Disiplinlerarası Yen Araş Der, 3(1), 23-30, 2023

İstanbul Arel Üniversitesi Lisansüstü Eğitim Enstitüsü
Disiplinlerarası Yenilik Araştırmaları Dergisi

Sağlıkta veri kalitesi ve veri madenciliği uygulamaları

Data quality and data mining applications in healthcare

Ahmet KOÇAK [1], Mehmet Ali ERGÜN [2]*

[1] Sağlık Bakanlığı Teftiş Kurulu Başkanlığı, Sağlık Bakanlığı, Ankara, Türkiye.
ahmet.kocak3@saglik.gov.tr

[2] Dahili Tıp Bilimleri, Tıp Fakültesi, Gazi Üniversitesi, Ankara, Türkiye.
aliergun@gazi.edu.tr

Geliş Tarihi/Received: 14.08.2022 Bölüm/Section: Sağlık Bilimleri/Sağlık Yönetimi

Kabul Tarihi/Accepted: 01.11.2022 Derleme/Review

Özet

Veri günümüzde çok sık karşılaşılan bir terimdir. Verinin doğru kullanımı doğru değerlendirmeyi sağlar. Bu da kaynakların verimli kullanımını, verilen hizmetin kalitesinin artmasını sağlamaktadır. Verinin en çok toplandığı alanların başında sağlık sektörü gelmektedir. Sağlık hizmet sunumunun maddi ve manevi yükü ağırdır. Bu hizmetin en iyi şekilde verilmesi, kaynakların doğru kullanılması ile yakın ilişkilidir. Sağlık verilerinden anlamlı sonuçların çıkarılarak hekimlere, hemşirelere ve sağlık yöneticileri gibi sağlık sektörü çalışanlarına yön gösterecek bilgilerin sağlanması sağlık verilerinin büyüklüğü düşünüldüğünde ancak veri madenciliği metotları ile mümkündür. Sağlık sektörünün insan hayatını direkt etkileyen bir doğası olması sebebi ile sağlıkta kullanılan verilerin kalitesinin en üst düzeyde olması beklenmektedir. Bu çalışmada veri kalitesini ve veri madenciliğini bütüncül olarak ele almıştır. Uygulama örnekleri aracılığıyla veri madenciliği ile sağlık sektöründe ne tür çalışmalar yapılabileceğine dair genel bir bakış açısı sağlanmıştır.

Anahtar Kelimeler: Sağlık, veri, veri madenciliği, veri kalitesi.

Abstract

Data is a common term recently. Correct use of data ensures correct evaluation, which ensures efficient use of resources and increases the quality of the service provided. The health sector is one of the areas where data is collected the most. The financial and moral burden of health service delivery is heavy. Providing this service in the best way is closely related to the correct use of resources. By extracting meaningful results from health data, providing information that will guide health sector workers such as physicians, nurses and health managers is only possible with data mining methods, considering the size of health data. Since the health sector has a nature that directly affects human life, the quality of data used in health is expected to be at the highest level. In this study, data quality and data mining were handled holistically. A general viewpoint on what kind of studies can be done in the health sector with data mining has been provided through application examples.

Keywords: Health, data, data mining, data quality.

1. Giriş

İnternetin gelişimi ve kullanımının yaygınlaşması, veri tabanlarının kapasitesinin artması, iletişim araçlarının çeşitlenmesi ve kullanımının sıklaşması veri teriminin önemini artırmıştır [1]. Verinin öneminin artması ile veri kalitesi kavramı da son dönemde araştırma konularından biri haline gelmiştir [2]. Veri kalitesi, bilgiyi oluşturan nicel veya nitel parçaların uygulanacak işleme, karara, amaca uygunluğunun durumudur [1]. Veri madenciliği ise büyük veriden yararlı bilginin çıkarılmasıdır [3]. Veri madenciliğinin uygulama alanlarına bankacılık, finans vb. alanlar öncülük etse de son dönemde

* Yazışılan yazar/Corresponding author: Ahmet KOÇAK

[1] orcid.org/0000-0003-0754-7773; [2] orcid.org/0000-0001-9696-0433

DOI: http://dx.doi.org/10.56723/dyad.1161993

sağlık alanında veri madenciliği çalışmaları artarak devam etmektedir [4]. Veri madenciliği metotları beş ana yöntemle bilgiyi keşfeder. Bunlar genel çıkarım, ilişki kuralları, sınıflandırma ve kümeleme metotları, tahmin algoritmaları ve aykırılık tespitleridir [3].

Veri madenciliği uygulamalarının kalitesi veri kalitesi ile doğrudan ilişkidir. Veri kalitesi ise verinin uygulanacak işleme ve amaca uygunluğu ile eksiksizlik, tamlık, doğruluk, kesinlik gibi birçok kavramı içerir [5]. Bu çalışma ile veri kalitesi ve veri madenciliği hakkında literatürde yer alan bilgiler özetlenerek veri madenciliği uygulamaları ile sağlık alanında sunulan faydalı çalışma örnekleri gösterilmiş ve bu doğrultuda yeni çalışmalara ışık tutulması amaçlanmıştır.

