VERİ, BÜYÜK VERİ VE İŞLETMECİLİK

Data, Big Data and Business Administration

Gönderim Tarihi: 09.05.2016
Kabul Tarihi: 25.05.2016

Suat ATAN*

Öz: Büyük veri verinin miktarının çokluğu ve çeşitliliği çağrışımını yapsa da özünde verinin değerinin yeniden keşfi sonrasında geleneksel veri analizi perspektifi yerine yeni araç ve yaklaşımlarla aslında verinin yeniden keşfedilmesini ifade eden trendin adıdır. Büyük veri işletmeler için çok ciddi bir kaynak teşkil edebilir hatta işletmelerin bizzat işi haline dönüşebilir. Bu çalışmada Büyük veri trendine kadar gerçekleşen son gelişmeler ve işletmeler ile ilişkisi ele alınacaktır.

Anahtar Kelimeler: Veri Madenciliği, Makine Öğrenmesi, Dev Veri.

Abstract: The term of big data connotates the abundance of data and properties of it. However, after re-discovering the value of data, the term of big data reflects the new trend which includes new approaches and tools rather than the sophisticated methods. On the other words, big data is the constant rediscovering of data. The big data can provide fruitful resources for the corporates. Even it can be the core task of a corporate. By the way, in this study, the recent advances in the field and case studies have been discussed. The perspectives of corporates to data have also been evaluated.

Keywords: Data Mining, Machine Learning, Big Data.

* Bilgi Sistemleri Uzmanı, Tarım ve Kırsal Kalkınmayı Destekleme Kurumu/Bilgi Sistemleri Koordinatörlüğü, Ankara Üniversitesi/İşletme Bölümü/Doktora Öğrencisi,
e-posta: suatatan@suatatan.com, web: http://blog.suatatan.com

GİRİŞ

Veri kavramı bilgisayarların iş ve özel yaşamın içerisinde girmesinden hemen sonraki dönemde formlara girilen basit verilerin toplanması ve bu verilerin günlük pratik amaçlarla kullanımı ile daha sık anılır hale gelmiştir. Bu dönemde verinin varoluş nedeni sadece ona ulaşmaktır. Burada "dönem" olarak kast edilen zaman dilimi özel bir tarihe işaret etmemektedir. Nitekim aynı şartlar altında farklı birey ve işletmeler için bu durum geçici veya kalıcı olarak devam edebilir. Verilerin kontrolü özellikle miktarının sürekli artması ile tüm veri sahipleri için gitgide güç duruma gelmiştir. Ancak veri kavramının geldiği bu yeni nokta sadece miktarla sınırlı değildir. Verilerin toplanması, kaynakları, işlenmesi ve değerlendirilmesi gibi konularda da ilgili çalışmalarda daha az değinilen ancak Büyük veri trendinin temellerini teşkil eden bu noktalar da önem arz etmektedir. Bu çalışmanın birinci hedefi büyük verinin bilgisayar bilimleri ile ilgili çalışmalar dışında az vurgulanan noktalarını ele almaktadır. İkinci hedef ise; büyük veri trendinin işletmeler için vadettiği faydaları değerlendirmektir. Bu faydalar işletmenin ürettiği verilerin çokluğundan ziyade ilgili verilere dair sahip olduğu bilinç düzeyi ile ilintili olmaktadır. Üçüncü hedef; Büyük veri trendine kadar veri bilimi ile ilgili olarak gerek pratik gerekse akademik amaçlarla ilk defa dâhil olacak araştırmacılar için kullanışlı temel kavram ve yaklaşımları genel hatları tanıtmaktır. Bu yaklaşımların literatürde farklı disiplinler altında ele alınıyor olması ve sürekli güncellenmesi nedeniyle takibi güç olmaktadır. Bu amaçla, özellikle veri madenciliği ve büyük verinin temellerini teşkil eden kavram ve yaklaşımlar işletmecilik perspektifinden ele alınacaktır.

VERİ-İŞLETME İLİŞKİLERİNİN DURUMU: VERİ BİLİNÇ DÜZEYİ

Verinin değeri ve potansiyelinin tam olarak ortaya koymak için, işletmelerin ürettikleri ve maruz kaldıkları verilerle arasındaki ilişkilerin düzeyini ele almak faydalı olabilir. Bu ilişkileri kategorize eden aşağıdaki sıralama, bu ilişkilerin gelişim düzeylerini göstermektedir. Her işletme kendi içerisinde farklı dönemlerde farklı düzeylerde varlığını idame ettiriyor olabilir. Örneğin Büyük verinin yoğun bir şekilde tartışıldığı günümüzde, genel olarak işletmelerin en azından işletme içerisinde elde edilen verilerden istifade ederek, istatistiksel izleme yapabildiği düşünülebilir. Ancak bazı işletmeler için düzey hala kayıt ve izleme düzeyindedir. İşletmelerin veri ile ilişkilerindeki bu düzeyler bu çalışmada "veri bilinç düzeyi(VBD)" olarak adlandırılmıştır. İşletmenin hangi VBD içerisinde olduğu o işletmenin büyüklüğü, kaynakları ve yetenekleri ile ilgilidir. Bu nedenle her bir düzeyi münferit olarak değerlendirmek doğru olacaktır. Ancak her üst düzey perspektif önceki düzeylerin aşılmış olduğunu da zımnen ifade eder.

VBD 1: Kayıt ve Basit İzleme Perspektifi

Verilerin herhangi bir teknoloji ile bilgisayarlarda uzun süreli olarak kolayca depolanabilmesinin mümkün olması ile işletmeler kimi zaman basılı belgelerle eş zamanlı olarak, kimi zaman basılı belgelerden bağımsız olarak tüm verilerini bilgisayarlara kaydetmeye başladılar. Bu VBD düzeyindeki bir işletmede kayıt işleminin temel amacı bizzat verinin kendisine erişimdir. Örneğin bir işletme stok giriş çıkış bilgisini temel muhasebe amacı dışında da mevcut fiziksel varlıklarını izlemek için kullanıyor olsun. Bu durumda verinin her bir girdisi, örneğin tek bir stok kaydı bile gerektiğinde pratik amaçlarla kontrol edilebilmektedir. Bu VBD'de tarih ve veri girdisi ilişkisi en kritik bilgiyi teşkil eder. Hangi işlemin, hangi gün ve ne miktarda yapıldığı bilgisi günlük fiziksel defter tutma alışkanlığının dijital hale getirilmiş halinden başka bir şey değildir. Bu düzeyde, bir bütün olarak verinin sadece toplam, periyodik alt toplam, genel eğilimi gibi bilgileri kullanılmaktadır. Aynı şekilde veri basit sıralama ve filtreleme ile değerlendirilebilmektedir. MS Excel, LibreOffice Calc gibi programlar temelde tam da bu amaca hizmet ederler. Bu düzeydeki işletmelerde verilerin detaylı istatistiklerinin izlenmesi ve değerlendirilmesi genellikle söz konusu değildir.

