Veri Analizinde İstatistik mi Veri Madenciliği mi?

İlkim Ecem EMRE, Çiğdem SELÇUKCAN EROL
Enformatik Bölümü, İstanbul Üniversitesi, İstanbul, Türkiye
ecememre@gmail.com, cigdems@istanbul.edu.tr
(Geliş/Received:01.08.2016; Kabul/Accepted:20.04.2017)
DOI: 10.17671/gazibtd.309297

***Özet*—** Günümüzün teknolojik gelişmeleri, veri boyutlarının ve çeşitliliğinin artması klasik istatistiksel yöntemlerin yetersiz kalabileceği durumları da beraberinde getirmektedir. Veri analizinde yıllardır kullanılan istatistik, son yıllarda veri madenciliği ile yan yana yer almaktadır. Veri analizi çalışmalarında kullanılan istatistik ve veri madenciliği yöntemleri birçok farklı araştırma alanında kendilerine yer bulabilmektedir. Özellikle sağlık alanında veri analizinde istatistik sıklıkla kullanılmakta, veri madenciliği yöntemleri ise son yıllarda araştırmacılar tarafından fark edilmektedir.

Bu çalışmanın amacı, her iki alanın incelenerek aralarındaki farklılıkların ve benzerliklerin ortaya konmasıdır. Özellikle sağlık alanında veri madenciliğinin çok yaygın olarak bilinmemesi ve sağlık çalışanlarının genellikle istatistiksel yöntemlere bağlı kalıyor olması çalışmanın çıkış noktasını oluşturmaktadır. Çalışma kapsamında veri madenciliği ve istatistiksel yöntemlerle ilgili olarak bir literatür araştırması yapılmıştır. Sonuç olarak bu iki alan kavramsal açıdan incelenmiş, aralarındaki farklar ve benzerlikler özetlenerek ortaya konmaya çalışılmıştır. İlerleyen zamanlarda çalışmanın sağlık verisi ile uygulaması yapılarak genişletilmesi umulmaktadır.

***Anahtar Kelimeler*—** Veri Madenciliği, İstatistik, Veri Analizi

Statistics or Data Mining for Data Analysis

***Abstract*—** Today's technological advances, increase of volume and diversity of data, brings inadequate status of classical statistical methods with it. Statistics, which is being used for data analysis for many years, is located side by side with data mining in recent. Statistics and data mining methods can find place in many different research fields. Especially in health sector statistics is being used frequently, whereas data mining methods is started to get attention of researchers in recent years.

The aim of this study is to observe similarities and differences between two fields. Starting point of the study is that while statistical methods are being frequently used in health field, data mining methods is not known widespread in the same field. In this study, literature research is made about data mining and statistical methods. As a result of the research both fields are observed conceptually, differences and similarities between fields are tried to be summarized in the study. In future it is planned to expand the scope this study with application of data mining and statistical methods to the same health data.

***Keywords*—** Data Mining, Statistics, Data Analysis

1. GİRİŞ (INTRODUCTION)

Günümüzde teknolojinin gelişimi ve veri boyutlarındaki hızlı artış farklı kavramlarında hayatımıza girmesine neden olmuştur. Bilgi iletişim teknolojilerinin hayatın hemen her alanına girmesi, hızlı gerçekleşen teknolojik gelişmeler veri boyutlarındaki ve türlerindeki artışı tetiklemektedir. Bu nedenle veri yığınları arasından anlamlı, işe yarar bilgi elde edilebilmesi açısından veri analizi her geçen gün daha fazla önem kazanmaktadır. Veri analizinde, veriyi anlamlandırma ve kanıta dayandırma bilimi olan istatistik, araştırmacıları sınırlandırabilmekte ve büyük boyutlu veri kümeleri karşısında yetersiz kalabilmektedir. İstatistiksel yöntemler birçok alanda kullanılsa da veri analizinde temelini istatistiksel yöntemlerden alan veri madenciliği kavramı ortaya çıkarak, gerek yapısal gerek yapısal olmayan farklı tipte ve büyük boyuttaki verinin analiz edilebilmesine olanak sağlamaktadır. Veri madenciliği ve istatistik birbirine benzer hedeflere hizmet etse de aralarında bazı farklılıklar bulunmaktadır. Bu farklılıklar ve benzerlikler bu çalışma kapsamında incelenmiştir.

2. KAVRAMLAR (CONCEPTS)

2.1.İstatistik (Statistics)

Bilgiyi kanıta dayandırma ihtiyacından ortaya çıkan istatistik [1] yüzyıllardır veri analizinde ve veriyi anlamlandırmada kullanılmaktadır. Kökeni veri madenciliğinden çok daha eskilere dayanan istatistik "veriden öğrenme, belirsizliği ölçme, kontrol etme

bilimidir" [2]. Değişkenler arasındaki ilişkileri inceleyen, veri setini çeşitli hesaplamalarla analiz edip bulguları özetleyen, örneklemden çıkartılan sonuçların veri kütlesi için genelleştirmeye yarayan bir bilim dalı olarak istatistik veri madenciliğinden çok daha eski yıllarda ortaya çıkmış ve veri analiz etmede yüzyıllar boyunca kullanılmıştır. İstatistiğin tarihine bakıldığında 16-17. yüzyıllara kadar dayanan çalışmalar olduğu görülmektedir [3].

