

VERİ MADENCİLİĞİ

T.C.
Dokuz Eylül Üniversitesi
İktisadi ve İdari Bilimler Fakültesi
Yönetim Bilişim Sistemleri Bölümü
E – Kitap Serisi

T.C.
Dokuz Eylül Üniversitesi
İktisadi ve İdari Bilimler Fakültesi
Yönetim Bilisim Sistemleri Bölümü
E – Kitap Serisi

VERİ

MADENCİLİĞİ

Bölüm Başkanı: Prof.Dr. Vahap TECİM

Editör: Prof.Dr. Kaan YARALIOĞLU

Tasarım: Aylin DİZDAROĞLU

İzmir, 2014

İÇERİK

VERİ MADENCİLİĞİ İLE İLGİLİ KAVRAMLAR
Gizem YÜKSEL

VERİ MADENCİLİĞİ SÜRECİ
Aylin DİZDAROĞLU

VERİ MEDENCİLİĞİ YÖNTEMLERİ
Cem YILDIZ

VERİ MADENCİLİĞİ UYGULAMALARI
Gülcihan YÜKSEL

VERİ MADENCİLİĞİ UYGULAMA PROGRAMLARI
Mustafa YURTSEVER

ULAŞIM TÜRÜ SEÇİMİNİ ETKİLEYEN FAKTÖRLERİN
BELİRLENMESİNDE VERİ MADENCİLİĞİ YÖNTEMİNİN
KULLANILMASI
Uzm. L.Özge ORAL





ÖNSÖZ

Elbette ki karar vericilerin donanımları, karar yetenekleri ve kullandıkları her anlamdaki argüman geleceğe dönük doğru kararların verilmesinde önem arz etmektedir. Ancak bir karar verici için bu olgular ne kadar hatasız bir şekilde bir araya getirilse getirilsin geçmişe yönelik veri olmadan da her şey eksik kalacaktır. Ve bu anlamda hemen her dönemde geçmişe ilişkin veriler geleceğe yönelik doğru kararların verilmesinde birincil öneme sahip olmuştur.

Ancak artık, gelişen teknoloji ve buna paralel olarak insan ihtiyaçlarındaki değişimler, sadece verileri doğru ve hızlı toplamayı ve analiz etmeyi yeterli kılmamakta, karar verici için veriler arasındaki ilişkileri de yorumlamak mutlak bir zorunluluk haline gelmektedir. Veri Madenciliği olarak isimlendirilen yöntemler bütünü bu ihtiyaç ve zorunluluğu karşılamak üzere geliştirilmiştir.

Bu anlamda Veri Madenciliği, çok büyük miktardaki verilerin içindeki ilişkileri inceleyerek aralarındaki bağlantıyı bulmaya yardımcı olan ve veri tabanı sistemleri içerisinde gizli kalmış bilgilerin çekilmesini sağlayan veri analizi tekniği olarak tanımlanabilir.

Veri Madenciliğinin gittikçe artan önemine karşısında bu çalışma, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Yönetim Bilişim Sistemleri Anabilim Dalı programında yer alan Veri Madenciliği dersi için hazırlanan sunumlar bir araya getirilerek oluşturulmuştur.

Bu çalışmanın içeriğini oluşturan sunumlar bir anlam bütünlüğü oluşturacak şekilde düzenlenmiştir. Ancak her ne kadar dil ve ifade birliğine dikkat edilmeye çalışılsa da sunumlardaki bazı tekrarlar birbirleriyle çelişmemek şartıyla, söz konusu anlam bütünlüğünü bozmamak adına çalışmadan çıkarılmamıştır.



Bu alıřmada yer alan sunumlar Aylin DİZDAROĐLU, Cem YILDIZ, Gizem YÜKSEL, Gülcihan řENTÜRK ve Mustafa YURTSEVER tarafından hazırlanmıřtır. Ayrıca bilgi ve deneyimlerini bizlerle paylařan Dokuz Eylül Üniversitesi Uzaktan Eėitim Merkezi uzmanı Sayın Özge ORAL'ın sunumuna da bu alıřmada yer verilmiřtir.

Bu alıřmanın, Veri Madenciliėi üzerine alıřacak arařtırmacılara yararlı olmasını diliyoruz.

Prof. Dr. Kaan YARALIOĐLU

VERİ MADENCİLİĞİ İLE İLGİLİ KAVRAMLAR



Gizem
YÜKSEL

GİRİŞ

Bilgisayar sistemleri ile üretilen veriler bir anlam ifade etmek amacıyla belirli bir amaç doğrultusunda işlenmektedir. Bu amaca yönelik işlenen veriler ile bilgi kavramı oluşturulur. Bu şekilde bilgiye dayalı olarak hareket etmek kayıpları önler ve ileriye yönelik durumsal öngörüler ile tedbir almamızı sağlayabilir. İşte bu noktada büyük miktarlardaki verileri işleyebilen, hamveriyi bilgiye dönüştüren ve anlamlı hale getiren tekniklerden yararlanmak gerekir ve bunların başında veri madenciliği gelmektedir. Veri madenciliği, büyük miktarlardaki verinin içinden geleceği tahmin edilmesinde yardımcı olacak anlamlı ve yararlı bağlantı ve kuralların bilgisayar programlarının aracılığıyla aranması ve analizidir (Savaş vd., 2012: 2). Bu açıklamaya dayanarak veri madenciliği özel bir uzmanlık alanıdır denilebilir. Özel bir uzmanlık alanıyla ilgili ilk aşılması gereken bariyerde, kavramlardır denilebilir. Çünkü jargonunu bilmediğiniz bir alan, keşfedilmemiş bir orman gibidir.

Bu çalışmada veri madenciliği konusunu tam olarak anlamak ve dilini öğrenebilmek amacıyla veri madenciliği konusu içerisinde geçen bütün kavramlar alfabetik bir sıra ile incelenmiştir.

VERİ MADENCİLİĞİ İLE İLİŞKİLİ KAVRAMLAR

A

Adli analiz: Adli analiz, normal olmayan ya da sıra dışı veri elemanlarını bulmak için, çıkarılmış desenleri uygulama işlemidir (Usgurlu, 4).

Algoritma: bir programın anlaşılması veya program geliştirmede yardımcı olmaya yönelik gösterim teknikleridir (Özdamar, 2002: 72).

Alternatif Varsayım: Boş varsayım kabul edilmezse başka varsayımlar doğar. Bunlara da alternatif varsayımlar denir (Sıramkaya, 2005: 19).

Analitik Veri: Analitik veri, diğer bir ifade ile çözümsel ya da çözümlemeli veri, farklı olaylar sonucu operasyonel sistemlerde ortaya çıkan işlenmemiş veri veya veri gruplarının belirli bir sorun ya da soruya çözüm sağlamak amacıyla bir araya getirilmiş durumudur (Kolay, 2006: 25).

Anlamlılık Düzeyi: Araştırmacı tarafından testten önce belirlenen bir değerdir ve α ile gösterilir. (Genellikle 0.05 olarak alınır daha hassas bir araştırma için 0.01 veya 0.001 de alınabilir.)

Apriori Algoritması: Apriori algoritmasını, sık geçen örüntülerin yakalanmasında kullanılan temel bir algoritmadır. Bu algoritma özünde yinelemeli bir yaklaşım sunar. Örüntüleri ya da diğer bir deyişle sık geçen öge kümelerini bulmak için birçok kez veritabanını taramak gerekir (Kolay, 2006: 49).

Ardışık Analiz: Ardışık analiz ise birbiriyle ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır (Akbulut, 2006: 25). (“Çadır alan müşterilerin %10’u bir ay içerisinde sırt çantası almaktadır.”)

Artık Veri: Artık veri, problemde istenilen sonucu elde etmek için kullanılan örneklem kümesindeki gereksiz niteliklerdir. Bu durum pek çok işlem sırasında karşımıza çıkabilir (Savaş vd., 2012:6).

Aşağıdan Yukarıya Çözümleme: Bu yaklaşımda ise belirli varsayımlar denenir (Sıramkaya, 2005: 5).

B

Bağımlılık Analizi: Bağımlılık analizi veri öğeleri veya oluşumları arasında anlamlı bir bağımlılığı (birlikteliği) tanımlayan bir modelin bulunmasından meydana gelir (Tüzüntürk, 2010: 76).

Bağımlı Değişkenler: Bağımsız değişkenlere tepki olarak değiştiği gözlenen değişkenlerdir (www.wikipedia.org).

Bağımsız Değişkenler: Bağımlı değişkenlerde bir değişime neden olmak için bilinçli veya kasti bir şekilde manipüle edilen değişkenlerdir (www.wikipedia.org).

