkileri tespit etmeyi ve ardından bu veri kümelerinden kurallar oluşturmayı amaçlamaktadır [10]. Örneğin; "çağrı bekletme" nin "çağrı görüntüleme" ile ilişkili olduğu bir ilişkilendirme kuralı, "çağrı bekletme" hizmetine abone olması durumunda müşterinin "çağrı görüntüleme" hizmetine de abone olma olasılığının yüksek olduğunu söyler.

(iii) Sınıflandırma: Sınıflandırma, veri kümelerini hedef sınıflara ayırır. Sınıflandırma teknikleri, mevcut her bir veri örneği için hedef sınıfları tahmin eder. Örneğin; sınıflandırma tekniklerinden yararlanılarak, bir hasta, hastalık modellerine dayanılarak“ yüksek riskli” veya “düşük riskli” olarak sınıflandırılabilir. Bu yaklaşımda sınıflar bilinir, dolayısıyla bir tür denetimli öğrenme olduğu söylenebilir. İkili ve çok düzeyli olmak üzere iki yöntemin mevcut olduğu sınıflandırma görevinde, veri seti, eğitim ve test veri kümelerine bölünmüştür.

Dahası, sınıflandırıcı, eğitici veri seti yardımıyla eğitilir ve ardından sınıflandırıcının doğruluğu test veri seti üzerinde test edilir. Veri madenciliğinin sınıflandırma görevi çoğunlukla sağlık hizmetleri endüstrilerinde kullanılmaktadır [6]. Genellikle farklı hastalıkların tedavi maliyetini tahmin etmek için, sınıflandırma görevinden yararlanılır [11].

(iv) Kümeleme: Sınıflandırma ve kümeleme arasında ince bir fark vardır. Sınıflandırma denetimli öğrenme, kümeleme ise denetimsiz öğrenme yöntemidir. Sınıflandırma seviyelendirilen sınıf bilgisine sahip iken, kümelemede seviyelendirilen sınıfa ilişkin bilgi mevcut değildir. Kümelemede benzer veriler aynı kümeye yerleştirilirken, farklı veriler başka bir kümeye yerleştirilir [12]. Kümeleme, verileri bölümlemek/parçalara ayırmak için çok az bilgiye ihtiyaç duyar veya hiç

bilgi gerektirmez. Önce kümeleri tanımlamayı, ardından da kümelere yeni bir örnek atanmasını gerektiriyor olması, kümelemenin dezavantajıdır [13].

(v) Trend analizi: Literatürde, zamana bağlı birçok veri ile karşılaşılabilir. Bir şirketin satışları, bir müşterinin kredi kartı işlemleri ve hisse senedi fiyatları gibi örneklerin tamamı zaman serisi verileridir. Bu tür veriler, "zaman" özelliğine sahip nesneler olarak görülebilir. Zaman boyutu ile beraber verilerde örüntüler ve düzenlilikler bulmak zordur. Trend analizi bu kalıpları keşfeder.[9].

(vi) Regresyon: Regresyon analizi, iki ya da daha çok değişken arasındaki ilişkiyi ölçmek için kullanılan analiz metodudur. Eğer bu analiz tek bir değişken kullanılarak yapılıyorsa tek değişkenli regresyon, birden çok değişken kullanılıyorsa çok değişkenli regresyon analizi olarak adlandırılır. Regresyon analizi ile değişkenler arasındaki ilişkinin varlığı, eğer ilişki var ise bunun gücü hakkında bilgi edinilebilir. [14]

A. Sağlık Hizmetleri için Veri Madenciliği

Sağlık hizmetleri sektörlerinin verilere olan bağımlılığı her geçen gün artmaktadır [15]. Tıp biliminin en önemli görevi, herhangi bir hastalığın teşhisi ve hastaların tedavisidir. Son zamanlarda, tedavi esnasında ortaya çıkan maliyetin düşürülmesi ve tedavinin etkinliğinin artırılması amacıyla, doktorların el yazısı notları elektronik kayıtlara dönüştürülmüştür [16].

Ek olarak, sağlık hizmetlerinde veri madenciliği uygulamaları aşağıdaki kategoriler altında incelebilir:

a. Hastalıkların tanısı ve tahmini –Sağlık sektörü söz konusu olduğunda son derece önemli olan, hastalıkların tanısı ve prognozu (öngörüsü) [17], veri madenciliğinin sağlık hizmetlerine yönelik kullanımının başlıca amaçlarındandır. Sağlık hizmetleri için veri madenciliğine başvurulması, doktorların kendileri tarafından sağlanan sağlık hizmetlerini iyileştirmesine yardımcı olmuştur [15]. Hasta için yanlış bir tedavi seçilerek zaman ve para israf edilemez, öyle ki bu aynı zamanda hastanın sağlığına da zarar verebilir[18].

b. Hastanelerin derecelendirilmesi – Veri madenciliği teknikleri, derecelendirilmeleri için çeşitli hastanelerin tüm ayrıntılarını inceler[19]. Çeşitli hastaneler, ciddi hastalıkları olan hastaları idare etme kapasitelerine göre, kuruluşlar tarafından sıralanırlar. Yani, daha yüksek sıradaki hastaneler, en yüksek önceliğe

sahip yüksek riskli hastaların tedavisi için daha uygundur. Öte yandan, risk faktörü daha düşük olan düşük dereceli hastanelerde durum böyle değildir.

c. Daha iyi tedavi teknikleri – Veri madenciliği teknikleri yardımıyla, hem doktor hem de hasta, tüm tedavi tekniklerini karşılaştırarak en iyi tedavi opsiyonunu seçebilir. En iyi yöntemi seçmek için, tedavileri hem etkinlik hem de maliyet açısından değerlendirebilirler. Yine veri madenciliği sayesinde, çeşitli tedavilerin yan etkileri de bilinerek, hastalar için olası riskler azaltılır [6].

d. Etkili Tedavi Teknikleri– Veri madenciliği; tedavilerin nedenleri, semptomları, yan etkileri ve maliyetleri gibi unsurlar karşılaştırılarak, tedavilerin etkinliğinin analiz edilmesi için kullanılır. Örneğin; aynı hastalıktan muzdarip ancak tedavilerinde farklı ilaçlar kullanılan farklı hastaların tedavi sonuçları karşılaştırılabilir. Bu sayede, hastanın hem sağlığı hem de bütçesi açısından, hangi tedavinin daha etkili olacağı anlaşılabilir [20].

e. Hastalara sunulan daha kaliteli hizmetler – Teknolojideki ilerleme ile paralel olarak, dijitalleştirilmiş biçimde depolanan çok sayıda veriye halihazırda sahibiz. Veri madenciliği, bu devasa tıbbi veriler üzerinde uygulandığında; ilginç, bilinmeyen örüntülerin birçoğunu ayıklamaya yardımcı olabilir. Bu örüntülerin yardımıyla, hastalara sağlanan hizmetlerin ve bakımın kalitesi artırılabilir. Veri madenciliği, ayrıca, daha iyi tedavi edilebilmeleri adına hastaların gereksinimlerinin bilinmesine de yardımcı olur [6]. Milley, buna ek olarak, veri madenciliğinin, sağlık kuruluşları tarafından verilen hizmetlerinin geliştirilmesi için belirli hasta ihtiyaçlarının analizine yardımcı olabileceğini ifade etmektedir [21].

f. Hastanelerde enfeksiyon kontrolü – Hastane enfeksiyonları her yıl milyonlarca hastayı etkilemekte olup, ilaca dirençli enfeksiyonların sayısı oldukça fazladır [22]. Veri madenciliği aracılığıyla, enfeksiyon kontrolü verilerindeki bir takım düzensiz örüntülerin belirlenmesi için, enfeksiyon incelemesi yapılır [15]. Enfeksiyon kontrolü için, bu örüntüler, alanda yetkin biri tarafından daha fazla incelenir. Enfeksiyon kontrol verilerinde bilinmeyen kalıpları keşfetmek için veri madenciliği tekniklerinden yararlanan bu tip bir gözetim sistemi, Alabama Üniversitesi'nde uygulanmıştır [23].

g. Yüksek riskli hastaların saptanması – American Healthways (günümüzdeki adıyla Tivity Health, Amerika Birleşik Devletleri'nde kurulmuş bir şirket); diyabet hastalığı

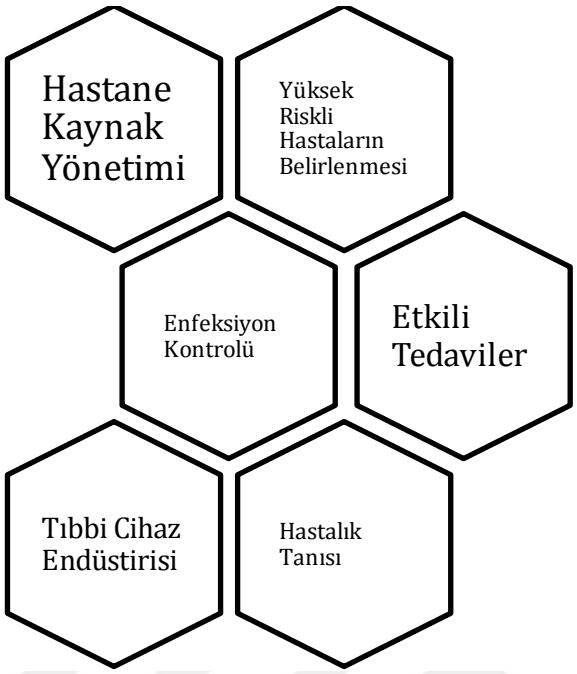
yönetim hizmetleri sunan hastanelerin, diyabetik hastaların yaşam kalitesini iyileştirmesine ve masraflarını düşürmesine yardımcı olmaktadır. Yüksek riskli ve düşük riskli hastaları ayırt etmek için, American Healthways, tahmine dayalı (öngörücü) modelleme tekniğini kullanmıştır. Tahmine dayalı modelleme tekniği ile sağlıklarıyla ilgili daha fazla endişe duyan yüksek riskli hastalar, sağlık hizmeti sağlayıcıları tarafından tespit edilmiştir [24].

h. Sigorta dolandırıcılığı ve kötüye kullanımının azaltılması – Sağlık sigortası; hastalar, doktorlar, hastaneler gibi aktörler tarafından sıradışı taleplerin örüntülerini belirlemek için bir model oluşturur [25]. 1998 yılında, Texas Medicaid Fraud and Abuse Detection System, veri madenciliği teknikleriyle dolandırıcılık ve suistimali tespit ederek milyon dolarlık tasarruf sağlamıştır [26].

i. Hastane kaynaklarının doğru yönetimi – Hastane kaynaklarının yönetimi, sağlık endüstrisinde önemli bir görev olup, veri madenciliği, hastane kaynaklarının yönetimi için bir model oluşturur. Group Health Cooperative (Amerika Birleşik Devletleri'nde kurulmuş kar amacı gütmeyen bir sağlık kuruluşu), veri madenciliğinden faydalanarak, hastanelere daha düşük maliyetlerle hizmet sağlar [27]. Blue Cross (sağlık sigortası sağlayıcı bir şirket), veri madenciliği yardımıyla maliyetleri düşürerek ve çıktıları iyileştirerek etkili bir hastalık yönetimi yürütür [28].

j. Tıbbi cihaz endüstrisi– Sağlık sektörü tıbbi cihazlar olmadan düşünülemez. Mobil iletişim ve ucuz kablosuz biyo-sensörler, hastaların önemli belirtilerini incelemek adına güvenli bir yöntem sağlayan mobil sağlık uygulamalarına ilişkin en önemli unsurdur [29]. Sonuç olarak, sağlık hizmetlerinde veri madenciliğinin başarısı, tamamen şeffaf ve organize sağlık hizmeti verilerinin mevcudiyetine bağlıdır. Bundan dolayı, sağlık endüstrileri, daha sonra uygun şekilde çıkarılabilmesi için verilerin nasıl toplanacağını ve depolanacağını dikkate almalıdır [30].

Sağlık hizmetlerinde veri madenciliği uygulamaları Şekil 2.5' de verilmiştir.



Şekil 2. 5 Sağlık Hizmetlerinde Veri Madenciliği Uygulamaları

Bu tez çalışmasında hastalık tanısı üzerinde durulmuştur. Üroloji branşında görülen hastalıkların teşhisine yönelik kullanılabilecek veri madenciliği algoritmaları kıyaslanarak sonuçlar yorumlanmıştır.

2.2.2. Sağlıkta Bilgi Yönetimi ve Veri Madenciliği

Son yıllarda, tıbbi sağlık hizmetleri artan bir ilgi ve popülerlik kazanıyor. Moleküler, biyomedikal teknikler, tıbbi görüntüleme ve hastaların tıbbi kayıtları gibi teknolojik gelişmeler nedeniyle, her gün büyük miktarda tıbbi veri üretilmektedir. Klinik uygulamalardan bireysel araştırmalara kadar, söz konusu tıbbi veriler; hasta kayıtları, laboratuvar raporları gibi tıbbi bilgilerin dijitalleştirilmesinin ardından, yüzlerce özel veri halka açık veritabanında saklanmaya başlamıştır. Bugün, veri toplama hızı veri çıkarma hızından çok daha yüksektir. Bu nedenle, işe yarar olması için bu verilerin iyi organize edilmesi ve saklanması gerekmektedir. Bu büyük tıbbi veri havuzlarını işlemek ve ondan yararlı örüntüler elde etmek için, yeni bilgi teknolojisi tekniklerine ihtiyaç vardır.

20. yüzyılda, psikoloji ve bilişsel bilimlerle birlikte yönetim, bilgi yönetiminin evrimine yol açmıştır [31]. 'Bilgi yönetimi' terimi 80'lerde ortaya çıkmış olup, akademik disiplin 1995'te geliştirilmiştir [32]. Aslında bilgi yönetimi, performansı en üst düzeye çıkarmak adına bilgiyi toplamak, yönetmek, kullanmak, analiz etmek, paylaşmak ve keşfetmek için yönetimsel bir yaklaşımdır [33]. Bilgiyi neyin oluşturduğuna dair bir tanım mevcut olmamakla birlikte, soyut ve çıkarımsal olmasının yanında, bilgi, hipotez üretmeyi ve karar vermeyi desteklemek için gereklidir. Yakın geçmişte, araştırmacılar, bilgi yönetiminin organizasyonel ve operasyonel performans üzerinde olumlu etkileri olduğunu ortaya koyan çalışmalar yapmıştır [34, 35]. Orzano ve ark. tarafından yapılan çalışmada önerilen bir bilgi yönetimi modeli, sağlık hizmetleri endüstrilerine ilişkin önemli bilgiler vermiş olup, bilgi yönetimi süreçlerinin daha iyi organizasyon öğrenimi ve karar vermeyi sağladığı ve bunun da, daha iyi bir organizasyon performansına yol açtığı ifade edilmiştir. Biyomedikal bilginin açık hale getirilmesi adına; verilerin depolanmasını, kurtarılmasını, paylaşılmasını ve yönetimini desteklemek için bilgi yönetimi metodolojileri ve tekniklerinden yararlanılmıştır. Son zamanlarda hem bilimsel hem de ticari alanlarda kullanılmakta olan bilgi yönetiminin, şirketlerdeki uygulamaları için birçok hedef ve zorluk söz konusudur. Bilgi yönetimi, performanslarını artırabilir, riskleri değerlendirebilir, ortaklık geliştirmeye yardımcı olabilir, yönetimi organize edebilir ve ekonomik değerlerini artırabilir [37]. Bilgi yönetimi konusuna, T.D. Wilson'ın getirdiği bazı eleştiriler mevcuttur [38]. Fakat, şirketlerin ve

organizasyonların bilgi yönetimine gerçekten ihtiyaç duyması nedeniyle, bu eleştiriler bilgi yönetimine yenik düşebilir.

Bilgi yönetimi yöntemleri ve teknikleri; insanlar ve teknoloji, gereksinimleri ortaya çıkarma ve değer ölçümü olmak üzere üç ayrı kategori altında incelenebilir. Günümüzde, bu çerçevede, teknik perspektifler kadar beşeri perspektifler de göz önünde bulundurulmaktadır. Beşeri perspektiflerin, temelde, motivasyon ve benimseme ile alakalı olduğu söylenebilir. Yalnızca teknolojik fayda uğruna değil, aynı zamanda organizasyonu/şirketi etkileyeceğinden dolayı, çalışanlar, bilgi yönetiminden yararlanmaları adına finansal veya finansal olmayan teşvikler aracılığıyla motive edilirler. Teşviklerin yanı sıra, hem çalışan hem de şirket için bir kazan-kazan sistemi ve bir kazan-kaybet ödül sisteminin mevcudiyeti önerilmiştir [39]. Bilgi motivasyonuyla ilgili bir diğer konu ise, bilginin benimsenmesidir. Bunun sebebi, insanların bilgi yönetimini kullanmaya hazır olmayışıdır [40]. Temelde bilgi keşfi olarak bilinen daha geniş bir olasılığın temel bir adımı olan veri madenciliği, farklı alanlarda kullanılmaktadır (örneğin; farklı biyolojik bilgilerin, ilaçlara ve hasta bakımına dair bilgilerin keşfi gibi). Modellerin istatistiki analizinde kullanılan veri madenciliğine, tıp alanında da, bir teknik olarak sıklıkla başvurulmaktadır[27]. Veri madenciliğinin temel amacı, bir dizi veriyi ya da ham veriyi analiz ederek, yeni ve işe yarar kalıpları belirlemek ve ortaya çıkarmaktır [41]. Yapay sinir ağları, karar ağaçları, bulanık kümeler, destek vektör makineleri, bayes ağları ve genetik algoritmalar gibi çeşitli veri madenciliği teknikleri, sistem ve kullanıcılar tarafından bilinmeyen bilgi ve kalıpları keşfetmek için kullanılır [42, 33]. Biyomedikal veri madenciliğinde, hasta verileri "bireysel olarak tanımlanabilir" olmamalıdır. Bu, hiçbir kaydın hasta hakkında yeterli veri sunmaması anlamına gelmektedir. Bu sayede hiç kimse hastayı teşhis edemez [2].

2.3. Veri Madenciliği Algoritmaları

Veri madenciliği, tıbbi verilerin işlenmesi için çeşitli teknikler kullanır. Aslında, nitelik seçimi için, veri madenciliği teknikleri kullanılmaktadır. Sınıflandırma için gerekli olan minimum bir öznitelik alt kümesini seçmek gerekir. Özellik kümesinin gereğinden fazla olması durumu verimliliği düşürebilir. Öznitelik seçimi, tıbbi teşhisin alanına giren bir sorundur [43]. Öznitelik alt kümesi üretimi, veri ön işlemenin bir adımı olan, veri azaltma olarak da bilinmektedir [44]. Bunun dışında, öznitelik seçimi, modelin doğruluğunu maksimize etmek için gereken özelliklerin sayısını en aza indirir. Ayrıca, özellik kümesinin gerektirdiği alanı azaltmaya yardımcı olur. Özellik kümesinde bulunabilecek fazlalık gürültü verimliliğini ortadan kaldırır ve bu sayede veri madenciliği algoritmasının verimliliği artar [45].

Özellik seçim süreci, dört temel aşamadan oluşmaktadır: alt küme oluşumu, alt kümenin değerlendirilmesi, durdurma ölçütü olarak kullanılan bir seçim kriteri ve nihai alt küme özelliği [44]. Başlangıç adımı olarak, mevcut olan boş değerler ve fazlalıklar gibi bazı tutarsızlıklar ortadan kaldırıldıktan sonra, özellik kümesi araştırılır. Özellik kümesinin aranmasının ardından, alt küme oluşturma süreci başlar. Sonrasında, öznitelik değerlendiricisi, oluşturulan alt kümeyi değerlendirir [47]. Alt küme oluşturma ve değerlendirme aşaması, seçim/durdurma kriterleri sağlanana dek sürer. Öznitelik seçiminin amacı, uygun maliyetli ve verimli bir model üretmektir [46].

2.3.1. Yapay Sinir Ağları

Yapay Sinir ağları; "çoğunlukla doğrusal olmayan sınıflandırma için kullanılan, denetimli sınıflandırıcılardır. 'Nöron' adı verilen biyolojik sinir hücrelerinin çalışma dinamiklerini devralırlar". Hesaplamalı nöronlar ve biyolojik sinir hücreleri arasındaki farkı belirtmek için, bilgisayar mühendisliği ve veri bilimi alanında bu ağlar, "yapay sinir ağları" adıyla anılır [48].

20. yüzyılın başlarında geliştirilmiş olan yapay sinir ağları [48], tıp biliminin kullandığı en popüler veri modelleme algoritmalarından biridir. Karar ağaçlarının ve Destek Vektör Makinesi'nin icadından önce, yapay sinir ağları en iyi sınıflandırma algoritmasıydı [49].