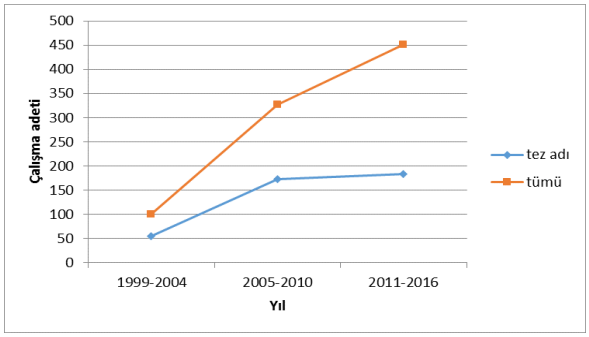
Ortalama, standart sapma gibi hesaplamalar, değişkenler arasındaki ilişkinin ortaya konması, sonuçların özetlenmesi, geleceğe yönelik tahminde bulunulması gibi çeşitli analizler için istatistikten biliminden yararlanılır. Buna bağlı olarak verinin olduğu her alanda veriyi analiz etmek için istatistikten yararlanılabileceğini söylemek yanlış olmayacaktır. Veri madenciliğinin kullanıldığı alanların hemen hepsinde istatistiksel yöntemler de kullanılabilmektedir. Kullanılacak alandan ziyade veri boyutu ve yapılmak istenen analizin niteliği hangi yöntemlerin kullanılacağı konusunda belirleyicidir. Önemli olan bu noktada veri ile ilgili nasıl bir analiz yapılacağına ve veri madenciliği veya istatistik yöntemlerinden hangilerinin kullanılacağına karar verilmesidir.

2.2. Veri Madenciliği (Data Mining)

Veri madenciliği veri miktarındaki artışla beraber çokça ilgi görmeye başlayan kavramlardan biridir. Toplamda 9000'den fazla kez alıntılanan Fayyad ve diğ. [4, 5] tarafından yazılan "*From Data Mining to Knowledge Discovery in Databases (KDD)*" ve "*The KDD Processs For Extracting Useful Knowledge From Volumes of Data*" makaleleri veri madenciliği ve veri tabanlarında bilgi keşfi konularıyla ilgili olarak 90'lı yıllarda oluşturulan en temel kaynaklardandır [4], [5]. Her iki makalede de hızla artan veri boyutlarından anlamlı bilgi elde edebilmek için, bilgisayar temelli yeni yöntemlere ihtiyaç duyulduğundan, veri tabanlarında bilgi keşfi ve veri madenciliği konularının araştırmacıların dikkatini çektiğinden bahsedilmektedir [4], [5]. Makalelerin yazılmış olduğu yılın üzerinden 20 yıl geçmesine rağmen bu konular hala daha güncelliğini korumakla beraber bugünün araştırmacıları için büyük öneme sahiptir. Farklı alanlardaki araştırmacılar açısından, veri madenciliğinin bilinirliği ve kullanımı 1996'dan bu yana gözle görülür şekilde artmış ve her geçen gün artmaya devam etmektedir.

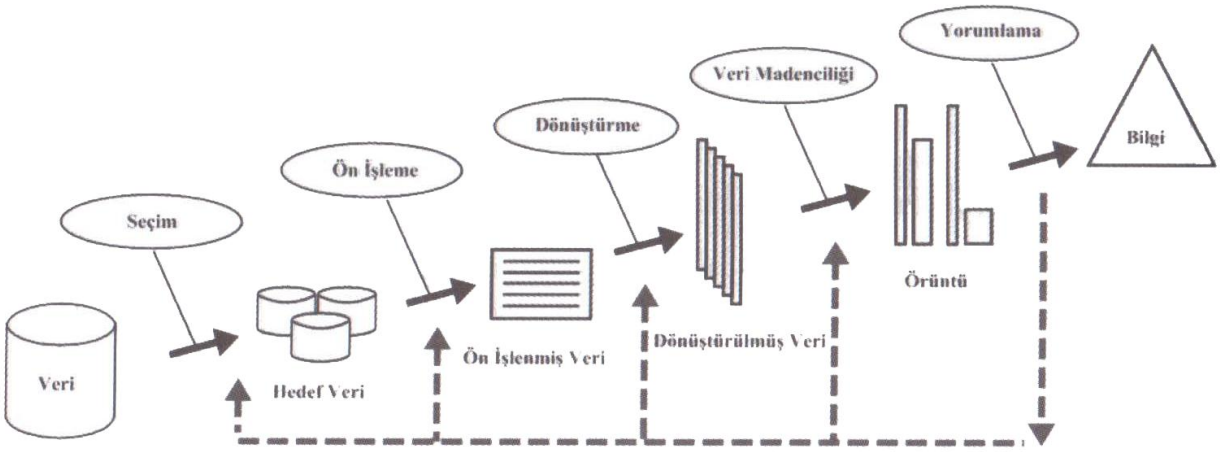
Fayyad ve diğ. [5] bilgisayar tabanlı veri madenciliği analizlerin konuşulmaya başlandığı 1960'lardan beri kötü bir üne sahip olduğunu belirtmiştir. Bir kişi eğer yeterince uzun süre veri setini ele alırsa içerisindeki örüntüleri bulabileceği düşünülmektedir. Ancak bu görüşün 1996'da ortaya konduğu göz önünde bulundurulursa o günden bugüne veri madenciliği ile ilgili düşüncelerin değişmiş olduğu ve alanla ilgili bu tip ön yargıların kırılmış olduğu görülmektedir. Ülkemizde yapılan çalışmalara bakıldığında veri madenciliği ile ilgili birçok yayın olduğu görülmektedir. Savaş ve diğ. [6] yaptıkları literatür çalışmasında Türkiye'de veri madenciliği alanında yapılmış çalışmalara yer vermişlerdir. Çalışma kapsamında 2003-2012 yılları arasında eğitim, mühendislik, bankacılık, tıp, ticaret ve telekomüniskasyon alanlarında yapılmış olan ve Türkiye'de yayımlanan 43 adet çalışmaya ait bilgiler verilmiştir.