Aynı şekilde bu veriler genellikle elle girilen ve fiziksel olarak doğrulanabilir, somut veya kolay anlaşılabilir bilgilerden meydana gelmektedir. Başka bir ifade ile bu veriler sensörler, uydular ve benzeri cihazlardan otomatik olarak üretilen soyut ve karmaşık verilerden meydana gelmemektedir.

Bu düzeyde tanımlı bir işletmenin verilerden elde edebildiği yegâne fayda verilerin operasyonel amaçlarla kullanımı ve basit istatistiksel sonuçları elde etmektir.

VBD 2: İstatistiksel İzleme Perspektifi

Bu düzey, verilerin tekil kayıtlar düzeyinde değil, bir araya gelerek teşkil ettiği bütünün özelliklerinin değerlendirilmesinin gerçekleşebildiği perspektiftir. Veri miktarı yatay ve dikey olarak arttıkça, - ya da verilerin adedi ve beher veririnin özelliğine dair bilgiler arttıkça – ortaya çıkan bütünün işletme için faydalı birçok gösterge üretmesi olasıdır.

Stok örneğinden devam edilecek olursa, işletme stok kayıtlarına ilişkin bilgileri operasyonel amaçlar için kullanmanın ötesinde bir adım daha ileriye giderek söz gelimi stok devir hızı gibi bir değeri ortaya koymak için kullandığında bu istatistiksel izleme düzeyi içinde kabul edilmelidir. Bu düzeydeki bir işletme artık fiziksel stok bilgisinin sayısal hali olan operasyonel kayıtların ötesinde soyut bir değer olan stok devir hızı gibi bir değeri de hesaplamakta

ve izlemektedir. Stok devir hızı kavramı veya benzeri bir kavramın muhasebe terimleri içerisinde var olması bunun her işletme için kullanıldığı anlamına gelmez. İşletme, çok küçük olduğundan stok devir hızı ile ilgilenmesi pratik olmayabilir bu durumda ilgili veriyi de bu bağlamda dikkate almayacaktır.

İşletme için bu faydalı göstergeler ön tanımlı olabileceği gibi (muhasebe tablolarındaki oranlar; genel toplamlar, ortalamalar gibi) işletmenin kendi ihtiyaçları doğrultusunda tasarladığı özel göstergeler olabilir. Örneğin bir işletme, personelin mesai devamlılığının kayıtlarını tuttuğu veri tabanından, herhangi bir zamanda işe geç gelen kişilerin listesi gibi bir listeye bakarak bu kişileri uyarmak yerine yıl içinde ortalama devamlılığı izleyerek daha gerçekçi bir devamlılık takibi yapabilir. VBD 2 düzeyindeki işletmelerde, bu işletmelerin daha verimli kılınmasını sağlayacak başka birçok göstergeler oluşturulabilir. Bu düzeydeki işletme için verilerin tek başlarına değil bir araya gelerek oluşturdukları kavramsal göstergeler işletmenin iç görü elde etmesine yönelik bir değer oluşturmaya henüz başlamıştır.

VBD 3: Tahmin Perspektifi

VBD 2 düzeyindeki işletmeler mevcut verileri statik olarak ele alarak işletme hakkında genel değerlendirmeler gerçekleştirmektedirler. Ancak aynı veriler eğer zaman bağlamlı olarak da kayıt altına alınmaktaysa bu verilerden geleceğe dair tahmin ve öngörüler de elde edilebilir. Bu tahminleri elde edecek ve değerlendirecek işletmeler VBD 3 düzeyindedirler. Bu düzeydeki işletmeler verileri statik değil dinamik olarak ele almaktadırlar. Başka bir değimle verilerin zaman değerleri vardır ve isabetli tahminler için verilerin en güncel hallerine ihtiyaç duyulmaktadır. Aynı zamanda verilerin düzenli aralıklarla tutulması da önem arz eder.

Bu perspektife sahip işletmeler zaman serisi şeklinde gösterilebilen, yüksek frekanslı verilerdeki mevsimsel veya dönemsel dalgalanmalarla trend gibi bilgileri izleyip bunları kullanarak da gerekli tedbirleri alabilirler. Bu düzeye kadar işletmelerin tuttukları verilerin boyut ve içeriği çok farklı değildir. Ancak amaçlar farklılaşmaktadır. Doğal olarak işletmeler büyüdükçe tahmin ve öngörü perspektifine yaklaşmaları beklenecektir.

Tahmin perspektifi ifadesi genel olarak geleceğe atıf yapmakla birlikte, geçmişe dönük ya da anlık tahmin yapılabilmesi de olgusunu da kapsar.

VBD 4: Veri Madenciliği Perspektifi

Veri madenciliği perspektifi, verilerin işletme için basit bir kayıttan ya da bütünsel olarak bir değer oluşturmaktan ziyade verilerin kendi başına gerçek bir

değer olarak ele alındığı düzeyi yakalamasını ifade eder. Bu durum, verinin çokluğundan ziyade verilerin barındırdığı gizli ve değerli bilgilerin varlığının mümkün olduğundan haberdar olmayı ve bu bilgilerin kullanımının işletmeyi rakiplerine karşı daha güçlü kılacağını bilmeyi gerektirir.

