T.C.

İstanbul Üniversitesi

Sosyal Bilimler Enstitüsü

İşletme Anabilim Dalı

Sayısal Yöntemler Bilim Dalı

Yüksek Lisans Tezi

# VERİ MADENCİLİĞİ VE VERİ MADENCİLİĞİNDE KULLANILAN K-MEANS ALGORİTMASININ ÖĞRENCİ VERİ TABANINDA UYGULANMASI



Şenol Zafer ERDOĞAN

2501010612

Tez Danışmanı : Yrd. Doç. Dr. Mehpare TİMOR

İstanbul, 2004

T.C
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ

# TEZ ONAYI

Enstitümüz **SAYISAL YÖNTEMLER** Bilim Dalında **2501010612** numaralı **ŞENOL ZAFER ERDOĞAN'ın** hazırladığı **"VERİ MADENCİLİĞİNDE KULLANILAN K-MEANS ALGORİTMASININ ÖĞRENCİ VERİ TABANINDA UYGULANMASI"** konulu **YÜKSEK LİSANS/** ~~**DOKTORA TEZİ**~~ ile ilgili **TEZ SAVUNMA SINAVI,** Lisansüstü Öğretim Yönetmeliği'nin 10.Maddesi uyarınca **14/07/2004 Çarşamba** günü saat **10.30'da** yapılmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin .....Kabulü.....'ne* **OYBİRLİĞİ /**~~**OYÇOKLUĞUYLA**~~ karar verilmiştir.

| JÜRİ ÜYESİ | KANAATİ(*) | İMZA |
|---|---|---|
| PROF.DR.ÖNER ESEN | Kabulü | |
| PROF.DR.ŞABAN EREN | Kabulü | |
| DOÇ.DR.ERHAN ÖZDEMİR | Kabulü | |
| YRD.DOÇ.DR.MEHPARE TİMOR | Kabulü | |
| YRD.DOÇ.DR.ALP BARAY | Kabulü | |

# ÖZ

Bilgi miktarının büyük oranlarda arttığı bu bilgi çağında büyük hacimlerdeki verilerden anlamlı bilgilerin elde edilmesi bir süreç gerektirmektedir. Bu sürecin en önemli adımı ise veri madenciliğidir.

Bu çalışmada, bilginin ortaya çıkarılması sürecinin tamamı araştırılmış ve incelenmiştir. Yine aynı şekilde sürecin en önemli adımı olan veri madenciliği adımı da ayrıntılı olarak incelenmiştir.

Tez içerisinde bilginin keşfi süreci aşamalarıyla açıklanmıştır. Sürecin en önemli adımı olan veri madenciliği adımı ayrıntılı olarak ele alınmış ve veri madenciliği tekniklerinden bahsedilmiştir. Veri madenciliği tekniklerinden olan kümeleme analizi ayrıntılı olarak incelenmiştir.

Tezin son kısmında veri madenciliği algoritmalarından biri olan k-means algoritması öğrenci veri tabanına uygulanmış ve sonuçlar ortaya konmuştur.



# ABSTRACT

As the rate of information growth, increases in this information age, extracting meaningful information from huge volumes of data requires a certain process. The most important step of this process is data mining.

In this study, a comprehensive survey of the knowledge process is presented and the data mining step is investigated, as the most important step of this process.

Steps of the knowledge discovery process are explained and discussed. Data mining is presented as the most important step in this process and data mining

techniques are reviewed. Cluster analysis which is one of the data mining techniques is investigated in detail.

In the last part of thesis, one of the data mining algorithms, the k-means algorithm, is applied to a student database and results are shown.

# ÖNSÖZ

Bu çalışmada, veri madenciliği ve veri madenciliği süreçleri detaylı bir şekilde incelenmiştir. Veri madenciliği içerisinde bulunan tekniklerden bahsedilmiştir. Bu süreç içerisinde uygulanabilecek bir analiz olan kümeleme analizinden bahsedilmiştir. Tez içerisinde veri madenciliğinin farklı bir yönden uygulaması yapılmıştır. Üniversiteye giren öğrencilerin yapısı incelenmiş ve veri madenciliği süreci sonunda grupların oluştuğu görülmüştür ve bu sonuçlar ışığında öneriler sunulmuştur.

Bu çalışmanın hazırlanmasında büyük desteğini gördüğüm tez danışmanım Yrd. Doç. Dr. Mehpare TİMOR' a, uygulama sırasında değerli yardımlarından dolayı Maltepe Üniversitesi Öğrenci İşleri Daire Başkanı Sacide DURUKAN' a, Maltepe Üniversitesi Bilgi İşlem vekili Fatih Yücalar' a ve bana her zaman destek olan aileme teşekkürü bir borç bilirim.

# İÇİNDEKİLER

## İKİNCİ BÖLÜM
## KÜMELEME ANALİZİ

## ÜÇÜNCÜ BÖLÜM
## UYGULAMA: BÖLÜMLEME METODU ALGORİTMALARINDAN K-MEANS ALGORİTMASININ ÖĞRENCİ VERİ TABANINA UYGULANMASI



## DÖRDÜNCÜ BÖLÜM
## SONUÇ

# TABLOLAR LİSTESİ

# ŞEKİLLER LİSTESİ

# KISALTMALAR LİSTESİ

| | |
|---|---|
| **BIRCH** | : Balanced Iterative Reducing and Clustering Using Hierarchies |
| **CF** | : Clustering Feature |
| **CHAMELEON** | : Hierarchical Clustering Using Dynamic Modeling |
| **CLARA** | : Clustering Large Applications |
| **CLARANS** | : Clustering Algorithm based on Randomized Search |
| **CRM** | : Customer Relationship Management |
| **CURE** | : Clustering Using Represantatives |
| **I / O** | : Input / Output |
| **JDBC** | : Java Database Connection |
| **KDD** | : Knowledge Discovery in Databases |
| **LS** | : Linear Sum |
| **ODBC** | : Open Database Connection |
| **OLAP** | : OnLine Analytical Processing |
| **MOLAP** | : Multidimensional OnLine Analytical Processing |
| **PAM** | : Partitioning Around Medoids |
| **RAM** | : Random Access Memory |
| **ROCK** | : Robust Clustering Algorithm |
| **ROLAP** | : Relational OnLine Analytical Processing |
| **VLDM** | : Very Large Data Bases |
| **VTBK** | : Veri Tabanlarında Bilgi Keşfi |
| **YSA** | : Yapay Sinir Ağları |

# GİRİŞ

Geçmiş yıllardan bugüne kadar geçen zaman peryoduna bakarsak insanlar yaşadıkları tecrübeleri veya kazandıkları bilgileri diğer insanlarla paylaşmaya ve aktarmaya çalışmışlardır. Bu kazanımları elde eden insanlarda geleceğe yönelik olarak farklı bakış açılarına sahip olmuşlardır.

Bilgi ve tecrübelerin diğer insanlara aktarımında daha öncelerden beri kağıt ortamları kullanılmıştır. Günümüzde ise bu bilgilerin büyük bir kısmı artık dijital ortamlarda saklanmaktadır.

Dijital ortamlarda saklanan bilgi miktarı gün geçtikçe büyük oranlarda artmakta ve adeta veri dağlarına dönüşmektedir. Veri dağlarına dönüşen bu veriler veri tabanlarında saklanmaktadır. Bu denli büyüklüklere ulaşan bu veri dağlarında anlamlı bilgilerin ortaya çıkartılması gerekmektedir.

Büyük hacimlardeki verilerden anlam taşıyan bilgilerin çıkartılması bir süreç gerektirmektedir. Bu sürecin tamamı veri tabanlarında bilgi keşfi olarak adlandırılmaktadır. Bu sürecin en önemli adımı ise veri madenciliğidir.

Süreç sonunda varolan eldeki veriler ışığı altında bazı bağıntılar, örüntüler veya kurallar elde edilmesiyle geleceğe yönelik tahminlerin yapılması veya kararların alınması sağlanacaktır.

Günümüzde veri madenciliğinin kullanımı her gün yeni sektörlerinde dahil olmasıyla artmaktadır. Veri madenciliğinin kullanımının artmasında ki önemli unsurlardan birisi de sürecin tamamının bilgisayar ortamına yazılan programlar

vasıtasıyla tamamlanmasıdır. Bu şekilde son kullanıcılar tarafından süreç daha kolay gerçekleşebilmektedir.

Bilginin ortaya çıkartılması ve bilginin elde edilmesinin önemli olduğu bu zamanda, veri madenciliğinin önemi daha da iyi anlaşılmaktadır.

# BÖLÜM 1

# VERİ MADENCİLİĞİ

## 1.1. VERİ MADENCİLİĞİ TANIMI

Teknoloji her geçen gün ilerlemekte ve gelişmektedir. Teknoloji, hayatımızın hemen hemen her alanına girmiştir. Özellikle verinin dijital ortamlarda saklanmasıyla beraber pek çok elektronik cihaz günlük hayatımızda kullanılmaya başlamıştır. Günümüzde bankacılık sektöründe, alışveriş merkezlerinde, sağlık sektöründe, sigorta sektöründe, vb. yerlerde elektronik cihazların kullanılmasıyla beraber saklanan veri miktarı heyecan verici boyutlara ulaşmıştır.

Verilerin dijital ortamda saklanmaya başlaması ile birlikte, yeryüzündeki bilgi miktarının her 20 ayda bir kendini iki katına çıkardığı günümüzde, veri tabanlarının sayısı da benzer, hatta daha yüksek bir oranda artmaktadır.[1]

Veri miktarının inanılmaz boyutlara ulaşması ve veri tabanlarının artan kullanımlarıyla beraber, bu toplanan yığınlardan anlamlı bilgiler çıkarmak insan gücü ile hem daha yavaş hem daha pahalı hem de daha öznel olacaktır. Bu noktadan hareketle yapılan çalışmalar sonucunda, ***Veri Tabanlarında Bilgi Keşfi-VTBK (Knowledge Discovery in Databases- KDD)*** adı altında yeni bir kavram ortaya çıkmıştır.[2]

---

[1] Alper Vahaplar, Dr. Mustafa İnceoğlu, **"Veri Madenciliği ve Elektronik Ticaret"**, (Çevirimiçi) http://inet-tr.org.tr/inetconf7/bildiriler/78.doc, 01 Kasım 2003.

[2] Haldun Akpınar, **"Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği"**, İstanbul, İ.Ü. İşletme Fakültesi Dergisi, Sayı:1, Nisan 2000, S: 1-22.

VTBK süreci içerisinde büyük önemi bulunan modelin kurulması ve değerlendirilmesi aşamalarına genel olarak *veri madenciliği* adı verilmektedir. Bu önemden dolayı birçok kaynakta VTBK ile veri madenciliği eş anlamlı olarak kullanılmaktadır.[3] VTBK sürecinin merkezinde bulunan veri madenciliği teknikleri bilginin keşfi ve çıkarılması uygulamalarını gerçekleştirmektedir.

Veri madenciliği, dünyanın önde gelen araştırma ve danışmanlık firmalarının açıkladığı bazı rakamlara göre gelecekte oldukça popüler bir konu olacaktır. Gartner Group araştırma şirketi gelecek on yıl içinde hedef pazarlarda veri madenciliği kullanımının %80 'lere ulaşacağı tahmininde bulunuyor.[4] Yine Gartner Group tarafından yapılan araştırmada gelecek 5 yılın ilk 5 teknolojisi listesinde veri madenciliği ve yapay zeka yer almaktadır.[5]

Veri madenciliğinin tanımı ile ilgili pek çok farklı görüş ortaya konulmuştur. Kabul gören görüşlerden bazıları aşağıda verilmiştir.

Veri madenciliği, organizasyonların karar aşamaları için yeni bilgiler üreten ya da gelecekle ilgili tahminler ve planlar yapmamızı sağlayan bir dizi teknikler ve anlayışlar bütünüdür.[6]

Veri madenciliği, büyük veritabanlarından, çok net olmayan, üstü kapalı, önceden bilinmeyen ancak potansiyel olarak kullanışlı olabilecek bilginin çıkarılmasıdır.[7]

Veri madenciliği, veri yığınları içindeki örüntüleri ve ilişkileri ortaya çıkaran, çeşitli veri analiz araçlarının kullanıldığı bir süreçtir.[8]

---

[3] A.e.
[4] Kurt Thearling, **"An Introduction to Data Mining"**, (Çevrimiçi) http://www.thearling.com/text/dmwhite/dmwhite.htm, 01 Aralık 2003.
[5] Melikşah Karakaş, **"Veri Madenciliği Üzerine"**, (Çevrimiçi) http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=132, 30Kasım 2003.
[6] A.e.
[7] U.M. Fayyad, G. Piatesky-Shapiro, P. Smyth, R. Uthurusamy, **"Advances in data mining and knowledge discovery"**, USA, MIT Press, 1994.

Veri madenciliği, büyük veri tabanlarında örüntülerin, birlikteliklerin, anormalliklerin, ve çeşitli yapıların yarı otomatik bir sistem ile keşfidir.[9]

Veri madenciliğinin zaman içindeki gelişim adımları aşağıdaki tablo 1'de gösterilmektedir.

| **Gelişim Adımları** | **Ticari Sorular** | **Geçerli Teknolojiler** | **Üretici Firmalar** |
|---|---|---|---|
| Veri Toplanması (1960) | Geçmiş beş yıldaki toplam gelirim nedir? | Bilgisayarlar, Teypler, Diskler | IBM, CDC |
| Veri Erişim (1980) | Geçen mart ayında İngiltere'de ki şube ne sattı? | İlişkisel veritabanları, Yapısal sorgulama dili, ODBC | Oracle, Sysbase, Informix, IBM, Microsoft |
| Veri Ambarı ve Karar Destek Sistemleri (1990) | Geçen mart ayında New England'daki şube satışları ne kadardı?(Boston bolgesi dahil) | OLAP, Çok boyutlu veri tabanları, veri ambarı | Pilot, Comshare, Arbor, Cognos, Microstrategy |
| Veri Madenciliği (Bugün) | Boston'da satışların gelecek ay ne kadar olmasını bekliyorsunuz?Neden? | İleri algoritmalar, Çok işlemcili bilgisayarlar, büyük veritabanları | Pilot, Lockheed, IBM, SGI |

**Tablo 1. Veri Madenciliği 'nin Gelişim Adımları**[10]

## 1.2. VERİ MADENCİLİĞİ UYGULAMA ALANLARI

Günümüzde veri madenciliği farklı farklı alanlarda uygulanmaya başlamıştır. Astronomi, biyoloji, finans, pazarlama, sigorta, tıp, kimya, sosyal bilimler, web madenciliği ve bir çok farklı alanlarda uygulanmaktadır.

---

[8] Two Crows Corporation, **"Introduction to Data Mining and Knowledge Discovery"**, USA, Two Crows Corporation, 1999.

[9] Robert L. Grossman, Chandrika Kamath, Vipin Kumar, **"Data Mining For Scientific And Engineering Approach"**, USA, Kluwer Academic Publishers, Ekim 2001.

[10] Thearling, **"An Introduction to Data Mining"**, s.1.

Veri madenciliğinin kullanıldığı alanların başlıcaları aşağıdaki kategoriler altında toplanmıştır.

*a) Pazarlama :*

Bu alana baktığımız zaman en önemli uygulamaların, farklı müşteri gruplarını belirlemek ve bu müşteri gruplarının davranışlarını tahmin etmeye yönelik olarak analiz çalışmaları yapmak olduğunu görüyoruz. American Express veri madenciliği tekniklerini kullanarak %10 - %15 arasında kredi kartı kullanımını arttırmıştır.[11]

Müşterilerin satın alma örüntülerinin belirlenmesi, müşterilerin demografik özellikleri, mevcut müşterilerin elde tutulması, pazar sepet analizi *(Market Basket Analysis)*, müşteri ilişkileri yönetimi *(CRM-Customer Relationship Management)*, müşteri değerlendirme, satış tahmini, hedef pazar analizi, müşteriler arası benzerliklerin saptanması gibi alanlarda da veri madenciliği teknikleri uygulanmaktadır.

*b) Biyoloji, Tıp ve Genetik :*

DNA sıraları içerisinde genlerin tespiti, gen haritalarının analizi, kanserli hücrelerin tespiti, genetik hastalıkların tespiti gibi alanlarda veri madenciliği uygulanmaktadır.

*c) Banka ve Sigorta :*

Finansal tablolar arasında korelasyon tespiti, kredi kartı dolandırıcılıklarının tespiti, kredi kartı taleplerinin değerlendirilmesi, sigorta dolandırıcılıklarının tespiti, riskli müşteri örüntülerinin belirlenmesi, yeni

[11] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, **"From Data Mining to Knowledge Discovery in Databases"**, USA, AAAI Pres, 1996.

poliçe talep edebilecek müşterilerin tahmini gibi alanlarda veri madenciliği teknikleri uygulanmaktadır.

Kredi kartı ödemelerini aksatan, gecikmeli yapan veya hiç yapmayan insanların özellikleri incelenerek bu tip sonuçları verebilecek muhtemel kişilerin belirlenmesinde kullanılmaktadır.

*d) Web Madenciliği (Web Mining) :*

İnternet ve web üzerindeki dosyalar hem hacim hem de karmaşıklık olarak hızla artmaktadır. Web madenciliği veri madenciliği tekniklerini kullanarak World Wide Web 'de bulunan dosya ve servislerden otomatik olarak örüntüler bulur ve öngörülmeyen bilgilere ulaşır.[12] Böylece veriye ulaşım süresinin azaltılması amaçlanır.

*e) Yüzey Analizi ve Coğrafi Bilgi Sistemleri :*

Bölgelerin coğrafi özelliklerine göre sınıflandırılması, kentlerde yerleşim yerlerinin belirlenmesi, kentlerde suç oranının tespiti, otomatik para makinelerinin yerlerinin tespiti, otobüs duraklarının yerlerinin belirlenmesi gibi durumlar için veri madenciliği uygulanmaktadır.

## 1.3. VERİ AMBARI

### 1.3.1. VERİ AMBARI TANIMI

Bilgisayarın kullanılmaya başlandığı her sektörde, veri tabanlarında bulunan veri miktarı gün geçtikçe inanılmaz boyutlarda artmaktadır. Bu hızlı ilerlemeyle beraber **Veri Ambarı *(Data Warehouse)*** kavramı ortaya çıkmıştır. 1990'lı yılların

[12] Oren Etzioni, **"The World Wide Web:Quagmire or Gold Mire"**, USA, Communications of The ACM, Kasım 1996, S:65-68.

başından itibaren şirketlerin sorunlarına genel amaçlı çözümler üretmek için veri ambarı kullanılmaya başlandı.[13] Veri ambarı sektörünün toplam büyüklüğü (yazılım ve donanım da dahil) 1998 yılında 8 milyar dolar olarak tahmin edilmiştir.[14]

Veri ambarı, şirketlerin analiz ve raporlama işlemlerini yapması için yönetim bilgilerinin tek ve tutarlı bir şekilde tutulduğu bir veritabanıdır.[15]

Veri ambarı, karar verme sürecine yardımcı olan, konu tabanlı, birleştirilmiş, zamana bağlı, verilerin sabit olduğu veriler topluluğudur.[16]

Rakiplerinin önüne geçirecek kritik kararların alınmasına yardımcı olacak ve şirket performansının ölçümünü sağlayacak bilgilerin toplanmasını sağlar ve bununla beraber trendleri ve örüntüleri uzun bir zaman periyodu içinde izleyerek maliyetleri azaltabilir.[17]

Veri ambarının taşıması gereken özellikleri aşağıdaki gibi açıklayabiliriz.

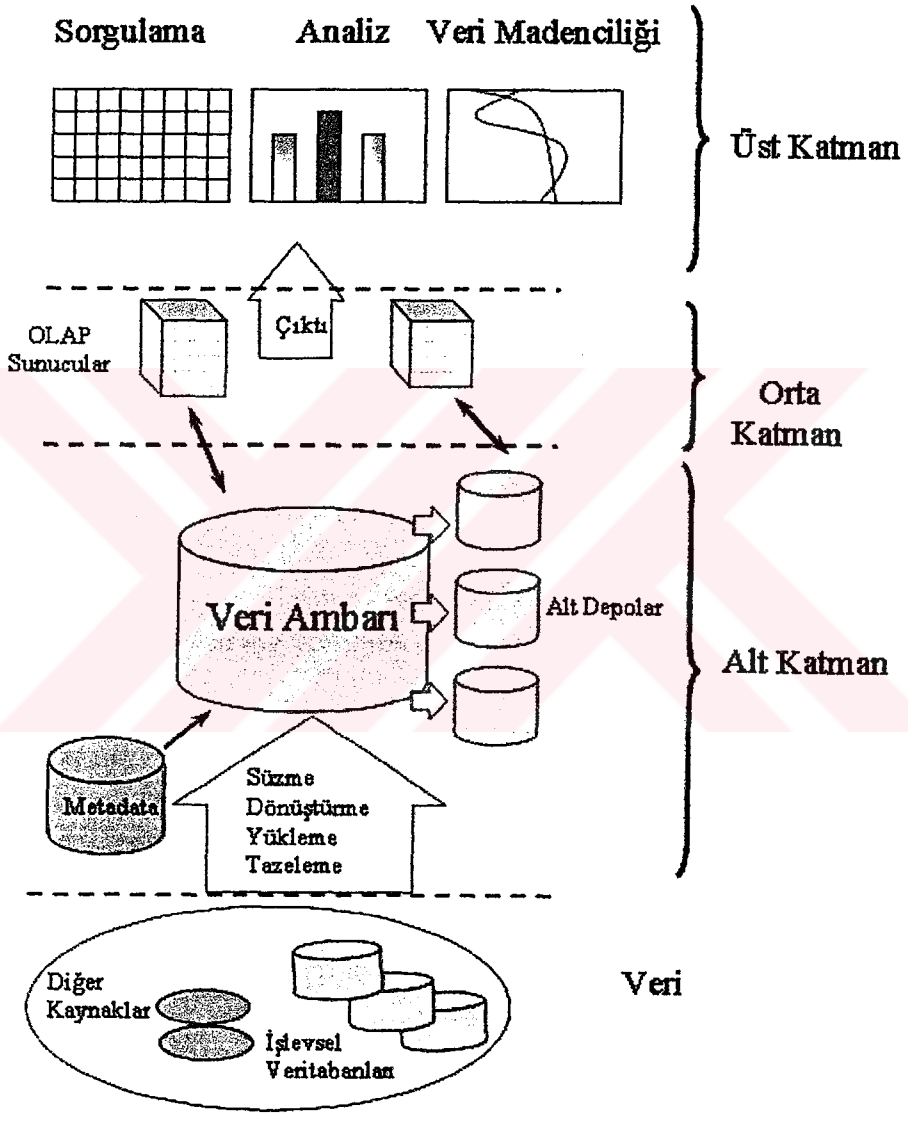
- Veri ambarları satış verileri veya müşteri bilgileri gibi belirli konularda veriler içerirler.
- Veri ambarları birçok farklı kaynaktan gelen verilerin toplanmasıyla oluşur. İçerisinde ilişkisel veri tabanları, düz metin dosyaları bulunabilir.
- Veriler veri ambarlarına belli periyotlarla eklenir. Örneğin aylara göre son on yılın satış bilgileri.
- Veri ambarları işlevsel veri tabanlarında olduğu gibi sürekli olarak güncellenmez. Veriler sabit olarak veri ambarına kayıt edilir.