Veri madenciliği ile ilgili çalışmaların yıllar içerisinde arttığını göstermek adına ayrıca Yükseköğretim Kurulu Başkanlığı'na ait Tez Merkezi web sitesinde de bir tarama yapılmıştır [7]. "Veri madenciliği" anahtar kelimeleri iki farklı filtre ile aranmıştır. İlk aramada veri madenciliği tez adları içerisinde, ikinci taramada ise aynı kavram herhangi bir kriter belirtmeden tüm alanlar içerisinde aranmıştır. Şekil 1'de arama sonuçları ve buna bağlı olarak veri madenciliği ile ilgili lisansüstü çalışmaların sayısı verilmiştir.



Şekil 1. Yıllara göre veri madenciliği lisansüstü çalışmaları
(Thesis studies about data mining by years)

Bilgi ve iletişim teknolojilerindeki gelişmeler çok hızlı bir şekilde gerçekleşmekte ve bu durum farklı kaynaklardan elde edilen farklı türlerde (yapısal, yarı yapısal ve yapısal olmayan) verinin birikmesine sebep olmaktadır. Bu veri, günümüzde büyük veri ile ifade edilen bir hale gelmiştir. Verinin hızlı artışı insanların bu veri yığınlarını yorumlayabilmelerini zorlaştırmaktadır. Bu yüzden büyük hacimli verinin analizi ya da yorumlanması için araştırmacılar veri madenciliği alanına yönelmektedir [8].



Şekil 2. Veri tabanlarında bilgi keşfi [5, 10]
(Knowledge discovery in databases)

Veri tabanlarında bilgi keşfi (*Knowledge Discovery in Databases - KDD*) kavramı adından da anlaşılacağı gibi büyük veri tabanlarında aşamalı bir şekilde bilgi keşfedilmesi sürecini ifade eder. Veri madenciliği ise bu keşif sürecinin belki de en önemli / kilit rol oynayan aşamasıdır [5]. Bir çok farklı alandan beslenen veri madenciliği; istatistik, makine öğrenmesi, yapay zeka, örüntü tanıma alanlarındaki gelişmelerden de beslenmektedir [9]. Akpınar'a [10] göre veri madenciliği yöntemleri istatistik, yapay zeka ve makine öğrenmesinden yararlanmaktadır.

Fayyad ve diğ. [5] veri tabanlarında bilgi keşfi sürecini "verideki geçerli, özgün, potansiyel olarak kullanışlı ve anlaşılabilir desenleri belirlemeye yarayan önemli bir süreç olarak" tanımlamıştır (Şekil 2). Veri madenciliği ile ilgili çalışan araştırmacılar ise bu tanımlamayı veri madenciliği kavramını tanımlamak için kullanmışlardır. Bu noktada KDD ve veri madenciliğinin tanımının birbirleri yerine kullanılabildiği görülmektedir. Kuonen [1], Kumar ve Bhardwaj [9] ve Kapoor [11] çalışmalarında; Fayyad ve diğ. [5]' nin KDD için yapmış olduğu tanımı veri madenciliğini tanımlamak için kullanmıştır. Veri madenciliği, veri tabanlarında bilgi keşfi sürecinin bir parçası olmasına rağmen tüm sürecin anlamlı sonuçlara varması açısından belki en önemli aşamasıdır. Elbette diğer süreçlerin analizler açısından önemi göz ardı edilemez ancak KDD içinde en fazla öne çıkan aşama veri madenciliğidir. Bu sebepten dolayı KDD için Fayyad ve diğ. [5] tarafından yapılan tanımlamanın, zaman içinde farklı araştırmacılar tarafından veri madenciliğinin tanımlanması için kullanmaya başlanması normal karşılanmaktadır.

