DOI: 10.13114/MJH.2019.443

Geliş Tarihi: 12.12.2018
Kabul Tarihi: 24.05.2019

Mediterranean Journal of Humanities
mjh.akdeniz.edu.tr
IX/1 (2019) 13-26

Büyük Veri: Sosyal Bilimler ile Eğitim Bilimlerinde Kullanımı ve Uygulama Alanları

Big Data: Its Use and application area in Social Sciences and Educational Sciences

Gökhan AKSU [*]
C. Oktay GÜZELLER[**]

Öz: Bu çalışmanın amacı özellikle son on yılda üzerinde en fazla konuşulan konulardan biri olan büyük verinin ne olduğu ve sosyal bilimlerde büyük verinin nasıl ele alındığının belirlenmesidir. Bu amaç kapsamında öncelikle büyük veriyi tanımlarken veri hacmi (volume), veri çeşitliliği (variety) ve veri akış hızı (velocity) kavramlarının baş harfleri alınarak yapılan 3V tanımı ve daha sonra verinin gerçekliği (veracity) ile verinin değeri (value) kavramlarının eklenmesiyle ortaya çıkan 5V ifadesinin ne olduğu açıklanmıştır. Bununla birlikte çalışmada büyük veri kaynakları, büyük veri türleri ve klasik anlamda ele alınan veri ile büyük veri arasındaki farklılıkların neler olduğu açıklanmıştır. Bunun yanında büyük veri analizinde kullanılan program ve yazılımların neler olduğu, farklı ortamlardaki büyük verilerin nasıl ele alınacağına ilişkin kavramsal olarak neler yapıldığı açıklanmaya çalışılmıştır. Çalışmada son olarak sosyal bilimlerde ve eğitim bilimlerinde büyük verinin nasıl tanımlandığı, hangi tür verilerin büyük veri olarak kabul edilebileceği, büyük verilerin analizinde güncel yazılımların ve programların neler olduğu açıklanmıştır. Çalışma özellikle sosyal bilimlerde ve eğitim bilimlerinde çalışma yapacak araştırmacılara büyük verinin ne olduğu, nasıl ele alınması gerektiği ve büyük veri uygulamalarında kullanılan yazılımlara ilişkin önemli bilgiler sunulmaktadır.

***Anahtar sözcükler*:** Büyük Veri, Veri Madenciliği, Veri Kaynakları, Veri Türleri, Metin Madenciliği

Abstract: The aim of this study is to determine what big data is, one of the most talked about topics in the last decade, and how big data is handled in the social sciences. Within the scope of this aim, the definition of large data, the volume of data (volume), data variety and variety of data velocity concepts of the 3V definition and then the verbs of the data (veracity) with the addition of value (value) 5V is explained. However, in this study, big data sources, large data types and the data discussed in the classical sense, and the differences between big data are explained. In addition, the attempt is made to explain what are the programs and the software that is used in big data analysis, and what is done conceptually about how big data in different environments will be handled. Finally, it is explained how big data is defined in the social sciences and educational sciences, what kind of data can be accepted as big data, the current software and programs in analyzing big data. The study presents important information about what is big data, how it should be handled, and the software used in large data applications, especially for researchers who will work in the social sciences and educational sciences.

Keywords: Big data, data mining, data sources, data types, text data mining.

[*] Dr., Adnan Menderes Üniversitesi, Aydın Meslek Yüksekokulu, Aydın. gokhanaksu1983@hotmail.com, https://orcid.org/0000-0003-2563-6112
[**] Prof. Dr., Akdeniz Üniversitesi, Turizm Fakültesi, Antalya. cemguzeller@gmail.com, https://orcid.org/0000-0002-2700-3565

Giriş

Günümüzde birçok kuruluş çok büyük miktardaki veri toplayıp, depolar ve bu verileri analiz etmektedir. Büyük veri (big data) ifadesi ilk olarak, Roger ve Magoulas (2005) tarafından geleneksel veri yönetimi tekniklerinin karmaşıklığı ve büyüklüğü nedeniyle yönetemediği ve işleyemediği büyük miktarda veriyi tanımlamak amacıyla kullanılmıştır (Hadi *et al.* 2015).

Bazıları tarafından büyük veri elde edilmesi kolay olmayan ve ilk olarak bakıldığında karmaşık olarak görülen verilerdir. Bu nedenle onları elde etmek, depolamak, yönetmek ve işlemek oldukça zordur (Laney 2001). Başka bir grup tarafından büyük veri hacim, hız ve çeşitlilik anlamlarına gelen 3V olarak tanımlanmaktadır (Vozabal 2016). Bu tanıma göre veriler genel olarak verinin hacmi (Volume), veri akış hızı (Velocity) ve elde ediliş şeklinin çeşitliliği (Variety) nedeniyle "*büyük veri*" olarak adlandırılır (Watson 2014). 3V olarak tanımlanan başka bir yaklaşımda ise yüksek hacim (High volume) verinin miktar olarak çokluğu, yüksek hız (High velocity) verinin oluşturulma veya elde edilme oranı, yüksek çeşitlilik (High variety) farklı türdeki verileri ifade etmektedir (Russom 2011). Başka bir tanımda büyük veri, yüksek hacimli, yüksek hızlı ve/veya çok çeşitli olan verileri tanımlamak için kullanılan bir terimdir ve onu elde etmek, depolamak ve analiz etmek için yeni teknolojiler ve teknikler gerektirir (Mills *et al.* 2012; Sicular 2013).

Büyük verinin ne olduğu ve nasıl ortaya çıktığını anlamak için son 30 yılda verilerin nasıl depolandığını ve bu verileri yönetmek için kullanılan veri tabanlarının nasıl bir gelişim gösterdiği Fig. 1'de gösterilmiştir.

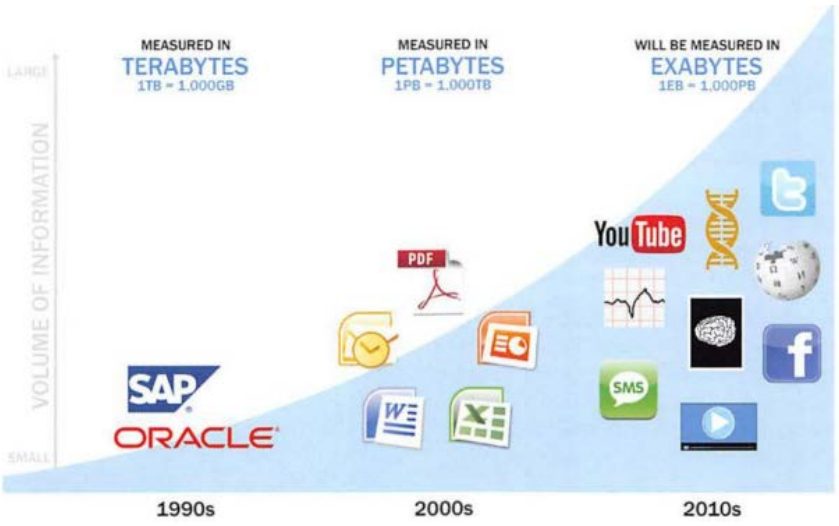

Fig. 1. Veri Evrimi ve Büyük Veri Kaynaklarının Yükselişi

Figür 1'de görüldüğü üzere 1990'lı yıllarda en büyük veri boyutu terabayt iken, 2000'li yıllarda petabayt ve 2010'lı yıllarda exabayt boyutuna ulaştığı belirtilmektedir. Günümüzde exabaytın 1000 katı olan zetabayt ve zetabaytın 1000 katı olan yotabayt olmak üzere depolanan veri miktarının sürekli artış gösterdiği bilinmektedir. Örnek olması bakımından 1 yottabayt veriyi terabayt büyüklüğündeki sabit disklerde saklamak için, her biri bir şehir büyüklüğünde olan bir milyon veri merkezine ihtiyaç duyacaktır ve bu da alan olarak en az bir şehir büyüklüğünde bir alanı kapsayacaktır (Pence 2015). Figür 2'de veri hızı, veri çeşitliliği ve veri hacmi arasındaki ilişkiler görsel olarak açıklanmaya çalışılmıştır.