2. Veri kalitesi

Veri kalitesi çok boyutlu bir kavram olup içerisinde doğruluk, tamlık, tutarlılık, güvenirlilik, geçerlilik gibi birçok parametresi olan bir kavramdır [6]. Veri kalitesinin yüksek olması bilgi sistemlerinin en temel ihtiyacıdır. Veri kalitesinin yüksek olması doğru karar vermede, fırsatları yakalamakta oldukça etkilidir. Veri kalitesi düşüklüğü yanlış aksiyon gösterilmesine, müşteri kayıplarına, fırsatların kaçmasına ve yanlış kararlara sebep olabilmektedir [7]. Sağlık sektöründe ise veri kalitesinin önemi daha kritiktir. Sağlık verisinin veri kalitesinin düşüklüğü can kayıplarına, sakatlıklara, kaynakların yanlış yerlere aktarılmasına neden olabilir. İngiltere'de yapılan çalışmada kalitesi düşük verilerden kaynaklı olarak yılda ortalama 71.000 kişinin öldüğü tahmin edilmektedir [8]. Gartner'in 2017 yılındaki araştırması, düşük veri kalitesinin yıllık ortalama 15 milyon ABD dolar kaybına sebep olduğunu göstermiştir [2]. Çalışmalar veri kalitesinin düşüklüğünün maddi ve manevi birçok zarara sebep olabileceğini göstermiştir. Sağlık alanında tutulan kayıtların kalitesinin yüksek olması hem hastaya gereksiz tıbbi işlem yapılmasını engelleyecek, hem de sağlık personellerine zaman ve iş gücü kazandıracaktır. Bunun yanı sıra israfı önleyerek mali kazanç sağlanacaktır. Örneğin; hasta hekime başvurduğunda hastanın sağlık kayıtlarında, kısa bir süre önce çekilmiş Manyetik Rezonans (MR) görüntüsü ve görüntülemeye ilişkin bilgiler varsa hekim tekrar hasta için MR çekilmesini talep etmeyerek mükerrer işlemi engellemiş olacaktır. Bu süreçte en kritik kavram veriye güvendir. Bu güven veri kalitesinin yüksekliğiyle iç içedir.

2.1. Veri kalite standartları

Veri kalite standartları ulusal ve uluslararası olmak üzere ikiye ayrılabilir. Uluslararası veri kalitesi standartları olarak ISO/IEC 25012 [ISO/IEC,2008], ISO/IEC 25024'e [ISO/IEC,2015], EUROSTAT veri kalitesi standartları gösterilebilir. Ulusal bazda Kanada İstatistik Kurumu, Hollanda İstatistik Kurumu, Türkiye İstatistik Kurumu gibi yerel standartlar uygulayan kurumlar da mevcuttur. Bununla birlikte bilim insanlarının, örneğin Wang, Strong ve Redman'ın tanımlamış olduğu veri kalite standartları vardır. ISO/IEC 25012 [ISO/IEC,2008], ISO/IEC 25024'e [ISO/IEC,2015] veri kalitesi kriterleri tamlık, doğruluk, uyumluluk ve tutarlılık kriterleriyle incelenmiştir [9]. EUROSTAT veri kalite kriterleri uygunluk, doğruluk, güncellik, erişilebilirlik, kıyaslanabilirlik, tutarlılık ve bütünlük kriterleriyle değerlendirilmiştir [10]. Ulusal bazda uygulanan veri kalite standartlarında uluslararası kriterler baz alınmıştır [11]. Wang ve Strong ise veri kalitesini doğruluk, bütünlük, tutarlılık, güncellik, yorumlanabilirlik, anlaşılabilirlik, akla uygunluk, saygınlık, tarafsızlık, erişilebilirlik, güvenlik, katma değer, açık sunum, uygun büyüklükte veri olarak tanımlamıştır [12].

Redman ise doğruluk, bütünlük, tutarlılık, temsil gücü, yaygınlık, değerlendirilebilirlik, uygunluk, ideal veri büyüklüğü, aktarılabilirlik-taşınabilirlik, veri sözlüğü, elde edilebilirlik ve minimum gereksiz veri olarak tanımlamıştır [13]. Birçok farklı kriter kullanılmakla birlikte bahsedilen bütün kriterler önemlidir. Verinin tamlığı ve doğruluğu ise veri kalitesinin temelini oluşturur.

2.2. Sağlıkta veri kalite standartları

Sağlıktaki veri kalitesini artırmak ve uluslararası bir dilin gelişmesini sağlamak amacı ile birçok sağlık veri standardı uygulaması vardır. Bunlardan en yaygın olarak kullanılanları;

Uluslararası Hastalık Sınıflandırması [International Classificcation of Diseases (ICD)], ilk olarak ölüm nedeni sınıflandırması ile başlayan ve daha sonra veri toplama standartlaşma süreci ile genişletilen bir sınıflandırma metodudur. Sınıflandırmanın son versiyonu olan ICD-11 2022 yılında yürürlüğe girmiştir [14].

Anatomik Terapötik Kimyasal [Anatomic Therapeutic Chemical Classification (ATC)] Sınıflandırması, ilaçların etken maddesinin vücudun hangi noktasına etki ettiği ile kimyasal özelliklerine göre beş farklı basamakta gruplandırıldığı standart biçimidir [15].

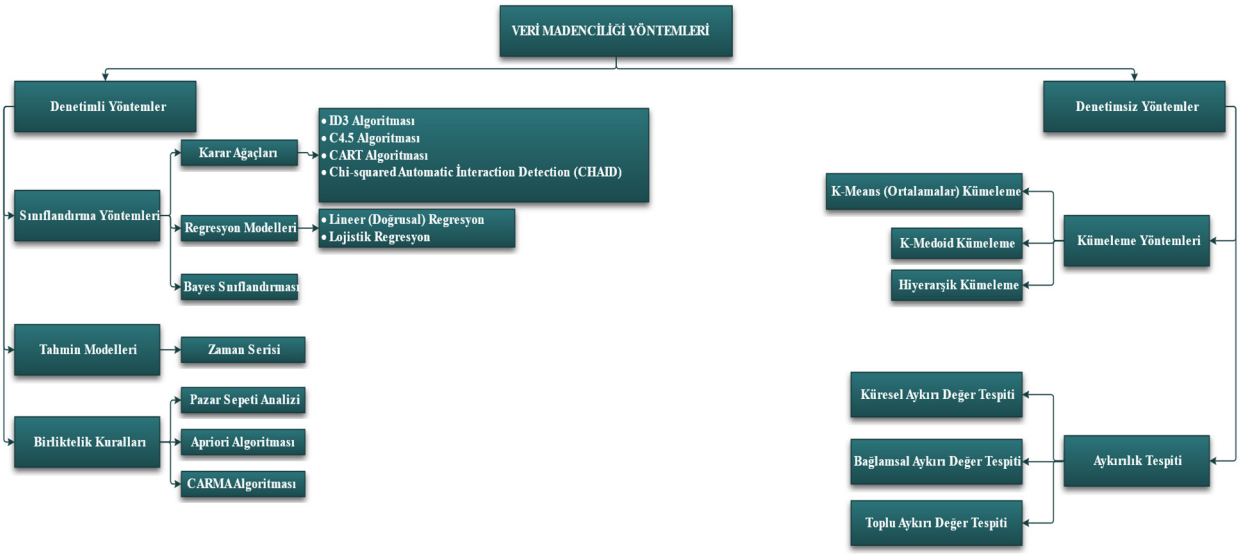
Sağlık Uygulamaları Tebliği [SUT], 24.03.2013 tarihinde yürürlüğe girmiş ve ilgili tebliğ SUT kodları ile tıbbi malzemenin adı, fiyatı vb. konularda standartlaşma sağlamıştır [16].