Veri madenciliğinin genel bir tanımı yapılacak olursa, büyük hacim ve farklı tipteki veriden anlamlı bir takım desenler/örüntüler oluşturmayı, veri arasındaki ilişkileri keşfetmeyi ve veriden bilgi elde etmeyi amaçlayan bir alandır. Tıpkı bir maden keşfi yapar gibi veri yığınları arasından da "değerli, anlamlı" olan bilginin bulunmasını sağlar.

Özkan [12] veri madenciliğini "*Büyük ölçekli veriler arasından değeri olan bir bilgiyi elde etme işidir.*" şeklinde tanımlamıştır.

Genel geçer bir tanımı olmamakla beraber veri madenciliği için birbirine benzer birçok tanım yapılmaktadır:

"*Veri madenciliği büyük hacimli veriler arasından bilgiyi çekip çıkarmaktır. Başka bir deyişle veri madenciliği, veri tabanlarındaki büyük ve karmaşık veriler arasındaki bilinmeyen örüntüleri, değerli yapıları ve ilginç ilişkilerin keşfedilmesi bilimidir.*" [13]

"*Veri madenciliğindeki esas amaç veri setinden bilgi elde etmek ve bunu anlaşılabilir bir yapı halinde kullanılmak üzere ortaya koymaktır.*" [11]

"*Veri madenciliğinin amacı, geçmiş faaliyetlerin analizini temel alarak gelecekteki davranışların tahminine yönelik karar-verme modelleri yaratmaktır.*" [14]

Özetlemek gerekirse farklı kaynaklarda yapılan farklı tanımlar aslında birbirine çok yakın noktalara parmak basmaktadırlar. Veri madenciliği de istatistik gibi çeşitli analizler yoluyla veriden öğrenmeye ve bilgi elde etmeye yarayan bir alandır. Esas olarak büyük boyutlu veri tabanlarındaki karmaşık yapılı verinin arasındaki örüntülerin bulunmasını, anlamlı ve kullanışlı olacak bilgilerin keşfedilmesini, elde edilen bilginin anlaşılabilir bir şekilde ortaya konmasını sağlar. Bunun yanında veri madenciliği ile istatistiksel analizlerde olduğu gibi geleceğe dönük tahminler yapmak da mümkündür [1], [4], [5], [8], [12], [13], [15], [16].

Veri madenciliği birçok farklı araştırma alanında kullanılmaktadır (Tablo 1).

Tablo 1. Veri madenciliğinin kullanım alanları
(Data mining application areas)

Benzerlikler	Detaylar	
	Uygulama	*Kaynak*
Sağlık / Tıp	Hastalıkların tahmini ve analizi Hastalıkların etkilerinin analizi İlaç yan etkilerinin tespiti Hastane yönetimi	[14], [17], [18]
Biyoenformatik	Biyolojik veri analizi	[19]
Astronomi	Gök cisimlerinin analizi/sınıflandırılması	[4]
	İklim modellemeleri	[9]
Telekomünikasyon	Müşteri kayıp analizi Müşteriye özel kampanya analizi	[6]
Üretim	Hata tespiti ve tahmini	[4]
Yatırım	Portfolyo yönetimi	[4]
Pazarlama	Müşteri davranış tahminleri Uygun kampanya geliştirme Farklı müşteri gruplarının analizi Pazar sepet analizi	[4], [15]
Bankacılık	Dolandırıcılık tespiti Para aklama işlemlerinin tespiti Müşteri gruplarının belirlenmesi	[4], [15]
Sigortacılık	Dolandırıcılık tespiti Müşteri tahmini	[15]
Yapısal olmayan veri analizi	Sosyal medya verisi Web sitelerinin analizi	[20]
Eğitim	Başarı faktörlerinin belirlenmesi	[6], [21]

3. YÖNTEM (METHOD)

Bu çalışmanın oluşturulması için literatür taraması yapılmıştır. 1996 ve 2015 yılları arasında yayımlanan konferans bildirisi, makale ve kitap olmak üzere, Türkçe, İngilizce ve Almanca dillerinde yazılmış toplam 24 kaynak incelenmiştir. İnternet üzerinden Google arama motorunda yapılan taramalarda "veri madenciliği", "data mining", "istatistik", "veri madenciliği ve istatistik", "veri madenciliği ve istatistiğin farkları", "differences between data mining and statistics", "unterschiede zwischen statistik und data mining" anahtar kelimeleri kullanılmış ve çıkan sonuçlar taranarak aralarından çalışma için uygun bulunan 24 kaynak incelenmiştir.

