Uludağ Üniversitesi
İktisadi ve İdari Bilimler Fakültesi Dergisi
Cilt XXIX, Sayı 1, 2010, s. 65-90

VERİ MADENCİLİĞİ VE İSTATİSTİK

Selim TÜZÜNTÜRK*

Özet

Bu çalışmanın konusu veri madenciliği ve istatistik ile ilgilidir. Bu çalışmanın amacı veri madenciliği ve istatistik arasındaki ilişkinin açıkça ortaya konmasıdır. Bu çalışma ile veri madenciliği süreci sunulmuştur, böylece veri madenciliğinin istatistik ile ayrı tutulamaz olduğu vurgulanmıştır. Teorik veri madenciliği açıklamalarına ilave olarak, OLAP küpleri ile küçük bir veri madenciliği uygulaması da yapılmıştır.

Anahtar Kelimeler: *Veri Madenciliği, İstatistik, OLAP Küpleri.*

Abstract

The subject of this study is about data mining and statistics. The aim of this study is to state clearly the relationship between data mining and statistics. By this study the process of data mining is presented, thus it is stressed that data mining can not be excluded from statistics. In addition to theoretic data mining explanations, a small data mining application is also applied with OLAP cubes.

Key Words: *Data Mining, Statistics, OLAP Cubes.*

1.GİRİŞ

Bilgisayara işlenebilen herhangi bir sayı veya metin **veri** olarak tanımlanır[1]. **Bilgi** (information) verinin düzenlenmiş şeklidir (Gürsakal, 2001: 48). Veriler bir araya getirilerek düzenlendiğinde, bilgiye dönüşürler (Oğuzlar ve Tüzüntürk, 2008). Veriler arasındaki düzenler, çağrışımlar veya ilişkiler bilgi sağlayabilir. Örneğin: bir firmanın satış işlemleri ile ilgili

* *Arş. Gör., Uludağ Üniversitesi, İİBF, Ekonometri Bölümü.*

1 İstatistiksel olarak veri yapıları için bkz. Gürsakal (2001), Gürsakal (2007), Oğuzlar (2007).

veriler, hangi ürünlerin satıldığı ve bu ürünlerin ne zaman satıldığı hakkında bilgi sağlayabilir. Bilginin daha öz bir şekle dönüştürülmesi ile elde edilen bir türü daha vardır. Gürsakal (2001) bu daha öz şekli **öz bilgi** (knowledge) olarak tanımlamaktadır. Tarihi düzenler (örüntüler) ve gelecek eğilimler için bilgi öz bilgiye dönüştürülebilir. Örneğin: Süpermarket satışlarına ilişkin özet bilgi tüketicilerin satın alma davranışına ilişkin öz bilginin sağlanması için promosyona ait çabalar ışığında analiz edilebilir. Özetle, öz bilgi bilginin hacim olarak iyice küçülmüş ancak kullanım değeri çok artmış bir türüdür (Oğuzlar, 2004: 4).

Günümüz gelişmiş toplumlarında verinin, bilginin yönetiminde daha çok öz bilgi ile ilgilenilmektedir (Gürsakal, 2001: 48). Bilgi teknolojilerinin gelişimi ve gündelik hayatın her aşamasında kullanılabilir hale gelmesiyle beraber, her alanda oldukça büyük miktarda veri birikmeye başlamıştır. Böylece, banka, üniversite, okul, seyahat şirketi, hastane, devlet dairesi vb. kuruluşların çalışıp işleyebilmesi için kayıt altında tutmak durumunda olduğu çeşitli veriler veritabanlarında[2] depolanmıştır. Verilerin hafızadaki durumları, veritabanı yaratmak ve yönetmek, kullanıcıların erişimleri, verilerin yönetilmesi, yedeklerin alınması gibi işlemleri düzenleyen sistemlerin (veritabanı yönetim sistemlerin) artan kullanımı ve hacimlerindeki olağanüstü artış, kuruluşların elde toplanan bu verilerden nasıl faydalanabileceği problemi ile karşı karşıya bırakmıştır. Geleneksel sorgu ve raporlama araçlarının veri yığınları karşısında yetersiz kalması **Veri Tabanlarında Öz Bilgi Keşfi** (Knowledge Discovery in Databases) adı altında, sürekli ve yeni arayışlara yöneltmiştir. Veri tabanlarında öz bilgi keşfi süreci içerisinde, model kurulması ve değerlendirilmesi aşamalarından meydana gelen **Veri Madenciliği** (Data Mining) en önemli kesimi oluşturmaktadır (Akpınar, 2000: 1). Veri madenciliği Veri Tabanlarında Öz bilgi Keşfi'nin özünü oluşturan keşif kısmının gerçekleştiği adım olarak alınabileceği gibi bağımsız süreç olarak da değerlendirilmektedir (Koyuncugil, 2007: 1).

Bu çalışmanın konusu veri madenciliğinin istatistik ile olan ilişkisidir. Bu çalışmanın temel amacı öz bilgi keşfi sürecinde önemli bir yere sahip olan veri madenciliğinin tanımlanması ve istatistik bilimi ile olan ilişkisinin ortaya konulmasıdır. Bu çalışmada veri madenciliği uygulaması yapacak araştırmacılara faydası olacağı düşüncesi ile veri madenciliği istatistik ekseninde genel bir çerçeve sunulma çabası içinde olunmuştur. Çalışmanın ikinci bölümünde veri madenciliği tanımı, süreci, aşamaları, uygulama alanları, yerli ve yabancı yakın dönem literatürü ve istatistik ile olan ilişkisi üzerinde durulmuştur. Üçüncü bölümde veri madenciliği

[2] Veri tabanı, bir kuruluşun uygulama programlarının kullanıldığı veriler bütünüdür (Şentürk, 2006: 4).

problem tiplerine değinilmiştir. Dördüncü bölümde veri madenciliği teknikleri ifade edilmiştir. Beşinci bölümde küçük bir uygulama yapılmıştır. Son bölümde sonuç ve genel değerlendirmeler yer almaktadır.

2. VERİ MADENCİLİĞİ

Genel olarak, veri madenciliği[3] verilerin farklı bir bakış açısından analiz edilmesi ve kullanışlı bilgi halinde özetlenme sürecidir. Teknik olarak veri madenciliği, büyük ve birbiriyle ilişkili veri tabanları içinde düzinelerce alan arasında korelasyonlar ve düzenler bulma sürecidir. Veri madenciliği üzerinde fikir birliğine varılmış ortak bir tanım yoktur. Veri madenciliği ile ilgili bazı tanımlar şöyledir[4]:

- Veri madenciliği büyük hacimli verilerden öz bilgi'nin çıkarılması sürecidir (Ganesh, 2002:1). Bir başka ifade ile veri madenciliği büyük ve karmaşık verilerde beklenmeyen patikaların, değerli yapıların ve ilginç ilişkilerin keşfedilmesi bilimidir.
- Veri madenciliği büyük veri tabanlarındaki gizli bilgi ve yapıyı açıklamak için, çok sayıda veri analizi aracını kullanan bir süreçtir (Oğuzlar, 2004: 4).
- Kuonen (2004) veri madenciliğini iş kararlarının alınabileceği doğru, alışılmamış, faydalı ve anlaşılabilir örüntüler veya modeller olarak tanımlamaktadır.
- Bilgisayar teknolojilerinin sağlamış olduğu çok hızlı veri işleme ve yüksek hacimde veri depolama imkânları yardımıyla ve farklı disiplinlerin katkısıyla sağlanan araçlarla, sahip olunan çok büyük hacimlerdeki veriden, karar vericinin etkin ve daha fazla bilgiye dayalı karar vermesinde kullanabilmesi amacıyla önceden bilinmeyen, gizli, örtük, klasik metotlarla ortaya çıkarılması güç, faydalı, ilginç, anlaşılabilir; ilişki, örüntü, bağıntı veya trendlerin otomatik veya yarı otomatik bir şekilde ortaya çıkarılması olarak tanımlanır (Şentürk, 2006: 4).
- Veri madenciliği genel anlamda, büyük miktarda veri içerisinden, gizli kalmış, değerli, kullanılabilir bilgilerin açığa çıkarılmasıdır (Koyuncugil, 2007: 1).

---

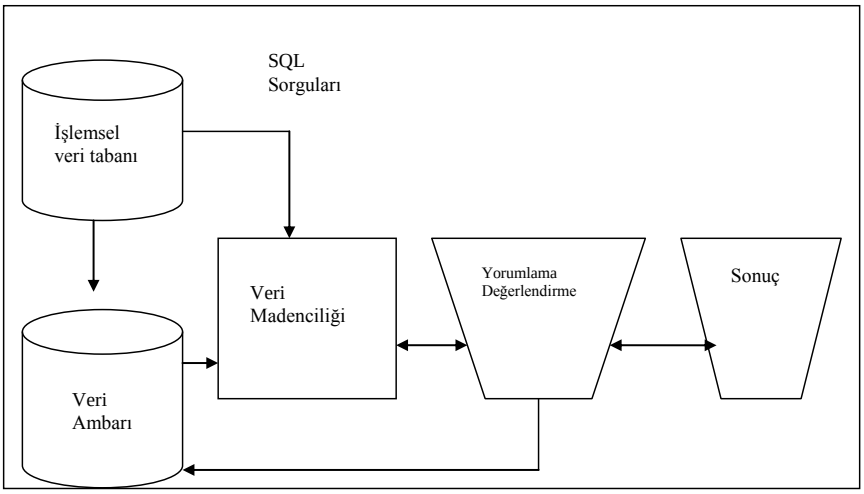
[3] Veri madenciliği bazen veri veya öz bilgi keşfi olarak ta adlandırılmaktadır.