Amaçlarının benzer olmasından ötürü istatistik alanı ile ona göre daha genç sayılabilecek veri madenciliği alanı genellikle birbirine karıştırılmaktadır. Aynı nedenlerle, veri madenciliği istatistik biliminin alt kolu gibi düşünülmektedir. Ancak bu düşünce gerçekçi değildir. İstatistik alanı daha biçimsel ve eski köklere sahiptir. Veri madenciliği alanı ise istatistikten farklı olarak başta bilgisayar bilimleri olmak üzere birçok alandan beslenir (Hand, 1999: 17). Bunun yanında, veri madenciliği terimindeki "madencilik" teriminden de anlaşılacağı üzere, veri madenciliği büyük miktardaki verilerin ele alınmasını ve buna ulaşana değin "maden" dışındaki anlamlı değeri olmayan verileri eleme sürecini çağrıştırır. Aynı şekilde, görselleştirme ve makine öğrenmesi gibi kritik araçlar da veri madenciliği ile ilişkilidir. Öte yandan, söz gelimi veri madenciliğinde kullanılan "karar ağacı", kNN algoritması gibi algoritmaların ise istatistikle ilgisi neredeyse yoktur. Bu bağlamda veri madenciliğini istatistiğin alt dalı olarak ele almak doğru bir yaklaşım olmayacaktır. Buna göre VBD 4 düzeyi de VBD 3 düzeyine göre verilerin kurumsal olarak ele alınması bağlamında daha yüksel bir yetenek ve birikimi ifade eder.

Veri madenciliği perspektifindeki bir işletmede verinin operasyonel değerinden daha fazla değere sahip olduğunun bilinci hâkimdir. Bu tür işletmelerde artık veri tali bir iş olmaktan çıkmış olup ana iş kollarından biri haline dönüşmüştür. Bu nedenle bu tür işletmelerde "veri bilimci" ve "veri analisti" gibi pozisyonlarda işe alımlar gerçekleştirilmektedir.

İşletmelerde veri madenciliği çerçevesinde kullanılmakta olan temel yöntemler genel hatları ile aşağıda incelenmektedir:

Korelasyon

Bir veri seti içerisinde herhangi iki parametrenin birbiri ile ilişkisini ifade eder. Eğer korelasyon değeri sıfır ise bu iki parametre birbirinden tamamen bağımsızdır. Korelasyon arttıkça bu iki parametre arasındaki bir bağımlılık ilişkisinden söz edilir. Bu değer maksimum 1 olabilir ki iki veri arasında ilişkinin %100 olduğunu gösterir. Ancak korelasyonun var olması ilişkiyi gerektirmez. Başka bir deyimle iki parametre arasında korelasyon varlığına dayanarak bir parametreye esas verinin diğerini etkilediği sonucu çıkarılamaz. Ancak korelasyonun varlığı parametreler arası ilişkinin daha detaylı olarak değerlendirilebilmesi için önemli bir işaret teşkil eder. Tespit edilebilen ilişkiler işletmele-

rin bu ilişkilere göre yeni yaklaşımlar geliştirmelerini ve sorunlarına çözümler bulabilmelerini sağlar.

Aykırı değerlerin tespiti

Korelasyon da dâhil olmak üzere veri bünyesindeki birçok ilişkinin varlığının tespit edilebilmesi için aykırı değerlerin elenmesi gerekmektedir. İnsanlara ait boy ve vücut ağırlıklarının yer aldığı bir tabloda, çok uzun bir boy ölçüsüne karşılık aşırı düşük bir vücut ağırlığına sahip istisnai ve az sayıda gözlemler aykırı değerler için örnek olarak gösterilebilir. Gözlem sayısının az olduğu bir veri setinde aykırı değerler hemen göze çarpabilir ve dolayısıyla elle eleme yapmak mümkündür. Ancak veri seti çok büyüdüğü durumlarda ya da özellikle verilerin teker teker incelendiği takdirde dahi aykırılık bağlamında üzerine yorum yapılmasının güç olduğu soyut olduğu durumlarda mevcut gözlemler içerisinden aykırı değerlerin elenmesi kolay olmayacaktır. Bu tür durumlarda aykırı değerleri tespit etmek için bir takım özel yöntemler kullanılır. Bu yöntemler ve çeşitli yaklaşımlar yardımı ile verilen bir gözlemler seti içinde istisnai olma ihtimali yüksek gözlemler tespit edebilir. Bu metotlardan en yaygın olanı *Thompson Tau* testidir.

Bu test tek parametreye dayalı olarak aykırı değerlerin seçilebilmesini sağlar (Dieck, 2007: 169).Boy ve vücut ağırlığı örneğinde, sadece boya ve sadece vücut ağırlığına göre ayrı ayrı aykırı değerler bu test yardımı ile ortaya çıkarılabilir. Ancak zaman zaman ayrı ayrı ele alındığında aykırı olmayan ancak iki parametre bir arada ele alındığında aykırılık teşkil eden veriler olabilir. Örneğin bir gözlemde boy çok uzun da olsa aykırı olmayabilir ancak bu boya göre ağırlık çok düşük olabilir bu durumda bu gözlem aykırı olacaktır.

Böyle bir durum birden fazla parametrenin mevcut olduğu durumlara örnektir. Birden fazla parametreye bakılarak aykırılık tespiti için ise *Mahalonobis* aykırılık testi (Varmuza ve Filzmoser, 2009: 47) gibi testler kullanılır.

Genellikle aykırı değerin tespiti verilerin gerçek analizine girmeden önceki ön işlem prosedürü gibi ele alınmaktadır. Ancak aykırı değer tespiti bizzat veri madenciliği aracı olarak da kullanılabilir. Özellikle usulsüzlük/dolandırıcılık şüphesi barındıran parasal hareketlerin tespiti (Ganji, 2012: 1035) ve şüpheli kayıtlar (Phua, Lee vd., 2012: 1005) gibi verilerin tespitinde aykırılık tespiti metotları faydalı sonuçlar sağlamaktadır.

Görselleştirme

Veri görselleştirme de veri madenciliğinde önemli değere sahip olan araçlardan biridir. Veri görselleştirme verilere ait birçok istatistiksel özelliğin görsel

olarak ve hızlıca sunulabilmesini ve anlaşılabilmesini sağlayan yöntemler bütünüdür. Tek parametreli bir verinin zaman içindeki artışı veya vektörel düzlemdeki dağılımı, iki parametreli verinin noktasal dağılım grafiğinde ilişkilerinin izlenmesi, üç parametreli veriler için de üç boyutlu düzlemde gösterim mümkündür. Kategorik verilerin pasta diyagramda gösterimi, histogramlar, mum grafikler, verilerin coğrafi haritalar üzerinde gösterimi, infografikler gibi birçok yöntem görselleştirme başlığı altındadır. Görselleştirme ile veri ne kadar büyük ve ilişkiler ne kadar karmaşık olursa olsun elde edilen sonuçlar herkes tarafından anlaşılabilecek duruma getirebilmektedir.