Mantıksal Gözlem Tanımlayıcı Adları ve Kodları [Logical Observation Identifiers Names and Codes (LOINC)], sağlık alanında yapılan ölçümlerin, gözlemlerin belgelere aktarılmasına yönelik standarttır [17].

Küresel Tıbbi Cihaz Sınıflandırması [Global Medical Devices Nomenclature (GMDN)], medikal cihaz sınıflandırmasında kullanılan Avrupa Standartlaştırma Kuruluşu tarafından oluşturulan bir standarttır. Jenerik isim ile 5 basamaklı koddan oluşur [18]. Mevcut standartlar ulusal ve uluslararası sağlık alanında ortak bir dil sağlar. Sağlık alanında uygulanan veri kalite standartları gerek klinik karar destek sistemleri gerekse sağlık harcamalarının faturalandırılması için önemlidir.

3. Veri madenciliği

21. yüzyılı tanımlarken ilk akla gelen bilgi çağı olduğudur. Ancak bu aslında yanlış bir ifadedir. Doğru olan ifade veri çağıdır. Bilgisayar ve internet ortamlarında terabaytlarla, petabaytlarla ifade edilen veriler bulunmaktadır. Verinin kaynağında ise toplum bilimleri, mühendislik, tıp, güvenlik gibi birçok farklı alan vardır. Veri hacmindeki artışın nedeni gelişen veri toplama ve depolama araçları ile olmuştur. Bu verilerin işlenip insanlara planlama, tahmin vb. işlemler yaptırabilecek bilgileri sağlama işleminde veri madenciliği yöntemleri uygulanmaktadır [19]. Veri madenciliği büyük hacimli verilerden öz bilgiye ulaşma sürecidir [20]. Başka bir tanıma göre büyük veri tabanlarında makine öğrenme yöntemi uygulamasıdır. Veri madenciliği uygulamaları veriden anlam çıkarmayı, tahmin yapmayı, yeni bir bilgi ortaya koymayı, sınıflandırma ve nesneler arasındaki ilişkileri tespit etmeyi sağlar [21].



Şekil 1. Veri madenciliği yöntemleri [3], [22], [23], [24].

Şekil 1'de belirtildiği üzere veri madenciliği yöntemleri denetimli ve denetimsiz yöntemler olarak iki gruba ayrılır. Denetimli yöntemler sınıflandırma yöntemleri, tahmin modelleri, birliktelik kurallarını içerir. Denetimsiz yöntemler ise kümeleme yöntemleri ve aykırılık tespiti yöntemlerini içerir [3], [22], [23], [24].

3.1. Denetimli yöntemler

3.1.1. Sınıflandırma yöntemleri

Farklı şeyler arasında farklı karakteristiklerin aynı özelliklerinin belirlenerek gruplandırılmasıdır [3]. Sınıflandırma modelleri iki temel adımda çalışır. Birinci adımda sınıflandırıcı veri sınıfını öncelikle veride tanımlar. Bu adım veri kümesinde test verisinde öğrenme aşaması olarak tanımlanır. İkinci adım ise test verisinde öğrenilen sınıfın eğitim veri setinde kullanılmasıdır. Sınıflandırma yöntemlerinde karar ağaçları, regresyon modelleri, naive bayes, yapay sinir ağları vb. algoritmalar kullanılır [22]. Örneğin; laboratuvar bulgularından hastalık tanısı sınıflandırılabilir.

3.1.2. Tahmin modelleri

Zaman serisi verilerinde mevcut verilere göre geleceği tahmin etmektir [3]. Örneğin; 2030 yılında kalp hastası sayısı geçmiş verilerden çıkarılabilir.

3.1.3. Birliktelik kuralları

Birliktelik analizi bir değişkenin başka bir değişkene olan bağlılığını ortaya koyar [3]. Birliktelik kuralları veri madenciliğinin tanımlayıcı alanlarından biridir. Büyük veriyi oluşturan elemanlar arasındaki ilişkinin tespitinde kullanılır. Geçmiş veriler analiz edilerek geleceğe yönelik çalışmalarda kullanılır [23]. Örneğin; Diyabet hastalığı ile görme kaybı arasındaki ilişkiyi ortaya koyabilir.

