I. KISIM - PART I

VERİ

DATA

1. BÖLÜM / CHAPTER 1

VERİ BİLİMİ

DATA SCIENCE

Serra ÇELİK*

*Dr., İstanbul Üniversitesi, Enformatik Bölümü, İstanbul, Türkiye
E-mail: serra.celik@istanbul.edu.tr

DOI: 10.26650/B/ET07.2021.003.01

ÖZ

Günümüzde veri kelimesi hemen her alanda kendine yer bulmuştur. Hemen hemen her alanda toplanması, saklanması ve işlenmesi hayati önem taşımaktadır. Veri tek başına hiçbir şey ifade etmez iken ancak işlenmeye başladığında konuşmaya başlar. Önemli olan nasıl işleneceğini bilmektir. Amaç; veri analizi yoluyla karar vermeyi geliştirmeye odaklanmaktır. Bu bağlamda bölümde ilk önce veri ve veri bilimi tanımları yapılmış, veri bilimi, makine öğrenmesi ve veri madenciliği arasındaki ilişki irdelenmiştir. Ardından veri bilimi problemleri ele alınmıştır. Veri bilimi problemleri tanımlayıcı, öngörücü ve pekiştirmeli olarak üç ana başlıkta toplanabilmekle beraber bu bölümde ilk ikisi incelenmiştir. Veri bilimini bir süreç olarak tanımlamak yanlış olmayacaktır. Bu süreç özetle, hedef seçimi, veri ön işleme, model oluşturma ve sonuçların analizi aşamalarından oluşur. Hedef seçimi veri bilimi sürecinin en önemli aşamasıdır. İhtiyacı anlamaya ve problem çözmeye odaklanılmalıdır. Veri ön işleme ise sürecin en çok zaman alan ve en çok dikkatli olunması gereken aşamasıdır. Uygun veri setinin elde edilerek amaca ulaşmada kullanılacak yöntemlere hazır hale getirilmesi bu aşamada gerçekleştirilir. Model oluşturma aşaması ön işlenmiş veriye bilgi çıkarım algoritmalarının uygulanması olup son aşama olan sonuçların analizi aşamasında algoritma sonuçları değerlendirilir. Bazı veri bilimi problemleri tanımlayıcı ya da öngörücü olarak sınıflandırılamayabilir. Standart dışı problemler olarak adlandırılan bu tarz problemler; türev problemler ve hibrit problemler olarak iki sınıfta incelenebilir. Bölümde bu problemlere örnekler de verilmiştir. Veri biliminde önyargı ve etik gibi insanlığı doğrudan ilgilendiren konulara da bölümde yer verilmiş olup, tıbbi veri bilimi ile bölüm sonlandırılmıştır. Bu bağlamda bu bölümde verinin anlamlı bilgiye dönüşmesi üzerine bir yolculuğa çıkılacak olup bu çalışma, veri bilimi ile yeni tanışacaklar için bir rehber olma özelliği taşımaktadır.

Anahtar Kelimeler: Veri, veri bilimi, tıbbi veri bilimi

ABSTRACT

Today, data have found and secured its place in almost every field. It is vital to collect, store, and process data in nearly every area. Although data alone mean nothing, data only begin to speak when processed. The most important thing is to know how to process data. The goal of this chapter is to focus on improving decision-making through data analysis. In this context, this chapter first defines data and data science, and discusses the relationship among data science, machine learning, and data mining. Then, it explores data science problems. Data science

problems can be grouped under three main titles, namely descriptive, predictive, and reinforcement, of which the first two are examined in this chapter. It would not be wrong to define data science as a process. This process comprises goal selection, data preprocessing, model construction, and analysis of results. Goal selection is the most crucial step of the data science process. The focus should be on understanding the need and solving problems. In contrast, data preprocessing is the most time consuming procedure and must be monitored throughout the process. At this step, the appropriate data set is obtained and prepared for the methods to be used in achieving the goal. The model construction step is the application of information extraction algorithms to the pre-processed data. The algorithms results are evaluated in the final step, which is the analysis of the results. Some data science problems may not be classified as either descriptive or predictive. These problems are called non-standard problems, which can be examined in two classes, namely derivative and hybrid problems. This chapter provides the examples of these problems. It also includes issues directly related to humanity, such as bias and ethics in data science. The chapter has been finalized with medical data science. In summary, this chapter takes a journey to transform data into meaningful information, and will guide those who will be newly discovering data science.

Keywords: Data, data science, medical data science

1. Veri ve Veri Bilimi

Dilimizde "Veri" olarak karşılık bulan "Data" kelimesinin orijinini incelediğimizde Latince "verilen şey" anlamına gelen "datum" kelimesinden geldiği görülmektedir. Veri (data) kelimesi 1500'lü yılların başında kullanılmaya başlanmış, modern anlamda kullanılışı ise elektronik işlemciler sayesinde (veri girişi, veri işleme ve çıktı verisi üretme özelliklerine sahipliklerinden dolayı) 1940'lı ve 1950'li yıllarda gerçekleşmiştir (Stanton, 2013). Günümüzde veri, yüzyılın yakıtı olarak (Economist, 2017) ya da yeni bir sermaye biçimi olarak (Corea, 2019) tanımlanmaktadır. İşletmeler için rekabet avantajı sağlayabildiği gibi sağlık alanında hastalıkların kısa sürede doğru teşhisine yardımcı olabilir. Eğitim sisteminde öğrencilerin başarılarını artırabileceği gibi tarımda ürün kalitesini geliştirebilir. Burada problemimiz bu saydıklarımızın nasıl gerçekleşeceğidir. Veri tek başına hiçbir şey ifade etmez. Witten ve diğerlerine (2011) göre akıllıca analiz edilmiş veri değerli bir kaynaktır. Veri, ancak işlenmeye başladığında konuşmaya başlar. Bu işlenme durumu ise "Veri Bilimi" dediğimiz bir süreç ile nihai sonucuna ulaşabilecektir. SINTEF (2013) araştırmasına göre Dünya'da üretilen verinin yüzde doksanı son iki yılda üretilmiştir. Her geçen yıl daha da fazla veri üretilmektedir. Önemli olan bu veri yığını içinde doğru veriyi yakalayıp anlamlı sonuçlar üretecek süreçleri yönetmektir.

Veri kaydedilmiş gerçekler olarak ifade edilirse, enformasyon verinin altında yatan örüntüler ya da beklentiler setidir. Bilgi (*knowledge*) beklentiler setinin birikimi, bilgelik (*wisdom*) ise bilgiye bağlı değer olarak tanımlanabilir (Witten, Frank ve Hall, 2011). Verinin bilgiye dönüşmesi ve kararları nasıl etkilediğine örnek olarak Plüton'un 2006 yılında gezegenlikten