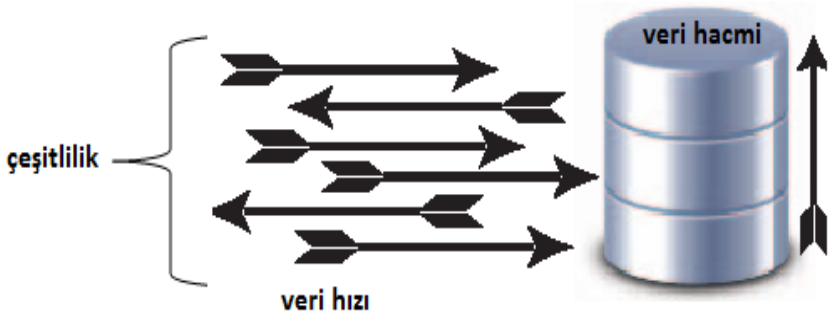

Fig. 2. Veri Hızı, Çeşitliliği ve Hacmi Arasındaki İlişkiler

Gulatieri (2014) bireyin sıralı DNA'sındaki verilerin yaklaşık olarak 750 MB olduğunu ve Amerika Birleşik Devletleri'nin toplam nüfusu için 222 Petabayt büyüklüğünde bir depolama alanı gerekeceğini belirtmektedir. Örnek olması bakımından yeni ortaya çıkan bir hastalığa teşhis koymak amacıyla sadece bir kişinin genomu analiz edilse bile bu 222 petabaytlık veri setinde örüntülerin ortaya çıkarılması anlamına gelmektedir. Bu sebeple daha önce IBM tarafından yapılan 3V tanımına Karmaşıklığı (Complexity) eklemek gerektiği belirtmektedir (Pence 2015). Ancak popüler olan başka bir bakış açısıyla son dönemde ham verinin önemli özelliklerini koruma anlamında Gerçeklik (Veracity) ve toplanan verinin amaçlanan sürece veya tahmine ilişkin getireceği Katma Değer (Value) olmak üzere 5V tanımı daha çok kullanılmaktadır (Hadi *et al.* 2015). Büyük verinin daha iyi anlaşılması bakımından 5V tanımında yer alan her bir kavram aşağıda ayrıntılı olarak açıklanmıştır.

Veri hacmi (Volume): İstenilen sonuçları elde etmek için manipüle edilen ve analiz edilen veri miktarını ifade eder. Figür 1'de gösterildiği üzere verinin hacmi sürekli artmakta ve bu nedenle üst sınırı yoktur. Bu büyük hacimli veri setlerini yönetmek için birçok teknoloji ortaya çıksa da şuan exabaytın ötesine geçilememiştir (Patgiri & Ahmet 2016).

Veri hızı (Velocity): Her geçen gün sahip olunan verinin boyutu katlanarak artmakta ve bu hız daha büyük bir veri tabanı oluşmasına sebep olmaktadır. Sıkıştırma teknolojisi kullanılsa bile verinin hacmi sürekli artış göstermektedir (Tole 2013). Bu nedenle büyük veride hız daima verinin hacmine göre tanımlanır. Büyük verideki hız, esas olarak büyüme hızı ve transfer hızı olmak üzere iki kavramla ilişkilidir. Bu iki hız gereksinimi birbirinden farklıdır. Sadece 2009 yılında Facebook'un 1 Petabayt ve Google'un 15 Exabaytlık depolama alanına sahip olduğu ve 2005 yılında bir milyar olan internet kullanıcısının 2015 yılında üç milyar olduğu düşünüldüğünde büyüme hızındaki artış açıkça görülmektedir.

Veri çeşitliliği (Variety): Saklanan, analiz edilen ve kullanılan verinin türü anlamına gelir. Yapılandırılmış, yarı-yapılandırılmış ve yapılandırılmamış olmak üzere üç farklı şekilde ele alınmaktadır. Yapılandırılmış veri, sekmeli bir veri biçimini, yarı yapılandırılmış veri günlük verileri veya XML verilerini ve yapılandırılmamış veri videoları, görüntüleri, bilimsel bilgileri içerir (Laney 2001). Depolanan ve analiz edilen verilerin türü değişkenlik gösterebileceği gibi genel olarak konum koordinatlarından, video dosyalarından, tarayıcılardan gönderilen verilerden ve simülasyonlardan oluşabilir.

Katma Değer (Value): Depolanan verilerin kalitesi ve daha fazla kullanılması ile alakalıdır. Kendi içinde aslında değersiz gibi görünen muazzam büyüklükteki verilerin içinden faydalı bilgiler çıkarmaktadır (Khan *et al.* 2014). Her gün internetin temel protokollerini içeren TCP/IP kayıtlarına cep telefonları çağrı kayıtlarından büyük miktarda veri depolanmaktadır. Önemli olan bu kayıtlardan hangilerinin değerli olup olmadığıdır. Eğer düzgün bir şekilde yönetilemiyorsa ve sonuçlardan iyileşme için öngörü sunamıyorsa bu kadar büyük veriyi depolamanın

bir anlamı olmayacaktır (Tole 2013).

Verinin Gerçekliği (Veracity): Eldeki verinin doğruluğu, gerçekliği ve anlamlılığı olarak ifade edilmektedir. Büyük veriler üzerinde bir işlem yapmak istediğimizde, büyük hacimli veriler sorunlu hale gelebilmektedir. Asıl soru elde edilen sonuçların doğru olduğuna nasıl inanmamız gerektiği ile ilgilidir (Patgiri & Ahmet 2016). Verinin bütünlüğü, veri doğruluğu, veri kaynağının güvenilir olması ve verinin güvenilir olması veri gerçekliği ile ilgili hususlardan bazılarıdır. Veri gerçekliği için önerilen ölçütlerden biri bilgi kazanç miktarıdır (entropi) ve nesnellik, doğruluk ve güvenirlik gibi ölçütleri bünyesinde barındırmaktadır (Paryani 2012). Bu aşamada verinin gerçekliği veri madenciliğinde incelenmesi gereken önemli bir özelliktir ve eğer veriler doğru bir kaynaktan elde edilmemişse veya güvenilmez ise büyük sorunlar ortaya çıkabilmektedir. Örnek olarak bir pazarlama şirketi veri gerçekliği konusunda şüpheli veriler ile gerçekleştirdiği analiz sonucunda gerçekte var olmayan ilişkileri tespit ederek hatalı kararlar alabilecektir (Trifu & Ivan 2014).

Büyük Veri ve Küçük Veri Arasındaki Farklar

Büyük verinin ne anlama geldiğini daha iyi anlamak amacıyla geleneksel yöntemlerde kullanılan verinin farklı alt başlıklardaki anlamı ve büyük veri ile arasındaki farkların neler olduğu Tablo 1'de gösterilmiştir.