---

[13] J. Pokorny, **"Data Warehouses: A Modelling Perspective"**, Slovenia, Proceedings of the 7.International Conference on Information Systems Development, Plenum Press, 1998.
[14] A. Sen-V.S. Jacob, **"Industrial Strength Data Warehousing"**, Communications of the ACM,1998.
[15] B. Love, **"Enterprise Information Technologies"**, Van Nostrand Reinhold, 1993.
[16] W. Inmon, **"What is A Data Warehouse?"**, Prism Tech Topic, Vol:1, 1992.
[17] J. Han, M. Kamber, **"Data Mining Concepts and Techniques"**, Morgan Kaufmann Publishers Inc., s. 63 Ağustos 2001.

## 1.3.2. VERİ AMBARI MİMARİSİ

Veri ambarı alt katman, orta katman ve üst katmandan oluşan 3 katmanlı bir mimariye sahiptir. Şekil 1'de görülen bu katmanlar şunlardır:




**Şekil 1. Veri Ambarının Mimarisi**[18]

[18] A.e., s. 46.

*a. Alt Katman:*

Alt katman ilişkisel bir veri tabanı sistemidir. İşlevsel veri tabanlarından ve dış kaynaklardan gelen veriler geçit yolu (gateway) olarak bilinen uygulama programları arayüzleri tarafından çekilir. Geçit yolu programlarının en bilinenleri Sun Microsystems firmasının *JDBC (Java Database Connection)* ve Microsoft firmasının *ODBC (Open Database Connection)* ürünleridir.

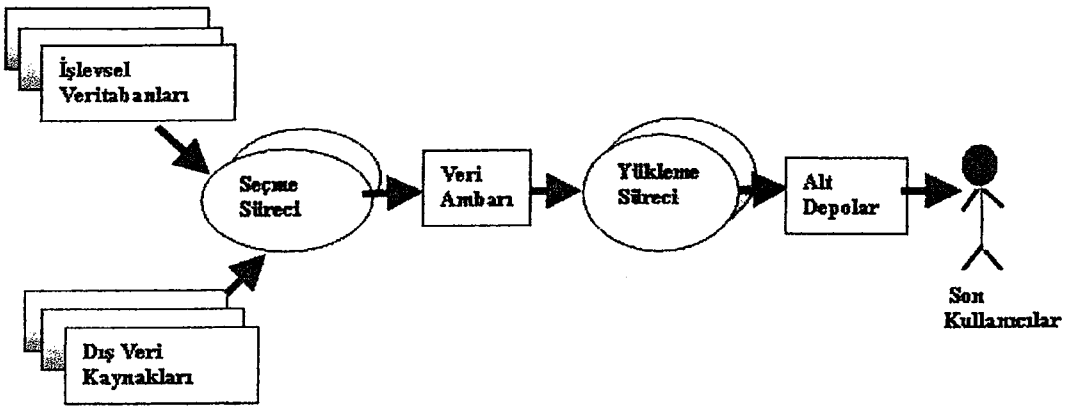
*b. Orta Katman:*

Orta katman ya ilişkisel OLAP *(ROLAP- Relational OnLine Analytical Processing)* yada çok boyutlu OLAP *(MOLAP – Multidimensionnal OLAP)* modelini kullanarak başlayan bir OLAP *(On Line Analytical Processing)* sunucudur. Alt katmandan gelen veriler, ROLAP yada MOLAP modellerinden birinin kullanılmasıyla raporlama, analiz ve veri madenciliği işlemleri için verileri anlamlı bir hale getirilir.

*c. Üst Katman:*

Üst katman sorgulama, raporlama araçları, analiz araçları ve veri madenciliği araçlarını içeren bir istemcidir.

## 1.3.3. VERİ AMBARI SÜRECİNDEKİ UNSURLAR[19]




**Şekil 2. Veri Ambarı Süreci**

*a. İşlevsel Sistemler (Operational Systems):*

Ticari işlemler sonucunda oluşan verilerin detaylı bir şekilde kaydını tutan sistemlerdir.Karar destek sistemleri için gerekli olan verilerin çoğunu bu sistemler oluşturur.

*b. Dış Kaynaklar (External Process):*

Veri ambarı oluşturmak ve yapılacak analizlere destek sağlamak için sık sık dış kaynaklardan veri alınmaktadır. Örneğin ekonomik veriler, nüfus verileri gibi.

*c. Seçme/Çıkarma Süreci (Extract Process):*

Veriler düzenli ve tutarlı bir yapı içinde veri ambarlarında saklanır. Veri, farklı kaynaklardan çıkarılır, birleştirilir ve uygun bir biçimde veri ambarlarında saklanır.

---

[19] Daniel L. Moody, **"From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design"**, (Çevrimiçi) http://sunsite.informatik.rwth-aachen.de/publications/ceur-ws/vol-28/paper5.pdf, 2000.

*d. Veri Ambarı (Data Warehouse):*

Karar destek sistemi için gerekli olan verilerin merkezi durumundadır. Burada alt depolar (Data Mart) yaratılır.

*e. Yükleme Süreci (Load Process):*

Veri ambarlarında bulunan veriler alt depolara (Data Mart) ayrılırlar. Bu süreçte bir dağıtım fonksiyonu kullanılır.

*f. Alt Depolar (Data Marts):*

Burayı veri ambarının dağıtım yeri gibi düşünebiliriz. Burada son kullanıcıların analiz işlemlerini yapabilmesi için verinin uygun biçime getirilmesi sağlanır. Alt depolara kullanıcı gruplarının isteklerine göre sürekli olarak yeni bir biçim verilir.

*g. Son Kullanıcılar :*

Kullanıcı dostu *(user friendly)* sorgulama araçları kullanılarak, alt depolar içerisinde ki kayıtlı veriler üzerinde analiz ve sorgulama yapılır.

## 1.3.4. İŞLEVSEL VERİ TABANI İLE VERİ AMBARI ARASINDAKİ FARKLAR

İşlevsel veri tabanları anlık işlemleri ve sorgulamaları yerine getirmektedir. Bu sistemler kurumun satın alma, envanter, üretim, muhasebe gibi günlük ve hız gerektiren işlemlerini yerine getirir. Veri ambarları ise veri analizi ve karar oluşturma gibi işlemlerde kullanıcılara hizmet verir. Bu sistemler farklı kullanıcıların farklı isteklerine cevap verebilmek için verileri farklı biçimlerde de

düzenleyip sunabilirler. Bu iki sistemin arasındaki başlıca ayrımları aşağıdaki şekilde açıklanabilir:

- İşlevsel veri tabanları müşteri tabanlı sistemlerdir. Kullanıcılar tarafından günlük işlemler ve sorgulamalar için kullanılırlar. Veri ambarları ise pazar tabanlı sistemlerdir. Analistler, müdürler tarafından uzun süreli veri analizi için kullanılır.

- İşlevsel veri tabanları günlük verileri kullanırlar. Bu veri tabanlarındaki veriler detaylı bir şekilde saklanmaktadır ve karar alma sürecinde çok kolaylıkla kullanılırlar. Veri ambarları yüksek oranda bir hacme ulaşmış ve güncelliğini yitirmiş verileri saklarlar. Geleceğe yönelik öngörüler elde etmek amacıyla kullanılırlar.

- İşlevsel veri tabanları sadece bir kurumun yada bir bölümün güncel verileriyle ilgilenir, farklı kurumlardan gelebilecek veriler yada güncel olmayan eski verilerle ilgilenmez. Veri ambarlarında ise farklı kurumlardan, farklı veri kaynaklarından gelebilecek verilerle de ilgilenir.

- İşlevsel veri tabanları okuma/yazma amaçlı, veri ambarları ise sadece okuma amaçlı kullanılırlar.

## 1.4. VERİ TABANLARINDA BİLGİ KEŞFİ

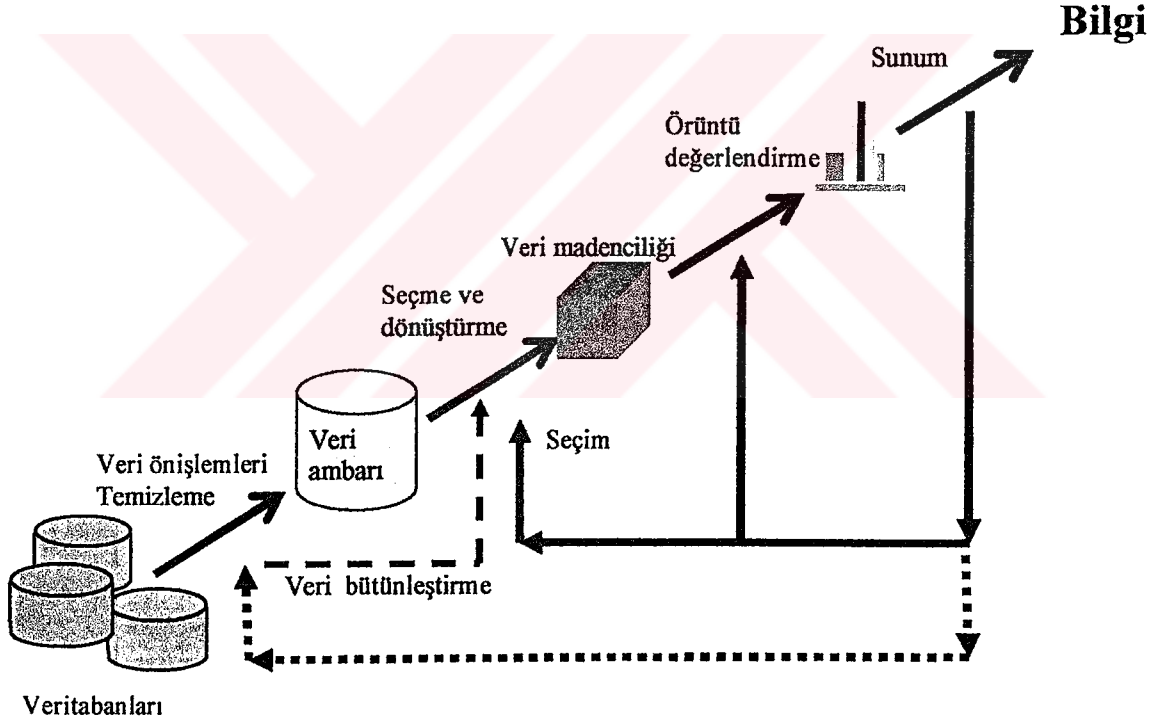
Veri madenciliği terimi en çok istatistikçiler, veri analistleri ve yönetim bilgi sistemleri grupları içerisinde kullanılmaktadır. Bu terim, veritabanları ile uğraşan gruplar tarafından da sıkça kullanılmaya başlanmıştır.

Veri tabanlarında bilgi keşfi (VTBK) terimi ilk olarak **Piatetsky-Shapiro** tarafından **Knowledge Discovery in Databases (KDD)** atölye çalışmasında

*(workshop)* 1989 yılında ortaya atılmıştır.[20] Bu terim yapay zeka ve makine öğrenmesi *(machine-learning)* alanlarında popüler olmuştur.

Veri tabanlarında bilgi keşfi terimini veri yığınları içerisinden yararlı bilgileri keşfetme ve çıkarma sürecinin tamamı olarak tanımlayabiliriz. Veri madenciliği ise bu sürecin önemli bir adımını göstermektedir.

Veri tabanlarında bilgi keşfi sadece veri madenciliğinin olduğu bir süreç değildir. Aşağıdaki tabloda gösterildiği gibi 5 aşamadan oluşan bir süreçtir.




**Şekil 3. Veri tabanlarında bilgi keşfi aşamaları**[21]

[20] Usama Fayyad, **"From Data Mining to Knowledge Discovery in Databases"**, S:37.
[21] J. Han, **"Data Mining Concepts and Techniques"**, s.6.

### 1.4.1. VERİ ÖN İŞLEMLERİ (DATA PREPROCESSING)

Bu aşamada seçilecek olan veri yığınları veri madenciliği aşamasında istenen modelin kurulması için önemlidir. Veri madenciliği aşamasında modelin kurulamaması durumunda, daha önceden seçilmiş olan veri yığınlarının yeniden düzenlenmesi gerekecektir.

Farklı kaynaklardan gelebilecek olan verilerin tek bir veri ambarında toplanabilmesi için genelleme, normalizasyon ve uyumluluk işlemleri yapılır.

Veri ön işlemleri aşaması toplama, değer biçme, birleştirme ve temizleme adımlarından meydana gelmektedir.

#### 1.4.1.1. TOPLAMA (COLLECTION)

İstenen amaca uygun olarak gerekli olan verilerin ve bu verilerin elde edileceği kaynakların belirlenmesi adımıdır. Verilerin elde edileceği kaynaklar ya kuruluşun kendi veri tabanlarıdır yada farklı kurumların veri tabanlarıdır.

#### 1.4.1.2. DEĞER BİÇME (ASSESSMENT)

Farklı kaynaklardan veri toplanması sebebiyle verilerin biçimleri arasında uyumsuzluklar olabilir. Bu uyumsuzlukların nedenleri arasında verilerin farklı zamanlara ait olması yada veri giriş biçimlerinin (bir veri tabanında cinsiyet bilgileri e/k olarak başka bir veritabanında 0/1 olarak girilmesi) kaynaklar arasında farklı olması gösterilebilir. Veriler üzerinde iyi sonuçlar alınabilmesi için bu veri uyumsuzluklarının giderilmesi gerekmektedir. Bu adımda veriler arasındaki uyum sağlanmaya çalışılmaktadır.

### 1.4.1.3. BİRLEŞTİRME VE TEMİZLEME (CONSOLIDATION AND CLEANING)

Veri kaynakları içinde alanlar içerisinde kayıp veriler olabilir yada bir önceki adımdan gelen ve halen çözülememiş sorunlar olabilir. Bu sorunlar mümkün olduğu ölçüde giderilmeye çalışılır ve tüm veriler bir veri tabanında toplanır.

## 1.4.2. VERİ SEÇME VE DÖNÜŞTÜRME (DATA SELECTION AND TRANSFORMATION)

Kullanılacak olan veri madenciliği algoritmasına göre verilerin gösteriliş şekilleri de önemli olmaktadır. Örneğin bir algoritmanın uygulanmasında değişken değerlerinin evet/hayır olması; başka bir algoritmada da değişken değerlerinin yüksek/orta/düşük olması kullanılan algoritmanın etkinliğini arttırabilir.

## 1.4.3. VERİ MADENCİLİĞİ (DATA MINING)

Veriler üzerinde anlamlı ve yararlı örüntüler çıkarabilmek için kullanılacak olan modelin ve algoritmanın belirlenmesi bu aşamada gerçekleşmektedir. VTBK sürecinin en önemli aşaması burasıdır. Hedefler doğrultusunda belirli veri madenciliği tekniği seçilir ve uygulanır.

## 1.4.4. ÖRÜNTÜ DEĞERLENDİRME (PATTERN EVALUATION)

Ortaya çıkarılan örüntülerin değerlendirilmesi bu aşamada yapılmaktadır. Yapılabilecek muhtemel değişiklikler ile daha önceki adımlar tekrarlanarak yeni çıkan durum tekrar değerlendirilebilir. Bu aşamada, ilginçlik *(interestingness)* ölçüm yöntemleri kullanılarak bulunan verilerin ne kadar yararlı olduğu tespit edilir.

### 1.4.5. BİLGİ SUNUMU (KNOWLEDGE PRESENTATION)

Farklı görselleştirme ve raporlama araçları kullanılarak elde edilen veriler ilgili kişilere sunulur.

VTBK süreci, kendi içerisinde birden fazla döngü yaparak veya aşamalar arasında atlamalar yaparak sürekli tekrarlanabilir. Günümüzde, VTBK sürecinin bir aşaması olan veri madenciliği daha çok ön plana çıkmaktadır, fakat sürecin diğer tüm aşamaları da veri madenciliği kadar bu sürecin önemli bir parçasıdır.[22]

## 1.5. VERİ MADENCİLİĞİ TEKNİKLERİ

### 1.5.1. SEPET ANALİZİ

Pazar sepeti analizi pazarlamada kullanılan en genel ve yararlı veri madenciliği tekniğidir.[23] Bu teknikte müşterilerin satınalma alışkanlıkları hakkında bilgi toplanmaktadır. Müşteriler mağaza içerisinde gezinirken hangi ürün gruplarını satın aldıkları ve hangi ürünlerin beraberce satın alındıkları bilgisine ulaşılabilmektedir. Bu bilgileri oluşturabilmek için müşterilerin mağazadaki satınalma işlemlerinde ortaya çıkan ve bilgisayar ortamında kayıt edilen veriler üzerinde pazar sepeti analizi yapılmaktadır.

Pazar sepeti analizi, satış verilerinden hareket ederek müşterilerin davranışlarını analiz eder. Çıkardığı sonuçları etkili promosyon ve reklam kampanyası için kullanır.[24]

---

[22] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, R. Uthurusamy, **"Advances in Data Mining and Knowledge Discovery"**, MIT Pres, 1994.
[23] G. Linoff, M.J.A. Berry, **"Data Mining Techniques for Marketing Sales and Customer Support"**, Wiley Computer Publishing, New York, USA, 1997.
[24] Hua Zhu, **"On-Line Analytical of Association Rules"**, University of Science and Technology of China, 1995.

Sepet analizi tekniği daha önceden veri tabanlarında tutulmuş olan veri yığınlarından hareketle hangi ürünlerin hangi ürünlerle beraber satıldığını, hangi ürünlerin promosyona girmesi gerektiğini gibi bilgileri ortaya çıkarır.[25]

Sepet analizi tekniği daha yoğun olarak satış alanında kullanılsa da bunun dışında birçok alanda daha uygulanmaktadır.

- Kartlar ile yapılacak olan alışverişlerin işlenmesi ve müşterilerin yapacakları potansiyel harcama kalemlerinin bulunması
- Sigortacılıkta ortaya çıkan değişik tarzdaki işlemlerin dolandırıcılık olup olmadığının belirlenmesi ve yapılacak soruşturmaya ışık tutması
- Hastaların sağlık kayıtlarından hareketle önerilen çeşitli tedavi kombinasyonlarından doğan farklı gelişmelerin görülmesi
- Reklam stratejilerinin belirlenmesi.
- Mağaza içerisinde oluşturulacak olan raf tasarımının satışları arttıracak şekilde oluşturulması.

Analiz sonucunda satın alınan ürün grupları akla uygun sonuçlar verebilir. Örneğin, bir fast-food lokantasında hamburger ile kola beraber satın alınmıştır. Bazı durumlarda ise tam aksine, akla uygun sonuçlar vermeyebilir. Örneğin, perşembe günleri çocuk bezi ile bira beraberce satın alınmaktadır. Bu durumu incelediğimiz zaman, evli çiftlerin hafta sonunu evde geçirmek istemeleri ve bu süre içerisinde de rahatlarını bozmamak için gerekli olabilecek muhtemel bira ve çocuk bezini haftasonu gelmeden almak istemeleri sonucu ortaya çıkmıştır.

Pazar sepeti analizini yerine getirebilmek için öncelikle satış işlemlerinin listesine sahip olunması gereklidir ve her bir işlemde nelerin satın alındığı bilinmesi gereklidir. Aşağıda bununla ilgili bir örnek verilmiştir.