[4] Diğer başka tanımlar için bkz. Friedman (1997).

- Veri madenciliği ve öz bilgi keşfi, verilerde daha önceden bilinmeyen, anlamlı ve değerli bilgiler elde etme işlemidir (Yıldırım, Uludağ ve Görür, 2007).

2.1. Veri Madenciliği Süreci ve Aşamaları

Şekil 1 veri madenciliği sürecini görsel olarak özetlemektedir. Burada işlemsel veri tabanı, veri ambarı, veri madenciliği, yorumlama değerlendirme ve sonuç arasında çeşitli bağlar bulunmaktadır.



Kaynak: Oğuzlar, 2004, s. 5.

Şekil 1.
Basit Bir Veri Madenciliği Süreci

Veri tabanlarında öz bilgi keşfi sürecinde sırasıyla 5 temel aşama izlenmektedir (Akpınar, 2000: 6):

1. Problemin tanımlanması:

Uygulamanın hangi işletme amacı için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili işletme amacı işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalıdır.

2. Verilerin hazırlanması (Ön İşleme**)**

Her veri analizi işi yeni bir veri setlerinin toplanması, betimlenmesi ve temizlenmesiyle başlar. Bu süreçten sonra, veriler analiz edilebilir ve sonuçlara ulaşılır (Dasu ve Johnson, 2003:1). Veri kalitesi veri madenciliğinde anahtar bir konudur. Veri madenciliğinde güvenilirliğin arttırılması için, veri ön işleme yapılmalıdır (Oğuzlar, 2003: 70). Verilerin hazırlanması aşaması şu adımlardan oluşmaktadır:

- Verilerin Toplanması

Tanımlanan problem için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır (Akpınar, 2000: 6).

- Verilerin Temizlenmesi

Eksik verilerin tamamlanması, aykırı değerlerin teşhis edilmesi amacıyla gürültünün[5] düzeltilmesi ve verilerdeki tutarsızlıkların giderilmesi gibi işlemlerden oluşmaktadır.

- Verilerin Birleştirilmesi

Bu aşamada farklı veri tabanlarındaki verilerin tek bir çatı altında (veri ambarında) birleştirilmesi işlemidir. Veriye sahip olma, verinin işlenmesi, iletilmesi ve depolama kapasitesindeki ilerlemeler işletmelerin sahip oldukları çeşitli veri tabanlarının veri ambarlarında birleştirilmesine olanak tanımıştır. En genel anlamında, veri ambarı farklı kaynaklarda tutulan verilerin ortak bir çatı altında birleştirilerek, verilerin zaman boyutunda birbiri ile konuşmasını sağlayan, tutarlı ve doğru verilerin yer aldığı sistemdir (Şentürk, 2006: 10)[6].

- Verilerin Dönüştürülmesi

Dönüştürme işlemi, verilerin veri madenciliği için uygun formlara dönüştürülmesidir. Veri dönüştürme, düzeltme, birleştirme, genelleştirme ve normalleştirme gibi işlemlerin bir veya bir kaçını içerir.

- Verilerin İndirgenmesi

Büyük hacimli veri kümesinden daha küçük hacimli veri kümesinin elde edilmesidir.

3. Modelin Kurulması ve Değerlendirilmesi (Veri Madenciliği)

Tanımlanan problem için en uygun modelin bulunabilmesi için, çok sayıda modelin denenmesi gerekebilir. Bu nedenle, veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varıncaya kadar yinelenen bir süreçtir.

4. Modelin Kullanılması

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama yapılabileceği gibi, bir başka uygulamanın alt parçası olarak da kullanılır.

---

[5] Veriler genellikle hata içermektedir ve bu hatalar toplu olarak gürültü olarak adlandırılmaktadır (Oğuzlar, 2003: 68).

[6] Data Mart olarak isimlendirilen işleve özel veri tabanlarına veri aktarımı ile de veri madenciliği işlemlerine başlanılması mümkündür (Akpınar, 2000: 1). Data Mart hacim olarak veri ambarlarına göre daha küçük yapıdadır. Data Martlar küçük boyutlu (1–10 GB) bölümsel ambarlardır (Şentürk, 2006: 8).

5. Modelin İzlenmesi

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modelin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirebilir.

2.2. Uygulama Alanları

Veri madenciliğinin amacı, geçmiş faaliyetlerin analizini temel alarak gelecekteki davranışların tahminine yönelik karar-verme modelleri yaratmaktır (Koyuncugil, 2007: 1).

Veri madenciliği müşteri ilişkilerinde başlamıştır (Kuonen, 2004:3). Veri madenciliği organizasyonel hedeflerin başarılmasında çok geniş kullanım alanına sahiptir (Ganesh, 2002:1). Ayrıca, bankacılıkta finansal göstergelere ilişkin gizli ilişkilerin bulunmasında, pazarlamada müşterilerin satın alma örüntülerinin belirlenmesinde ve sigortacılıkta ise riskli müşterilerin örüntülerinin belirlenmesinde veri madenciliği uygulamalarına çok sık rastlanılmaktadır.

Günümüzde, veri madenciliği firmalar tarafından öncelikle müşteri odaklı olarak (finansal, iletişimsel ve pazarlama) kullanılmaktadır. Veri madenciliği firmalara fiyat, üretim planlaması, personel becerileri gibi **iç faktörleri** belirlemelerine olanak tanımaktadır. Ayrıca, ekonomik göstergeler, rekabet ve pazarın yapısı gibi **dış faktörleri** belirlemelerine olanak tanımaktadır. Böylece, firmaların satışları, müşterilerinin tatmini ve şirket karları üzerindeki olumlu ya da olumsuz etkiler belirlenebilmektedir. Sonuçta, öz bilgiyi elde etme ve veriler içindeki detayları görebilme olanağı sağlanmaktadır. Bu çerçevede, veri madenciliği birçok alanda uygulanabilmektedir. Veri madenciliğinin kullanıldığı başlıca alanlar şöyle özetlenebilir (Akpınar, 2000: 1–22; Koyuncugil, 2007: 1–20):

- Sağlık Sektörü
- Telekomünikasyon
- Finans (Bankacılık, Borsa)
- Pazarlama
- Sigortacılık
- Astronomi
- Biyoloji
- Tıp

Türkiye'de 2000'li yıllarda yapılmış çalışmalar incelendiğinde, çok sayıda çalışma yapıldığı görülmektedir. Bu çalışmalardan bazıları kronolojik olarak şöyle sıralanabilir: Akpınar (2000) kredi değerlendirmesi konusunda Chaid Algoritması ile Karar Ağacı Analizi uygulaması, Alpaydın (2000) finans kurumu üzerinde karar ağacı analizi, Özmen (2001) kredi kartı

kampanya yönetimi konusunda diskriminant analizi, lojistik regresyon analizi ve Ki-kare analizi uygulaması, Emel, Taşkın ve Kılıçarslan (2004) çelik üretim sürecinin sinir ağları ile analizi uygulaması, Emel ve Taşkın (2005a) satışların karar ağaçları analizi ile uygulaması, Baykasoğlu (2005) çimento sektöründe yapay sinir ağları ve regresyon analizi uygulaması, Emel ve Taşkın (2005b) pazarlama stratejilerinde birliktelik kuralları analizi uygulaması, Özçakır ve Çamurcu (2007) yiyecek sektöründe birliktelik kuralları analizi uygulaması, Daş, Türkoğlu ve Poyraz (2007) web kayıt dosyalarında ilginç örüntülerin keşfedilmesi, Yıldırım, Uludağ ve Görür (2007) hastane bilgi sistemleri uygulaması, Koyuncugil (2007) sermaye piyasalarında uygulaması, Duru ve Canbay (2007) deprem verilerinin doğrusal regresyon analizi uygulaması, Vahaplar ve İnceoğlu (2008) elektronik ticaret uygulaması, Bozkır, Gök ve Sezer (2008) Üniversite öğrencilerinin internet kullanımı üzerine karar ağaçları, kümeleme ve birliktelik kuralları analizleri uygulamaları, Ata ve Seyrek (2009) imalat firmaları üzerine karar ağaçları ve sinir ağları analizleri uygulaması, Albayrak ve Yılmaz (2009) İMKB verileri üzerine karar ağaçları analizi uygulaması, Özkul ve Pektekin (2009) muhasebe yolsuzluklarının tespiti, Gürbüz, Özbakır ve Yapıcı (2009) havayolu işletmesine ait raporlar ile ilgili uygulama yapmıştır.