Bayes Teoremi: Bayes kuralı (Bayes rule), bir veri grubunda bir özelliğin olasılığını tahmin etme yöntemidir; belirli bir veri değerinde çeşitli varsayımların olasılığını araştırır (Sıramkaya, 2005: 18).

Belirsizlik: Yanlılıkların şiddeti ve verideki gürültünün derecesi ile ilgilidir (Albayrak, 2008: 63).

Bilgi Sistemleri: Bilgi sistemleri, yapay sistemlerdir. Bu sistemler karar verme prosesine yardımcı olmak amacıyla tasarlanmışlardır. Bilgi sistemi; karar verme, işlemlerin kontrolü, problemlerin çözümü ve yeni ürünler ya da hizmetler oluşturmada örgütlerin gereksinim duydukları olan bilgiyi üretir (Tahirov, 2009: 123).

Bilgi Tabanı: Alakalı verilerin saklandığı ortamdır.

Bilgisayar Ağları: Yeni nesil internet, çok yüksek hızları kullanmayı sağlamaktadır. Böyle bir bilgisayar ağı ortamı oluştuktan sonra, dağıtık verileri analiz etmek ve farklı algoritmaları kullanmak mümkün olacaktır (Savaş vd., 2012:6). Bundan 10 yıl önceki bilgisayar ağları teknolojisinde hayal edemediklerimizi artık kullanabiliyoruz. Buna bağlı olarak, veri madenciliğine uygun ağların tasarımı da yapılmaktadır (Akbulut, 2006: 5-6).

Bilimsel Hesaplamalar: Günümüz bilim adamları ve mühendisleri, simülasyonu, bilimin üçüncü yolu olarak görmekteler. Veri madenciliği ve bilgi keşfi, teori, deney ve simülasyonu birbirine bağlamada önemli bir rol almaktadır (Savaş vd., 2012:6).

Birliktelik Analizi: Birliktelik analizi, bir veri kümesindeki kayıtlar arasındaki bağlantıları arayan denetimsiz (unsupervised) veri madenciliği şeklidir (Akbulut, 2006: 25). (“Düşük yağlı peynir ve yağsız süt alan müşteriler %85 olasılıkla diyet süt alırlar.”)

Boş Varsayım: Belirli bir veride sınanacak olan başlangıç varsayımına boş varsayım (null hypothesis) denir (Sıramkaya, 2005: 19).

Boş Veri: Bir veri tabanında boş değer, birincil anahtarda yer almayan herhangi bir niteliğin değeri olabilir. Boş değer, tanımı gereği kendisi de dâhil olmak üzere hiçbir değere eşit olmayan değerdir (Savaş vd., 2012:6).

Bootstrapping: Bootstrapping küçük veri kümeleri için modelin hata düzeyinin tahmininde kullanılan bir başka tekniktir (Akbulut, 2006: 17).

Bölümleme: Veri madenciliği problem tiplerinden bölümleme verinin merak uyandıran ve anlamlı alt gruplara veya sınıflara ayrılmasını amaçlar (Tüzüntürk, 2010: 73).

D

Değişken: Kitle üyelerinin herbirinin belirli bir kesitinin özelliklerinden herbiridir. (<http://istatistik.yasar.edu.tr/wp-content/uploads/2011/10/TSTAT11.pdf>) Bir durumdan diğerine farklı değerler alabilen özelliklerdir.

Denetimli Öğrenme (Supervised Learning): Denetimli öğrenimde amaç; girdi ve çıktı değerleri arasındaki ilişkiyi öğrenmektir. Böylelikle yeni bir girdi değeri için bir çıktı değeri tahminlenebilir (Özdamar, 2002: 10). (Sınıflama algoritmaları, regresyonun dahil olduğu klasik istatistik teknikleri ve yapay sinir ağlarıdır). Denetimli öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunmasıdır. Öğrenme süreci tamamlandığında, tanımlanan kurallar

verilen yeni örneklerle uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir (Akbulut, 2006: 16).

Denetimsiz Öğrenme (Unsupervised Learning): Denetimsiz öğrenimde amaç, girdi değerlerine en uygun gösterim şeklini bulmaktır (Özdamar, 2002: 10). Denetimsiz öğrenimde, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır (Akbulut, 2006: 16).

Destekli Öğrenim (Reinforcement Learning): Destekli öğrenim, hayvanların öğrenme şeklini modelleyen öğrenim yönetimidir. Öğrenci eylemde bulunur, eylemi nedeniyle çevresindeki değişimlerden sadece geribesleme ile bilgi alır (Özdamar, 2002: 10).

Dinamik Veri: Kurumsal çevrim içi veri tabanları dinamiktir ve içeriği sürekli olarak değişir. Bu durum, bilgi keşfi metotları için önemli sakıncalar doğurmaktadır (Savaş vd., 2012:6).

Diskriminant Analizi: Diskriminant analizi, bir dizi gözlemi önceden tanımlanmış sınıflara atayan bir tekniktir. Model, ait oldukları sınıf bilinen gözlem kümesi üzerine kurulur (Akbulut, 2006: 21).

Doğrudan Veri Madenciliği: Bu yöntemde, örneğin hekim, belirli bazı bilgilerin edinilmesiyle (örneğin tıkalı damarların bulunması) ilgilidir (Sıramkaya, 2005: 9).

Dolaylı Veri Madenciliği: Sistemde hiçbir kısıtlamanın ve kullanıcının bulacağı bilgi ya da ne tür bir bilgiyle ilgilenildiğine ilişkin bir beklentinin olmadığı en genel senaryo budur. Bu aynı zamanda en güç yöntemdir (Sıramkaya, 2005: 10).

Donanım: bir enformasyon sisteminde girdi, işleme ve çıktı faaliyetleri için kullanılan fiziksel ekipmandır (Loudon and Loudon, 2011: 20). Veri madenciliği, sayısal ve istatistiksel olarak büyük veri kümeleri üzerinde yoğun işlemler yapmayı gerektirir (Akbulut, 2006: 5). Gelişen bellek ve işlem hızı kapasitesi sayesinde, birkaç yıl önce madencilik yapılamayan veriler üzerinde çalışmayı mümkün hale getirmiştir (Savaş vd., 2012:6).

E

Eksik Veri: Veri kümesinin büyüklüğünden ya da doğasından kaynaklanmaktadır. Eksik veriler olduğunda yapılması gerekenler şunlardır (Savaş vd., 2012:6):

- Eksik veri içeren kayıt veya kayıtlar çıkarılabilir.
- Değişkenin ortalaması eksik verilerin yerine kullanılabilir.

- Var olan verilere dayalı olarak en uygun değer kullanılabilir.

Eksik veriler, yapılacak olan istatistiksel analizlerde önemli problemler yaratmaktadır. Çünkü istatistiksel analizler ve bu analizlerin yapılmasına olanak veren ilgili paket programlar, verilerin tümünün var olduğu durumlar için geliştirilmiştir (Savaş vd., 2012: 7).

Enformasyon(Bilgi): İnsanlar için anlamlı ve kullanılabilir bir şekle getirilen veriye enformasyon demektir (Loudon and Loudon, 2011: 15).

F

Farklı Tipteki Veri: Gerçek hayattaki uygulamalar makine öğreniminde olduğu gibi yalnızca sembolik veya kategorik veri türleri değil, fakat aynı zamanda tamsayı, kesirli sayılar, çoklu ortam verisi, coğrafi bilgi içeren veri gibi farklı tipteki veriler üzerinde işlem yapılmasını gerektirir (Savaş vd., 2012: 7).

G

Genetik Algoritma: karmaşık düzenli problemlerin çözümünü gerçekleştirmek amacıyla, kromozomların yeni diziler üretme esasını temel alan, sezgisel bir araştırma yöntemidir (Özçakar, 1998: 62). Genetik algoritmalar, doğal seçim ilkelerine dayanan bir arama ve optimizasyon yöntemidir (Emel ve Taşkın, 2002: 130).

Gürültü ve Kayıp Değerler: Veri özellikleri ya da sınıflarındaki hatalara gürültü adı verilir (Akbulut, 2006: 12). Büyük veri tabanlarında pek çok niteliğin değeri yanlış olabilir. Veri toplanması esnasında oluşan hatalara ölçümden kaynaklanan hatalar da dâhil olmaktadır. Bu hataların sonucu olarak birçok niteliğin değeri yanlış olabilir ve bu yanlışlardan dolayı veri madenciliği amacına tam olarak ulaşmayabilir (Savaş vd., 2012: 7).

H

Ham Veri: İşlenmemiş veridir.

Hareketli Olmayan Veriler: Veri ambarına aktarılan veriler bir daha değiştirilmeyecek, sadece erişimin ve yeni veri yüklenmesinin mümkün olacağı verilerdir (Kolay, 2006: 27).