---

[25] Ahmet Cüneyd Tantuğ, **"Veri Madenciliği ve Demetleme"**, Yüksek Lisans Tezi, İTÜ, Mayıs 2002.

| İşlem No: | Satın Alınan Ürünler |
|---|---|
| 1 | Soğuk pizza, kola, süt |
| 2 | Süt, patates cipsi |
| 3 | Kola, soğuk pizza |
| 4 | Süt, kraker |
| 5 | Kola, kraker |

**Tablo 2. Örnek satış kayıtları**

Tabloda her müşteri farklı bir sepet oluşturmakta ve gruplar arasında bir ilişki bulunmamaktadır. Pazar sepeti analizinde ilk adım verileri, ürünlerin birbirleriyle olma sıklığına göre bir tablo içine yerleştirmektir. Buna göre aşağıdaki şekilde bir tablo oluşturulur.

| | Soğuk pizza | Süt | Kola | Patates cipsi | Kraker |
|---|---|---|---|---|---|
| **Soğuk pizza** | 2 | 1 | 2 | 0 | 0 |
| **Süt** | 1 | 3 | 1 | 1 | 1 |
| **Kola** | 2 | 1 | 3 | 0 | 1 |
| **Patates cipsi** | 0 | 1 | 0 | 1 | 0 |
| **Kraker** | 0 | 1 | 1 | 0 | 2 |

**Tablo 3. Ürün birliktelik tablosu**

Bu birliktelik tablosu ilk bakışta şunları göstermektedir:

- Soğuk pizza ve kolanın birlikte satılma olasılığı diğer tüm ikili ürün gruplarından daha fazladır.

- Soğuk pizza hiçbir zaman patates cipsi yada kraker ile beraber satılmamıştır.

Tablodaki 5 kayıttan 2'sinde soğuk pizza ile kola birlikte satılmıştır. "Eğer soğuk pizza ise kola alır" kuralının doğru olma ihtimali %40'dır. Ayrıca soğuk pizza

içeren tüm kayıtlarda kolanında alınmış olması çok yüksek oranda güvenilirlik sağlamaktadır. Bu yüzden "eğer soğuk pizza ise kola" kuralının güvenilirliği %100'dür. Bu kuralın tersi için ise "eğer kola ise soğuk pizza" güvenilirlik değeri o kadar yüksek değildir. Bu kuralın güvenilirliği ise %66'dır.

Pazar sepeti analizinin başarılı olduğu noktalar:

- Kolay ve anlaşılır sonuçlar üretir.
- Değişik boyutlardaki veriler üzerinde çalışabilir.
- Analiz için gerekli olan adımlar diğer yöntemlere göre (genetik algoritmalar, yapay sinir ağları vb.) daha basittir.

Pazar sepeti analizinin başarısız olduğu noktalar:

- Boyut büyüdükçe gerekli hesaplamalar üstel olarak artmaktadır.
- Pazar sepeti analizinde kullanılacak ürün gruplarının belirlenmesi sırasında biraz bilgi kaybı olabilir fakat bu durum analizin boyutlarını küçültebilir.
- Kayıtlarda çok az rastlanan ürünleri yok sayar. Bu teknik en doğru sonucu, tüm ürünlerin kayıtlar içinde aynı frekansta göründüğü durumlarda üretmektedir.

## 1.5.2. KÜMELEME ANALİZİ (CLUSTER ANALYSIS)

Kümeleme analizi nesneler setinin benzer özellikte olanlarının aynı kümeler içinde toplandığı bir gruplama sürecidir. Bir küme, aynı küme içindeki diğer nesneyle benzer özellikleri gösteren nesneler topluluğudur. Benzer özellikleri göstermeyen nesnelerde farklı kümelerde gruplanmaktadır.[26]

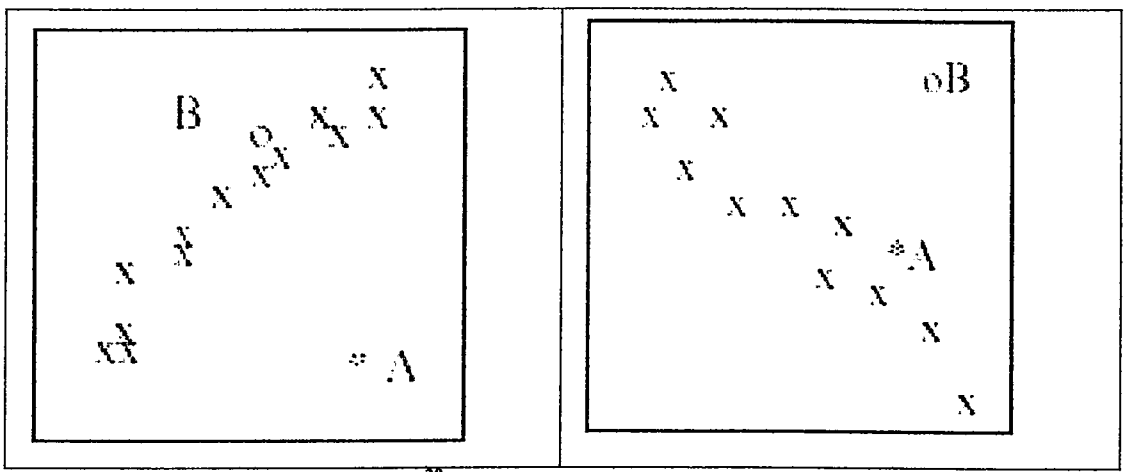
[26] Han, a.g.e., s.380.

Analiz süreci başlamadan önce veriler belli sınıflara ayrılmazlar, veriler dağılımlarına göre oluşan sınıflara ayrılmaktadır. Kümeleme analizinde temel amaç küme içi benzerliği maksimum yapmak, kümeler arası benzerliği ise minimum yapmaktır.

Kümeleme analizi veri madenciliği yanında pazarlama, görüntü işleme, örüntü tanıma, biyoloji, coğrafi bilgi istemleri, web doküman sınıflandırma gibi pek çok farklı alanda kullanılmaktadır. Kümeleme analizi tezin 2. bölümünde ayrıntılı olarak açıklanmıştır.

## 1.5.3. SIRADIŞI ANALİZİ (OUTLIER ANALYSIS)

Veri madenciliği tekniklerinin kullanıldığı veri setlerinin içinde diğer verilerden göze çarpıcı bir şekilde ayrılan veriler bulunabilir. Bu verilerin varlığı ortaya çıkacak sonuçlar üzerinde değişikliklere sebep olur. Sıkça görülen bir davranışı göstermeyen veya veri modeline uymayan, diğer verilerden çok daha fazla farklılık gösteren verilere sıradışı veri adı verilir. Bu tür verileri ortaya çıkarma sürecine de sıradışılık analizi denir.[27]



**Şekil 4. Sıradışı verilerin görünümü**[28]

[27] A.e., s.381.

[28] Charu C. Aggarwal, Philip S. Yu, **"Outlier Detection for High Dimensional Data"**, IBM T.J. Watson Research Center, NY, 2001.

Sıradışı veriler okuma, kayıt etme, ölçüm, uygulama veya hesaplama sırasında oluşan hatalardan dolayı oluşmaktadır. Örneğin, bir insanın yaşı programa girilirken 44 yerine 445 olarak yazılabilir. Bu durumda kullanıcının veriyi yanlış bir şekilde girmesiyle veri, diğer verilerden ayrılacak ve sıradışı durumuna düşecektir. Alternatif bir şekilde sıradışılık doğal bir verinin sonucu olarak da ortaya çıkabilir.Örneğin, firmanın icra kurulu başkanının ücreti diğer çalışanlarının ücretlerine göre doğal olarak çok farklı görülebilir.

Veri madenciliği algoritmalarının çoğu sıradışı verilerin etkisini minimuma indirmeyi veya tamamen ortadan kaldırmayı amaçlamaktadır.[29]

Sıradışı analizi, n adet veri ve k adet beklenen sıradışı veri olmak üzere, diğer verilerden oldukça farklı veya tutarsız olan k verinin bulunmasıdır.[30] Sıradışı analizinde 2 alt problem ortaya çıkmaktadır. Birincisi, veri setindeki hangi verilerin tutarsız olarak kabul edileceği, ikincisi tanımlanan şekilde sıradışı verileri ortaya çıkaracak etkili metodun bulunmasıdır.

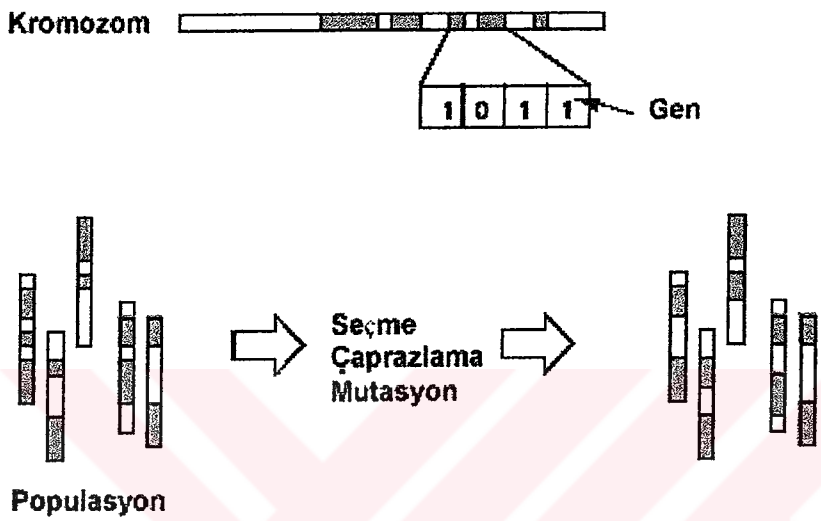
Sıradışı analizi geniş bir uygulama alanına sahiptir. Kredi kartlarının olağandışı kullanımının tespiti, telekomünikasyon servislerindeki olağandışılığın tespiti gibi dolandırıcılık tespitinde kullanılmaktadır. Ayrıca uç noktalardaki düşük ve yüksek gelire sahip müşterilerin harcama davranışlarını belirlemek için, çeşitli tıbbi tedavilerde olağandışı sonuçları bulmak için kullanılmaktadır.[31]

## 1.5.4. GENETİK ALGORİTMALAR (GENETIC ALGORITHMS)

Biyolojik işlemlerden kaynağını alan bir arama ve optimizasyon tekniği olan genetik algoritmalar karmaşık ve zor problemlerin hızlı ve kolay bir şekilde çözümlenmesinde etkili bir yöntemdir. Genetik algoritmaların temel ilkeleri 1960' lı

[29] A.e. s.381.
[30] A.e. s.381.
[31] Aggarwal, 2001.

yıllarda John Holland tarafından ortaya konmuştur. Olasılık kurallarına göre çalışan genetik algoritmalar, yalnızca amaç fonksiyonuna gerek duyarlar. Çözüm uzayının tamamını değil sadece belirli bir kısmını incelerler. Bu sayede etkin bir arama yaparak çözüme kısa sürede ulaşırlar.[32]




**Şekil 5. Genetik Algortiması**[33]

Genetik algoritmalar problemlerin çözümünde etkili bir yöntem olduğundan dolayı pek çok uygulama alanında kullanılmıştır. Bazı uygulama alanlarını şunlardır:

*a) Optimizasyon:*

Genetik algoritmaların kullanıldığı optimizasyon problemlerini fonksiyon ve birleşi (combinatorial) optimizasyonu olarak iki sınıfa ayırabiliriz.

Problemlerde optimize edilecek olan amaç fonksiyonunun süreksiz olması durumunda süreksiz olan noktalarda fonksiyonun türevi alınamayacaktır. Bu durumda da türeve dayalı optimizasyon algoritmaları kullanılamaz. Genetik

[32] D.E. Goldberg, "**Genetic Algorithms in Search, Optimization and ,machine Learning**", USA, 1989.
[33] A.e.

algoritmalar ise türeve ihtiyaç duymadığından bu tip problemlerin çözümünde önemli bir üstünlüğe sahiptir.[34]

Birleşi optimizasyon problemlerin de ise istenen amaçlara ulaşmak için sınırlı kaynakların etkin tahsisi hedeflenmektedir. Bu sınırlı kaynaklar genellikle işgücü, zaman, tedarik veya finans ile ilgilidir. Gezgin satıcı problemi, yerleşim tasarımı problemi, atama problemi gibi örneklerde bu optimizasyon sınıfı kullanılmaktadır.[35]

*b) Finans:*

Genetik algoritmalar amaç fonksiyonu odaklı olduğundan dolayı finans problemlerinde hisse senedi fiyatlarındaki değişim kalıplarını tahmin etmede ve kaynak tahsisi belirleme ve sermaye tahsisi planlarını oluşturmada kullanılmaktadır. Bununla birlikte müşteri kredi değerliliğini analiz etme, kredi kartları puanlama, piyasalar ile ilgili tahminler oluşturma gibi sıklıkla uygulandığı alanlarda bulunmaktadır.

*c) Pazarlama:*

Veri madenciliği teknikleri içerisinde önemli bir yere sahip olan genetik algoritmalar pazarlama alanında da çok fazla kullanılmaktadır. Tüketici verilerini analiz etmek, tüketici verilerinden yola çıkarak pazarlama stratejileri oluşturmak, müşteri profillerini ortaya çıkarmak pazarlama içindeki önemli uygulama alanlarıdır.

---

[34] C.L. Karr, M.L. Freeman, **"Industrial Applications of Genetic Algorithms"**, CRC Press, USA, 1999.

[35] Gül Gökay Emel, Çağatan Taşkın, **"Genetik Algoritmalar ve Uygulama Alanları"**, Bursa, Uludağ Üniversitesi İİBF Fakültesi Dergisi Cilt XXI, Sayı:1, S:129-152, 2002.

*d) Üretim:*

Üretim alanı içindeki çok farklı alanlarda yoğun olarak genetik algoritmalar kullanılmaktadır. Montaj hattı içinde her iş istasyonundaki toplam işlem zamanını minimize etmeyi amaçlayan problemin çözümü, tesis yerleşimi problemlerinde kaynakların belirli kısıtlar içinde optimum performans sağlayacak şekilde yerleşimi genetik algoritmalar ile gerçekleştirilebilmektedir.Yine aynı şekilde n elemanın n farklı göreve atanması, üretim sistemlerinin verimliliğinin arttırılması problemlerinde genetik algoritmalar kullanılmaktadır.

## 1.5.5. YAPAY SİNİR AĞLARI (NEURAL NETWORKS)

Yapay sinir ağları, biyolojik sinir sisteminden faydalanarak oluşturulan modellerdir. 1943 yılında bir nörobiyolojist olan Warren McCulloch ve bir istatistikçi olan Walter Pitts 'in "*Sinir Aktivitesindeki Düşüncelere Ait Bir Mantıksal Hesap*" başlıklı makalesiyle yapay sinir ağları hakkındaki ilk çalışmalar başlamıştır. John Von Neumann ve Marvin Minsky ile devam eden yapay sinir ağları çalışmaları Kohonen, Grossberg ve Hopfield gibi araştırmacıların katkılarıyla hızla gelişmiştir. 1987 yılında yapılan ilk yapay sinir ağları sempozyumundan sonra yapay sinir ağları uygulamaları yaygınlaşmıştır.

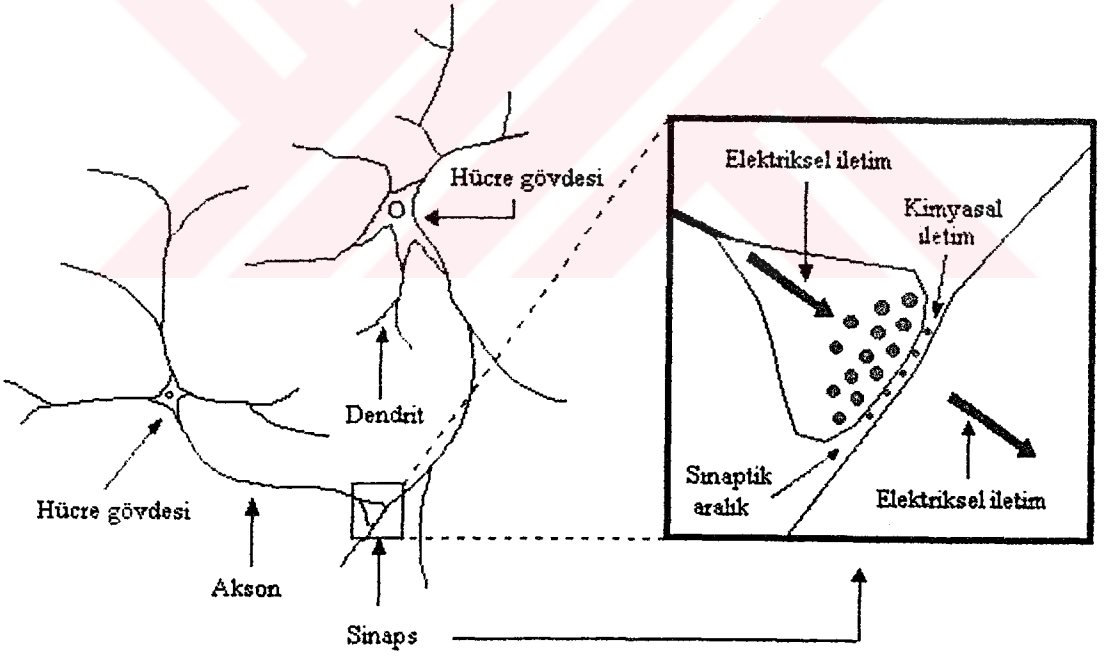
Biyolojik sinir sistemi bilgiyi alan, yorumlayan ve uygun bir karar üreten beyinin bulunduğu 3 katmandan oluşmaktadır. Alıcı sinirler, iç veya dış ortamlardan algıladıkları uyarıları elektriksel sinyallere dönüştürerek beyne iletirler. Tepki sinirleri ise beynin ürettiği elektriksel sinyalleri çıktı olarak uygun tepkilere dönüştürürler. Sinir hücrelerine *nöron* adı verilir. Nöron, gövde(soma), gövdeye giren sinyal alıcı lifler(dentrit), gövdeden çıkan sinyal iletici lifler(akson) olmak üzere 3 kısımdan oluşur.[36] Yapay sinir ağları insan beyninin çalışma prensiplerini

[36] Şeref Sağıroğlu, **"Şifrelemede Nörol Yaklaşımlar"**, (Çevrimiçi) http://arsiv.emo.org.tr/Kartus01/SEMPOZYUMLAR/iletisimteknolojilericalistayi/makale_pdf/24.pdf , 17.04.2004.

örnek olarak alınıp geliştirildiğinden dolayı yapısal olarak benzerlikler taşımaktadırlar. Bu benzerlikler aşağıdaki tablo 4.'de verilmektedir.

| SİNİR SİSTEMİ | YSA SİSTEMİ |
|---|---|
| Nöron | İşlem elemanı |
| Dendrit | Toplama fonksiyonu |
| Hücre gövdesi | Transfer fonksiyonu |
| Aksonlar | Eleman çıkışı |
| Sinapslar | Ağırlıklar |

**Tablo 4. Sinir sistemi ile YSA' nın benzerlikleri.**

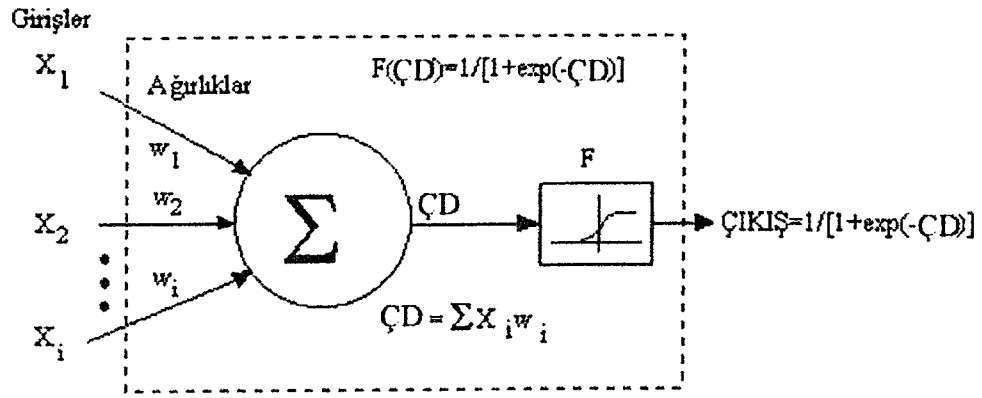



**Şekil 6. Nöron yapısı**[37]

Yapay sinir ağları içindeki nöron yapısına gelen girişler dış kaynaklardan veya diğer işlem elemanlarından gelen işaretlerdir. Bu girişler kuvvetli veya zayıf

[37] Haldun Akpınar, **"Yapay sinir ağları gelişimi ve yapılarının incelenmesi"**, İ.Ü. İşletme Fakültesi Dergisi, C:23, S:1 /Nisan 1994, s.41-78.

olabilir. Buna göre de ağırlıkları da farklılık gösterecektir. YSA' da giriş değerlerine önce toplama fonksiyonu uygulanır ve bir çıkış değeri (İEÇ) bulunur. Bu çıkış değeri öğrenme eğrisine uygulanır. Sonuçta ortaya çıkan değer ağın çıkış değeri olabilir.




**Şekil 7. Yapay Sinir Ağı** [38]

Yapay sinir ağları, birçok farklı alanda uygulanmaktadır.Özellikle mühendislik problemlerinin çözümünde sıkça kullanılmaktadır. Yapay sinir ağlarının kullanıldığı bazı alanlar aşağıda sıralanmaktadır.[39]

- *Endüstriyel Uygulamalar*

  Kimyasal proseslerin dinamik modellenmesi, cep telefonlarında ses ile çalışabilme, otomobillerde otomatik rehber sisteminin geliştirilmesi, gezgin satıcı problemi, işlerin makinalara atanması ve çizelgeleme.

- *Finansal Uygulamalar*

  Makro ekonomik tahminler, borsa benzetim çalışmaları endekslerinin tahmin edilmesi, kredi kartı hilelerinin tespiti, banka kredilerinin değerlendirilmesi, döviz kuru tahminleri, risk analizleri.

---

[38] A.e.