Yapay sinir ağlarının avantajları aşağıdaki gibi sıralanabilir: [48]

1. Eğitim için gürültülü verileri düzgün bir şekilde işleyebilir.

2. Girdi ve çıktı arasında karmaşık ilişkiler üretebilir. Herhangi bir dış yardım olmadan kendi özelliklerine göre verileri analiz edebilir ve düzenleyebilir.

3. Kümeleme ve prototip oluşturma için çeşitli yapay sinir ağları kullanılabilir.

Yapay sinir ağları'nın kullanılmasındaki temel amaç, örüntü tanıma ve sınıflandırma görevlerini yerine getirmektir [50]. Yapay sinir ağı sistemi, bir insan beyni örnek alınarak modellenmiştir. İnsan beyni, birbirine bağlı milyonlarca nörondan oluşmaktadır. Yapay nöronları, benzer biçimde birbirine bağlayan sinir ağında, her bağlantının bağlı ağırlığı vardır. Uyarlanır doğası nedeniyle, ağırlıkları ayarlayarak hatayı en aza indirmeye yardımcı olur [3]. Bu nöronlar, çıktı fonksiyonunu üretmek için paralel olarak birlikte çalışırlar. Öğrenme aşamasında, ağ, girdinin doğru sınıf etiketini tahmin etmek için, ağırlıkları ayarlayarak öğrenecektir. Yapay Sinir Ağları, basit modelleme yöntemlerinden farklı olarak, doğrusal olmayan ilişkiyi tahmin edebilmelerinden kaynaklı ek bir avantaja sahiptir [51]. Tıbbi verilerin analizinde önemli bir rol oynayan yapay sinir ağlarının bu alandaki uygulamaları; doku sınıflandırması, hastalık tahmini ve ilaç geliştirmeyi içermektedir. Bazı çalışmalarda yapay sinir ağı kullanılarak, kalp hastalıklarının tahmini yapılabilir olduğu tespit edilmiştir [52]. Aşağıda, yaygın olarak kullanılan yapay sinir ağ mimarilerine yer verilmektedir:

i. Çok Katmanlı Yapay Sinir Ağı (MLNN): Bu tür yapay sinir ağları, doğrusal olmayan kümeler için sınıflandırma problemini çözdüğü gizli katmanları kullanır [53]. Bu gizli katmanlar genellikle hiper düzlemler olarak yorumlanırlar. Bu tür sinir ağları, farklı veri kategorilerini sınıflandırmak için kullanılmaktadır.

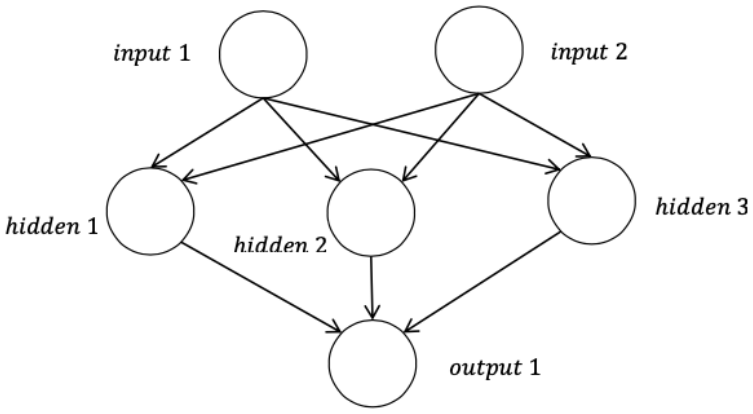
ii. Polinomiyal Yapay Sinir Ağları (PNN): Polinomiyal yapay sinir ağları, çok değişkenli polinom eşlemeleri üreten çok katmanlı algılayıcılar gibi birimlere benzeyen nöronlara sahiptir.

Şekil 2.6'de tek gizli katmanlı yapay sinir ağı gösterilmiştir.

Giriş 1 (Input 1)
Giriş 2 (Input 2)

Girdi 1 (Hidden 1)
Girdi 2 (Hidden 2)
Girdi 3 (Hidden 3)

Çıkış 1 (Output 1)

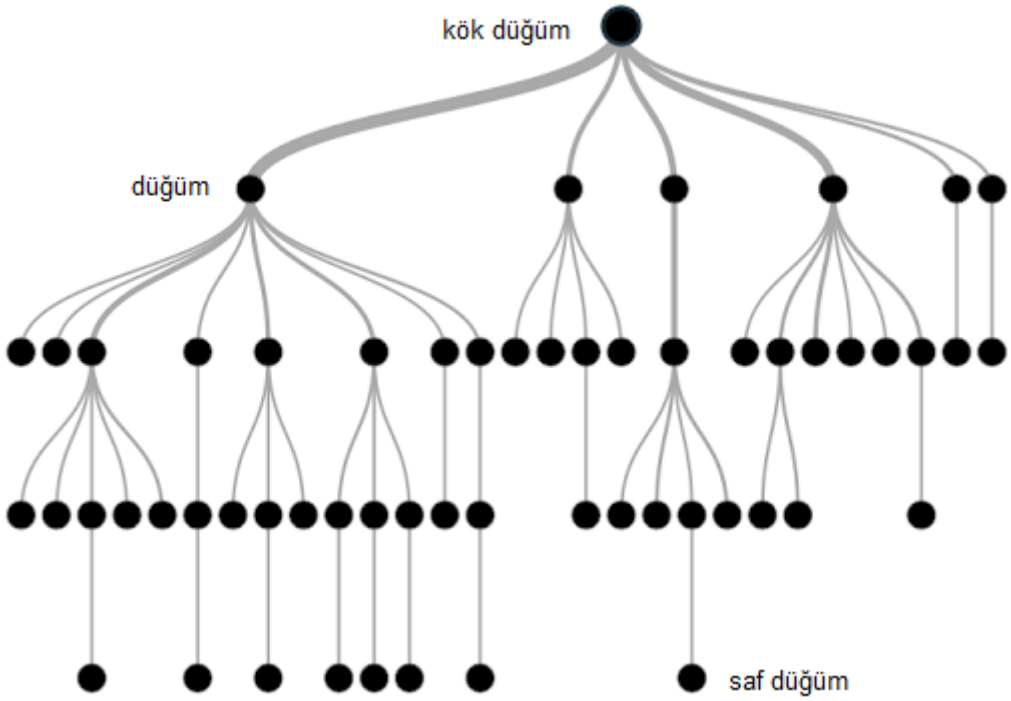


Şekil 2. 6 Tek gizli katmanlı yapay sinir ağı.

2.3.2. Karar Ağacı

Karar ağacı algoritması, koşullu bir olasılık algoritmasıdır. Karar ağacı, eğitim verilerinin analizinden elde edilen kurallara göre veri sınıflandırmasını test eder. Karar ağacı, karar için bir nedene ulaşmak isteyen süreçlerde çok kullanışlıdır. Pek çok makine öğrenimi algoritmasında sadece sınıflandırma yapılır ve sınıflandırma nedeni belli bir sınıfa açıklanmaz ve bazen açıklanamaz, ancak karar ağacı algoritmasında sınıflandırma insanlar tarafından anlaşılabilir kurallara dayandığından karar vermenin nedeni de anlaşılabilir. Örneğin, profesyonel pazarlamada, farklı grupları ve başarılı pazarlama için ihtiyaçlarını bilmek çok faydalıdır [54].

Şekil 2.7'de örnek bir karar ağacı algoritma sonucu verilmiştir.



Şekil 2. 7 Örnek bir karar ağacı algoritma sonucu

Karar ağacı, dallardan ve düğümlerden oluşur. İç düğüm (internal), alt iki dala ayrılırken saf (terminal) düğüm en son yerdir ve hiç alt düğümü yoktur.[107] Terminal olmayan her düğüm, bir veri öğesindeki bir testi veya koşulu temsil eder. Karar ağaçları, örnekleri terminal olmayandan terminal düğümlerine doğru sıralayarak sınıflandırır [54]. Hangi dalın seçileceği, tamamen testin sonucuna bağlıdır. Örneğin, tıbbi geri kabul için bir karar ağacı olduğunu düşünelim. Bu ağacın yardımıyla, bir hastanın yeniden hastaneye yatması gerekip gerekmediğine karar verebiliriz [3]. Temelde çeşitli avantajlar ve dezavantajların ve her seçeneğin potansiyel değerinin görsel bir temsilini oluşturan karar ağaçları [55], genellikle yöneylem araştırması analizinde koşullu olasılıkların hesaplanması için kullanılır. Karar ağaçları yardımıyla en uygun alternatifler seçilebilir ve maksimum bilgi kazanımına dayalı olarak kök düğümden yaprak düğüme geçişin olduğu görülür, bu geçişte benzersiz sınıf ayrımına işaret eder [57]. Veri madenciliğinin diğer bazı uygulamalarında (örneğin pazarlama), bir tahminin doğruluğu, ihtiyaç duyulan tek şey olabilir. Örüntünün işleyişine dair bilgi sahibi olmak önemli olmayabilir.

Örneğin; bir pazarlama profesyoneli bir pazarlama kampanyası başlatmak istediğinde, müşteri segmentlerinin genel tanımlarına ihtiyaç duyacaktır. Karar ağacı algoritması, bu tür uygulamalar için oldukça elverişlidir [58].

2.3.3. Bulanık Kümeler

Bulanık kümeler ve bulanık mantık, veri madenciliğinde genellikle belirsizliği temsil etmek ve işlemek için kullanılan en iyi metodolojidir. Eksik ve gürültülü veriyle başa çıkmanın en iyi yöntemlerinden biri olan bulanık kümeler ve bulanık mantık [51], önerilen uzman sistemin uygulanması için gereklidir. Bu, belirsiz verileri işlemeye yardımcı olan bulanık küme teorisi, Zadeh [59] tarafından tanıtılmıştır. Bulanık mantık yardımıyla, herhangi bir özel durumun herhangi bir kümeye düşme olasılığı hesaplanabilir ve ardından bu değere dayalı olarak kararlar alınabilir [60].

2.3.4. Destek Vektör Makinesi

İlk kez 1999 yılında ortaya atılmış olan Destek vektör makinesi (DVM) kavramı [61-62], diğer tüm algoritmalara kıyasla en doğru sonuçları sağlar. Bir sınıflandırma tekniği olup, istatistiksel öğrenme teorisi temelinde çalışır [62-63]. DVM, çeşitli çekirdekler için evrensel bir yaklaşım olarak kullanılmıştır [64]. Öğrenme verilerinin alt kümesine destek vektörü adı verilir ve destek vektör makineleri bunun yardımıyla tanımlanır. Yerel minimum değerin olmaması, DVM'nin temel özelliklerinden biridir.

DVM modeli, eğitim verilerinin bir temsili olup, destek vektörlerinin yardımıyla, yoğunlaştırılmış veriler elde edilebilir [65]. DVM, iki farklı sınıfın örnekleri arasındaki marjı en üst düzeye çıkaran optimal bir ayırıcı hiper düzlem bulur. İkili sınıflandırmaya ilişkin problemlere yönelik geliştirilmiş olan DVM, daha sonra çok sınıflı problemlerle ilgili de kolayca genişletilebilir niteliktedir. DVM'nin popülerlik kazanmasındaki en önemli nedenlerden biri de budur [66-67]. Bir ikili sınıflandırma görevinde, hiper düzlem, iki çıktı arasındaki bölünmedir. Görevler için faydalı olması adına tekli ve çoklu hiper-düzlemler oluşturabilir. DVM'leri uygulamak için iki yöntem mevcuttur. İlki, matematiksel programlamayı içermektedir. İkinci yöntem ise, çekirdek işlevlerini kullanır. Hiper düzlem kullanmanın temel görevi, veri noktaları arasındaki ayrımın maksimize edilecek olmasıdır [3]. Gürültülü verilerde, iki farklı sınıfın örnekleri arasındaki marj maksimize edilerek hata en aza indirilir ve hiper düzlem, ayırma alanının merkez çizgisi olarak tanımlanır.

DVM'lerin iki türü vardır. İlki, veri noktalarını doğrusal karar sınırı yardımıyla ayıran doğrusal DVM'lerdir. Bunlar, rahatlıkla ikiye ayrılabilen veri kümelerinde iyi performans sergiler. Verilerin sınıflandırılmasını tahmin ederken maksimum genelleme yapabildiğinden dolayı en güçlü algoritma olduğu bilinir[45]. DVM, kapakçık sınıflandırması/kalp atışı gibi ikili sınıflandırma problemlerinde kesinlik sağlar [68-70].

2.3.5. Bayes Ağı

Bayes ağı, belirsiz alan hakkındaki bilgileri temsil eden özel bir ağ türü olup, olasılıklı grafik modellerin (GM'ler) alanına aittir. Bayes ağında; düğümler değişkenleri, çeşitli kenarlar da bu değişkenler arasındaki olasılıksal bağımlılıkları temsil etmektedir [71-73]. Bayes ağı, her değişken için iki tür bilgi tanımlar [74].

2.3.6. Genetik Algoritma

Genetik algoritma, genetiğe ve seleksiyona dayalı bir arama ve optimizasyon tekniğidir. Bu algoritmalar, temelde, örüntüleri bulmaktan ziyade veri madenciliği algoritmalarının öğrenme süreci için bir rehber görevi gören sinir kümelerinde kullanılır. Ayrıca, değişkenler ve aralarındaki bağımlılıklar hakkında hipotez formüle etmek için, veri madenciliğinde ilişkilendirme kuralları ya da başka bir biçiminde kullanılırlar. Genetik algoritmanın temel fikri, şema teorisinde sunulan diğer çözümlerin iyi kısımlarını - doğanın canlıların DNA'larını birleştirerek yaptığı gibi - birleştirerek çok daha iyi bir çözüm elde edilebilir [75]. Genetik bir algoritmada, belirli seçim kuralları altında uygunluğun en üst düzeye çıkarıldığı bir duruma evrimleşen birçok bireyden oluşan bir popülasyon mevcuttur [76]. Başlangıçta, her kuralın bir problemin çözümünü temsil edeceği şekilde, rastgele bir kurallar topluluğu oluşturulur. Sonrasında, genellikle en güçlü kurallar olan ebeveyn olarak kural çiftleri seçilir ve daha sonra yavru üretmek için bu kural çiftleri birleştirilir [77]. Genetik bir algoritma temelde seçim, geçiş ve mutasyon olmak üzere üç operatörden oluşur. Seçimde, yeni bir nesil yetiştirmek için uygunluk esasına göre bir dizi seçilir. Ardından, çaprazlama ile daha iyi yavrular üretmek için bu uygun ve iyi diziler birleştirir. Mutasyon, daha sonra, bir diziyi yerel olarak değiştirir ve böylelikle genetik çeşitlilik, bir popülasyonun bir neslinden diğerine devam eder. Her nesilde, algoritmanın sonlandırılması için popülasyon değerlendirilir. Sonlandırma

kriterleri karşılanmadığı takdirde, üç operatör tarafından tekrar çalıştırılır ve ardından yeniden değerlendirilir.

2.3.7. Naive Bayes Algoritması

Bayes teoremi veri madenciliği ve makina öğreniminin temel dinamiklerinden biridir. $x = \{x_1, ... x_n\}$ biçiminde bir veri seti olduğunda, öğrenme görevi, gözlemlenen kümenin meydana getirdiği dağılımın özelliklerini açığa çıkarmaktır. Bayes kuralı; veri setinin koşul olarak verildiği bir özelliğin olasılığını tahmin etme yöntemidir. Yöntem bunu yaparken h_1 ve h_2 hipotezlerinin h_1 /\ h_2 ifadesinin gerçekleşemeyeceği hallerde açığa çıktığını ve x_i'nin gözlemlenebilir bir olay olduğunu varsaymaktadır.

$$P(h_1|x_i) = \frac{P(x_i|h_1)P(h_1)}{P(x_i|h_1)P(h_1) + P(x_i|h_2)P(h_2)} \tag{2.1}$$

h_1 hipotezi ile bağlantılı önsel olasılık $P(h_1)$ iken, $P(h_1|x_i)$ ise sonsal olasılıktır. x_i veri değerinin ortaya çıkma olasılığı $P(x_i)$ ve bir hipotez verildiğinde gözlemin onu tahmin ettiği koşullu olasılık ise $P(h_1|x_i)$'dur. *m* farklı hipotez söz konusu olduğu durumlar için (2.2) uygulanır ve bu durumda (2.1) de, en genel haliyle (2.3)' e eşittir.

$$P(x_i) = \sum_{j}^{m} P(x_i|h_j)P(h_i) \tag{2.2}$$

$$P(h_1|x_i) = \frac{P(x_i|h_1)P(h_1)}{P(x_i)} \tag{2.3}$$

Naive Bayes algoritması, Bayes koşullu olasılık kuralının gözlemlenen bir örneğin özellikleri arasında uygulanması temeline dayanan doğrusal bir sınıflandırma algoritmasıdır. Naive Bayes algoritmasında, her özelliğin sınıflandırma sonucuna katkısı eşittir. Veri setindeki bir özelliğin var olup olmaması, diğer özelliklerin var olup olmamasını kesinlikle etkilememektedir. Bu algoritma, sınıflandırmaya ilişkin gerekli parametreleri tahmin etmek için yalnızca değişkenlerin varyansına ve ortalamasına ihtiyaç duymaktadır. Bundan dolayı, öğrenme şeması kontrol

edildiğinden, eğitim adımında sınıflandırıcı için genellikle az miktarda veri yeterlidir [71].

Bir değişkenler grubu sınıflandırmasında, eğitim kümesinden elde edilen koşullu ve önsel olasılıklar kullanılır [71]. t_i değişkenler grubunun, $\{x_{i1}, x_{i2},\ldots,x_{ip}\}$ şeklinde gösterilen p bağımsız özellik değerine sahip olduğu varsayılsın, $P(x_{ik}|C_j)$ bilindiğinden dolayı, her x_{ik} özelliği ve C_j sınıfı için $P(t_i|C_j)$ aşağıdaki gibi tahmin edilir:

$$P(t_i|C_j) = \prod_{K=1}^{P} P(x_{ik}|C_j) \tag{2.4}$$

Naive Bayes algoritması'nın avantajları aşağıdaki gibi sıralanabilir: [72]

1. Büyük veri kümeleri için de hızlı ve doğrudur.

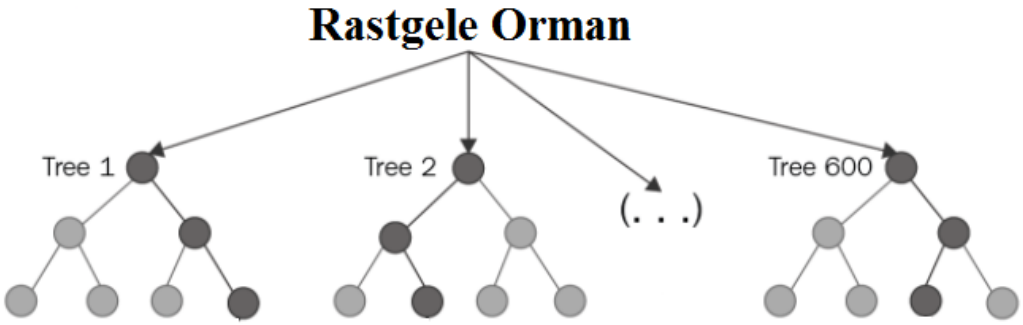
2. Hesaplamaları kolaylaştırır.

Naive Bayes'in en önemli dezavantajı, aralarında bağımlılığın olduğu bazı durumlarda değişkenler, doğru sonuçlar vermez. [72].

2.3.8. Rastgele Orman (Random Forest)

Rastgele Orman; ilk olarak Adele Cutler ve Leo Breiman tarafından, denetimli bir grup sınıflandırma algoritması biçiminde geliştirilmiştir. Aslında bu algoritma isminde geçen "Orman" sözcüğü karar ağacı algoritmasıyla farkını anlatır. Bu algoritmada tek bir karar ağacı yerine, orman gibi bir grup ikili karar ağacı kullanılmaktadır. "Rastgele" sözcüğü, bu ormanın, daha şekilsel bir biçimde, ön yükleme ile değiştirilen temel eğitim kümesinin rasgele alt kümelerinden oluşturulduğunu gösterir. Eğitim kümesi boyutuna göre, ondan binlerceye kadar ayrı karar ağacından rastgele bir orman oluşturulabilir. Bu nedenle karar ağacı algoritmasının avantajları ve dezavantajları bu algoritma içinde geçerlidir.
Rastgele orman sınıflandırıcıları, doğrusal olmayan sınıflandırma görevlerinin çoğunu yerine getirebilir. Rastgele Orman, her vektöre ilişkin değişken sayısı çok fazla olduğunda ve eğitim kümesi çok geniş olduğunda en ideal algoritmadır [54-

55].Rastgele Orman algoritması iki aşamadan oluşur. Eğitim aşaması olarak da adlandırılan ilk aşamada, etiketli temel kümenin rastgele alt kümelerinden orman yapısı meydana getirilir. Bir N vektör popülasyonun bir sınıflandırma çalışmasında elimizde olduğunu varsayalım ve hedefimiz 'n' sayıda karar ağacı barındıran rastgele bir orman oluşturmak olsun. İlk olarak, orman yapısını meydana getirmek amacıyla, temel veri setinden ’n’ sayıdaki alt kümeleri rastgele alıyoruz. ‘Yerine koyma tekniği ile her yeni adımda alt kümelerin seçimi gerçekleştirilir. Bu nedenle, alt kümeler benzer olabilir. Böylece, rastgele seçilen ’n’ altkümeye bağlı olarak ormandaki her karar ağacı büyür. İkinci bir parametre olarak, değişken sayısı sabit tutulabilir. Bundan dolayı, her alt kümede aynı tutulmasına rağmen alt kümeler arasındaki değişkenler farklı olabilir. Her iki halde de, karar ağaçlarının en iyi bölünme ölçütleri rastgele seçilen altküme vektörleri ile belirlenmektedir. Ağaçlar büyütülürken her ağacın tamamen büyümesini sağlamak için, karar ağaçları budanmaz. Şekil 2.8’de örnek bir rastgele orman algoritması verilmiştir.