Hipotez: Bilimsel yöntemde olaylar arasında ilişkiler kurmak ve olayları bir nedene bağlamak üzere tasarlanan ve geçerli sayılan bir önermedir (www.wikipedia.org).

I

ISODATA: 80'li yılların ortalarında bulunan, hala kullanılan, özellikle görsel veriler üzerinde yapılan işlemlerde sıkça kullanılan bir iyileştirme işlemidir (Özdamar, 2002: 19).

İstatistik: İstatistik verilerin toplanması, sınıflandırılması, özetlenmesi, grafik ve tablolarla sunulması, analiz edilerek ana kütle hakkında anlamlı bilgiler elde edilmesi ve yorumlar yapılmasıdır (Özmen, 2). Veri madenciliğinde ulaşılmak istenen amaç aslında istatistik biliminin amacı ile aynı doğrultudadır. Verilerden bilgiyi keşfetmek (Özmen, 2).

K

Kaba Kümeler: Kaba Küme (KK) kuramı, kümenin tek olarak elemanları ile tanımlandığı ve kümenin elemanları hakkında ilave hiçbir bilginin bulunmadığı klasik küme kuramının aksine, bir kümenin tanımlanması için başlangıçta evrenin elemanları hakkında bazı bilgilere gereksinim olduğu varsayımına dayanan yaklaşımdır (Aydoğan ve Gencer, 18).

Karar Ağacı: Karar Ağacı olası tüm eylem seçeneklerini, bu eylem seçeneklerine etkisi olabilecek tüm olası faktörleri ve tüm bu faktörlere dayanan her bir olası sonucu, verilere bağlı olarak değerlendiren, çizgi, kare, daire gibi geometrik semboller kullanımı yoluyla karar vericiye problemi anlamada kolaylık sağlayan grafiksel bir teknik olarak tanımlanabilir (Lezki ve Er, 2010: 235). (CHAID, C&RT , ID3, C4.5)

Karar Destek Sistemleri: KDS'leri karar vermeyi kolaylaştırmak ve daha etkili hale getirmek için tasarlanan değişik model, simülasyon ve uygulamaları kapsamı içinde bulunduran sistemlerdir (Tahirov, 2009: 126).

Karar Noktası: Karar ağacı üzerinde karar değişkenini temsil eder ve kare biçiminde gösterilen şekildir. Karar noktasında en az iki seçenek arasından bir seçim yapılarak karar verilir. İl karar noktası kök düğüm olarak adlandırılır.

Keşif: Keşif, ne olabileceği konusunda önceden belirlenmiş bir fikir ya da hipotez olmadan, veri tabanı içerisinde gizli desenleri arama işlemidir (Usgurlu, 3).

Kitle: İlgilenilen belli bir büyüklüğe ilişkin eksiksiz sayısal bilgi kümesidir (<http://istatistik.yasar.edu.tr/wp-content/uploads/2011/10/TSTAT11.pdf>) .

Kullanıcı Arayüzü: İstenen bilginin girileceği ve bilgisayarın geri döndürdüğü cevapların sunulacağı kısımdır.

Kümeleme: Kümeleme, verideki benzer kayıtların gruplandırılmasını sağlayan bir tekniktir (Akbulut, 2006: 24). Nesnelerin kendilerini ya da diğer nesnelerle olan ilişkilerini tarif eden bilgileri kullanarak nesneleri gruplara ayırma işlemine kümeleme denir (Özdamar, 2002: 11).

Korelasyon: Korelasyon analizinde ise iki değişkene arasındaki ilişkinin yönü ve şiddeti hesaplanır. Fakat bu ilişki bir neden-sonuç ilişkisi olmak zorunda değildir. Örneğin, horozların sabah ötmeleriyle, güneşin doğması arasında kusursuz doğrusal pozitif korelasyon ilişki vardır. Ancak bu ilişki güneşi horozların doğmasını sağladığını göstermez (Eymen, 2007: 92).

M

Makine Öğrenimi: Kullandığı algoritmaları öğrenim yöntemlerine göre sınıflandırır (Özdamar, 2002: 10). (denetimli, denetimsiz, destekli öğrenim)

Medoid: Kümenin orta noktasıdır ve gerçek bir noktadır (Özdamar, 2002: 13).

Merkeze Dayalı Ayırıcı Küme Teknikleri: Optimal bir kritere göre iteratif olarak veri setini bölümlere ayıran tekniklerdir (Özdamar, 2002: 13). (En sık kullanılanları k-means ve k-medoid yaklaşımlarıdır.)

Metadata: Metadata her veri elementinin anlamını, hangi elementlerin hangileriyle nasıl ilişkili olduğunu ve kaynak verisi ile erişilecek veri gibi bilgileri içermektedir (Usgurlu, 5).

Modelin Kurulması ve Değerlendirilmesi: Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir (Savaş vd., 2012: 8).

Modelin Kullanılması: Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir (Savaş vd., 2012: 8).

Modelin İzlenmesi: Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve yeniden düzenlenmesini gerektirecektir (Savaş vd., 2012: 8).

Model Oluşturma: Machine learning ile kural üretme, sinir ağları ve gerileme (regression) (Sıramkaya, 2005: 9).

N

Naive Bayes: Naive Bayes, hedef değişkenle bağımsız değişkenler arasındaki ilişkiyi analiz eden tahminci ve tanımlayıcı bir sınıflama algoritmasıdır (Akbulut, 2006: 21).

O

Olasılık: Olasılık bir şeyin olmasının veya olmamasının matematiksel değeri veya olabilirlik yüzdesi, değeridir (www.wikipedia.org).

Örnek(lem): Kitleden seçilen bir alt kümedir (<http://istatistik.yasar.edu.tr/wp-content/uploads/2011/10/TSTAT11.pdf>).

Örüntü: Örüntü, bir nesne ve ya olay kümesindeki elemanların ardışık olarak düzenli bir biçimde birbirlerini takip ederek yenilenmesi olarak tanımlanabilir. En basit örnek bir hafta içerisinde ardışık olarak günler ve ya bir sene içerisinde ardışık olarak gelen aylardır. <http://www.nkfu.com/oruntu-nedir-oruntu-ornekleri/>

Örüntü Algılama: Normal EKG dalgalarındaki sapmaların algılanmasından, alışveriş ya da yolculuk örüntülerine kadar değişebilir (Sıramkaya, 2005: 9).

Öz Bilgi: Bilginin hacim olarak iyice küçülmüş ancak kullanım değeri çok artmış bir türüdür (Tüzüntürk, 2010: 66).

Özetlemeye Dayalı Modeller: Özetlemede amaç, veriyi az sayıda özellik/öznelikle karakterize etmektir (Sıramkaya, 2005: 18).

P

Parametre: Kitleye ilişkin sayısal bir ölçüt –ortalama, varyans, oran) (<http://istatistik.yasar.edu.tr/wp-content/uploads/2011/10/TSTAT11.pdf>) (ortalama, varyans, oran)

Parametrik Model: Girdi ile çıktı arasındaki ilişkiyi, bazı değişkenlerin belirlenmediği cebirsel eşitlikleri kullanarak açıklar. Bu belirlenmemiş değişkenler, girdi örnekleri sağlanarak belirlenir (Sıramkaya, 2005: 16).

Parametrik Olmayan Model: Burada, modeli belirlemek için hiçbir eşitlik kullanılmaz. Bu, modelleme işleminin, eldeki veriye uyumlu hale getirilebileceği anlamına gelir (Sıramkaya, 2005: 16-17). Veriye göre bir model oluşturulur. (sinir ağları, karar ağaçları, genetik algoritmalar)

Popülasyon: en basit anlamda birbiriyle etkileşim içinde olan bireylerden oluşan bir topluluktur (Özdamar, 2002: 52).

Problemin Tanımlanması: Veri madenciliği çalışmalarında başarılı olmanın en önemli şartı, projenin hangi işletme amacı için yapılacağını ve elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceğinin tanımlanmasıdır (Savaş vd., 2012: 8).

R

R: Grafikler, istatistiksel hesaplamalar, veri analizleri için geliştirilmiş bir programdır. S diline benzer bir GNU projesidir (Dener vd., 4).

RBF Ağları: Girdi, gizli ve çıktı düğümlerinden oluşan üç katmanlı en basit yapay sinir ağıdır (Özdamar, 2002: 48).

Regresyon Analizi: Regresyon analizi, bir ya da daha fazla bağımsız değişken ile hedef değişken arasındaki ilişkiyi matematiksel olarak modelleyen bir yöntemdir (Akbulut, 2006: 21). Örneğin “yemek yeme” ile “kilo alma” arasındaki ilişki regresyon analizi ile ölçülebilir (Eymen, 2007: 92).