[39] Prof. Dr. Ercan ÖZTEMEL, **"Yapay Sinir Ağları"**, İstanbul, Papatya Yayıncılık, Ağustos 2003, s.203-206.

- *Askeri Uygulamalar*

  Hedef tanıma ve takip sistemleri, yeni sensörlerin performans analizleri, radar ve görüntü işleme, mayın detektörleri.

- *Sağlık Uygulamaları*

  Solunum hastalıklarının teşhisi, tıbbi resim işleme, üroloji uygulamaları, hastalıkların teşhisi ve resimlerden tanınması, EEG ve ECG analizleri.

# BÖLÜM 2

# KÜMELEME ANALİZİ

## 2.1. KÜMELEME ANALİZİ TANIMI

Büyük hacimlere sahip olan veri setlerini homojen olarak ayırma işlemi, veri madenciliği içerisindeki önemli işlemlerden birisidir. Kümeleme analizi, veriler setinin benzer özellikli olanlarının aynı sınıflar içinde toplandığı gruplama sürecidir.[40] Bir başka şekilde ifade edersek kümeleme analizi, aynı karakteristik özelliklere sahip olan bireylerin saptanması ve benzerliklerine göre sınıflandırılmasını (gruplandırılmasını) sağlayan çok değişkenli bir analiz tekniğidir.[41] Ortaya çıkan grupların her birine küme adı verilmektedir. Bir küme, aynı küme içindeki diğer verilerle benzer özellikler gösteren veriler topluluğudur ve benzer özellikleri göstermeyen verilerde diğer kümelerde bulunmaktadır. Kümeleme analizi ile yoğun ve seyrek bölgeler tanımlanabilir ve bundan dolayı verinin özelliklerine bakarak ilginç korelasyonlar ve dağılım örüntüleri ortaya çıkabilir.

Kümeleme analizi, gözetimsiz sınıflama (unsupervised classification) yöntemidir ve önceden tanımlanmış sınıflandırma yöntemi değildir. Gözetimsiz sınıflamada temel amaç, başlangıçta verilen ve henüz sınıflandırılmamış bir kümeyi anlamlı alt kümeler oluşturacak şekilde gruplamaktır. Kümeleme işlemi yeni gelen verinin özelliklerine göre şekillenir. Gözetimli sınıflandırma yönteminde ise bize verilen veriler önceden sınıflandırılmışlardır.

---

[40] Han, s. 348.

[41] Gülçin Tunalı MENTEŞ, **"Faktör ve Kümeleme Analizi Yardımıyla Bankacılık Ürün ve Hizmetlerinin Araştırılması Üzerine Bir Uygulama"**, Doktora Tezi, İstanbul, İ.Ü.Sosyal Bilimler Enstitüsü, 2000, s.55.

Kümeleme analizi, istatistiğin bir dalı gibi uzun yıllar farklı alanlarda kullanıldı. S-Plus, SPSS ve SAS gibi istatistik analiz yapan yazılım paketlerinde K-Means ve K-Medoids gibi kümeleme yöntemleri yoğun olarak kullanılmaktadır. Kümeleme analizinin kullanıldığı alanlar şu şekildedir:

- *Pazarlama:* Kümeleme analizi, müşterilerin ayrı gruplara ayrılmasını ve satınalma örüntülerine bakarak karakterlerine göre gruplanmasını sağlar.

- *Biyoloji:* Biyoloji alanında bitki ve hayvan gruplarını ortaya çıkarılması benzer fonksiyonları gösteren genlerin kategorize edilmesinde kullanılır.

- *Coğrafya:* Konumsal verilerden yararlanarak bölgeler arasındaki benzerliklere göre bölgelerin gruplandırılmasında ve yerleşim yerlerine götürülecek mal ve hizmetler için ideal yerler belirlemede kullanılır.

- *İnternet:* Web üzerindeki belgelerin sınıflandırılmasında kullanılmaktadır.

- *Şehir Planlaması:* Konumlarına, değerlerine ve türlerine göre binaların gruplandırılmasında kullanılmaktadır.

- *Deprem Araştırmaları*:Sürekli fay hatları belirlenerek deprem merkezlerinin gruplandırılmasında kullanılmaktadır.

- Önceden tespit edilmiş kümeler üzerinde sınıflama yapma ve kategorize etme gibi işlemler yapan diğer algoritmalar için bir ön işlem adımını sağlar.

## 2.2. KÜMELEME ANALİZİNİN GEREKSİNİMLERİ

- *Ölçeklenebilirlik:* Çoğu kümeleme algoritmaları 200 veriden daha az veri içeren veri setlerinde iyi çalışmaktadır. Bununla beraber büyük bir veri

tabanında milyonlarca veri bulunabilir. Kümeleme algoritmaları bu büyüklükteki veri tabanlarında da uygulanabilmelidir.

- *Farklı Özelliklerle Uygulanabilme:* Kümeleme algoritmalarının birçoğu sayısal verileri kümeleme için tasarlanmıştır. Bazı uygulamalarda ikili, nominal, düzenli ve karışık veri tipleri kullanılmaktadır ve bu gibi farklı veri tiplerin de kümeleme analizi kullanılması gerekebilir. Kümeleme algoritmaları farklı veri tiplerinde de uygulanabilmelidir.

- *Farklı Biçimdeki Kümelerin Keşfi:* Kümeleme algoritmalarının çoğu Öklid veya Manhattan uzaklık ölçüleri temelinde kümeleri belirlerler. Bu ölçüleri kullanan algoritmalar küçük boyutlu veya yoğunluklu küresel kümeleri bulmaya yönelirler fakat bir küme farklı şekilde olabilir. Önemli olan keyfi şekillerdeki kümeleri tespit edebilen algoritmaların geliştirilebilmesidir.

## 2.3. KÜMELEME ANALİZİ VERİ TÜRLERİ

Kümeleme analizinde veriler matris formuna getirilir. Matris formu bilgisayar ortamında hesaplama yapabilmek için en uygun veri yapısıdır. Kümeleme işleminde temel olarak iki matris grubu kullanılır.[42]

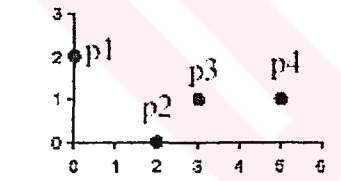
- *Veri Matrisi (Data Matrix):*

Veriler çok boyutlu uzayda bir nokta olarak temsil edilirler. Her boyut ayrı bir özelliği temsil etmektedir. Veri setleri m satır n sütun ile ***(m x n)*** boyutunda bir veri matrisi oluşturur. **m** her veriyi , **n** ise her özelliği temsil etmektedir. Veri, kullanılmadan önce veri matrisine çevrilmektedir. Bu çevrim işleminin bir sebebi de farklı ölçekler üzerinde farklı özellikler hesaplanabilmektedir. (Santimetre, kilogram gibi.)

[42] Han, s.338.

- *Yakınlık Matrisi (Proximity Matrix):*

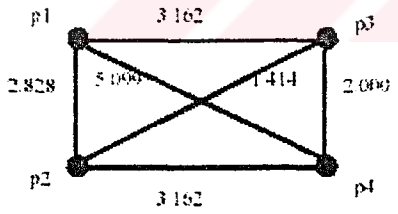
Kümeleme algoritmalarının çoğu (**S**) benzerlik matrisi (similarity matrix) ve (**D**) farklılık matrisini (dissimilarity matrix) kullanır. Her iki matris genellikle (**P**) yakınlık matrisi olarak ifade edilir. **P** yakınlık matrisi (**m x m**) lik tüm benzerlik ve farklılıkları içerir. $X_i$ ve $X_j$ , i. ve j. veriler olmak üzere yakınlık matrisin i. satır ve j. sütun girişi ($s_{ij}$) benzerlik değeri veya ($d_{ij}$) farklılık değeridir.[43] Kolaylık olması sebebiyle $s_{ij}$ veya $d_{ij}$ yerine $p_{ij}$ sembolü kullanılmaktadır. Aşağıdaki şekillerde veri matrisi ve yakınlık matrisi bir örnek üzerinde gösterilmektedir.




a) Noktalar

| point | x | y |
|---|---|---|
| p1 | 0 | 2 |
| p2 | 2 | 0 |
| p3 | 3 | 1 |
| p4 | 5 | 1 |

b) Veri Matrisi




d) Uzaklık Grafiği

| | p1 | p2 | p3 | p4 |
|---|---|---|---|---|
| p1 | 0.000 | 2.828 | 3.162 | 5.099 |
| p2 | 2.828 | 0.000 | 1.414 | 3.162 |
| p3 | 3.162 | 1.414 | 0.000 | 2.000 |
| p4 | 5.099 | 3.162 | 2.000 | 0.000 |

c) Yakınlık Matrisi

**Şekil 8. Dört nokta için veri ve yakınlık matrisi**[44]

[43] Michael Steinbach-Levent Ertöz-Vipin Kumar, **"The Challenges of Clustering High Dimensional Data"**, t.y.
[44] A.e.

## 2.4. KÜMELEME ANALİZİNDE KULLANILAN UZAKLIK ÖLÇÜLERİ

Kümeleme analizinde benzerlik ve farklılık kavramları kullanılmaktadır. Veriler arasındaki benzerlik veya farklılığın hangi şartlara göre belirlendiği önemli bir konudur. Benzerlik veya farklılığı belirlerken kullanılan değişken değerleri aynı standartlarda olmalıdır. Bir niteliği tanımlayan ölçüm değerleri aynı birimlerde olmalıdır. Örneğin uzunluk ölçüm verileri ile işlem yapılıyorsa verilerin bir bölümünün santimetre diğer bölümünün metre olması kümeleme analizinin başarısız olmasına sebep olabilir.

X veri matrisi, n veri ve p değişken sayısını göstermek üzere:

$$X=\begin{bmatrix} x_{11} \cdots x_{12} \cdots x_{1p} \\ \cdots \cdots \cdots \cdots \cdots \\ x_{21} \cdots x_{22} \cdots x_{2p} \\ \cdots \cdots \cdots \cdots \cdots \\ x_{n1} \cdots x_{n2} \cdots x_{np} \end{bmatrix}$$ Veri Matrisi

$x_i$ ve $x_j$ gözlem vektörleri arasındaki $d(x_i, x_j)= d_{ij}$ uzaklığının şu şartları yerine getirmesi gerekmektedir.

- $d(x_i, x_j) \geq 0$
- $d(x_i, x_j) = 0$; $i = j$ iken
- $d(x_i, x_j) = d(x_j, x_i)$
- $d(x_i, x_j) \leq d(x_i, x_1) + d(x_1, x_j)$

(n x n) uzaklık matrisi aşağıdaki şekilde gösterilmektedir.

$$D_{ij} = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{1n} \\ d_{21} & 0 & d_{23} & d_{2n} \\ d31 & d_{32} & 0 & d_{3n} \\ : & : & : & \\ d_{n1} & d_{n2} & d_{n3} & \ldots \quad 0 \end{bmatrix}$$

$$d(x_i, x_j) = d(x_j, x_i)$$

Eşitliği olduğu için uzaklık matrisinin altında kalan sayılar ile üstünde kalan sayılar simetriktir. Bu yüzden uzaklık matrisine tek yönlü (one-mode) matris denir.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

**Şekil 9. Tek yönlü matris**[45]

Kümeleme sürecinde verileri gruplarken verilerin ne kadar benzer yada farklı olduğunu tayin edebilmek için birtakım uzaklık ölçülerine ihtiyaç vardır. En sık kullanılan yöntem veri çiftleri arasında ki mesafe bakımından benzerliğin ölçülmesidir. Birbirleriyle arasındaki uzaklığın az olduğu veriler birbirlerine daha çok benzerken, birbirinden uzak olanlar daha az benzerler.

Uzaklık veya komşuluk mesafesini hesaplamada aşağıdaki uzaklık formülleri kullanılır.

[45] W. Ras Zbigniew, **"ITCS 6155/8155-900 Knowledge-Based Systems"**, 2002, s.1.

1. *Öklit Uzaklığı (Euclidian Distance):*

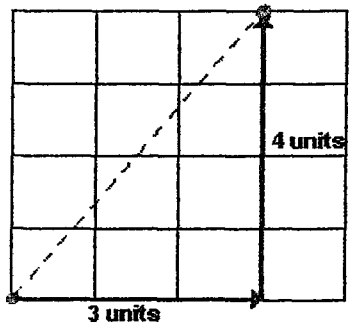
Uzaklık ölçüleri içerisinde en çok kullanılan uzaklık ölçüsüdür. İki boyutlu düzlemlerde kolaylıkla kullanılabilir. Boyut sayısı arttıkça hesaplama süresi de artmaktadır.

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2}$$

Formüldeki i = ($x_{i1}$, $x_{i2}$,..., $x_{ip}$ ) ve j = ($x_{j1}$, $x_{j2}$,...,$x_{jp}$) değerleri **p** boyutlu veri nesnelerini temsil etmektedir. Öklit uzaklık ölçüsü, değişkenlerin birbirinden bağımsız olduğunu varsaymaktadır fakat çoğu veri matrisinde değişkenler arasında önemli korelasyonlar vardır. Öklit uzaklık ölçüsünün uygulanabilmesi için verilerin en azından aralıklı ölçekte ölçülmüş olması gerekir. Aralık ölçekli değişkenler doğrusal bir ölçek üzerinde temsil edilebilen değişkenlerdir.

2. *Manhattan Uzaklığı (Manhattan Distance):*

Öklit uzaklığı ile benzerlikler taşımaktadır. Manhattan uzaklığı, iki vektörün toplamıdır.




**Şekil 10. Manhattan Uzaklık ölçüsü.**[46]

[46] (Çevrimiçi): http://ucl.ac.uk/oncology/MicroCore/HTML_resource/Distance_detailed_popup.htm, Erişim tarihi 22.03.2004.

3. *Minkowski Uzaklığı (Minkowski Distance):*

Minkowski uzaklık ölçüsü yukarıda anlatılan diğer iki uzaklık ölçüsünün genel halidir. Formülde yer alan q bir tamsayı olmak üzere q=1 için Manhattan uzaklığını ifade etmektedir. q=2 için ise Öklit uzaklığını belirtmektedir.

$$d(i,j) = (|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + ... + |x_{ip} - x_{jp}|^q)^{1/q}$$

q değişkenin değeri artmaya başladıkça daha hassas uzaklık ölçümleri elde edilebilir.

4. *Mahalanobis Uzaklığı (Mahalanobis Distance):*

Diskriminant analizi içerisindeki olayların analizinde kullanılır. Mahalanobis uzaklığı, özellik uzayındaki (attribute space) her grubun merkez noktası ile veri arasındaki uzaklıktır. Veri her grup için bir mahalanobis uzaklık değerine sahip olacaktır. En küçük uzaklık değerine sahip olan gruba veri eklenecektir. Böylece uzaklığa göre en çok benzer olduğu kümeye dahil olmuş olacaktır.

$$D_m^2 = (x_m - \bar{x})' S^{-1} (x_m - \bar{x})$$

m : örnekteki eleman sayısı ,

$x_m$ : m 'ninci değer ,

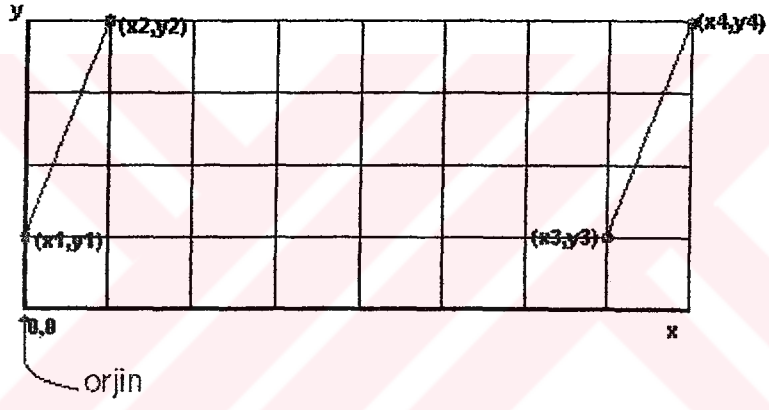
$\bar{x}$ : ortalama vektörü ,

S : örneğin varyans-kovaryans matrisi göstermektedir.

5. *Canberra Uzaklığı (Canberra Distance):*

Canberra uzaklık ölçüsü, benzer düzlemsel değerler arasındaki oranların düzeltilmesini yapmaktadır. Bu uzaklık ölçüsü, iki nokta arasındaki uzaklık değeriyle beraber bu noktaların orjin noktasıyla olan ilişkisini de dikkate alır.

Canberra uzaklık ölçüsü, biyoinformatik (anormal durumların tespiti), bilgisayar saldırı tespiti (*computer intrusion detection*), yapay sinir ağları (*neural networks*) gibi alanlarda kullanılmaktadır.[47]




**Şekil 11. Canberra uzaklık ölçüsü.** [48]

Yukarıdaki şekilde ( $x_1$ , $y_1$ ) ve ( $x_2$ , $y_2$ ) noktaları arasındaki Canberra uzaklığı 0.6 olarak bulunmuştur. Bununla beraber ( $x_3$ , $y_3$ ) ve ( $x_4$ , $y_4$ ) noktaları arasındaki uzaklık 0.666 olarak hesaplanmıştır. Aslında geometrik olarak düşünüldüğünde bu iki uzaklık değeri aynıdır. Bu farklılık çok boyutlu uzayda daha açık görülmektedir.

[47] (Çevrimiçi): http://ucl.ac.uk/oncology/MicroCore/HTML_resource/Distance_detailed_popup.htm, Erişim tarihi 22.03.2004.
[48] A.e.

## 2.5. KÜMELEME İŞLEMİ SÜRECİ

### 2.5.1. ÖRÜNTÜ SEÇİMİ

Bu adım kapsamında demet sayısının belirlenmesi, örüntü kümesinin büyüklüğünün belirlenmesi sağlanmaktadır. Bunun yanında kümeleme analizinde kullanılacak algoritma için kayıt tipleri ve ölçeklerinin seçimi bu adımda belirlenmektedir.

### 2.5.2. ÖZELLİK SEÇİMİ (FEATURE SELECTION)

Kümeleme işleminde en uygun sonuçları verecek özelliklerin seçimi ve kümeleme algoritmasında kullanılmak üzere en uygun özelliklerin (kayıt alanlarının) seçimi bu adımda yapılmaktadır. Kümeleme süreci bir veri ambarı üzerinde gerçekleşiyorsa bu durumda farklı tablolar üzerinde bulunan özellikler (kayıt alanları) bir araya getirilerek kümeleme için kullanılacak verileri oluştururlar.

### 2.5.3. BENZERLİK YÖNTEMİ SEÇİMİ

Kümeleme analizinde en önemli konulardan biri, veriler arasındaki benzerliklerin yada farklılıkların ölçülmesinde kullanılacak olan uzaklık fonksiyonunun belirlenmesidir. Farklı çevreler tarafından farklı uzaklık fonksiyonları kullanılmıştır.[49] Veriler arasındaki uzaklıkların belirlenmesinde en çok kullanılan uzaklık fonksiyonu Öklit uzaklık *(Euclid Distance)* fonksiyonudur. Bunun yanında yukarıda anlatılanlar gibi pek çok farklı uzaklık fonksiyonu da bulunmaktadır.

[49] A.K. Jain, R.C. Dubes, **"Algorithms for Clustering Data"**, Prentice-Hall Advanced Reference Series, Prentice-Hall Inc., New Jersey, 1988.

## 2.5.4. KÜMELEME İŞLEMİ

Bu adımda büyük hacimlerde bulunan veri setleri benzer özelliklerine göre ayrı kümelere ayrılırlar. Benzer karakteristikleri gösteren veriler aynı grup içerisinde toplanırlar. Kümeleme işleminde kullanılan farklı kümeleme teknikleri bulunmaktadır.

Kümeleme teknikleri hiyerarşik olabilir. Hiyerarşik kümeleme bir hiyerarşik yapı yada ağaç benzeri bir yapı ile ifade edilebilir. Birleştirici kümeleme *(agglomerative clustering)* ayrı bir veri setindeki her bir veri ile başlar. Kümeler verilerin büyük veya daha büyük kümelere gruplanmasıyla oluşur. Bu teknik, tüm verilerin tek bir kümenin üyesi oluncaya kadar devam eder. Bölünebilir kümeleme *(divisive analysis)* tüm verilerin tek bir küme içerisinde gruplanmasıyla başlar. Kümeler her bir verinin ayrı bir kümede oluncaya kadar devam eder yani veri sayısı kadar küme sayısı oluşur.

Diğer bir kümeleme tekniği hiyerarşik olmayan kümeleme yöntemleridir. Bu yöntem içerisinde bir küme merkezi seçilir ve merkezden daha önceden öngörülmüş eşik değeriyle tüm veriler biraraya gruplanır. Sonra yeni bir küme merkezi seçilir ve teknik, kümelenmiş noktalar için tekrarlanır.[50]

## 2.6. KÜMELEME METODLARI

Kümeleme analizi sürecinde kullanılan birçok farklı kümeleme metodu bulunmaktadır. Bu metodların kullanılmasında verilerin türü ve uygulamanın amacı önemli bir unsur olmaktadır. Bu bölümde Jwai Han'nın kitabı referans alınarak en çok kullanılan kümeleme metodlarından bahsedilecektir.

---

[50] Yrd. Doç. Dr. Vedat Pazarlıoğlu, **"Kümeleme Analizleri"**, (Çevrimiçi) http://www.angelfire.com/ia/selcukkoc/ka.html , 12.Mart 2004.