Şekil 2. 8 Örnek bir rastgele orman algoritması

Test aşaması olarak da adlandırılan ikinci aşamada, oluşturulmuş orman yapısı kullanılarak ormanın daha önce hiç görmediği etiketlenmemiş bir vektörün sınıfı tahmin edilir. Ormandaki karar ağaçlarının her biri ile etiketlenmemiş vektör sınıflandırılır. Her bir bölge ağacının sınıflandırma sonucu oylama olarak adlandırılır. Sonuç olarak, nihai sınıflandırma işlemi, tek bir karar ağacına dayalı olarak değil ormandaki oy çoğunluğu göz önünde bulunrularak yapılır [55].

Bu algoritmalar sağlık hizmetlerinde de yaygın olarak makine öğrenme teknikleri olarak karşımıza çıkmaktadır.

Sağlık Hizmetlerinde Makine Öğrenme Teknikleri

Makine öğrenimi alanında çok sayıda araştırma mevcut olup, bu araştırmalar çoğunlukla uygulamaya yöneliktir. Makine öğrenimi araştırmaları sağlık alanında yaygın olarak kullanılmaktadır. Makine öğrenimi yöntemleri, araştırmadaki artışın ilerlemelere yol açacağı alanları belirleyebilir. Algoritmik çözümlerin sunulamadığı ve biçimsel kodların veya uygulama alanı hakkında bilgi tanımının yetersiz olduğu durumlarda, makine öğrenimi yöntemleri ortaya çıkar. Makine öğreniminin birçok yöntemi bulunmakla birlikte, bunlar genel olarak, öğrenme aşamasında manipülasyonun doğasına göre sembolik ve alt-sembolik olarak sınıflandırılabilir [78]. Söz konusu sembolik öğrenme yöntemi olduğunda, karar ağaçlarında olduğu gibi, gerekli bilgi ve yapılan çıkarım seviyesi farklıdır [79]. Diğer yandan, genetik algoritmalar [80] ve yapay sinir ağları [81] alt-sembolik sınıflandırma yöntemlerine örnektir.

Sağlık alanında makine öğrenimi yöntemlerinden bahsedecek olursak; bu tekniklerin ve araçların hastalıkların teşhis ve prognozunda (öngörüsünde), hastalığın ilerlemesine yönelik tahminlerde veya tıbbi bilginin çıkarılmasında yardımcı olduğu söylenebilir. Tümevarımsal öğrenme gibi sembolik yöntemler, uzman sistemlere öğrenme ve bilgi yönetimi eklemek için kullanılır [82]. Makine öğrenimi araçları; eksik değerler, rastgele gürültü veya sadece birkaç hasta kaydı gibi, tıbbi alanın bazı karakteristik özelliklerini ele almaya yardımcı olur [83]. Sinir ağları gibi alt-sembolik öğrenme yöntemleri, bu veri kümelerini idare edebildiklerinden dolayı, karar vermeyi geliştirmeye de yardımcı olurlar [84]. Tıbbi teşhiste önemli bir uygulama, göz ardı edilemeyecek bir yardım sağlayan tıbbi görüntülemelerin yorumlanmasıdır [85]. Gerçekten de, sağlık hizmetleri alanının bilgisayar sistemlerine olan bağımlılığının artması ile makine öğrenimi yöntemleri birçok durumda hekime önemli ölçüde yardımcı olabilmekte ve gerçek zamanlı tanı konmasını sağlayabilmektedir.

Makine öğrenimi, tıbbi kararlar almanın yanı sıra tıbbi karar verme sistemlerinin verimliliğini ve kalitesini de iyileştirir [86]. Bir tıp uzmanının bir sistemden elde edilen sonuçları ne derece iyi kavrayabileceği ve kullanabileceği, büyük ölçüde, kullanılan makine öğrenimi yöntemlerine bağlıdır. EKG tanısı için çalışan birçok

araştırmacı tıbbi uzman, tıbbi uzman sisteminin bilgilerini geliştirmeye yönelik makine öğrenme teknikleri uygulamıştır.

Daha önce veri madenciliğinden bahsederken özellikle sağlık hizmetleri ve sağlık sektöründeki uygulamalarından bahsedilmiştir. Ancak veri madenciliği uygulamaları sadece sağlık alanı ile kısıtlı değildir. Bu bölümde diğer alanlardaki uygulamalarından da bahsedilecektir.

2.4. Veri Madenciliği Uygulamaları

Veri çözümlemesine yönelik veri madenciliğinin gücünden kaynaklı olarak, veri madenciliğinin, gerçek hayatta farklı alanlarda çok çeşitli uygulamaları mevcuttur [100–102]. Bu uygulamalarda bir veya daha fazla veri madenciliği görevi, tekniği ve yöntemi uygulanmaktadır [6, 15]. Veri madenciliğinin gerçek hayattaki çeşitli uygulamalarına aşağıdaki bölümlerde yer verilmiştir.

2.4.1. Telekomünikasyon Sektörü

Telekom/mobil servis sağlayıcıları tarafından (i) pazarlama kampanyası, (ii) müşteriyi elde tutma, (iii) müşteriler için, müşteri segmentasyonuna dayalı paket oluşturma, (iv) iletişim altyapısının optimum kullanımı gibi unsurlara yönelik strateji oluşturulması ve tasarlanmasında veri madenciliğinden yararlanılmaktadır. Mobil hizmet sağlayıcıları, sınıflandırma ve kümeleme kullanımı sayesinde, doğrudan pazarlamayı teşvik edecek pazarlama kampanyaları için strateji geliştirebilirler. Kümeleme ve ardından sınıflandırma yardımı ile müşteriler, hareketlerinin tahmin edilebilmesi için çeşitli gruplara ayrılabilir. Müşterilerin servis sağlayıcı tarafından elde tutulabilmeleri adına, spesifik pazarlama stratejileri ve paketleri formüle edilebilir ve tasarlanabilir. Tanımlanan müşteri gruplarına göre, özel paketler de, bu çeşitli müşteri gruplarının ihtiyaçlarına/gereksinimlerine uygun şekilde formüle edilebilmektedir. Paketlerin tasarlanmasında birliktelik çözümlemesinden de yararlanılabilir. Ağ kullanım modeli, yetersiz kullanılan ve aşırı kullanılan ağ altyapısını ayırt etmek için veri madenciliğine başvurularak analiz edilebilir. Böylece altyapının geneli en verimli şekilde kullanılabilir ve/veya ihtiyaca göre geliştirilebilir [87, 88].

2.4.2. Perakende Sektörü

Sektör oyuncuları ve süper market sahipleri, veri madenciliğinden yararlanabilir. Veri madenciliği yardımıyla; (i) müşterilerin satın alma davranışları, (ii) pazar-sepet analizi, (iii) müşterilerin seçimleri, (iv) ürünlerin raflara yerleştirilmesi, (v) etkili tekliflerin/kuponların/indirimlerin sunulması ve (vi) müşteri segmentasyonu'na dair tahminde bulunabilirler. Müşterilerin satın alma davranışlarını ve pazar sepet analizini öngörmek için birliktelik çözümlemesi kullanılır. Birliktelik kullanılarak, satış verilerinden, sağlanan destek ve güven seviyesine dayalı yaygın öge kümeleri ortaya çıkarılabilir. Böylece, bu yaygın öğe kümeleri yakınlara yerleştirilerek satışlarda artış elde edilebilir. Satış verilerinin kümeleme kullanılarak analiz edilmesi ile ürünlerin yerleştirilmesi için en elverişli konum (yani raflar) belirlenebilir ve satışların artırılması için en uygun teklifler tespit edilebilir. Satış verileri, kümeleme ve/veya sınıflandırma kullanılarak çeşitli müşteri segmentlerini keşfetmek için de analiz edilebilir. Farklı pazarlama kampanyaları ve promosyonlar, ya da teklifler, ortaya çıkan müşteri segmentlerine uygun biçimde özelleştirilebilir. Çok az satın alma gerçekleştirdiği halde çok harcayan müşteri ile çok sık satın alım yapan ancak küçük miktarlar harcayan müşteriye yaklaşım farklı olacaktır [89-91].

2.4.3. Finansal Veri Analizi

Finans ve bankacılık sektörlerindeki finansal veriler, sistematik veri analizini ve veri madenciliğini kolaylaştırır. Finansal veri analizinde veri madenciliği; (i) kredi ödeme tahmini, (ii) müşteri kredi politikası analizi, (iii) hedefe yönelik pazarlama için müşteri segmentasyonu ve (iv) kara para aklanmasının tespiti için kullanılabilmektedir. Öznitelik sıralaması ve öznitelik seçimi sayesinde, veri madenciliği ile müşterilere ait (i) kredi geçmişi, (ii) ödeme-gelir oranı ve (iii) kredi vadesi'nin tespit edilmesi için müşterilerin ödeme geçmişleri analiz edilebilir. Bu tahmin, bankaların/finans kuruluşlarının kredi verme politikalarını oluşturmalarında ve müşterilere puanlarına göre kredi vermelerinde yardımcı olacaktır. Günümüzde bankalar ve finans kuruluşları, müşterilere kredi vermeden önce, veri madenciliğine dayalı CIBIL puanını kontrol etmektedir [92-95]. CIBIL puan, kredi geçmişinizin üç basamaklı sayısal bir özetidir. Puan, CIBIL raporunda bulunan kredi geçmişi kullanılarak elde edilir.

2.4.4. Sağlık sektörü

Sağlık hizmetleri sektöründe, veri madenciliğinin yararlı olduğu kanıtlanmış farklı alanlar mevcuttur. Bu alanlar müşteri ilişkilerinin iyileştirilmesi, öngörücü tıbbın geliştirilmesi, dolandırıcılık ve ihmalin tespit edilmesi ve genel sağlık yönetimini de kapsamakta olup, bazıları, sistem içerisinde belli öğelerin doğruluğunu kesin olarak belirlemeye yöneliktir.

Özellikle sağlık sektöründe tek tek verimliliklerin artırılması, hasta bakımının artırılmasında ve daha fazla hastanın yaşamı kurtarılarak maliyetlerin düşürülmesinde veri madenciliğinin işe yarar olduğu görülmektedir.

Son yıllarda, sağlık hizmetleri sektöründe; (i) kronik hastalıkların tanımlanması ve analiz edilmesi, (ii) etkili tedaviler için olası nedenlerin ve ilaçların tespit edilmesi ve ortaya çıkarılması, (iii) hastalığın yayılmaya eğilimli olduğu yüksek riskli bölgelerin takip edilmesi, (iii) hastalığın yayılmasını azaltmaya yönelik programların tasarlanması, (iv) hastaların rejyonlarının (bölgelerinin) tanımlanması gibi amaçlarla veri madenciliğinden yararlanılmaktadır.

Sağlık hizmetleri sektöründe, görüntüleme ve laboratuvar testlerine ait veriler ve raporlar; kümeleme, sınıflandırma, ilişkilendirme ve aykırı değer algılama gibi veri madenciliği görevleri ile analiz edilir. Kronik hastalık semptomlarının, bu hastalıkların olası nedenlerinin ve tedavilerinin tanımlanması, tespit edilmesi ve öngörülmesi için bu görevlerden yararlanılır. Bu şekilde, söz konusu hastalıklar için etkin tedavi yöntemleri mümkün olabilir. Hastalığın yayılma ihtimalinin yüksek olduğu riskli rejyonların belirlenmesi ve takip edilmesi adına, analiz daha da genişletilebilir. Bu tip bir analizden yola çıkarak; hastalıklara ve önlemlerine dair toplumu bilgilendiren bölgesel kampanyalar tasarlanabilir. Hastalık semptomlarının, hastalığa yol açan nedenlerin ve tedavide kullanılan ilaçların sürekli karşılaştırılmasını mümkün kılan veri madenciliği sayesinde, etkin tedavi yöntemlerini ve bunların beraberinde getirdiği yan etkileri tespit etmeye yönelik veri analizleri gerçekleştirilebilir [96-97].

2.4.5. Dolandırıcılık tespiti ve suç önleme

Aykırılıklar veya aykırı değerler, büyük miktarda veri işlenerek yapılan veri madenciliği yoluyla da tespit edilebilir. Aykırı değerler, verilerdeki seyrek kalıpların ortaya çıkarılmasıyla belirlenebilir. Yaygın olmayan örüntüler genellikle sahtekarlık/suç faaliyetine aittir. Dolayısıyla, aykırı değer tespiti ve/veya yaygın örüntü madenciliği yardımı ile olası dolandırıcılıklar tespit ve tahmin edilmekte, böylelikte suçların ortaya çıkışı önlenebilmektedir [98].

2.4.6. Müşteri İlişkileri Yönetimi

Başarılı müşteri ilişkileri yönetimi, daha uygun müşterilere hitap edilmesi ve müşterilerin daha iyi elde tutulması ile elde edilebilir. Veri madenciliği; (i) veritabanı pazarlamasının, (ii) müşteri edinme ve müşteri koruma kampanyalarının tanımlanması ve tahmini yoluyla müşteri ilişki yönetimini güçlendirebilir [99-100].

2.4.7. Tavsiye sistemleri

Tavsiye sistemleri, paydaşlara veri madenciliğinden yararlanan kullanıcıların ilgisini çekebilecek çeşitli öneriler sunar. Bu sistemler, kullanıcı için bir öğeyi tahmin etmeye yönelik kullanıcı işlemlerini, kullanıcı profillerini, anahtar kelimeleri, öğeler arasındaki ortak özellikleri inceler. Tavsiye sistemlerinde; makine öğrenimi, istatistik, bilgi alma gibi birçok veri madenciliği tekniği kullanılmaktadır. Örneğin pazarlama alanında, tavsiye sistemi; kullanıcı tarafından geçmişte sorgulanan ürünlerle benzer olan ürünlerin veya zevklerin doğrultusunda, kullanıcının zevkleri ile benzerlik taşıyan diğer müşterilerin tercihlerine bakarak, ürün önerisinde bulunabilmektedir [95].

2.4.8 Çevrimiçi pazarlama / E-ticaret

Çevrimiçi pazarlama ve e-ticaret alanlarında faaliyet gösteren çeşitli büyük markalar/satıcılar da işlerini büyütmek için veri madenciliğinden yararlanmaktadır. Örneğin: (i) e-ticaret satıcıları, web üzerinde metin madenciliği kullanarak ürünün en düşük fiyatını keşfeder, (ii) büyük fast-food zinciri satıcıları, müşteri deneyimlerini geliştirmek için veri madenciliği kullanarak müşterilerin sipariş modelleri, bekleme süreleri, sipariş büyüklükleri gibi unsurları inceler, (iii) çevrimiçi medya servis sağlayıcıları ayrıca bir dizi veya filmi müşteriler arasında nasıl popüler hale getireceklerini öğrenmek için veri madenciliğinden yararlanırlar [95].

3. YÖNTEM

Veri madenciliği yöntemleri; Birliktelik kuralları, sınıflandırma ve tahmin, kümeleme analizi ve aykırılık analizi olarak 4 farklı yönteme bölünmektedir.

1.Birliktelik Kuralları: Bu yöntem, veri madenciliğinde en iyi bilinenlerdendir. Büyük veri tabanlarındaki ilişkili değişkenleri ve bu değişkenlerin aralarındaki ilişkinin büyüklüğünü belilemek için başvurulan bir metoddur.

2.Sınıflandırma ve Tahmin: Bir gözlemin niteliklerinin gelecekteki veri yönelimlerinin açıklanması amacıyla incelenmesi ve önceden belirlenmiş bir sınıfa bu gözlemin atanmasıdır. Rastgele orman, Lojistik regresyon ve Navie Bayes sınıflandırmada kullanılan temel algoritmalardandır.

3.Kümeleme Analizi: Kümeleme yöntemindeki amaç; işlenebilir duruma getirmek için dağılmış haldeki verilerin özellikleri doğrultusunda birleştirilmesidir. Sınıflandırmadan farkı kümeler önceden saptanmamıştır. Kümelemede yaygın olarak kullanılan algoritmalara K-Metoids ile K-Means örnek olarak verilebilir.

4.Aykırılık Analizi: Verilerdeki aykırı değerlerin ve uç değerlerin, algoritmalar ile denetim yapılarak ortaya çıkarılması sürecidir. Bu tarz olağan dışı veriler kayıt etme, okuma ve ölçüm gibi hatalardan dolayı meydana gelmektedir. Veri madenciliğindeki algoritmalar, bu olağandışı verileri tamamen ortadan kaldırmayı ya da en aza indirmeyi hedeflemektedir.

Bu çalışmada amaç veri madenciliğinde en sık kullanılan yöntemlerden sınıflandırma yöntemlerini incelemek ve bu sınıflandırma algoritmalarını kullanarak WEKA yazılımı aracılığı ile üroloji branşı verilerinin analizini yapmak ve tıp doktorlarının da kullanabileceği faydalı sonuçlar elde etmektir.

3.1. Sınıflandırma

Önemli sınıfları tanımlayan modelleri çıkarmak veya gelecekteki veri eğilimlerini tahmin etmek için kullanılabilecek iki veri analiz biçimi vardır. Bu iki yöntem aşağıdaki gibidir:

- Sınıflandırma
- Tahmin

Sınıflandırma modelleri, kategorik sınıf etiketlerini öngörür ve tahmin modelleri sürekli önemli fonksiyonları tahmin eder. Örneğin, banka kredisi uygulamalarını güvenli veya riskli olarak sınıflandırmak için bir sınıflandırma modeli veya gelir ve

meslekleri göz önüne alındığında potansiyel müşterilerin bilgisayar donanımı üzerindeki harcamalarını dolar cinsinden tahmin etmek için bir tahmin modeli oluşturabiliriz [1].

Burada önemli faktör, algoritmaların başarısını en iyi şekilde ölçmektir. Yaratılan modelin başarı oranını ölçmek için birçok faktör kullanılabilir. Veri madenciliği araştırmalarında en doğru modelin hangisi olduğunu anlamak için doğruluk, kesinlik ve duyarlılık gibi ölçütlerden faydalanılır. Yalnızca doğruluk kriteri kullanılarak model seçiminin yapılması bilimsel çalışmalarda yeterli görülmemektedir. Doğruluk ölçütü ile birlikte duyarlılık ve kesinlik ölçütlerinin de modele karar verme sürecinde dikkate alınması gerekir [1].

Bu bölümde, makine öğrenme algoritmalarının başarılarını belirlemede kullanılan ölçütlerden bahsedilecektir. İlk olarak doğruluk, duyarlılık ve kesinliğin hesaplanmasını anlayabilmek için hata(karışıklık) matrisinden bahsetmek gerekir [1]. Çizelge 3.1'de hata matrisi verilmiştir.

Çizelge 3 1 Hata(Karışıklık) Matrisi

	Tahmin Edilen Değer	
Gerçek Değer	Doğru Pozitif (DP)	Yanlış Negatif (YN)
	Yanlış Pozitif (YP)	Doğru Negatif (DN)

Model tarafından doğru olarak tahmin edilen gözlemler Doğru Pozitif ile Doğru Negatif kısımlarında, yanlış tahmin edilenler ise Yanlış Pozitif ve Yanlış Negatif altında toplanır. Bu terimlerin açıklamalarını inceleyecek olursak:

Doğru Pozitif: Bunlar gerçek değeri 1 ve tahmin ettiğimiz değerin de 1 olduğu örneklerdir.

Doğru Negatif: Bunlar gerçek değeri 0 ve tahmin ettiğimiz değerin de 0 olduğu örneklerdir.

Yanlış Pozitif: Bunlar gerçek değeri 0 ancak tahmin ettiğimiz değerin 1 olduğu örneklerdir.

Yanlış Negatif: Bunlar gerçek değeri 1 ancak tahmin ettiğimiz değerin 0 olduğu örneklerdir.