Uluslar arası alanda da çok sayıda veri madenciliği uygulaması bulunmaktadır. Son dönemde yapılmış çalışmalardan bazıları şöyledir: Kusiak ve Smith (2007) ürün tasarımında ve üretim sistemleri, Chien ve Chen (2008) beşeri sermaye, Liao, Chen ve Hsu (2009) spor firmaları, Ogwueleka (2009) bankacılık sektöründe müşteri ilişkileri, Zhang, Ma, Zhang ve Wang (2009) elektrik mühendisliği, Kumar ve Uma (2009) öğrenci performansı, Naveh, Sariri ve Zadeh (2009) elektrik santrali jeneratörü, Çiflikli ve Özyirmidokuz (2010) halı üretimi, Gervilla, Cajal, Roca ve Palmer (2010) alkol tüketimi, Liang (2010) otomotiv sektöründe müşteri değeri, Ayesha, Mustafa, Sattar ve Khan (2010) yüksek eğitim sistemi, Rebbapragada, Basu ve Semple (2010) gelir yönetimi, Deshpande, Gogolak ve Smith (2010) ilaç güvenliği, Srinivas ve Harding (2010) mimari, Seng ve Cheng (2010) iş kararı, Abdelmelek, Saidane ve Trabelsi (2010) yağların biyo-çözünürlüğü, Glasgow ve Kaboli (2010) ilaç vakaları ile ilgili çalışmalar yapmıştır.

2.3. Veri Madenciliğinin İstatistik ile Olan İlişkisi

İstatistiksel tekniklerin özellikle verilerin indirgenmesi ve modellenmesi gibi temel veri ön işleme aşamalarında ve çıktıların değerlendirilmesi veya yorumlanmasında faydaları gözlenmektedir.

Veri madenciliği istatistik ile birçok yönden çok yakın ilişki içindedir (Zhao ve Luan 2006: 10). Veri madenciliği ile istatistiğin ortak özelliği "veriden öğrenilmesi"(Ganesh, 2002:1) veya "verinin bilgiye dönüştürülmesi"dir (Kuonen, 2004:5). Her iki yaklaşım da verilerin anlamını çözmek ile ilgilenir. Her iki araç belirsizliklerin üstesinden gelmek ve gelecekteki olaylar hakkında bilgi vermek için bulunmuştur. Veri madenciliği ve istatistiğin her ikisi de bir olayı etkileyen önemli faktörleri belirlemek ve türetilen modeller ile gelecekteki olayları daha iyi öngörmek ile ilgilenmektedir.

Veri madenciliğinin istatistik ile benzer yönlerinin yanında birçok farklı yönleri de mevcuttur. Zhao ve Luan (2006) dört temel farklılıktan bahsetmektedir. Bunlar; teorinin rolü, genellenebilirlik, hipotez testi ve güven düzeyidir. İstatistik teori ile ortak yaşama ilişkisi içindedir.

Teorinin rehberliği olmadan gözlemlerin ve olayların tamamı bunaltıcı olabilir. İstatistiksel analizler apriori bilgiye dayalı teori ile başlar ve teorinin onaylanması veya ret edilmesi hakkında kanıt arar. İstatistik doğası gereği doğrulayıcı bir süreçtir. Öte yandan veri madenciliği teorinin doğrulanması üzerine odaklanmaz. Bu bilgisayarın otomatik olarak örüntüleri bulacağı veya öngörü yapacağı anlamına gelmez. Tam tersi, veri madenciliği analistten açık talimatlar bekler. Bununla birlikte, istatistik ile karşılaştırıldığında veri madenciliği değişkenler arasındaki ilişkiler hakkında daha az varsayımlarla sınırlandırılmıştır.

Veri madenciliği tümden gelim, istatistik tüme varım ile ilgilenir. İstatistiksel araştırmanın ilgilendiği nadiren örneklemin kendisidir. Araştırmacılar çekilen herhangi bir örneklem ile sadece ana kütle hakkında çıkarım yapmak için ilgilenir. İstatistik bireyselliğin üstesinden gelmesi için tasarlanmış değildir. İstatistik özneler arasındaki benzerliklerin bulunması için tasarlanmıştır. Veri madenciliğinin amacı daha detaylı, çok özel ve yerel bilginin toplanmasıdır.

Hipotez testinin veri madenciliği için özel bir anlamı yoktur, çünkü veri madenciliğine bir teori veya hipotez ile veya özel bir ana kütlenin sonuçlarının genelleştirilmesinin vurgulanması ile başlanılmaz. Veri madenciliğinde büyük hacimli veri kullanılmaktadır ve hatta çoğu zaman ana kütle ile çalışılmaktadır. Bu nedenle istatistikte kullanılan anlamlılık düzeyi de çıkarım yapmaktaki ilintisini kaybeder.

Diğer bazı önemli farklılıklar şöyledir:

- İstatistiksel araştırmalarda veriler akıldaki belirli sorular için toplanır ve bu sorulara yanıt bulmak için analiz edilir. İstatistiksel deney tasarımı ve alan araştırması gibi alt disiplinler veri toplamak için en iyi yollarla ilgili ipuçlarını sağlarlar. Veri

madenciliğinde ise veriler, veri madenciliği uygulamak için değil diğer bazı amaçlar için kullanılır (Oğuzlar, 2003: 69).

- Klasik istatistiksel uygulamalar ve veri madenciliği arasındaki en temel farklılık, veri kümesinin büyüklüğüdür (Oğuzlar, 2004: 12). 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.
- Veri madenciliği uygulamalarında veri kümesinin ön işlemlerden geçirilmesi veri kalitesi açısından oldukça önemlidir. Kaliteli veri doğru çıktıların elde edilmesini sağlayacaktır. Ancak, veri madenciliği analizlerinde verilerin ön işlemlerden geçirilmesi veri kümesinin büyük olması sebebiyle çok zaman alıcı bir durumdur. Çeşitli istatistiksel araştırmalar göz önüne alındığında, birçok durumda verilerin ön işlemlere tabi tutulması üzerinde pek durulmadan doğrudan analize geçilmektedir. Bu veri madenciliğini istatistikten ayıran önemli bir özelliktir.

3. VERİ MADENCİLİĞİ PROBLEM TİPLERİ[7]

Veri madenciliğinde çeşitli problem tipleri çözümlenmeye çalışılır. Çözümlenmeye çalışılan bu problem tiplerine bir veya birkaç istatistiksel teknik kullanılarak çözümler üretilir. Kümeleme teknikleri, diskriminant analizi, korelasyon ve regresyon analizi gibi teknikler veri madenciliği problem tiplerinde kullanılan önemli istatistiksel tekniklerden bazılarıdır.

Bir veri madenciliği projesi işletme problemini bir arada çözen farklı problem tiplerinin genellikle bir birleşimini gerektirir. Veri madenciliğinde belli başlı 6 tane problem tipi vardır (http://www.crisp-dm.org/CRISPWP-0800.pdf, erişim tarihi 29.10.2008):

3.1. Veri tanımlama ve özetleme

Veri tanımlama ve özetlemenin amacı verinin basit bir biçimde özelliklerinin az ve öz olarak tanımlanmasıdır. Böylece veri yapısı gözden geçirilir. Veri tanımlama ve özetleme tek başına bir veri madenciliği projesinin amacı olabilir.

Veri madenciliği sürecinin başlangıcında analizin amacı ve verinin niteliği kesin olarak bilinmeyebilir. Basit betimsel istatistikler ve görselleştirme teknikleri ile keşifsel veri analizi yapılır ve bu teknikler verinin niteliğinin anlaşılmasına ve gizli kalmış bilgilerin anlaşılmasına

[7] http://www.crisp-dm.org/CRISPWP-0800.pdf, (Erişim tarihi 29.10.2008).

yardımcı olur. Veri tanımlama ve özetleme genellikle diğer veri madenciliği problemleri ile bir arada gerçekleşir. Özetleme nihai sonuçların sunulmasında önemli bir rol oynar. Diğer veri madenciliği problem tiplerinin sonuçları (örneğin: kavram tanımları veya kestirim modelleri) daha yüksek bir kavramsal düzeyde verinin özetlenmesi olarak düşünülebilir.

Birçok raporlama sistemi, istatistiksel paketler, OLAP (On Line Analytical Processing) veri tanımlama ve özetlemeyi içerir. Geleneksel doküman tabanlı raporlardan farklı olarak OLAP küpleri, OLAP analiz aracı ile oluşturulan farklı yönlerden bakıldığında farklı konulara dikkat çeken özel formatlı ve çok indeksli bilgi kaynaklarıdır[8]. OLAP üzerinde görüş birliğine varılan ortak özellik çok boyutlu bir veri analizi olmasıdır[9]. OLAP Küpleri sahip olunan bilgilerin faydaya dönüştürülmesinde duruma özel hazırlanmış periyodik raporların hazırlanmasında kullanılmaktadır. OLAP Analiz aracı ile sürekli veri girişi yapılan bilgi işletmenin / kurumun işleyişine uygun şekilde toparlanarak, kolay, çabuk, esnek, hiyerarşik ve performansı yüksek analizler yapılması sağlanır ve bilgi işletmenin / kurumun hedefleri doğrultusunda faydaya çevrilir[10]. Bu kaynaklara küp adı verilmesi farklı renklerde yüzleri olan bir küpün farklı kenarlarında bakıldığında farklı görünümlerin ortaya çıkmasıyla karşılaştırılabilir. Bu bağlamda küpün farklı kenarları analizi yapılacak bölge, müşteri ürün gibi alanlara, küpün ilgili yüzünün rengi de analiz yapılan sayısal toplam miktar ve tutarlara benzetilir. OLAP uygulaması sayesinde büyük miktarlarda verinin anlık raporlanması, müşterinin teknik desteğe ihtiyaç duymadan bilgiyi farklı boyutları ile değerlendirmesi, yeni rapor ihtiyaçlarının hızlı bir şekilde karşılanması, farklı ortamlara XML ile kolay veri aktarımı, Excel ve Pivot (özet) tablolar ile uyum, kolayca PDF ve Excel'e aktarım mümkün olur.