Tablo 1. Veri ve Büyük Veri Arasındaki Kavramsal Farklılıklar

	Veri	Büyük veri
Amaç	Genellikle belirli bir soruyu cevaplamak veya belirli bir hedefe ulaşmak için tasarlanmıştır.	Projenin çıktısının tam olarak ne olduğu bilinmez. Genellikle bir amaç için gerçekleştirilir ancak amaç esnektir.
Depolama	Genel olarak küçük veriler bir kurumda, genellikle bir bilgisayarda veya bir dosyada bulunur.	Büyük veriler şirket ağı genelinde veya İnternet'ten elde edilir. Genellikle, her yerde elde edilebilen bu veriler birden çok sunucuda tutulmaktadır.
Veri yapısı	Genellikle yapılandırılmış verilerdir. Veriler yaygın olarak tek bir disiplin veya onun alt alanları ile ilgilidir. Veriler tek tip kayıtlar veya elektronik tablolar şeklindedir.	Metin belgeleri, görüntüler, sesler ve fiziksel nesneler gibi yapılandırılmamış veya yarı yapılandırılmış verileridir. Veride farklı disiplinlere ilişkin bilgiler yer almaktadır.
Veri hazırlama	Çoğu durumda, veriler kendi kullanıcıları tarafından kendi amaçları için hazırlanmaktadır.	Veriler birçok farklı kaynaktan elde edilmektedir. Verilerde yer alan kişiler genellikle bu verileri kullanan kişiler değildir.
Veri ömrü	Veriler genellikle belirli bir süre (akademik yıl) için kayıt altına alınmaktadır. Verilerin kullanılması veya analizinden kısa bir süre sonra bu veriler atılmaktadır.	Büyük veriler genellikle kalıcı olarak saklanması gereken verileri içerir. Projenin maçına veya sonucuna bakılmaksızın sürekli veri tabanlarında saklı tutulmaktadır.
Ölçme	Genel olarak veriler bir deneysel işlem ile ölçülmektedir.	Birçok farklı veri türü birçok farklı elektronik formatta ölçülebilmektedir.
Tekrar edilebilirlik	Projeler tipik olarak tekrarlanabilir. Eğer veri kalitesinde bir sorun varsa, tüm proje tekrar edilebilir.	Projenin tekrarlanması pek mümkün değildir. Kalitesi kötü olan verinin bulunması ve onun yerine daha iyisini koyma düşüncesi iyimserlikten başka bir şey değildir.
Maliyet	Projenin maliyeti sınırlıdır ve enstitü veya bireyler bu maliyetten kendi imkanlarıyla kurtulabilmektedirler.	Büyük verilerin yer aldığı projeler gerçekten pahalı olabilmektedir. Büyük veriler ile yapılmış bir projenin başarısızlıkla sonuçlanması

		gerçekten iflasa neden olabilir.
Veri görünüşü	Veriler genellikle elektronik bir tabloda veya veri tabanında satır ve sütunlarda tanımlanmaktadır.	Verilere ulaşmak küçük veriye kıyasla daha zordur. Verinin yapısı veya içeriği anlaşılmayabilir. Verilere ulaşmak içebakış (introspection) adı verilen özel bir teknikle mümkün olmaktadır.
Analiz	Çoğu durumda projedeki tüm veriler birlikte analize dahil edilmektedir.	Büyük veriler adım adım ilerleyen teknikler yardımıyla analiz edilmektedir. Bu aşamada farklı yöntemlerle yeniden analiz yapılabilir, veriler gözden geçirilir, azaltılır, normalleştirilir, dönüştürülür, görselleştirilir ve yorumlanır.

Büyük Veri Kaynakları

Büyük veri birçok farklı kaynaktan elde edilebilmektedir. Genel olarak makinelerden, insan etkileşimlerinden ve bilgi işleme merkezlerinden elde edildiği belirtilse de sağlıkla ilgili kayıt altına alınan veriler, sosyal medyada paylaşılan resim ve videolar, cep telefonlarının konum ve görüşme kayıtları, internette yapılan aramalara ilişkin kayıtlar büyük veri örnekleridir (Vozabal 2016). Bir web sitesinde gezinirken her fare tıklaması sonucunda, müşterilerin satın alma davranışlarını daha iyi anlamak ve ürünleri müşterinin ihtiyaçlarına göre etkin bir şekilde sunmak amacıyla Web günlük dosyaları depolanabilir ve bu veriler analiz edilebilir. Başka bir örnekte Facebook ve Twitter gibi sosyal medya kaynakları muazzam miktarda yorum veya paylaşım üretmektedir. Bu veriler insanların yeni ürün tanıtımları hakkında ne düşündüğünü anlamak için depolanıp sonrasında analiz edilebilir. Bir başka örnek akıllı sayaçlar gibi makinelerin üretmiş olduğu verilerdir. Bu sayaçlar, müşterilerle paylaşılabilecek elektrik, su veya gaz tüketimi hakkında sürekli veri akışı sağlar ve müşterilerine çamaşır yıkamak gibi enerji tüketiminin bir kısmını yoğun olmayan saatlere taşımaya motiveye yönelik fiyatlandırma planları ile birleştirilir. Bir diğer örnek ise GPS yardımıyla konum bilgilerine ilişkin depolanan verilerdir. Bu sayede arkadaşlarınızın konumlarını bilmeniz ve yakındaki mağazalar veya restoranlar hakkında bilgi sahibi olmanız sağlanmaktadır. Son örnek ise güvenlik sistemlerinde yüz tanıma veya taşıt tanıma gibi özellikler sayesinde büyük miktarda veriler depolanarak gerekli durumlar için görüntü, ses ve video verileri analiz edilebilir (Watson 2014).

Büyük Veri Türleri

Büyük veri genellikle, multimedya dosyaları (videolar, görüntüler ve sesler), metin dosyaları, coğrafi ve finansal veriler gibi yapılandırılmış veya yapılandırılmamış şekilde bulunmaktadır. Geleneksel verileri saklama ve analiz etme çabasına bakılmaksızın, büyük verilerin çoğu yapılandırılmış veya yarı yapılandırılmış olup, bu verileri depolamak, işlemek ve analiz etmek için farklı teknikler ve araçlar gereklidir. Bazı araştırmacılar verileri yapılandırılmış, yarı yapılandırılmış, kısmı yapılandırılmış ve yapılandırılmamış olmak üzere dört başlık altında ele alınmaktadır ve bunlardan yapılandırılmamış veya yarı yapılandırılmış olanları büyük veri olarak isimlendirilmektedir (Lafrate 2015).

Yapılandırılmamış veriler kolaylıkla sınıflandırılamayan verilerdir. Resimler, videolar, metin belgeleri (doc, pdf, txt, vb.) yapılandırılmamış veriye iyi birer örnektir. Bu verilerin kendilerine ait herhangi bir içsel yapısı bulunmamaktadır. Kısmi yapılandırılmış veriler ardışık olaylara ilişkin işlem geçmişlerine benzer. İnternet tarayıcınızdan önce arama motoru için bir web sitesine, sonra aradığınız içerikle ilgili başka bir siteye oradan da bir başka sayfaya yönlendirildiğinizde arka planda kaydı tutulan Tek biçimli kaynak konumlayıcı (Uniform Resource Locator-URL) adresleri kısmi yapılandırılmış verilere birer örnektir. Yarı yapılan-

dırılmış veriler katı bir modele sahip olmayıp esnek bir yapıya sahip olan kayıtları gösterir. Örnek olarak genişletilebilir işaretleme dili (Extensible Markup Language-XML) genel olarak ağaç yapısında olup hem insanlar hem de bilgi işlem sistemleri tarafından kolayca okunabilecek dokümanlar oluşturmaya yarayan bir araçtır. Yapılandırılmış veriler ise kendi formatını tanımlayan bir yapıya sahiptir. Genel olarak yapılandırılmış veriler nasıl depolanacaklarını (veri tipleri: Sayı, metin, vb.) nasıl işleneceklerini, nasıl erişilebileceğini ve nasıl kısıtlanacağını (olası değerler kümesi: Erkek, kadın) tanımlayan gruplara ayrılmaktadır. Yapısı nedeniyle, bu tür veriler kolayca depolanabilir, işlenebilir ve sorgulanabilir. Farklı amaçlarla kullanılan e-devlet, e-reçete, e-okul gibi veri tabanları yapılandırılmış verilere iyi birer örnektir (Vozabal 2016).

Yapılandırılmamış yapıda farklı çok sayıda bilgi depolanmaktadır. Yapılan bir çalışmada dünyadaki verilerin yaklaşık %80'inin yapılandırılmamış metin belgeleri olduğu ifade edilmektedir (Ramanathan & Meyyappan 2013). Bu yapılandırılmamış metinler bilgisayarlar tarafından kolayca kullanılamaz ve bu sebeple yapılandırılmamış metinlerden değerli bilgileri çıkarmak için bazı tekniklere ihtiyaç duyulmaktadır. Bu sayede elde edilen bilgiler daha sonra yapılandırılmış ve kısmi yapılandırılmış alanlar içeren metin veri tabanları biçiminde saklanabilmektedir. Bu aşamada oluşturulmuş olan büyük veri tabanlarından ilginç ve bir o kadar da değerli bilgileri elde etmek için kullanılan yöntem ve tekniklerin yer aldığı çalışma alanı metin madenciliği olarak tanımlanmaktadır (Pande & Khandelwal 2014).