Modelin başarısını ölçmek için doğruluk oranı çok kullanılan bir ölçüttür. Doğruluk oranı, modelin doğru tahminlerinin oranını belirler. Örneğin; 100 maçtan meydana gelen bir veri setinde 40 maçın sonucu doğru tahmin edilir ise, doğruluk oranımız %40 olur. Doğruluk ölçütü, eşit dağılım göstermeyen veri kümelerinde yalnız başına hiçbir zaman yeterli olmamaktadır. Mesela; 100 hastadan oluşan bir verimiz olsun ve bu kişilerden bazıları kanser hastası olsun ve bazıları da kanser hastası olmasın. Gerçekten kanser hastası olanlardan yalnızca 10'nuna kanser teşhisi konulsun. Bu ise aslında kanser olduğu halde teşhis edilemeyen kişilerin varlığını gösterir ve bu arzu edilmeyecek bir durumdur. Bu nedenle modellerin doğruluğunu belirlemede diğer ölçütler de kullanılmalıdır [1].

Aşağıda belirtilen ölçütler ile algoritmaların başarısı ölçülebilir.

- Ortalama Kök Kare Hata (Root mean squared error)
- Doğru Sınıflandırma Oranı (Correct classification rate/Accuracy)
- Kesinlik değeri (Precision)
- Kappa İstatistik Değeri (Kappa statistic)
- Ortalama Mutlak Hata (Mean absolute error)
- ROC Alan Değeri (ROC Area)
- Duyarlılık (Recall)

Bu bölümde algoritmaların sınıflandırma performanslarını değerlendirme ölçütleri hakkında bilgi verilecektir.

Bu ölçütlerin hesaplanmasında Çizelge 3.1'de verilen hata(karışıklık) matrisi kullanılır. Hata matrisi, tahminlerin doğruluğu hakkında bilgi veren 2x2'lik bir matristir. Sınıflandırma sonucu oluşan modelin gerçek değerleri ile tahmin edilen değerlerinin karşılaştırılmasına olanak sağlar.

Modellerin performanslarını değerlendirme ölçütleri aşağıdaki gibi verilebilir:

Doğruluk (Accuracy): Tüm doğru tahmin edilen değerlerin, bütün sonuçlara olan oranıdır [1].

$$\text{Doğruluk} = DP+DN \,/\, (DP+DN+YN+YP) \quad (3.1)$$

Duyarlılık (Sensitivity): Duyarlılık (DP Oranı) doğru tahmin edilen pozitif sınıf değerlerinin, tüm pozitif sınıf değerlerine oranıdır. Pozitif kararın doğru olma olasılığıdır [1].

$$\text{Duyarlılık} = \text{DP oranı} = DP / (DP + YN) \quad (3.2)$$

Özgüllük (Specificity): Doğru tahmin edilen negatif sınıf değerlerinin, tüm negatif sınıf değerlerine oranıdır. Negatif kararın doğru olma olasılığını gösterir [1].

$$\text{Özgüllük} = DN / (DN + YP) \quad (3.3)$$

Yanlış pozitiflerin oranı olan *YP Oranı= 1- Özgüllük* şeklinde hesaplanır.

ROC (Receiver Operating Characteristics): Bir tanı testine ait özgüllük ve duyarlılık değerleri arasındaki bağıntı için ROC eğrisi grafiksel gösterim sağlar. Doğru pozitif orana yani duyarlılığa karşın yanlış pozitif oranların (1-özgüllük) noktalanarak çizilmesiyle ROC eğrisi oluşturulur. DP oranı ile YP oranı arasındaki ilişkinin grafiksel olarak gösterimidir. Grafikte eğri altında kalan alan (AUC) değeri 1'e yaklaştıkça tanı değeri yükselmektedir. ROC için en iyi değer 1'dir, 0,5 değeri ise sınıflandırmanın başarılı olmadığını gösterir. ROC değerinin 1 olması, veride hiçbir yanlış tahmin olmadığı anlamına gelir.

MCC (Matthews Correlation Coefficient): Bu kriter, iki sınıflı veriler için diğer ölçütlere kıyasla daha doğru sonuçlar vermektedir. Dengeli dağılım göstermeyen veri kümelerinde dahi en doğru sonucu verir. MCC'nin formülü (3.4)'de verilmiştir. MCC, −1 ve +1 arasında değişen bir korelasyon katsayısı olup, gerçek sınıf ve tahmin edilen ikili sınıflandırmalar arasındaki korelasyonu gösterir. Rasgele sınıflandırma olduğunda kriter 0 değerini verir, 1 değeri elde edildiğinde mevcut gerçek değerler ile sınıflandırıcının verdiği kararların tamamen birbirine ters olduğu anlamına gelir. +1 değeri ise sınıflandırma başarısının tam doğru olduğu anlamına gelir. Hesaplanan değerin 1'e yakın olması doğru bir sınıflandırma yapıldığını göstermektedir[1]:

$$MCC = (DP \times DN) - (YP \times YN) / \sqrt{(DP+YP)\times(DP+YN)\times(DN+YP)\times(DN+YN)} \quad (3.4)$$

Duyarlılık (Recall): Doğru tahmin edilen pozitif sınıf değerinin, bütün gerçek pozitif sınıf değerlerine oranıdır.

$$\text{Duyarlılık}=DP/(DP+YN) \quad (3.5)$$

NOT: Geri çağırma ve duyarlılık (DP Oranı) aynı değerleri vermektedir. Uygulama bölümündeki algoritmalar için elde edilen tablolarda bu değer görülmektedir.

Kesinlik (Precision): Doğru tahmin edilen pozitif sınıf değerinin, bütün pozitif olarak tahmin edilen sınıf değerlerine oranıdır [1].

$$\text{Kesinlik} = DP / (DP+YP) \quad (3.6)$$

PRC (Precision-recall Curve): İki önemli performans ölçütü olan kesinlik ve duyarlılık arasındaki bağıntı bu performans ölçütü ile incelenir. Bu iki değer her bir kesim noktasında hesaplanarak bir eğri oluşturulur. Eğri oluşturulurken Y ekseninde kesinlik ve X ekseninde duyarlılık değerleri yer alır. PRC eğrisi bu iki değerin kesişiminden elde edilen noktaların birleştirilmesi ile oluşturulur. 1'e yakın PCR değerleri tahminin doğruluğunu gösterir [1].

F-Ölçütü (F-Measure): Duyarlılık ve kesinlik karşılaştırma kriterleri, yalnız başlarına anlamlı bir karşılaştırma sonucu çıkarmamıza yetmeyebilir. Bu durumda, bu iki kriteri birlikte değerlendirmek daha doğru sonuçlara götürebilir. F-ölçütü bu amaçla tanımlanmıştır. Bu ölçüt, duyarlılık ve kesinliğin harmonik ortalamasıdır.

$$\text{F-Ölçütü} = (2 \times \text{Duyarlılık} \times \text{Kesinlik})/(\text{Duyarlılık}+\text{Kesinlik}) \quad (3.7)$$

Kesinlik, duyarlılık ve F-ölçüt değerleri 1' e yaklaştıkça modelin başarısı artar.

4. BULGULAR

Veri madenciliğinin tıp bilişiminde kullanımı güncel konulardan biridir ve giderek daha da önemini arttırmaktadır. Veri madenciliği, yapay zekanın en dikkat çeken uygulamalarındandır ve sistemlerin planlı bir deneyimi olmadığı halde deneyimden otomatik olarak öğrenmesini ve ilerlemesini sağlar. Veri madenciliği'nde, verilere kendileri için ulaşabilen ve bu verileri öğrenebilen bilgisayar programları geliştirmek üzerinde odaklanılır. Veri madenciliği algoritmaları ile bir tahmin modeli oluşturulabilir, örneğin bu tezde veri madenciliği algoritmaları üroloji hastalıklarını tahmin etmek için kullanılmıştır.

Veri madenciliği üç ana bölümden oluşmaktadır: Veri, veriye ait özellikler ve algoritmalar. Veri madenciliği'nde kullanılan çok fazla algoritma mecuttur ve her geçen gün de yeni algoritmalar eklenmektedir. Bu tez çalışmasında sınıflandırma algoritmaları içinde yaygın olarak bilinen sekiz algoritma incelenmiştir.

4.1. Veri Seti Hakkında Bilgiler

Analizde kullanılacak veri setinden bu bölümde bahsedilecektir. Bu tez çalışmasında, üroloji branşına ait dört farklı hastaneden alınan, etik ilkelerine uygun, 1985 hastaya ilişkin verilerle WEKA programında sınıflandırma işlemleri yapılmıştır. Veriler, ilk olarak hastanelerden toplanarak veri birleştirme, veri temizleme ve önişleme süreçlerinde işleme tutulmuşlardır. Veriler üzerinde bir takım düzenlemeler (ön işlemler) yapılarak model kurmaya hazır hale getirilmiştir. Bu işlemlerin ardından 175909 veriden 1985 hastaya ilişkin veri ortaya çıkmış olup analize hazır hale getirilmiştir. Sınıflandırma işlemine geçmeden önce kullanılan veri setine ilişkin bilgiler verilecektir:

Bu uygulamada 5 farklı aşama vardır:

- Veri setini toplamak
- Verileri biçimlendirmek
- Modeli oluşturmak
- En iyi modeli belirlemek
- En iyi modeli üroloji hastalıklarını tahmin etmek için kullanmak

Bu çalışmanın amacı yaş, cinsiyet, lokasyon ve semptom bilgilerine dayanarak üroloji hastalıklarını teşhis etmektir.

4.1.1.Kullanılan Veri Seti ve Veri Setinin Düzenlenmesi

Kullanılan veri seti 18 adet teşhis ana grubunu bulundurmaktadır. Bu teşhislere ilişkin 45 adet semptom bulunmaktadır. Hastalara ilişkin doğru teşhis koymada hastalarda görülen semptomlar önemli bir bilgi içermektedir.

Yaş verileri 6 kategoriye ayrılmış olup 1'den 6'ya kadar etiketlenmiştir.

Cinsiyet verileri 2 kategoriye ayrılmış olup, erkek 1 ve kız 2 olarak etiketleme yapılmıştır.

Konum verileri 3 kategoriye ayrılmıştır: 0: Konum Bilinmiyor, 1: Kentsel ve 2: Kırsal olmak üzere etiketleme yapılmıştır. Aşağıda Çizelge 4.1.'de veri seti detaylı bir şekilde anlatılmıştır.

Verinin temizlenmesi ve dönüşümünden sonraki aşama modelleme aşamasıdır. Sınıflandırma için kullanılan modeller veri seti üzerinde uygulanarak doğruluğu en yüksek olan model seçilecektir.

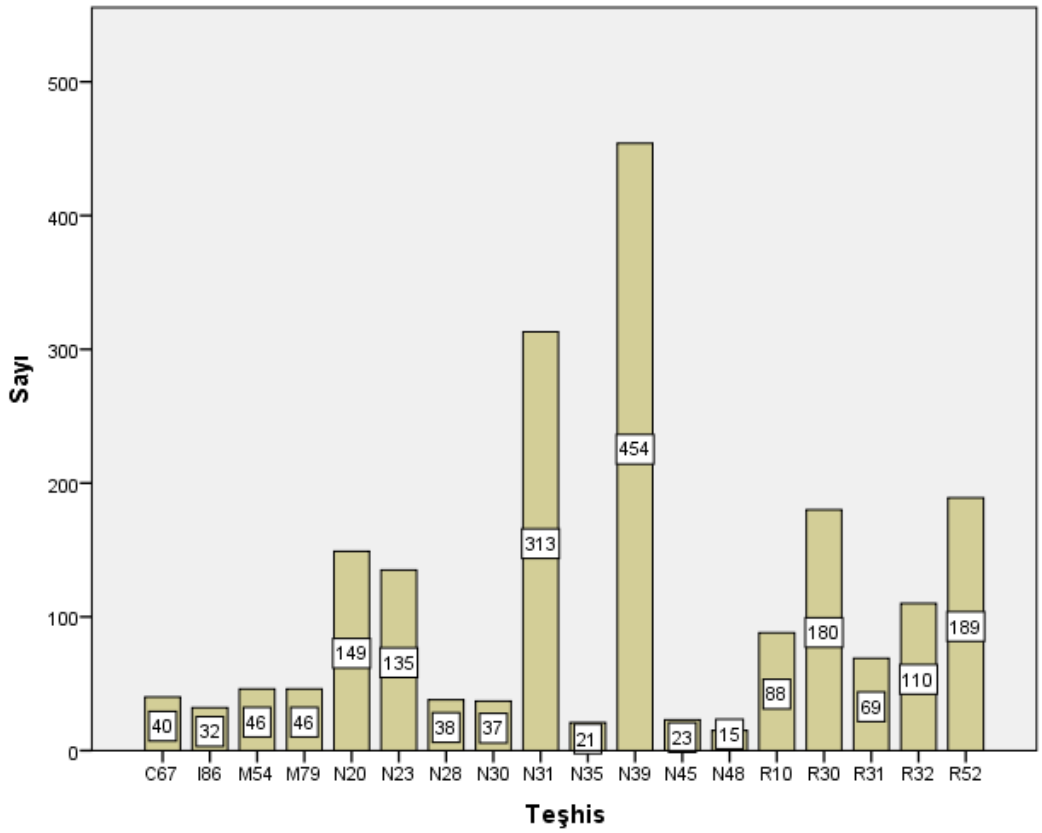
Çizelge 4. 1 Veri seti değişkenleri

Yaş	Cinsiyet	Lokasyon	Semptom	Teşhis
1:0-10 çocukluk	1: Erkek	0:Konum bilinmiyor	45 adet semptom bulunmaktadır.	18 adet teşhis bulunmaktadır.
2:11-24 gençlik çağı	2: Kadın	1:Kentsel		
3:25-40 yetişkinlik çağı		2: Kırsal		
4: 41-64				
5:65-79 yaşlılık dönemi				
6: 80-üstü				

Çizelge 4.2'de 18 adet teşhisin görüldüğü hasta sayısı, bu teşhisleri gösteren teşhis numaraları ve teşhislerin neler olduğuna ilişkin bilgiler verilmektedir.

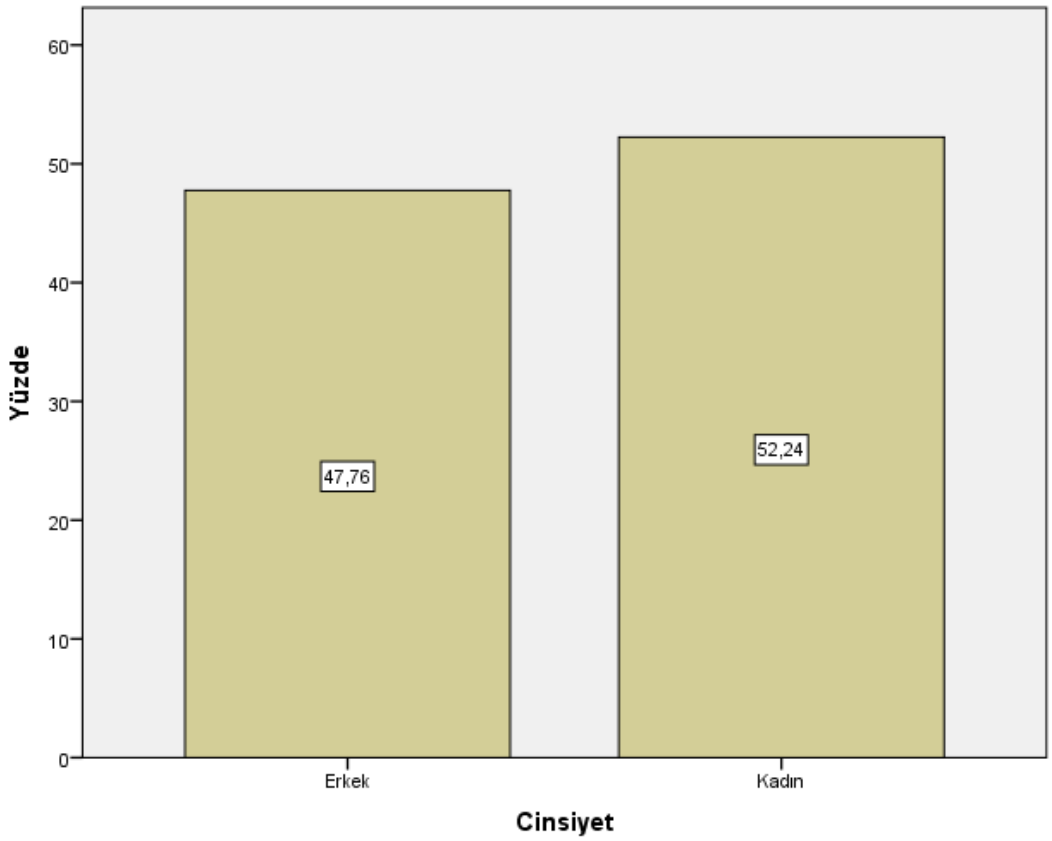
Çizelge 4. 2 Teşhis listesi

NO	SAYI	TESHIS	TESHIS ADI
1	135	N23	Renal kolik, tanımlanmamış
2	149	N20	Böbrek ve üreter taşı
3	454	N39	Üriner sistemin diğer bozuklukları
4	37	N30	Sistit
5	21	N35	Üretra darlığı
6	69	R31	Hematüri, tanımlanmamış
7	313	N31	Mesanenin nöromusküler fonksiyon bozukluğu, başka yerde sınıflanmamış
8	40	C67	Mesane malign neoplazmı
9	32	I86	Variköz venler, diğer yerlerin
10	110	R32	Üriner inkontinans, tanımlanmamış
11	88	R10	Abdominal ve pelvik ağrı
12	180	R30	İşemeyle birlikte ağrı
13	38	N28	Böbrek ve üreterin diğer bozuklukları, başka yerde sınıflanmamış
14	46	M79	Diğer yumuşak doku bozuklukları, başka yerde sınıflanmamış
15	46	M54	Dorsalji
16	15	N48	Penisin diğer bozuklukları
17	189	R52	Ağrı, başka yerde sınıflanmamış
18	23	N45	Orşit ve epididimit



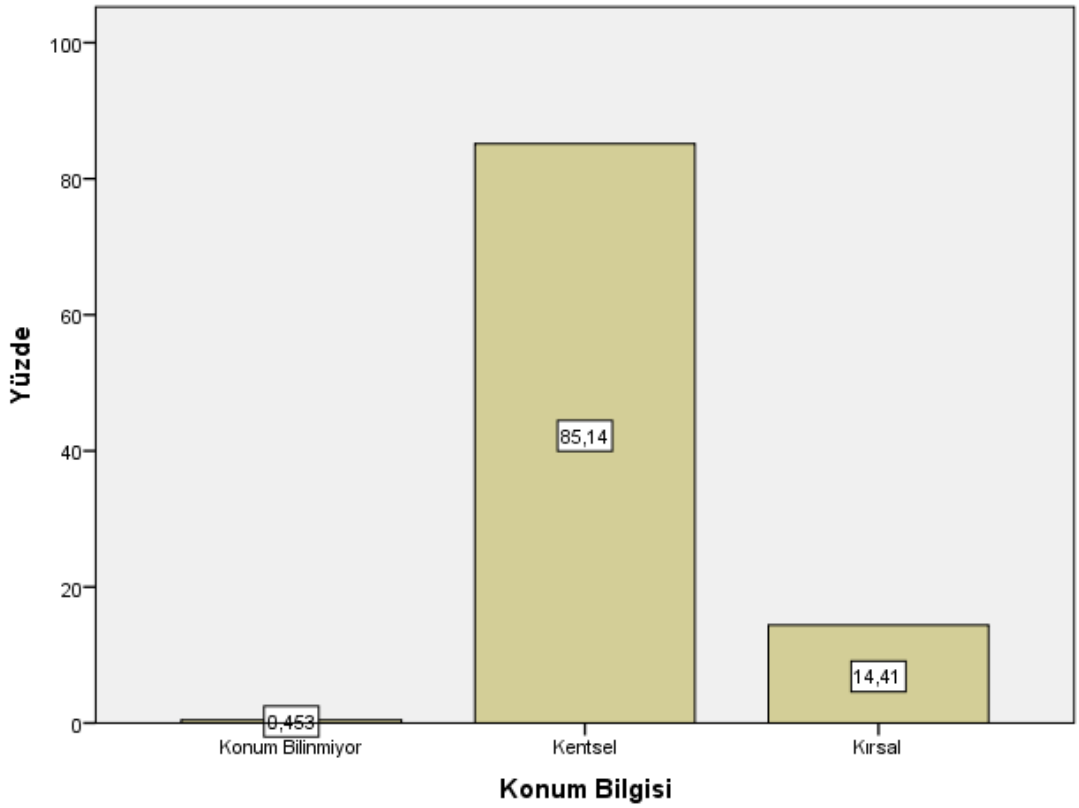
Şekil 4. 1 Hastaların teşhis bilgilerinin frekans dağılım grafiği

Şekil 4.1'de hastaların teşhis bilgileri frekans dağılım grafiği verilmiştir. Şekil 4.1 ve Çizelge 4.2 birlikte incelendiğinde "N39 Üriner sistemin diğer bozuklukları" teşhisinin en sık görülen teşhis olduğu görülmektedir. Hastaların 454 (%22,87) tanesine bu teşhis konmuştur.