Özetle OLAP Küpleri yüksek hacimli veriler üzerinde hızlı ve esnek raporlar alabilmeyi sağlayan analiz servislerinin bir yapısıdır. OLAP Küpleri sayesinde çeşitli karar ağaçları ve karar mekanizmaları oluşturarak bir nevi veri madenciliği yapmak mümkündür[11].

3.2. Bölümleme

Veri madenciliği problem tiplerinden bölümleme verinin merak uyandıran ve anlamlı alt gruplara veya sınıflara ayrılmasını amaçlar. Bir alt grubun bütün üyeleri ortak özellikleri paylaşır. Bölümleme kendi kendine bir

---

8 OLAP Küpleri örneği için bkz. Bayram (2009).

9 http://www.cengaver.net/?q=node/14(Erişim tarihi 29.10.2008).

10 http://www.excel.web.tr/f50/olap-kupleri-ve-excel-t15412.html(Erişim tarihi 29.10.2008).

11 http://www.btakademi.com/egitim/egitimler/?id=16(Erişim tarihi 29.10.2008).

veri madenciliği problem tipi olabilir. Böylece, bölümlerin keşfedilmesi veri madenciliğinin asıl amacı olacaktır. Bununla birlikte, çoğu zaman bölümleme diğer problem tiplerinin çözülmesine yönelik bir adımdır. Böylece, amaç veri hacminin idare edilebilmesinin sağlanması veya analiz edilmesi daha kolay olan homojen veri altkümelerin bulunması olabilir. Genellikle büyük veri setlerinde çeşitli etkiler birbirini örter ve enteresan örüntüleri gizler. Bu durumda, uygun bölümleme işi kolaylaştırır. Bölümlemede kullanılan teknikler şunlardır:

- Kümeleme Teknikleri
- Sinir Ağları
- Görselleştirme

3.3. Kavram Tanımları

Kavram tanımı kavramların veya sınıfların anlaşılabilir bir tanımını amaçlar. Amaç, yüksek kestirim doğruluğu ile modellerin eksiksiz geliştirilmesi değildir, iç yüzünün kavranmasının sağlanmasıdır.

Kavram tanımının bölümleme ve sınıflamanın her ikisi ile yakın bir ilişkisi vardır. Bölümle herhangi bir anlaşılabilir tanımı olmaksızın bir kavrama veya sınıfa ait olan nesnelerin sıralanması ile sonuçlanır. Genellikle, kavram tanımlaması gerçekleştirilmeden önce bölümleme gerçekleştirilir. Bazı teknikler, örneğin kavramsal kümeleme teknikleri, bölümleme ve kavram tanımı aynı anda gerçekleştirir. Kavram tanımlarında kullanılan teknikler şunlardır:

- Kural Koyma Metotları
- Kavramsal Kümeleme

3.4. Sınıflama

Sınıflama farklı sınıflara ait bazı nitelik veya özelliklerle tanımlanan bir nesneler setinin olduğunu varsayar. Amaç, önceden görünmeyen ve etiketlenmemiş nesnelerin doğru sınıf etiketine atayan sınıflama modellerinin (bazen sınıflandırıcı olarak adlandırılır) kurulmasıdır. Sınıflama modelleri genellikle kestirim modellemesinde kullanılmaktadır. Sınıf etiketleri önceden verilebilir, örneğin kullanıcı tarafından veya bölümlemeden türetilerek tanımlanabilir.

Sınıflama geniş bir alanda çeşitli uygulamaların yapıldığı en önemli veri madenciliği problem tiplerinden biridir. Birçok veri madenciliği problemleri sınıflama problemine dönüştürülebilir. Sınıflamanın hemen hemen bütün diğer problem tipleri ile bağlantıları vardır. Sınıflamada kullanılan teknikler şunlardır:

- Diskriminant Analizi
- Kural Koyma Metotları
- Karar Ağacı Öğrenme
- Sinir Ağları
- k En yakın Komşuluk
- Durum bazında muhakeme
- Genetik Algoritmalar

3.5. Kestirim

Geniş bir alanda uygulamaların gerçekleştirildiği bir başka önemli problem tipi kestirimdir. Kestirim sınıflamaya çok benzerdir. Tek farkı kestirimde hedef nitelik (sınıf) niteliksel bir vasıf değil, sürekli bir vasıftır. Kestirimin amacı gözle görülmeyen nesneler için hedef niteliklerin sayısal değerlerinin bulunmasıdır. Literatürde bu problem tipi bazen regresyon olarak adlandırılmaktadır. Eğer kestirim zaman serisi verileri ile alakalı ise o zaman bu çoğu kez öngörü olarak adlandırılmaktadır. Kestirimde kullanılan teknikler şunlardır:

- Regresyon Analizi
- Regresyon Ağaçları
- Sinir Ağları
- k En Yakın Komşuluk
- Box-Jenkins Metotları
- Genetik Algoritmalar

3.6. Bağımlılık Analizi

Bağımlılık analizi veri öğeleri veya oluşumları arasında anlamlı bir bağımlılığı (birlikteliği) tanımlayan bir modelin bulunmasından meydana gelir. Bağımlılıklar diğer veri öğelerinin verilen bilgisine bağlı bir veri öğesinin değerinin öngörülmesinde kullanılabilir. Birliktelik kuralları son dönemlerde çok popüler hale gelmiş bağımlılıkların özel bir halidir. Birliktelik kuralları veri öğelerinin benzerliklerini tanımlar.

Uygulamalarda çoğu kez bağımlılık analizi bölümleme ile beraber gerçekleşir. Bağımlılık analizinde kullanılan teknikler şunlardır:

- Korelasyon Analizi
- Regresyon Analizi
- Birliktelik Kuralları
- Bayezyen Ağlar
- Tüme varım mantık programlaması
- Görselleştirme teknikleri

4. VERİ MADENCİLİĞİ TEKNİKLERİ

Veri Madenciliği problem tiplerinin çözümünde istatistiksel teknikler oldukça önemli işlevler görür. Özünde, veri madenciliği tekniklerinin çoğu istatistiksel teknikleri kullanır. Birçok alanda çok sık kullanılan Regresyon Analizi, Korelasyon Analizi bu tekniklerden bazılarıdır.

4.1. Kümeleme Analizi

Kümeleme Analizi çok değişkenli istatistiksel tekniklerden bir tanesidir[12] ve Veri Madenciliği'nde kullanılan önemli bir istatistiksel tekniktir. Kümeleme Analizi **X** veri matrisinde

Değişkenler

$$X' = \begin{bmatrix} x_{11} & \cdot & \cdot & x_{1p} \\ \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & x_{np} \end{bmatrix}_{np} \text{ Gözlemler}$$

yer alan birimleri, değişkenleri ya da birimleri ve değişkenleri birbiri ile benzer olan alt kümelere (grup, sınıf) ayırmaya yardımcı olan yöntemler topluluğudur (Özdamar, 2004a: 279). Kümeleme analizinin 2 temel amacı vardır, bunlar; veri indirgemek ve özet bilgiler elde etmektir. Kümeleme işlemi uzaklıklar veya benzerlikler ile yapılır. Nicel veriler için beş adet uzaklık[13] vardır ve herhangi biri hesaplanarak kullanılabilir. Bu uzaklıklar şunlardır:

1. Öklit Uzaklığı

$$d(i, j) = \sqrt{\sum_{k=1}^{P} (X_{ik} - X_{jk})^2}$$

2. Karesel Öklit Uzaklığı

$$d(i, j) = \sum_{k=1}^{P} (X_{ik} - X_{jk})^2$$

---

[12] Temel bileşenler analizi, faktör analizi, diskriminant analizi, çok boyutlu ölçekleme analizi, lojistik regresyon analizi çok değişkenli analiz teknikleri diğer bazı önemli çok değişkenli analiz teknikleridir.

[13] Ayrıntılar için bkz. Tatlıdil, 2002: 331, Özdamar, 2004a: 279–351.

3. Mahalanobis Uzaklığı

$$d(x_i, x_j) = D^2 = (\bar{x}_i - \bar{x}_j)' S^{-1} (\bar{x}_i - \bar{x}_j)$$

4. Hotelling Uzaklığı

$$T^2 = \frac{n_1 n_2}{n} (\bar{x}_i - \bar{x}_j)' S^{-1} (\bar{x}_i - \bar{x}_j)$$

5. Manhattan Uzaklığı

$$d_M(i, j) = \sum_{k=1}^{P} \left| X_{ik} - X_{jk} \right|$$

Uygulamada değişkenlerin tümünün nicel değişken olması mümkün değildir, bir kısmı nitel değişkenler olabilir. Bu durumda nitel veriler için de uzaklıkların hesaplanması gerekir. Böyle durumlarda uzaklıklar şöyle hesaplanır:

$$d(x_i, x_j) = \frac{1}{p} \sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right| \qquad w_k = \begin{cases} 1 & \text{Nicel Veri ise} \\ 1/\text{k.değ. dağ.} & \text{Nitel Veri ise} \end{cases}$$

Uzaklık matrisi "D" kullanılarak temelde iki kümeleme yöntemi vardır. Birinci yöntem hiyerarşik yöntemdir. Küme sayısına karar verilmemiş ise bu yöntem kullanılır.