3.2. Denetimsiz yöntemler

3.2.1. Kümeleme yöntemleri

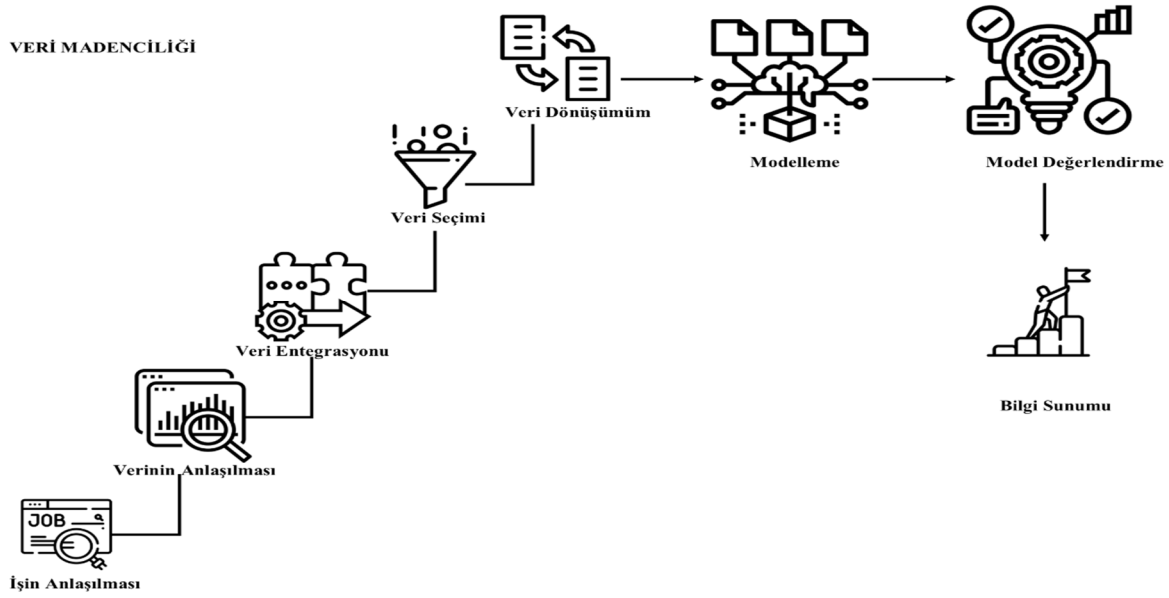
Kümeleme yöntemleri, tahmin edilecek bir sınıf olmadığında, verideki elemanlar doğal gruplara bölünerek uygulanır. Bu kümeler, örneklerin alındığı alanda çalışmakta olan bir mekanizmayı, bazı örneklerin diğer örneklere göre daha güçlü bir benzerlik göstermesiyle yansıtır [24]. Benzerliği veri setindeki elemanların uzaklık ölçütlerine göre gruplandırır [3].

Örneğin; Antibiyotik reçete etme veri setinde aşırı antibiyotik reçete eden hekimler ve düşük antibiyotik reçete eden hekimler kümesi oluşturulabilir.

3.2.2. Aykırılık tespiti

Veri setindeki uç değerlerin veya anormal durumların tespitidir [3]. Aykırı değer tespiti, beklentiden çok farklı davranışlara sahip veri nesnelerini bulma işlemidir. Bu tür nesnelere aykırı değerler veya anomaliler denir [25]. Örneğin; 80 yaşında kadın hastanın doğum yapamayacağı büyük veride işaretlenebilir.

Veri madenciliği sekiz basamakta yürütülen bir süreçtir [4].



Şekil 2. Veri madenciliği süreci [4].

Şekil 2'de belirtildiği üzere veri madenciliği işin anlaşılması ile başlayıp, uygulanan çalışmanın ikincil kişilerle paylaşılmasını içeren bilgi sunumu aşamasıyla sonlanan bir süreçtir. Süreçteki aşamalar belirlenmiştir. İşin anlaşılması, işin amacının, hedefinin ve kapsamının tanımlandığı aşamadır. Verinin anlaşılması, eksik ve kirli verinin anlaşılıp veri kalitesini yükseltmek için eksiklerin ve kirliliklerin iyileştirildiği aşamadır. Veri entegrasyonu, çeşitli veri kaynaklarının birleştirildiği aşamadır. Veri seçimi, analizde kullanılacak verilerin seçiminin yapıldığı aşamadır. Veri dönüşümü, verilerin normalize edildiği aşamadır. Modelleme, çalışmaya uygun olan modelin seçimin yapıldığı aşamadır. Modelin değerlendirilmesi uygulanan modelin doğruluğunun, başarısının değerlendirildiği aşamadır. Bilgi sunumu yapılan çalışmalar sonucunda elde edilen bilgilerin paylaşılması aşamasıdır [4].

4. Sağlık sektöründe veri madenciliği uygulama örnekleri

Sağlık sektöründe veri madenciliği son dönemde oldukça yaygın bir konudur. Veri madenciliği uygulamaları ile ilaç kullanım önerileri, hastalık teşhis sistemleri, yolsuzluk tespitleri gibi birçok farklı konuda çalışma yapılmıştır. Yapılan çalışmalar ile sınırlı olan sağlık kaynaklarının kullanımının verimliliği artmakla birlikte, minimum hata payına sahip olan sağlık sektörü için doğru karar verme mekanizmaları kurularak sakatlıkların, ölümlerin önüne geçilebilecektir.

Hong, Lu M, Lu C ve Zhu [26], akciğer kanseri veri seti üzerinden veri madenciliği çalışması uygulamıştır. Veri seti olarak Nanjing Göğüs Hastanesindeki akciğer kanseri teşhisi alan hastaların tıbbi kayıtları kullanılmıştır. Ocak 2018 ile Aralık 2018 arasında gerçekleşen akciğer kanseri tanı ve tedavi planı veri tabanındaki veriler alınmıştır. Medcase Ver5.2 klinik araştırma destek platformu ile ilaç ve 287 teşhis ve tedavi şemaları çıkarılmış ve klinik akciğer kanseri için bir veri tabanı kurulmuştur. Kurulan veri tabanı ile elde edilen bilgiler kanser hastalarının tedavilerinde doğru karar verme sürecinde kullanılabilecektir [26].