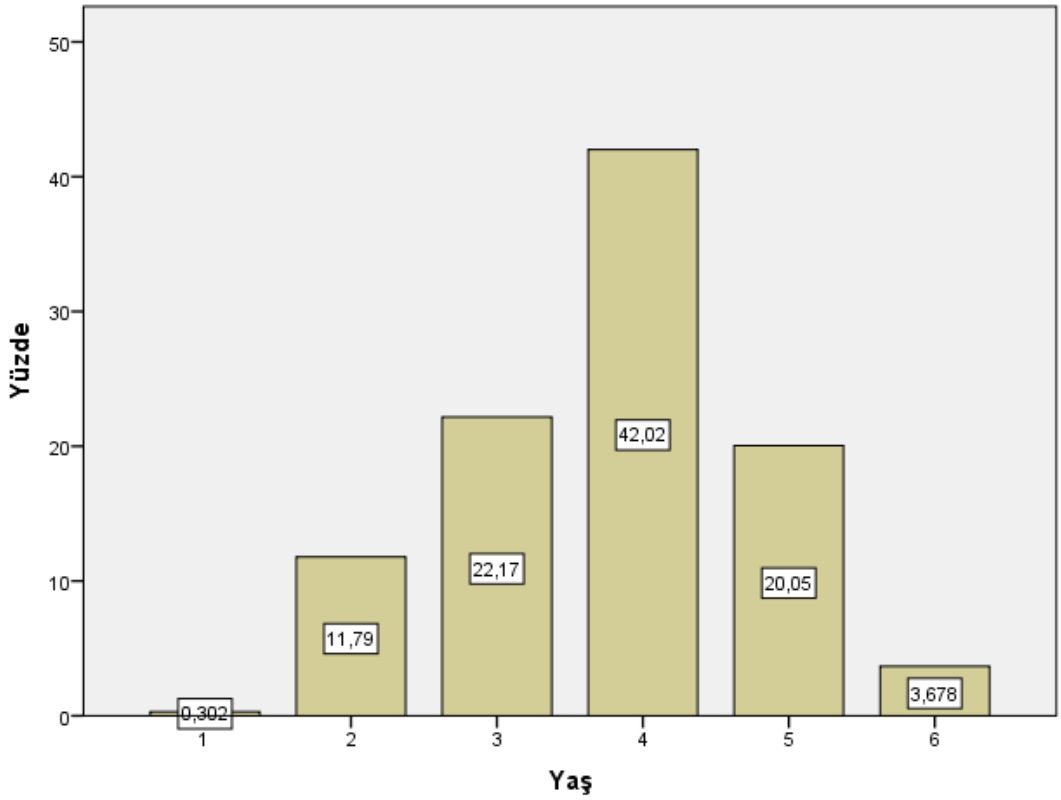
Şekil 4. 2 Hastaların cinsiyetinin frekans dağılım grafiği

Şekil 4.2'de hastaların cinsiyetinin frekans dağılım grafiği verilmiştir. Hastaların %47,76'sını erkek ve %52,24'ünü kadınlar oluşturmaktadır. Sayılara bakıldığında kadınların sayısının daha fazla olduğu görülmektedir.



Şekil 4. 3 Hastaların konumunun frekans dağılım grafiği

Şekil 4.3'de hastaların konumlarının frekans dağılım grafiği verilmiştir. Hastaların büyük çoğunluğunun kentlerde yaşadığı görülmektedir. Konum bilinmiyor, kentsel ve kırsal olarak grafik dağılmaktadır.



Şekil 4. 4 Hastaların yaş frekans dağılım grafiği

Şekil 4.4'de hastaların yaş frekans dağılım grafiği verilmiştir. Şekil 4.4'de görüldüğü gibi; çalışmaya katılan 1985 hastanın %0,3'ü 0-10, %11,79'u 11-24, %22,17'si 25-40, %42,02'si 41-64, %20,05'i 65-79 ve %3,68'i 80 ve üzeri yaş aralığındaki hastalardır. Özet olarak, yaş dağılımına baktığımız zaman 6 gruba ayrılmıştır.0-10 aralığı en az görülen teşhis grubudur.41-64 yaş aralığı ise en çok teşhisin görüldüğü gruptur.

4.2. Analiz Sonuçları

Veri madenciliğinde yaygın kullanılan yöntemlerden biri sınıflandırmadır. Veri setindeki herbir gözlemin bir seri özelliği mecuttur ve bu özelliklerden biri de sınıf bilgisidir. Bir öğrenme algoritmasına dayanan sınıflandırmada, sınıf bilgisi bilinen eğitim kümesindeki gözlemler öğrenme amacıyla kullanılarak bir model meydana getirilir. Bu oluşturulmuş modelin başarısı, eğitim kümesinde bulunmayan gözlemler (test kümesi) ile denenerek ölçülür. Bu model kullanılarak sınıf bilgisi bilinmeyen bir gözleme ilişkin bir sınıf belirlenebilir. Verinin temizlenmesi ve dönüşümünden sonraki aşama modellemedir. En yüksek doğruluğa sahip model, değişik modeller veri seti üzerinde denendikten sonra seçilir.

Bu bölümde sınıflandırma algoritmalarının sonuçları ayrı ayrı olarak verilmiştir. Bu algoritmalar ile elde edilen modelleri karşılaştırmak için performans ölçme kriterleri kullanılmıştır. Bu çalışmada 10-katlı bağımsız geçerlilik sınaması yöntemi kullanılmıştır.

Sınıflandırma algoritmalarından Rasgele ağaç, Rastgele orman, Çok katmanlı algılayıcılar, K-en yakın komşu, Kstar, Lojistik regresyon, ZeroR ve Naive Bayes algoritmaları ile analizler yapılmıştır. Bu bölümde her bir algoritma ile sınıflandırma analizleri yapılmış olup çıkan sonuçlar sonraki alt bölümlerde verilmiştir. Daha sonra en başarılı model için elde edilen sonuçlara ilişkin yorumlar detaylı olarak verilmiştir.

4.2.1. Rasgele Ağaç (Random Tree) Algoritması Sonuçları

Çizelge 4.3'de görüldüğü gibi, Rasgele ağaç algoritması ortalama % 65,44 doğruluk ile hastalığı teşhis edebilir. Tabloda doğruluk oranına ek olarak, ortalama mutlak hata, ortalama kök kare hatası, kappa istatistiği, bağıl mutlak hata, kök göreceli kare hatası, kesinlik, duyarlılık, F ölçütü ve ROC alan değeri gibi diğer ölçütlere de yer verilmiştir. Bu değerleri elde etmemiz için 10-katlı bağımsız geçerlilik sınaması yöntemi kullanılmıştır.

Çizelge 4. 3 Rasgele Ağaç algoritması sonuçları

Doğru Sınıflandırma Oranı	% 65,4408
Kappa istatistiği	0,6105
Ortalama Mutlak Hata	0,0399
Ortalama Kök Kare Hatası	0,185
Bağıl Mutlak Hata	% 40,4648
Kök Göreceli Kare Hatası	% 83,3568
Kesinlik	0,654
Duyarlılık	0,654
F-Ölçütü	0,653
ROC Alan Değeri	0,837

Çizelge 4.3'de görülebileceği gibi, Rasgele ağaç algoritmasının doğru sınıflandırma oranı %65,44 olarak belirlenmiştir. Çizelge'de doğruluk oranıyla beraber, duyarlılık ve kesinlik gibi başka ölçütler de verilmiştir. Rasgele ağaç algoritması %65,4 kesinlik ve % 65,4 duyarlılık oranlarına sahiptir.

4.2.2. Rastgele Orman (Random Forest) Algoritması Sonuçları

Çizelge 4.4'de görüldüğü gibi, Rastgele orman algoritması ortalama %73,2 doğruluk ile hastalığı teşhis edebilir. Tabloda doğruluk oranına ek olarak, ortalama mutlak hata, ortalama kök kare hatası, kappa istatistiği, bağıl mutlak hata, kök göreceli kare hatası, kesinlik, duyarlılık, F ölçütü ve ROC alan değeri gibi diğer ölçütlere de yer verilmiştir. Bu değerleri elde etmemiz için 10-katlı bağımsız geçerlilik sınaması yöntemi kullanılmıştır.

Çizelge 4. 4 Rastgele Orman algoritması sonuçları

Doğru Sınıflandırma Oranı	%73,2
Kappa istatistiği	0,6963
Ortalama Mutlak Hata	0,0435
Ortalama Kök Kare Hatası	0,1458
Bağıl Mutlak Hata	%44,1458
Kök Göreceli Kare Hatası	%65,6856
Kesinlik	0,728
Duyarlılık	0,732
F-Ölçütü	0,727
ROC Alan Değeri	0,950

Çizelge 4.4'de görülebileceği gibi, Rastgele orman algoritmasının doğru sınıflandırma oranı %73,2 olarak belirlenmiştir. Çizelge'de doğruluk oranıyla beraber, duyarlılık ve kesinlik gibi başka ölçütler de verilmiştir. Rastgele orman algoritması %72,8 kesinlik ve %73,2 duyarlılık oranlarına sahiptir.

4.2.3. ZeroR Algoritması Sonuçları

Çizelge 4.5'de görüldüğü gibi, ZeroR algoritması ortalama %22,87 doğruluk ile hastalığı teşhis edebilir. Tabloda doğruluk oranına ek olarak, ortalama mutlak hata, ortalama kök kare hatası, kappa istatistiği, bağıl mutlak hata, kök göreceli kare hatası, kesinlik, duyarlılık, F ölçütü ve ROC alan değeri gibi diğer ölçütlere de yer verilmiştir. Bu değerleri elde etmemiz için 10-katlı bağımsız geçerlilik sınaması yöntemi kullanılmıştır.

Çizelge 4. 5 ZeroR algoritması sonuçları

Doğru Sınıflandırma Oranı	%22,8715
Kappa istatistiği	0
Ortalama Mutlak Hata	0,0985
Ortalama Kök Kare Hatası	0,2219
Bağıl Mutlak Hata	%100
Kök Göreceli Kare Hatası	%100
Kesinlik	-
Duyarlılık	0,229
F-Ölçütü	-
ROC Alan Değeri	0,493

Çizelge 4.5'de görüldüğü gibi, ZeroR algoritmasının doğru sınıflandırma oranı %22,87 olarak belirlenmiştir. Çizelge'de doğruluk oranıyla beraber, duyarlılık ve kesinlik gibi başka ölçütler de verilmiştir. ZeroR algoritması %22,9 duyarlılık ve %49,3 ROC alan değeri oranlarına sahiptir.

4.2.4. KStar Algoritması Sonuçları

Çizelge 4.6'da görüldüğü gibi, KStar algoritması ortalama %70,93 doğruluk ile hastalığı teşhis edebilir. Tabloda doğruluk oranına ek olarak, ortalama mutlak hata, ortalama kök kare hatası, kappa istatistiği, bağıl mutlak hata, kök göreceli kare hatası, kesinlik, duyarlılık, F ölçütü ve ROC alan değeri gibi diğer ölçütlere de yer verilmiştir. Bu değerleri elde etmemiz için 10-katlı bağımsız geçerlilik sınaması yöntemi kullanılmıştır.

Çizelge 4. 6 KStar algoritması sonuçları

Doğru Sınıflandırma Oranı	%70,932
Kappa istatistiği	0,6706
Ortalama Mutlak Hata	0,0434
Ortalama Kök Kare Hatası	0,1522
Bağıl Mutlak Hata	% 44,0688
Kök Göreceli Kare Hatası	% 68,6068
Kesinlik	0,703
Duyarlılık	0,709
F-Ölçütü	0,702
ROC Alan Değeri	0,950

Çizelge 4.6'da görüldüğü gibi, KStar algoritmasının doğru sınıflandırma oranı %70,93 olarak belirlenmiştir. Çizelge'de doğruluk oranıyla beraber, duyarlılık ve kesinlik gibi başka ölçütler de verilmiştir. KStar algoritması %70,3 kesinlik ve %70,9 duyarlılık oranlarına sahiptir.

4.2.5. K-En Yakın Komşu (IBK) Algoritması Sonuçları

Çizelge 4.7'de görüldüğü gibi, k-en yakın komşu algoritması ortalama %69,12 doğruluk ile hastalığı teşhis edebilir. Tabloda doğruluk oranına ek olarak, ortalama mutlak hata, ortalama kök kare hatası, kappa istatistiği, bağıl mutlak hata, kök göreceli kare hatası, kesinlik, duyarlılık, F ölçütü ve ROC alan değeri gibi diğer ölçütlere de yer verilmiştir. Bu değerleri elde etmemiz için 10-katlı bağımsız geçerlilik sınaması yöntemi kullanılmıştır.

Çizelge 4. 7 K-En Yakın Komşu Algoritması sonuçları

Doğru Sınıflandırma Oranı	%69,1184
Kappa istatistiği	0,6513
Ortalama Mutlak Hata	0,0383
Ortalama Kök Kare Hatası	0,1648
Bağıl Mutlak Hata	%38,8949
Kök Göreceli Kare Hatası	%74,2764
Kesinlik	0,694
Duyarlılık	0,691
F-Ölçütü	0,691
ROC Alan Değeri	0,896

Çizelge 4.7'de görüldüğü gibi, k-en yakın komşu algoritmasının doğru sınıflandırma oranı %69,12 olarak belirlenmiştir. Çizelge'de doğruluk oranıyla beraber, duyarlılık ve kesinlik gibi başka ölçütler de verilmiştir. K-en yakın komşu algoritması %69,4 kesinlik ve %69,1 duyarlılık oranlarına sahiptir.

4.2.6. Çok Katmanlı Algılayıcılar (MultilayerPerception) Algoritması Sonuçları

Çizelge 4.8'de görüldüğü gibi, Çok katmanlı algılayıcılar algoritması ortalama %72,19 doğruluk ile hastalığı teşhis edebilir. Tabloda doğruluk oranına ek olarak, ortalama mutlak hata, ortalama kök kare hatası, kappa istatistiği, bağıl mutlak hata, kök göreceli kare hatası, kesinlik, duyarlılık, F ölçütü ve ROC alan değeri gibi diğer ölçütlere de yer verilmiştir. Bu değerleri elde etmemiz için 10-katlı bağımsız geçerlilik sınaması yöntemi kullanılmıştır.

Çizelge 4. 8 Çok Katmanlı Algılayıcılar algoritması sonuçları

Doğru Sınıflandırma Oranı	%72,1914
Kappa istatistiği	0,6862
Ortalama Mutlak Hata	0,0339
Ortalama Kök Kare Hatası	0,16
Bağıl Mutlak Hata	%34,3806
Kök Göreceli Kare Hatası	%72,1128
Kesinlik	0,720
Duyarlılık	0,722
F-Ölçütü	0,720
ROC Alan Değeri	0,945

Çizelge 4.8'de görüldüğü gibi, çok katmanlı algılayıcılar algoritmasının doğru sınıflandırma oranı %72,19 olarak belirlenmiştir. Çizelge'de doğruluk oranıyla beraber, duyarlılık ve kesinlik gibi başka ölçütler de verilmiştir. Çok katmanlı algılayıcılar algoritması %72 kesinlik ve %72,2 duyarlılık oranlarına sahiptir.

4.2.7. NaiveBayes Algoritması Sonuçları

Çizelge 4.9'da görüldüğü gibi, Naive Bayes algoritması ortalama %66,22 doğruluk ile hastalığı teşhis edebilir. Tabloda doğruluk oranına ek olarak, ortalama mutlak hata, ortalama kök kare hatası, kappa istatistiği, bağıl mutlak hata, kök göreceli kare hatası, kesinlik, duyarlılık, F ölçütü ve ROC alan değeri gibi diğer ölçütlere de yer verilmiştir. Bu değerleri elde etmemiz için 10-katlı bağımsız geçerlilik sınaması yöntemi kullanılmıştır.

Çizelge 4. 9 Naive Bayes algoritması sonuçları

Doğru Sınıflandırma Oranı	%66,2222
Kappa istatistiği	0,6151
Ortalama Mutlak Hata	0,0434
Ortalama Kök Kare Hatası	0,1631
Bağıl Mutlak Hata	%44,0316
Kök Göreceli Kare Hatası	%73,5376
Kesinlik	-
Duyarlılık	0,662
F-Ölçütü	-
ROC Alan Değeri	0,952

Çizelge 4.9'da görüldüğü gibi, Naive Bayes algoritmasının doğru sınıflandırma oranı %66,22 olarak belirlenmiştir. Çizelge'de doğruluk oranıyla beraber, duyarlılık ve kesinlik gibi başka ölçütler de verilmiştir. Naive Bayes algoritması %66,2 duyarlılık ve %95,2 ROC alan değeri oranlarına sahiptir.

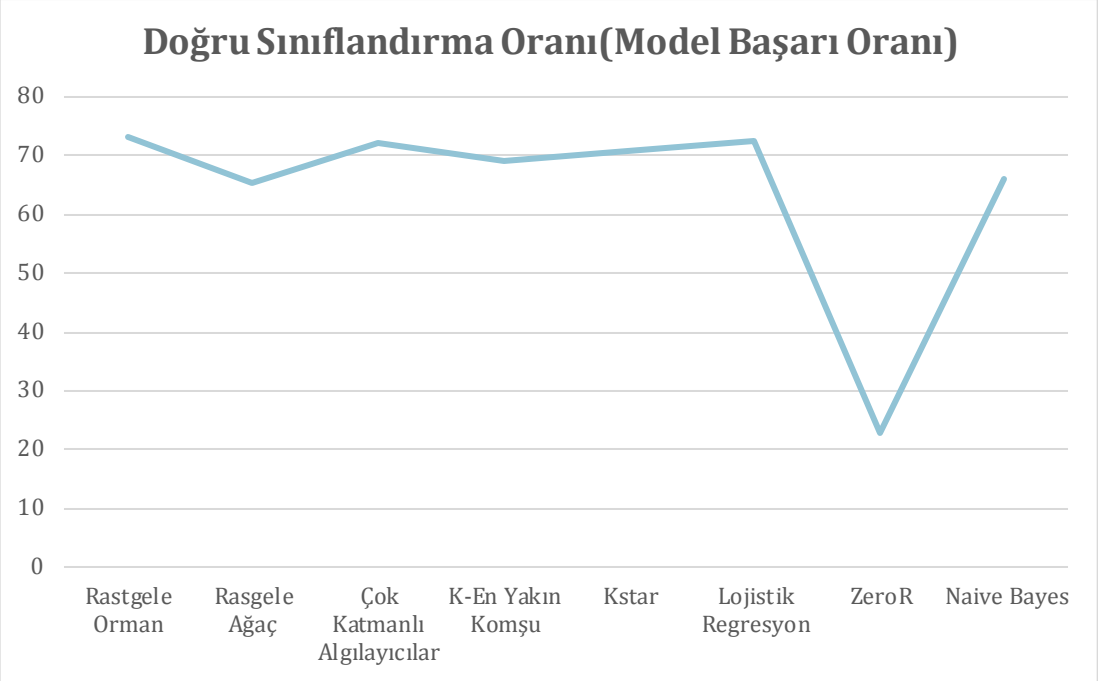
4.2.8. Lojistik Regresyon (Simple Logistic) Sonuçları

Çizelge 4.10'da görüldüğü gibi, Lojistik regresyon analizi ile ortalama %72,6 doğruluk ile hastalık teşhis edilebilir. Tabloda doğruluk oranına ek olarak, ortalama mutlak hata, ortalama kök kare hatası, kappa istatistiği, bağıl mutlak hata, kök göreceli kare hatası, kesinlik, duyarlılık, F ölçütü ve ROC alan değeri gibi diğer ölçütlere de yer verilmiştir. Bu değerleri elde etmemiz için 10-katlı bağımsız geçerlilik sınaması yöntemi kullanılmıştır.

Çizelge 4. 10 Lojistik Regresyon sonuçları

Doğru Sınıflandırma Oranı	%72,6
Kappa istatistiği	0,6908
Ortalama Mutlak Hata	0,0393
Ortalama Kök Kare Hatası	0,1443
Bağıl Mutlak Hata	%39,91
Kök Göreceli Kare Hatası	%65,02
Kesinlik	0,727
Duyarlılık	0,726
F-Ölçütü	0,724
ROC Alan Değeri	0,964

Çizelge 4.10'da görüldüğü gibi, Lojistik regresyon analizinin doğru sınıflandırma oranı %72,6 olarak belirlenmiştir. Çizelge'de doğruluk oranıyla beraber, duyarlılık ve kesinlik gibi başka ölçütler de verilmiştir. Lojistik regresyon analizi sonuçları %72,7 kesinlik ve %72,6 duyarlılık oranlarına sahiptir.