ve de güneş sistemi üyeliğinden çıkarılmasını verebiliriz. Öncelikle gezegen tanımına bakmakta fayda vardır. Eski Yunanda Merkür, Venüs, Mars, Jüpiter, Satürn, Ay ve Güneş gezegen olarak tanımlanmaktaydı. Gök cisminin gökyüzünde hareket etmesi ve parlak olması gezegen olarak tanımlanması için yeterliydi. Dünya bu listede yoktu ve gezegenler birbirinden tamamen farklıydı. Ay'ın Satürn'den ne kadar farklı olduğu bilinmiyordu çünkü tek teknoloji olan evreni gözlemlemek insan gözüyle sınırlıydı. Astronomlar teleskopu icad edince güneş sistemi tanımı yeniden düzenlendi. Güneş ve Ay gezegenlikten çıkarılarak Jüpiter, Satürn, Merkür, Venüs ve Mars'ın yanına Dünya da gezegen olarak eklendi. Gezegen tanımı da değişti. Gök cisminin Güneş'in etrafında dönmesi gezegen olarak sınıflandırılması için yeterliydi. Teleskoplar daha da büyüyüp, teknolojileri geliştikçe yeni gezegenler de keşfedildi. Uranüs, Neptün ve Plüton sırasıyla 1781, 1846 ve 1930 yıllarında keşfedildiler. Gözlemler devam ettikçe elde edilen veriler ışığında gezegenlerin birbirlerinden çok farklı oldukları hatta Plüton'un Jüpiter'in dört uydusundan dahi çok küçük olduğu görüldü. 1801 yılında astronomlar Mars ve Jüpiter arasında yeni bir gezegen buldu: Ceres. Sonraki yıl aynı bölgede yeni bir gezegen daha bulundu: Pallas. Sonraki yıllarda aynı bölgede iki gezegen daha keşfedildi: Juno ve Vesta. Güneş Sistemindeki gezegen sayısı artmıştı. Astronomlar gözlemlere devam ettikçe daha fazla gezegen keşfettiler ve bunların gezegen olarak tanımlanamayacağı görüşünde uzlaşılarak yeni bir kategori ortaya çıkardılar: Asteroid kuşağındaki asteroidler. Plüton'a gelindiğinde ise, astronomlar 2005 yılında Plüton'un bulunduğu bölgede ona benzer buzlu yapıda ve büyüklükte olan Eris'i keşfetti. Güneş sistemi tekrar kategorize edildi ve içinde Plüton'un da olduğu uzak gök cisimler Kuiper Kuşağı olarak gruplandırıldı. Plüton da Güneş sistemindeki gezegenlerden daha çok Kuiper kuşağındaki gök cisimlerine benzediğinden dolayı güneş sistemi üyeliğinden çıkarıldı (Grey, 2012). O ve onun benzeri olan gezegenler de cüce gezegen olarak sınıflandırıldı. Bu örnekte de görüleceği üzere elimizdeki veriye göre bilgi çıkarma süreci sonucu kararlarımız ortaya çıkmaktadır. Gelişen teknolojiyle daha doğru, daha ayrıntılı ve daha da çok veri toplayabiliyoruz. Önemli olan bu veriyi doğru bir şekilde yönetip kararlarımıza doğru bir şekilde entegre edebilmektir.

Verinin nasıl kaydedildiğini inceleyecek olursak, ilk Sümerler zamanında MÖ 4000 yıllarında varlık ve vergi kayıtları için tabletlerin veri kaydedici olarak kullanıldığı görülmektedir. Yazılı bir dilleri olmaksızın tabletlerden farklı olarak İnkalar MS 1200-1500 yılları arasında özellikle askeri amaçlı veri kaydı tutabilmişlerdir. İnkalar, karmaşık bir mesaj verisini kaydetmek için renkli ve de düğümlü pamuk kordonlardan oluşmuş ip topluluğu olan *quipu*'lar ile veriyi kayda alabilmişlerdir. Kordonun türü, rengi, düğümlerin biçimi ise farklı anlamlar taşımıştır (Beynon-Davies, 2007). Veri toplama ve depolamada en önemli dönüm noktası ise

Edgar F. Codd'un 1970 yılında veri tabanlarında verinin nasıl saklanacağı, dizinleneceği ve de çekileceğini sunarak, ilişkisel veri modelini açıkladığı yayındır. Son 150 yılda, elektronik sensörlerin gelişimi, verinin dijitalleşmesi ve bilgisayarın icadıyla toplanan ve depolanan verinin miktarında çok büyük artışlar yaşanmaktadır (Kelleher ve Tierney, 2018). Dünya'daki veri tabanlarında depolanan veri miktarı ise her 20 ayda bir iki katına çıkmaktadır (Witten ve diğerleri, 2011).

Verinin bilgiye dönüşme yolculuğuna göz atıldığında, veri üretiminin hızlandığı 1980'li yılların sonunda, veri tabanlarından bilgi keşfi (*Knowledge Discovery in Databases*) olarak adlandırılan yeni bir araştırma alanın doğduğu görülmektedir (Han, 2011). İlerleyen yıllarda ise veri tabanlarından bilgi keşfi süreci alanında farklı isimler ortaya çıkmıştır. Bunlardan bazıları; Bilgi Keşfi (*Knowledge Discovery*), Veri Keşfi (*Data Discovery*), Enformasyon Keşfi (*Discovery Information*), Bilgi Çıkarımı (*Knowledge Extraction*), Veri Çıkarımı (*Data Extraction*), Örüntü Keşfi (*Pattern Discovery),* Veri Madenciliği (*Data Mining*), Veri Bilimi (*Data Science*) olarak sayılabilmektedir (Fernández ve diğerleri, 2018). Son yıllarda ise yaygın olarak kabul gören isimler Veri Madenciliği ve Veri Bilimidir (Aggarwal, 2015; Steele, Chandler ve Reddy, 2017).

Nisbet (2018) veri bilimi tanımını yapmadan önce istatistiksel modelleme, veri madenciliği ve bilgi keşfinin tanımı yapmayı tercih eder:

- İstatistiksel modelleme: öngörücü değişkenlere (*predictor variables*) dayalı olarak bir sonucu veya olayı gruplandırmak veya tahmin etmek için parametrik istatistiksel algoritmaların kullanılması,
- Veri madenciliği: büyük, gürültülü ve dağınık veri setlerindeki veri öğeleri arasındaki zayıf ilişki modellerini bulmak için (tıbbi tanı koyma, kar artırma, sahtecilik tespiti gibi alanlarda faydayı artıracak) öğrenme algoritmalarının kullanılması,
- Bilgi keşfi: tüm veri erişimi, veri keşfi, veri hazırlama, modelleme, model dağıtımı ve model izleme sürecidir.

Bu tanımların ışığında veri bilimi; bilgi keşfinin bir yandan analitik veri merkezlerinin veri mimarisine genişletilmesi, diğer yandan da oldukça gelişmiş makine öğrenme algoritmalarıyla karmaşık görüntü, konuşma ve metin analizleri yapılabilmesi olarak tanımlanır (Nisbet, 2018). Stanton (2013) veri bilimini, büyük bilgi yığınlarının derlenmesi, hazırlanması, analizi, görselleştirilmesi, yönetimi ve korunmasıyla ilgilenen bir çalışma alanı olarak tanımlamıştır. Pathak (2014) ise veriden içgörü çıkartma metodolojisini veri bilimi olarak tanımlamaktadır.

Igual ve Seguí (2017) de benzer şekilde veri bilimini, eyleme geçirilebilir öngörülerin verilerden çıkarılabileceği bir metodoloji olarak tanımlar ki bu iş zekası veya keşfedici istatistik gibi veri analizine yönelik önceki yaklaşımlar açısından ince ama önemli bir farktır. Dinov (2018) ise veri bilimini (1) teorik, hesaplamalı, deneysel ve biyososyal (*biosocial*) alanlar arasında köprü görevi gören, (2) farklı kaynaklardan büyük miktarlarda karışık, uyumsuz ve dinamik veriyle baş edebilen, (3) yarı otomatik karar destek sistemleri oluşturabilen algoritmalar, yöntemler, araçlar ve hizmetler geliştiren disiplinler ötesi bir alan olarak tanımlamaktadır. Kelleher ve Tierney (2018)'a göre veri bilimi, büyük veri setlerinden kullanışlı örüntüler çıkartmak için ilkeler, problem tanımları, algoritmalar ve süreçleri kapsar. Günümüzde Veri Bilimi; matematik, istatistik, enformatik, bilgisayar bilimleri gibi çoğu alandan teori ve yöntemleri içine alan, veri madenciliği ve veri tabanlarından bilgi keşfi terimlerini genelleştiren yeni bir disiplin olarak tanımlanmaktadır (Fernández ve diğerleri, 2018).