## 2.6.1. HİYERARŞİK METODLAR

Hiyerarşik kümeleme methodları, verileri ağaç yapısı şeklindeki gruplar içerisinde kümeleyerek çalışırlar. Hiyerarşik kümeleme methodları aşağıdan yukarıya yada yukarıdan aşağıya hiyerarşik ayrılmaya bağlı olarak birleştirici kümeleme (*agglomerative nesting*) ve ayrıştırıcı kümeleme (*divisive analysis*) olarak iki sınıfa ayrılırlar. Hiyerarşik methodlar, küme sayısını belirten k değerine ihtiyaç duymazlar, fakat ağaç yapısını oluşturma işleminin ne zaman durduracağını belirten eşik değerini bilmeye ihtiyaç duyarlar.

Hiyerarşik kümeleme methodları, yorumlama ve okumada kolaylık sağlaması bakımından avantajlı olmasına rağmen, sağlam ve basit olmaması ve az güvenilir olması sebebiyle de dezavantajlıdır. Buna rağmen uygulamalarda hiyerarşik olmayan methodlara göre daha fazla kullanılmaktadır.[51]

### 2.6.1.1. BİRLEŞTİRİCİ KÜMELEME (AGGLOMERATIVE NESTİNG)

1990 yılında Kaufman ve Rousseuw tarafından ortaya çıkmıştır.[52] Bu methodun çalışmasında aşağıdan yukarıya doğru bir yol izlenir. Başlangıçta her veri, kendisi bir küme olacak şekilde yerleşir ve algoritma bu şekilde çalışmaya başlar. Benzer özellik gösteren veriler birleşerek daha büyük kümeleri oluştururlar. Bu işlemler sonlandırma koşulu gerçekleşene kadar tekrar edilir. Koşul verilmediği durumda sonunda tüm veriler birleşerek tek bir küme altında toplanırlar.

---

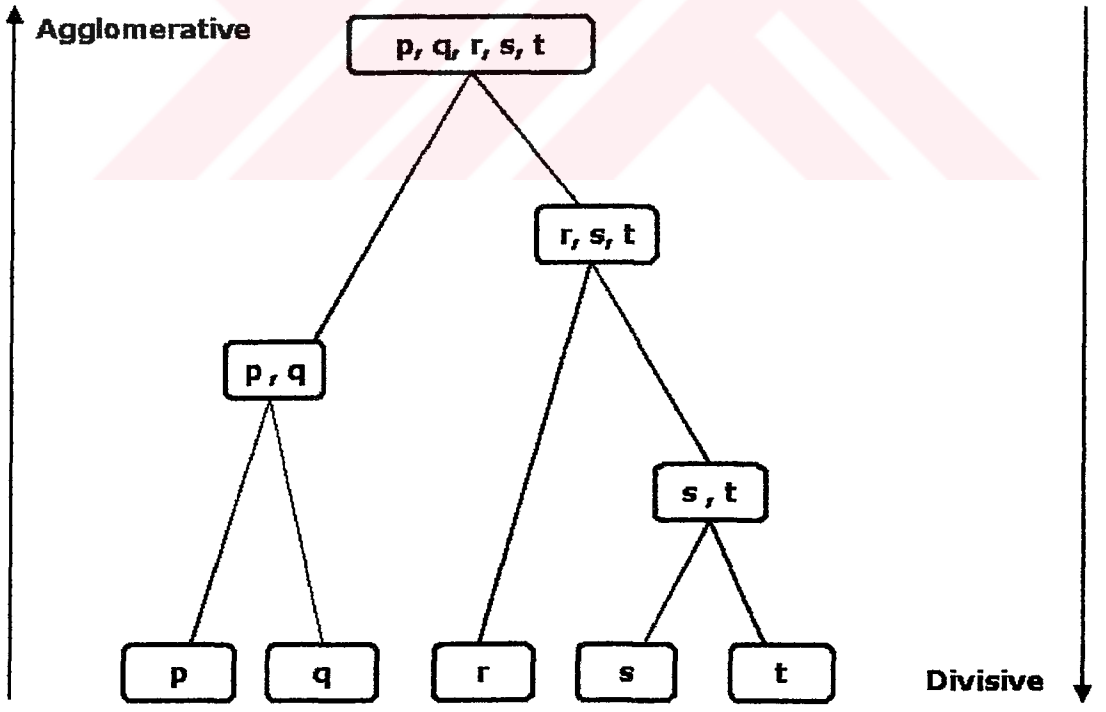
[51] David A. Aaker, George S. Day, "**Marketing Research**", John Wiley & Sons, Third Edition, New York, 1986, S.481.

[52] L. Kaufman, P. J. Rousseuw, "**Finding Groups in Data: An Introduction to Cluster Analysis**", John Wiley & Sons, New York, 1975.

*Birleştirici Kümeleme Algoritması Süreci*[53]

1. *Her veri nesnesini ayrı bir küme olarak ata.(n adet veri olmak üzere)*
2. *Kümeler arasındaki uzaklık değerlerini oluştur.*
3. *Uzaklık değerlerini kullanarak bir uzaklık matrisi oluştur.*
4. *En yakın uzaklık değerine sahip küme çiftini araştır.*
5. *Bu çifti matristen çıkart ve birleştir.*
6. *Bu yeni kümeden diğer tüm kümelere olan uzaklık değerlerini yeniden bul ve matrisi güncelle.*
7. *4, 5 ve 6 nolu adımlar n-1 kez tekrarlanır.*

Hiyerarşik kümeleme methodları, ***dendogram*** adı verilen bir yapıyı kullanırlar. Şekil 12' de bir dendogram yapısı gösterilmektedir.




**Şekil 12. Dendogram Yapısı.**[54]

[53] Çevrimiçi, http://www.predictivepatterns.com/docs/ WebSiteDocs/ Introduction/Tutorials/Tutorial_Use_Case_Scenarios.htm , 02 Ocak 2004.

## 2.6.1.1.1. BİRLEŞTİRİCİ KÜMELEME TEKNİKLERİ

### 2.6.1.1.1.1. TEK BAĞLANTI YÖNTEMİ (SINGLE LINKAGE)

En yakın komşuluk methodu olarak da ismi geçen bu teknik en bilinen ve en basit birleştirici kümeleme tekniğidir. Veri çiftleri arasındaki en yakın uzaklık değeri alınarak iki benzer kümenin bulunması amaçlanır. En yakın değeri veren iki küme birleştirilir. Bundan sonra da diğer kümelerin bu yeni oluşan küme ile birleşmeleriyle işlemler devam eder. En sonunda da tüm veriler tek bir kümede toplanırlar.

Tek bağlantı yöntemi D (r, s) olarak gösterilir.

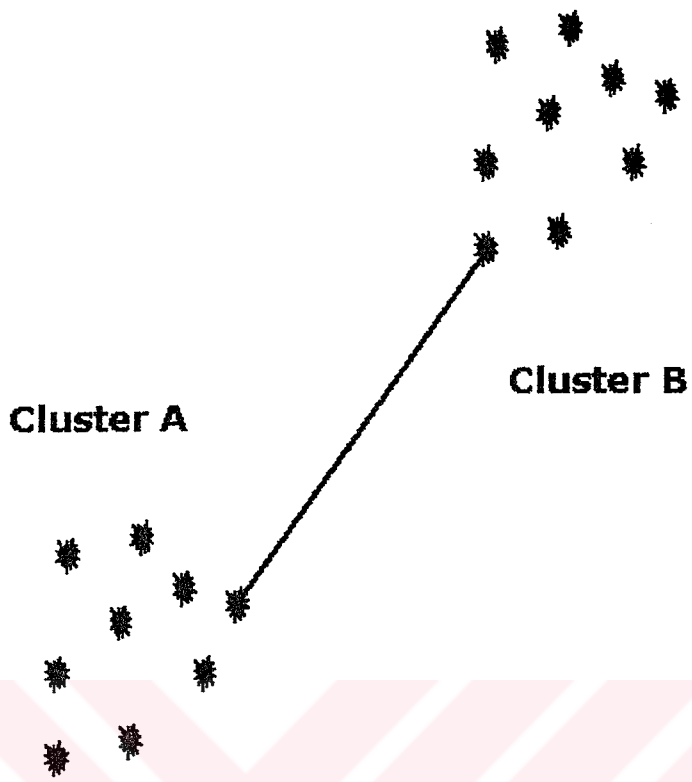
D (r, s) = Min { d (i, j) : i verisi r kümesinde ve j verisi s kümesinde }

Burada her (i, j) veri çifti uzaklık değeri hesaplanır. Bu uzaklıklardan minimumu r ve s kümeleri arasındaki uzaklık değeri olarak alınır. Eğer r ve s kümeleri birleşmiş iseler, bundan sonraki birleşmeler için kullanılacak olan uzaklık değeri D (r, s) olacaktır.

Hiyerarşik kümelemenin her adımında minimum uzaklık değerine sahip D(r, s) kümeler birleştirilir.

---

[54] (Çevrimiçi) http://www.resample.com/xlminer/help/hcist/hcist_intro.htm, 01 Ocak 2004.




**Şekil 13. Tek Bağlantı Yöntemi.**[55]

## 2.1.1.1.2. TAM BAĞLANTI YÖNTEMİ (COMPLETE LINKAGE)

Tek bağlantı yönteminin tam tersi olan bu yöntem ***En Uzak Komşuluk Yöntemi*** olarak bilinmektedir. Her kümeden alınan veri çiftleri arasındaki en yüksek uzaklık değeri kümeler arasındaki uzaklık değeri olarak alınır.

Tam bağlantı yöntemi D (r, s) olarak gösterilir.

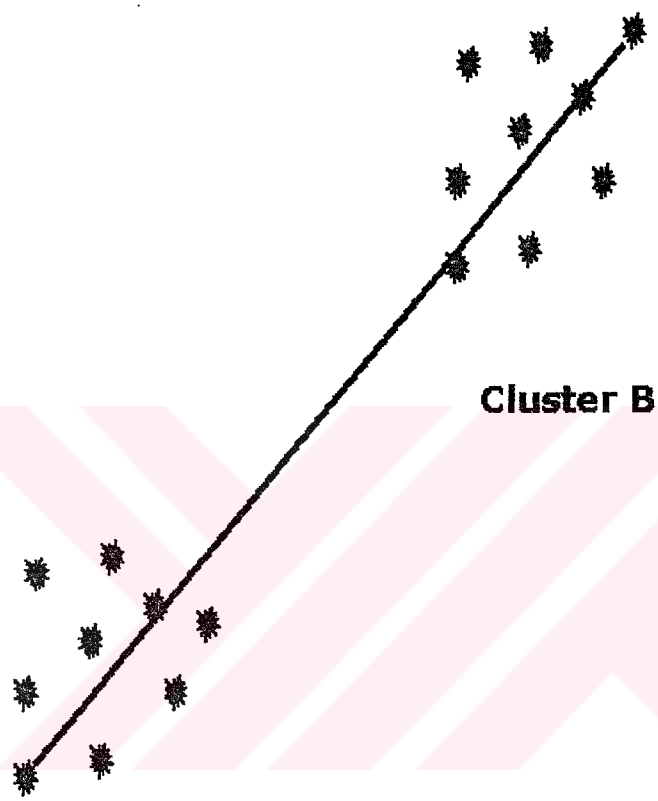
D (r, s) =Max { d(i, j) : i verisi r kümesinde ve j verisi s kümesinde }

Burada veri çiftleri arasındaki uzaklık değerleri hesaplanır. Bu uzaklık değerleri arasındaki en büyük değer r ve s kümeleri arasındaki uzaklık değeri olarak alınır.

---

[55] A.e.

Hiyerarşik kümelemenin her adımında D (r, s) değerlerinin en küçüğünü veren kümeler birleştirilir.




**Şekil 14. Tam Bağlantı Yöntemi.**[56]

## 2.1.1.1.3. ORTALAMA BAĞLANTI YÖNTEMİ (AVERAGE LINKAGE)

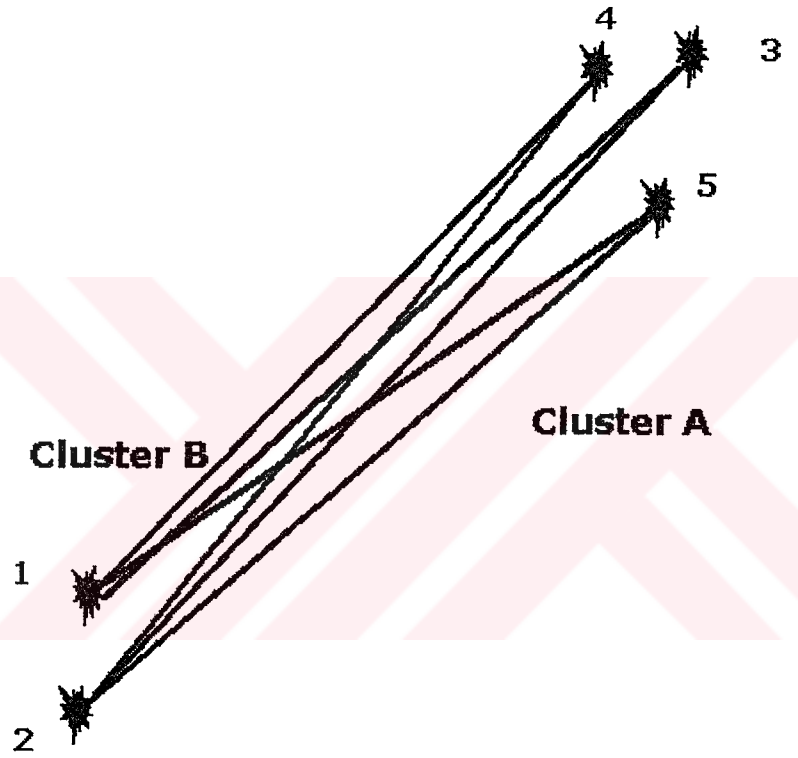
Ortalama bağlantı yönteminde, iki küme arasındaki uzaklık, birinci kümedeki verilerin ikinci kümedeki verilere olan uzaklıkların ortalamasıdır.

Ortalama bağlantı yöntemi D (r, s) olarak gösterilir.

[56] A.e.

$$D(r, s) = T_{rs} / (N_r * N_s)$$

T , r ve s kümelerindeki çiftler arasındaki uzaklıkların toplamını ifade etmektedir. $N_r$ ve $N_s$ , kümenin boyutunu göstermektedir.




**Şekil 15. Ortalama Bağlantı Yöntemi.**[57]

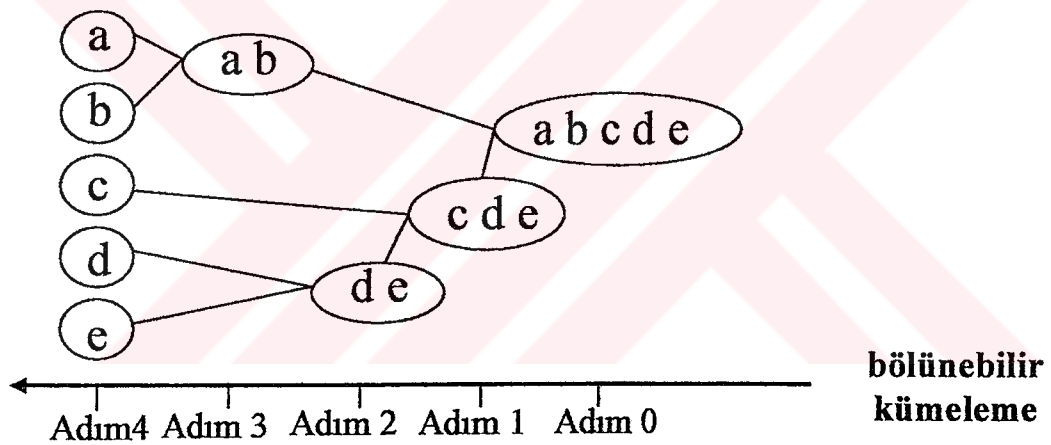
Tek bağlantı yöntemi sağlıklı sonuç vermesine rağmen işlemlerinin uzun sürmesi yüzünden pratik değildir. Tam bağlantı yöntemi aynı küme içerisindeki verilerin uzaklıklarının belli bir değerden küçük olması durumunda tüm kümelerin sağlıklı oluşturulmasını garanti edemez. Son yıllarda sıkça kullanılmaya başlanan

[57] A.e.

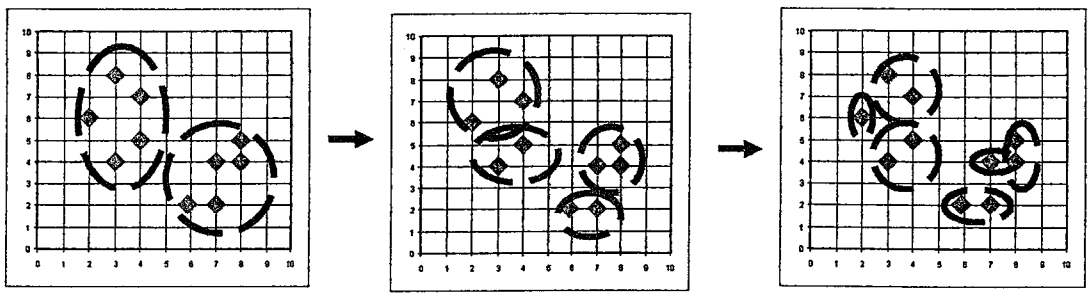
ortalama bağlantı yöntemi, bu iki uç teknik arasında sonuçlar vermesi nedeniyle bir alternatif olarak kabul edilmektedir.[58]

## 2.6.1.2. BÖLÜNEBİLİR KÜMELEME (DIVISIVE ANALYSIS)

Bölünebilir kümeleme algoritması 1990 yılında Kaufman ve Rousseuw tarafından sunulmuştur. Bölünebilir hiyerarşik yöntemler birleştirici hiyerarşik yöntemlere göre tam ters şekilde çalışır. Başlangıçtaki tek küme, iki alt kümeye, bu kümeler de birbirine benzemeyen diğer alt kümelere bölünürler. Bu işlemler veri sayısı kadar alt küme oluşana kadar devam eder.




**Şekil 16. Bölünebilir Hiyerarşik Kümeleme Yöntemi.**[59]



**Şekil 17. Divisive Analysis.**[60]

[58] Menteş, a.g.e. s.63.
[59] Hairong Huang, Yuming Mu, Ying Wei, Stanly Yong, **"Clustering Analysis.ppt"**, (Çevrimiçi) https://netfiles.uiuc.edu/dgs/www/stat478/projects/19, 10 Ocak 2004.
[60] J. Han, **"Data Mining Concepts and Techniques"**, s.357.

### 2.6.1.3. BIRCH (BALANCED ITERATIVE REDUCING AND CLUSTERING USING HIERARCHIES)

Birch algoritması 1996 yılında Sigmod konferansında Zhang, Ramakrishnan ve Linvy tarafından duyurulmuştur. Küme niteleyici *(clustering feature)* ve küme niteleyici ağacı *(clustering feature tree)* Birch algoritmasının çekirdeğini oluşturmaktadır. Küme niteleyici, bir küme içindeki d boyutlu N veri noktası için 3 parametreyle tanımlanan bir vektördür.[61]

CF = (N, LS, SS) , N : Küme içindeki veri sayısı.

LS : Küme içindeki N verinin doğrusal toplamları (linear sum).

$$LS = \sum_{i=1}^{N} x_i$$

SS : Küme içindeki N verinin karelerinin doğrusal toplamı.

$$SS = \sum_{i=1}^{N} x_i^2$$

$CF_1 = (N_1, LS_1, SS_1)$ ve $CF_2 = (N_2, LS_2, SS_2)$ şeklinde iki ayrı küme düşünülürse bu iki ayrı küme birleştirilerek yeni bir küme vektörü yaratılabilir.

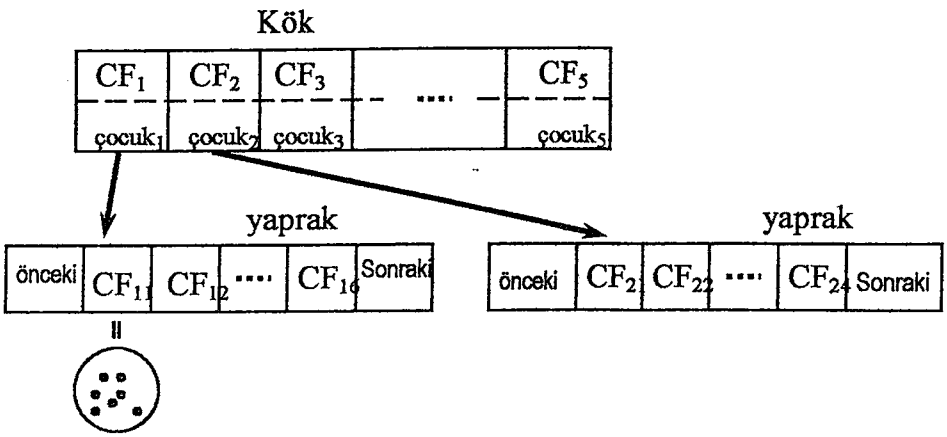
$CF_1 + CF_2 = (N_1 + N_2 , LS_1 + LS_2 , SS_1 + SS_2 )$

*Küme Niteleyeci Ağacı (Clustering Feature Tree – CF Tree) :*

CF ağacı iki parametreden oluşur. Bu parametreler dallanma katsayısı *(branching factor)* B ve eşik *(threshold)* T değeridir. Dallanma katsayısı B, ağaçta en fazla kaç adet yaprak olacağını belirtir. Eşik değeri T ise, yaprak içinde oluşacak küme sayısının çapının ne olacağını belirler. Bulunan CF değerleri kullanılarak CF ağaçları oluşturulur. CF değerleri ana bellekte veritabanlarına göre daha az yer kapladığı için buralarda tutulurlar.