Risk: Risk, belirli bir zaman aralığında, hedeflenen bir sonuca ulaşamama, kayba ya da zarara uğrama olasılığıdır. http://www.kirbas.com/pdf/pdf_ciktisi.php?id=410

S

Sayı Tahmini (Estimation): Bir bağımsız değişkenin (parameter) tahmin etme işlemidir. Bu işlem, ortalama, değişken (variance), standart sapma gibi istatistiksel değerleri tahmin etmek için yapılır. Veri madenciliğinde, eksik değerlerin tahmini için kullanılır (Sıramkaya, 2005: 18).



Sınıflama (Classification): Sınıflama, verinin önceden belirlenen çıktılarına uygun olarak ayrıştırılmasını sağlayan bir tekniktir (Akbulut, 2006: 20). (sınıflama algoritmaları şunlardır, Diskriminant analizi, Naive Bayes, Karar ağaçları, Sinir ağları, Kaba kümeler, Genetik algoritma, Regresyon analizi).

Sınırlı Bilgi: Veri tabanları genel olarak veri madenciliği dışındaki amaçlar için hazırlanmışlardır. Öğrenme görevini kolaylaştıracak bazı özellikler bulunmayabilir (Albayrak, 2008: 63).

Standart Sapma: Bir çalışma grubundaki her bir verinin ortalamaya göre ne kadar uzaklıkta olduğunu, bir diğer deyişle dağılımın ne yaygınlıkta olduğunu gösteren bir ölçüdür (Özbek ve Keskin, 2007: 65).

T

Tahmin Edici Modeller: Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Örneğin, bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler kredi alan müşterinin özellikleri, bağımlı değişken değeri ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun olarak kurulan model, daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılmaktadır (Özekers ve Çamurcu, 2002: 2).

Tanımlayıcı Model: Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. X-Y aralığında geliri ve iki veya daha fazla arabası olan çocuklu aileler ile, çocuğu olmayan ve geliri X-Y aralığından düşük olan ailelerin satın alma örüntülerinin birbirlerine benzerlik gösterdiğinin belirlenmesi tanımlayıcı modellere bir örnektir (Özekers ve Çamurcu, 2002: 2).

Ticari Eğilimler: Günümüzde, işletmeler rekabet ortamında varlıklarını koruyabilmek için daha hızlı hareket etmeli, daha yüksek kalitede hizmet sunmalı, bütün bunları yaparken de minimum maliyeti ve en az insan gücünü göz önünde bulundurmalıdır (Savaş vd., 2012:6). Bu tip hedef ve kısıtların yer aldığı iş dünyasında veri madenciliği, temel teknolojilerden biri haline gelmiştir. Çünkü veri madenciliği sayesinde müşterilerin ve

müşteri faaliyetlerinin yarattığı fırsatlar daha kolay tespit edilebilmekte ve riskler daha açık görülebilmektedir (Akbulut, 2006: 6).

Tümdengelim: Tümdengelim ise, bir takım yasa, ilke yada kurallardan hareket ederek özel bir olayı, durumu yada örnekleri inceleme ve açıklama yöntemidir. Bu niteliği ile tümevarım tam karşıttır. Öğretmende bu yöntemden herhangi bir ilkenin, kuralın yada formülün ne ölçüde gerçekçi olduğunu araştırmada yararlanır (<http://www.mevzuatdergisi.com/2003/01a/03.htm>) .

Tümevarım: Tümevarım, özel durumlarda genel bir sonuca erişmek için yapılan, usavurmaya dayalı bir çalışma ve tartışma yöntemidir. Bu yöntem öğretimde örneklerden, sorunlardan, olaylardan ve özel durumlardan hareket ederek genel sonuçlara, kurallara yada kanılara varmak için kullanılır (<http://www.mevzuatdergisi.com/2003/01a/03.htm>) .

V

Varsayım Deneme ve Varsayımı Daha İyi Hale Getirme: Kullanıcı bazı varsayımlar üretir ve bunun sistem tarafından doğrulanmasını, değiştirilmesini ve daha uygun olabilecek yenilerinin önerilmesini bekler (Sıramkaya, 2005: 10).

Varsayım Sınama: Bu teknik, önce bir varsayım oluşturup sonra da bu varsayımı veri üzerinde sınayarak gözlenen veriyi açıklayan bir model bulmaya çalışır. Bu, önce herhangi bir tahminde bulunmadan gerçek veriden bir model oluşturan birçok veri madenciliği yönteminin tersidir (Sıramkaya, 2005: 19).

Veri: Örgütlerde veya fiziksel çevresinde gerçekleşen olayları temsil eden insanların anlayabileceği ve kullanabileceği şekilde ayarlanıp düzenlemeden önceki ham gerçekler yığınıdır (Loudon and Loudon, 2011: 15).

Veri Ambarları: Veri ambarı, organizasyonun ihtiyaçları ile uyumlu büyük miktarlardaki verinin kolay erişilebilir bir yapıda tutulmasını sağlayan bilgisayar tabanlı depolama sistemleridir (Akbulut, 2006: 3).

Veri Birleştirme: Bu aşamada farklı veri tabanlarındaki verilerin tek bir çatı altında (veri ambarında) birleştirilmesi işlemidir (Tüzüntürk, 2010: 69).

Veri Dönüştürme: Kullanılacak model ve algoritma çerçevesinde verilerin tanımlama veya gösterim şeklinin de değiştirilmesi gerekebilir. Örneğin kredi riski uygulamasında



iş tiplerinin, gelir seviyesi ve yaş gibi değişkenlerin kodlanarak gruplanmasının gerekliliği söz konusudur (Albayrak, 2008: 67).

Veri İndirgeme: Büyük hacimli veri kümesinden daha küçük hacimli veri kümesinin elde edilmesidir (Tüzüntürk, 2010: 69).

Veri Kaynakları: Veri kaynağı, bir araştırmanın, bir tartışmanın, bir muhakemenin temeli olan ana ögenin kaynağıdır. Kitaplar, Makaleler, Veri dosyaları, İlişkisel veritabanı, Veri ambarları, Gelişmiş veritabanları, Film, Elektronik kaynaklar, Dergiler, Danışma Kaynakları, Sözlükler, Ansiklopediler, Almanaklar, İstatistiki kaynaklar, Atlaslar, Biyografik kaynaklar, Kronolojiler vb veri kaynakları olarak kullanılabilir (Berkay vd, 4-5).

Veri Madenciliği: Veri madenciliği; veri ambarlarındaki tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarmak, bunları karar vermek ve eylem planını gerçekleştirmek için kullanma sürecidir (Özmen, 2).

Veri Patlaması: Otomatik veri toplama araçları ve veri tabanı teknolojilerindeki gelişme, veritabanlarında, veri ambarlarında ve diğer bilgi depolarında çok miktarda bilgi depolanması sonucunu doğurmuştur (Sıramkaya, 2005: 4).

Veri Pazarı (Datamart): Datamartlar küçük boyutlu (1-10 GB) bölümsel ambarlardır (Usgurlu, 5).

Veri Tabanı: Veritabanı en genel tanımıyla, kullanım amacına uygun olarak düzenlenmiş veriler topluluğudur. Daha ayrıntılı açıklamak gerekirse Veri tabanı (Database), birbirleriyle ilişkisi olan verilerin tutulduğu, kullanım amacına uygun olarak düzenlenmiş veriler topluluğunun mantıksal ve fiziksel olarak tanımlarının olduğu ve bunların sayısal ortamlarda saklandığı ve gerektiğinde tekrar bir erişime olanak sağlayan, büyük boyutlarda veriler barındıran bilgi depolarıdır (Berkay vd, 4).

Veri Tanımlama ve Özetleme: Veri tanımlama ve özetlemenin amacı verinin basit bir biçimde özelliklerinin az ve öz olarak tanımlanmasıdır (Tüzüntürk, 2010: 73).

Veri Tabanı Teknikleri: Alakalı verileri işlenebilecek hale getirecek olan tekniklerdir.

Veri Temizleme: Eksik verilerin tamamlanması, aykırı değerlerin teşhis edilmesi amacıyla gürültünün düzeltilmesi ve verilerdeki tutarsızlıkların giderilmesi gibi işlemlerden oluşmaktadır (Tüzüntürk, 2010: 69).

Veri Toplama: Tanımlanan problem için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır. Verilerin toplanmasında kuruluşun kendi veri kaynakları dışında, nüfus sayımı, merkez bankası, kara listesi gibi



veri tabanlarından veya veri pazarlayan kuruluşların veri tabanlarından faydalanılabilir (Albayrak, 2008: 66).

Veri Yönetim Teknolojisi: Fiziksel depolama ortamında verinin düzenlenmesini yöneten yazılımı içermektedir (Loudon and Loudon, 2011: 20).