Kaur, Doja ve Ahmad [27], kanser hastaları üzerine yapmış olduğu çalışmada veri madenciliği ve makine öğrenme yöntemleri ile kanser hastalarının sağ kalım sürelerinin hesaplanmasına odaklanılmıştır. ABD merkezli SEER veri seti kullanılmıştır. Meme ve akciğer kanserlerine odaklanılmıştır. Eksik veriler mod atama yöntemi ile normalleştirilmiştir. Veri madenciliği yöntemi olarak lojistik regresyon, destek vektör makineleri, karar ağaçları, Bayes sınıflandırma yöntemleri, rastgele orman, en yakın komşular gibi birçok teknik kullanılarak hibrit bir model uygulanmıştır. Uygulanan yöntemlerin karışıklık matrisi ile doğruluğu test edilmiş ve yöntemlerin benzer sonuçlar verdiği görülmüştür. Meme kanseri hastaların yaşam süreleri tahmin edilmiştir [27].

Kirlioğlu ve Aşuk [28], sağlık sigortalarının yolsuzluk ve usulsüzlük tespitine yönelik veri madenciliği çalışması yapmıştır. Çalışmalarında 2001-2009 arasındaki 808348 kayıtlı veriyi kullanmışlardır. Anomali tespitine dayanan yöntemlerle belirlenen kişiler kümeleme yöntemi ile yeni giren kayıtlarla anomalilerin olduğu küme işaretlenmiştir. Çalışma sağlık sigorta şirketlerinin dolandırıcılara yüksek sigorta tazminatları ödemesine engel olmaktadır [28].

Santos, Malheiros, Cavalheiro ve Oliveira [29], halk sağlığı karar vericilerine beyin tümörleri hakkında analitik bilgi sağlamak için bir veri madenciliği sistemi kurmayı amaçlamıştır. Kanser hastalarının getirmiş olduğu finansal yük itibariyle halk sağlığı açısından beyin tümörleri hakkında bilgilendirici sistem kurulmuştur. Brezilya sağlık sistemleri beyin tümörü verileri kullanılmıştır. Veri madenciliği yöntemi olarak kümeleme, sınıflandırma ve birliktelik analizi yöntemleri kullanılmıştır. Kurulan sistem, beyin tümörleri hakkında uzman olmayan kişilere beyin tümörü tedavisi hakkında doğru bilgi sağlamakta başarılı bulunmuştur [29].

Costa ve diğerleri, çoklu biyobelirteçlerin kombinasyonu ile oral skuamöz hücreli karsinom teşhisi koymak için bir veri madenciliği yaklaşımı ortaya koymuştur [30]. Çalışmada tükürük örnekleri, cinsiyet, yaş ve sigara alışkanlığı gibi kriterler rastgele orman algoritması ile kullanılmıştır. Test grubu olarak 27 hastanın, kontrol grubu olarak 41 hastanın sonuçları kullanılmıştır. Çalışma %80'in üzerinde doğruluk vermiştir. Bu çalışma doğru oral skuamöz hücreli karsinom teşhisi koymak için kullanılabilecektir [30].

Wang ve diğerleri, ilaç ve tanı arasındaki uyumsuzlukları tespit etmek için bir çalışma yapmıştır [31]. Çalışmada Tayvan'da 2011 ile Aralık 2015 arasında ayakta tedavi alan hastaların kayıtları kullanılmıştır. Eğitim ve test verisi kullanılmıştır. Toplam 23 milyon satır veri kullanılmıştır. Çalışmada ilaç ve tanı arasındaki ilişkide 3,120,449 satır veride sınıflandırma yöntemleri kullanılarak yüksek ilişki kurulmuştur. ICD-9 CM (International Classification of Diseases, Ninth Revision, Clinical Modification) ve ATC (Anatomical Therapeutic Chemical) kodları üzerinden eşleştirme yapılmıştır. 1000 eşleşme sonucu için uzman iki eczacı ile çalışmanın sonucu değerlendirilmiştir. Sonuçlarda yüksek başarı elde edilmiştir. Yapılan çalışma klinik rehberlerde ve ilaç endikasyon bilgilendirilmesinde kullanılabilecektir [31].

Parviainen, Vázquez-Arias, Arrebola ve Martín-Peinado, kırsal bölgedeki maden alanlarının insan sağlığı riskleri üzerine bir çalışma yapmıştır [32]. Potansiyel fitotoksisite ve insan sağlığına odaklanarak 52 farklı bölgedeki toprak örneklerinin kimyasal bileşenleri incelenmiştir. Fitotoksisite biyo-tahlili ile toprak ve sağlık ilişkisi üzerine kanserojen ve kanserojen olmayan iki sınıf modellenmiştir. İnsan sağlığı risk değerlendirmesi kanser ölüm oranlarıyla karşılaştırılmıştır. 52 farklı bölgedeki kanser alt türleri ile risk tanımlanmıştır. As ve Pb 'ye maruz kalma durumu, kanser türleri ile ilişkili bulunmuştur. Açıklanan kimyasal ve risk değerlendirme sonuçları göz önüne alındığında toprakların agrega malzemeleriyle kaplanması tavsiye edilmiştir. Çalışma halk sağlığı açısından çevresel faktörlerin düzenlenebileceğini göstermiştir [32].

Aljumah, Ahamad ve Siddiqui, genç ve yaşlı diyabet hastalarının tedavisi üzerine veri madenciliği çalışması yapmıştır [33]. Bu çalışma regresyona dayalı bir veri madenciliği tekniği kullanılarak diyabet tedavisinin tahmine dayalı analizine odaklanmaktadır. Oracle Data Miner (ODM) yazılımı üzerinden Destek Vektör Makine algoritması kullanılmıştır. Veri seti, farklı tedavi türlerinin etkinliğini belirlemek için analiz edilmiştir. Beş farklı yaş grubu gençten yaşlıya gruplara ayrılmıştır. Gruplar için tedavi tercihleri belirlenmesi amaçlanmıştır. Genç yaş grubundaki hastalarda ilaç tedavisinin geciktirilebileceği sonucuna varılmıştır. Yaşlı hastalar için ilaç tedavisinin gecikmeden başlaması gerektiği anlaşılmıştır. Bu çalışma yaşlı hastaların hastalık seyrinin değerlendirilmesinde ve tedavi planlanmasında kullanılabilecektir [33].