---

[61] Tian Zhang, Raghu Ramakrishnan, Miron Linvy, **"BIRCH: An Efficient Data Clustering Method for Very Large Databases"**, SIGMOD, 1996.




**Şekil 18. CF Ağacı yapısı.**[62]

*CF Ağacına Ekleme İşlemi:*

- *Uygun yaprağı belirleme:* Kökten başlayarak CF ağacından aşağıya tekrarlı bir şekilde inerek seçilen uzaklık ölçütlerine göre en yakın çocuk düğümü (child node) seçilmektedir.

- *Yaprağı değiştirme:* Yaprak düğüme ulaştığı zaman $L_1$ olarak adlandırılan en yakın yaprak girişini bulur. Eşik değerini geçip geçmediği test edilir. Eğer test geçilirse CF vektörü bu yeni durumu yansıtmak için güncellenir ve yeni giriş yaprağa eklenir. Yaprakta boşluk varsa bu yeni giriş yapılır aksi durumda yaprak düğüm ikiye bölünür. Bölünme işlemi en uzak çiftlerin seçilmesiyle olur. Kalan girişler ise en yakın kriterine göre yeniden dağıtılırlar.

[62] J. Han, **"Data Mining Concepts and Techniques"**, s.357.

- *Yaprağın adresini değiştirme:* Yaprağa giriş yapıldıktan sonra CF bilgisi güncellenmelidir.

*BIRCH algoritmasının işleyişi:*

- Tüm veri taranır ve veri tabanındaki N adet veri kadar alt kümeler oluşturulur. Bu alt kümelerin her biri için CF değeri hesaplanır. Bu CF değerleri veri tabanlarına göre daha az yer kapladığı için ana bellekte saklanırlar. Hesaplanan CF değerleri kullanılarak CF ağaçları oluşturulur. CF ağaçları için iki parametre söz konusudur. Bunlar ağaçta en fazla kaç adet yaprak olacağını belirleyen dallanma katsayısı (branching factor) ve yapraklarda oluşturulacak olan küme sayısının çapının en fazla ne kadar olacağını belirleyen eşik değeri *(treshold)* 'dir.

- Veri tabanının yapısı, oluşturulan CF ağacı ile ortaya çıkmaktadır. Kullanılacak olan bölümleme algoritması yada hiyerarşik algoritma ile kümeleme işlemi yerine getirilmektedir.

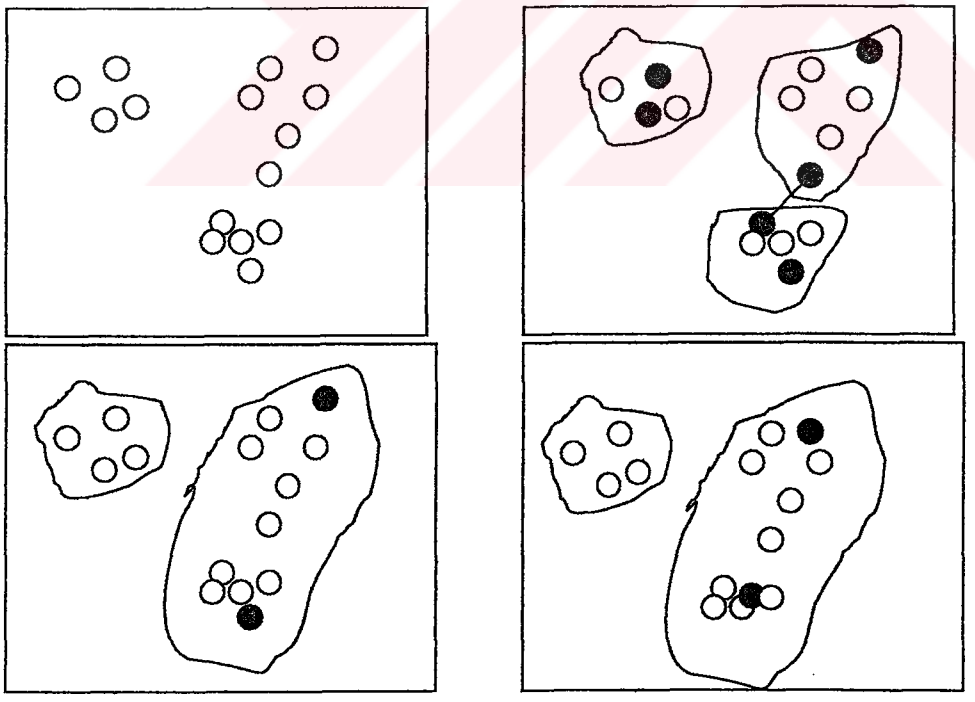
Orijinal veri seti sadece bir defa taranır fakat ağaç yapısı defalarca taranabilir. BIRCH algoritmasının sağladığı bazı faydalar bulunmaktadır. Bu algoritma, çalısması sırasında ana belleği kullandığından I / O (Input / Output – Giriş / Çıkış) miktarını azaltır ve performansı arttırır. Veri setini tek bir defa tarayarak veri setinin modelini oluşturabilir ayrıca çok fazla işlem gücü gerektirmez. Bunun yanında bu algoritma sadece sayısal verilerde kullanılmaktadır ve sadece dairesel kümeleri ortaya çıkarabilmektedir.[63]

[63] Pavel Berkhin, **"Survey of Clustering Data Mining Techniques"**, Accruse Software Inc., California, USA, 2002.

## 2.6.1.4. CURE (CLUSTERING USING REPRESANTATIVES)

CURE algoritması 1998 yılında SIGMOD konferansında Guha, Rastogi ve Shim tarafından sunulmuştur.[64] CURE algoritması, dairesel olmayan kümelerin bulunması konusunda diğer hiyerarşik algoritmaların zayıflıklarını gidermek amacıyla ortaya çıkmıştır.

CURE algoritmasında tek bir merkez nokta yada bir kümeyi temsil eden bir veri kullanmak yerine uzayda sabit sayıda tanımlayıcı nokta seçilir. Bir kümenin tanımlayıcı noktaları ilk olarak iyi dağılmış verilerden seçilerek oluşturulur ve sonra belirli bir fonksiyon yada daraltma faktörüyle bunlar küme merkezine doğru daraltılır yada taşınır. Algoritmanın her adımında, tanımlayıcı nokta çifti en yakın iki küme birleştirilir. CURE algoritması BIRCH algoritmasında olduğu gibi rastgele örneklem alarak veritabanını modeller.



**Şekil 19.CURE Algoritması.**[65]

[64] S. Guha-R. Rastogi-K. Shim, **"CURE: An Efficient Clustering Algorithm for Large Databases"**, ACM SIGMOD, Seattle, USA, 1998.
[65] A.e.

CURE algoritması aşağıdaki adımlardan oluşur.

- *Orijinal veri setinden S sayıda rastgele örneklem seçilir.*
- *Veri seti, S örneklem sayısı kadar bölüme ayrılır.*
- *Veri setindeki her bölüm üzerinde kümeleme işlemi yapılır. Örneklem kümesinin her elemanı küme merkezi olacak şekilde alt kümeler oluşturulur.*
- *Kümeleme sürecinde yeterli büyüklüğe erişememiş kümeler ortadan kaldırılır.*
- *Büyük kümelerin yaratılması için daha küçük olan alt kümelerin merkez noktaları dikkate alınarak kümeleme süreci tekrar başlatılır. Süreç sonunda daha büyük ve küresel olmayan kümeler yaratılmış olur.*
- *İstenen küme büyüklüğüne ulaşılana kadar CURE algoritması elde edilen yeni küme noktaları üzerinden devam eder.*

CURE algoritması yüksek kalitede kümeler oluşturur, farklı boyutlarda ve karmaşık şekillerde kümeler yaratabilir. CURE algoritması kategorik veriler üzerinde kullanılamaz. Bu eksiğini gidermek için ROCK (Robust Clustering Algorithm) algoritması yaratılmıştır. Bu algoritma ile CURE algoritması benzer bir çalışma mantığına sahiptir.

## 2.6.1.5. CHAMELEON ( HIERARCHICAL CLUSTERING USING DYNAMIC MODELING )

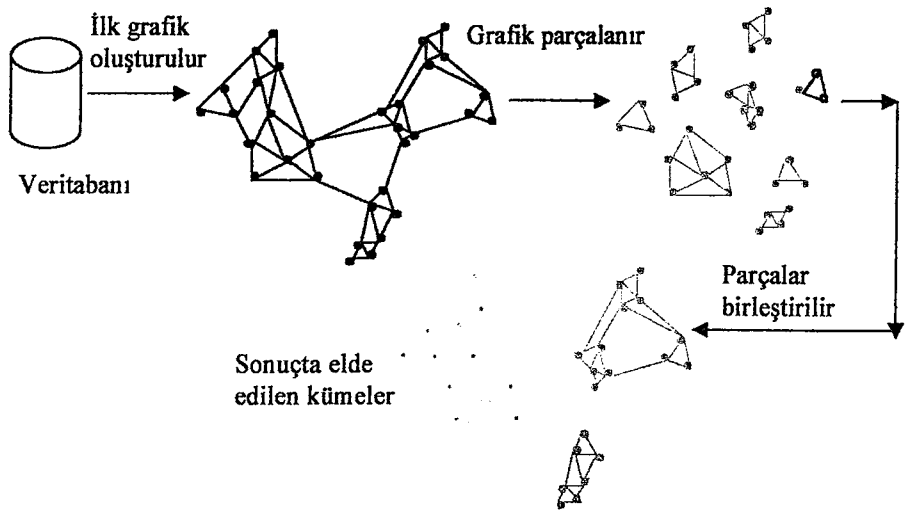
Chameleon algoritması 1999 yılında Karypis, Han ve Kumar tarafından sunulmuştur. Chameleon, hiyerarşik kümeleme içerisinde dinamik bir kümeleme yapısı ortaya çıkarmaya çalışan bir kümeleme algoritmasıdır.[66]

Kümeleme sürecinde, eğer iki küme arasındaki yakınlık derecesi yüksek ise bu iki küme birleştirilir. Bu algoritma dinamik modelli bir yapıya sahiptir ve bu birleştirme sürecinde doğal ve homojen kümelerin keşfini de kolaylaştır. Chameleon algoritması CURE ve ROCK kümeleme algoritmalarının zayıflıklarından doğmuştur. CURE algoritması, iki farklı kümedeki verilerin ilişkileri hakkındaki bilgileri dikkate almamaktadır, ROCK algoritması ise iki kümenin ilişkisini vurgularken aralarındaki yakınlık ile ilgili bilgileri dikkate almaz.

Chameleon algoritmasının çalışma şekli şekil 20'de gösterilmektedir. Öncelikle Chameleon algoritması veri nesnelerini çok sayıda küçük alt kümelere bölen bir grafik bölümleme algoritması kullanır. Sonraki adımda alt kümeler tekrarlı bir şekilde birleştirici hiyerarşik kümeleme algoritması kullanılarak birleştirilerek gerçek kümeler bulunmaya çalışılır.[67] Alt kümelerin birleşerek yeni kümeler oluşturmasında iki unsur önem taşımaktadır. Birincisi, iki kümenin birbirine ne kadar benzer olduğunu gösteren bağıl bağlanabilirlik (*Relative Interconnectivity*) ölçüsü, ikincisi ise iki kümenin birbirine olan yakınlık durumunu gösteren bağıl yakınlık (*Relative Closeness*) ölçüsüdür. Kümeleme süreci dinamik bir yapıya sahiptir ve açıkladığımız iki unsura göre kümeler birleştirilir. Kümeleme sürecinde en uygun küme sonuçları ortaya çıkana kadar ayrılma ve birleşme işlemleri devam eder.

---

[66] Han, a.g.e. s. 361-362.
[67] A.e. s. 361-362.




**Şekil 20. CHAMELEON algoritmasının çalışma yapısı [68].**

## 2.6.2. BÖLÜMLEME METODLARI (PARTITIONING METHODS)

Bölümleme metodları, veri tabanında bulunan n adet veriyi, belirlenmiş olan k sayıda kümeye ayırmaya çalışmaktadır ($k \leq n$). Farklılık fonksiyonuna (*dissimilarity function*) göre veritabanındaki veriler kümelere ayrılırlar.[69]

Bölümleme metodları küçük ve orta ölçekli veri tabanlarında küresel şekilli kümeleri bulmada iyi sonuçlar vermektedir.[70] Bölümleme algoritmalarının genel problemi başlangıçta verilen k giriş parametresine bağlı olması ve düzgün şekilli olmayan kümeleri bulamamasıdır.[71]

Bölümleme methodlarına k-means,k-medoids ve clara-clarans algoritmalarını örnek verebiliriz.

---

[68] A.e. s. 361-362.
[69] Jean Fen Ju Hou, **"Clustering with Obstacle Entities"**, Master tezi, Kanada, Simon Fraser Univerity Computing Science, 1999.
[70] **"Clustering Algorithms Classification"**, Computer Science Departman, Trinity College.
[71] Pavel Berkhin, **"Survey of Clustering Data Mining Techniques"**, California, USA, 2002.

### 2.6.2.1. K-MEANS ALGORİTMASI

K-means algoritması 1967 yılında MacQueen tarafından sunulmuştur. Uzun yıllar boyunca pek çok uygulama alanında yoğun olarak kullanılan bir kümeleme algoritmasıdır. Bu algoritmada *k* sayıda küme oluşmaktadır ve her küme içerisinde bulunan verilerin ağırlıklı ortalamaları sonucu bir değer ortaya çıkmaktadır. Küme içerisinde bu değere en yakın olan nokta değeri küme merkezi (*centroid*) olarak kabul edilmektedir.[72]

K-means algoritması öncelikle *k* sayıda rasgele nokta belirler. Bu noktalar ilk küme merkezlerini temsil etmektedir. Daha sonra gelen her veri değeri merkez noktaya en yakın olduğu kümeye dahil edilir. Eklendiği küme elemanlarının ağırlıklı ortalamaları tekrar hesaplanarak yeni bir küme merkezi değeri bulunur ve bu yeni değer bundan sonraki kümeleme işlemlerinde bu kümeyi temsil eder.

K-means algoritmasının çalışma şekli şöyledir:

1. *K sayıda rasgele küme merkezi belirle.*
2. *döngü.*
3. *veriyi, hangi kümenin ortalamasına en yakınsa o kümeye dahil et.*
4. *küme ortalamasını tekrar hesaplayarak yeni küme merkezini belirle.*
5. *küme üyeliklerinde değişiklikler biti mi? Hayır ise 2. adıma geri dön. Evet ise dur.*

[72] A.e.

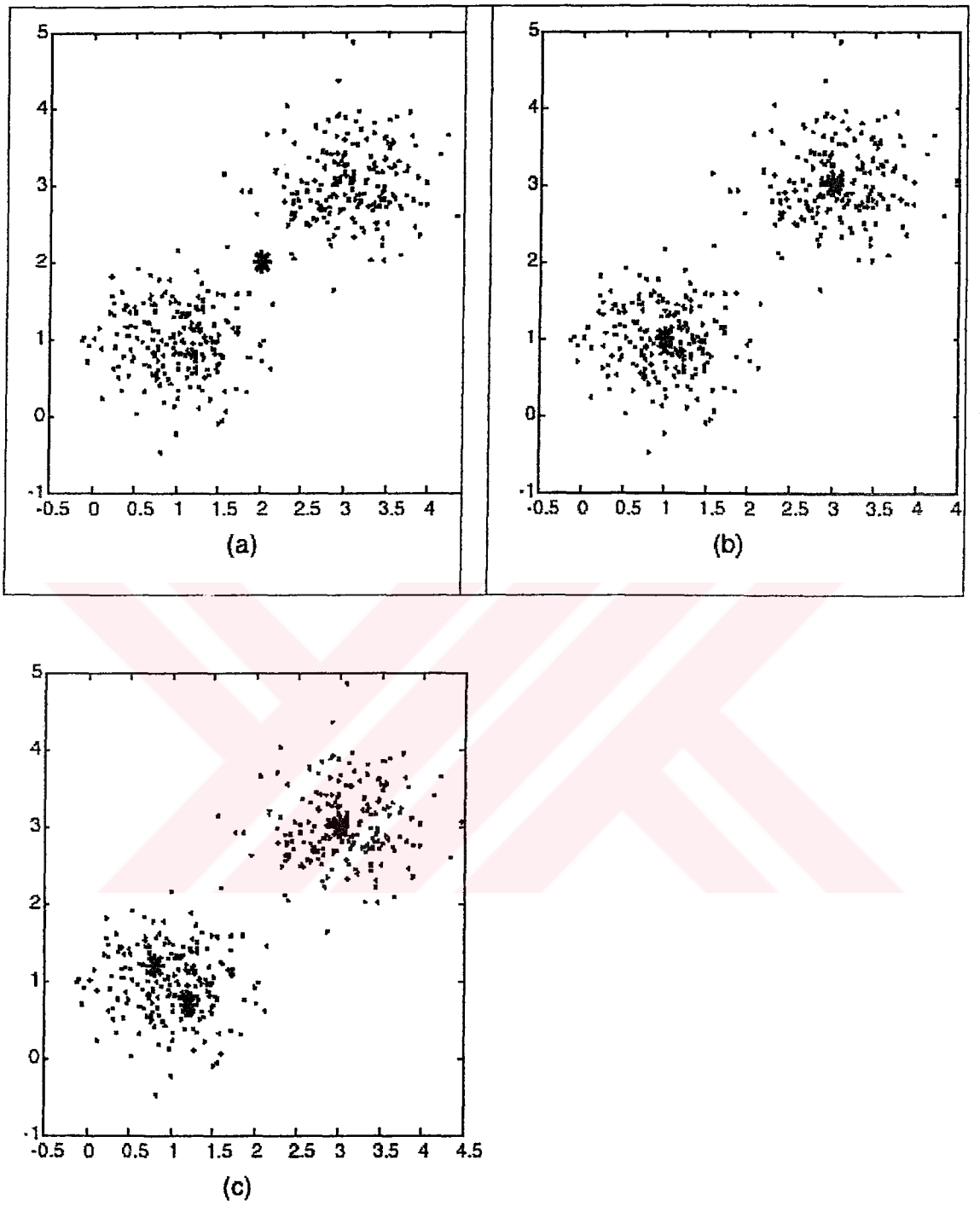
K-means algoritmasının zayıflıklarını aşağıdaki gibi sıralayabiliriz:[73]

- Algoritmanın başında giriş parametresi olarak bir $k$ sayısına ihtiyacı vardır. Elde edilecek olan sonuçlar $k$ sayısına göre değişkenlik gösterebilmektedir.

- Aşırı gürültü ve istisna veriler algoritmayla hesaplanan ortalamayı değiştirdiği için k-means algoritması gürültü ve istisnaya aşırı duyarlıdır.

- K-means algoritması sadece sayısal veriler ile kullanılabilmektedir. Kategorik verilerin kümelenmesi için k-means algoritması bir çözüm sunamamaktadır.

[73] Alexander Sturn, "**Cluster Analysis for Large Scale Gene Expression Studies**", Master Tezi, The Institute for Genomic Research (TIGR), USA, 2000.




**Şekil 21. K-means algoritmasının sonuçları. (a) k = 1; (b) k = 2; (c) k = 3.**[74]

[74] A.e.

## 2.6.2.2. CLARA VE CLARANS ALGORİTMALARI

K-medoids ve PAM gibi bölümleme algoritmaları küçük veri tabanlarında etkili bir şekilde çalışmaktadır fakat büyük veri tabanlarında bu denli başarılı çalışamamaktadır. Büyük veri tabanlarında etkili sonuçlar alabilmek için Kaufman ve Rousseeuw tarafından 1990 yılında CLARA (Clustering LARge Application) algoritması ortaya çıkmıştır.[75]

CLARA algoritmasında tüm veri setini işlem sürecine almak yerine içerisinden küçük bir kısmını alarak tüm veriyi temsil ediyor kabul edilerek seçim yapılır. Bu örneklem üzerinde daha sonra PAM algoritması uygulanır. Eğer örneklem rasgele olarak seçilirse orijinal veri tabanına temsil gücü olarak daha yakın olabilir. CLARA algoritması bu şekilde veri tabanından birçok farklı örneklem çıkarır ve PAM algoritmasını her örneklem üzerinde uygular. Bu sürecin sonunda en iyi PAM sonucunu veren örneklemden elde edilen sonuç çıktı olarak verilir. Bu şekilde CLARA algoritması PAM algoritmasına göre daha büyük veri tabanlarıyla çalışır.

CLARA algoritmasının etkisi örneklem büyüklüğüne bağlıdır ve örneklem seçimi yeterince bağımsız değilse seçilen örneklem veri tabanını yeterince temsil edemeyeceği için yanlış sonuçlar verebilir.[76]

1994 yılında Ng. Raymond ve J. Han tarafından VLDB'94 konferansında CLARANS (CLustering Algorithm based on RANdomized Search) algoritması ortaya çıkmıştır.[77] CLARANS algoritması CLARA algoritmasından farklı olarak herhangi bir zamandaki herhangi bir örneklemeyle kendisini sınırlamaz. CLARA algoritması aramanın her adımında sabit bir örneklem kullanır, buna karşın CLARANS algoritması aramanın her adımında değişen rasgele örneklem kullanır.