4. BULGULAR (FINDINGS)

4.1. İstatistik ve Veri Madenciliği Arasındaki Benzerlikler (Similarities Between Data Mining and Statistics)

İstatistik ve veri madenciliği kavramsal olarak farklı alanlar olsa da ikisinin ortak özellikleri vardır (Tablo 2). Her iki alan da veriden öğrenme, verinin bilgiye dönüştürülmesi, veriyi analiz etme, verinin anlamını çözme, belirsizlikleri ortadan kaldırma, olayı etkileyen faktörleri belirleme, ön görüde bulunma amaçlarını yerine getirirler [1], [13], [16], [22]. Genel amaçlarına bakıldığında her iki alanın da odaklandığı nokta veriden öğrenme ya da veriyi bilgiyi dönüştürmedir. Bu açıdan, her ikisinin de edindiği misyon birbirine çok yakındır.

Tablo 2. İstatistik ve veri madenciliğinin benzerlikleri
(Similarities between data mining and statistics)

Benzerlikler
Veriden öğrenme / verinin bilgiye dönüştürülmesi
Veriyi analiz etme / verinin anlamını çözme
Öngörüde bulunma / tahmin yapma
Belirsizlikleri ortadan kaldırma
Olayı etkileyen faktörleri belirleme
Veri ön işleme
Kullanılan analiz çeşitleri

[1], [9], [13], [15], [16], [22]

Veri madenciliği yöntemlerinin birçoğu temelinde istatistiksel yöntemlere dayanır. Başka bir deyişle veri madenciliğinin temelini istatistik oluşturur ve istatistik bilimi olmadan veri madenciliğinden söz etmek mümkün olamaz [9]. Bazı araştırmacılara göre istatistiksel bakış açısıyla bakıldığında veri madenciliği yöntemleri daha esnek yöntemler olarak tanımlanmış ve veri madenciliği istatistikteki çok değişkenli analizlerle eş değer tutulmuştur [1].

Kullanılan analizler incelendiğinde; kümeleme, diskriminant, regresyon, korelasyon analizlerinin her iki alanda da kullanılan ortak yöntemler olduğu görülmektedir [16]. Aslında veri madenciliği istatistikte var olan bu analizleri kullanmakta ve bu açıdan bakıldığında veri madenciliği, yukarıda da bahsedildiği gibi çok değişkenli istatistiksel analizler olarak değerlendirilebilmektedir. Yine veri madenciliğinde sıklıkla kullanılan karar ağaçları veya birliktelik analizleri de çok değişkenli istatistik analizleri ile eş tutulmaktadır [13].

Her iki alan için kullanılan yöntemler dışında sahip olunan bir diğer benzerlik ise veri ön işleme aşamasıdır. Ön işleme aşaması, verinin analize girmeden önce işlenmeye uygun hale getirilmesini sağlarken, analizin doğruluğu ve sonuçların kalitesi için de önemli rol oynamaktadır. Analiz sonuçlarının doğruluğu için analize giren verinin de temiz olması gerekir [16]. Akpınar [15] ön işlem aşamasının veri tabanlarında bilgi keşfi sürecinin %50 ile %80' ini oluşturabileceğini belirtmiştir.

4.2. İstatistik ve Veri Madenciliği Arasındaki Farklılıklar (Differences Between Data Mining and Statistics)

İstatistik ve veri madenciliği arasında esas amaçları birbirine çok yakın olsa da önemli farklılıklar olduğu görülmektedir (Tablo 3).

Tablo 3. İstatistik ve veri madenciliğinin farklılıkları
(Differences between data mining and statistics)

Farklılıklar	Detaylar	
	İstatistik	*Veri Madenciliği*
Kavramsal [22]	Kökeni eskiye dayanan, başlı başına bir bilim dalıdır	İstatistiğin alt dalı değildir.
Veri seti büyüklüğü [12], [13], [16]	Görece daha küçük boyutlu veri	Milyonlarca, milyarlarca veri, çok fazla değişken
Örneklem büyüklüğü [16]	Veri setinden seçilen bir küme	Veri setinin tamamı
Verinin toplanış amacı [13], [16]	Belli bir amaç için toplanır	Birincil amaç veri madenciliği uygulamak değildir. (genellikle)
Hipotezin varlığı [16], [23], [24]	Hipotez var.	Hipotez yok/olmayabilir.
Benimsenen yaklaşım [16]	Tümevarım	Tümdengelim
Bilgisayar kullanımı [22]	Bilgisayarsız analiz mümkündür.	Bilgisayarsız düşünülemez.

Kavramsal açıdan her iki kavramın birbirinden bağımsız olduğunu bilmek önemlidir. Veri madenciliği ve istatistik birbirlerinden farklı alanlardır ve veri madenciliği istatistiğin bir alt dalı ya da parçası değildir. Temelinde istatistiğe dayanan birçok veri madenciliği yöntemi / algoritması vardır ancak bunlar sebebiyle veri madenciliğini istatistiğin bir alt dalı olarak görmek yanlıştır [22].