1. Tek Bağıntı (En yakın komşuluk)

$$d_{k(i,j)} = \text{Min}(d_{ki}, d_{kj})$$

2. Tam Bağıntı (En uzak komşuluk)

$$d_{k(i,j)} = \text{Max}(d_{ki}, d_{kj})$$

İkinci yöntem ise hiyerarşik olmayan yöntemdir. Küme sayısına karar verilmiş ise bu yöntem kullanılır.

1. k-ortalama tekniği

$$W_n = \frac{1}{n} \sum_{i=1}^{n} \min \left| x_i - a_{jn} \right|^2$$

2. En çok olabilirlik

Gözlemler en büyük olabilirlik değeri verecek biçimde önceden belirlenen bir kümeye atanır.

Küme sayısının belirlenmesi için

$$k = (n/2)^{1/2}$$

formülünden yararlanılır.

4.2. Yapay Sinir Ağları

Genel anlamıyla yapay sinir ağları, beynin bir işlevini yerine getirme yöntemini modellemek için tasarlanan bir sistemdir. Yapay sinir ağı, yapay sinir hücrelerinin birbirleri ile çeşitli şekilde bağlanmasından oluşur ve genellikle katmanlar şeklinde düzenlenir (Oğuzlar, 2004: 57). Tek tabaka ya da tek eleman içeren bazı başarılı ağlar oluşturulabilmesine rağmen çoğu uygulamalar en az üç tabaka içeren ağlara ihtiyaç duymaktadır (Şentürk, 2006: 28). Bunlar:

1. Girdi tabakası: Dışarıdan veri alan nöronları içermektedir.
2. Çıktı tabakası: Çıktıları dışarı ileten nöronları içeren tabakalardır.
3. Gizli tabaka: Girdi ve Çıktı tabakaları arasında birden fazla gizli tabaka bulunabilir. Bu gizli tabakalar çok sayıda nöron içermektedir ve bu tamamen nöronlar ağ içindeki diğer nöronlarla bağlantılıdırlar.

İleri beslemeli çok katmanlı modeller en sık kullanılan yapay sinir ağı modelleridir. Bundan başka geri beslemeli yapay sinir ağı modelleri de mevcuttur. İleri beslemeli sinir ağlarında, hücreler katmanlar şeklinde düzenlenir ve bir katmandaki hücrelerin çıktıları bir sonraki katmana ağırlıklar üzerinden giriş olarak verilir. Geri beslemeli sistemde ise, en az bir hücrenin çıkışı kendisine ya da diğer hücrelere giriş olarak verilir ve genellikle geri besleme bir geciktirme elemanı üstünden yapılır. Yapay sinir ağlarının güçlü yönleri şunlardır:

- Çok sayıda gürültülü girdi verileri içeren veri kümelerinde iyi sonuçlar verir.
- Sayısal ve kategorik çıktıların ele alınıp tahmin edilmesine olanak tanır.
- Veri kümesinde zaman faktörünün gerekli olduğu uygulamalarda da kullanılır.
- Farklı alanlara iyi uyum gösterir.

4.3. Diskriminant Analizi

Diskriminant Analizi Veri Madenciliği'nde kullanılan kategorik bağımlı değişken(ler) ile sürekli bağımsız değişkenler arasındaki ilişkileri tahmin etmeyi amaçlayan çok değişkenli istatistiksel tekniklerden biridir.

Diskriminant Analizi'nin temel görevler şunlardır: a-Gruplanmış durumları birbirinden ayırmayı sağlayan fonksiyonları bulmak. Bu fonksiyonlarla gruplar arası ayırıma en fazla etki eden ayırıcı değişkenleri belirlemek. Bu değişkenlerin bağımlı değişkendeki değişimi açıklama gücünü belirlemek. b-Ayırıcı fonksiyon aracılığı ile yeni gözlenen birimleri sınıflama hatası minimum olacak şekilde bir gruba ayırmak.

Diskriminant Analizi'nin bazı varsayımları vardır, bunlar[14]:

1. Bağımsız değişkenlerin çoklu normal dağılıma sahip olması
2. Grupların kovaryans matrislerinin eşit olması
3. Değişkenlerin ortalamaları ve varyansları arasında korelasyon bulunmaması
4. Bağımsız değişkenler arasında çoklu doğrusal bağlantı probleminin olmaması
5. X değişkenler matrisi gereksiz değişken içermemesi

Diskriminant Analizi'nin aşamaları sırasıyla şöyledir: Değişkenler arasındaki korelasyonlar incelenir. Çoklu normal dağılım varsayımının sağlanıp sağlanmadığı (PP Plot veya QQ Plot grafikleri ile) araştırılır. Grup kovaryans matrislerinin türdeşliği Box M test istatistiğine bakılarak karar verilir. Grup kovaryansları ise Doğrusal Diskriminant Analizi uygulanır[15]. Diskriminant fonksiyonu $y_i = b_0 X_{j1} + b_1 X_{j2} + ... + b_p X_{jp}$ tahmin edilir[16]. Tahmin edilen fonksiyonun önemli olup olmadığı Wilk's Lambda istatistiği ile sınanır. Sınıflandırma tablosu ile diskriminant fonksiyonunun doğru sınıflandırma ve hatalı sınıflandırma sayıları ve yüzdeleri elde edilir. Standardize edilmiş kanonik diskriminant fonksiyonu katsayıları elde edilir. Bu katsayılara bakılarak bağımlı değişkenin kategorilere ayrılmasında önemli olan değişkenler belirlenir.

4.4. Karar Ağaçları

Herhangi bir konuda "karar" kelimesinin kullanılması mümkün iki veya daha fazla hareket tarzı arasından birinin seçimini ifade eder. Eğer tek hareket tarzı varsa seçim yoktur, "karar" teriminin uygulanması veya "karar alma" işlemi söz konusu olmaz.

Karar alacak her şahıs sonucun iyi olmasını bekler. Kötü sonuçla tamamlanan bir karar alma işlemi ilgili şâhısa pişmanlık verir, uygun bir seçim yapılmamış olduğunu açıklığıyla gösterir. Ayrıca mali yönden bakılırsa iyi bir karar kazançla, kötü bir karar kayıpla sonuçlanır (Ural, 1973: 4) . Her karar eyleminde en fazla altı eleman bulunur. Bunlar:

1. Karar veren: Mevcut seçeneklerden bir tercih yapan kişi veya grubu yansıtır.
2. Amaç: Karar verenin faaliyetleri ile elde edeceği amaçlardır.

---

[14] Özdamar, 2004a: 357–358.

[15] Aksi halde karesel diskriminant analizinin uygulanması uygun olacaktır.

[16] i=1….g veya g-1 ve j=1…n olmak üzere.

3. Karar Kriteri: Karar veren veya yöneticinin seçimini oluşturmada kullandığı değer sistemidir. Gelir, kar ve faydanın maksimizasyonu; maliyet, gider v.b. minimizasyonunu kapsayacaktır.
4. Seçenekler/Stratejiler (S): Karar verenin seçebileceği farklı alternatif faaliyetleridir. Seçenekler, karar verenin kontrolü altındaki kaynaklara bağlıdır ve kontrol edilebilir değişkenlerdir.
5. Olaylar (N): Karar verenin kontrolü altında olmayan faktörlerdir. Karar verenin seçenek tercihini etkileyen çevreyi olaylar yansıtacaktır.
6. Sonuç: Her bir seçenek veya olaydan ortaya çıkan değeri yansıtır.

Seçenek, olay, sonuç değerlerini kapsayan bir tabloya karar matrisi adı verilir(Halaç, 1991: 25–26). En iyi karar kriterini seçmek için olasılık mantığının grafik gösterimi ise karar ağaçları ile yapılır.

4.5. Regresyon Analizi

Regresyon analizi Veri Madenciliği'nde kullanılan bir diğer önemli istatistiksel tekniktir. Y bağımlı ve X_k (i=1…k) bağımsız değişken(ler) arasındaki sebep-sonuç ilişkilerini matematiksel model olarak ortaya koyan yönteme **regresyon** adı verilir (Özdamar, 2004b: 527). Basit doğrusal regresyon modeli Y bağımlı değişkeni ile X bağımsız değişkeni arasındaki ilişkiyi tahmin etmek üzere kurulan bir modeldir. Y bağımlı ve X_k (i=1…k) bağımsız değişkenleri arasındaki ilişkiyi tahmin etmek üzere kurulan bir modele çoklu doğrusal regresyon modeli denir.

Basit doğrusal regresyon modeli şöyle gösterilir:

$$Y = \alpha + \beta X + u$$

Bu modelde Y bağımlı değişken, X bağımsız değişkendir. α ve β tahmin edilecek olan katsayılardır. u ise hata terimidir. Model tahmin edildiğinde tahmin edilen katsayılar $\hat{\alpha}$ ve $\hat{\beta}$ olarak ifade edilir. Katsayıların tahmin edilmesi için alternatif yaklaşımlar bulunmasına karşın Olağan En Küçük Kareler Tahmin Yöntemi en sık kullanılanıdır.