---

[75] Han, s. 354.
[76] A.e., s. 354.
[77] N.G. Raymond, J.Han **"Efficient and Effective clustering method for spatial data mining"** Int. Conference Very Large Data Bases (VLDB'94), Santiago, Şili (Eylül 1994) 144-155

## 2.6.2.3. K-MEDOIDS ALGORİTMASI

K-medoids algoritması Kaufman ve Rousseeuw tarafından 1987 yılında ortaya çıkmıştır. Bu algoritma, k-means algoritmasının ortadan kaldıramadığı gürültü ve istisna verilerin etkisini gidermeyi amaçlamaktadır.[78]

K-medoids algoritmasında *medoid (merkez nokta)* olarak ifade edilen bir kavramdan bahsedilmektedir. Merkez nokta, bir kümeyi temsil eden veri setinden seçilen bir nesnedir. Algoritma *k* tane kümeyi temsil eden *k* tane merkez nokta seçer. Daha sonra, geriye kalan nesnelerin her biri en yakın merkez noktaya atanarak kümeler oluşturulur.[79]

K-medoids algoritmasında kümeyi temsil edecek nokta, kümedeki verilerin ortalaması ile belirlenmemektedir. Kümenin en merkez noktasındaki eleman küme merkezi olarak alınmaktadır. Bu şekilde istisna verilerin küme merkezinin kenarlara doğru kaydırılması engellenmektedir.

Giriş değeri olarak K-medoids algoritmasına *k* küme sayısının belirtilmesi gerekmektedir. Bu durumda kullanıcının sürekli denetleme yaparak iyi bir tahminde bulunması gerekmektedir.

---

[78] D.P. Mercer, **"Clustering Large Datasets"**, Linacre College, Ekim 2003.
[79] A.e.

# BÖLÜM 3

# BÖLÜMLEME METODU ALGORİTMALARINDAN K-MEANS ALGORİTMASININ ÖĞRENCİ VERİTABANINA UYGULANMASI

## 3.1. UYGULAMANIN AMACI

Son yıllarda veri madenciliği yöntem ve tekniklerinin kullanım alanları hızla artmaktadır. Bununla beraber bu süreçlerin sonunda çıkan önemli sonuçlardan biriside örüntü tanımadır *(Pattern Recognition)*. Birden fazla özelliklerin dikkate alınması sonucu ortaya çıkan sonuçlarla birtakım özellikleri taşıyan örüntüler meydana gelmektedir. Bu sonuçlara dayanarak ileriki bir zamanda karşılaşılabilecek olan yeni durumun hangi örüntü grubunda olabileceği sorusuna da cevap verilebilecektir.

Uygulamamızın amacı ise tezin başından beri açıklamaya çalıştığımız veri madenciliği ve süreçleri ile bu süreç içerisinde yer alan kümeleme analizinin gerçek hayatta bir uygulamasını göstermektir. Uygulama kapsamında kullanılacak olan veriler öğrenci verileridir.

Uygulamamızın amaçlarından bir diğeride öğrenci verilerinden hareket ederek benzer özelliklere sahip olan öğrenci örüntülerini elde edebilmektir. Böylece ortaya çıkan örüntülerin özelliklerinden faydalanılarak karşımıza gelebilecek bir öğrencinin, sahip olduğu özelliklerinden hareket ederek hangi gruba girebileceği tahmini yapılabilecektir.

Bu uygulamada veri madenciliği süreci içerisinde yer alan algoritmalardan K-means algoritması kullanılmaktadır. Bu algoritmanın kullanılmasının sebebi uzun yıllar boyunca çok geniş bir alanda yaygın olarak kullanılmasıdır. Bu uygulamada, K-means algoritmasını gerçek hayatta elde edilmiş olan verilerin üzerinde uygulayacağız.

## 3.2. UYGULAMA KONUSUNUN TESPİTİ

Tezin kapsamını oluşturan veri madenciliği ve kümeleme analizini ticari kurumların verileri üzerinde uygulanması düşünülmüştür. Fakat yapılan tüm çalışmalara rağmen bu tür verilerin elde edilmesi sağlanamamıştır. Bu nedenle aşağıdaki noktalar dikkate alınarak uygulama konusu tespit edilmiştir:

- Verilerin elde edilme kolaylığı.
- Verilerin veri tabanları içerisinde saklanmış olması ve bu veri tabanları içerisinde düzenli olarak bulunması.
- Uygulama sonucu elde edilecek olan sonuçların, verilerin elde edildiği kurum gibi diğer benzer kurumlara farklı bir bakış açısı getirmesi.
- Uygulamanın sonuçları açık olarak verildiği için verilerin elde edildiği kuruma bir takım bilgilerinin ortaya çıkmasından dolayı zarar gelmemesi.

Bu noktaların dikkate alınmasıyla beraber Maltepe Üniversitesi'nin öğrenci işleri veritabanından elde edilmiş olan öğrenci kayıtları kullanılmıştır.

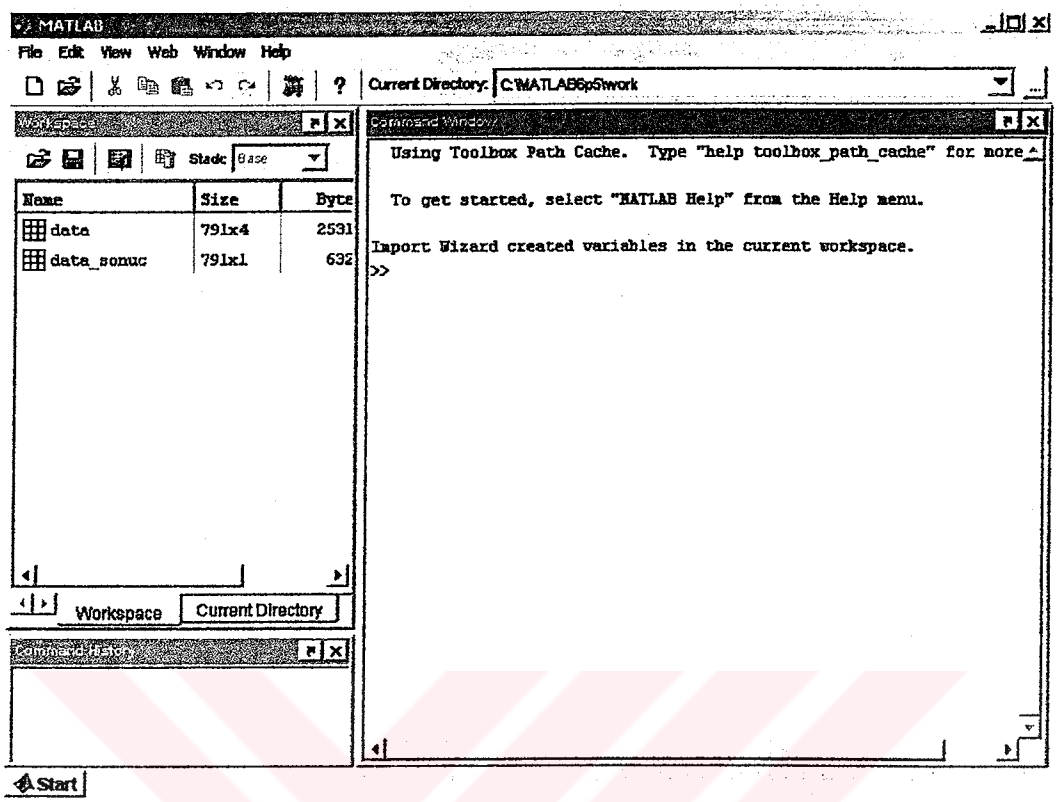
## 3.3. UYGULAMA GELİŞTİRME ORTAMI

Uygulamanın gerçekleştirileceği ortamları belirlerken birtakım noktalar göz önüne alınmıştır. Bu noktaları aşağıdaki gibi ifade edebiliriz:

- Uygulamanın gerçekleşme süresinin kısalığı.
- Kullanılacak olan programlama dilinin bu uygulama için uygunluğu.
- Kullanılacak olan veritabanı yönetim sisteminin uygulanmanın hazırlanacağı programlama dili ile olan uyumu.
- Birtakım sınırlılıklar sebebiyle kullanılacak bilgisayarların fiziki özellikleri.

Uygulamanın gerçekleştirileceği programlama dili olarak Matlab yazılımı seçilmiştir. Bunun sebeblerini aşağıdaki gibi açıklayabiliriz:

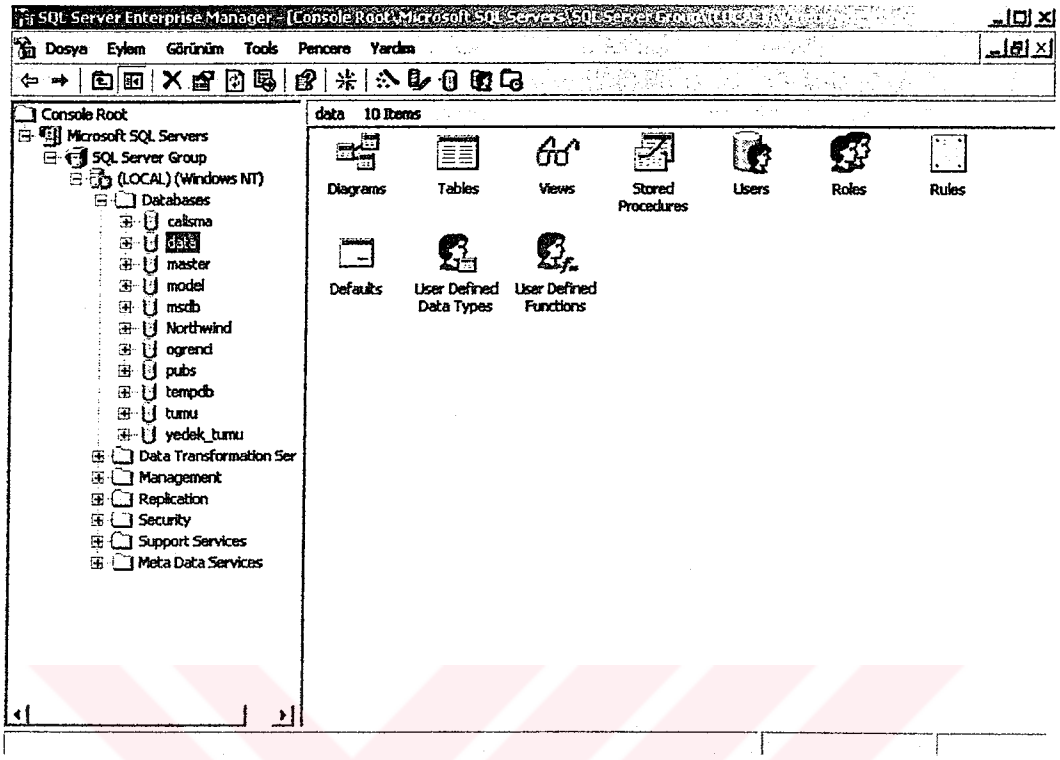
- Uygulamamızda kullanılacak olan K-means algoritmasının bu yazılım ortamı içerisinde hazır fonksiyon olarak yer alması.
- Verilerin tutulduğu veri tabanı yönetim sistemi ile Matlab yazılımının uyum içerisinde kullanılabilmesi.
- Uygulanmasının hazırlanması sırasında birçok hazır bileşene sahip olması.




**Şekil 22. Matlab 6.5 yazılım ortamı.**

Uygulamanın gerçekleştirilmesinde kullanılan veri tabanı yönetim sistemi Microsoft SQL Server 2000' dir. Veri tabanı yönetim sistemi olarak seçilmesinin nedenlerini aşağıdaki gibi açıklayabiliriz:

- Üniversitenin öğrenci işleri daire başkanlığından alınacak olan veriler Microsoft SQL Server 2000 veri tabanı yönetim sisteminde tutulmaktadır.
- Matlab 6.5 yazılım ortamıyla beraber uyum içerisinde kullanılabilmesi.
- Büyük hacimdeki verilerin saklanması, yönetilmesi ve kullanılmasındaki performansı ve verimliliği.




**Şekil 23. Microsoft SQL Server 2000 ortamı.**

Uygulamanın gerçekleştirildiği bilgisayar ortamı olarak Pentium IV 2400 Mhz tek işlemcili 256 MegaByte anabelleğe (RAM – Random Access Memory) sahip bir bilgisayar kullanılmıştır. Uygulamanın tümü tek bir bilgisayar üzerinde (Standalone) olarak gerçekleştirilmiştir.

Uygulamada farklı tablolarda bulunan verilerin birleştirilmesi sırasında Borland firmasının Delphi 6.0 görsel programlama dili kullanılmıştır.

## 3.4. VERİLERİN YAPISI

Öğrenci işleri daire başkanlığı 2003-2004 eğitim yılı başından itibaren yeni bir öğrenci işleri programına geçtiği için yeni öğrenci işleri programında tüm öğrencilerin ayrıntılı bilgileri henüz yer almamaktadır. Yeni öğrenci işleri programında şu an için sadece 2003 yılında Maltepe Üniversitesi'ni kazanan

öğrencilerin ayrıntılı tüm bilgileri yer almaktadır. Buna göre veri tabanında 3876 kayıt yer almaktadır.

Veri tabanında öğrenci bilgileri 2 farklı tabloda yer almaktadır. Bu tablolar Students ve Students_Log tablolarıdır. Students tablosunda öğrencinin genel bilgileri, aile bilgileri, nüfus bilgileri ve üniversite bilgileri yer almaktadır. Students_Log tablosunda ise öğrencinin not bilgileri yer almaktadır.

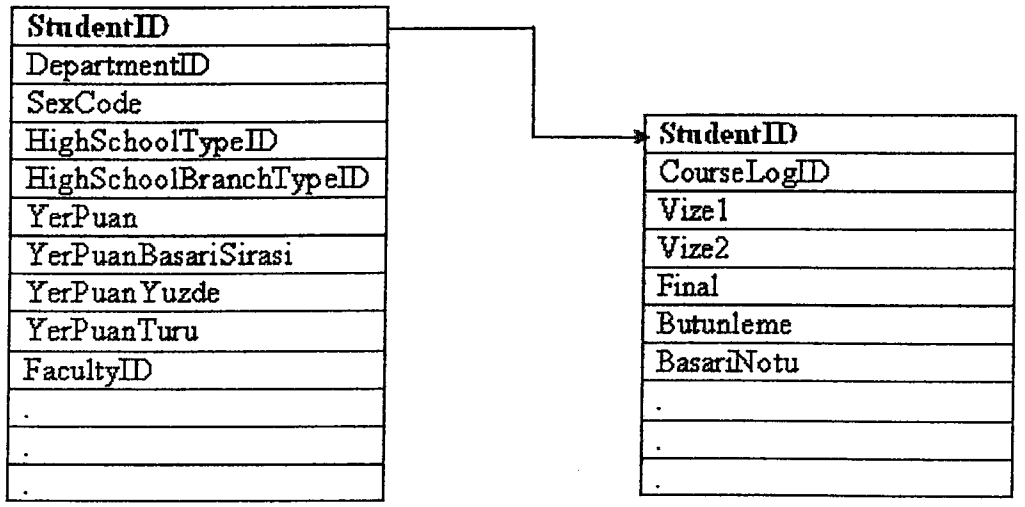
Bu iki tablonun tüm alanları *(Fields)* ve veri türleri EK-1 ve EK-2' de verilmiştir.

## 3.5. UYGULAMANIN VERİ MADENCİLİĞİ SÜREÇLERİ

Öğrenci veri tabanı üzerine uyguladığımız veri madenciliği, tezin başından beri açıklamaya çalıştığımız veri madenciliği süreçlerine uygun olarak gerçekleştirilmiştir. Her süreçte uygulama üzerine gerçekleştirilen işlemler ayrıntılı bir şekilde açıklanmaya çalışılmıştır.

### 3.5.1. VERİ TOPLAMA VE BİRLEŞTİRME

Bölüm 3.4 'de belirtildiği gibi öğrenci verileri Students ve Students_Log olmak üzere 2 farklı tabloda tutulmaktadır. Uygulamanın bu adımında Students tablosunda bulunan öğrenci bilgileri ile Students_Log tablosunda tutulan öğrenci notları birleştirilmektedir. Aşağıda verilen şekillerde olduğu gibi öğrencilerin bilgileri ve öğrencilerin notları StudentsID alanı kullanılarak birleştirilmektedir. Bu iki tablonun birleştirilmesi sırasında Delphi programlama dili kullanılmıştır.




**Şekil 24. Öğrenci ve öğrenci notlar tablo yapısı**[80]

Her iki tablonun birleştirilmesinden sonra aşağıdaki şekil 25' deki veri tablosu meydana çıkmıştır.

| StudentID |
|---|
| DepartmentID |
| SexCode |
| HighSchoolTypeID |
| HighSchoolBranchTypeID |
| YerPuan |
| YerPuanBasariSirasi |
| YerPuanYuzde |
| YerPuanTuru |
| FacultyID |
| CourseLogID |
| Vize1 |
| Vize2 |
| Final |
| Butunleme |
| BasariNotu |
| . |
| . |

**Şekil 25. Birleştirilmiş öğrenci tablo yapısı**

[80] Şekil 24.'de verilmiş olan iki tablonun tüm alanlarının EK-1 ve EK-2' de verilmiş olduğundan dolayı tekrar yazılmasına gerek görülmemiştir.

### 3.5.2. VERİ SEÇME VE TEMİZLEME

Birleştirme adımının gerçekleştiği yeni tabloda bazı temizleme işlemlerinin yapılması gerekmektedir. Bu adımları aşağıdaki gibi açıklayabiliriz:

- 2003-2004 yılında kayıt yaptırmış olan öğrencilerin bazılarının YerPuanYuzde değerleri bulunmamaktadır. Bunlar özel yetenek sınavı ile okula girmişlerdir. Bu tipteki öğrenciler öğrenci tablosundan silinmişlerdir.
- Tabloda yer alan bazı öğrenciler ne vize sınavlarına ne de final sınavlarına katılmamışlardır. Dolayısıyla başarı durumları olmamaktadır. Bu tipteki öğrencilerde öğrenci tablosundan silinmişlerdir.
- Uygulamada kullanılacak olan K-means algoritması aşırı uç değerlerden etkilendiğinden dolayı uç değerler tablodan silinmiştir.

Uygulamanın bu adımında veri madenciliğinde kullanılacak olan alanların seçilmesi ve bunların ayrı bir tablo olarak yaratılması gerçekleşmiştir. Birleşme işleminden sonra ortaya çıkan öğrenci tablosundan veri madenciliği işleminde kullanılmak üzere BasariNotu, YerPuan Turu, SexCode, HighSchoolTypeId ve FacultyId alanları alınarak yeni bir tablo yaratılmıştır.Bu yeni tablo aşağıdaki gibidir:

| YerPuanYuzde |
|---|
| BasariNotu |
| SexCode |
| HighSchoolTypeId |
| FacultyId |

**Tablo 5. Veri madenciliğinde kullanılacak alanlar tablosu.**

Bu adımda bahsedilen veri seçme ve temizleme işlemlerinden sonra veri madenciliğinde kullanılacak olan tablo hazırlanmış oldu. Bu işlemlerden sonra tabloda 722 satır kayıt içermektedir. Veri madenciliği için az bir miktar olmakla birlikte uygulamada kullanılacak veri miktarı yapılan önişlemler sonucunda bu sayıya düşmüştür. Öğrenci işlerindeki kullanılmaya başlanan yeni programda sonraki senelerde gelecek öğrenci bilgilerinin tam olarak eklenmesiyle beraber bu veri miktarı önemli hacimlere ulaşacaktır.

### 3.5.3. VERİ MADENCİLİĞİ

Uygulamanın bu adımında, daha önceki adımlarda hazırlanmış olan veri, veri madenciliği işlemine sokulmaktadır. Veri madenciliği sürecine tabloda yer alan 5 alan girmektedir. Bu alanlar ÖSS sınavından aldığı puana göre yüzdelik sırası, üniversitesindeki derslerinden aldığı başarı notu, öğrencinin cinsiyeti, lise tipi ve fakülte tipidir.

Bu beş alanı ölçek birimlerinden arındırabilmek için standart hale sokulmuş ve daha sonra da veri madenciliği uygulanmıştır. Algoritma sırasında öklit uzaklık ölçüsü kullanılmıştır. K-means algoritmasının en önemli eksikliği olan küme sayısının bilinmemesi yüzünden deneme yoluyla optimum ayrılmış durumda olan kümeler tespit edilebilmektedir.

K-means algoritmasının uygulanması sonucunda aşağıdaki tabloda verilen küme merkezleri ortaya çıkmıştır.