Analizlerde kullanılan veri setlerinin büyüklüğü açısından bakıldığında da her iki alan farklılık göstermektedir. Geleneksel olarak kullanılan istatistiksel yöntemler çok büyük hacimli veri setleri, büyük veri gibi kavramlar ile karşılaştığında, verinin büyüklüğü ve çeşitliliği açısından yetersiz kalabilmektedir. Bu sebeple veri madenciliği yöntemleri öne çıkmakta ve büyük veri setlerinin analizinde kullanılmaktadır. Veri madenciliği yöntemleri için kolaylıkla analiz edilebilecek veri setleri istatistiksel yöntemlerle analiz edilemeyecek büyüklükte olabilir [13].

"*Bir istatistikçi için büyük veri kümesi birkaç yüz veya bin veri kümesi içerir. Veri madenciliği ile uğraşanlar için, milyon veya milyar veri beklenmeyen bir durum değildir.*" [16]

"*Veri madenciliğinde ise milyonlarca ve hatta milyarlarca veri ve çok fazla değişken ile ilgilenilir.*" [12]

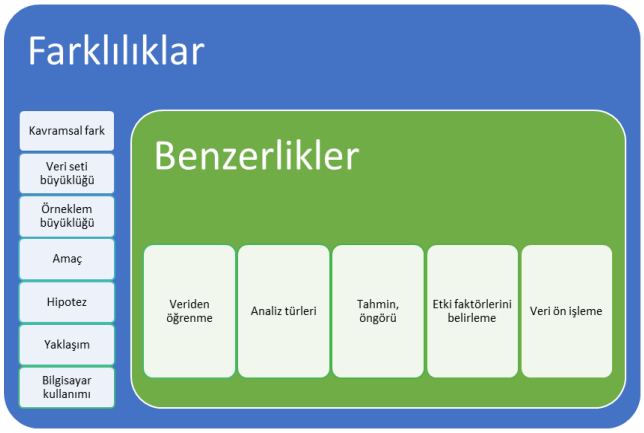
Veri setinin büyüklüğü gibi örneklemin büyüklüğü de farklıdır [16]. Örneklem, istatistiksel analizlerde ana kütleyi temsil eden daha ufak bir kütledir. Veri madenciliğinde ise örneklem veri setinden seçilen daha ufak bir veri seti değil çoğunlukla eldeki veri setinin tamamıdır. Bu sebeple veri madenciliği analizleri bilgisayar kullanımı olmadan düşünülemez [22]. Çünkü eldeki veri yığını farklı tipte veri içermekle beraber çok fazla sayıda veriyi içerir ve veri madenciliği algoritmalarını, karmaşık ve büyük hacimli veri yığınlarına bilgisayar olmadan uygulamak neredeyse imkansızdır. Oysa istatistiksel analizlerde veri seti veri madenciliğindekine oranla daha küçük olduğundan bilgisayar kullanmadan da analizler yürütülebilir. Bu duruma başka bir açıdan bakılacak olursa istatistiğin kökenleri veri madenciliğinden ve bilgisayarların gelişiminden daha eskilere dayandığından, zaten istatistiksel analizler uzunca bir süre bilgisayar kullanımı olmadan yapılmıştır [22].

Analizi yapılacak verinin toplanış biçimi veya amacı açısından bakıldığında da veri madenciliği ve istatistiğin yaklaşımı farklılık göstermektedir. İstatistik çalışmaları kapsamındaki veri belirli bir amaç için veya sorudan yola çıkılarak toplanır [13], [16]. Bu duruma bağlı olarak başka bir fark da göz önüne alınmalıdır; o da hipotezin varlığıdır.

Önceden belirlenmiş bir soruyla yola çıkılan istatistiksel araştırmalarda, analiz ve sonuç aşamalarından önce kurulan bir hipotez mevcuttur. Veri madenciliğinde ise analizden önce tanımlanmış bir hipotezin varlığından söz edilemez [23]. Hipotezden yola çıkıldığı için istatistiksel analizlerde apriori bilgiden yola çıkılarak analize başlandığı söylenebilir. Apriori bilgi deneyden veya herhangi bir deneyimden bağımsız olarak var olan bilgi olarak adlandırılabilir [25]. Bu şekilde yola çıkılarak analiz sonucunda ortaya konan hipotezin doğruluğu veya yanlışlığı ispatlanmış olur. Veri madenciliği analizlerinin ise böyle bir iddiası yoktur. Çünkü analize başlamadan önce herhangi bir hipotez belirlenmez. Aksine analiz sonucunda elde edilen bulgular incelenir ve yorumlanır. Veri madenciliği ve istatistikte bu sebeplerden ötürü hipotezin rolü farklılık gösterir [16], [24].