4.6. Box-Jenkins Metotları

Box-Jenkins Metotları denildiğinde akla ilk gelen zaman serileri analizleridir. Zaman serisi analizlerinde değişkenler tek tek ele alınır. Değişkenlerin zaman içerisinde oluşan değerlerinin -ki buna veri üretme süreci (data generating process) denir- belirli bir kalıba (modele) göre meydana geldiği varsayılır ve uygun kalıbın(modelin) belirlenmesi amaçlanır. Bu modellerden bazıları: AR, MA, ARMA, Rassal Yürüyüş,

Kayan Rassal Yürüyüş, ARIMA'dır. Böylece değişkenin gelecekteki alacağı değerlerin sağlıklı bir biçimde öngörülebileceği varsayılır.

4.7. Korelasyon Analizi

Korelasyon Analizi Veri Madenciliği'nde kullanılan bir diğer önemli istatistiksel tekniktir. İstatistikte değişkenler arasındaki ilişkinin derecesini gösteren katsayıya **korelasyon katsayısı** denir (Gürsakal, 2002: 305). Nicel veriler arasındaki korelasyon katsayısı basit korelasyon katsayısı ile hesaplanırken nitel veriler arasındaki korelasyon ise sıra korelasyon katsayısı ile hesaplanmaktadır[17].

4.8. Birliktelik Kuralları

Birliktelik kuralları özel bir kural formunda verilir (Oğuzlar, 2004: 46). Bu özel form, sol ve sağ kısım olmak üzere birbirine bağlı iki kısımdan oluşur. Sol veya sağ kısımlarda yapılan iş veya nesneler yer alır ve veriler arasındaki ilişkiler, eğer-sonra ifadeleri ile gösterilir. Bir birliktelik kuralı için simgeler ile örnek **X⇒Y**, nesneler ile **kot pantolon⇒ kazak** veya yapılan herhangi ile **araştırma yapmak⇒ internet kullanımı** biçiminde örnekler verilebilir. Buradaki **Eğer** bölümü ile ilişkili durumlar öncül ve **sonra** bölümü ile ilişkili durumlar ise sonuç gösterilir.

Herhangi bir birliktelik kuralında destek ve güven olmak üzere iki önemli kavram vardır. **Destek**, öncül ve sonuç bölümlerinde yer alan öğeleri içeren işlem sayısının veri tabanındaki toplam işlem sayısına oranıdır. **Güven** ise, öncül ve sonuç bölümlerinde yer alan öğeleri içeren işlem sayısının öncül bölümünde yer alan işlem sayısına oranıdır.

5. SOSYO-EKONOMİK GELİŞMİŞLİK ÜZERİNE KÜÇÜK BİR UYGULAMA

İkinci bölümde bahsedildiği gibi, OLAP Küpleri, özet tablo ve özet grafik görünümlerinin kullanılması ile uygulayıcılara kolayca göz atabilecekleri, ulaşabilecekleri bilgileri sağlar[18]. Veri tanımlama ve özetleme **tek başına bir veri madenciliği projesinin amacı olabilir** (http://www.crisp-dm.org/CRISPWP-0800.pdf, erişim tarihi 29.10.2008). Bir veri ambarınızın olması, OLAP'a ihtiyacınız olmadığı anlamına gelemez. Veri ambarları ve OLAP birbirlerini tamamlarlar. Veri ambarları

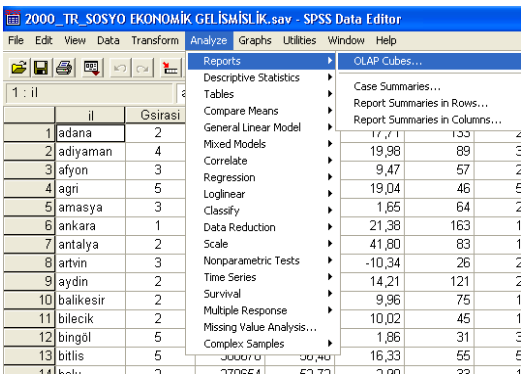
---

[17] Korelasyon katsayılarının hesaplanmasında kullanılan formüller için bkz. Gürsakal, 2002; Serper, 1996.

[18] http://office.microsoft.com/tr-tr/projectserver/HA100749111055.aspx (Erişim tarihi 29.10.2008).

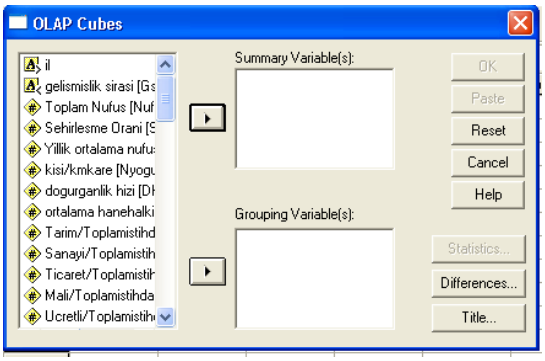
verileri barındırmaya yarar. OLAP ise, bu yığın halinde duran verileri anlamlı hale getirip analizler yapmaya yarar[19]. Bu çalışmada, Sosyo-ekonomik gelişmişlik üzerine küçük bir uygulama yapılmıştır ve veri tanımlama ve özetleme **bir veri madenciliği projesinin amacı** olarak ele alınmıştır. OLAP Analizi bağlamında çalışmanın bu bölümünde küçük bir uygulamaya yer verilecektir.

Bu çalışmanın konusu, Türkiye'nin 81 ilinin sosyoekonomik gelişmişliği üzerine 2000 yılında TUİK tarafından yapılan anket verilerin analizidir. Çalışmanın amacı ise, Türkiye'nin 81 ilinin sosyoekonomik gelişmişliği ile ilgili verileri kullanarak OLAP analizi yapılmasıdır. Kullanılan veriler Türkiye'nin 81 ilinin sosyoekonomik gelişmişliği üzerine 2000 yılında TUİK tarafından yapılan ankette yer alan 59 değişkenden[20] toplam veri sayısı 4779 (59*81=4779)'tır. SPSS 13.0 Paket programında aşağıdaki ekrandaki biçimde



EKRAN 1

Analyze / Reports / OLAP Cubes seçimi yapıldığında,



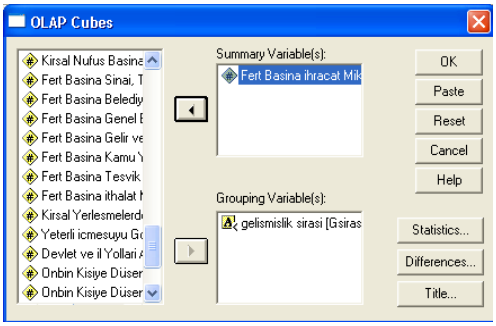
EKRAN 2

---

19 http://www.technoface.com/root/olap.aspx(Erişim tarihi 29.10.2008).

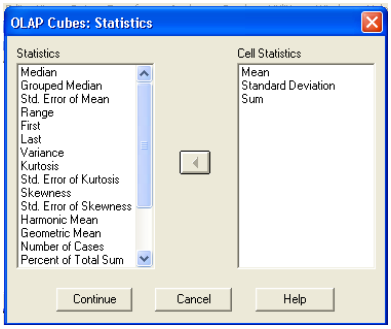
20 Değişken isimleri Ek'te verilmektedir.

OLAP CUBES diyalog kutusu ekranda belirir. Ekran 2'nin sol tarafında değişkenler görülmektedir. Summary Variable(s) kısmında raporlanacak değişken girilir. Grouping Variable(s) kısmına ise özeti yapılacak değişken ile ilgili grup değişkeni girilir. Örneğimizde, Summary Variable(s) kısmında raporlanacak değişkene Fert Başına ihracat Miktarın seçilsin. Grouping Variable(s) kısmına ise illerin gelişmişlik sırasını girelim[21].



EKRAN 3

Daha sonra **Statistics** seçeneği seçildiğinde,



EKRAN 4

Ekran 4'teki biçimde OLAP Küplerine ilişkin özet raporlanabilecek istatistikler belirlenebilir. Ekranın sol tarafında yer alan istatistikler aradaki ok ile sol tarafa aktarılarak ilgili raporlama yapılabilir. Şu anda sadece ortalama (mean), standart sapma(standart deviation) ve toplam(sum) istatistiklerin raporlanması istendiği için bu şıklar işaretlenmiştir. OK / OK tıklandığında,

---

[21] Gelişmişlik sırası TUİK tarafından yapılan Kümeleme Analizi sonucunda belirlenmiş olup, en gelişmiş illerden gelişmemiş illere doğru 1'den, 2, 3, 4 ve 5'e doğru sıralanmaktadır.