Metin Madenciliği (Text Mining)

Büyük veri analizi alan yazında veri analitiği olarak isimlendirilmektedir. Bu kavram farklı modelleri ve diğer faydalı bilgileri keşfetmek için büyük veri setlerini toplama, organize etme ve analiz etme sürecini ifade eder. Normalden farklı, daha karmaşık ve büyük bir büyük ölçekli veri setlerinden değerli bilgileri keşfetmek için yeni uygulama yöntemlerini içeren bir dizi teknoloji ve tekniğin genel adıdır. Bu süreçte temel olarak yeni problemler veya çözülmemiş eski problemler daha iyi ve etkili yollarla çözülmeye odaklanılmaktadır (Chen & Zhang 2014). Araştırmacıların veri madenciliği yöntemleriyle yazılı belgeler arasındaki ilişkileri ve örüntüleri keşfetmek amacıyla gerçekleştirdikleri analizlere metin madenciliği denilmektedir (Dang & Ahmad 2014). Başka bir tanımda metin madenciliği yapısal olmayan ve düzensiz haldeki elektronik metin yığınlarından; önceden bilinmeyen, potansiyel olarak kullanışlı, yapısal ve düzenli veri elde etme sürecidir (Vidhya & Aghila 2010). Metin Madenciliği, veri madenciliğinin bir parçası olarak düşünülmesine rağmen, alışılagelen veri madenciliği yöntemlerinden biraz farklıdır. İki veri madenciliği yöntemi arasındaki temel farklılık, metin madenciliğinde örüntülerin olay tabanlı veri tabanlarından daha çok, doğal dil metinlerinden çıkarılmasıdır.

Metin analizinde kullanılacak verilerin tamamı veya büyük çoğunluğu veri tabanlarında yazılı olarak bulunmaktadır. Ancak bu metinlerden önemli bilgilerin etkili bir şekilde elde edilmesi geleneksel yöntemlerle pek mümkün değildir. Bunun için büyük metin verilerini analiz etmek amacıyla bazı otomatik araçlar tasarlanmaktadır. Bu aşamada kurallar temelinde varlıklar arasında ilişkilendirme, tahmin kuralları, örüntüleri keşfetme, gibi birçok farklı yöntem kullanır. Bu süreçte yapılandırılmamış veya yarı yapılandırılmış metin verilerinden tüm bilgiler alınır ve önemli bilgiler açığa çıkarıldıktan sonra kategorilere ayrıştırılmaktadır. Son aşamada metinlerden elde edilen ve kategorilere ayrılan önemli bilgiler karar almak için gerekli yerlere rapor edilmektedir. En temel istatistiklerden frekans ve yüzde değerleri kullanılarak metin içinde en çok tekrar eden kelimeler belirlenerek metin madenciliği yöntemleri uygulanabilmektedir (Zanini & Dhawan 2015). Metin madenciliği, doğal dil metninden anlamlı bilgiler çıkarmaya çalışan ve yeni gelişen bir çalışma alanıdır (Kalra 2013). Metin madenciliği veri madenciliğine benzemektedir, ancak veri madenciliği yöntemleri veri tabanlarından yapılandırılmış verileri analiz etmek için tasarlanmıştır, ancak metin madenciliği bunların dışında e-posta, metin bel-

geleri ve HTML dosyaları gibi yapılandırılmamış veya yarı yapılandırılmış veri kümeleriyle de çalışabilmektedir (Elmasri & Navathe 2011). Sosyal bilimler ve eğitim alanından bir örnek olması bakımından metin madenciliğinde kullanılan güncel programlardan biri olan CiteSpace yazılımı ile Web of Science veri tabanında yer alan tüm dergiler içerik analizine dahil edilerek sorgulama yapıldığı taktirde yıllar itibariyle hangi konuların trend olduğu ve hangi alanda çalışmaların yoğunlaştığı kolaylıkla görülebilecektir (Chen 2016).

Metin madenciliği beş adımda gerçekleşmektedir. Birinci aşamada yapılandırılmamış verilerden bilgi toplanır. İkinci aşamada toplanan bilgiler yapılandırılmış veriye dönüştürülür. Üçüncü aşamada yapılandırılmış veriler içindeki örüntüler ortaya çıkarılır. Dördüncü aşamada ortaya çıkarılan bu örüntüler analiz edilir. Beşinci ve son aşamada analiz edilen örüntülerden değerli bilgiler çıkarılır ve bu bilgiler veri tabanında kayıt altına alınır. Metin madenciliğinin beş adımlık süreci Figür 3'te gösterilmiştir (Behera & Kumar 2015).

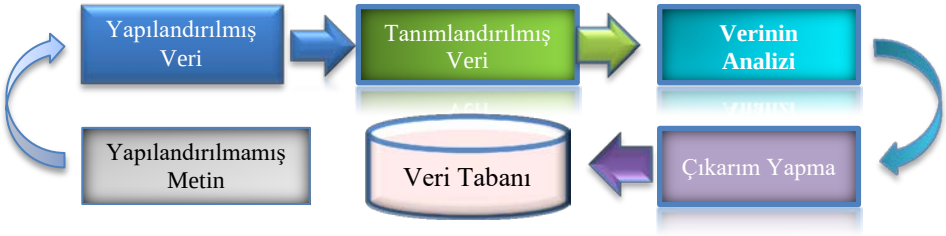

Fig. 3. Metin Madenciliği Süreci

Zohar (2002) metin madenciliği işlemlerini ve her bir süreçte yapılması gerekenleri altı adımda tanımlamıştır. Buna göre farklı ortamlardan elde edilen metinler ilk olarak ön işleme tabi tutulmaktadır. Bu aşamada sözcükler anlamsal ve söz dizimsel olarak ayrıştırılmakta ve bir sonraki metin dönüşümü aşamasına geçilmektedir. Özellik üretimi olarak tanımlanan bu aşamada ortak sözcüklerden kök bulma ve kelime sıklık analizinde önemsiz görülen bazı edat vb. kelimelerin sorgu sürecinin dışında bırakılması (stopword) sağlanır. Özellik seçimi aşamasında frekans ve yüzde gibi basit betimsel istatistiklerden yararlanılır. Veri madenciliği yöntemlerinin uygulandığı bir sonraki aşamada sınıflandırma ve kümeleme yöntemleri işe koşulmaktadır. Son aşama olan yorumlama ve değerlendirmede analiz sonuçları paylaşılmaktadır. Metin Madenciliği işlemleri ve içerdikleri yaklaşımlar Figür 4'te gösterilmiştir.