İki alan arasındaki diğer fark ise analiz yapılırken benimsenen yaklaşımdır. Tümevarım ya da tümdengelim yaklaşımlarından hangisinin benimsendiği farklı kaynaklarda farklı şekillerde açıklanmıştır. Tüzüntürk [16], veri madenciliği yöntemlerinin kullanıldığı analizlerde tümdengelim, istatistiksel analizlerde tümevarım yaklaşımlarının benimsendiğinden bahsetmektedir. Veri madenciliği analizlerinde veri setinin tamamının analiz edilmesi ile daha özel/yerel bilgiye ulaşılır. Bu durumda genelden özele doğru bir bilgi keşfi olduğu yani tümdengelim yaklaşımıyla ilerlendiği söylenmektedir. İstatistiksel analizlerde ise tam tersine, tümevarım yaklaşımı ile hareket edilir çünkü ana kütle içinden seçilen örneklemin analizinden yola çıkılarak elde edilen sonuçlar tüm ana kütle/veri seti için genellenir. Bu yaklaşımda özel olandan genel olana doğru bir yönelme vardır. Ganesh [13] ise veri madenciliğinin genellikle asıl toplanma amacından bağımsız olan ikincil analizler yapmada kullanıldığını belirtmiştir. Bu sebeple tümevarım yaklaşımıyla hareket edildiğini söylemektedir. İstatistiksel analizler ise ona göre hipotezden yola çıkılarak tümdengelim yöntemiyle ilerlemektedir. Bu çalışma kapsamında; veri madenciliğinin tümdengelim, istatistiksel yöntemlerin tümevarım yöntemlerine uygun olarak yapıldığı görüşü hakim olmaktadır.

Çalışma kapsamında elde edilen benzerlik ve farklılık Şekil 3'te özetlenerek sunulmuştur:



Şekil 3. Farklılıklar ve benzerlikler
(Differences and similarities)

5. TARTIŞMA VE SONUÇ (DISCUSSION AND CONCLUSION)

Günümüzdeki teknolojik gelişmeler ve veri boyutlarındaki artış, istatistiksel yöntemlerin tek başına yetersiz kalabilmesine sebep olmaktadır. Bu gelişmeler sonucu ortaya çıkan veri madenciliği kavramı büyük veri boyutlarında analiz yapılmasını sağlamaktadır. Bu çalışma kapsamında veri madenciliği ve istatistik alanlarının birbirleriyle nasıl bir ilişki içerisinde olduğu ortaya konmaya çalışılmıştır. Çalışmanın, her iki alan arasındaki ilişki, benzerlik ve farklılıklar hakkında genel bir çerçeve çizerek araştırmacılara yardımcı olması umulmaktadır.

Veri madenciliği kavramı ayrı bir alan olsa da temeli istatistiksel yöntemlere dayanmaktadır. Veriden öğrenme, anlamlı bilgi elde etme, öz bilgi keşfi, bilgi keşfi, örüntüleri keşfetme, değerli bilgiye ulaşma, değişkenler arasındaki ilişkileri keşfetme gibi farklı şekillerde adlandırılabilecek ve analizlere dayalı işlemler, istatistik veya veri madenciliği yöntemleri kullanarak yapılabilmektedir. Bu noktada eldeki veri ile nasıl bir analiz yapılacağının belirlenmesinde analizi yapacak kişinin belirlemesi gerekecektir. Hangi yöntemlerin kullanılacağını belirlerken veri kümesinin büyüklüğünü ve yapısını bilmek; analiz yöntemleri arasındaki farklılıklara hakim olmak önemlidir. Veri madenciliği ve istatistik arasında farklılıklar olmasına rağmen iki alan birbirlerinden bağımsız düşünülemeyecek alanlardır. Çünkü her iki alan da "veriden öğrenme" ve "veriden bilgi elde etme" amaçları için kullanılırlar. Veri madenciliğindeki gelişmelerin istatistiksel yöntemlerin kullanımına, istatistiksel yöntemlerdeki gelişmelerin ise veri madenciliğinin gelişimine katkı sağlayacağı ve bu iki alanın birbirine doğru bir büyüme içinde olduğu düşünülmektedir [1].

İstatistiksel analizler halen veri analizinde kullanılmakta olan analiz yöntemleridir. Ancak teknolojinin gelişmesi, çok çeşitli veri kaynaklarının oluşması, veri üretiminin artması ve veri türlerinin çeşitliliği analiz yöntemlerinin de gelişmesini ve çeşitlenmesini sağlamaktadır. Buna paralel olarak araştırma alanları arasındaki keskin sınırlar ortadan kalkmakta ve her farklı alan birbirinden beslenmektedir. Veri madenciliğinin de beslendiği temel alanlardan biri istatistiktir. Temellerini istatistikten alan veri madenciliği yöntemleri analizlerde istatistiksel yöntemlerin sınırlılıklarını ortadan kaldırmakta ve günümüzün veri bolluğunda her alandan araştırmacılara daha geniş kapsamlı ve esnek analizler yapma imkanı sunmaktadır. "*Günün gereksinimlerine göre kendini güncelleyerek gelişimini sürdüren veri madenciliği*" [10] her geçen gün farklı araştırma alanlarında araştırmacılar tarafından ilgi görmektedir. Tıp gibi, araştırmacıların istatistiksel yöntemlere bağlı kaldığı alanlarda ilerleyen zamanlarda veri madenciliği yöntemlerinin daha fazla ilgi göreceği düşünülmektedir. Elektronik ortama geçildikten sonraki süreç düşünüldüğünde veri tabanlarında her geçen gün hastalara ait daha fazla veri birikmektedir. Bu verinin analiz edilmesi yoluyla, hastalıklar hakkında öngörüde bulunulabilmesi, çeşitli belirtilere ya da özelliklere göre hastalıkların tahmin edilmesi, hastalıklara yönelik önleyici tedbirlerin alınabilmesi gibi araştırma başlıklarında veri madenciliği yöntemlerinin ileride kullanılacağı öngörülmektedir.