Veri bilimi unsurlarının çoğu makine öğrenmesi ve veri madenciliği gibi ilgili alanlarda geliştirilmiştir. Aslında veri bilimi, makine öğrenmesi ve veri madenciliği terimleri genellikle birbirlerinin yerine kullanılır. Bu disiplinler arasındaki ortak nokta, verilerin analizi yoluyla karar vermeyi geliştirmeye odaklanmaktır. Bununla birlikte, veri biliminin diğer alanlara göre kapsamı daha geniştir (Kelleher ve Tierney, 2018). Makine öğrenmesi, performans geliştirmek ya da doğru öngörüler elde etmek için deneyimleri kullanan hesaplama yöntemleri olarak tanımlanabilir (Bach, 2018). Makine öğrenmesi veriden örüntü çıkarımı için tasarım ve algoritmalara odaklanır. Veri madenciliği genellikle yapılandırılmış verilerin analizini ele alır ve ticari uygulamalar ile ilgilenir. Veri bilimi, makine öğrenmesi ve veri madenciliğinin tüm konularını dikkate almakla birlikte, yapılandırılmamış sosyal medya ve web verilerinin yakalanması, temizlenmesi ve dönüştürülmesi ile büyük, yapılandırılmamış veri kümelerini depolamak ve işlemek için büyük veri teknolojilerinin kullanılması ve veri etiği ile de ilgilenir (Kelleher ve Tierney, 2018).

2. Veri Bilimi Problemleri

Veri biliminde her problem türü için uygun algoritma saptamak oldukça zordur. Bunlara rağmen, çözülmüş (hedef) probleme ve veri türüne göre uygulanabilen en iyi algoritma türlerini bulmada bir rehber oluşturmak mümkündür. Problemlerin ilk ve genel olarak sınıflandırılması, tanımlayıcı problemler (*descriptive problems* – gözetimsiz öğrenme) ve öngörücü problemler (*predictive problems* - gözetimli öğrenme) arasında ayrım yapmamıza neden olacaktır (Fernández ve diğerleri, 2018). Gözetimli öğrenme, olası tüm girdi setlerini genelleştirmek için, etiketlenmiş örneklerden oluşan bir eğitim setinden öğrenen algoritmalar iken gözetimsiz öğrenme, etiketsiz örneklerin oluşturduğu bir eğitim setinden öğrenen algoritmalardır (Igual

ve Seguí, 2017). Tanımlayıcı modelde amaç, değişkenler arasındaki ilişkileri tanımlayarak veri yapısı hakkında daha fazla bilgi edinmek iken öngörücü modelde amaç, diğer değişkenler göze alınarak bir değişkenin değerini öngörmektir (Pathak, 2014).

Gözetimli ve gözetimsiz öğrenmeden farklı olarak üçüncü bir öğrenme problemimiz ise Pekiştirmeli Öğrenmedir (*Reinforcement Learning*). Bir çözümün kalitesi hakkında bilgi sağlayan ancak onu nasıl geliştireceğine dair değil, eleştiriden güç alarak öğrenen algoritmalardır. Çözüm alanını tekrar tekrar keşfederek geliştirilmiş çözümler elde edilir (Igual ve Seguí, 2017). Bu kitap bölümünde gözetimli ve gözetimsiz öğrenme ile ilgili bilgiler verilmiş, pekiştirmeli öğrenme kapsam dışı bırakılmıştır.

2.1. Tanımlayıcı Problemler

Burada amaç çalışılan verinin bir tanımını yapabilmektir. Bu tarz problemlere örnek olarak bir organizasyonun müşterilerinin özellikleri hakkında bilgi sahibi olmak, birlikte satın alınan ürünleri bulmak ya da benzer semptomlara sahip hastalıkların belirlenmesi gösterilebilir. Bu problemlerin amacı kaynak veri setini tanımlamaktır. Tanımlayıcı problemi iki şekilde detaylandırmak mümkündür: kümeleme analizi (*Clustering Analysis*) ve birliktelik analizi (*Association Analysis*). Kümeleme analizinde amaç kaynak popülasyonda homojen grupları bulmaktır. Birliktelik analizinde amaç ise veri tabanında değişken değerleri ilişkileri belirlemektir. Market sepeti analizi en bilinen analizdir. (Fernández ve diğerleri, 2018).

2.2. Öngörücü Problemler

Veri bilimi, davranışları öngörmek için uygulanacak modeli belirlemede de problemlere sahiptir. Bu türdeki problemler öngörme problemleri olarak anılmakta olup Yapay Zeka (*Artificial Intelligence*) çevrelerinde ise gözetimli öğrenme ismini almaktadır. Gözetimli öğrenme ismini almasının nedeni ise araştırmacının istenen cevabı sağlamasındandır. Öngörülen değişken, kategorik değişken (ürünü satın aldı ya da almadı) olabileceği gibi nümerik de (ödemeyi geciktirme olasılığı) olabilmektedir. Modelin öngördüğü değişken türlerindeki bu ayrım öngörücü problemleri sınıflandırma problemleri (*Classification Problems*) ve regresyon problemleri (*Regression Problems*) olarak ayırt etmemize neden olur. Sınıflandırma problemlerinde öngörü değişkeni sonlu sayıda değere sahip kategorik değişkendir. Regresyon problemleri ise öngörülen değişkenin nümerik olduğu problemlerdir. Bazı semptomları tanımlanma ya da bir hastalık gösterme olasılığının olması ya da olmaması örnek verilebilir (Fernández ve diğerleri, 2018).

3. Veri Bilimi Süreci

Veri tabanlarından bilgi keşfini bir süreç olarak tanımlamak yanlış olmayacaktır. Problem tanımlanır, problemi çözmek için uygun veri seçilir, veri temizlenir, veri dönüşümü gerçekleştirilir, modelleme yöntemleri uygulanır, oluşturulan modellerden bilgi elde edilir ve de bu bilgi kullanılır. Veri biliminin esas özelliği ise bu süreçte, verideki gizli örüntüler ve bilginin tümevarımsal olarak keşif olasılığıdır. Otomatik eğilim ve davranış tahmini ve önceden bilinmeyen örüntülerin otomatik keşfi gibi örnekler verilebilir. Veri bilimi, dört ana aşamayı içeren bir süreç olarak tanımlanabilir. Bu adımlar; hedefleri seçme, veri ön işleme, model oluşturma, sonuçları analiz etme olarak (Şekil 1) sayılabilmektedir (Fernández ve diğerleri, 2018).

Şekil 1. Veri Bilimi Süreci

Kelleher ve Tierney (2018), veri bilimi sürecindeki aktiviteleri bir piramid olarak tanımlamıştır. Şekil 2'de görüldüğü üzere piramidin her seviyesinde işlenen veri miktarı ve her katmanda erişilen bilgi farklıdır. Veri yakalama ve üretiminden veri ön işleme ve toplamaya, veri anlama ve keşfine, makine öğrenmesi kullanarak model keşfi ve model oluşturmaya kadar veri bilimi süreci resmedilmiştir.

Şekil 2. Veri Bilimi Piramidi, Kelleher ve Tierney (2018)

3.1. Hedef Seçimi

Bu aşamada mevcut problem iyi çalışılmalı ve projenin hedefine karar verilmedir. Probleme doğru yaklaşım ile veri kaynaklarını keşfetmek kolaylaşır ve en uygun algoritmalar uygulanabilir. Kötü bir yaklaşım ise yanlış sonuçlar elde edilmesine neden olur (Fernández ve diğerleri, 2018). Veri bilimi matematik ve istatistik gibi diğer alanlardan farklıdır. Veri bilimi uygulamalı bir alan olup veri kullanıcılarının ihtiyaçlarını anlamaya ve problemlerini çözmeye odaklanılmalıdır. Bir problemi çözmeden önce problemi tanımlamak gerekir (Stanton, 2013).