Kılınç, çalışmasında veri madenciliği teknikleri kullanarak Manyetik Rezonans (MR) görüntü verileri ile Alzheimer hastalığı teşhisi koymayı amaçlamıştır [34]. Veri seti olarak 55-91 arası yaş grubunda toplam 319 hastanın (175 alzheimer hastası, 144 kontrol grubu) verisi kullanılmıştır. Çalışma Waikato Environment for Knowledge Analysis (WEKA) platformunda gerçekleştirilmiştir. BayesNet, K-NN, Random Tree, NaiveBayes, J48, Decision Table, Rastgele Orman gibi sınıflandırma algoritmaları kullanılmıştır. Rastgele orman %85 doğruluk ile en iyi sonucu vermiştir. Çalışma ile Alzheimer hastalığına doğru teşhis konulacak klinik karar destek sistemi oluşturulabilecektir [34].

Hassani, çalışmasında farklı antibiyotiklerin birlikte kullanımının hastalığın tedavisini olumlu yönde etkilediği bilgisi ile antibiyotiklerin birlikte kullanımı ve antibiyotik doz optimizasyonunu veri madenciliği yöntemleri ile tespit etmeyi amaçlamıştır [35]. Yapılan çalışmalarda E. Coli veri setinde çalışmanın performansını test etmiştir. Çalışmada genetik ve sınıflandırma algoritmaları kullanılmıştır. Kimyasal genomiğe dayanan GRASP algoritması %94 doğruluk ile E. Coli veri setinde en iyi sonucu vermiştir. Bu çalışma ile hastalar için doğru antibiyotiklerin doğru dozda alınması sağlanabilecektir [35].

Çalışmalarda uygulanan veri madenciliği yöntemleri Tablo 1'de, yapılan çalışmanın kaynak referans kodu, çalışmanın orijinal adı ve çalışmada uygulanan veri madenciliği yöntemleri başlıkları ile belirtilmiştir.

Tablo 1. Sağlık sektöründe uygulanan veri madenciliği çalışmaları ve yöntemleri

ÇALIŞMA REFERANS NO	YAZARLAR	ÇALIŞMA ADI	ÇALIŞMA YILI	ÇALIŞMADA KULLANILAN VERİ MADENCİLİĞİ YÖNTEMİ
26	Hong M, Lu M, Lu C, Zhu Y.	Association analysis of the clinical medical case-set based on the data mining in lung cancer	2022	Sınıflandırma Yöntemleri ve Birliktelik Analizi
27	Kaur I, Doja MN, Ahmad T	Data mining and machine learning in cancer survival research: an overview and future recommendations	2022	Sınıflandırma Yöntemleri
28	Kirlidog M, Aşuk C	A fraud detection approach with data mining in health insurance	2012	Anomali Tespiti
29	Santos RS, Malheiros SMF, Cavalheiro S, de Oliveira JMP	A data mining system for providing analytical information on brain tumors to public health decision makers	2013	Kümeleme Yöntemleri, Sınıflandırma Yöntemleri ve Birliktelik Analizi
30	da Costa NL, de Sá Alves M, de Sá Rodrigues N, Bandeira CM, Oliveira Alves MG, Mendes MA	Finding the combination of multiple biomarkers to diagnose oral squamous cell carcinoma	2022	Sınıflandırma Yöntemleri
31	Wang C-H, Nguyen PA, Li YC, Islam MM, Poly TN, Tran Q-V	Improved diagnosis-medication association mining to reduce pseudo-associations	2021	Sınıflandırma Yöntemleri
32	Parviainen A, Vázquez-Arias A, Arrebola JP, Martín-Peinado FJ	Human health risks associated with urban soils in mining areas	2022	Sınıflandırma Yöntemleri
33	Aljumah AA, Ahamad MG, Siddiqui MK	Diabetes health care in young and old patients	2013	Sınıflandırma Yöntemleri
34	Kılınç Ü.	Classification of brain MR image data using data mining techniques	2019	Sınıflandırma Yöntemleri
35	Hassani M	Predicting drug synergy using data mining	2016	Sınıflandırma Yöntemleri

5. Sonuç ve değerlendirme

Veriyi iyi kullanan işletmeler kaynakların verimliliğini ve etkinliğini artırabilir, doğru stratejiler kurabilir ve hedefler verebilir. Sağlık sektörü, kişinin anne rahmine düşmesinden ölümüne kadar olan süreçte birçok veri elde edilen bir sektördür. Bu veriler devasa boyutlara ulaşabilmektedir. Bu verilerin kalitesinin sistematik bir biçimde iyileştirilmesi önemlidir. Çünkü sağlıkta veri kalitesinin düşüklüğü yanlış kararlara sebep olabilir. Bu yanlış kararlar sakatlıklara, hatta ölüme neden olabilmektedir. Veri kalitesinin yüksek olması sürekli bir hizmet olan sağlık sektörü için hastanın, hastalık

geçmişinin doğru bilinip, bu bilgi ile doğru teşhis ve tedavi planlanması için kritiktir. Doğru teşhis ve tedavi hastanın yaşam kalitesini etkileyecektir. Örneğin acil servise başvurmuş bilinci kapalı bir hastanın geçmiş verilerinden hangi ilaçlara alerjisinin olduğunun bilinmesi veya kan ihtiyacı olan bir hastanın kan grubunun önceki sağlık kayıtların doğru olarak yer alması tedavinin en hızlı ve doğru şekilde başlamasını sağlayacaktır. Bununla birlikte hatalı sağlık verileri olan hastaya alerjisi olan bir ilaç vermek veya hastanın kan grubuna uyumlu olmayan bir kan transferi yapmak istenmeyen sonuçlara sebep olabilecektir.