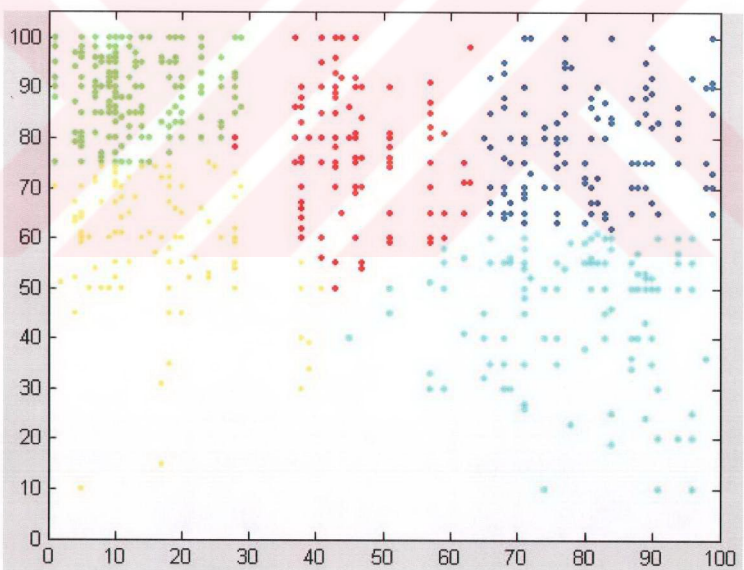
| Kümeler | YerPuanYuzde | BasariNotu | Cinsiyet | HighSchoolTypeID | FacultyID |
|---|---|---|---|---|---|
| 1 | 12.4774 | 89.5350 | 8.1070 | 6.4650 | 2.5844 |
| 2 | 16.3113 | 59.0472 | 6.9811 | 5.1981 | 3.0189 |
| 3 | 46.7851 | 77.1240 | 6.9008 | 5.7190 | 3.2149 |
| 4 | 80.1095 | 78.0146 | 6.6788 | 3.2774 | 3.8321 |
| 5 | 79.3565 | 44.8870 | 6.3043 | 3.1391 | 4.1304 |

**Tablo 6. Küme merkezleri.**

Veri madenciliği işlemi MatLab 6.5 yazılımı üzerinde gerçekleştirilmiştir. Uygulamanın sonuçlarının grafik gösterimi için MatLab yazılımı içinde bulunan *MapToolBox* eklentisi kullanılmıştır. Veri tabanında bulunan ve işleme giren her noktanın içinde bulunduğu kümeye göre grafik üzerinde renkli olarak gösterilmesi için MatLab içerisinde ***grafik.m*** isimli bir kod yazılmış ve kodları EK-3' de verilmiştir.

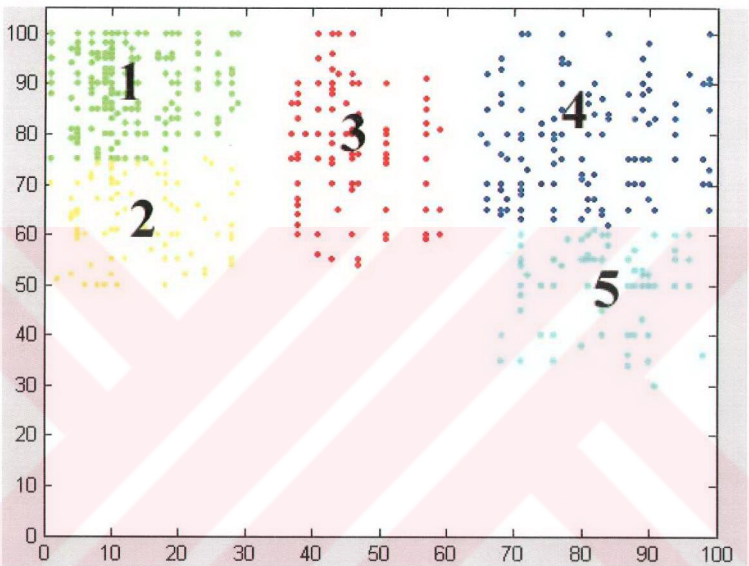
## 3.5.4. BİLGİ SUNUMU

Şekil 26' da veri madenciliği süreci sonunda ortaya çıkan kümeler gösterilmektedir.




**Şekil 26. Kümelerin grafik üzerinde gösterimi. (x-ÖSS yerleşme yüzdesi, y- Başarı notu)**

Yukarıdaki kümeler içerisinde bulunan istisnaların çıkarılmasından sonra aşağıdaki şekil 27 ortaya çıkarılmıştır. Kümeler y ekseninde başarı notu ve x ekseninde ÖSS yerleşme yüzdelerini belirten 2 boyutlu bir grafikte gösterilmiştir.




**Şekil 27. İstisnaların çıkarılmasından sonraki kümelerin gösterimi. (x-ÖSS yerleşme yüzdesi, y- Başarı notu)**

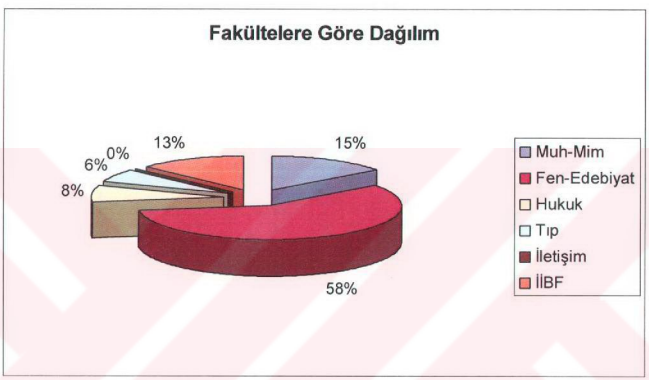
Elde edilen kümelere dahil olan kayıtların bir kaç özellik bakımından da istatistik olarak karşılaştırılması aşağıdaki gibidir.

Kayıtların dahil oldukları kümelerdeki yüzdelik dilimlerine göre dağılımları aşağıdaki şekillerde verilmektedir.

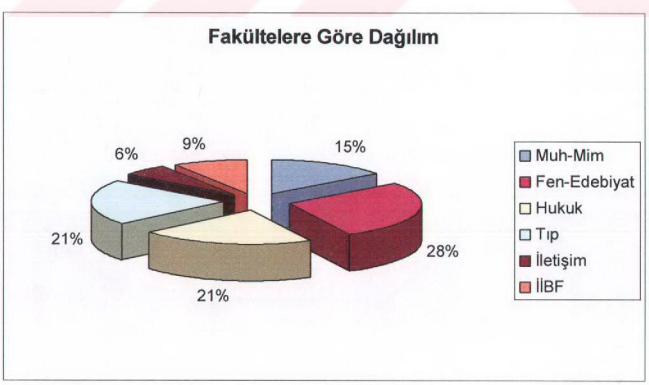
| Yerleşme Yüzdeleri | Küme1 | Küme2 | Küme3 | Küme4 | Küme5 |
|---|---|---|---|---|---|
| **% 0 – 34** | % 100 | % 93 | % 3 | % 0 | % 0 |
| **% 35 – 65** | % 0 | % 7 | % 97 | % 15 | % 1 |
| **% 66 - 100** | % 0 | %0 | % 0 | % 85 | % 99 |

**Tablo7. Kümeler içerisindeki yerleşme yüzdelerine göre dağılım.**

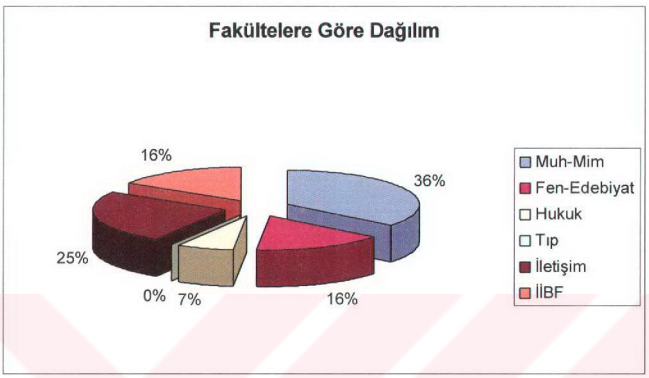
Kayıtların dahil oldukları kümelerdeki fakültelere göre dağılımları aşağıdaki şekillerde verilmektedir.



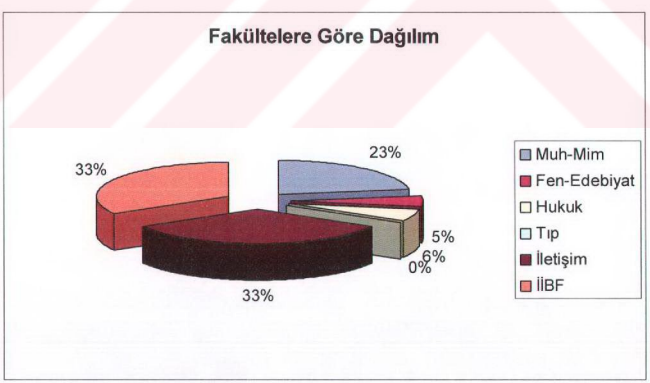

**Şekil 28. Küme 1'in fakültelere göre dağılımı .**



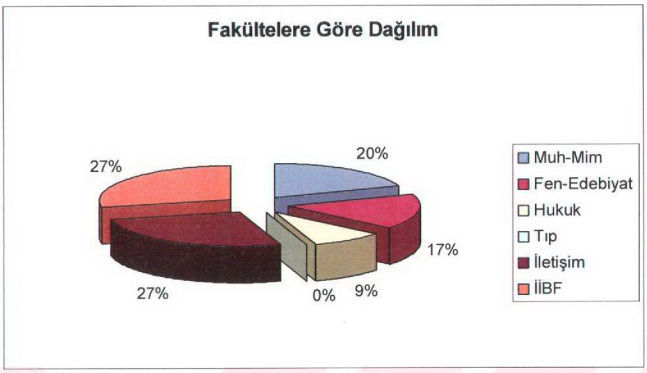

**Şekil 29. Küme 2'nin fakültelere göre dağılımı.**




**Şekil 30. Küme 3'ün fakültelere göre dağılımı.**




**Şekil 31. Küme 4'ün fakültelere göre dağılımı.**




**Şekil 32. Küme 5'in fakültelere göre dağılımı.**

Ortaya çıkan sonuçların ardından kümelerin özelliklerini genel bir yapıda incelersek :

- Kümelerin ortaya çıkarılmasında kullanılan alanlar genel anlamda başarılı olmuştur. Üniversite sınavından aldıkları yüzdelik dilimleri ile üniversite döneminde elde ettikleri başarı notları kümelerin başarılı bir şekilde oluşmasına ve ayrılmasını sağlamıştır.

- Ortaya çıkan kümeler öğrencilerin ÖSS sınavında elde ettikleri yüzdelik dilimlere göre benzer gruplar içerisinde bulunmasını sağlamıştır.

- Küme 1 ve küme 2 yüzdelik dilimlere göre benzerlik taşımaktadır, aynı şekilde küme 4 ve küme5' de benzer özellikleri taşımaktadır. Küme 3 ise tek başına diğer kümelerden ayrı durmaktadır.

- En başarılı küme olarak gözüken küme 1 'in fakültelere göre dağılımına bakarsak en büyük payı Fen-Edebiyat Fakültesinin aldığını görmekteyiz.

Bunun sebebi ise bu fakülteye gelen öğrencilerin çoğunluğu burslu olarak yüksek puanlar almış öğrencilerdir.

- En başarısız küme olarak gözüken küme5' in fakültelere göre dağılımı gözönüne alınırsa en büyük payın İİBF (İktisadi İdari Bilimler Fakültesi) ve İletişim fakültesine ait olduğu görülmektedir. Bu fakültelere gelen öğrencilerin büyük bir kısmı düşük puanlara sahiptir. Sınavda daha az puan almış öğrencilerdir.

# BÖLÜM 4

# SONUÇ

Günümüzde teknolojik gelişmeler tüm sektörleri etkilemiş ve bu sektörler içerisinde yoğun olarak kullanılmaya başlanmıştır. Bu teknolojik gelişmelerin bir parçası olan veri madenciliği, tez kapsamında farklı metodlarıyla beraber incelenmiştir.

Daha çok ticari sektörlerde yoğun olarak kullanılan veri madenciliğini farklı bir sektörde kullanılmasıyla farklı bir bakış açısı getirilmeye çalışılmıştır.

Eğitim sektöründe gerçek veriler üzerinde yapılan veri madenciliği uygulaması ile elde edilmiş olan sonuçlar ile öğrenciler ve üniversiteler açısından değerlendirme imkanı yaratılmış olmaktadır.

Elde edilen sonuçlar ile beraber benzer özellikteki öğrenciler aynı kümeler içerisinde toplanmışlardır.

Veri madenciliği, tez çalışmasının içerisinde de belirtildiği gibi özellikle ticari amaçlar için kullanılmaktadır. Bu uygulama da ise ticari amaçların haricinde farklı amaçlar içinde kullanım alanının olabileceği gösterilmektedir.

Veri madenciliğinin farklı açılardan eğitim sektörüne uygulamalarıyla beraber geleceğe yönelik gerek öğrenci tarafında gerekse eğitim kurumları tarafında bazı yeni uygulamalar ve kararların alınmasında yol göstericilik açısından faydalar

sağlayacağı ve bu şekilde geleceğe yönelik sağlam değerlendirmeler yapılabileceği düşünülmektedir.

Öğrencilerin tüm eğitim dönemleri boyunca elde edilen veriler ışığında yapılacak olan veri madenciliği çalışmalarıyla farklı durumlar araştırılacağı ve eğitim öğretime yönelik yeni uygulama kararları alınabileceği düşünülmektedir.

# KAYNAKÇA


Vahaplar, Alper; İnceoğlu, Mustafa : **"Veri Madenciliği ve Elektronik Ticaret"**, (Çevrimiçi ) http://inet-tr.org.tr/inetconf7/bildiriler/78.doc , 01 Kasım 2003

Akpınar, Haldun : **"Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği"**, İstanbul, İ.Ü. İşletme Fakültesi Dergisi, Sayı:1, Nisan 2000, s.1-22.

Thearling, Kurt : **"An Introduction to Data Mining"**, (Çevrimiçi) http://www.thearling.com/text/dmwhite/dmwhite.htm, 01 Aralık 2003.

Karakaş, Melikşah : **"Veri Madenciliği Üzerine"**, (Çevrimiçi) http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=132, 30Kasım 2003.

Fayyad, U.M.; Piatesky-Shapiro, G.; Smyth, P.; Uthurusamy, R. : **"Advances in data mining and knowledge discovery"**, USA, MIT Press, 1994.

Two Crows Corporation : **"Introduction to Data Mining and Knowledge Discovery"**, USA, Two Crows Corporation, 1999, ISBN:1-892095-02-5

Grossman, Robert L; Kamath, Chandrika; Kumar, Vipin : **"Data Mining For Scientific And Engineering Approach"**, USA, Kluwer Academic Publishers, Ekim 2001, ISBN:1-4020-0033-2.

Fayyad, U.M.; Piatesky-Shapiro, G.; Smyth, P. : **"From Data Mining to Knowledge Discovery in Databases"**, USA, AAAI Pres, 1996.

Etzioni, Oren : **"The World Wide Web:Quagmire or Gold Mire"**, USA, Communications of The ACM, Kasım 1996, S:65-68.

Pokorny, J : **"Data Warehouses: A Modelling Perspective"**, Slovenia, Proceedings of the 7.International Conference on Information Systems Development, Plenum Press, 1998.

| | |
|---|---|
| Sen, A.; Jacob, V.S. | : **"Industrial Strength Data Warehousing"**, Communications of the ACM, 1998. |
| Love, B. | : **"Enterprise Information Technologies"**, Van Nostrand Reinhold, 1993. |
| Inmon, W. | : **"What is A Data Warehouse?"**, Prism Tech Topic, Vol:1, 1992. |
| Han, J; Kamber, W. | : **"Data Mining Concepts and Techniques"**, Morgan Kaufmann Publishers Inc., s. 63 Ağustos 2001. |
| Moody, Daniel L. | : **"From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design"**, Melbourne, Avusturalya, 2000. |
| Linoff, G.; Berry, M.J.A. | : **"Data Mining Techniques for Marketing Sales and Customer Support"**, Wiley Computer Publishing, New York, USA, 1997. |
| Zhu, Hua | : **"On-Line Analytical of Association Rules"**, University of Science and Technology of China, 1995. |
| Tantuğ, Ahmet Cüneyt | : **"Veri Madenciliği ve Demetleme"**, Yüksek Lisans Tezi, İTÜ, Mayıs 2002. |
| Aggarwal,Charu C.; Yu, Philip S. | : **"Outlier Detection for High Dimensional Data"**, IBM T.J. Watson Research Center, NY, 2001. |
| Goldberg, D.E. | : **"Genetic Algorithms in Search, Optimization and ,machine Learning"**, USA, 1989. |
| Karr, C.L.; Freeman M.L. | : **"Industrial Applications of Genetic Algorithms"**, CRC Press, USA, 1999. |
| Emel, Gül Gökay; Taşkın, Çağatay | : **"Genetik Algoritmalar ve Uygulama Alanları"**, Bursa, Uludağ Üniversitesi İİBF Fakültesi Dergisi Cilt XXI, Sayı:1, S:129-152, 2002. |

Öztemel, Ercan : **"Yapay Sinir Ağları"**, s.203-206,İstanbul, Papatya Yayıncılık, Ağustos 2003.

Menteş, Gülçin Tunalı : **"Faktör ve Kümeleme Analizi Yardımıyla Bankacılık Ürün ve Hizmetlerinin Araştırılması Üzerine Bir Uygulama"**, Doktora Tezi, İstanbul, İ.Ü.Sosyal Bilimler Enstitüsü, 2000.

Steinbach, Michael; Ertöz, Levent; Kumar, Vipin : **"The Challenges of Clustering High Dimensional Data"**, t.y.

Çevrimiçi :(Çevrimiçi)http://ucl.ac.uk/oncology/MicroCore/HTML_resource/Distance_detailed_popup.htm, Erişim tarihi 22.03.2004.

Jain, A.K.; Dubes, R.C. : **"Algorithms for Clustering Data"**, Prentice-Hall Advanced Reference Series, Prentice-Hall Inc., New Jerset, 1988.

Pazarlıoğlu, Vedat : **"Kümeleme Analizleri"**, İzmir, 1999.

Aaker, David A.; Day, George S. : **"Marketing Research"**, John Wiley & Sons, Third Edition, New York, 1986, S.481.

Kaufman, L.; Rousseuw, P.J. : **"Finding Groups in Data: An Introduction to Cluster Analysis"**, John Wiley & Sons, New York, 1975.

Çevrimiçi : (Çevrimiçi), http://www.predictivepatterns.com/docs/WebSiteDocs/Introduction/Tutorials/Tutorial_Use_Case_Scenarios.htm /

Zhang, Tian; Ramakrishnan, Raghu; Linvy, Miron : **"BIRCH: An Efficient Data Clustering Method for Very Large Databases"**, SIGMOD, 1996.

Berkhin, Pavel : **"Survey of Clustering Data Mining Techniques"**, Accruse Software Inc., California, USA, 2002.

Guha, S.; Rastogi, R.; Shim, K. : **"CURE: An Efficient Clustering Algorithm for Large Databases"**, ACM SIGMOD, Seattle, USA, 1998.

Hou, Jean Fen Ju : **"Clustering with Obstacle Entities"**, Master tezi, Kanada, Simon Fraser Univerity Computing Science, 1999.

Computer Science Departman : **"Clustering Algorithms Classification"**, Computer Science Departman, Trinity College.

Sturn, Alexander : **"Cluster Analysis for Large Scale Gene Expression Studies"**, Master Tezi, The Institute for Genomic Research (TIGR), USA, 2000.

Sağıroğlu, Şeref : **"Şifrelemede Nörol Yaklaşımlar"**, (Çevrimiçi) http://arsiv.emo.org.tr/Kartus01/SEMPOZYUMLAR/iletisimteknolojilericalistayi/makale_pdf/24.pdf, 17.04.2004.

Akpınar, Haldun : **"Yapay sinir ağları gelişimi ve yapılarının incelenmesi"**, İ.Ü. İşletme Fakültesi Dergisi, C:23, S:1 /Nisan 1994, s.41-78.

Raymond, N.G.; Han, J. : **"Efficient and Effective clustering method for spatial data mining."** Int. Conference Very Large Data Bases (VLDB'94), Santiago, Şili (Eylül 1994) 144-155

**EK- 1**

## STUDENTS TABLO YAPISI

| ALAN ADI | VERI TURU |
|---|---|
| StudentID | float |
| StatusID | float |
| Grade | float |
| LangStatusID | float |
| LangPoint | float |
| YearRegister | float |
| DepartmentID | float |
| StudentNo | nvarchar |
| OrderNo | float |
| PreferenceNo | float |
| SexCode | float |
| HighSchoolID | float |
| HighSchoolTypeID | float |
| HighSchoolBranchTypeID | float |
| YerPuan | float |
| YerPuanBasariSirasi | float |
| YerPuanYuzde | float |
| YerPuanTuru | float |
| FacultyID | float |
| HighSchoolBranchTypeID1 | float |
| HighSchoolBranchTypeCode | float |

**EK-2**

## STUDENTS_LOG TABLOSU

| ALAN ADI | VERİ TÜRÜ |
|---|---|
| StudentLogID | float |
| StudentID | float |
| CourseLogID | float |
| Vize1 | float |
| Vize2 | float |
| Final | float |
| Butunleme | float |
| BasariNotu | float |
| IsMuaf | float |
| CourseStatus | float |
| NumRepeat | float |
| OldStudentLogID | float |
| DevamsizliktanKaldi | float |
| Vize1Onaylandi | float |
| Vize2Onaylandi | float |
| FinalOnaylandi | float |
| ButunlemeOnaylandi | float |

## EK-3

## GRAFIK.M MATLAB DOSYASI

```
k=0;

for i=1:size(data,1),


  if k+1==ans(i,1)
  plot(data(i, 1), data(i, 2), 'b.');hold on;
  elseif  k+2==ans(i)
  plot(data(i, 1), data(i, 2), 'g.');hold on;
  elseif  k+3==ans(i,1)
  plot(data(i, 1), data(i, 2), 'r.');hold on;
  elseif  k+4==ans(i,1)
  plot(data(i, 1), data(i, 2), 'y.');hold on;
  elseif  k+5==ans(i,1)
  plot(data(i, 1), data(i, 2), 'c.');hold on;
  elseif  k+6==ans(i,1)
  plot(data(i, 1), data(i, 2), 'm.');hold on;
  elseif  k+7==ans(i,1)
  plot(data(i, 1), data(i, 2), 'b.');hold on;
  elseif  -1==ans(i,1)

end
end
```