3.2. Veri Ön İşleme

Baesens, Vlasselaer ve Verbeke (2015) veri ön işlemeyi çöp içeri çöp dışarı (*garbage in, garbage out*) prensibiyle açıklamaktadır. Bu prensibe göre dağınık veriler, dağınık analitik modeller üretir. Daha fazla analize devam etmeden önce her veri ön işleme adımının dikkatlice yürütülmesi son derece önemlidir. En ufak bir hata bile, daha ileri analizler için veriyi tamamen kullanılamaz hale getirebilir ve sonuçlar geçersizleşir, hiçbir şekilde kullanılamaz.

Tutarsızlıklar, kayıp veri, uç değerler ve gürültü, tüm veri kümelerinin özelliklerindendir. Tamamlanmamış veri farklı nedenler için üretilmiştir örneğin ilgili nitelikler her zaman mevcut değildir ya da sahip olunan veri hatalıdır. Veri seti kayıp değerleri doldurarak, uç değerleri belirleyerek, tutarsızlıkları çözerek temizlenir. Veri, analiz için hazır olduğunda uygulanacak algoritmaya bağlı olarak kategorik ya da nümerik özellikte niteliklere ihtiyacı olmasına göre veri dönüşümü gerçekleştirilir. Veri ön işleme aşaması en fazla zaman alan aşamadır. Veri seçimi, verinin hazırlanması, verinin dönüştürülmesi ve veri indirgeme olarak dört aşamadan oluşmaktadır (Fernández ve diğerleri, 2018).

Veri seçimi: Bu aşamada iç ya da dış veri kaynakları tanımlanır ve gerekli alt veri setleri çözülecek problem tipine ve hedefe bağlı olarak seçilir. Seçilecek veri kümesi, değişkenler (nitelikler) aracılığıyla belirlenecek bir dizi örnekten oluşacağı için, her bir değişkeni anlamak için her değişkenle ilişkili meta verilerin (veri hakkında veri) analiz edilmesi gerekecektir. Meta veri, yalnızca değişkenleri tanımlamakla kalmamalı, aynı zamanda veri türleri, potansiyel değerleri, orijinal kaynakları, biçimleri ve diğer özellikleri hakkında da bilgi sağlamalıdır (Fernández ve diğerleri, 2018). Veri, en iyi değişken türlerinden anlaşılabilmektedir. Değişken türlerini sürekli ve kategorik gibi iki gruba ayırabileceğimiz gibi (Baesens ve diğerleri, 2015), nicel (*quantitative*) ve nitel (*qualitative*) olarak da ikiye ayırabiliriz. Kendi içlerinde de Tablo1'de görüldüğü gibi alt bölümlere ayrılabilmektedirler (Fernández ve diğerleri, 2018).

Tablo 1. Veri Türleri (Fernández ve diğerleri, 2018)	
Nicel	**Nitel**
Kesikli (*discrete*) *İşyerindeki çalışan sayısı, evdeki oda sayısı, ...*	Nominal *Cinsiyet, medeni durum, ...*
Sürekli (*continuous*) *Boy, ağırlık, maaş,...*	Ordinal *Değerlere göre sıralama (yüksek, orta, düşük),...*

İstatistik açısından yaklaştığımızda veri türlerini dört ölçek altında tanımlamaktayız. Nominal (Latince nomen, isim) ölçek (veri), basitçe gözlemlerin kategorilere ayrılmasıdır. Değerler arasında nümerik kıyaslamalar yoktur. Ordinal ölçek, nesnelerin ölçülen özelliğinden daha fazla veya daha azına sahip olduğuna göre nesnelerin sıralamasıdır. Aralık ölçek, ölçüm verilerinin artışlar arasında eşit adımlar içeren bir hiyerarşiyle sıralandığı bir ölçektir. Bu, ölçümlerin nicel karşılaştırmasını yapabileceği anlamına gelir (Johansson, 2016). Sıfır noktası ya da orijinden eşit aralıklarla ölçülen nümerik değerler alan değişkenlerdir (Bramer, 2016). Oran ölçek değişkenler aralık ölçeğe benzemekle birlikte farkı, sıfır noktası ölçümün olmadığını ifade eder (Bramer, 2016).

Genelde çok fazla değişken türü bulunmaktadır. Bramer (2016) veri türlerini nominal, ikili, ordinal, tam sayı, aralık ölçek ve oran ölçek değişkenler olarak altıya ayırmaktadır. İkili değişken (*binary variable*); nominal değişkenin iki mümkün değer (doğru ise 1, yanlış ise 0 değeri gibi) alabilen özel bir halidir. Tam sayı değişkeni (*integer variable*) aslında kesikli değişkendir, tam sayı değerleri alır. Sürekli ve kategorik olarak sınıflamayı tercih ettiğimizde, tam sayı, aralık ölçek ve oran ölçek değişkenleri sürekli türde değişkenler; nominal, ikili ve ordinal değişkenler de kategorik sınıfa ait değişkenlerdir (Bramer, 2016).

Verileri yapılarına göre de sınıflamamız mümkündür. Yapılandırılmış ve Yapılandırılmamış Veri olarak iki sınıfa ayrılabilmektedir. Bazı veri setleri -bir veri tabanındaki tablolar gibi- oldukça iyi yapılandırılmıştır. Veri genelde bir matris ile tanımlanır. Satırlar birimleri, kayıtları temsil ederken; sütunlar bu birimlerin özelliklerini temsil eder. Yapılandırılmamış veride ise, bu yapıdan farklı olarak, çoğu kayıt daha heterojen şekildedir. Görüntü ve bağlantılar içeren büyük bir metin topluluğu ya da kişisel tıbbi kayıtlarda görünen notların ve test sonuçlarının karmaşık bir karışımıdır (Skiena, 2017). Yapılandırılmış veri kolaylıkla saklanır, organize edilebilir, sıralanabilir, diğer yapılandırılmış verilerle birleştirilebilir. Veri bilimi sürecinin uygulanması daha kolaydır çünkü hali hazırda analizler için uygun formata dönüştürülmeye hazırdır (Kelleher ve Tierney, 2018). İşlem verisi (*transactional data*), abonelik verisi (*subscription -account- data*), sosyodemografik veri (*sociodemographic information*), veri havuzu (datapoolers) yapılandırılmış verilere örneklerdir. Yapılandırılmamış veri ise, veri

kümesindeki her örneğin kendi iç yapısına sahip olabileceği ve bu yapının her örnek için aynı olması gerekmeyen verilerdir. Örneğin bir web sayfaları setinde her web sayfası yapısı birbirinden farklı olacaktır. Yapılandırılmamış veriler, yapılandırılmış verilerden daha yaygındır. Genellikle, makine öğrenmesi yöntemleriyle yapılandırılmamış veriden yapılandırılmış veri elde edilmekle birlikte, veri dönüştürme zaman alıcı bir süreçtir (Kelleher ve Tierney, 2018). Yapılandırılmamış veriye örnek olarak sosyal medya paylaşımları, e-posta metinleri, günlük kayıtları, çağrı merkezi ses kayıtları, video kayıtları verilebilir.

Veri, boyutlarına göre de sınıflandırılabilmektedir. Temelde daha fazla veri (büyük veri) daha iyi sonuçlar üretir. Ancak uygulamada büyük veriyle çalışmak bazı zorluklar getirir. Veri büyüdükçe analiz zamanı artar ve de görselleştirilmesi zorlaşır. En basit modellerde bile veriye uyumsuzluk ortaya çıkar (Skiena, 2017). Gerçek dünya verisi, tutarsızlıklar, eksiklikler, tekrarlama, birleştirme ve diğer birçok sorun nedeniyle kirli olabilir (Baesens ve diğerleri, 2015). Modelleme aşamasına geçmeden çeşitli veri filtreleme yöntemleri uygulayarak veri temizlenip yönetilebilir bir boyuta indirgenmelidir (Baesens ve diğerleri, 2015).