Şekil 4. 5 Uygulanan algoritmaların doğru sınıflandırma oranı

Uygulanan algoritmların doğru sınıflandırma oranı (model başarı oranı) Şekil 4.5'de verilmiştir. Şekil incelendiğinde, ZeroR algoritmasının en düşük performansa sahip olduğu görülmektedir. Rastgele orman, Çok katmanlı algılayıcılar ve Lojistik regresyon algoritmalarının birbirine yakın performansları ile öne çıktığı söylenebilir.

Çizelge 4. 11 Uygulanan algoritmaların detaylı sonuçları

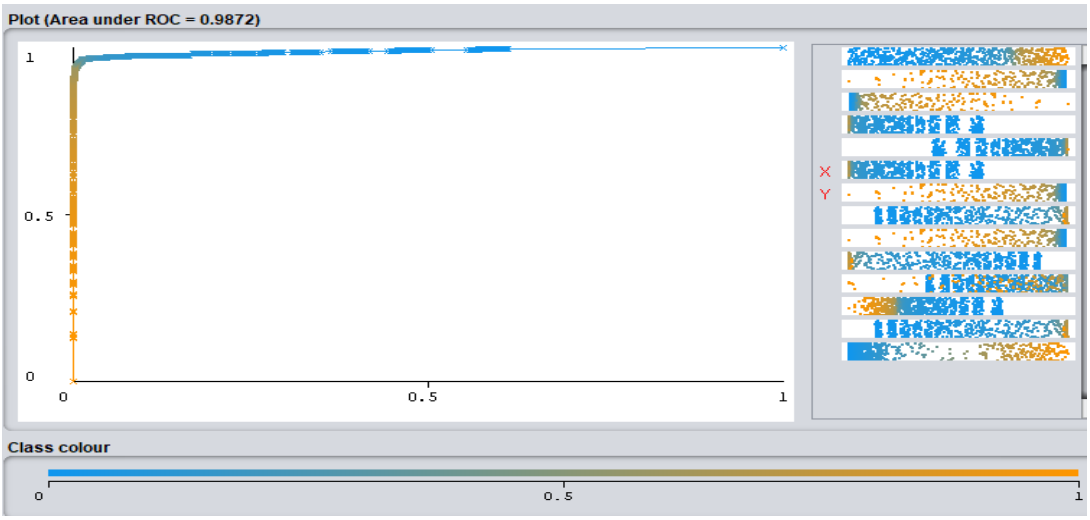
	Rastgele Orman	Rasgele Ağaç	Çok Katmanlı Algılayıcılar	K-En Yakın Komşu	Kstar	Lojistik Regresyon	ZeroR	Naive Bayes
Doğru Sınıflandırma Oranı (Model Başarı oranı)	%73,2	% 65,4	%72,191	%69	%71	%72,59	%22,9	%66,22
Kesinlik değeri	0,728	0,654	0,720	0,694	0,703	0,727	-	-
Kappa İstatistik Değeri	0,6963	0,6105	0,6862	0,651	0,671	0,6908	0	0,6151
Ortalama Mutlak Hata	0,0435	0,0399	0,0339	0,038	0,044	0,0393	0,099	0,0434
Ortalama Kök Kare Hatası	0,1458	0,185	0,16	0,165	0,152	0,1443	0,222	0,1631
ROC Alan Değeri	0,950	0,837	0,945	0,896	0,950	0,964	0,493	0,952
F Ölçütü	0,727	0,653	0,720	0,691	0,702	0,724	-	-

Weka paket programında sırasıyla Rastgele orman, Rasgele ağaç, Çok katmanlı algılayıcılar, K-en yakın komşu, KStar, Lojistik Regresyon, ZeroR ve NaiveBayes algoritmaları uygulanmış ve elde edilen sonuçlar Çizelge 4.11'de görülmektedir.

Çizelge 4.11'de görüldüğü gibi en yüksek başarı oranını veren algoritma %73,2 oranıyla Rastgele orman algoritması olmuştur. Bu algoritmayı takiben en başarılı algoritmalar %72,59 başarı oranıyla Lojistik regresyon ve %72,2'lik başarı oranıyla Çok katmanlı algılayıcılar algoritması olmuştur.

Kappa, beklenen doğruluk ile gözlenen doğruluğu karşılaştırmaya yarayan, mevcut sınıflara yapılan atamalar arasında uyuşma olup olmadığını gösteren bir istatik değeridir. 0,6963 oranıyla Rastgele orman algoritması ve 0,6908 oranıyla Lojistik regresyon algoritmaları Kappa değerleri açısından öne çıkan algoritmalardır. Bunları Çok katmanlı algılayıcılar algoritması takip etmektedir. Kappa değerinin 0,61 – 0,80 arasında olması önemli derecede uyuşma olduğunu göstermektedir.[106]

Ayrıca model başarısı değerlendirme kriterlerinden ROC alan değerinin, kesinlik, f ölçütü ve duyarlılık değerlerinin de 1'e yakın olması istenir. Çizelge 4.11'e bakıldığı zaman ZeroR algoritması dışındakilerinin ROC alan değerlerinin 1'e genelde oldukça yakın olduğu görülmektedir. En çok görülen teşhis "N39 Üriner sistemin diğer bozuklukları" için en başarılı algoritma olan Rastgele orman algoritmasının ROC eğrisi çizimi Şekil 4.6'da verilmiştir.



Şekil 4. 6 Rastgele orman algoritmasının N39 teşhisi için ROC eğrisi çizimi

5.SONUÇ

Günümüzde verideki artıştan ötürü bilgiye ulaşmada yaşanan sıkıntılardan ötürü veri madenciliği ortaya cıkmıştır. Bilgisayar programları olmadan veri madenciliği yöntemleri uygulanamaz. Karar ağaçları, veri kümeleme, apriori yöntemi, bayes sınıflandırıcılar vb. çeşitli algoritmalar bu programlarda vardır. Elde edilen verilerden algoritmalar uygulanarak bilgi elde edinimi sağlanır. Bu algoritmaları uygulamak için çeşitli programlar mevcuttur. Bu tez çalışmasında açık kaynak kodlu WEKA programı kullanılmıştır. Günümüzde WEKA erişime açık ve pratik bir program olduğu için yaygın bir biçimde tercih edilmektedir.

Veri madenciliğinin en popüler uygulamalarına son yıllarda özellikle sağlıkta ve tıpta rastlanmaktadır. Büyük boyutlardaki veriden, gizlenmiş, faydalı, kullanılabilir bilgileri elde etmek ve stratejik kararlar vermede veri madenciliği kullanılır. Veri madenciliği, verilerin analizine dayalı karar verme modelleri oluşturan bir yöntemdir. Bu nedenle veri madenciliğinin, karar vermede destekçi olan bir yöntem olarak sağlık hizmetleri sunarken, sağlık kurumlarını yönetirken ve sağlık politikaları oluştururken kullanılmasıyla, sağlıkçıların doğru ve yararlı kararlar vermelerine destek sağlanır. Literatürde başta kanser ve kalp hastalıkları olmak üzere birçok hastalık teşhisine yönelik yapılan tez ve makaleler mevcuttur.

Bu tez çalışmasında, hastanelerin üroloji branşına başvuran hastalardan elde edilen gerçek bir veri seti üzerinde, WEKA veri madenciliği yazılımı ile 8 adet sınıflandırma algoritması uygulanmıştır. Böylece, 18 farklı hastalık için hangi algoritma ya da algoritmaların sınıflandırmada en iyi sonuç verdiği bulunmaya çalışılmıştır. 1985 hasta için en çok görülen hastalık teşhisi N39 Üriner sistemin diğer bozukluklarıdır. Analizdeki hastaların çoğunluğunu kadınlar oluşturmakta, ancak, kadın ve erkek sayısı arasında çok büyük bir fark yoktur. Bu hastalardan elde edinilen bilgiler ışığında 41-64 yaş aralığındaki kişilerde üroloji ile ilgili rahatsızlıkların daha sık görüldüğü gözlenmiştir.

Oluşturulan modellerin doğru sınıflandırma oranlarına bakıldığında %73,2 ile Rastgele orman algoritmasının en yüksek sonucu verdiği, sırasıyla %72,59 başarı oranıyla Lojistik regresyon ve %72,2 oranıyla Çok katmanlı algılayıcılar algoritmalarının başarılı olduğu görülmüştür.

ZeroR algoritması ise %22,87'lik doğru sınıflama oranıyla en düşük doğru sınıflandırma başarısı gösteren algoritma olmuştur. ZeroR algoritması için elde edilen sonuç literatür ile uyuşmaktadır. Bu algoritma diğer algoritmalar ile kıyaslandığında genellikle daha başarısız sonuçlar vermektedir.

Rastgele orman algoritmasına ilişkin her bir teşhis için elde edilen analiz sonuçlarına ve karışıklık matrisine bakıldığında; algoritmanın başta en sık görülen teşhis N39 olmak üzere diğer teşhislerin birçoğu için de başarılı doğru sınıflandırma sonuçları verdiği görülmüştür.

Model sonuçlarında en başarılı algoritma çıkan Rastgele orman algoritması ile onun en yakın takipçileri Lojistik regresyon ve Çok katmanlı algılayıcılar algoritmaları kullanılarak; ilerleyen çalışmalarda üroloji hastalıklarının teşhisine yönelik bir uygulama geliştirilebilir. Böylece sağlık çalışanlarına teşhiste fikir vermesi ve iş yüklerinin azaltılması, erken teşhis ile hastalıkların önceden bulunarak tedavi sürelerinin daha kısa olması mümkün olabilir.

Benzer çalışmalar farklı tıp branşındaki hastalardan elde edilen veri kümeleri üzerinde de yapılabilir ve farklı veri madenciliği araçlarında ve farklı algoritmalar kullanılarak karşılaştırmalar genişletilebilir. Sağlık yöneticilerinin, sağlık kurumlarının verimli, daha etkin ve kaliteli yönetilmesi amacıyla eldeki verilerden en başarılı şekilde faydalanan ve karar süreçlerine yardım sağlayacak sistemlere ihtiyaçları vardır. Günümüzde sağlık alanında yapay zekâ, sağlık uzmanlarının en hatasız ve yeni bilgiyi elde etmesini, en ideal ve objektif çözümler üretmesine yol açacak karar vermede destek bir araç halini almaktadır.

6. KAYNAKLAR

[1] Silahtaroğlu, Gökhan. "Veri madenciliği." Papatya Yayınları, İstanbul (2008).

[2] Sayad, S., Data Mining. Erişim Tarihi: 29. 08. 2018. http://chem eng.utoronto.ca/~datamining/dmc/data_mining.htm (2018).

[3] Bardak, T. and Sözen, E., 2018, October. Veri madenciliği ve Önemi. 6th ASM International Congress of Agriculture and Environment, Proceeding Book. 2018

[4] Savaş, Serkan, Nurettin Topaloğlu, and Mithat Yılmaz. "Veri madenciliği ve Türkiye'deki uygulama örnekleri." (2012).

[5] Gorunescu, Florin. Data Mining: Concepts, models and techniques. Vol. 12. Springer Science & Business Media, 2011.

[6] Emre, İlkim Ecem, and Çiğdem Selçukcan Erol. "Veri Analizinde İstatistik mi Veri Madenciliği mi?." Internatıonal Journal Of Informatıcs Technologıes 10.2 (2017):161.

[7] Mohammed Abdul Khalid, Sateesh kumar Pradhan, G.N.Dash, F.A.Mazarbhuiya, A survey of data mining techniques on medical data for finding temporally frequent diseases", International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2013.

[8] S.D.Gheware, A.S.Kejkar, S.M.Tondare, Data Mining: Task, Tools, Techniques and Applications, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 10, October 2014.

[9] Yongjian Fu, Data Mining: Tasks, Techniques and Applications http://academic.csuohio.edu/fuy/Pub/pot97.pdf

[10] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2002.

[11] G. Beller, J. Nucl. Cardiol. "The rising cost of health care in the United States: is it making the United States globally noncompetitive?" vol. 15, no. 4, pp. 481-482, 2008.

[12] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2005.

[13] Gosain, A.; Kumar, A., "Analysis of health care data using different data mining techniques," Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009, International Conference on, vol. no., pp.1,6, 22-24 July 2009.

[14] "Regresyon Analizi." Wikipedia, Wikimedia Foundation, 16 Dec. 2020, tr.wikipedia.org/wiki/Regresyon_analizi.

[15] A. S. Elmaghraby, et al. Data Mining from multimedia patient records. 6, 2006.

[16] Nada Lavrac, BlažZupan, "Data Mining in Medicine" in Data Mining and Knowledge Discovery Handbook, 2005.

[17] Soni J, Ansari U, Sharma D, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, March 2011.

[18] Naren Ramakrishnan, David Hanauer, Benjamin J. Keller, Mining Electronic Health Records, IEEE Computer 43(10): 77-81, 2010.

[19] O. Mary K, Mat, "Applications of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, August 2004.

[20] Hian Chye K, Gerald T, Data mining applications in healthcare, Journal of healthcare information management: JHIM.19 (2): 64-72, (2005).

[21] A. Milley, "Healthcare and data mining", Health Management Technology, vol. 21, no. 8, pp. 44-47, 2000.

[22] Gaynes R, Richards C, Edwards J, et al. Feeding back surveillance data to prevent hospital-acquired infections. Emerg Infect Dis 2001;7:2 95- 298, 2001.

[23] Brosette SE, Spragre AP, Jones WT, Moser SA. A data mining system for infection control surveillance. Methods Inf Med,39: 303-310, 2000.

[24] M. Ridinger, "American Healthways uses SAS to improve patient care", DM Review, vol. 12, no.139, 2002.

[25] M. Durairaj, V.Ranjani, Data mining applications in healthcare sector: A Study, International Journal Of Scientific & Technology Research Volume 2, Issue 10, ISSN 2277-8616, October 2013.

[26] Anonymous. Texas Medicaid Fraud and Abuse Detection System recovers $2.2 million, wins national award. Health Management Technology, vol. 20, no. 10, 1999.

[27] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, 2005.

[28] B. K. Schuerenberg, "An information excavation", Health Data Management, vol. 11, no. 6, pp. 80-82, 2003.

[29] P. D. Haghighi et. al., Mobile Data Mining for Intelligent Healthcare Support, IEEE xplore, 2009.

[30] Neelamadhab Padhy, Pragnyaban Mishra and Rasmita Panigrahi, The survey of data mining applications and feature scope, Asian Journal Of Computer Science And Information Technology 2: 4, 68– 77, 2012.

[31] Wiig, K, Knowledge Management: An Emerging Discipline Rooted in a Long History Knowledge Management (pp. 352): ButterworthHeinemann, 1999.

[32] Stankosky M, Creating the Discipline of Knowledge Management: Butterworth-Heinemann, 2005.

[33] Chen H and Chau M. "Web Mining: Machine Learning for Web Applications," Annual Review of Information Science and Technology, 38, 289-329, 2004.

[34] Chen C-J & Huang J-W. Strategic human resource practices and innovation performance - The mediating role of knowledge management capacity. Journal of Business Research 62: 104-114, 2009.

[35] Fugate BS, Stank TP & Mentzer JT. Linking improved knowledge management to operational and organizational performance. Journal of Operations Management in Press, Corrected Proof, 2008.

[36] Orzano A.J, McInerney CR, Scharf D, Tallia AF & Crabtree BF. A knowledge management model: Implications for enhancing quality in health care. Journal of the American Society for Information Science & Technology 59: 489-505, 2008.

[37] Christo El Morr and Julien Subercaze, Knowledge Management in Health care, DOI: 10.4018/978-1-61520-670-4.ch023, pp 490-510.

[38] Wilson T. D, The nonsense of knowledge management, Information Research, 8(1), 2002.

[39] Zand D.E. The Leadership Triad: Knowledge, Trust and Power, New York: Oxford University Press, 1997.

[40] Sussman S. W & Siegal W. S, Informational Influence in Organizations: An Integrated Approach to Knowledge Adoption, Information Systems Research, 14(1), 47-65, 2003.

[41] Fayyad U. M, Piatetsky-Shapiro G and Smyth P, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, 17(3), 37-54, 1996.

[42] Dunham M. H, Data Mining: Introductory and Advanced Topics, New Jersey, USA: Prentice Hall, 2002.

[43] Jihoon Yang and Vasant Honavar, Feature subset selection using Genetic Algorithm, IEEE Intelligent Systems, 1998.

[44] Hany M. Harb, Abeer S. Desuky, Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization, International Journal of Computer Applications (0975 – 8887) Volume 104 – No.5, October 2014.

[45] G. Ravi Kumar, Dr. G.A.Ramachandra, K.Nagamani, An Efficient Feature Selection System to Integrating DVM with Genetic Algorithm for Large Medical Datasets International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014.

[46] V.Sangeetha, J.Preethi, M.Sreeshakthy, Survey on Medical Data Cluster analysis using Feature Selection and Neural Networks, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 11, November 2014.

[47] Megha Aggarwal, Amrita, Performance Analysis Of Different Feature Selection Methods In Intrusion Detection, International Journal Of Scientific & Technology Research Volume 2, Issue 6, June 2013.

[48] Anderson J. A and Davis J., An introduction to neural networks, MIT, Cambride, 1995.

[49] Obenshain M. K, Application of data mining techniques to healthcare data Infect. Control Hosp. Epidemiol, 25(8):690–695, 2004.

[50] M. H. Dunham, “Data mining introductory and advanced topics”, Upper Saddle River, NJ: Pearson Education, Inc., 2003.

[51] A. Shameem Fathima, D. Manimegalai and Nisar Hundewale, A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, ISSN (Online): 1694- 0814, November 2011.

[52] K. Usha Rani, Analysis Of Heart Diseases Dataset Using Neural Network Approach, International Journal Of Data Mining &Knowledge Management Process (Ijdkp) Vol.1, No.5, September 2011.

[53] Haykin. S, Neural Networks: A Comprehensive Foundation, Prentice Hall, 1999.

[54] Emina Alickovic and Abdulhamit Subasi, Data Mining Techniques for Medical Data Classification, The International Arab Conference on Information Technology (ACIT), 2011.

[55] S. Anto, Dr.S.Chandramathi, Supervised Machine Learning Approaches for Medical Data Set Classification - A Review, IJCST Vol. 2, Issue 4, Oct - Dec 2011.

[56] Goharian & Grossman, Data Mining Classification, Illinois Institute of Technology, 2003. http://ir.iit.edu/~nazli/cs422/CS422-Slides/DMClassification.

[57] Apte & S.M. Weiss, Data Mining with Decision Trees and Decision Rules, T.J. Watson Research Center, http://www.research.ibm.com/dar/papers/pdf/fgcsapteweissue_with_cov er.pdf, 1997.

[58] V.Gayathri, M.Chanda Mona, S.Banu Chitra, A survey of data mining techniques on medical diagnosis and research. V.Gayathri, M.Chanda Mona, S.Banu Chitra, International Journal of Data Engineering (IJDE) Singaporean Journal of Scientific Research (SJSR) Vol.6.No.6 2014.

[59] L.A. Zadeh, “Some reflection on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent system”, Soft computing, vol. 2, 1998.

[60] Kalyani Mali & Samayita Bhattacharya., Soft computing on Medical - Data (SCOM) for a Countrywide Medical System using Data Mining and Cloud

Computing Features, Global Journal of Computer Science and Technology Cloud and Distributed, Volume 13 Issue 3 Version 1.0 Year 2013

[61] V. Vapnik, "Statistical Learning Theory", Wiley, ISBN: 978-0-471- 03003-4, 1998.

[62] V. Vapnik, "The support vector method of function estimation", AT & T Labs – Research, John Wiley and Sons, New York, USA, 1998.

[63] N. Cristianini and J.Shawe-taylor, An Introduction To Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, 1995.

[64] Hammer, B. and Gersmann, K., "A Note On theUniversal Approximation Capability of SupportVector Machines", Neural Process Lett 17, pp. 1061 - 1085, 2003.

[65] Vapnik, V.N., "The Nature of Statistical LearningTheory", Springer, New York, 2005.

[66] N. Chistianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines, and other kernel-based learning methods", Cambridge University Press, 2000.

[67] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines", Cambridge University Press, 2000.

[68] Argyro Kampouraki, Christophoros Nikou, George Manis, "Robustness of Support Vector Machine-based Classification of Heart Rate Signals", Proceedings of the 28th IEEE, EMBS Annual International Conference, New York, USA, Aug30-sep 3,2006, 1995.

[69] Samjin Choi, "Detection of valvular heart disorders using wavelet packet decomposition and support vector machine, Elsevier", Expert Systems with Applications, 35, pp 1679-1687, 2008.

[70] Ilias Maglogiannis, Euripidis Loukis, Elias Zafiropoulos, Antonis Stasis, "Support vector machine based identification of heart valve diseases using heart sounds", Elsevier, Computer Methods and Programs in Biomedicine ,95, pp. 47-61, 2009.