Bunun yanı sıra kaliteli verinin doğru kullanımı ile hekim, hemşire ve diğer sağlık personellerinin iş gücünün doğru planlanmasına, doğru kararlar verilmesine katkıda bulunulur. Veri madenciliği yöntemleri ile büyük miktardaki sağlık verilerinde anlamlı birçok bilgi edinilir. Bu bilgiler sağlık çalışanlarına ilaç öneri sistemleri, hastalık teşhisi, sağlıkta yolsuzluğun önlenmesi, yatak kapasite planlaması gibi birçok alanda yardımcı olur. Klinik karar destek sistemleri oluşturularak hatalı karar vermelerin önüne geçilebilecek, tecrübe veya bilgi eksikliğinden doğabilecek hatalar bertaraf edilebilecektir. Ayrıca uygulanan tedavi yöntemlerinin etkinliği belirlenip sağlık sektöründeki bilimsel çalışmalara katkı sağlayacaktır. İnsan hayatında vazgeçilemeyecek olan sağlık sektörü veri madenciliği yöntemleri ile desteklenerek kaliteli hizmet sunumu sağlanabilecektir.

Sağlık sektöründe veri madenciliği kullanımına ilişkin Santos ve diğerlerinin yapmış olduğu derleme çalışmada da sağlık alanında yaşanılan zorlukların her geçen gün arttığına değinilmiştir [36]. Doğru tedavi yöntemlerinin veri madenciliği yöntemleri ile tespit edilebileceğine ve bu zorlukların aşılabileceği vurgulanmıştır. Öncelikli olarak veri madenciliği süreçleri aktarılmıştır. Daha sonra sağlık alanında yapılan doğru tedaviye ilişkin veri madenciliği çalışmaları aktarılarak birkaç farklı model seçilebileceği gösterilmiştir. Çalışmada veri madenciliği için kullanılabilecek araçlar tanıtılmıştır [36].

Rojas ve diğerleri, yapmış olduğu çalışmada sağlık sektöründe kullanılan veri madenciliği yöntemlerini incelemiştir [37]. 74 makale incelemiş olup hastane bilgi sistemleri için analiz stratejilerinde en ortak kullanılan yöntemler belirlenmiştir. Makalenin faydalı bir geniş bakış açısı sağlayabileceği, bu alanda yürütülen makalede yer verilen çalışma örneklerinde daha sonra yapılabilecek çalışmalar bakımından uygun metodun seçiminde faydalı olabileceği değerlendirilmiştir [37].

Srivastava ve diğerleri, 2004-2020 yılları arasında Parkinson hastalığının tanı ve tedavi yöntemlerine ilişkin kullanılan 159 veri madenciliği çalışmasını incelemiştir [38]. Parkinson hastalığına ilişkin erken tanı tahmini ve Parkinson hastalığı tedavi yönetimi uygulamalarını değerlendirmiştir [38].

Karataş ve diğerleri, gelişen teknolojik imkanlar ile oluşan büyük verinin sağlık sektöründe sağlık kaynaklarının etkin yönetimi, klinik bakım süreçleri, hizmet planlaması, sağlık hizmetlerinin sunumu ve değerlendirilmesi için kullanabilecek çalışmaları incelemiştir [39]. E-sağlık sistemleri, akıllı uygulamalar, hastalık bazlı çalışmaları incelemiştir [39].

Bu çalışmada diğer çalışmalardan farklı olarak veri kalitesi ve veri madenciliğini bütüncül olarak ele almıştır. Veri madenciliğinin en kritik aşamalarından biri veri temizliği kavramı veri kalitesi ile yakın ilişkilidir. Veri kalitesi kavramının doğru anlaşılması veri madenciliği çalışmalarının kalitesini doğrudan etkileyecektir. Bu çalışma genel bir bakış açısı sağlamakla birlikte sağlık alanında yapılabilecek yeni çalışmalar açısından örnek çalışma içerikleri ile gelecek çalışmalara ışık tutacaktır.

6. Yazar katkı beyanı

Makalenin kapsam, içerik, yazım ve düzenleme kısmı Ahmet KOÇAK tarafından üstlenilmiş olup makalenin ilerleme metodolojisini Mehmet Ali ERGÜN üstlenmiştir.

7. Etik kurul onayı ve çıkar çatışması beyanı

Çalışmada herhangi bir etik komisyon onayına ihtiyaç duyulmamıştır.

8. Kaynaklar

[1] Doger Ş. Veri Kalitesinde Eksik Veri Sorunlarının Derin Öğrenme Yöntemi İle Çözülmesi: Üretici Çekişmeli Ağlar İle Bir Uygulama. Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi, İzmir, Türkiye, 2020.
[2] Liu Q, Feng G, Zhao X, Wang W. "Minimizing the data quality problem of information systems: A process-based method". *Decision Support Systems*, 137, 113381, 2020.
[3] Shi G. *Chapter 1.* Data mining and knowledge discovery for geoscientists*,* 1-22, Elsevier, 2013.
[4] Han J, Pei J, Kamber M. *Data Mining: Concepts and Techniques,* The Morgan Kaufmann Series in Data Management Systems Book, 1-38, 2012.
[5] McCord SE, Welty JL, Courtwright J, Dillon C, Traynor A, Burnett SH et al. *Ten practical questions to improve data quality. Rangelands*, 44(1), 17-28, 2022.
[6] Gualo F, Rodriguez M, Verdugo J, Caballero I, Piattini M. "Data quality certification using ISO/IEC 25012: Industrial experiences". *Journal of Systems and Software*, 176, 110938, 2021.
[7] Olson JE. *Chapter 1. Data quality: the accuracy dimension*, 3-23, Elsevier, 2003.
[8] Daneshkohan A, Alimoradi M, Ahmadi M, Alipour J. "Data quality and data use in primary health care: A case study from Iran". *Informatics in Medicine Unlocked*, 28, 100855, 2022.