Veri biliminde en kritik konulardan biri doğru veri setini bulmaktır. İhtiyacım olan veriye kim sahip? Benim kullanmama izin verirler mi? Veriye nasıl ulaşabilirim? Soruları sorularak uygun yöntemler tercih edilir (Skiena, 2017). Veri toplama yöntemlerini Avlama, Kazıma ve Günlükleme olarak üç grupta toplamak mümkündür:

- Avlama (*Hunting*): Şirketler oldukça önemli veri kaynaklarıdır. Ancak verilerini paylaşmakta cömert davranmazlar. Bunun iki nedeni vardır. İlki iş kaynaklı olarak rakiplerine yardım etme korkusu, ikinci ise mahremiyet kaynaklı müşterilerini rahatsız etme korkusudur. Buna karşın belli kısıtlamalar ile verilere ulaşma imkanı bulunmaktadır.
- Kazıma (*Scraping*): Web sayfaları genelde değerli metin ve nicel veri içerir. İhtiyaç duyulan konularda bilgiler (veri) web sitelerinden kazıma programları ile elde edilebilmektedir.
- Günlükleme (*Logging*): Bir web hizmetine ya da iletişim cihazına erişim, günlük denilen log aktiviteleri veri kaynağını oluşturacaktır. Özellikle internet of things (IoT) ile sensörlerin olduğu her alanda kolaylıkla veri toplama imkanı bulunmaktadır.

Veri birçok farklı formatta saklanıyor olabilir. En iyi hesaplanabilir formattaki verinin sahip olduğu özellikler; bilgisayarlar için ayrıştırması kolaydır, kişiler için okuması kolaydır ve diğer araç ve sistemler ile çokça kullanılabilir olmasıdır. En çok kullanılan veri formatları ise CSV (*comma separated value*), XML (*eXtensible Markup Language*), SQL (*structured query language),* JSON (*JavaScript Object Notation*) olarak verilebilir (Skiena, 2017).

Veri hazırlanması: Veri elde edildiğinde analizler için hazırlanması gerekir. Veride bazı niteliklere ait farklı isimlerle tekrarlayan birimler ya da farklı biçimlerde birimler olabilecektir. Bu problemler verinin farklı kaynaklardan gelmesi ya da aynı yöntemle saklanmamasından kaynaklanabilmektedir. Amaç seçilen verinin kaliteli olmasını sağlamaktır. Veri ne kadar iyi anlaşılır ve de hazırlanırsa veri bilimi süreci de o kadar başarılı sonuçlanır. Bu aşamada tanımlayıcı istatistik ve keşfedici veri analizi yardımcı olacaktır (Igual ve Seguí, 2017).

Tanımlayıcı istatistik yöntemler ile kategorik değişkenler için frekans dağılımlarını gözlemleyerek, histogram gibi görselleştirme yöntemleri kullanılarak kayıp veriler, uç değerler gözlemlenebilir. Nicel özellikte değişkenler için ise minimum-maksimum değerler, ortalama, varyans, mod, medyan gibi merkezi eğilim ve saçılım ölçülerine bakılarak veri yapısı anlaşılabilir (Fernández ve diğerleri, 2018).

Keşfedici veri analizi ile ise, örnek dağılımı görselleştirilir ve özetlenir, böylece ana kütle dağılımı hakkında varsayımlar yapmamıza izin verir (Igual ve Seguí, 2017). Bu aşamada genellikle görselleştirme yöntemleri kullanılmaktadır. Kutu grafikleri, pasta grafikleri, histogramlar ve QQ grafikleri veride değişken dağılımları ve de değişkenler arasındaki ilişkiler hakkında bilgi verebilecektir. (Fernández ve diğerleri, 2018). Baesens ve diğerlerine (2015) göre veri görselleştirme, veriyi en iyi anlama yollarından biridir.

Bramer (2016), veri hazırlama aşamasında değişkenin özelliklerine odaklanarak dikkat edilmesi gerekenleri dört maddede özetlemiştir:

1. Nümerik bir değişken sadece altı farklı değer alıyor alabilir. Bu durumda değişken, sürekli değişken yerine kategorik değişken olarak tanımlanabilir.
2. Bir değişkenin tüm değerleri aynı olabilir. Bu durumda değişken dikkate alınmaz.
3. Bir değişkenin biri dışında tüm değerleri aynı olabilir. Bu değerin hata mı yoksa gerçekten sahip olduğu değer mi olduğu araştırılır.
4. Bazı değerler normal değişken aralığının dışında olabilir. Bu değerler araştırılır.

Veri hazırlamada önemli olan diğer konu ise veriden gürültünün temizlenmesidir. Bir değişkendeki tesadüfi bir hata ya da varyans gürültü olarak tanımlanmaktadır. Gürültü ile etkilenen değişkenler ise uç değerler olarak veri setinde karşılanmaktadır. Uç değerler insan hatası olarak ortaya çıkabileceği gibi operasyonel değişikliklerden dolayı da ortaya çıkabilir. İleri yöntemler ile uç değerler belirlenir, ortandan kaldırılır ya da düzeltilir (Fernández ve diğerleri, 2018).

Gerçek dünya veri setlerinde sıklıkla karşılaşılan bir diğer konu ise bazı değişkenlerin bir değere sahip olmadığı kayıp veri durumudur. Verileri kaydetmek için kullanılan ekipman arızalanabilir ya da elde edilemeyen bilgiler ortaya çıkabilir (bir hastanın tüm bilgileri gibi). Örneğin bazı tıbbi veriler sadece kadın hastalar ya da belli bir yaşın üzerindeki hastalar için anlamlı olabilir. Bu noktada erkek hastalarım o değişkenle ilgili veri hanesi boş olacaktır. Kayıp verilerle baş edebilmek için birçok yöntem mevcuttur. Bu örnek için hastaları iki ya da daha fazla gruba ayırmak bir çözüm olabilir (Bramer, 2016). Diğer çözümlerden biri de kayıp verileri göz ardı etmektir. Bütün kayıp veriye sahip birimler elimine edilir (Fernández ve diğerleri, 2018). Yöntemin avantajı hatalı veriden kaçınmış olması, dezavantajı ise veriden elde edilen sonuçların geçerliliğinin hasar görmesidir (Bramer, 2016). Bir diğer yöntem de istatistik yardımıyla ortalama ya da korelasyon yardımıyla yeni değerler üretilmesidir (Fernández ve diğerleri, 2018). Nitelik sayısını azaltma da kayıp verilerle baş etmede kullanılan bir yöntem olup boyut indirgeme olarak bu süreçte karşımıza çıkar (Bramer, 2016).

Veri dönüşümü: Problem tipi ve mevcut veri tipi analiz edildiğinde, uygulanacak algoritma(lar) seçilmelidir. Her algoritma girdi verisinde farklı formata ihtiyaç duyar. Seçilen algoritmaya göre düzenlemeler yapılarak veri dönüşümü sağlanır. Bazı algoritmalar nicel özellikte veri yapısına ihtiyaç duyarken bazı algoritmalar kategorik değişkenlerle çalışabilmektedir (Fernández ve diğerleri, 2018).

Veri indirgeme: Orijinal verinin çoğu özelliği muhafaza edilerek daha küçük hacimde temsili bir veri seti oluşturularak yöntemler uygulanır. Buradaki hedef sonraki veri madenciliği algoritmalarının orijinal veriler yerine azaltılmış verilere uygulandığında aynı (veya neredeyse aynı) sonucu üretebilecek bir mekanizma sağlamaktır (Fernández ve diğerleri, 2018).