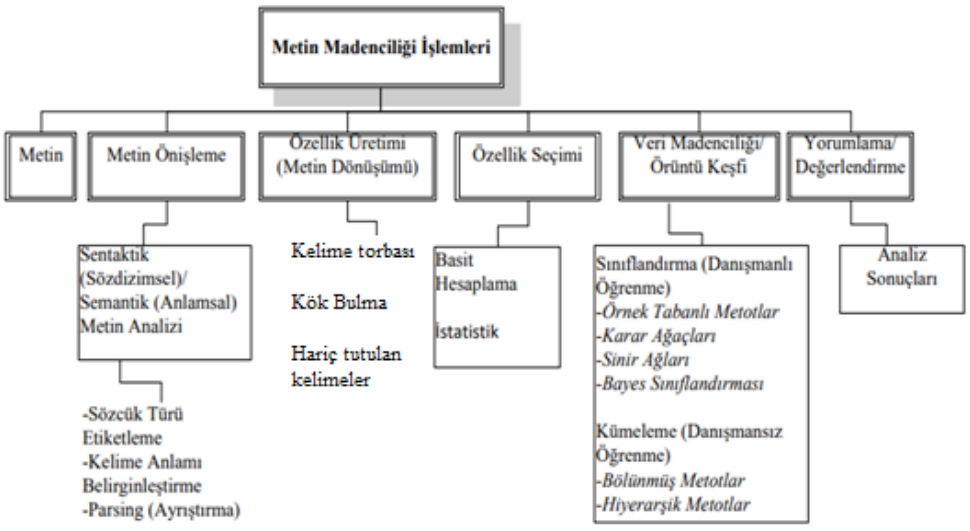

Fig. 4. Metin Madenciliği İşlemleri Akış Şeması

Metin madenciliğinin uygulama alanlarından biri olan sosyal medya analizinde veriler bloklar, uygulamalar veya sosyal medya programlarından toplanmakta ve karar alma amacıyla kullanılmaktadır. Günümüzde sosyal medya reklamcılık ve ürün pazarlama yöntemleriyle gerçek zamanlı müşteri seçimlerini, niyetlerini ve duygularını anlamak için en iyi platformlardan biridir. İnternet üzerinden alışveriş yapılan sitelerden ebay.com arama, tüketici önerileri ve yorumlar için 7.5 ve 40 petabayt büyüklüğünde iki tane depolama alanı kullanmaktadır (Verma *et al.* 2016). Sosyal medyanın analizinin uygulama alanları: Davranış analizi, lokasyon bazlı etkileşim analizi, tavsiye sistemi geliştirme, ilişki-bağlantı tahmini, müşteri etkileşimi ve Pazarlama, medya kullanımı, güvenlik ve sosyal çalışmalar başlıklarında toplanmaktadır (Stieglitz *et al.* 2018).

İçerik tabanlı analiz sosyal medyanın arka planında depolanan verilerin analizi anlamına gelmektedir. Örneğin Facebook kullanıcıları kendilerine ait fotoğraf, video ve kendileriyle ilgili bilgileri Facebook'un depolama alanına kaydetmektedir. İçerik tabanlı analizde kullanıcıların aradığı kelimeler ile benzer kelimeler ve kullanıcının profilindeki içerikle eşleşen öğeler ile öneriler sunulmaktadır. Örneğin bir kullanıcı özellikleri daha önceden bilinen bir ürün satın aldığında, kendisine satın aldığı ürünle eşleşen özelliklere sahip diğer ürünler önerilecektir. Kullanıcının satın aldığı ürün ile önerilen ürün ne kadar çok eşleşiyorsa bu analizin hassaslığı olarak adlandırılmaktadır (Doreswamy 2012).

Diğer Büyük Veri Analizleri

Sayısal ve metin olarak büyük miktardaki verilerin haricinde ses kayıtları ve video görüntüleri de analiz edilmektedir. Ses analizinde veriler ses dosyası formatında sıkıştırılarak ses sinyallerinden anlam ve bilgilerin çıkarılması amaçlanmaktadır. Ses analizinde, verilen ses dosyalarının gönderenin gönderdiği formatta veya uygun formatta olup olmadığını kontrol etmek için kullanılır. Örneğin gözetleme yaklaşımında (surveillance application) toplumda işlenen suçların tespiti için ses kayıtlarının sistematik olarak analiz edilmektedir. Ses dosyalarının analizinde şüpheli veya tehdit içeren kelimelerin kullanılması durumunda uygulama acil koduyla uyarı vermekte ve ilgili ses dosyası ya da o dosyanın kaynağı gözetim altına alınmaktadır (Verma *et al.* 2016). Bir başka uygulamada ise cam kırılması, çığlık, silah sesi ve yardım sesi kritik seslerin yakalanarak ani müdahalelerde bulunulabilmektedir (Mishra & Sharma 2016). Sosyal bilimlerde ve eğitim alanında sıklıkla kullanılan R paket programı da diğer birçok teknolojide olduğu gibi, ses analizi için TuneR ve audio gibi paketleri sayesinde ses nesnelerinin kaydedilmesini, alınmasını, değiştirilmesini ve dışa aktarılmasını mümkün kılmaktadır.

Video analizinde kapalı devre kameralar ile kayıt altına alınan görüntüler daha sonra kullanılmak üzere saklanır ancak videolar çok fazla bilgi içerir ve diğer dosya türlerine göre daha fazla yer kaplamaktadırlar. Videoların çoğunda gerekli olmadığı müddetçe önemli bilgi olmaması sebebiyle büyük verinin hacim boyutunda yer almaktadır (Verma *et al.* 2016). Bir kaza anının analiz edilmesinde, iş yerlerindeki hırsızlık olayının aydınlatılmasında kapalı devre kameralardan elde edilen görüntüler kullanılmaktadır. Açık sistemli kamera kayıtlarının analizi ile trafiğin yoğun olduğu noktalar belirlenerek sürücüler farklı güzergâhlara yönlendirebilir. Başka bir örnekte ise yüz tanıma tekniği ile aranan bir kişinin kolaylıkla yakalanması ve yetkililere teslim edilmesi video analizi ile gerçekleşmektedir (BSİA 2016).

Sosyal Bilimlerde ve Eğitim Bilimlerinde Büyük Veri

İlk olarak 1990'lı yılların başında kullanılmaya başlanan veri madenciliği yöntemlerinde veriler üzerinde belirli bir model oluşturmak amacıyla bilgi keşfi algoritmalarının uygulanması yoluyla analizler gerçekleştirilmiştir. Yıllar boyunca, kümeleme, birliktelik, sınıflandırma algoritmaları, regresyon modelleri, öngörücü yöntemler ve faktör analizi veri madenciliği araştırmalarına egemen olan temel yaklaşımlar olarak kabul edilmiştir (Daniel 2015). Büyük veri geleneksel

anlamda ele alınan verilerden çok daha büyük veriler olarak tanımlanmaktadır. Eğitimde ve sosyal bilimlerde çok büyük ifadesi sırasıyla öğrenci gözlem sayısı, gözlem sıklığı ve gözlem türü sayısı anlamına gelmektedir (Laney 2001; Ray 2013). Eğitim alanında yapılan çalışmalar için büyük verinin faydaları farklı veri kümeleri birleştirildiğinde ortaya çıkmaktadır (Perry & Klopfer 2014). Örneğin okullar tarafından kayıt altına alınan demografik, davranışsal, duyuşsal ve akademik veriler yıllık veya yıllar itibariyle çok fazla miktarda veri içermektedir. Büyük veri olarak kabul edilen bu veriler özellikle ulusal ve uluslararası öğrenci karşılaştırmalarında sıklıkla kullanılmaktadır (Steinkuehler 2017). OECD tarafından düzenlenen ve uluslararası geniş ölçekli sınavlardan biri olan PISA (Programme for International Student Assessment) sınavı ile öğrenci ve okula ilişkin oldukça fazla verinin içerisinden anlamlı sonuçlar çıkarmak amaçlandığından TIMSS ve PISA gibi sınavlardan elde edilen veriler büyük veri olarak tanımlanmaktadır (Vaitsis *et al.* 2016). Yurt dışında Eğitim ve Test Servisi (ETS) tarafından yapılan lisansüstü kayıt sınavları ve Akademik Yeterlik Testi (SAT) uzun yıllar boyunca çok fazla öğrenciden elde edilen bilgilere sahip olması nedeniyle büyük veri olarak kabul edilmektedir (Pardos 2017).

Baker ve Yacef (2009) eğitimde kullanılan veri madenciliği türlerini dört ana başlık altında toplamaktadır. Bunlar:

- Tahminleme
- Örüntüleri/yapıları keşfetme
- İlişkileri keşfetme
- Bireyleri yönetmek için verilerin özünü anlama

Tahminleme aşamasında kullanılabilecek yöntemlerin sınıflama, regresyon ve örtük/gizil bilgi kestirimi; örüntü keşfi aşamasında kümeleme, faktör analizi, alan yapısı keşfi ve ağ analizi; ilişkileri keşfetme aşamasında ise birliktelik kuralları, ilişkisel veri madenciliği, sıralı/ardışık örüntü madenciliği ve nedensel veri madenciliği en çok tercih edilen yöntemlerdir.