[9] Rajan NS, Gouripeddi R, Mo P, Madsen RK, Facelli JC. "Towards a content agnostic computable knowledge repository for data quality assessment". *Computer Methods and Programs in Biomedicine*, 177, 193-201, 2019.
[10] UNECE *Sustainable development GOALS*. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2000/11/metis/crp.3.e.pdf (14.12.2022).
[11] Türkiye İstatistik Kurumu [TUİK]. "TÜİK Kalite Güvence Çerçevesi Belgesi". https://www.tuik.gov.tr/Kurumsal/PDF_Detay (23.08.2022).
[12] Wang RY & Strong DM. "Beyond accuracy: What data quality means to data consumers". *Journal of management information systems*, 12(4), 5-33, 1996.
[13] Redman TC. *Data quality for the information age*. Artech House, Inc. 1997.
[14] Dünya Sağlık Örgütü [DSÖ]. "Uluslararası Hastalık Sınıflandırılması, 2022". https://www.who.int/classifications/classification-of-diseases (20.05.2022).
[15] Dünya Sağlık Örgütü [DSÖ]. "Anatomik Terapotik Kimyasal Kodu [ATC] 1948". https://www.who.int/classifications/classification-of-diseases (06.06.2022).
[16] T.C. Çalışma ve Sosyal Güvenlik Bakanlığı Sosyal Güvenlik Kurumu. "Sağlık Uygulama Tebliği, 2013". https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=17229&MevzuatTur=9&MevzuatTertip=5 (21.05.2022).
[17] Mantıksal Gözlem Tanımlayıcı Adları ve Kodları [LOINC]. https://loinc.org/ (21.05.2022).
[18] Küresel Medikal Cihaz Sınıflandırma (GMDN). https://www.gmdnagency.org/ (21.05.2022).
[19] Han J, Kamber M. and Pei J. *Chapter 1 Introduction. Data Mining: Concepts and Techniques*. Third Edition, 1-38, The Morgan Kaufmann Series in Data Management Systems Book, 2012.
[20] Karimi HA. *Big Data: techniques and technologies in geoinformatics*, 2, Crc Press, 2014.
[21] Zhao Y, Cen Y. *Data mining applications with R. Academic Press*, 35-7, 2013.
[22] Losarwar V, Joshi DM. "Data preprocessing in web usage mining". *International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July* (pp. 15-16). Chapter 3, 88-113, 2012.
[23] Bekki, A. *Sağlık Alanında İstatistik*, T.C. Anadolu Üniversitesi Yayını No:3238, 104-106, 2019.
[24] Frank E, Hall MA. *Chapter 4.* Data mining: practical machine learning tools and techniques, 124-127, Morgan Kaufmann, 2011.
[25] Han J, Pei J, Kamber, M. *Chapter 12. Data mining: concepts and techniques*, 543-550, Elsevier, 2011.
[26] Hong M, Lu M, Lu C, Zhu Y. "Association analysis of the clinical medical case-set based on the data mining in lung cancer". *Asian Journal of Surgery*, 45(5), 1158-1159, 2022.
[27] Kaur I, Doja MN, Ahmad T. "Data mining and machine learning in cancer survival research: An overview and future recommendations". *Journal of Biomedical Informatics*, 128, 104026, 2022.
[28] Kirlidog M, Aşuk C. "A Fraud Detection Approach with Data Mining in Health Insurance". *Procedia - Social and Behavioral Sciences*, 62, 989-94, 2012.
[29] Santos RS, Malheiros SMF, Cavalheiro S, de Oliveira JMP. "A data mining system for providing analytical information on brain tumors to public health decision makers". *Computer Methods and Programs in Biomedicine*, 109(3), 269-82, 2013.
[30] da Costa NL, de Sá Alves M, de Sá Rodrigues N, Bandeira CM, Oliveira Alves MG, Mendes MA, et al. "Finding the combination of multiple biomarkers to diagnose oral squamous cell carcinoma – *A data mining approach". Computers in Biology and Medicine*, 143, 105296, 2022.
[31] Wang C-H, Nguyen PA, Li YC, Islam MM, Poly TN, Tran Q-V, et al. "Improved diagnosis-medication association mining to reduce pseudo-associations". *Computer Methods and Programs in Biomedicine*, 207, 106181, 2021.
[32] Parviainen A, Vázquez-Arias A, Arrebola JP, Martín-Peinado FJ. "Human health risks associated with urban soils in mining areas". *Environmental Research*, 206, 112514, 2022.
[33] Aljumah AA, Ahamad MG, Siddiqui MK. "Application of data mining: Diabetes health care in young and old patients". *Journal of King Saud University-Computer and Information Sciences*, 25(2), 127-36, 2013.
[34] Kılınç Ü. Classification of brain MR image data using data mining techniques. Yüksek Lisans Tezi, Adana Bilim ve Teknoloji Üniversitesi, Adana, Türkiye, 2019.
[35] Hassani M. Predicting drug synergy using data mining. Doktora Tezi, Sabancı Üniversitesi, İstanbul, Türkiye, 2016.
[36] Santos-Pereira, Judith, Le Gruenwald, and Jorge Bernardino. "Top data mining tools for the healthcare industry". *Journal of King Saud University-Computer and Information Sciences,* 34(8), 4968-4982*,* 2022.
[37] Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. "Process mining in healthcare: A literature review". *Journal of biomedical informatics*, 61, 224-236, 2016.
[38] Srivastava AK, Jeberson K, and Jeberson W. "A systematic review on data mining application in Parkinson's disease". *Neuroscience Informatics*, 100064, 2022.
[39] Karatas M, Eriskin L, Deveci M, Pamucar D, Garg H. "Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives". *Expert Systems with Applications*, 116912, 2022.