Bu çalışmanın ilerleyen aşamalarında aynı veri seti üzerinde hem istatistiksel yöntemlerin hem de veri madenciliği yöntemlerinin uygulamasının yapılması ve sonuçların/yöntemlerin karşılaştırılması planlanmaktadır.

KAYNAKLAR (REFERENCES)

[1] D. Kuonen, "Data Mining and Statistics: What is the Connection?," The Data Administration Newletter, 2004.

[2] M. Davidian and T. A. Louis, "Why Statistics?," Science, 336, 12, 2012.

[3] İnterner: J. Aldrich, "Figures from the History of Probability & Statistics," Ekim 2012. http://www.economics.soton.ac.uk/staff/aldrich/Figures.htm, 31.05.2016.

[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Mag., 17 (2), 37, 1996.

[5] U. Fayyad, Piatetsky-Shapiro, and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," Communıcatıons Of The ACM, 39(11), 27–34, 1996.

[6] S. Savaş, N. Topaloğlu, and M. Yılmaz, "Veri madenciliği ve Türkiye'deki Uygulama Örnekleri," İstanbul. Ticaret Üniversitesi Fen Bilimleri Dergisi, 11(21), 1–23, 2012.

[7] İnternet: YÖK, "Ulusal Tez Merkezi," 2016. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp, 29.06.2016.

[8] M. E. Balaban and E. Kartal, Veri Madenciliği ve Makine Öğrenmesi, 1. baskı, Çağlayan Kitapevi, İstanbul, 2015.

[9] D. Kumar and D. Bhardwaj, "Rise of Data Mining: Current and Future Application Areas," International Journal of

Computer Science Issues, 8(5), 256–260, 2011.

[10] H. Akpınar, Data: Veri Madenciliği Veri Analizi, 1.baskı, Papatya Yayıncılık Eğitim, İstanbul, 2014.

[11] A. Kapoor, "Data Mining: Past, Present and Future Scenario," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 3(1), 95–99, 2014.

[12] Y. Özkan, Veri Madenciliği Yöntemleri, 1.baskı, Papatya Yayıncılık Eğitim, İstanbul, 2008.

[13] S. Ganesh, "Data mining: Should it be included in the statistics curriculum?", The 6th International Conference on Teaching Statistics (ICOTS 6), Cape Town, Güney Afrika, 2002.

[14] A. S. Koyuncugil and N. Özgülbaş, "Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları," Bilişim Teknolojileri Dergisi, 2(2), Mayıs 2009.

[15] H. Akpınar, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği," İÜ İşletme Fakültesi Dergisi, 29(1), 1–22, 2000.

[16] S. Tüzüntürk, "Veri Madenciliği ve İstatistik," Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,29(1), 65-90, 2010.

[17] S. Güllüoğlu, "Tıp ve Sağlık Hizmetlerinde Veri Madenciliği Çalışmaları: Kanser Teşhisine Yönelik Bir Ön Çalışma," AJIT-E AJIT-e: Online Academic Journal of Information Technology, 2(5), 2011.

[18] E. Kaya, M. Bulun, and A. Arslan, "Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları,", Akademik Bilişim, Çukurova Üniversitesi, Adana, 3-5 Şubat, 2003.

[19] Ç. Erol and Y. Özkan, "Temel Biyoloji ve Mikrodizi," in Biyoenformatik DNA Mikrodizi Veri Madenciliği, 2.baskı, Papatya Yayıncılık Eğitim, İstanbul, 51–81, 2015.

[20] M. Ö. Dolgun, T. Güzel Özdemir, and D. Oğuz, "Veri Madenciliğinde Yapısal Olmayan Verinin Analizi: Metin ve Web Madenciliği," İstatistikçiler Dergisi, 2, 48–58, 2009.

[21] Ç. Kurt and O. A. Erdem, "Öğrenci Başarısını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle İncelenmesi," Politeknik. Dergisi., 15(2), 111-116, 2012.

[22] D. Meschenmoser, "Data Mining und Statistik: Gemeinsamkeiten und Unterschiede," Şubat 2004.

[23] H. Mannila, "Data Mining: Machine Learning, Statistics and Databases," SSDBM '96 Proceedings of the Eighth International Conference on Scientific and Statistical Database Management, Stockholm, 2–9, 1996.

[24] C.-M. Zhao and J. L. Vice, "Data mining: Going beyond traditional statistics," New Directions for Institutional Research., 131, 7–16, Eylül 2006.

[25] İnternet: J. S. Baehr, "A Priori and A Posteriori", http://www.iep.utm.edu/apriori/, 22.06.2016.