Verilerin Hazırlanması (Ön İşleme): Verilerin hazırlanması aşaması kendi içerisinde toplama, birleştirme ve temizleme, dönüştürme adımlarından meydana gelmektedir (Albayrak, 2008: 66).

Veritabanı Boyutu: Veri tabanı boyutları büyük bir hızla artmaktadır. Veri tabanı algoritması çok sayıda küçük örnekleme ele alabilecek biçimde geliştirilmiştir. Aynı algoritmaların yüzlerce kat büyük örneklerde kullanılabilmesi için çok dikkat gerekmektedir (Savaş vd., 2012: 7).

Veritabanı Yönetim Sistemleri: Bir veri tabanını oluşturmak, saklamak, çoğaltmak, güncellemek ve yönetmek için kullanılan programlara Veri Tabanı Yönetme Sistemi (DBMS – Database Management Systems) adı verilir (Usgurlu, 2).

W

WEKA: WEKA bir proje olarak başlayıp bugün dünya üzerinde birçok insan tarafından kullanılmaya başlanan bir Veri Madenciliği uygulaması geliştirme programıdır. WEKA java platformu üzerinde geliştirilmiş açık kodlu bir programdır (Dener vd., 4).

Y

Yapay Sinir Ağları (YSA): YSA, beyindeki sinirlerin çalışmasını taklit ederek sistemlere öğrenme, genelleme yapma, hatırlama gibi yetenekler kazandırmayı amaçlayan bilgi işleme sistemidir (Saraç, 2004: 9).

Yazılım: Bir enformasyon sisteminde bilgisayar donanımı bileşenlerini kontrol ve koordine eden detaylı, daha önceden programlanmış komutlardır (Loudon and Loudon, 2011: 20).

YALE: YALE üniversitesi bilim adamları tarafından Java dili kullanılarak geliştirilmiş bir yazılımdır.

YSA Ara Katman: Girdi katmanından gelen bilgiler işlenerek çıktı katmanına gönderilirler. Bu bilgilerin işlenmesi ara katmanlarda gerçekleştirilir. Bir ağ içinde birden fazla ara katman olabilir (Saraç, 2004: 21).

YSA Çıktı Katmanı: Bu katmandaki proses elemanları ara katmandan gelen bilgileri işleyerek ağı girdi katmanından sunulan girdi seti için üretmesi gereken çıktıyı üretirler. Üretilen çıktı dış dünyaya gönderilir (Saraç, 2004: 21).

YSA Girdi Katmanı: Bu katmandaki proses elemanları dış dünyadan bilgileri alarak ara katmanlara transfer ederler. Bazı ağlarda girdi katmanında herhangi bir bilgi işleme olmaz (Saraç, 2004: 21).

YSA Geri Beslemeli Ağlar: Bir geri beslemeli sinir ağı, çıkış ve ara katlardaki çıkışların, giriş birimlerine veya önceki ara katmanlara geri beslendiği bir ağ yapısıdır (Saraç, 2004: 28).

YSA İleri Beslemeli Ağlar: İleri beslemeli bir ağda işlemci elemanlar (İE) genellikle katmanlara ayrılmışlardır. İşaretler, giriş katmanından çıkış katmanına doğru tek yönlü bağlantılarla iletilir (Saraç, 2004: 26).

Yukarıdan Aşağıya Çözümleme: Bu yaklaşımda, gizli eğilimleri ve grupları bulma amacıyla ham verinin çözümlemesi yapılır (Sıramkaya, 2005: 5).

Z

Zaman Serileri: Zaman değişkeniyle ilişkili bir değişken hakkında, elde edilen gözlem değerlerini zamana göre sıralanmış olarak gösteren serilere, “zaman serisi” denir (<http://web.sakarya.edu.tr/~adurmus/statistik/acikogretim/unite14.pdf>).

SONUÇ

Veri madenciliği; veri ambarlarındaki tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarmak, bunları karar vermek ve eylem planını gerçekleştirmek için kullanma sürecidir (Özmen, 2). Bu kavramın farklı alanlarda kullanılması nedeniyle süreçte birçok teknik ve yöntem uygulanmaktadır. Bu alanlarda, verilerin analiz edilmesi ve anlamlı bilgiye ulaşılması veri madenciliği ile mümkün olabilmektedir. Bu nedenle veri madenciliği ile ilgili kavramların iyi bilinmesi ileride verilerin daha kolay düzenlenmesi ve değerlendirilmesi açısından son derece önemlidir.

Bu çalışmada veri madenciliği ile ilgili ilk adımı oluşturan kavramlar alfabetik bir sıra ile açıklanmaya çalışılmış ve bu kavramların anlaşılması ile veri madenciliği konusunun yabancıliğin aşılmasına yardımcı olmaya çalışılmıştır.

Kaynakça

Kitaplar

Loudon, C. K.; Loudon, P. J.; **Yönetim Bilişim Sistemleri**, Nobel Akademik Yayıncılık, 2011, Ankara

U. Erman E., **Spss 15.0 Veri Analiz Yöntemleri**, İstatistik Merkez Yayın, 2007

Tezler

Akbulut, S.; **Veri Madenciliği Teknikleri İle Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu**, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2006

Albayrak, M.; **EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci İle Tespiti**, Sakarya Üniversitesi Fen bilimleri Enstitüsü, Doktora Tezi, 2008

Kolay, G.; **İşletmelerde Bilgi Sistemleri Verimliliğini Arttırmada Veri Madenciliği Yöntemi: Bir Simülasyon Çalışması**, Zonguldak Karaelmas Üniversitesi Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 2006

Lezki, Ş.; Er, F.; **Tatil Yeri Kararının Verilmesinde Karar Ağacı ve Etki Diyagramının Uygulanması**, Anatolia: Turizm Araştırmaları Dergisi, S.2, 2010

Özdamar, Ö., E.; **Veri Madenciliğinde Kullanılan Teknikler ve Bir Uygulama**, Mimar Sinan Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2002

Sıramkaya, E.; **Veri Madenciliğinde Bulanık Mantık Uygulaması**, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2005

Usgurlu, Ü., B.; **Veritabanı, Veri Madenciliği, Veri Ambarı, Veri Pazarı**, Başkent Üniversitesi Bilgisayar Mühendisliği

Saraç, T.; **Yapay Sinir Ağları**, Gazi Üniversitesi Endüstri Mühendisliği Bölümü, Seminer Projesi, 2004

Makaleler

- Emel, G., Gökay; Taşkın, Ç.; **“Genetik Algoritmalar ve Uygulama Alanları”**, Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, S.1, 2002
- Özbek, H.; Keskin, S.; **“Standart Sapma mı Yoksa Standart Hata mı?”**, Van Tıp Dergisi, S.2, 2007
- Özçakar, N.; **“Genetik Algoritmalar”**, İ.Ü. İşletme Fakültesi Dergisi, S.1, 1998
- Özekes, S.; Çamurcu, A. Y.; **“Veri Madenciliğinde Sınıflama ve Kestirim Uygulaması”**, Marmara Üniversitesi Fen Bilimleri Dergisi, S.18, 2002
- Savaş, S.; Topaloğlu N.; Yılmaz, M.; **“Veri Madenciliği ve Türkiye’deki Uygulama Örnekleri”**, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, S.21, 2012
- Tüzüntürk, S.; **“Veri Madenciliği ve İstatistik”**, Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, S.1, 2010