Büyük verinin faydalı olarak kullanımına ilişkin bir diğer örnek ise öğrenme analitiği olarak tanımlanan çalışma alanıdır. Öğrenme analitiği öğrenenlerle ve öğrenme süreçleriyle ilgili verilerinin nasıl analiz edileceği ve öğrenme sistemlerinin kanıta dayalı nasıl geliştirilmesi ile ilgilenen bir alan olarak ortaya çıkmıştır (Shum 2012). Birçok ticari kurum pazar paylarını ve gelirlerini artırmak için iş analitiğini geliştirmiştir. Örneğin Amazon şirketi müşterilerinin alışveriş eğilimlerini kullanarak güçlü bir öneri motoru geliştirmiştir. Facebook ve Google gibi birçok küresel şirket kullanıcıların Web üzerindeki etkinliklerini analiz ederek ticari sürekliliklerini sağlayabilmek için kullanıcı verilerini toplayan ve bu bağlamda öneriler sunan algoritmalar geliştirmişlerdir (del Blanco *et al.* 2013). Bu açıdan bakıldığında yüksekokullardaki ve üniversitelerdeki büyük veriler, öğretim ve araştırma için çok yüksek bir değere sahiptir. Buradan elde edilen veriler yardımıyla eğitimin kalitesi belirlenebileceği gibi süreç içinde öğrenme güçlüğü yaşayan öğrenciler daha kolay belirlenerek başarı üzerinde etkili duyuşsal ve davranışsal özellikler ortaya çıkarılabilmektedir (Meng & Meng 2014). Bununla birlikte üniversiteler ve okullar da mevcut durumları hakkında bilgi sahibi olacak ve gelecekteki durumları hakkında tahminde bulunabileceklerdir (Bienkowski *et al.* 2012). Yüksek eğitimde büyük verinin kimler için ve hangi amaçla kullanılabileceğine ilişkin yapılan tanımda öğretmenler, öğrenciler ve karar alıcılar olmak üzere üç farklı grup için yararlı bilgiler elde edilmektedir. Figür 5'te veri madenciliğinin paydaşları ve bunların hangi amaçla büyük veriyi kullandıkları gösterilmektedir.

Fig 5. Yükseköğretimde Büyük Veri Kullanıcıları ve Fırsatları

Yükseköğretimde büyük veri araçlarını kullanma yollarından biri öğrencilerin bireysel performans ve beceri seviyelerini analiz etmek ve kendi özel öğrenme ihtiyaçlarını karşılayan kişiselleştirilmiş bir öğrenme deneyimi oluşturmaktır. Büyük veriler etkili olarak kullanıldığında, kurumların öğrenme deneyimini geliştirmesine ve öğrenciyi de geliştirmesine yardımcı olabilirken, okulu bırakma oranlarını düşürür ve mezuniyet sayılarını artırabilmektedir (Daniel 2015). Bunların yanında öğrencileri sistemdeki mevcut programlar ile karşılaştırarak hem öğrencilere hem de ailelere en iyi okulu ve programı bulmaları için yardımcı olmaktadır. Eğitim alanında büyük veriler sayesinde daha gelişmiş öğretim programlarıyla öğrencilerin performansları ve öğrenme yetenekleri geliştirilerek dersler daha bireysel hale gelmektedir (Sharma *et al.* 2017).

Bunlara ek olarak Siemens ve Gasevic (2012) büyük verinin çevrimiçi öğrenmelerde kolayca uygulanabilir olduğunu belirtmektedir. Son yıllarda eğitim ortamlarında uzaktan eğitim uygulamalarıyla öğrenme ve öğretmeye katkıda bulunabilecek çok sayıda veri elde edilmektedir. Bu yeni veriler aynı zamanda sosyal ağların da yardımıyla, farklı geçmişlere sahip öğrenciler arasında ilişki kurmaya ve dersle ilgili temel kavramları anlamalarına yardımcı olmaktadır. Çevrimiçi öğrenme ortamlarından elde edilen veriler sayesinde öğrencilerin öğrenme becerileri geliştirebilmekte ve onların geleneksel eğitimden daha verimli sonuçlara ulaşmaları sağlanabilmektedir (West 2012).

Ülkemizde MEB tarafından kurulması planlanan ölçme değerlendirme merkezleri, Uluslararası Öğrenci Değerlendirme Programı (PISA) ile Uluslararası Fen ve Matematik Eğilimleri Araştırması'na (TIMSS) rakip olacak "*Eğitimde Öğrenci Gelişimini İzleme Sistemi*" ile PISA'ya göre daha geniş bir katılımın sağlandığı "*Akademik Becerilerin İzlenmesi ve Değerlendirilmesi*" (ABİDE) çalışmasını illerde yürütecektir. Bu kapsamda örgün eğitim kapsamında öğrencilerin akademik gelişimlerinin izlenebilmesi, öğrenme eksikliklerinin tespit edilmesi için öğrenci, veli ve eğitimcilere geri bildirim verilmesine olanak sağlanacaktır.

Veri Madenciliğinde Kullanılan En Popüler Yazılımlar

Sin ve Muthu (2015) büyük verilerin analizinde kullanılabilecek açık kaynak kodlu programların MongoDB, Hadoop, MapReduce, Orange ve Weka olduğunu belirtmektedir. Bunun yanın-

da Rapid Miner, Knime, Sisense, SSDT, Apache, Oracle, IBM SPSS, R Programing ve Python gibi yazılımlar da veri madenciliği alanında kullanılan popüler araçlar olarak kabul edilmektedir (DataFlair, 2018). Veri madenciliğinde kullanılan yazılımların ne tür verileri analiz edebildiği ve arkada planında nasıl bir algoritma olduğu aşağıda açıklanmıştır.

MongoDB: Bir çapraz platform belge odaklı veri tabanı yönetim sistemidir. Mongo'da tablo tabanlı mimari yerine JSON benzeri nesneler kullanır.

Hadoop: Basit programlama modelleri kullanarak büyük veri kümelerinin ağa bağlı bilgisayar kümelerinde dağıtılmasını sağlayan bir sistemdir.

MapReduce: Farklı ortamlardaki büyük miktardaki veriyi aynı anda analize tabi tutmaktadır.

Orange: Büyük verilerin işlenmesi ve madenciliği için Python tabanlı bir araçtır. Eklentileri çok çeşitli olmakla birlikte sürükle ve bırak işlevleriyle kullanımı kolay bir ara yüze sahiptir. Veri görselleştirmesine en iyi şekilde yardımcı olur ve bileşen tabanlı bir yazılımdır.

Weka: Büyük miktarda veriyi işlemek için JAVA tabanlı bir araçtır. Örüntü keşfi, sınıflama ve tahminleme analizlerinde kullanılabilecek çok çeşitli algoritmalara sahiptir.

Rapid Miner: Kendisiyle aynı adı taşıyan şirket tarafından geliştirilen en iyi tahmine dayalı analiz sistemlerinden biridir. JAVA programlama dilinde yazılmıştır. Derin öğrenme, metin madenciliği, makine öğrenmesi ve tahmine dayalı analizler yapılabilmektedir.

Knime: Kendisiyle aynı adı taşıyan şirket tarafından geliştirilen veri analitiği ve raporlama için en iyi platformlardan biridir. Yazılımın içinde birlikte gömülü çeşitli makine öğrenmesi ve veri madenciliği yöntemleri bulunmaktadır. Yazılım genel olarak ilaç araştırmalarında yaygın olarak kullanılmaktadır. Ayrıca müşteri veri analizi, finansal veri analizi ve iş zekası için iyi bir performans sergilemektedir.

Sisense: Kendisiyle aynı adı taşıyan şirket tarafından geliştirilen lisanslı bir yazılımdır. Özellikle bir kurum ya da kuruluş içinde raporlama söz konusu olduğunda son derece yararlı ve en uygun yazılımdır. Hem küçük ölçekli hem de büyük ölçekli verileri kolaylıkla analiz edebilmektedir. Ortak bir depo oluşturmak için çeşitli kaynaklardan gelen verileri birleştirmeyi sağlar ve ayrıca raporlama için farklı birimler arasında paylaşılabilen zengin raporlar oluşturmaktadır.