Sınıflandırma

Veri madenciliğinde kullanılabilen ve birçok işletme problemlerinde kullanılabilecek özelliklerden birisi de sınıflandırmadır. Sınıflandırma algoritmaları sayesinde verilerin daha önce veri tabanı tasarlanırken filtreleme amacıyla konmamış parametreler olmasa bile, verilerin mevcut özellikleri ve parametreleri kullanılarak sınıflandırmaya olanak verir.

Tablo 1: Koşullu Olasılık Formülasyonu

Sembol	Anlamı	Örnek
B	Bir durum	Gerçekten kanser olma hali
P(B)	B durumun gerçekleşme olasılığı	Gerçekten kanser olma olasılığı
A	Başka bir durum	Kanser testinde pozitif çıkma
P(A)	A durumunun gerçekleşme olasılığı	Kanser testinde pozitif çıkma ihtimali
$P(A \cap B) = P(B \mid A)$	A durumu ve B durumunun bir arada var olma olasılığı (A durumu kesin iken B durumuna rastlanma oranı)	Hem Kanser testinden pozitif çıkıp hem de gerçekten kanser olma ihtimali (Gerçekten kanser olan hastalardan kanser testinde pozitif çıkma oranı)
$P(A \mid B)$	B durumu söz konusu iken A durumunun gerçekleşme ihtimali	Kanser testi pozitif çıkan bir hastanın gerçekten de kanser olma ihtimali: $P(A \cap B)$.

Makine Öğrenmesi

Makine öğrenmesi de yakın zamanda veri madenciliği alanında yaygınlaşmaya başlamış özel yöntemlerden biridir. Makine öğrenmesi kabaca; bilgisayarların mevcut veri setlerindeki yapıları keşfederek, verileri değerlendirebilecek hatta tahmin yapabilecek duruma getirilmesi sürecidir. Bu süreç sonunda algoritmalar çok değerli çıkarımlar yapılabilmektedir. Sadece tek tip makine öğrenmesinden ziyade farklı durumlarda kullanılabilecek çok sayıda makine

öğrenme algoritması vardır. Makine öğrenmesinin temel boyutlarını ortaya koyması bakımından özellikle işletme ile ilgili veri madenciliği çalınmalarında kullanılan algoritmaların çalışma şekli sık kullanılan algoritmalardan biri olan Naive Bayes algoritması özelinde, aşağıda ifade edilmiştir:

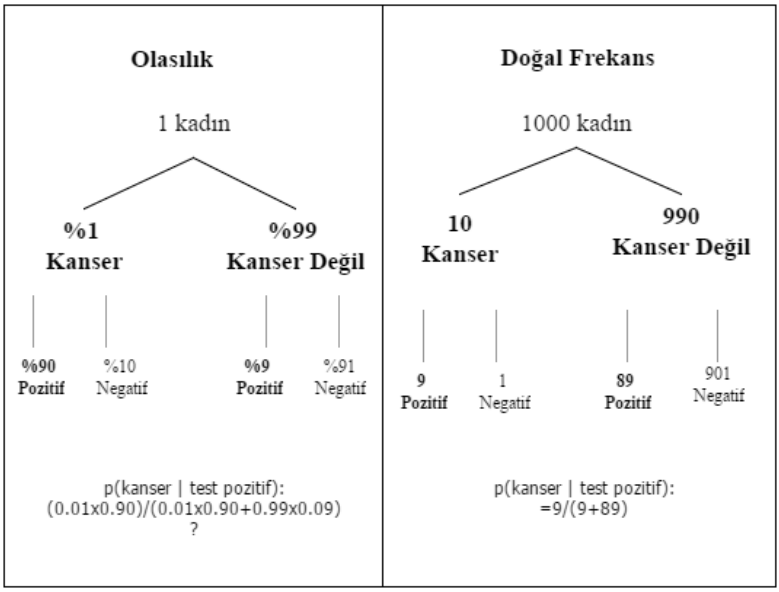
Naive Bayes algoritmasının mantıksal temelleri 18. yüzyılda Thomas Bayes tarafından ortaya konulan yaklaşımlara dayanmaktadır. Bu metotlar genel olarak olayların olasılıklarının değerlendirilmesi ve her yeni ek bilgi ışığında yeniden ele alınmasını sağlayan metotlardır (Lantz, 2013: 90). Bayes Teoremi kapsamındaki durumsal olasılık kavramı bu algoritmaya temel teşkil etmektedir. Bu kavram hem formülü hem de sıkça verilen medikal test örneği ile birlikte aşağıdaki senaryo ile ifade edilebilir. 1000 kişinin dâhil olduğu bir kanser tarama testi olsun:

P(A|B) ifadesi Bayes teorisinin özeti gibidir. Bu ifade böyle bir kanser testinde pozitif sonuca rağmen kanser olmama ihtimalinin varlığını da vurgular. Aynı şekilde P(B|A) ifadesi de gerçekten kanser olduğu halde kanser testinden pozitif çıkan hastaların oranına gönderme yaparken, kanser olduğu halde testte kanser değilmiş gibi gözüken hastaların varlığını ifade eder. Sonuç olarak Bayes'in koşullu olasılık formülü yukarıda anılan P(B), P(A) ve P (A ∩B) değerleri yardımı ile P (A|B) değerinin hesaplanmasını temin eder. Başka bir deyimle, geçmiş istatistikleri belirli olan bir hastalık tarama testinin güvenilirliğini ortaya koyar. Bayes koşullu olasılık formülü aşağıdaki gibidir (Lantz, 2013):

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

Bu formülün olasılık değerlerine göre hesaplanması zaman zaman karmaşık hale gelmektedir. Bu amaçla Gerg Gigerenzer, kullanılacak değerlerin frekans yüzdesi olarak değil doğal frekans olarak alınmasını önermektedir. Çalışmada örnek olarak 1000 kadının katıldığı bir kanser testinde 10'unun gerçekten kanser olduğu halde bunlardan 9'unun testte pozitif çıktığı (kanserli olarak teşhis edildiği) 1'inin ise negatif çıktığı (kanserli olmadığı teşhisi) örneğinden yola çıkılmaktadır. Bu örneğe göre kalan 990 hasta gerçekte kanser değildir. Bu durumda bu testin güvenilirliği ya da testin pozitif çıktığı durumda gerçekten kanser olma olasılığı formüle göre (bkz: sağ blok) 9/(9+89)=%9,1 olacaktır. Doğal frekans yerine yüzdesel frekansın kullanılması halinde de (bkz: sol blok) aynı değer elde edilecektir. (Gigerenzer, 2014: 247). Bulunan bu değerler ise kanser testi pozitif çıkan bir hastanın gerçekte kanser olma olasılığını ortaya koymaktadır.