[71] Friedman N., Geiger, D.Goldszmidt M, "Bayesian network classifiers. Machine Learning 29: pp. 131-163, 1997.

[72] Friedman N., Koller D, "Being Bayesian About Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks", Machine Learning 50(1): pp. 95-125, 2003.

[73] Finn V. Jensen, An Introduction to Bayesian Networks, Springer, New York, 1996.

[74] Sebe N., Ira Cohen, Ashutosh Garg and Thomas Huang S. "Machine Learning in Computer Vision", Springer, Netherlands, pp. 130-133, 2005.

[75] Ankita Agarwal, "Secret Key Encryption algorithm using genetic algorithm", vol.-2, no.-4, ISSN: 2277 128X, IJARCSSE, pp. 57-61, April 2012.

[76] Jihoon Yang and Vasant Honavar. Feature subset selection using Genetic Algorithm. IEEE Intelligent Systems, 1998.

[77] Sang Jun Lee, Keng Siau, A review of data mining techniques, Industrial Management & Data Systems, 101/1, MCB University Press [ISSN 0263-5577], 2001.

[78] George D., Magoulas and Andriana Prentza. Machine Learning in Medical Applications, Proceeding machine learning and its applications: Advance lectures, pp. 300-307, 2001.

[79] Quinlan J.R, "Induction of decision trees", Machine Learning, 1, 1, 81- 106, 1986.

[80] Goldberg D, Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, 1989.

[81] Rumelhart D.E and Mc Clelland, J.L. (eds.), Parallel Distributed Processing, Vol. 1: Foundations. MIT Press, Cambridge, MA: MIT Press, 1986.

[82] Bourlas Ph, Sgouros N, Papakonstantinou G and Tsanakas P, "Towards a knowledge acquisition and management system for ECG diagnosis", In Proceedings of 13th International Congress Medical Informatics EuropeMIE96, Copenhagen, 1996.

[83] Zupan B., Halter J.A and Bohanec M., “Qualitative model approach to computer assisted reasoning in physiology”, In Proceedings of Intelligent Data Analysis in Medicine and Pharmacology-IDAMAP98, Brighton, UK, 1998.

[84] Akay Y.M, Akay M, Welkowitz W and Kostis J.B, "Noninvasive detection of coronary artery disease using wavelet-based fuzzy neural networks", IEEE Engineering in Medicine and Biology, 761-764, 1994.

[85] Coppini G, Poli R and Valli G, "Recovery Of The 3-D Shape Of The Left Ventricle From Echo Cardio Graphic Images", IEEE Transactions on Medical Imaging, 14, 301-317, 1995.

[86] Moustakis V and Charissis G, "Machine Learning And Medical Decision Making". In Proceedings of Workshop on Machine Learning in Medical Applications, Advance Course in Artificial Intelligence- ACAI99, Chania, Greece, 1-19, 1996.

[87] Odabaş, Özlem. Veri Madenciliği Teknikleri Ile Telekom Sektöründe Ayrılan Müşteri Analizi. MS Thesis. İstanbul Ticaret Üniversitesi, 2017.

[88] Oralhan, Burcu. Veri Madenciliği Yaklaşımı İle Telekomunikasyon Sektöründe Arıza Giderme Analizi. Business & Management Studies: An International Journal, 2020, 8.1: 1026-1043.

[89] Akdemir, Çağla. "Hilenin Veri Madenciliği İle Ortaya Çıkartılması ve Perakende Sektöründe Bir Uygulama." 2019.

[90] Namlı, Ersin, and Sümeyra Murat. "Müşteri Odaklı Pazarlama Stratejileri için Veri Madenciliği Teknikleri Kapsamında Perakende Sektöründe Kümeleme Analizi Uygulaması." Gümüshane University Electronic Journal of the Institute of Social Science/Gümüshane Üniversitesi Sosyal Bilimler Enstitüsü Elektronik Dergisi 10. 2019.

[91] Ismaıl, Gürsoy. Pay senedi fiyatlarını etkileyen değişkenlerin C4. 5 karar ağacı algoritması ile modellenmesi. 2019.

[92] Seyrek, İbrahim Halil; ATA, H. Ali. Veri Zarflama Analizi ve Veri Madenciliği ile Mevduat Bankalarında Etkinlik Ölçümü. Journal of BRSA Banking & Financial Markets, 2010, 4.2.

[93] Yakut, Emre; Elmas, Bekir. İşletmelerin Finansal Başarısızlığının Veri Madenciliği Ve Diskriminant Analizi Modelleri İle Tahmin Edilmesi. Journal of Economics & Administrative Sciences/Afyon Kocatepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 2013, 15.1.

[94] Albayrak, Ali Sait; Yılmaz, Sebnem Koltan. Veri Madenciliği: Karar Ağacı Algoritmaları Ve İmkb Verileri Üzerine Bir Uygulama. Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences, 2009, 14.1.

[95] Savaş, Serkan; Topaloğlu, Nurettin; YILMAZ, Mithat. Veri Madenciliği Ve Türkiye'deki Uygulama Örnekleri. 2012.

[96] Koyuncugıl, Ali Serhan; Özgülbaş, Nermin. Veri Madenciliği: Tıp Ve Sağlık Hizmetlerinde Kullanımı Ve Uygulamaları. International Journal Of Informatics Technologies, 2009, 2.2.

[97] Yıldırım, Pınar; Uludağ, Mahmut; GÖRÜR, Abdülkadir. Hastane Bilgi Sistemlerinde Veri Madenciliği. Çanakkale On Sekiz Mart Üniversitesi Akademik Bilişim, 2008.

[98] Orakcı, Merve, et al. Suç Analizinde Veri Madenciliği Teknikleri Ve Makine Öğrenmesi Algoritmalarının Kullanılması. 2019.

[99] Ezerçe, Ayşegül. Müşteri İlişkileri Yönetimi (CRM) ve Veri Madenciliği (Data Mining): Tekstil Sektöründe Bir Uygulama. 2008.

[100] Yiğit, D. Müşteri İlişkileri Yönetimi, Veri Madenciliğinin Müşteri İlişkileri Yönetiminde Kullanımı ve Bir Uygulama. Galatasaray Üniversitesi Sosyal Bilimler Enstitüsü, 2017.

[101] Idrıs, Nashreen Md, et al. Feature Selection And Risk Prediction For Patients With Coronary Artery Disease Using Data Mining. Medical & Biological Engineering & Computing, 2020, 1-18.

[102] Wu, Chien-Ting, et al. Applying Data Mining Techniques for Predicting Prognosis in Patients with Rheumatoid Arthritis. In: Healthcare. Multidisciplinary Digital Publishing Institute, 2020. p. 85.

[103] Van Loo, Hanna M., et al. Data mining algorithm predicts a range of adverse outcomes in major depression. Journal of Affective Disorders, 2020, 276: 945-953.

[104] MOSAYEBI, Alireza, et al. Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. PloS one, 2020, 15.10: e0237658.

[105] Aksu, G. and Doğan, N. "Veri Madenciliğinde Kullanılan Güncel Bir Analiz Programı: WEKA." Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi 10.1 (2019): 80-95.

[106] Bilgen, Ö. B., & Doğan, N. (2017). Puanlayıcılar Arası Güvenirlik Belirleme Tekniklerinin Karşılaştırılması. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8(1), 63-78.

[107] Yılmaz, Hülya. Random Forests Yönteminde Kayıp Veri Probleminin İncelenmesi ve Sağlık Alanında Bir Uygulama. Yüksek Lisans Tezi. Eskişehir Osmangazi Üniversitesi, 2014.

EKLER

EK-I WEKA YAZILIMI HAKKINDA

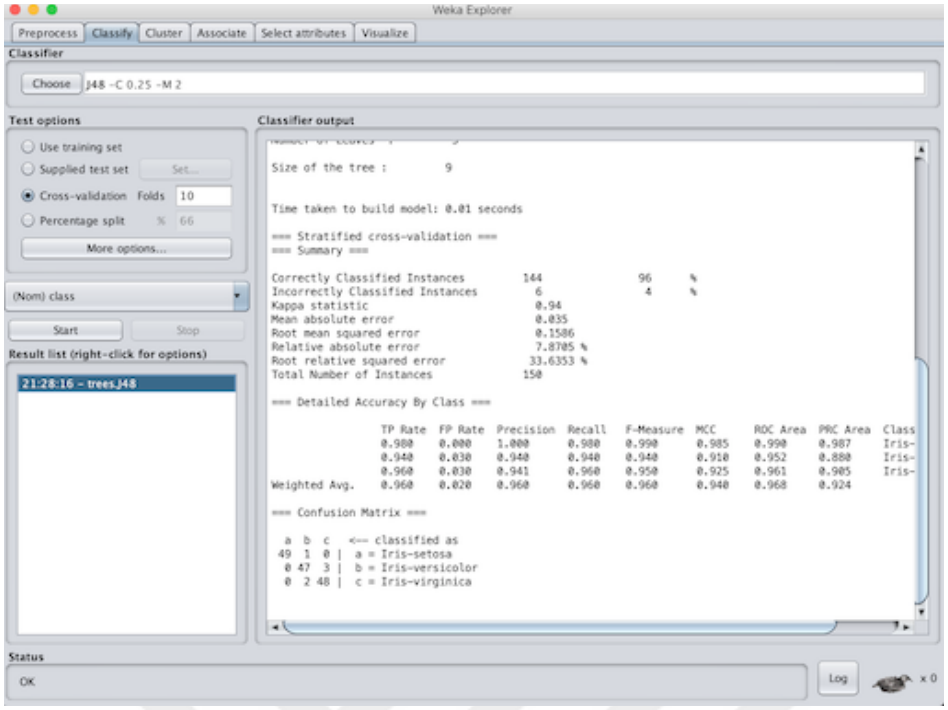
1993 yılında, Yeni Zelanda'daki Waikato Üniversitesi, Weka'nın orijinal versiyonunu geliştirmeye başladı.

Weka, açık kaynaklı bir yazılım, veri ön işleme, çeşitli veri madenciliği ve makine öğrenimi algoritmalarının uygulanması ve görselleştirmesi için araçlar sağlar, böylece makine öğrenimi tekniklerini geliştirebilir ve bunları gerçek dünyadaki veri madenciliği problemlerine uygulayabilirsiniz [105].

Weka, veri madenciliği görevleri için makine öğrenimi algoritmalarından oluşan bir koleksiyondur. Algoritmaları ya kendi Java kodunuz ile çağırabilir ya da direkt bir veri setine uygulanabilirsiniz. Weka'da, sınıflandırma, veri ön işleme, regresyon, kümeleme, görselleştirme ve ilişkilendirme kuralları için araçlar mevcuttur. Ayrıca yeni makine öğrenimi şemaları ortaya çıkarmak için de çok idealdir.

Weka, aşağıdakiler için bir araç koleksiyonudur:

- Regresyon
- Kümeleme
- Bağlantı
- Veri ön işleme
- Sınıflandırma
- Görselleştirme



WEKA yazılımının görüntüsü

Weka'nın avantajları şunlardır:

- GNU Genel Kamu Lisansı altında ücretsiz kullanılabilirlik.
- Taşınabilirlik, tam olarak Java programlama dilinde uygulandığından ve bu nedenle neredeyse tüm modern hesaplama platformlarında çalıştığından.
- Veri ön işleme ve modelleme tekniklerinin kapsamlı bir koleksiyonu.
- Grafik kullanıcı ara yüzleri sayesinde kullanım kolaylığı.

EK-II ALGORİTMALARA İLİŞKİN ÖLÇÜTLERİN PERFORMANS DEĞERLENDİRME SONUÇLARI VE KARIŞIKLIK MATRİSLERİ

Rasgele ağaç algoritmasının sınıflara göre ayrıntılı doğruluk oranı ve diğer ölçütlere ilişkin performans değerlendirme sonuçları bu bölümde sunulmuştur.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,543	0,014	0,481	0,543	0,510	0,499	0,784	0,336	M79
0,472	0,054	0,467	0,472	0,470	0,416	0,715	0,303	R30
0,836	0,017	0,742	0,836	0,786	0,775	0,910	0,656	R32
0,430	0,046	0,406	0,430	0,417	0,374	0,753	0,261	N23
0,474	0,010	0,486	0,474	0,480	0,470	0,728	0,224	N28
0,435	0,015	0,400	0,435	0,417	0,403	0,758	0,227	M54
0,261	0,026	0,319	0,261	0,288	0,259	0,690	0,181	R10
0,863	0,032	0,889	0,863	0,876	0,840	0,918	0,815	N39
0,551	0,013	0,613	0,551	0,580	0,567	0,784	0,439	R31
0,400	0,004	0,429	0,400	0,414	0,410	0,697	0,189	N48
0,436	0,038	0,481	0,436	0,458	0,417	0,769	0,303	N20
0,783	0,003	0,783	0,783	0,783	0,780	0,890	0,643	N45
0,475	0,006	0,613	0,475	0,535	0,531	0,784	0,381	C67
0,750	0,005	0,727	0,750	0,738	0,734	0,873	0,647	I86
0,865	0,006	0,744	0,865	0,800	0,798	0,943	0,701	N30
0,762	0,005	0,615	0,762	0,681	0,681	0,878	0,471	N35
0,815	0,033	0,820	0,815	0,817	0,783	0,909	0,731	N31
0,598	0,052	0,549	0,598	0,572	0,526	0,835	0,396	R52
Weighted Avg. 0,654	0,032	0,654	0,654	0,653	0,621	0,837	0,529	

Rasgele ağaç algoritmasının detaylı karışıklık matrisi bu bölümde sunulmuştur.

```
 a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  <-- classified as
25  3  0  8  1  1  2  3  0  0  1  1  0  0  0  0  0  1 |  a = M79
 1 85  6  2  4  9 12 16  0  3  7  0  1  0  2  3  7 22 |  b = R30
 0  2 92  2  0  0  0  4  0  0  3  0  1  1  1  0  4  0 |  c = R32
12  2  3 58  2  2  9  4  0  0 20  0  0  1  2  0  1 19 |  d = N23
 2  4  1  2 18  0  0  1  0  0  3  0  1  0  0  0  2  4 |  e = N28
 2  2  1  5  1 20 10  0  0  0  1  0  0  0  0  1  2  1 |  f = M54
 0 20  3 11  0 11 23  2  1  2  4  0  2  0  0  1  6  2 |  g = R10
 3 11  6 10  2  0  4 392  0  1  6  0  1  1  6  3  3  5 |  h = N39
 1  1  0  0  0  0  2  1 38  0  1  0  0  0  0  0 24  1 |  i = R31
 0  3  1  0  0  0  0  1  0  6  0  0  1  0  0  0  2  1 |  j = N48
 1 13  1 27  3  1  3  4  1  1 65  0  2  0  0  0  2 25 |  k = N20
 0  0  0  0  0  0  0  1  0  0  1 18  0  1  0  0  0  2 |  l = N45
 2  3  0  1  2  0  0  1  0  0  4  0 19  1  0  2  1  4 |  m = C67
 0  1  0  0  0  0  0  0  0  0  0  1  1  0 24  0  0  0  5 |  n = I86
 0  3  0  0  0  0  1  1  0  0  0  0  0  0  0 32  0  0  0 |  o = N30
 1  1  0  0  0  0  0  0  0  0  0  0  0  2  0  0 16  1  0 |  p = N35
 0 10  9  1  1  2  5  5 22  1  1  0  0  0  0  0  0 255  1 |  q = N31
 2 18  1 16  3  4  1  5  0  0 17  3  1  4  0  0  1 113 |  r = R52
```

Rastgele orman algoritmasının sınıflara göre ayrıntılı doğruluk oranı ve diğer ölçütlere ilişkin performans değerlendirme sonuçları bu bölümde sunulmuştur.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,500	0,007	0,639	0,500	0,561	0,556	0,910	0,618	M79
0,600	0,040	0,600	0,600	0,600	0,560	0,908	0,629	R30
0,927	0,006	0,895	0,927	0,911	0,906	0,989	0,946	R32
0,415	0,040	0,431	0,415	0,423	0,381	0,881	0,374	N23
0,421	0,006	0,571	0,421	0,485	0,482	0,809	0,399	N28
0,435	0,007	0,588	0,435	0,500	0,496	0,907	0,512	M54
0,375	0,026	0,402	0,375	0,388	0,361	0,934	0,434	R10
0,960	0,014	0,954	0,960	0,957	0,944	0,987	0,976	N39
0,594	0,006	0,774	0,594	0,672	0,668	0,980	0,754	R31
0,400	0,002	0,667	0,400	0,500	0,514	0,917	0,467	N48
0,483	0,036	0,518	0,483	0,500	0,461	0,880	0,482	N20
0,609	0,002	0,778	0,609	0,683	0,685	0,947	0,787	N45
0,675	0,005	0,730	0,675	0,701	0,696	0,984	0,766	C67
0,844	0,005	0,730	0,844	0,783	0,781	0,979	0,882	I86
0,892	0,003	0,846	0,892	0,868	0,866	0,997	0,905	N30
0,714	0,001	0,938	0,714	0,811	0,817	0,998	0,895	N35
0,952	0,031	0,854	0,952	0,900	0,882	0,992	0,956	N31
0,667	0,056	0,555	0,667	0,606	0,563	0,933	0,601	R52
Weighted Avg. 0,732	0,025	0,728	0,732	0,727	0,706	0,950	0,750	

Rastgele orman algoritmasının detaylı karışıklık matrisi bu bölümde sunulmuştur.

=== Karışıklık Matrisi===

```
  a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   q   r  <-- classified as
 23   3   0  10   1   0   3   0   1   0   1   2   1   0   0   0   0   1 |  a = M79
  0 108   6   2   4   3  13   5   0   2   6   0   1   1   1   0   4  24 |  b = R30
  0   2 102   1   0   0   0   0   0   0   1   0   0   0   0   0   4   0 |  c = R32
  8   1   2  56   0   0  11   3   0   0  24   0   0   0   1   0   2  27 |  d = N23
  2   5   0   3  16   0   0   0   0   0   4   0   2   0   0   0   2   4 |  e = N28
  0   3   0   5   0  20  13   1   0   0   1   0   0   0   0   0   2   1 |  f = M54
  1  20   0   8   1  11  33   1   1   0   1   0   2   0   0   0   6   3 |  g = R10
  0   5   0   6   0   0   1 436   0   0   1   0   0   0   3   0   0   2 |  h = N39
  0   0   0   0   0   0   0   0  41   0   1   0   0   0   0   0  27   0 |  i = R31
  0   4   0   0   0   0   0   0   0   6   1   0   0   0   0   0   3   1 |  j = N48
  1  13   0  22   3   0   3   3   0   0  72   0   2   1   0   0   0  29 |  k = N20
  0   0   0   1   0   0   0   1   1   0   0  14   0   3   0   0   0   3 |  l = N45
  1   1   0   1   1   0   1   1   0   0   4   0  27   0   1   1   0   1 |  m = C67
  0   0   0   0   0   0   0   0   0   0   0   0   1   0  27   0   0   0   4 |  n = I86
  0   0   0   0   0   0   1   2   0   0   0   0   0   0   0  33   0   0   1 |  o = N30
  0   2   1   0   0   0   0   1   0   0   0   0   0   1   0   0  15   1   0 |  p = N35
  0   1   3   0   0   0   1   0   9   1   0   0   0   0   0   0   0 298   0 |  q = N31
  0  12   0  15   2   0   2   3   0   0  22   1   1   5   0   0   0 126 |  r = R52
```

ZeroR algoritmasının sınıflara göre ayrıntılı doğruluk oranı ve diğer ölçütlere ilişkin performans değerlendirme sonuçları bu bölümde sunulmuştur.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,000	0,000	?	0,000	?	?	0,473	0,022	M79
0,000	0,000	?	0,000	?	?	0,499	0,091	R30
0,000	0,000	?	0,000	?	?	0,499	0,055	R32
0,000	0,000	?	0,000	?	?	0,490	0,067	N23
0,000	0,000	?	0,000	?	?	0,478	0,018	N28
0,000	0,000	?	0,000	?	?	0,473	0,022	M54
0,000	0,000	?	0,000	?	?	0,490	0,043	R10
1,000	1,000	0,229	1,000	0,372	?	0,497	0,228	N39
0,000	0,000	?	0,000	?	?	0,493	0,034	R31
0,000	0,000	?	0,000	?	?	0,417	0,007	N48
0,000	0,000	?	0,000	?	?	0,496	0,074	N20
0,000	0,000	?	0,000	?	?	0,453	0,011	N45
0,000	0,000	?	0,000	?	?	0,499	0,020	C67
0,000	0,000	?	0,000	?	?	0,474	0,015	I86
0,000	0,000	?	0,000	?	?	0,470	0,017	N30
0,000	0,000	?	0,000	?	?	0,478	0,010	N35
0,000	0,000	?	0,000	?	?	0,496	0,157	N31
0,000	0,000	?	0,000	?	?	0,497	0,095	R52
Weighted Avg.	0,229	0,229	?	0,229	?	?	0,493	0,113

ZeroR algoritmasının karışıklık matrisi bu bölümde sunulmuştur.