SSDT (SQL Server Data Tools): Visual Studio'da veritabanı geliştirmenin tüm aşamalarını genişleten evrensel ve geri bildirimci bir modeldir. Microsoft tarafından veri analizi yapmak ve iş zekası çözümleri sunmak için geliştirilen yazılımda kullanıcılar doğrudan bir veri tabanıyla çalışabilir kurum içinden veya kurum dışında analizleri gerçekleştirebilirler. Kullanıcılar Visual Basic, C#, gibi programlama dilleri aracılığıyla yeni tablolar oluşturabilirken aynı zamanda mevcut tabloları düzenleyebilmektedirler.

Apache: Apache Foundation tarafından geliştirilen ve makine öğrenmesi algoritmaları kullanan programda genel olarak veri kümeleme, sınıflandırma ve işbirlikçi filtrelemeye odaklanılmaktadır. Apache Java tabanında yazılmıştır ve lineer cebir ve istatistik gibi matematiksel işlemleri gerçekleştirmek için JAVA kütüphanelerini kullanmaktadır.

Oracle: Aynı isme sahip şirketin geliştirdiği veri madenciliği yazılımı, veri sınıflandırma, tahminleme, regresyon ve analistlerin öngörüleri analiz etmesini, daha iyi tahminler yapmasını, en iyi müşterileri hedeflemesini, çapraz satış fırsatlarını tespit etmesini ve dolandırıcılığı tespit etmesini sağlayan algoritmaları içermektedir. Veri tabanındaki verilerin doğrudan "*sürükle ve bırak*" özelliği ile kullanıcılara daha iyi bir kullanım kolaylığı sağlamaktadır.

IBM SPSS Modeler: IBM'in sahip olduğu veri madenciliği ve metin analitikleri için tahmine dayalı modeller oluşturmak için kullanılan bir yazılım paketidir. İlk olarak SPSS tarafından üretilmiş ancak daha sonra IBM tarafından satın alınmıştır. SPSS Modeler, kullanıcıların prog-

ramlama gerekmeden veri madenciliği algoritmaları ile çalışmasına izin veren görsel bir arayüze sahiptir. Veri dönüşümleri sırasında karşılaşılan gereksiz karmaşıklıkları ortadan kaldırır ve tahmine dayalı modellerin kullanımını kolaylaştırır.

R Programing: Öncelikle C ve Fortran'da tabanlarında yazılmış olan R, istatistiksel hesaplama ve grafikleri gerçekleştirmek için ücretsiz bir yazılım ortamıdır. Kendi kütüphanesi içinde birçok paketi olan yazılım akademik ortamlarda araştırma, mühendislik ve endüstriyel uygulamalarda yaygın olarak kullanılır. Kullanım kolaylığı ve genişletilebilir olması R yazılımının popülerliğini son yıllarda büyük ölçüde artırmıştır. Yazılım veri madenciliğinin yanı sıra, doğrusal ve doğrusal olmayan modelleme, klasik istatistiksel testler, zaman serisi analizi, sınıflandırma, kümeleme ve diğerleri dahil olmak üzere istatistiksel ve grafiksel teknikler sunar.

Python: Özgür ve açık kaynaklı bir dil olarak oluşturulan yazılım kullanım kolaylığı açısından en sık R ile karşılaştırılmaktadır. Kullanıcılar kendileri veri setleri oluşturabilirken en karmaşık analizleri dakikalar içinde gerçekleştirebilmektedirler. Mevcut durumu görselleştirmede şirketler tarafından yaygın olarak kullanılan yazılımlardan biridir.

Sonuç ve Tartışma

İnovasyon çağını yaşadığımız bu günlerde farklı kaynaklardan elde edilen büyük miktarda verinin içinde yer alan değerli bilginin keşfedilmesi ve bu bilginin ileride alınacak kritik kararlar için kullanılması büyük önem taşımaktadır. Farklı kaynaklardan ve farklı ortamlardan elde edilen büyük veriler eğitim, turizm, pazar araştırması vb. birçok alanda önemli bir çalışma konusu haline gelmiştir. Bu aşamada elde edilen verilerin büyük veri olarak tanımlanabilmesi için farklı araştırmacılar tarafından yapılan ve geniş bir kesim tarafından kabul edilen 3V ve 5V tanımları üzerinden büyük verinin ne olduğu açıklanmıştır. Bunun yanında büyük veri kaynakları, büyük veri türleri ve klasik anlamda ele alınan veri ile büyük veri arasındaki farklılıkların neler olduğu açıklanarak araştırmacıların bu konuda bakış açısı kazanmaları amaçlanmıştır. Büyük verinin sağladığı birçok avantajın yanında bazı dezavantajları da beraberinde getirdiği araştırmacılar tarafından rapor edilmektedir.

Büyük verinin elde edilmesi ve farklı sebeplerle kullanılması bir dizi etik, sosyal ve politik karmaşıklıkla birlikte günümüzde modernleşmenin getirdiği bir dizi tehlikeyle birlikte güvensizlik sorununu da gündeme getirmektedir (Mills 2018). Sürekli artan veriler ve teknolojik gelişmelerle birlikte elde edilen verinin güvenliği ve gizliliği önemli bir konu haline gelmektedir. Farklı ortamlarda farklı kurumlar tarafından elde edilen büyük veriler kişisel mahremiyet ve güvenlikle ilgili sorunları da beraberinde getirmektedir. Hali hazırda farklı veri tabanlarında depolanan bu verilerin yanlış ve kötü niyetli kişilerin ellerine geçmesi durumunda hiç de istenmeyen sonuçlar ortaya çıkabilir. Bu nedenle elinde büyük veri bulunan kurum ve kuruluşların bu konuda çok dikkatli olması gerektikleri düşünülmektedir.

Çalışma kapsamında büyük veri ve uygulama alanlarına ilişkin tanımlamalar bir bütün olarak incelendiğinde özellikle MEB, ÖSYM, YÖK gibi resmi kurumların sahip oldukları büyük verileri etkin bir şekilde kullanarak üniversitelerin ve diğer eğitim kurumlarının hem kalitesini hem de başarısını artırmak için elde ettikleri önemli sonuçları paylaşmaları önerilmektedir. Bu noktada verinin gizliliği konusunda ortaya çıkabilecek problemleri engellemek amacıyla ilgili kurumda çalışan personelin büyük veri analizi ve veri madenciliği konularında eğitim alarak eldeki hazır veriyi kurumun dışına çıkarmadan analiz ederek elde ettikleri sonuçları kamuoyu ile paylaşabilecekleri düşünülmektedir. Özellikle veri madenciliğinde popüler olan programların ilgili personele öğretilerek ileride alınacak kararlar için istatistiksel delillere sahip olacağı düşünülmektedir.

Kişisel verilerin gerek resmi kurumlar gerekse ticari kuruluşlar tarafından kayıt altına alın-

ması, depolanması ve kullanımı günümüzde oldukça kolaylaşmıştır (Bainbridge 1997). Bilgisayar ve internet kullanımının yaygınlaşmaya başlamasıyla kişisel verilerin toplanması, depolanması, paylaşılmasını ve analiz edilmesini oldukça kolay bir hale gelirken bununla birlikte birtakım güvenlik tedbirlerinin alınmasını da zorunla hale gelmiştir (Kutlu & Kahraman 2017). Ülkemizde kişisel verileri işleyen gerçek ve tüzel kişilerin yükümlülükleri ile uyacakları usul ve esasların da düzenlenmesi 2016 yılında yürürlüğe giren 6698 sayılı "*Kişisel Verilerin Korunması Kanunu*" ile teminat altına alınmıştır. Büyük veriye sahip kurum ve kuruluşların kişisel verilerin korunması ve saklanması konusunda oldukça dikkatli olmaları gerektiği düşünülmektedir. Aksi taktirde toplumun büyük bir kısmının zor durumda kalacağı ve mağduriyetlerin yaşanacağı olaylar ile karşı karşıya kalınabilir.