İnternet

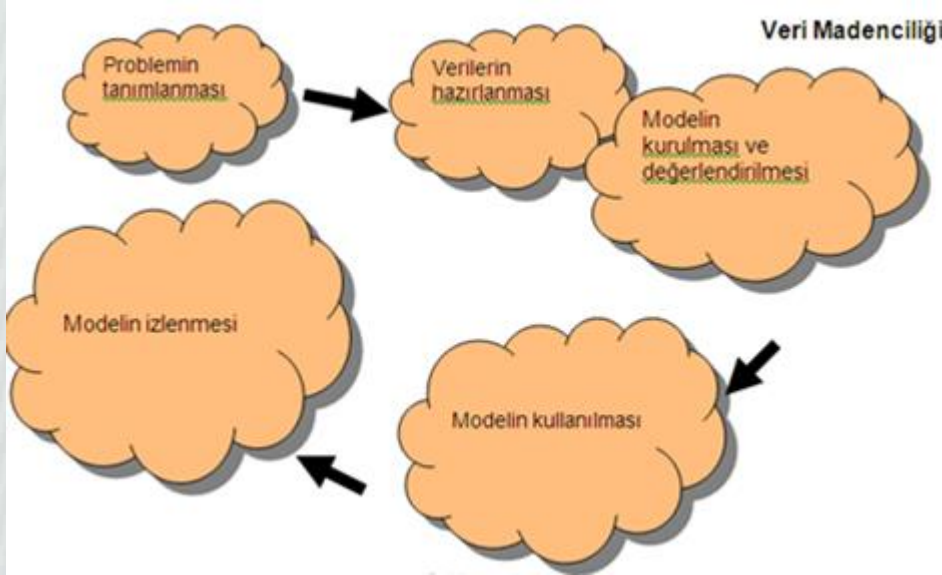
- <http://web.sakarya.edu.tr/~adurmus/statistik/acikogretim/unite14.pdf> (E.T. 20.04.2013)
- <http://www.mevzuatdergisi.com/2003/01a/03.htm> (E.T. 20.04.2013)
- http://www.kirbas.com/pdf/pdf_ciktisi.php?id=410 (E.T. 20.04.2013)
- <http://www.nkfu.com/oruntu-nedir-oruntu-ornekleri/> (E.T. 20.04.2013)
- <http://istatistik.yasar.edu.tr/wp-content/uploads/2011/10/TSTAT11.pdf> (E.T. 20.04.2013)
- http://tr.wikipedia.org/wiki/Ana_Sayfa (E.T. 20.04.2013)
- Aydoğan, Kızılkaya, Emel; Gencer, Cevriye; Kaba Küme Yaklaşımı Kullanılarak Veri Madenciliği Problemlerinde Sınıflandırma Amaçlı Yapılmış Olan Çalışmalar (http://w3.gazi.edu.tr/~ctemel/aydogan&gencer_2007.pdf, E.T. 20.04.2013)
- Can, Berkay, Mehmet; Çamur, Eren; Kuru, Mine; Özkan, Ömer; Rzyeva, Zeynep; Veri Kümelerinden Bilgi Keşfi: Veri Madenciliği (<http://tip.baskent.edu.tr/egitim/mezuniyetoncesi/calismagrp/ogrsmpzsnm14/14.P8.pdf>, E.T. 20.04.2013)
- Dener, Murat; Dörterler, Murat; Orman, Abdullah; Açık Kaynak Kodlu Veri Madenciliği Programları: Weka’da Örnek Uygulama (<http://akademik.maltepe.edu.tr/~kadirerdem/42.pdf>, E.T. 20.04.2013)
- Özmen, Şule; İş Hayatı Veri Madenciliği ile İstatistik Uygulamalarını Yeniden Keşfediyor



(http://www.suleozmen.com/teblig_sunumlar/3is_hayati_veri_madenciligi_istatistik.pdf
(E.T. 20.04.2013)

Tahirov, Arif; Bilgisayar Destekli Bilgi Sistemleri, 2009
(http://journal.qu.edu.az/article_pdf/1041_510.pdf, E.T. 20.04.2013)

VERİ MADENCİLİĞİ SÜRECİ



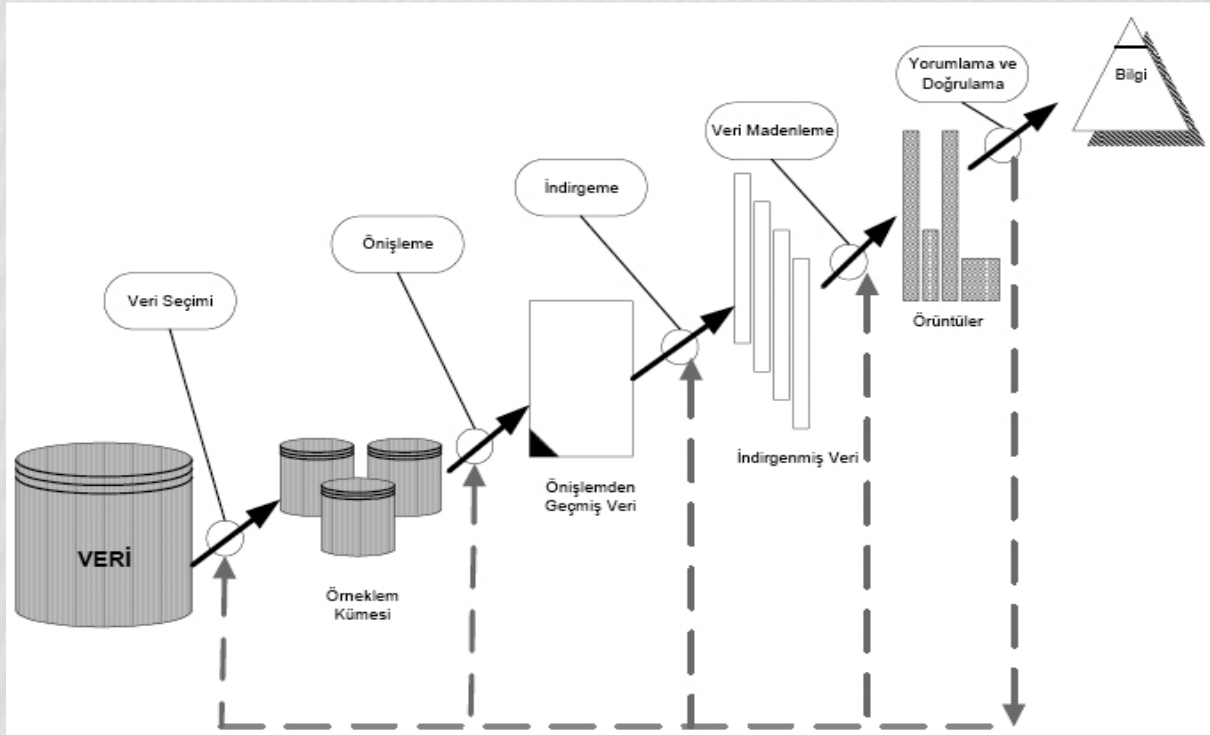
Aylin
DİZDAROĞLU

VERİ MADENCİLİĞİ SÜRECİ

Veri tabanı sistemlerinin artan kullanımı ve hacimlerindeki bu olağanüstü artış, organizasyonları elde toplanan bu verilerden nasıl faydalanılabileceği problemi ile karşı karşıya bırakmıştır. Geleneksel sorgu (Query) veya raporlama araçlarının veri yığınları karşısında yetersiz kalması, Veri Tabanlarında Bilgi Keşfi-VTBK (Knowledge Discovery in Databases) adı altında, sürekli ve yeni arayışlara neden olmaktadır.

Büyük miktarlarda verinin veri tabanlarında tutulduğu bilindiğine göre bu verilerin veri madenciliği teknikleriyle işlenmesine de veri tabanında bilgi keşfi denir. Yeni gelişmekte olan her araştırma dalında olduğu gibi, VTBK'nın tanımı ve faaliyet alanının ne olacağı konusunda farklı görüşler vardır. Bazı kaynaklara göre; VTBK daha geniş bir disiplin olarak görülmekte ve veri madenciliği terimi ise, sadece bilgi keşfi metotlarıyla uğraşan VTBK sürecinde yer alan bir adım olarak nitelendirilmektedir.

VTBK sürecinde yer alan adımlar Şekil 1'de gösterilmiştir:



Şekil 1. VTBK sürecinde yer alan adımlar



VTKB süreci dört adımdan oluşmaktadır:

1. Veri Seçimi: Bu adım veri kümelerinden sorguya uygun verilerin seçilmesidir.

Elde edilen verilere örneklem kümesi denmektedir.

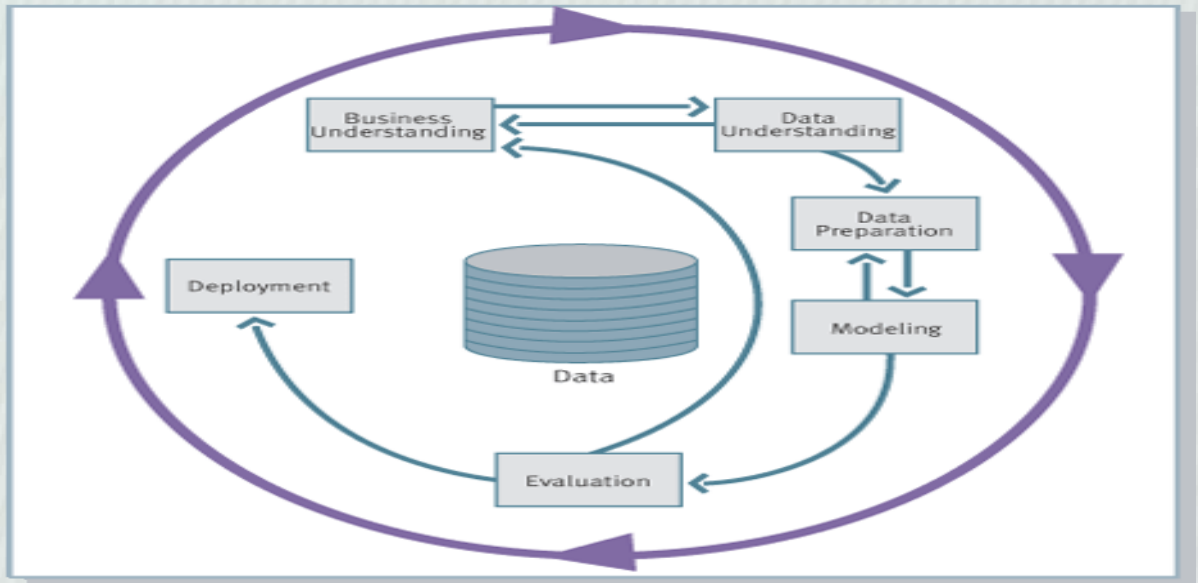
2. Veri Temizleme ve Ön İşleme: Örneklem kümesi elde edildikten sonra, örneklem kümesinde yer alan hatalı tutanakların çıkarıldığı ve eksik nitelik değerlerinin değiştirildiği aşamadır. Bu aşama seçilen veri madenciliği sorgusunun çalışma zamanını iyileştirir.