Şekil 1: Bayesyan Koşullu Olasılığın Doğal Frekansla Hesaplanması



Bayesyan koşullu olasılığın veri madenciliğinde kullanımı aynı formülün veri setleri üzerinde kullanılması ile mümkün olmaktadır. Kanser testi örneğinden gidilecek olursa aynı ilişkinin başka veriler üzerinde uygulaması aşağıdaki Tablo 2'de olduğu gibi özetlenebilir.

Buna göre tıpkı hasta örneğinde olduğu gibi örneğin, meyvelere ait şekil, ölçü ve renklerin yer aldığı üç satırlı bir tabloda bu meyvelerin gerçekte ne olduklarının yazılı olduğunu varsayalım. Naive Bayes algoritması dâhilinde, örneğin sarı ve 10 cm'den uzun olan meyvelerin %90 olasılıkla muz olduğu, turuncu, yuvarlak ve 7cm'dan büyük meyvelerin ise %70 olasılıkla portakal %30 olasılıkla greyfurt olduğunu tespit edilebilir. Bu algoritma artık ilgili meyvelerin ölçüsüne bakılarak yüzde kaç ihtimalle hangi meyve olduğuna karar verebilmektedir. Meyve örneği Naive Bayes algoritması ifade edilirken basitliğinden ötürü sıklıkla kullanılan örneklerden biridir. Kuşkusuz bu örnek meyveler yerine daha karmaşık problemlerde kullanılmaktadır. Örneğin hücrelerinin boyut, ölçü ve başka özelliklerine göre kanser hücresi olup olmadıkları Büyük veri içinden insan müdahalesi olmaksızın tespit edilebilmektedir. Aynı şekilde bireylerin yüzlerce kritere sahip tıbbi verisi (yaş, kilo, boy, kan değerleri) kullanılarak kalp hastası olma ihtimalleri yine bu algoritma ile hesaplanabilir. Burada bu algoritmanın temel istatistiksel yöntemlerden farkı, belirli bir öğrenme verisi üzerinden yapıları anlayarak sonraki daha sonra sunulan edilmemiş verileri öğrendiği yapılara göre etiketleyebilmesidir.

Tablo 2: Koşullu Olasılık Yaklaşımının Örnek Kullanım Alanları

	Hasta örneği	**Veri Madenciliği**	**Metin madenciliği**
Kategorize edilecek nesne	Hastaların Hücre Örnekleri	Meyveler	Metinler
Kategori	Hücre Kanser/Kanser Hücresi Değil	Meyvenin cinsi	Spam/Spam Değil
Nesnenin gerçek kategorisi	Otopsi Raporları (Kesin)	Görsel inceleme	Metnin insanlar tarafından okunarak değerlendirilmesi
Nesnenin saptanan kategorisi ne esas girdi veri	Test veya biyopsiden elde edilen veriler	Şekil, ölçü ve renk değerlerine göre tespit	Belirli kelimelerin tekrarları

Metinsel verilerde Naive Bayes algoritmasının kullanılması ise daha karmaşık biçimde işlemekle birlikte temel mantığı yukarıda anılan şekle benzer şekilde işlemektedir. Metinsel veriler önce el yordamıyla kategorize edilmekte daha sonra algoritma bu eğitim verisi üzerinden öğrenimini sağlayarak sonraki etiketsiz veriyi etiketleyebilmektedir. Naive Bayes algoritmasının metinsel verilerde en çok kullanıldığı alanlardan biri olan epostaların spam olmaması bu şekilde anlaşılmaktadır. Günümüzde kullanılan modern önde gelen Gmail, Outlook.com, Yandex gibi web tabanlı eposta servislerinin gibi servisler bu ve benzer makine öğrenmesi algoritmaları ile gelen bir mailin içeriğine göre spam klasörüne direkt olarak yollamaktadır. Kuşkusuz bazı durumlarda spam olmadığı halde spam klasörüne giden e-postalar olduğu gibi, spam olduğu halde ana gelen kutusuna düşen e-postalar da bulunmaktadır. Bu durumun olma olasılığı da koşullu olasılık çerçevesinde ele alınabilir.

Metinsel verilerde Naive Bayes algoritmasının kullanıldığı alanlardan bir diğeri ise metinsel verinin hangi dilde yazıldığının tespitidir. Doğal dillerin spesifik bazı özellikleri olmakla birlikte her metinde bu özellikler ortaya çıkmadığından doğal dillerin mantıksal olarak her yerde ve her zaman geçerli olan ayrıştırıcı tanımının yapılması mümkün değildir. Bu nedenle bir metin bloğunun hangi dilde yazıldığı da makine öğrenmesi algoritmaları ile sağlanabilmektedir. Metin madenciliğinin potansiyelini ve yönünü gösteren en ilginç çalışmalardan biri sosyal medyada bireyler tarafından yazılan metinlerden psikanaliz gerçekleştirme imkânıdır.

Bu kapsamda yapılan çalışmalardan birinde Twitter'da yazılan girdiler üzerinden, bu girdilerin yazarlarının intihar eğilimi hesaplanmaya çalışılmaktadır. Bu çalışmada SVM ve Lojistik Regresyon algoritmaları ile analiz edilen girdi yazarlarının psikolojisini ortaya çıkarmaktadır (O'Dea vd., 2015: 186).

Makine öğrenmesi anılan alanlar dışında, işletme veya kamusal verileri üzerinden yolsuzluk tespiti, arama motorları, resim, ses, el yazısı tanıma teknolo-

jileri gibi alanlarda da kullanılmaktadır. Özetle makine öğrenmesi, verilerin süpervize ve süpervize olmayan yöntemleri ile verilerden etkin bir biçimde değerli iç görüler elde edilmesini sağlayan güçlü yaklaşımlar bütünüdür.