3.3. Model Oluşturma

Modelleme aşaması bilgi keşfi için önemli bir adım olup ön işlenmiş veriye bilgi çıkarım algoritmaları bu aşamada uygulanır. Bir sonraki aşama olan sonuçların analizi aşamasıyla da ilişkilidir. Sonuçların analizi aşaması sıklıkla daha büyük veri ya da daha fazla değişken için ön işleme aşamasına dönmeyi gerektirebilir. Bu aşamada ne olacağı kazanılması hedeflenen amaca bağlıdır. Nihai sonuç veriyi karakterize etmek ya da hedef sürecin daha uzun ve daha zor olduğu bir model tahmini olabilecektir. Sonuçların analizi, sürecin en önemli adımlarından biridir.

Aynı tip problemi çözebilecek çok sayıda algoritma, çok fazla karışıklığa neden olabilecektir. Bu aşamada farklı veri analizi algoritmaları veriye uygulanır. Verideki örüntüler araştırılır. Seçilen algoritmaya bağlı olarak çıktıda farklı bir form elde edilir. Aynı algoritmaların

birden fazla kullanılması ya da farklı tipte algoritmaların kullanılması mümkündür (Fernández ve diğerleri, 2018).

Öngörülecek değişken bağımlı ya da hedef değişken olarak isimlendirilirken, öngörü için kullanılan değişkenler bağımsız ya da öngörü değişkenleri olarak isimlendirilir. Tanımlayıcı ve öngörücü modeller için aynı modelleme teknikleri kullanabilir. Ancak her model için altta yatan bir modelleme varsayımı vardır, örneğin, doğrusal modellerde, değişkenler arasında doğrusal bir ilişki olduğunu varsayıyoruz. Bu varsayımlar, modeli uygulayabileceğimiz ve sonuçlarını nasıl yorumlayacağımız açısından model üzerinde sınırlamalar getirmektedir. Tıbbi ilaç etiketi üzerindeki yan etki uyarılarını bilmek gibi modelleme varsayımlarının farkında olmak da önemlidir (Pathak, 2014).

Tanımlayıcı problemler kümeleme ve birliktelik analizi olarak karşımıza çıkmaktadır. Kümeleme yöntemlerinde farklı veri yapılarına farklı algoritmalar ile yaklaşılmaktadır. Kümeleme analizi, büyük veri setlerini daha homojen küçük veri gruplarına bölerek ya da uç değerlerin tespitinde kullanılarak bir veri ön işleme yöntemi görevi de görmektedir. Grup sayısı hakkında ön bilgiye sahip olunmadığı ve küçük veri setleriyle hiyerarşik kümeleme yöntemlerinden, küme sayısı hakkında bilgi sahibi olunduğunda k-ortalamalar yöntemi iyi sonuçlar vermektedir. Büyük veri setleri için DBSCAN, CLARA algoritmalarıyla, farklı ölçekteki karışık değişkenlere sahip veri setlerinde PAM algoritmasıyla çalışmak doğru sonuçlar verecektir. Diğer tanımlayıcı problem yöntemi olan birliktelik analizinin amacını ise Fernández ve diğerleri (2018), aynı işlem içinde diğer bileşenlerin varlığını işaret eden elementleri bulmak olarak açıklamıştır. Yöntemin sonucunda "X ise Y'dir" şeklinde kurallar ortaya çıkartmak hedeflenir. X kuralın öncülü (*antecedent*), Y ise kuralın ardılıdır (*consequent*). En sık kullanılan birliktelik algoritması Aprioridir. Elementlerin bütün mümkün kombinasyonlarının yineleme sayılarına dayalıdır.

Öngörücü problemler için sınıflandırma ve regresyon modelleri örnek verilebilmektedir. Sınıflandırma modellerinde klasik gözetimli öğrenme kullanılır. Veri noktalarını (birimleri, nesneleri, gözlemleri) birden çok kategoriye veya sınıfa sınıflandırmak için bir sınıflandırma modeli kullanılır. Bu durumda, hedef değişken sınıf etiketi olarak da adlandırılabilir. İkili sınıflandırma, sınıf etiketi iki değer alan özel bir sınıflandırma türüdür. Hedef değişkenleri ikiden fazla değer alan sınıflandırma modellerine ise çok sınıflı sınıflandırma denir (Pathak, 2014). Karar ağaçları (*decision trees*), kural çıkarım (*rule induction*), örneğe dayalı öğrenme (*instance-based learning*), lojistik regresyon (*logistic regressions*), destek vektör makineleri (*support vector machines*) ve yapay sinir ağları (*artifical neural networks*) sıklıkla kullanılan sınıflandırma yöntemleridir. Sınıflandırma, uygulamada yaygın olarak kullanılır; spam filt-

releme, konuşma ve el yazısı tanıma ve biyometrik kimlik doğrulama bunlardan birkaçıdır (Pathak, 2014). Regresyon modellerinde ise nümerik değerlerin öngörüsü için, doğrusal ya da doğrusal olmayan regresyon kullanılır (Fernández ve diğerleri, 2018). Regresyon, gerçek dünya varlıkları hakkında nasıl tahminlerde bulunacağıyla ilgilidir. Satışlar, fiyatlardaki değişimle nasıl ilişkilidir? Hangi zaman aralıklarında trafik sıkışıklığından kaçınılabilir? gibi sorulara cevap aranır (Igual ve Seguí, 2017). Regresyon analizi, değişkenler arasındaki ilişkilerin tanımlanmasını amaçlar. Bir dizi bağımsız veya öngörücü değişken göz önüne alındığında, amacımız bağımlı veya hedef bir değişkeni tahmin etmektir. Regresyon analizinde hedef değişken nümeriktir ve öngörücü değişkenler nümerik, kategorik veya ordinal olabilir (Pathak, 2014).

Daha önce de bahsedildiği üzere, günümüzde en fazla üretilen veri türü yapılandırılmamış veridir. Bu türdeki veri üzerine alanlar gelişmeye devam etmektedir. Ağ analizi (*Network Analysis*), Tavsiye Sistemleri (*Recommender Systems*), Doğal Dil İşleme (*Natural Language Processing*), Görüntü İşleme (*Image Processing*) gibi çalışma alanları tanımlayıcı ve öngörücü modeller ışığında metin, görüntü ve ses verileriyle baş etmeyi sağlamaktadır. Özellikle sağlık alanında görsel verilerin analizi hastalıkların daha erken evrelerde teşhisini sağlayabilmektedir.

3.4. Sonuçların Analizi

Önceki aşamada elde edilen sonuçların değerlendirilmesi bu aşamada gerçekleşir. Sonuçlar görselleştirildiğinde, eğer beklentileri karşılamıyorsa farklı parametrelerle algoritmalar tekrar çalıştırılmalı, hatta daha iyi sonuçlar için farklı algoritmalar denenmelidir. Burada veri biliminin yinelemeli bir süreç olduğu söylenebilmektedir. Bu aşamada elde edilen sonuçları nasıl kullanacağımız belirlenmelidir. Sonuçların uzman sistemlere entegre edilebilir ya da veri tabanı yönetim sisteminde karar alma sürecine katılabilir olması hedeflenmelidir (Fernández ve diğerleri, 2018).

Veri madenciliği verideki örüntüleri keşfetme süreci olarak tanımlanmaktadır. Süreç otomatik ya da çoğunlukla yarı otomatik olabilmektedir. Keşfedilen örüntüler avantaj (özellikle ekonomik) sağlayacak şekilde anlamlı olmalıdır. Yararlı örüntüler bize yeni veri üzerinde önemsiz öngörüler yapmamızı sağlar. Bir örüntünün ifadesi için iki uç vardır. Biri doğası gereği anlaşılamayan kara kutu (*black box*), diğeri de örüntü yapısını ortaya çıkaran şeffaf kutudur (*transparent box*). Fark çıkarılan modellerin gelecekteki kararlarda kullanılabilecek olup olmamasıdır. Bazı örüntüler karar yapısını açık bir yolda yakalar bu yüzden yapısal (*structural*) olarak adlandırılır. Diğer bir deyişle veri hakkında bir şeyler açıklamaya yardımcıdır. Verideki yapısal örüntüleri bulma ve tanımlamak için yöntemler vardır. Çoğu yöntem makine öğrenmesi olarak bilinen alanda geliştirilmiştir (Witten ve diğerleri, 2011).