3. Veri Madenciliği: veri temizleme ve ön işlemden geçen örneklem kümesine VM sorgusunun uygulanmasıdır. Örnek Veri madenciliği sorguları: kümeleme, sınıflandırma, ilişkilendirme, vb. sorgulardır.

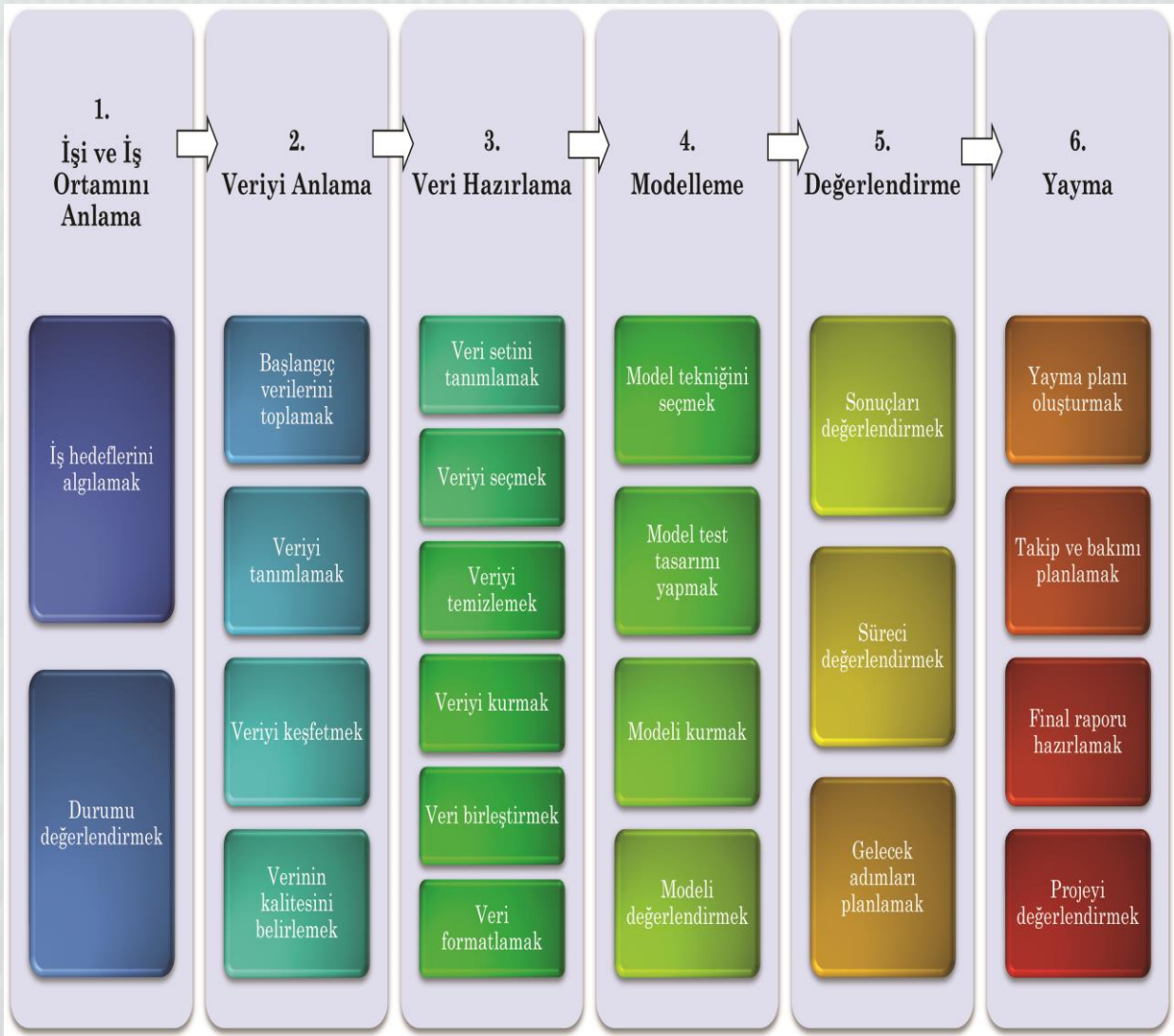
4. Yorumlama: Veri madenciliği sorgularından ortaya çıkan sonuçların yorumlanma kesimidir. Burada geçerlilik, yenilik, yararlılık ve basitlik açılarından üretilen sonuçlar yorumlanır.

Veri madenciliği, veri ambarlarındaki tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş verileri ortaya çıkarmak, bunları karar vermek ve gerçekleştirmek için kullanma sürecidir. Bu tanımdan yararlanarak veri madenciliğinin aynı zamanda bir istatistiksel süreç olduğunu da söylemek mümkündür.

Daha detaylı süreç tanımı ise veri madenciliğinin uluslararası düzeyde standardı olarak kabul edilmiş, CRISP-DM (CRoss Industry Process for Data Mining) ile yapılmaktadır. Veri madenciliği projelerinin hızlı, daha verimli ve daha az maliyetli gerçekleştirilmesi için geliştirilmiş olan bu süreç altı adımdan oluşmaktadır. Bu adımlar Şekil 2 ve Şekil 3’de gösterilmiştir.



Şekil 2. Süreç Adımları



Şekil 3. Süreç Adımları

1. İşi ve İş Ortamını Anlama (Business Understanding):

İlk adım veri madenciliği çalışmasının hangi amaç için yapılacağını net olarak tanımlanmasıdır. Amaç; cevap aranılan sorunun üzerine odaklanmalı, net biçimde ifade edilmeli ve sonuç değerlendirme kriterleri tanımlanmalıdır. Çalışma sonunda doğru cevaplanmış birçok yanlış soru elde edilmek istenmiyorsa, çalışmanın cevap aranılan soru ile uyumlu olması güvence altına alınmalıdır.

a. İş Hedeflerini Algılamak:

Çalışmanın temel amacının belirlenmesi ve bu amacın mümkün olduğunca ikincil amaçlardan ayıştırılarak net olarak tanımlanması aşamasıdır. Çalışma sonuçlarının değerlendirme kriterlerinin belirlenmesi de bu aşamada yapılması gereken diğer bir konudur.

b. Durumu Değerlendirmek:

Veri madenciliğinin temel amacı verim artırmaktır. Bu amaç elde edilecek sonuçlar kadar sürecin kendisi için de geçerlidir. Çalışma sonucunda elde edilecek faydayı değerlendirmek (yanlış kararların maliyetleri ve doğru kararların getirilerine ilişkin öngörüler) önemli bir gerekliliktir. Bu aşamada çalışma için gerekli kaynaklar, tahmini maliyet, mevcut kısıtlar, olası riskler, vb. değerlendirilerek elde edilecek faydanın boyutu ile karşılaştırılır.