BÜYÜK VERİ KAVRAMI

Şimdiye kadar ifade edilen dört perspektif işletmelerin veri kullanım ilgi düzeylerini ele alırken; veri, veriye esas teşkil eden olayların tali sonucu olan bir araç olarak ele alınan ve tahmin perspektifinde ise işletme için değeri bir ölçüde hissedilebilen bir olgu olarak ortaya çıkmaktadır. Sözgelimi, işletme satışlarının kaydı ve tarihi ve ürün cinsinin mevcut olduğu bir veri tabanında anılan dört perspektifin ötesine geçmek olası olmayabilir. Ancak verilerin miktarı, çeşitliliği ve bütünleşme kabiliyeti arttıkça, veri işletmenin finansal bir varlığı haline bile gelebilmektedir. Nitekim bu durumda veri basit bir araç olmaktan çıkıp işletme için sürekli değer sağlayan bir kaynak haline gelmektedir.

Büyük veri ile ilgili olarak akademik literatür ve endüstride bir çok farklı tanım yapılmaktadır). Bu konuda en yaygın tanımlardan biri Gartner adlı Amerikan Şirketine ait raporda yapılan ve 3V olarak kısaltılan tanımdır (Ward ve Barker, 2013: 1). 3V Volume (Hacim), Variety (Çeşit) ve Velocity (Hız) kavramlarının baş harflerinden oluşmaktadır. Bu kavramlar aşağıda detaylı olarak ele alınmıştır:

Hacim: Veri depolama ve işleme maliyetlerinin ucuzlaması daha fazla verinin işlenebilmesine olanak vermiştir. Öyle ki geçen son yarım yüzyıl içerisinde veri depolama maliyeti, kabaca, her iki yılda bir önceki döneme göre yarı yarıya ucuzlamıştır (Mayer-Schönberger ve Cukier, 2013: 101). 2013 yılı itibariyle dünya üzerindeki mevcut verilerin %90'ının son iki yıl içerisinde üretildiği tahmin edilmektedir (Hurwitz vd., 2013). Google'ın CEO'su Eric Schmidt'in 2010 yılında yaptığı açıklamaya göre insanlık artık her iki günde bir, insanlığın var oluşundan bugüne ortaya çıkan veri miktarı kadar veri üretmektedir. Örneğin 2012 yılında internet kullanıcıları 2 milyardan fazla bilgisayar ve cep telefonu ile 4 eksabayt veri ürettiler. Facebook kullanıcıları her ay 30 milyar adet girdi oluşturmuşlardır (Salminen ve Kaartemo, 2014). Bir verinin ne zaman Büyük veri kabul edileceğine dair tanım sürekli olarak farklı algılanmaktadır. 70'li yıllarda megabayt düzeyindeki bir veri "büyük" iken bugün terrabayt ve üzerindeki veriler için "büyük" denmektedir (Salminen ve Kaartemo, 2014). Başka bir deyimle milyonlarca satır veri içeren cigabayt düzeyindeki veri bile zaman zaman büyük veri sayılmamaktadır. Bu noktada büyük verinin hafızada işgal ettiği alandan ziyade, bu verinin işlenmesinin güçlüğüne büyük veri veya normal veri olarak nitelendirilmesi daha doğru bir yaklaşım olacaktır.

Genel olarak sadece bir bilgisayarda işlenmesi güç olan veri büyük veri olarak nitelendirilmektedir.

Cep telefonları, akıllı kol saatleri, akıllı gözlükler ve hatta otomobiller artık sahip oldukları onlarca farklı sensor sayesinde sürekli olarak veri üretilebilmektedir. Öte yandan büyük verinin kapsamın sadece yukarıda ifade edilen yapılandırılmış veri ile de sınırlı değildir. Yapılandırılmamış veri olarak nitelendirilen veriler de büyük verinin kapsamındadır.

Bir veriye ait daha önceden tanımlı model yoksa bu veri yapılandırılmamış veri olarak kabul edilmektedir. Bir şirkete ait e-postalar, video kayıtları, resimler, sesler yapılandırılmamış veridir. Bu verileri yönetmek ve analiz etmek güçtür. Yapılandırılmamış verilerin kritik olmasının nedeni işletmelerin önemli bilgilerinin yaklaşık olarak %80'inin yapılandırılmamış verilerden oluşmasıdır (Grimes, 2005).

Çeşit: Verilerin çok çeşitli format ve yerlerde olması olgusu da büyük verinin diğer özelliğidir. Büyük veri kavramı, veri tabanında kayıtlı girdilerden fazlasını ifade eder. E-postalar, resim formatında taranmış faturalar, veri tabanındaki gerçek işlem kayıtları, gerçek işlemlerin tali özelliklerini içeren loglar, müşteri iletişim merkezleri gibi merkezler için ses kayıtları gibi farklı formatlarda ve yerlerde depolanan verilerin tamamı bir işletme için değer teşkil eder. Tüm bu farklı veri türlerinin incelenmesi farklı teknoloji ve yaklaşımlar gerektirmektedir. Örneğin, ses dosyalarının metin dosyalarına çevrilmesi ses tanıma teknolojisi gibi özel bir alanı ilgilendirirken, metin dosyalarından hangi müşterinin olumlu, hangi müşterinin olumsuz görüş ifade ettiğini analiz etmek veri madenciliğinin alt dallarından biri olan metin madenciliğini ilgilendirmektedir.

Hız: Büyük veri kavramının son özelliği ise verilerin oluşma ve işlenme hızı ile ilgilidir. Geleneksel yaklaşımla bir alışveriş sitesi ele alındığında bu sitede sadece gerçekleşen satışlara dair kayıtlar önemli olarak ele alınacaktır. Ancak müşterilerin alışveriş sitesindeki davranışları, incelendiği ürünler, aynı ürünle ilgili sayfada kalma süresi gibi parametreler de müşterinin satın alma davranışının tahmin edilmesi için değerli bir kaynaktır. Diğer yandan bu kaynakta elde edilen verinin hemen değerlendirilerek kullanılması gerekmektedir nitekim bu müşteri dakikalar içinde farklı bir alışveriş sitesine yönlenebileceği gibi satın alma davranışından vaz geçebilir. Bu durumu aşmak için ise verileri eş zamanlı olarak analiz ederek müşteriye alternatifler sunan öneri modelleri geliştirilmelidir. Burada verilerin oluşmasından bir süre sonra analizi yerine hemen değerlendirilmesi olgusu hız kavramı ile ilintilidir. Bu analizlerin hızlı bir şekilde gerçekleştirilebilmesi olgusu da aynı şekilde büyük veri ile ilgili çalışmalardaki zorluk noktalarından biridir.