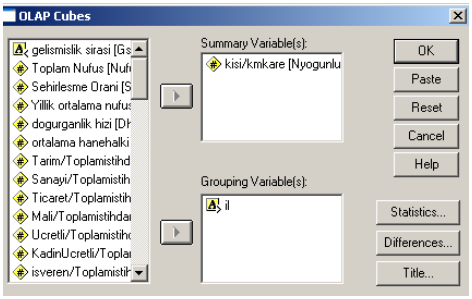
OLAP Cubes

gelismislik sirasi: Total

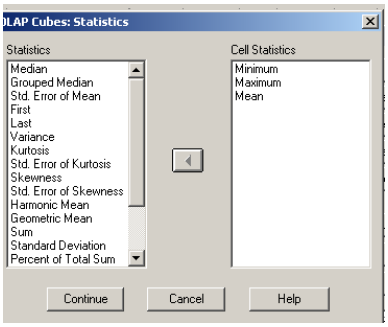
	Mean	Std. Deviation	Sum
Fert Basina ihracat Miktari / ABD dolari	1181,26	2045,958	73238

EKRAN 5

OLAP Küplerine ilişkin özet rapor elde edilir[22]. Burada örneğin toplam Türkiye'nin gelişmişlik sırasına göre kümelenen 5 kümenin toplamının (bir anlamda Türkiye'nin 2000 yılı fert başına ihracat miktarının) 73238 dolar olduğunu göstermektedir. Ekran 5'te yine bu beş kümenin fert başına ihracatının ortalaması 1181,26 ve standart sapmasının 2045,858 olduğu görülmektedir. Nüfus yoğunluğu değişkeni ile il bazında,



EKRAN 6

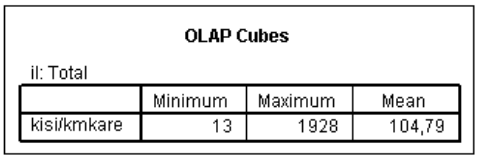


EKRAN 7

Ekran 6'daki ve 7'deki biçimde seçimler yapıldığında,

---

22 Amaca yönelik çeşitli raporlar yukarıda anlatılan yöntem çerçevesinde elde edilebilir.



OLAP Cubes

il: Total

	Minimum	Maximum	Mean
kisi/kmkare	13	1928	104,79

EKRAN 8

Ekran 8'e ulaşılır. Ekran 8'te kişi başı kilometre kareye en az 13 kişi, en çok 1928 kişi, ortalama olarak 105 kişi düştüğü görülmektedir.

6. SONUÇ

Bu çalışmada veri madenciliğinin istatistik ile olan ilişkisi irdelenmiştir. Çalışmanın genelinde veri madenciliği uygulaması yapacak araştırmacılara faydası olacağı düşüncesi ile veri madenciliği istatistik ekseninde genel bir çerçeve sunulma çabası içinde olunmuştur. Veri madenciliği tanımı, süreci, aşamaları, uygulama alanları, yerli ve yabancı yakın dönem literatürü ve veri madenciliğinin istatistik ile olan ilişkisi irdelenmiştir. Veri madenciliği problem tipleri ile veri madenciliği teknikleri ele alınmıştır. Bu çaba akabinde, sosyo-ekonomik gelişmişlik konusunda küçük bir OLAP küpleri uygulaması yapılmıştır.

Özünde veri madenciliğinin büyük veri tabanlarında gizli bilgi ve yapıyı açıklamak için çok sayıda veri analizi aracını kullanan bir süreç olduğu açıktır. Dolayısıyla, istatistik ile iç içedir. Veri madenciliğinin çeşitli problem tiplerine bir veya birkaç istatistiksel teknik kullanılarak çözümler üretilmeye çalışılır. Özetle, veri madenciliği analistlerinin istatistik konularına, istatistiksel tekniklere hâkim olması gerektiği söylenebilir. Bu çalışmada, bir veri madenciliği süreci, veri madenciliği aşamaları, problem tipleri ve istatistiksel teknikler düzenli bir biçimde ele alınmıştır. Herhangi bir veri madenciliği problem tipinde hangi istatistiksel tekniğin veya tekniklerin kullanılması gerektiği konusunda araştırmacıların yapacağı uygulamalarda faydalı olması nedeniyle harmanlanmış değerli bilgiler verilmiştir. Bu ve benzeri bilgilerin buluştuğu yer yine istatistik olmuştur. Sonuçta, istatistiğin veri madenciliği için "olmaz ise olmaz" bir durum halinde olduğu açıktır.

KAYNAKÇA

Abdelmelek, S. B., Saidane, S., Trabelsi, M. (2010) "Base Oils Biodegradability Prediction with Data Mining Techniques", *Algorithms,* Vol. 3, pp. 92–99.

Albayrak, A. S. ve Yılmaz, Ş. K. (2009) "Veri Madenciliği: Karar ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama", *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Dergisi,* Cilt 14, Sayı 1, s. 31–52.

Akpınar, H. (2000) "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, Cilt 29, Sayı 1/Nisan, s. 1–22.

Alpaydın, E. (2000) "Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri", *Bilişim 2000 Eğitim Semineri*.

Ata, H. A ve Seyrek, İ. H. (2009), "The Use of Datamining Techniques in Detecting Fraudulent Financial Statements: An Application on Manufacturing Firms", *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Dergisi,* Cilt 14, Sayı 2, s. 157–170.

Ayesha, S., Mustafa, T., Sattar, A. R. ve Khan, M. I. (2010), "Data Mining Model for Higher Education System", *European Journal of Scientific Research*, Vol. 43, No. 1, pp. 24–29.

Baykasoğlu, A. (2005) "Veri Madenciliği ve Çimento Sektöründe Bir Uygulama", *Akademik Bilişim Konferansı*, 2–4 Şubat Gaziantep Üniversitesi.

Bayram, N. (2009) *Sosyal Bilimlerde SPSS ile Veri Analizi*, Ezgi Kitabevi, Bursa.

Bozkır, A. S., Gök, B. ve Sezer E. (2008) "Üniversite Öğrencilerinin İnterneti Eğitimsel Amaçlar için Kullanmalarını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti", *Bilimde Modern Yöntemler Sempozyumu*, 15-17 Ekim 2008, Eskişehir Osmangazi Üniversitesi Kongre Merkezi, Eskişehir.

Chien, C. F. and Chen L. F. (2008), "Data Mining to Improve Personel Selection and Enhance Human Capital: A Case Study in High Technology Industry", *Expert Systems with Applications*, Vol. 34, pp. 280–290.

Çiflikli, C. and Özyirmidokuz, E. K. (2010), "Implementing a Data Mining Solution for Enhancing Carpet Manufacturing Prouctivity", *Knowledge-Based Systems*, In Pres.

CRISP-DM 1.0 Step-by-step data mining guide (http://www.crisp-dm.org/CRISPWP-0800.pdf, erişim tarihi 29.10.2008).

Dasu, T. ve Johnson, T. (2003) *Explarotary Data Mining and Data Cleaning*, John Wiley & Sons Publication, New Jersey, USA.

Daş, R. , Türkoğlu, İ. ve Poyraz, M. (2007) "Web Kayıt Dosyalarında İlginç Örüntülerin Keşfedilmesi", *Fırat Üniversitesi Fen ve Mühendislik Bilimleri Dergisi, Cilt 19, Sayı 4, s. 493–503.*

Deshpande, G., Gogolak, V., ve Smith, S. W. (2010) "Data Mining in Drug Safety", *Pharm Med.*, Vol. 24, No. 1., pp. 37–43.

Duru, N. ve Canbay, M. (2007) "Veri Madenciliği ile Deprem Verilerinin Analizi", *Uluslar arası Deprem Sempozyumu*, Kocaeli Üniversitesi, Kocaeli, 22–26 Ekim 2007.

Emel, G. ve Taşkın, Ç. (2005a) "Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması", *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi*, Cilt 6, Sayı 2, s. 221–236.

Emel, G. ve Taşkın, Ç. (2005b) "Pazarlama Stratejilerinin Oluşturulmasında bir Karar Destek Aracı: Birliktelik Kuralı Madenciliği" *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, Cilt 7, Sayı 3, 2005*.

Emel, G., Taşkın, Ç. ve Kılıçaslan, S. (2004) "Sinir Ağları Veri Madenciliği ile Çelik Üretim Sürecinde Bir Analiz", *Dokuz Eylül Üniversitesi İşletme Fakültesi Dergisi*, Cilt 5, Sayı 1, s. 205-225.

Friedman, J. H. (1997) "Data Mining and Statistics: What's the Connection?", http://www-stat.stanford.edu/~jhf/ (Erişim tarihi, 29.10.2008).

Ganesh, S. (2002) "Data Mining: Should it be included in the 'Statistics' cirriculum?", *The Sixt International Conference on Teaching Statistics*, Cape Town, South Africa, 7–12 July.

Gervilla, E., Cajal, B., Roca, J. ve Palmer, A. (2010) "Modelling Alcohol Consumption During Adolescence Using Zero Inflated Negative Binomial and Decision Trees", *The European Journal of Psychology Applied to Legal Context*, Vol 2, No. 2, pp. 145–159.

Glasgow, J. M. ve Kaboli, P. J. (2010) "Detecting adverse drug events through data mining", *Am J Health-Syst Pharm*, Vol. 67, pp. 317–320.

Gürbüz, Özbakır ve Yapıcı, (2009) "Türkiye'de Bir Havayolu İşletmesine ait Söküm Raporlarına İlişkin Veri Madenciliği Uygulaması", *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, Cilt 24, Sayı 1, s. 73–78.

Gürsakal, N. (2002) *Bilgisayar Uygulamalı İstatistik II*, Alfa Yayın Dağıtım, İstanbul.