Karışıklık Matrisi

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	<-- classified as
0	0	0	0	0	0	0	46	0	0	0	0	0	0	0	0	0	0	a = M79
0	0	0	0	0	0	0	180	0	0	0	0	0	0	0	0	0	0	b = R30
0	0	0	0	0	0	0	110	0	0	0	0	0	0	0	0	0	0	c = R32
0	0	0	0	0	0	0	135	0	0	0	0	0	0	0	0	0	0	d = N23
0	0	0	0	0	0	0	38	0	0	0	0	0	0	0	0	0	0	e = N28
0	0	0	0	0	0	0	46	0	0	0	0	0	0	0	0	0	0	f = M54
0	0	0	0	0	0	0	88	0	0	0	0	0	0	0	0	0	0	g = R10
0	0	0	0	0	0	0	454	0	0	0	0	0	0	0	0	0	0	h = N39
0	0	0	0	0	0	0	69	0	0	0	0	0	0	0	0	0	0	i = R31
0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	j = N48
0	0	0	0	0	0	0	149	0	0	0	0	0	0	0	0	0	0	k = N20
0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	l = N45
0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	m = C67
0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	n = I86
0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	o = N30
0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	p = N35
0	0	0	0	0	0	0	313	0	0	0	0	0	0	0	0	0	0	q = N31
0	0	0	0	0	0	0	189	0	0	0	0	0	0	0	0	0	0	r = R52

KStar algoritmasının sınıflara göre ayrıntılı doğruluk oranı ve diğer ölçütlere ilişkin performans değerlendirme sonuçları bu bölümde sunulmuştur.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,543	0,010	0,556	0,543	0,549	0,539	0,894	0,562	M79
0,539	0,037	0,591	0,539	0,564	0,523	0,913	0,584	R30
0,900	0,008	0,868	0,900	0,884	0,877	0,991	0,948	R32
0,437	0,037	0,461	0,437	0,449	0,410	0,897	0,403	N23
0,474	0,007	0,581	0,474	0,522	0,516	0,873	0,427	N28
0,326	0,009	0,469	0,326	0,385	0,379	0,926	0,463	M54
0,318	0,022	0,400	0,318	0,354	0,330	0,918	0,409	R10
0,921	0,023	0,923	0,921	0,922	0,899	0,985	0,957	N39
0,406	0,008	0,651	0,406	0,500	0,501	0,949	0,591	R31
0,400	0,005	0,400	0,400	0,400	0,395	0,960	0,436	N48
0,470	0,035	0,519	0,470	0,493	0,455	0,873	0,460	N20
0,652	0,002	0,789	0,652	0,714	0,715	0,946	0,681	N45
0,675	0,003	0,844	0,675	0,750	0,750	0,986	0,795	C67
0,781	0,005	0,714	0,781	0,746	0,743	0,990	0,866	I86
0,892	0,004	0,805	0,892	0,846	0,844	0,995	0,903	N30
0,762	0,002	0,800	0,762	0,780	0,778	0,998	0,892	N35
0,930	0,041	0,808	0,930	0,865	0,840	0,989	0,950	N31
0,730	0,061	0,556	0,730	0,632	0,594	0,930	0,574	R52
Weighted Avg. 0,709	0,029	0,703	0,709	0,702	0,678	0,950	0,729	

KStar algoritmasının karışıklık matrisi bu bölümde sunulmuştur.

```
 a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  <-- classified as
25  2  0  7  0  1  1  1  1  0  3  2  0  0  0  0  0  3 |  a = M79
 0 97  7  1  3  2 12 10  0  3  5  0  0  0  2  2 14 22 |  b = R30
 0  3 99  0  0  0  1  1  0  0  0  0  0  1  0  0  5  0 |  c = R32
 9  2  1 59  1  1  3  4  1  0 20  0  0  0  1  0  0 33 |  d = N23
 1  4  1  1 18  0  3  0  0  0  4  0  0  0  0  0  2  4 |  e = N28
 0  4  0  7  0 15 13  1  0  0  2  0  0  0  0  0  2  2 |  f = M54
 2 14  3  6  2 13 28  1  2  1  0  0  2  0  1  0 11  2 |  g = R10
 0  7  0  6  0  0  0 418  2  2  4  0  0  1  4  0  3  7 |  h = N39
 2  2  0  4  0  0  3  2 28  0  2  0  0  0  0  0 25  1 |  i = R31
 0  3  0  0  0  0  0  0  0  6  0  0  0  0  0  0  5  1 |  j = N48
 1 12  0 24  3  0  2  6  0  1 70  0  1  1  0  0  1 27 |  k = N20
 0  0  0  0  0  0  0  1  1  0  1 15  0  2  0  0  0  3 |  l = N45
 3  1  0  1  1  0  1  1  0  0  5  0 27  0  0  0  0  0 |  m = C67
 1  0  0  0  0  0  0  0  0  0  0  1  0 25  0  0  0  5 |  n = I86
 0  0  0  0  0  0  0  3  0  0  0  0  0  0 33  0  1  0 |  o = N30
 0  2  0  0  0  0  0  1  0  0  1  0  1  0  0 16  0  0 |  p = N35
 0  3  3  1  1  0  1  0  8  2  1  0  0  0  0  2 291  0 |  q = N31
 1  8  0 11  2  0  2  3  0  0 17  1  1  5  0  0  0 138 |  r = R52
```

K-en yakın komşu algoritmasının sınıflara göre ayrıntılı doğruluk oranı ve diğer ölçütlere ilişkin performans değerlendirme sonuçları bu bölümde sunulmuştur.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,587	0,017	0,450	0,587	0,509	0,501	0,828	0,480	M79
0,589	0,058	0,502	0,589	0,542	0,494	0,847	0,492	R30
0,864	0,008	0,864	0,864	0,864	0,856	0,973	0,895	R32
0,430	0,044	0,414	0,430	0,422	0,379	0,805	0,350	N23
0,447	0,009	0,500	0,447	0,472	0,463	0,779	0,325	N28
0,457	0,010	0,525	0,457	0,488	0,478	0,839	0,421	M54
0,295	0,023	0,377	0,295	0,331	0,307	0,763	0,282	R10
0,907	0,020	0,932	0,907	0,920	0,896	0,965	0,937	N39
0,449	0,013	0,564	0,449	0,500	0,487	0,833	0,453	R31
0,400	0,006	0,353	0,400	0,375	0,371	0,777	0,235	N48
0,463	0,036	0,507	0,463	0,484	0,445	0,790	0,385	N20
0,696	0,002	0,800	0,696	0,744	0,743	0,878	0,642	N45
0,575	0,002	0,852	0,575	0,687	0,695	0,879	0,673	C67
0,781	0,006	0,694	0,781	0,735	0,732	0,960	0,772	I86
0,838	0,004	0,795	0,838	0,816	0,812	0,981	0,851	N30
0,714	0,002	0,833	0,714	0,769	0,769	0,918	0,758	N35
0,879	0,029	0,851	0,879	0,865	0,839	0,972	0,901	N31
0,630	0,050	0,572	0,630	0,599	0,556	0,874	0,496	R52
Weighted Avg. 0,691	0,028	0,694	0,691	0,691	0,664	0,896	0,665	

K-en yakın komşu algoritmasının karışıklık matrisi bu bölümde sunulmuştur.

=== Karışıklık Matrisi===

```
  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  <-- classified as
 27  4  0  6  0  1  1  0  1  0  3  2  0  0  0  0  0  1 |  a = M79
  2 106  6  1  4  2 10 10  0  4  5  0  0  0  1  1  7 21 |  b = R30
  0  6 95  1  0  0  1  0  0  1  0  0  0  1  0  0  5  0 |  c = R32
 11  2  1 58  2  1  8  4  1  0 20  0  0  0  2  0  0 25 |  d = N23
  3  5  2  3 17  0  1  0  0  0  3  0  0  0  0  0  1  3 |  e = N28
  0  6  0  6  0 21 11  0  0  0  1  0  0  0  0  0  1  0 |  f = M54
  2 20  2 10  1 14 26  1  2  1  0  0  1  0  0  0  6  2 |  g = R10
  1 10  1  7  0  0  1 412  2  2  4  0  0  2  4  0  2  6 |  h = N39
  3  3  0  4  0  0  2  2 31  0  2  0  0  0  0  0 22  0 |  i = R31
  0  3  0  0  1  0  0  0  0  6  0  0  0  0  0  0  4  1 |  j = N48
  2 16  0 26  4  1  1  4  0  1 69  0  1  1  0  0  0 23 |  k = N20
  0  0  0  0  0  0  0  1  1  0  1 16  0  2  0  0  0  2 |  l = N45
  3  1  0  1  1  0  1  1  1  0  8  0 23  0  0  0  0  0 |  m = C67
  1  0  0  0  0  0  0  0  0  0  0  1  0 25  0  0  0  5 |  n = I86
  0  1  0  0  0  0  1  4  0  0  0  0  0  0 31  0  0  0 |  o = N30
  0  3  0  0  0  0  0  1  0  0  1  0  1  0  0 15  0  0 |  p = N35
  1  7  3  1  1  0  3  0 16  2  1  0  0  0  1  2 275  0 |  q = N31
  4 18  0 16  3  0  2  2  0  0 18  1  1  5  0  0  0 119 |  r = R52
```

Çok katmanlı algılayıcılar algoritmasının sınıflara göre ayrıntılı doğruluk oranı ve diğer ölçütlere ilişkin performans değerlendirme sonuçları bu bölümde sunulmuştur.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,500	0,013	0,469	0,500	0,484	0,472	0,895	0,476	M79
0,561	0,034	0,620	0,561	0,589	0,551	0,910	0,563	R30
0,891	0,012	0,810	0,891	0,848	0,840	0,987	0,862	R32
0,400	0,045	0,394	0,400	0,397	0,353	0,879	0,343	N23
0,474	0,009	0,500	0,474	0,486	0,477	0,876	0,355	N28
0,391	0,009	0,500	0,391	0,439	0,431	0,876	0,357	M54
0,398	0,025	0,422	0,398	0,409	0,383	0,869	0,348	R10
0,947	0,014	0,953	0,947	0,950	0,936	0,990	0,983	N39
0,681	0,009	0,723	0,681	0,701	0,691	0,969	0,730	R31
0,333	0,004	0,417	0,333	0,370	0,368	0,958	0,307	N48
0,497	0,034	0,540	0,497	0,517	0,481	0,894	0,499	N20
0,739	0,005	0,654	0,739	0,694	0,691	0,950	0,700	N45
0,675	0,005	0,730	0,675	0,701	0,696	0,948	0,768	C67
0,781	0,007	0,641	0,781	0,704	0,702	0,977	0,797	I86
0,865	0,006	0,727	0,865	0,790	0,789	0,983	0,768	N30
0,762	0,002	0,800	0,762	0,780	0,778	0,997	0,880	N35
0,917	0,017	0,911	0,917	0,914	0,898	0,985	0,949	N31
0,667	0,049	0,589	0,667	0,625	0,585	0,915	0,511	R52
Weighted Avg. 0,722	0,022	0,720	0,722	0,720	0,698	0,945	0,712	

Çok katmanlı algılayıcılar algoritmasının karışıklık matrisi bu bölümde sunulmuştur.

=== Karışıklık Matrisi===

```
 a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  <-- classified as
23  1  0 12  2  1  2  0  0  0  1  1  1  1  0  0  0  1 |  a = M79
 3 101  9  4  4  3 12  3  0  4  8  0  1  1  3  2  1 21 |  b = R30
 1  2 98  3  0  0  2  0  0  1  0  0  0  0  0  0  3  0 |  c = R32
12  3  3 54  1  3 11  5  0  0 19  0  0  0  1  0  0 23 |  d = N23
 1  2  2  3 18  0  2  0  0  0  2  0  2  0  0  0  0  6 |  e = N28
 4  4  1  5  1 18 10  0  0  0  1  0  0  1  0  0  0  1 |  f = M54
 2 17  3  7  1  8 35  4  0  1  3  0  1  0  1  1  0  4 |  g = R10
 0  3  0  9  0  0  1 430  0  0  2  0  0  1  6  0  0  2 |  h = N39
 0  0  0  0  0  0  0  0 47  0  0  0  0  0  0  0 22  0 |  i = R31
 0  3  0  1  1  0  1  1  0  5  2  0  1  0  0  0  0  0 |  j = N48
 2 10  1 24  2  2  5  0  0  0 74  1  1  2  0  1  1 23 |  k = N20
 1  0  0  0  0  0  0  0  0  0  0  0 17  1  3  0  0  0  1 |  l = N45
 0  3  0  0  1  0  0  1  0  0  7  0 27  0  0  0  1  0 |  m = C67
 0  0  0  0  0  0  0  0  0  0  0  0  2  0 25  0  0  0  5 |  n = I86
 0  1  0  1  0  0  0  3  0  0  0  0  0  0  0 32  0  0  0 |  o = N30
 0  1  1  0  0  0  0  1  0  0  0  0  0  1  0  0 16  0  1 |  p = N35
 0  1  3  0  1  0  1  0 18  1  0  0  0  0  0  1  0 287  0 |  q = N31
 0 11  0 14  4  1  1  3  0  0 18  5  1  5  0  0  0 126 |  r = R52
```

Naive Bayes algoritmasının sınıflara göre ayrıntılı doğruluk oranı ve diğer ölçütlere ilişkin performans değerlendirme sonuçları bu bölümde sunulmuştur.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,294	0,009	0,455	0,294	0,357	0,353	0,920	0,290	M79
0,540	0,069	0,447	0,540	0,489	0,434	0,921	0,532	R30
0,897	0,017	0,761	0,897	0,824	0,815	0,994	0,929	R32
0,211	0,031	0,286	0,211	0,242	0,207	0,904	0,340	N23
0,111	0,002	0,500	0,111	0,182	0,231	0,904	0,215	N28
0,071	0,003	0,333	0,071	0,118	0,147	0,902	0,241	M54
0,219	0,012	0,467	0,219	0,298	0,297	0,894	0,390	R10
0,914	0,002	0,993	0,914	0,952	0,940	0,990	0,983	N39
0,375	0,005	0,750	0,375	0,500	0,519	0,944	0,532	R31
0,000	0,000	?	0,000	?	?	0,963	0,313	N48
0,420	0,027	0,553	0,420	0,477	0,446	0,908	0,474	N20
0,636	0,009	0,538	0,636	0,583	0,578	0,962	0,681	N45
0,667	0,008	0,615	0,667	0,640	0,634	0,985	0,647	C67
0,273	0,003	0,600	0,273	0,375	0,398	0,992	0,698	I86
0,714	0,002	0,909	0,714	0,800	0,802	0,984	0,798	N30
0,400	0,000	1,000	0,400	0,571	0,631	1,000	1,000	N35
0,923	0,052	0,788	0,923	0,850	0,820	0,982	0,907	N31
0,806	0,121	0,403	0,806	0,538	0,511	0,904	0,529	R52
Weighted Avg. 0,662	0,033	?	0,662	?	?	0,952	0,697	

Naive Bayes algoritmasının karışıklık matrisi bu bölümde sunulmuştur.

=== Karışıklık Matrisi ===

```
  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  <-- classified as
  5  2  0  8  0  0  1  0  0  0  1  0  0  0  0  0  0  0 |  a = M79
  1 34  1  1  0  0  2  0  0  0  0  0  0  0  1  0  6 17 |  b = R30
  0  1 35  0  0  0  0  0  0  0  0  0  0  0  0  0  3  0 |  c = R32
  1  0  0  8  0  0  0  0  0  0 10  0  0  0  0  0  0 19 |  d = N23
  0  1  0  0  1  0  0  0  0  0  1  0  0  0  0  0  1  5 |  e = N28
  0  1  1  3  0  1  3  0  0  0  3  0  0  0  0  0  0  2 |  f = M54
  1 10  4  1  0  1  7  0  0  0  0  0  0  0  0  0  6  2 |  g = R10
  0  7  0  2  0  0  1 138  0  0  0  0  0  0  0  0  0  3 |  h = N39
  1  0  0  2  0  0  0  0  9  0  0  0  0  0  0  0 12  0 |  i = R31
  1  4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0 |  j = N48
  1  5  0  2  0  0  1  0  0  0 21  0  1  0  0  0  0 19 |  k = N20
  0  0  0  0  1  0  0  0  0  0  0  7  1  1  0  0  0  1 |  l = N45
  0  0  0  0  0  0  0  1  0  0  1  0  8  0  0  0  0  2 |  m = C67
  0  0  0  0  0  0  0  0  0  0  0  0  4  0  3  0  0  0  4 |  n = I86
  0  3  0  1  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0 |  o = N30
  0  0  0  0  0  0  0  0  0  0  0  0  0  3  0  0  2  0  0 |  p = N35
  0  0  5  0  0  1  0  0  3  0  0  0  0  0  0  0 108  0 |  q = N31
  0  8  0  0  0  0  0  0  0  0  0  1  2  0  1  0  0  0 50 |  r = R52
```

Lojistik regresyon analizinin sınıflara göre ayrıntılı doğruluk oranı ve diğer ölçütlere ilişkin performans değerlendirme sonuçları bu bölümde sunulmuştur.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,522	0,011	0,533	0,522	0,527	0,516	0,936	0,427	M79
0,617	0,051	0,547	0,617	0,580	0,536	0,928	0,557	R30
0,909	0,011	0,826	0,909	0,866	0,859	0,996	0,908	R32
0,467	0,038	0,474	0,467	0,470	0,432	0,922	0,405	N23
0,263	0,007	0,417	0,263	0,323	0,321	0,921	0,352	N28
0,304	0,009	0,438	0,304	0,359	0,352	0,950	0,462	M54
0,398	0,025	0,427	0,398	0,412	0,386	0,937	0,419	R10
0,938	0,007	0,977	0,938	0,957	0,945	0,993	0,987	N39
0,696	0,006	0,800	0,696	0,744	0,738	0,988	0,784	R31
0,467	0,005	0,438	0,467	0,452	0,448	0,945	0,330	N48
0,436	0,026	0,580	0,436	0,498	0,469	0,910	0,523	N20
0,783	0,004	0,692	0,783	0,735	0,733	0,991	0,714	N45
0,775	0,007	0,705	0,775	0,738	0,733	0,984	0,785	C67
0,844	0,006	0,711	0,844	0,771	0,770	0,998	0,891	I86
0,784	0,006	0,725	0,784	0,753	0,749	0,995	0,853	N30
0,810	0,001	0,944	0,810	0,872	0,873	0,991	0,830	N35
0,917	0,017	0,911	0,917	0,914	0,898	0,991	0,965	N31
0,683	0,062	0,538	0,683	0,601	0,559	0,940	0,576	R52
Weighted Avg. 0,726	0,022	0,727	0,726	0,724	0,703	0,964	0,739	

Lojistik regresyon analizinin karışıklık matrisi bu bölümde sunulmuştur.

=== Karışıklık Matrisi ===

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	<-- classified as
24	2	0	9	1	1	2	0	0	0	2	1	1	1	0	0	0	2	a = M79
3	111	8	6	5	1	12	0	0	5	5	0	3	0	2	0	0	19	b = R30
0	3	100	0	0	0	0	0	0	0	1	0	2	0	0	0	4	0	c = R32
9	7	1	63	1	2	2	3	0	0	17	0	0	0	1	0	0	29	d = N23
1	8	0	1	10	0	5	0	0	0	3	0	0	0	0	0	1	9	e = N28
2	2	1	6	1	14	15	0	0	0	2	0	0	0	0	0	0	3	f = M54
2	22	2	8	1	13	35	0	0	0	1	0	1	0	0	0	2	1	g = R10
1	6	0	6	0	0	1	426	0	1	1	0	1	0	7	1	0	3	h = N39
0	0	0	0	0	0	0	0	48	0	0	0	0	0	0	0	21	0	i = R31
0	2	0	0	0	0	2	0	0	7	2	0	2	0	0	0	0	0	j = N48
2	8	2	21	3	0	6	1	0	0	65	0	1	1	0	0	0	39	k = N20
0	1	0	0	0	0	0	0	0	0	0	18	0	1	1	0	0	2	l = N45
1	4	0	1	0	0	0	1	0	0	1	1	31	0	0	0	0	0	m = C67
0	0	0	0	0	0	0	0	0	0	0	0	1	0	27	0	0	0	4	n = I86
0	1	0	3	1	0	0	3	0	0	0	0	0	0	29	0	0	0	o = N30
0	0	0	0	0	0	0	1	0	0	0	1	2	0	0	17	0	0	p = N35
0	4	7	0	0	0	1	0	12	2	0	0	0	0	0	0	287	0	q = N31
0	22	0	9	1	1	1	1	0	1	12	4	0	8	0	0	0	129	r = R52