Başka bir tanıma göre büyük veri, işletmelerin içinde ve dışında geleneksel ve dijital kaynaklardan ortaya çıkan verilerin teşkil ettiği koleksiyondur. Öyle ki bu koleksiyon işletme için sürekli analiz ve keşif kaynağı olan bir kaynak sağlamaktadır (Salminen ve Kaartemo, 2014).

Geleneksel anlamı ile veri ile daha sonradan ortaya çıkan büyük veri arasındaki fark bilgisayar bilimleri bakımından 3V ile özetlenen yeni özelliklerdir. Geleneksel veri ile Büyük veri ayrımının işletmeler düzleminde farklı 3V tanımı ötesinde, işletme verilerinin daha önce pek yaygın olmayan ve yenilikçi bir biçimde kullanımında yatmaktadır.

Örneğin İngiliz Perakende devi Tesco 2009 yılında yaptığı bir araştırmada işletme içi bir veri olan et satışları ile işletme dışı bir veri olan hava sıcaklığı arasında ilginç bir ilişki keşfetmiştir. Londra bölgesinde her 3 derecelik hava sıcaklığı artışının et satışlarında %10 artmaya neden olduğu ortaya çıkmıştır. Bu ilişki barbekü kullanımına bağlı olarak değerlendirilmiş ve doğal görülmüştür. Ancak bazı başka sonuçlar da elde edilmiştir. Yine 3 derecelik hava sıcaklığı artışı marul satışlarında %15 artışla ilişkilendirilmiştir. Tesco Büyük veriden elde edilen bu çıkarımlar sayesinde 16 milyon sterlin tasarruf sağlamayı başarmıştır (Aksoy,2014: 98).

Turkcell müşterilerinin lokasyon verilerinden faydalanarak hangi profildeki müşterinin herhangi bir zaman diliminde nerede olduğu bilgisini anonim olarak kurumsal müşterileri ile paylaşmaktadır. Böylece kurumsal müşterilerden biri pazarlama stratejisini belirlemeye hangi müşteri profiline, nerede ve ne zaman ulaşacağı bilgisini bilerek başlamaktadır (Aksoy,2014: 99).

Büyük veri, geleneksel veri bakış açısını aynen devralmakla birlikte bazı ek özellikleri de içermektedir. Bu özellikler verinin kendisi ile ilgili olabileceği gibi depolanma ve işlenme şekli ile ilgili de olabilir. Bu özelliklerden bazıları aşağıdadır:

API Entegrasyonu: Açık programlama ara yüzleri (API'ler) çeşitli yazılımların insanların kullanımına sunulan ara yüzler dışında birbiri ile iletişim kurabilmelerini sağlamak amacıyla geliştirilen soyut ara yüzlerdir. Bu ara yüzler, onları kullanan uygulamaların kullanıcıları tarafından genellikle fark edilmezler ancak uygulamalara çok ciddi katkılar sağlarlar. Örneğin, ülkemizdeki bir çok kamu kurumunun uygulaması MERNİS (Merkezi Nüfus ve İdare Sistemi) adlı uygulamanın API'si ile personel veri tabanları arasında ilişki kurmaktadır. Aynı şekilde Navigasyon programları artık trafik verilerini de kullanmaya başlamıştır bu trafik verileri ise ilgili veri sunucu servislerin API'lerinden sağlanmaktadır. Trafik verisinin kullanımı, navigasyon programlarının sürücüleri daha az trafik bulunan alternatif güzergahlara yönlendirmesini sağlayabilmektedir.

API'ler büyük veri mimarisinin çekirdeğinde yer alacak özelliklerdendir. Sadece tek kaynaktaki verilere dayalı bir mimarinin büyük veri yaklaşımı altında ele alınması güçtür (Hurwitz, 2013)

Dağıtık Hesaplama: Geleneksel anlamı ile veriden söz edildiğinde, verilerin aynı yerde olduğu varsayımı zımnen de olsa yapılmış olmaktadır. Ancak büyük veri mimarisinde, kaynak gereksinimleri veya tabiatı itibariyle veriler dağıtık halde olmakta ve aynı şekilde işlenmeleri de birden fazla bilgisayarın bir arada organize olarak çalışmasını gerektirebilmektedir.

Yeni araçlar ve büyük veri ekosistemi: Bilgisayarlarda veri depolama süreci, verilerin düz metin dosyalarında derlenmesinden, ilişkisel veri tabanlarına dönüşmesi ile ciddi bir aşama kaydetmiştir. Birbirine bağlı tablolar düz veri dosyalardan sağlanan analizlerden daha fazlasının elde edilmesine olanak sağlamıştır. Büyük veri ise bu trendin devamı olarak yapılandırılmış olmayan veriler de dahil olmak üzere farklı kaynaklarda, farklı formatlarda ve çok büyük miktarlardaki verilerin işlenmesi için Hadoop, MapReduce, BigTable gibi araçların kullanımını gerektirmektedir. Bu araçlar anılan devasa verilerin geleneksel yöntemlerle işlenmesi yerine karmaşık ancak etkin algoritmalarla işlenebilmesine olanak vermektedir.

İşlenebilen farklı girdiler: Geleneksel anlayış içerisinde bir veri genellikle sayısal veya nominal kayıtlardan oluşmaktadır. Oysa dev veri ses, metin ve video gibi verilerin bile işlenebilmesine ve sorgulanabilmesine olanak veren büyük bir ekosistemi ifade eder. Özellikle metin madenciliği alanı veri madenciliğinin bir alt dalı olarak insanlar tarafından kullanılan doğal dille yazılmış metinler içerisinden anlamlı sayısal verilerin otomatize olarak elde edilebilmesine olanak verir. Metinsel veri artık finansal tahmin modellerinde kullanılabilen, geleneksel parametrelere göre yeni bir parametredir. Bu parametrelerin geleceğe tahmin modellerinde ek parametre olarak kullanılma imkânı da bulunmaktadır. Metin madenciliği aslında finansal piyasalar dışında, sosyal medyanın otomatik olarak izlenmesi ve müşteri görüşlerinin eş zamanlı olarak analiz edilebilmesi için de kullanılmakta ve başarılı sonuçlar sağlamaktadır

Büyük veriyi standart veriden ayıran bu özellikler sadece tanımlayıcı düzlemde kalmamaktadır. Bu durum verilere yönelik olarak özel bir perspektifin ortaya çıkmasını da sağlamıştır. Bu perspektif, Büyük verinin çeşitlilik imkânı sayesinde daha önce bir biri ile ilgisiz duran verileri bir arada kullanarak işletme için yalnızca içgörü değil aynı zamanda değer yaratılabilmesini sağlamıştır. Söz gelimi artık açık olarak sunulan hava durumu API'leri ile geleceğe dair hava tahminleri verisi anlık olarak alınabilmektedir. Bir lojistik işletmesi bu API'den aldığı verilerle bu lojistik operasyon planlarını birleştirerek olumsuz

hava koşullarına karşı otomatize edilmiş önlemler alabilir. Böyle bir durumda lojistik operasyonlara dair verilerle, hava durumu verisi ilgisiz gibi gözükse de bir arada kullanılarak işletme için değer yaratmaktadır.