KAYNAKÇA

Bainbridge D. I. (1997). "Processing Personal Data and the Data Protection Directive". *Journal of Information & Communications Technology Law* 6/1 (1997) 17-40.

Baker R. S. J. D. & Yacef K. (2009). "The State of Educational Data Mining in 2009: A Review and Future Visions". *Journal of Educational Data Mining* 1 (2009) 3-16.

Behera S. & Kumar N. V. (2015). "Filtering of Unstructured Text". *International Journal of Engineering Research and Development* 11/12 (2015) 45-49.

Bienkowski M., Feng M. & Means B. (2012). *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief* (2012) 1-64. Washington.

BSİA (2016). *An Introduction to Video Content Analysis – Industry Guide*. Issue 2. Form no: 262. (2016) 1-12.

Chen C. (2016). *Citespace: A Practical Guide for Mapping Scientific Literature*. New York 2016.

Chen C. L. & Zhang C. Y. (2014). "Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data". *Information Sciences* 275 (2014) 314-347.

Dang S. & Ahmad P. H. (2014). "Text Mining: Tecniques and Application". *IJETI International Journal of Engineering & Technology Innovations* 1/4 (2014) 22-25.

Daniel B. (2015). "Big Data and Analytics in Higher Education: Opportunities and Challenges". *British Journal of Educational Technology* 46/5 (2015) 904-920.

del Blanco Á., Serrano Á., Freire M., Martínez-Ortiz I. & Fernández-Manjón B. (2013). "E-Learning Standards and Learning Analytics. Can Data Collection Be Improved by Using Standard Data Models?". *Global Engineering Education Conference (EDUCON)* 2013 IEEE (2013) 1255-1261.

Doreswamy H. K. (2012). "Performance Evaluation of Predictive Classifiers for Knowledge Discovery from Engineering Materials Data Sets". *CIIT International Journal of Artificial Intelligent Systems and Machine Learning* 3/3 (2012) 162-168.

Elmasri R. & Navathe S. B. (2011). *Fundamentals of Data Base Systems*. Boston 2011.

Hadi J. H., Shnain A. H., Hadishaheed S. & Ahmad A. (2015). "Big Data and Five V's Characteristics". *International Journal of Advances in Electronics and Computer Science* 2/1 (2015) 16-23.

Kalra P. (2013). "Text Mining: Concepts, Process and Applications". *Journal of Global Research in Computer Science* 4/3 (2013) 36-39.

Khan M. A., Uddin M. F. & Gupta N. (2014). "Seven Vs of Big Data: Understanding Big Data to Extract Value". *2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1)* (2014) 3-5.

Kutlu Ö. & Kahraman S. (2017). "Türkiye'de Kişisel Verilerin Korunması Politikasının Analizi *Siyaset, Ekonomi ve Yönetim Araştırmaları Dergisi* 5/4 (2017) 45-62.

Lafrate F. (2015). *From Big Data to Small Data*. USA 2015.

Laney D. (2001). "3D Data Management: Controlling Data Volume, Velocity, and Variety". *Meta Group* 949 (2001) 1-4.

Meng L. & Meng L. Q. (2014). "Application of Big Data in Higher Education". *2nd International Conference on Teaching and Computational Science (ICTCS 2014)* (2014) 215-216.

Mills K. A. (2018). "What are the Threats and Potentials of Big Data for Qualitative Research?". *Qualitative Research* 18/6 (2018) 591-603.

Mills S., Lucas S., Irakliotis L., Ruppa M., Carlson T. & Perlowitz B. (2012). *Demystifying Big Data: A Practical Guide to Transforming the Business of Government.* Washington 2012. Available at http://breakinggov.com/documents/demystifying-big-data-a-practical-guide-to- transforming-the-bus/

Mishra C. & Sharma A. M. (2016). "A Review Paper on Voice Analytics". *International Journal of Science Technology and Management* 5/12 (2016) 247-257.

Pande V. C. & Khandelwal A. S. (2014). "A Survey of Different Text Mining Techniques". *IBMRD's Journal of Management and Research* 3/1 (2014) 125-133.

Pardos Z. A. (2017). "Big Data in Education and the Models That Love Them". *Current Opinion in Behavioral Sciences* 18 (2012) 107-113.

Paryani J. (2012). *A Case Study on Determining the Big Data Veracity: A Method to Compute the Relevance of Twitter Data*. Unpublished Master Thesis. Bachelor of Engineering in Computer Engineering, Pune University, Pune, India.

Patgiri R. & Ahmed A. (2016). "Big Data: The V's of the Game Changer Paradigm". *IEEE 18th International Conference on High Performance Computing and Communications*; *IEEE 14th International Conference on Smart City*; *IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS*) (2016) 17-24.

Pence H. E. (2015). "What is Big Data and Why is it Important?". *J. Educational Technology Systems* 43/2 (2015) 159-171.

Perry J. & Klopfer E. (2014). "UbiqBio: Adoptions and Outcomes of Mobile Biology Games in the Ecology of School". *Computers in the Schools* 31 (2014) 43-64.

Ramanathan V. & Meyyappan T. (2013). "Survey of Text Mining". *International Conference on Technology and Business and Management* (2013) 508-514.

Ray S. (2013). "Big Data in Education" *Gravity Issue* 20 (2013) 1-4.

Russom P. (2011). "*Big Data Analytics*". *TDWI Best Practices Report*. Seattle 2011. Source: http://tdwi.org/research/2011/09/best- practices-report-q4-big-data-analytics.aspx

Sharma M., Shazia A. & Husain S. (2017). "Implementation of Big Data Analytics in Education Industry". *Journal of Computer Engineering (IOSR-JCE)* 19/6 (2017) 36-39.

Shum B. (2012). *Learning Analytics*. UNESCO Policy Brief.

Sicular S. (2013). "*Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three 'V's*". Source: http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-threeparts-not-to-be-confused-with-three-vs/

Steinkuehler C. (2017). *Big Data in Education: Balancing the Benefits of Educational Research and Student Privacy*. National Academy of Education Report.

Stieglitz S., Mirbabaie M., Ross B. & Neuberger C. (2018). "Social Media Analytics Challenges in Topic Discovery, Data Collection, and Data Preparation". *International Journal of Information Management* 39 (2018) 156-168.

Tole A. A. (2013). "Big Data Challenges". *Database Systems Journal* 4/3 (2013) 31-40.

Trifu M. R. & Ivan M. L. (2014). « Big Data: Present and Future". *Database Systems Journal* 1 (2014) 32-41.

Vaitsis C., Hervatis V. & Zary N. (2016). « Introduction to Big Data in Education and Its Contribution to the Quality Improvement Processes". Ed. Ventura S. S., *Big Data Real-World Appl. Tech* (2016) 41-64.

Verma J. P., Agrawal S., Patel B. & Patel A. (2016). "Big Data Analytics: Challenges and Applications for Text, Audio, Video, And Social Media Data". *International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI)* 5/1 (2016) 41-51.

Vidhya K. A. & Aghila G. (2010). "Text Mining Process, Techniques and Tools: An Overview". *International Journal of Information Technology and Knowledge Management* 2/2 (2010) 613-622.

Vozabal M. (2016). *Tools and Methods for Big Data Analysis*. Unpublished Master Thesis. University of West Bohemia Faculty of Applied Sciences Department of Computer Science and Engineering, Univerzitní, Czech Republic 2016.

Watson H. J. (2014). "Tutorial: Big Data Analytics: Concepts, Technologies, and Applications". *Communications of the Association for Information Systems* 34 (2014) 65.

West D. M. (2012). Big Data for Education: Data Mining, Data Analytics, and Web Dashboards. 2012.

Zanini N. & Dhawan V. (2015). "Text Mining: An Introduction to Theory and Some Applications". *Research Matters* 19 (2015) 38-45.

Zohar E. (2002). *Introduction to Text Mining. Supercomputing.* Automated Learning Group National Center for Supercomputing Applications, University of Illinois.