4. Standart Olmayan Veri Bilimi Problemleri

Bazı veri bilimi problemleri tanımlayıcı ya da öngörücü olarak sınıflandırılamayabilir. Bunun sonucu olarak, standart dışı problemler olarak adlandırabileceğimiz, çokça bilinen problemlerden bahsetmekte yarar vardır. Türev problemler (*Derivative Problems*) ve Hibrit Problemler (*Hybrid Problems*) olarak incelenebilmektedir (Fernández ve diğerleri, 2018).

Türev problemler, orijinal veri bilimi probleminin genişletilmesine ya da kısıtlanmasına dayanır. Dengesiz Öğrenme (*Imbalanced Learning*) en sık karşılaşılan türev problemdir. Verinin hedef özellik üzerinde (cevap değişkeni) istisnai bir dağılıma sahip olduğu (cevap değişkeninin örnek sayısının diğer sınıflardan çok düşük olması durumunda) bir sınıflandırma problemi olan genişletilmiş denetimli bir öğrenme olarak tanımlanabilir (Chawla, 2005). Çoğu gerçek dünya örneğinde karşılaşılan bir durumdur. Diğer türev problemlere örnek olarak; Çoklu-örnek öğrenme (*Multi-instance Learning*), Çoklu-etiket sınıflandırma (*Multi-label Classification*), Veri akışı Öğrenme (*Data Stream Learning*) verilebilir.

Hibrit problemler ise öngörücü ve tanımlayıcı görevler arasında, sınıflandırma, kümeleme ya da örüntülerin birliktelikleri arasındaki sonuçlara odaklanmıştır. En sık kullanılan hibrit problem Transfer Öğrenmedir (*Transfer Learning*). Bir veya daha fazla orijinal kaynaktan bilgi çıkarımı yapıp, elde edilen bilgiyi farklı bir hedef göreve uygulamayı hedefler (Pan ve Yang, 2010). Geleneksel öğrenme algoritmaları, eğitim verilerinin ve test verilerinin aynı kaynaktan alındığını ve aşağı yukarı aynı dağılım ve nitelik uzayını koruduğunu varsayar. Ancak bu dağılım değişirse, bu yöntemlerin iyi çalışması için modeli yeniden oluşturması veya uyarlaması gerekir. Veri kümesi kaydırma problemi (data set shift problem) (Candela, Sugiyama, Schwaighofer, ve Lawrence, 2009) transfer öğrenimi ile yakından ilişkilidir. Diğer hibrit problemlere örnek olarak, Yarı Gözetimli Öğrenme (*Semi-supervised Learning*), Altgrup Keşfi (*Subgroup Discovery*) ve Ordinal Sınıflandırma/Regresyon (*Ordinal Classification/Regression*) verilebilmektedir (Fernández ve diğerleri, 2018).

5. Veri Biliminde Önyargı ve Etik

Veri biliminin hedefi veri setlerinden uygun genellemeleri ortaya çıkarıcı modeller oluşturarak anlamlı bilgiye sahip olmaktır. Genellemeler ortaya çıkarken de yanlı tahminlerde bulunabilen modeller de üretilebilmektedir ki bu da bizi veri biliminde önyargı konusuna getirmektedir. Bilgi sahibi olabilmek için kullandığı algoritmanın bir veri kümesinden üreteceği genelleme (veya model) için iki ana faktör katkıda bulunur. Birincisi, algoritmanın çalıştırıldığı veri kümesidir. Veri kümesi popülasyonu temsil etmiyorsa, algoritmanın oluşturduğu model doğru olmaz. Örneğin, daha önce bir bireyin BMI'sına dayanarak Tip 2 diyabet geliştirme

olasılığını öngören bir regresyon modeli geliştirilmiş olsun. Bu model A ülkesindeki erkeklerin bir veri setinden üretilmiş bir model olsun. Bunun sonucu olarak da model, kadın ya da farklı ülkedeki (etnik kökenli) bireylerde diyabeti öngörme amacında başarısız olacaktır. İkinci faktör ise makine öğrenmesi algoritmasının seçimidir. Çok fazla algoritma bulunmakta ve her biri aynı veri setini farklı şekilde genellemektedir (Kelleher ve Tierney, 2018).

2014 yılında Amazon, iş başvurularında başvuru sahiplerini özgeçmişlerindeki metne dayanarak tanımlama sürecini otomatikleştirmenin bir yolu olarak bir sistem geliştirmeye başladı. Ancak algoritmanın mühendislik rollerinde erkekleri kadınlara tercih ettiği ortaya çıktı. Amazon, algoritmanın adil olmadığını keşfettikten sonra sistemi terk etti. Diğer bir örnek Microsoft'un Tay Twitter botu. Tay, Twitter'da insanlarla etkileşime girerek öğrenen bir sohbet yapay zeka (chatbot) uygulamsıydı. Algoritma, konuşma için bir model oluşturmak üzere herkese açık verilerden beslendi aynı zamanda Twitter'da zaman içinde etkileşimlerden de öğrenmeye devam etti. Ne yazık ki, Tay'in yaşadığı etkileşimlerin hepsi olumlu değildi ve Tay modern toplumun önyargılarını öğrendi. Önyargı konusunda halen algoritmalar üzerine çalışılmaktadır. Özetle dört ana ön yargıdan bahsetmek mümkündür. Eğitim verisinde kullanılan verinin geneli yansıtmaması (*sample bias*), eğitim verisinde kullanılan verinin kültürel ya da diğer özelliklerin etkisinde kalması (*prejudice bias*), gözlemlemek veya ölçmek için kullanılan cihazla ilgili bir sorun oluştuğunda meydana gelen sistematik değer bozulması (*measurement bias*) ve son olarak da veriyle değil algoritmanın matematiği ile alakalı olan algoritma önyargısı (*algorithm bias*) en sık karşılaşılan önyargılardır (Ford, 2018).

Kişiler hakkındaki veri kullanımının etik etkileri vardır. İnsana uygulandığında veri madenciliği sıklıkla ayrımcıdır (kim kredi alacak, kim özel teklif alacak gibi). Bazı ayrımcılık çeşitleri ise ırksal, cinsel, dinsel olabilmektedir ki burada sadece etik değil illegal bir sorun var demektir. Aslında durum biraz karışıktır. Her şey uygulamaya bağlıdır. Tıbbi teşhis için cinsel ve ırksal bilgiyi kullanmak kesinlikle etiktir ancak aynı bilgiyi kredi ödeme davranışı için kullanmak etik değildir. Hassas bilgiler elense dahi ırksal ya da cinsel özellikler yerine geçebilecek değişkenlere dayalı modellerin oluşma riski vardır (ırksal bilgiler veri setinden çıkarılmış olsa da) (Witten ve diğerleri, 2011).

Yaygın şekilde kabul gören kişilerin, kişisel bilgilerinin nasıl kullanılacağını bilmesi gerektiğidir. Hangi veri hangi amaçla toplanmakta? Sorusuna yanıt kişilerin haklarını ihlal etmemeli onları manipüle edici yönde kullanılmamalıdır. Buna örnek 2014 ABD seçimlerinde, Cambridge Analytica'nın Facebook verilerini manipülasyon aracı olarak kullanarak seçime etki etmesi verilebilir (Budak, 2018).