Aslında ilişkisiz gibi duran veri setleri arasındaki ilişkileri inceleme olgusu Büyük veriden döneminden önce de var olmuştur. Hawthorne deneylerinde her ne kadar aralarında ciddi bir ilişki olmadığı düşünülse de ışık düzeyi ile çalışan verimliliği arasındaki ilişki gibi ilk bakışta zor tahmin edilebilir ilişkiler ele alınmıştır (Mayo, 1933: 55). Ancak bu durum günümüzde hemen hemen her işletmenin ihtiyaç duyduğu takdirde kullanabileceği hale gelmiştir.

Bu noktada vurgulanması gereken husus, büyük veri perspektifinin yukarıda anılan veri madenciliği perspektifinin daha geniş ve yönetilmesi güç hale gelmiş versiyonu olduğudur. Öyle ki veri madenciliğinde de bir düzeyde veri yönetimi problemi varsa da, büyük veride bu kapsamlı ve ayrışık başka bir inceleme sahası haline gelmiştir. Bu aşamada veriler artık hemen ele alınıp kullanılabilecek girdilerden ziyade, toplanması, derlenmesi, temizlenmesi ve ayrıştırılması ayrı bir göreve dönüşmüş durumdadır. Başka bir deyimle veri artık ticari bir meta haline gelmiştir ve ticari bir meta gibi stoklanması, yönetilmesi ve ele alınması gerekmektedir. Aynı zamanda tüm bu süreçlerin çok hızlı ve doğru şekilde gerçekleştirilmesi gerekmektedir.

DEĞERLENDİRME VE SONUÇ

Bu çalışmada işletmecilik perspektifinde veriye bakışın zaman içerisinde değişimi ele alınmış olup, ilgili teknolojilere göre işletmenin konumu değerlendirilmiştir. Bu konumu somutlaştırabilmek adına işletmenin veri karşısındaki bilinç düzeyi dört farklı düzeyde kategorize edilmiştir. Bu kategorizasyon, sektörde ve literatürde işletmelerin taleplerinin doğal dışavurumudur. İşletmeler bu bilinç düzeyleri boyunca veri kavramına farklı bakmaktadırlar. Günümüz itibariyle en yüksek bilinç düzeyi veri madenciliği yapabilen işletmelerin düzeyidir. Bu düzeydeki olanaklar ise bu başlık altında ele alınmış ve sık kullanılan bazı veri madenciliği araçları ele alınmıştır. Bu araçlardan en güçlü olanı ise makine öğrenmesidir. Makine öğrenmesi artık kendi içerisinde bir sistem olarak değerlendirilmelidir. Öte yandan, veri kavramı ile ilgili büyük veri kavramı da aslında çok miktarda verinin analizi şeklindeki bir anlayıştan çok daha fazlasını ihtiva eden. Büyük veri, verinin ticari bir meta kadar somut olarak ele alındığı, yönetilmesi güç ve analitiği sonucunda önemli çıkarımların yapılabilmesine olanak sağlayacak veri türüdür. Büyük veri bir dizi araç veya uygulanacak bir proje olarak değil bir strateji olarak ele alınmalıdır.

KAYNAKÇA

Aksoy, C. (Ocak-Şubat 2014). Müşteriye Daha Yakın Olmak. *Harvard Business Review Türkiye*, 96-101.

Dieck, R. (2007). *Measurement Uncertainty Methods and Applications, the Instrumentation*. (4. Bs.). New York: *Systems and Automation Society (ISA).*

Ganji, V. R. (2012). Credit Card Fraud Detection Using Anti k-Nearest Algorithm. *International Journal on Computer Science and Engineering*, *4*(6), 1035-1039.

Grimes, S. (2005). Structure, Models and Meaning. *InformationWeek*. 20 Mart 2016 tarihinde http://informationweek.com/software/business-intelligence/structure-models-and- meaning/59301538 adresinden erişildi.

Gigerenzer, G. (2014). *Risk Savvy: How to Make Good Decisions.* (1. B.s). New York: Penguin.

Hand, D. J. (1999). Statistics and Data Mining: Intersecting Disciplines. *ACM SIGKDD Explorations Newsletter*, *1*(1), 16–19.

Hurwitz, J., Nugent, A., Halper, F. ve Kaufman, M. (2013). *Big Data For Dummies.* (1. Bs.). New Jersey: John Wiley & Sons.

Lantz, B. (2013). *Machine Learning with R.* (1. Bs.). Birmingham: Packt Publishing Ltd.

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity and Variety*. META Group Araştırma Raporu, 20 Mayıs 2016 tarihinde https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf adresinden erişildi.

Mayer-Schönberger, V., ve Cukier, K. (2013). *Big data: A Revolution that will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

Mayo, E. (1933). *The Human Problems of an Industrial Civilization,*. New York: Routledge Taylor&Francis Group.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C. ve Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions*, *2*(2), 183–188.

Phua, C., Lee, V., Smith, K. ve Gayler, R. (2012). A Comprehensive Survey of Data Mining-based Fraud Detection Research. *Computers in Human Behavior*, *28*(3), 1002–1013.

Salminen, J. ve Kaartemo, V. (Ed.). (2014). *Big Data: Definitions, Business Logics, and Best Practices to Apply in Your Business*. New York: Amazon

Varmuza, K. ve Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics* (1.Bs.). Florida: CRC Press.

Ward, J. S., ve Barker, A. (2013). Undefined by data: a survey of big data definitions. *arXiv preprint* arXiv:1309.5821.