Gürsakal, N. (2001) *Sosyal Bilimlerde Araştırma Yöntemleri*, Uludağ Üniversitesi Basımevi, Bursa.

Gürsakal, N. (2007) *Betimsel İstatistik Minitab, Spss, Statistica, Excel Uygulamalı*, Nobel Yayın Dağıtım, Ankara.

Halaç, O. , (1991) *Kantitatif Karar Verme Teknikleri: Yöneylem Araştırması*, Evrim Yayınları, İstanbul.

Koyuncugil, A. S. (2007) "Veri Madenciliği ve Sermaye Piyasalarına Uygulaması", *Sermaye Piyasası Kurulu Araştırma Raporu*, Araştırma Dairesi, 28.02.2007 ASK/1.

Kumar, N. V. A. ve Uma, G. V. (2009) "Improving Academic Performance of Students by Appyling Data Mining Technique", *European Journal of Scientific Research*, Vol. 34., No. 4, pp. 526–534.

Kuonen, D. (2004) "Data Mining and Statistics: What is the Connection?", *The Data Administration Newsletter*, http://www.tdan.com/view-articles/5226/ (Erişim tarihi, 29.10.2008).

Kusiak, A. and Smith, M. (2007) "Data Mining in Design of Products and Production Systems", *Annual Reviews in Control*, Vol. 31, Issue 1, pp. 147–156.

Liang, Yi-Hui (2010) "Integration of data mining techniques to analyze customer value for the automotive maintenance industry", *Expert Systems with Applications*, Vol. 37, pp. 7489–7496.

Liao, S.H., Chen, J. L. ve Hsu, T.Y. (2009) "Ontology-Based Data Mining Appoach Implemented for Sport Marketing", *Expert Systems with Applications*, Vol. 36, Issue 8, pp. 11045–11056.

Naveh, I. M. H. Sariri, I. ve Zadeh, B. A. (2009) "An Approach of Fault Detection and Prediction in Boiler of Power Plant Using Data-Mining: a Case Study of Application of Artificial Neural Network Technique", *International Review of Modelling and Simulations*, Vol. 2, No. 4., pp. 458–464.

Ogwueleka, F. N. (2009) "Potential Value of Data Mining for Customer Relationship Marketing in the Banking Industry", *Advances in Natural and Applied Sciences*, Vol. 3, Issue 1, pp. 73–78.

Oğuzlar, A. (2003) "Veri Önişleme", *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, Sayı 21/ Temmuz-Aralık, s. 67–76.

Oğuzlar, A. (2004) *Veri Madenciliğine Giriş*, Ekin Kitabevi, Bursa.

Oğuzlar, A. (2007) *İstatistiksel Veri Analizi*, Ezgi Kitabevi, Bursa.

Oğuzlar, A. ve Tüzüntürk, S., (2008) "Metin Madenciliğinin İşletmeler Açısından Önemi ve FMEA İçin Küçük Bir Uygulama Örneği", *9. Türkiye Ekonometri ve İstatistik Kongresi*, Dokuz Eylül Üniversitesi, Kuşadası, 28-30 Mayıs 2008.

Özçakır, F. C. ve Çamurcu, A. Y. (2007) "Birliktelik Kuralı Yöntemi için Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması", *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, Sayı 12/2, s.21–37.

Özdamar, K. (2004 a) *Paket Programlar ile İstatistiksel Veri Analizi (Çok Değişkenli Analizler),* Kaan Kitabevi, Eskişehir.

Özdamar, K. (2004 b) *Paket Programlar ile İstatistiksel Veri Analizi,* Kaan Kitabevi, Eskişehir.

Özkul, F. U. ve Pektekin, P. (2009) "Muhasebe Yolsuzluklarının Tesbitinde Adli Muhasebecinin Rolü ve Veri Madenciliği Tekniklerinin Kullanılması", MODAV, Cilt 4, s.57–87.

Özmen, Ş. (2001) "İş Hayatı Veri Madenciliği ile İstatistik Uygulamalarını Yeniden Keşfediyor", *V. Ulusal Ekonometri ve İstatistik Sempozyumu*, Çukurova Üniversitesi, Adana, 19–22 Eylül, 2001.

Rebbapragada, S., Basu, A. ve Semple, J. (2010) "Data Mining and Revenue Management Methodologies in College Admission", *Communications of the ACM*, Vol. 53, No. 4, pp. 128–133.

Seng, Jia-Lang ve Cheng, T. C. (2010) "An analytic approach to select data mining for business decision" , *Expert Systems with Applications*, Vol. 37, pp. 8042–8057.

Serper, Ö. (1996) *Uygulamalı İstatistik 2*, Filiz Kitabevi, İstanbul.

Srinivas ve Harding, J. A. (2010) "A data mining integrated architecture fors hop flor control", *Proc. IMechE*, Vol. 222, Part B, pp. 605–624.

Şentürk, A. (2006) *Veri Madenciliği Kavram ve Teknikler*, Ekin Kitabevi, Bursa.

Tatlıdil, H. (2002) *Uygulamalı Çok Değişkenli İstatistiksel Analiz*, Akademi Matbaası, Ankara.

Ural, K., (1973) *İstatistik ve Karar Alma*, İstanbul Üniversitesi Yayınları, İstanbul.

Vahaplar, A. ve İnceoğlu M. "Veri Madenciliği ve Elektronik Ticaret", www.bayar.edu.tr/baum/dokümanlar (erişim tarihi, 29.10.2008).

Yıldırım, P., Uludağ, M. ve Görür, A. (2007), "Hastane Bilgi Sistemlerinde Veri Madenciliği", *Akademik Bilişim Kongresi,* Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, 30 Ocak-1 Şubat 2007.

Zhang, Y., Ma, J., Zhang J. ve Wang, Z. (2009), "Applications of Data Mining Theory in Electrical Engineering", *Engineering,* Vol. 1, pp. 79–83.

Zhao Chung-Mei ve Luan, J. (2006) "Data Mining: Going Beyond Traditional Statistics", *New Directions for Institutional Research*, No. 131, pp. 7–16.

EK 1 Sosyo-ekonomik Gelişmişlik Değişkenleri

1	Gelişmişlik sırası	31	Kişi başına imalat sanayi elektrik tüketimi / Kws
2	Toplam Nüfus	32	Kişi basına imalat sanayi katma değeri / Milyon TL
3	Şehirleşme Oranı	33	Kırsal nüfus başına tarımsal üretim / Milyon TL
4	Yıllık ortalama nüfus artış hızı	34	Tarımsal üretim değerinin Türkiye içindeki payı
5	Kişi/km kare	35	Daire sayısı
6	Doğurganlık hızı	36	Borulu su tesisatı bulunan daire oranı
7	Ortalama hane halkı büyüklüğü	37	GSYH içindeki payı / yüzde
8	Tarım/Toplam istihdam	38	Kişi Başı GSYIH
9	Sanayi/Toplam istihdam	39	Banka şube sayısı
10	Ticaret/Toplam istihdam	40	Kişi başına banka mevduatı / Milyon TL
11	Mali/Toplam istihdam	41	Toplam Banka Mevduatı içindeki Pay
12	Ücretli/Toplam istihdam	42	Toplam Banka Kredileri içindeki Pay
13	Kadın Ücretli/Toplam istihdam	43	Kırsal Nüfus Başına Tarımsal Kredi Miktarı / Milyon TL
14	İşveren/Toplam istihdam	44	Fert Başına Sınaî, Ticari ve Turizm Kredileri Miktarı / Milyon TL
15	Okuryazar/Toplam nüfus	45	Fert Başına Belediye Giderleri / Milyon TL
16	Okuryazar/Kadın nüfus	46	Fert Başına Genel Bütçe Gelirleri / Milyon TL
17	Üniversite mezunu/ 22 yas ustu nüfus	47	Fert Başına Gelir ve Kurumlar Vergisi Miktarı / Milyon TL
18	Okullaşma oranı	48	Fert Başına Kamu Yatırımları Miktarı / Milyon TL
19	Okullaşma oranı	49	Fert Başına Teşvik Belgeli Yatırım Tutarı / Milyon TL.
20	Okullaşma oranı	50	Fert Başına ihracat Miktarı / ABD doları
21	Bebek olum oranı	51	Fert Başına ithalat Miktarı / ABD doları
22	On bin kişiye düsen hekim sayısı	52	Kırsal Yerleşmelerde Asfalt Yol Oranı
23	On bin kişiye düsen diş hekimi sayısı	53	Yeterli içme suyu götürülen Nüfus Oranı
24	On bin kişiye düsen eczane sayısı	54	Devlet ve il Yolları Asfalt Yol Oranı
25	On bin kişiye düşen hastane yatak sayısı	55	On bin Kişiye Düsen özel Otomobil Sayısı
26	Organize sanayi bölgesi parsel sayısı	56	On bin Kişiye Düsen Motorlu Kara Taşıtı Sayısı
27	Küçük sanayi sitesindeki işyeri sayısı	57	Fert Başına Elektrik Tüketim Miktarı
28	İmalat sanayi işyeri sayısı	58	Fert Başına Telefon Kontör Değeri
29	İmalat sanayi ortalama çalışan sayısı	59	Yeşil Karta Sahip Nüfus Oranı
30	İmalat sanayi beygir gücü		