6. Tıbbi Veri Bilimi

Tıp alanında veri bilimi çalışmalarını incelediğimizde öngörücü modellerin son yıllarda kullanılmaya başlandığı görülmektedir. Doktorlar teşhis koyarken ya da tedaviye karar verirken deneyimlerini ve içgüdülerini kullanmaktadır. Kanıta dayalı tıp ya da hassas tıp (precision medicine) hareketi tıbbi kararların verilere dayanması gerektiğini, ideal olarak mevcut en iyi verileri bireysel bir hastanın öngörüsüne ve tercihlerine bağlayacağını savunmaktadır. Örneğin, hassas tıp söz konusu olduğunda, hızlı genom dizileme teknolojisi, hastalığa neden olan mutasyonları tanımlamak ve buna özgü uygun tedavileri tasarlamak ve seçmek için nadir hastalıklara sahip hastaların genomlarını analiz etmenin mümkün olduğu anlamına gelir (bireysel). Veri bilimini tıpta önemli hale getiren bir diğer faktör de sağlık hizmetlerinin maliyetidir. Veri bilimi (özellikle öngörüye dayalı modeller) bazı sağlık bakım süreçlerini otomatikleştirmek için kullanılabilir. Örneğin, antibiyotiklerin ve diğer ilaçların bebeklere ve yetişkinlere ne zaman uygulanması gerektiğine karar vermek için öngörü modelleri kullanılmakta ve bu yaklaşım ile birçok hayatın kurtarıldığı bilinmektedir (Kelleher ve Tierney, 2018).

Hasta tarafından giyilen, yutulan ya da hastaya implante edilebilen tıbbi sensörler, hastanın hayati belirtilerini, davranışlarını ve de organlarının gün boyunca nasıl çalıştığını sürekli olarak izlemek için geliştirilmektedir. Bu veriler sürekli olarak toplanmakta ve merkezi bir izleme sunucusuna geri gönderilmektedir. Burada izleme sunucusunda sağlık uzmanları tüm hastalar tarafından üretilen verilere erişmekte, durumlarını değerlendirmekte ve tedavinin ne gibi etkileri olduğunu anlayarak ve her bir hastanın sonuçlarını benzer durumları olan diğer hastaların sonuçlarıyla karşılaştırabilmektedir. Hastanın özellikleri ve vücudunun çeşitli ilaçlara nasıl tepki verdiğine bağlı olarak kişiselleştirilmiş tedavi programları geliştirilmektedir. Tıbbi veri biliminin geleceği ise, ilaçlar ve bunların etkileşimleri, daha verimli ve ayrıntılı izleme sistemlerinin tasarımı ve klinik çalışmalardan daha fazla bilgi edinilmesi konusunda sürekli yeni araştırmaların yapılması olarak gözükmektedir (Kelleher ve Tierney, 2018).

7. Sonuç

Günümüzde inanılmaz büyüklükte veri üretebiliyoruz. Ürettiğimiz veriyi, Veri Bilimi şemsiyesi altında toplayabileceğimiz özelleştirilmiş yöntemler aracılığıyla analiz edebiliyor, bilgiye dönüştürebiliyoruz. Bu bilgi içgörü (insight) şeklindedir ve süreç hakkında enformasyon sağlar. Kurumlar (şirketler, hastaneler, okullar,...) daha çok veri güdümlü hale gelmiştir. Veriden elde ettikleri içgörüleri de karar süreçlerine uygulamaktadırlar. Böylelikle Patil (2012)'in tanımladığı veri ürünü (data product) ortaya çıkar. Veri ürününe örnek Google flu trends verilebilir. Google, arama motoru sorgularını analiz ederek Centers for Disease Control

and Prevention (CDC)'den daha hızlı olarak influenzanın yayılım izini sürebilmiştir. Bölümde de üzerinde durulduğu üzere veri bilimi bir süreçtir ve amaç sürecin başında net bir şekilde ifade edilmelidir. Süreç sonunda da bir veri ürünü çıkması beklenmelidir. Unutulmaması gereken son bir cümle ise iyi kalitede veri olmaksızın bir veri bilimi projesinin başarılı olmasının düşünülemeyeceğidir.

Kaynakça / References

Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer Inc., Cham.

Bach, F. (2018). *Foundations of machine learning*, 2nd edition, The MIT Press Cambridge, Massachusetts London, England.

Baesens, B., Vlasselaer, V. V. ve Verbeke, W. (2015). F*raud analytics using descriptive, predictive, and social network techniques a guide to data science for fraud deteciton*, John Wiley & Sons, Inc.

Beynon-Davies, P. (2007). Informatics and the inca, *International Journal of Information Management*, 27 (2007) 306-318.

Bramer, M. (2016). *Principles of data mining*, 3rd edition, Springer.

Budak, B. (2018). Bilmeniz gerekenler: Cambridge Analytica hikayesi, Facebook ve büyük veri, Erişim tarihi 8 Ocak 2020, https://webrazzi.com/2018/03/22/cambridge-analytica-hikayesi-facebook-ve-buyuk-veri/

Candela, J. Q., Sugiyama, M., Schwaighofer, A. ve Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press, Cambridge.

Chawla, N. V. (2005). Data mining for imbalanced datasets: an overview. In: Maimon, O. Z., Rokach, L. (eds.) *Data Mining and Knowledge Discovery Handbook*, pp. 853–867. Springer, New York.

Corea, F. (2019). *An introduction to data everything you need to know about ai, big data and data science*, Springer.

Dinov, I. D. (2018). *Data science and predictive analytics: Biomedical and health applications using r*, Springer.

Economist. (2017). *The world's most valuable resource is no longer oil, but data*, https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data, May 6th 2017.

Fernández, A., García, S., Galar, M., Prati, R .C., Krawczyk, B. ve Herrera, F. (2018). L*earning from imbalanced data sets*, Springer Nature Switzerland.

Ford, G. (2018). 4 human-caused biases we need to fix for machine learning, Erişim tarihi 8 Ocak 2020, https://thenextweb.com/contributors/2018/10/27/4-human-caused-biases-machine-learning/

Grey, CGP. (2012). *Is pluto a planet?,* Erişim tarihi 8 Ocak 2020, https://www.youtube.com/watch?v=Z_2gbGXzFbs.

Han, J. (2011). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers Inc., San Francisco.

Igual, L. ve Seguí, S. (2017). *Introduction to data science*, Springer.

Johansson, L. G. (2016). *Philosophy of science for scientists*. Springer.

Junqué de Fortuny, E., Martens, D. ve Provost, F. (2013). Predictive modeling with big data: Is bigger really better?. *Big Data*, *1*(4), 215-226.

Kelleher, J. D. ve Tierney, B. (2018). Data Science, The MIT Press Cambridge, Massachusetts London.

Nisbet, R. (2018). H*andbook of statıstıcal analysis and data mining applications*, 2nd edition, Elsevier Inc.

Pan, S. J. ve Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng*. *22*(10), 1345–1359.

Pathak, M. A. (2014). *Beginning Data Science with R,* Springer.

Patil, D. J. (2012). *Data jujitsu: The art of turning data into product*. O'Reilly Radar.

SINTEF. (2013). Big Data, for better or worse: 90% of world's data generated over last two years. ScienceDaily. Erişim tarihi 8 Ocak 2020, www.sciencedaily.com/releases/2013/05/130522085217.htm

Skiena, S. S. (2017). *The data science design manual,* Springer.

Stanton, J. (2013). *Version 2: An introduction to data science*, Syracuse University's School of Information Studies, http://jsresearch.net/wiki/projects/teachdatascience

Steele, B., Chandler, J. ve Reddy, S. (2017). *Algorithms for data science*, 1st edition, Springer Berlin/Heidelberg.

Witten, I. H., Frank, E. ve Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*, 3rd edition, Elsevier Inc.