2. Veriyi Anlama (Data Understanding):

İkinci adım ilk verilerin toplanması, mevcut verilerin uygunluğunun değerlendirilmesi, modeli oluşturmak için gerekli farklı veri ihtiyaçlarının tespit edilmesi, sahip olunan kayıt sayısının yeterliliği gibi veri kalite ve yeterliliğine yönelik düşünce surecinden geçilmesi aşamasıdır. Hedef çalışmada kullanılacak verilere aşinalık kazanmaktır. Veriyi anlamak ile işi anlamak iç içe geçmiş alt süreçlerdir. İş anladıkça farklı verilere bakmak veya verilerin gösterdiklerini anlamak, verilere baktıkça iş ile ilgili farklı bakış açıları kazanmak mümkündür. Bu döngü kendi içinde devam ettikçe çalışmada kullanılacak verilerin netlik kazanır.

a. Başlangıç Verilerini Toplamak:

Proje kaynaklarında tanımlanmış olan başlangıç verilerinin toplanması aşamasıdır.

b. Veriyi Tanımlamak:

Toplanan verinin tanımlanması ve ihtiyaçları karşılama yeterliliğinin değerlendirilmesi aşamasıdır.

c. Veriyi Keşfetmek:

Başlangıç aşamasında toplanan veriler ile başlangıç hipotezlerinin oluşturulması, limitli bir şekilde veriden çıkarımlar yapılması aşamasıdır. Bu aşamada sonuca yönelik bilgilerin elde edilmesinden daha çok çalışmanın gerçekleştirilebilmesi için veri anlamında eksikliklerin tespit edilmesi amaçlanır.

d. Verinin Kalitesini Belirlemek:

Veri tam mı, doğru mu, hatalar içeriyor mu, hatalar içeriyorsa ne tür hatalar içeriyor, veride eksik bölümler var mı şeklindeki sorular ile verinin kalitesinin tespit edilmesi aşamasıdır.

3. Veri Hazırlama (Data Preparation):

Bu aşama başlangıç verilerinin, çalışmalara temel oluşturacak final verilere dönüştürülmesi aşamasıdır. Bu çalışmanın adımlarının belirgin bir sırası veya tekrar sayısı yoktur. Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olabilmektedir. Bu aşama karar vericinin veri keşfi sürecinin toplamı içerisindeki enerji ve zamanının % 50'sinden fazlasını harcamasına neden olmaktadır.

a. Veri Setini Tanımlamak:

Modelin kurulacağı, tanımlanan soru için gerekli olduğu düşünülen veri setinin (veriler ve bu verilerin toplanacağı veri kaynakları) belirlenmesi aşamasıdır. Bazı durumlarda kurum içinde büyük boyutlarda veri toplanıyor olmasına karşın, toplanmakta olan veriler tanımlanan soru için yeterli olmayıp, başka veri gruplarından eklemeler yapmak gerekebilir. Müşterilerinin yaptığı her işlemi kaydeden, hesaplarının düzeyini, kredi başvuru detaylarını tamamen bilen bir bankanın veri madenciliği çalışmaları için tüm verilere sahip olduğu düşünülebilir. Ancak yapılacak bir pazarlama kampanyasında müşterilerin bireysel ilgi alanları gibi demografik göstergelere göre değerlendirme yapılmak isteniyorsa bu verileri bankanın veri tabanlarındaki operasyonel verilerden sağlamak mümkün değildir. Çalışmalar öncesinde bu veriler doğrudan müşteriler ile temas kurularak ya da bir veri sağlayıcısından temin edilerek tamamlanmalıdır.

b. Veriyi Seçmek:

Yapılacak analizde kullanılacak verilerin belirlenmesi aşamasıdır. Değerlendirme sırasında verinin hedefler ile ilişkisine, kalitesine, teknik limitlere dikkat etmek gerekir. Verilerin değişken sayısı kadar kullanılan kayıt sayısı da önemlidir. Gereğinden az veri, çalışmayı eksik



bırakabileceği gibi, gereğinden fazla veri, veri kirliliğine ve sürecin uzamasına neden olabilecektir.

c. Veriyi Temizlemek:

Gürültülü ve tutarsız verileri çıkararak verinin kalitesini artırma aşamasıdır. Yanlış girişten veya istisnalardan kaynaklanan verilerin, değerlendirmeden çıkarılması tercih edilir. Bazı durumlarda çok büyük veri tabanı ile çalışmak yerine örnekleme yapılması uygun olabilir (seçilen örneklerin tüm popülasyonu temsil düzeyi önemlidir.) Verileri temizlemek sadece çıkarmak gibi düşünülmemelidir, bazı durumlarda eksik verilerin tamamlamak için modellemeler yapmak da söz konusudur.

d. Veriyi Kurmak:

Verileri tanımlayan mevcut değişkenlerde modifikasyonlar yaparak model için daha kullanılabilir değişken setleri oluşturma aşamasıdır. Örneğin müşterilere yapılan satışların aylık olarak kaydedildiği bir yapıda, müşterilerin ortalama sipariş verme sıklığını tanımlamak, mevcut değişkenlerden yeni bir değişken tanımlamaktır. Herhangi bir ayda alım yapmamış müşterilerin listede yer almaması analizlerde yanlış sonuçlar doğurabileceğinden alım yapmayan müşteriler için sıfır miktarlı kayıtlar eklemek yine bu aşamada sık yapılan uygulamalardan biridir.

e. Veri Birleştirmek:

Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olmaktadır. Bu uyumsuzlukların başlıcaları farklı zamanlara ait olmaları, güncelleme hataları, veri formatlarının farklı olması, kodlama farklılıkları (örneğin bir veri tabanında cinsiyet özelliğinin e/k, diğer bir veri tabanında 0/1 olarak kodlanması), farklı ölçü birimleri ve varsayım farklılıklarıdır. Bu adımda farklı kaynaklardan toplanan verilerin uyumsuzluklar mümkün olduğu ölçüde giderilerek, tek bir veri tabanında toplanması amaçlanır. Ancak burada çok dikkatli ve titiz davranmak gereklidir.

Dikkatin en çok gerektiği aşamalardandır. Bu aşamada yapılacak bir hata, ileriki aşamalarda daha büyük sorunlar yaratacaktır. Doğru sonuç alınacak veri madenciliği çalışmaları ancak doğru verilerin üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir.

f. Veri Formatlamak:

Veri seti oluşturulduktan sonra kullanılacak modele göre anlam değişikliği yapmayacak format düzenlemelerinin yapılması aşamasıdır.

4.Modelleme (Modeling):

Benzer veri madenciliği problemleri için birden çok çözüm tekniği olabilmektedir. Bazı teknikler verilerde spesifik ihtiyaçlar duyarlar. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir.

a. Model Tekniğini Seçmek:

Kullanılacak veri madenciliği fonksiyonun ve algoritmasının belirlenmesi aşamasıdır.

Genel olarak verilerin oluşturulma aşamasından itibaren bu konuda bir öngörünün oluşmuş olması gerekir.

b. Model Test Tasarımı Yapmak:

Modeli işletip sonuçları elde etmeye başlamadan önce, modelin kalitesini ve geçerliliğini test etmek gereklidir. Örneğin öngörü fonksiyonlarından sınıflandırma fonksiyonunda hata oranlarını kalite göstergesi olarak kullanılır. Veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenilmesi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenilmesi, öğrenim kümesi kullanılarak gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi belirlenir. Modelde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır.

Sınırlı miktarda veriye sahip olunması durumunda, kullanılabilecek bir yöntem, capraz geçerlilik testidir. Bu yöntemde veri kümesi rasgele iki eşit parçaya ayrılır. İlk aşamada bir parça üzerinde model eğitimi ve diğer parça üzerinde test işlemi; ikinci aşamada ise ikinci parça üzerinde model öğrenimi ve birinci parça üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır.

Model kuruluşu çalışmalarının sonucuna bağlı olarak, aynı teknikle farklı parametrelerin kullanıldığı veya başka algoritma ve araçların denendiği değişik modeller kurulabilir. Model kuruluş çalışmalarına başlamadan önce, hangi tekniğin en uygun olduğuna karar verebilmek güçtür. Bu nedenle farklı modeller kurarak, doğruluk derecelerine göre en uygun modeli



bulmak üzere denemeler yapılmasında yarar bulunmaktadır. Önemli bir diğer değerlendirme kriteri modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, bir çok kuruluş uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir.

Kaldıraç oranı ve grafiği, bir modelin sağladığı faydanın değerlendirilmesinde kullanılan önemli bir yardımcıdır. Örneğin kredi kartını muhtemelen iade edecek müşterilerin belirlenmesi amacını taşıyan bir uygulamada, kullanılan modelin belirlediği 100 kişinin 35'i gerçekten bir süre sonra kredi kartını iade ediyorsa ve tesadüfi olarak seçilen 100 müşterinin aynı zaman diliminde sadece 5'i kredi kartını iade ediyorsa kaldıraç oranı 7 olarak hesaplanmaktadır. Ancak kurulan modelin doğruluk derecesi ne denli yüksek olursa olsun, gerçek dünyayı tam anlamı ile modellediğini garanti edebilmek mümkün değildir.

c. Modeli Kurmak:

Model için kullanılacak algoritmanın/yöntemin/teknikğin hazırlanan veri üzerinde çalıştırılması aşamasıdır. Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak da kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilirler.

d. Modeli Değerlendirmek:

Başarı kriterleri, daha önceki tecrübeler ve test sonuçlarına göre modelin değerlendirilmesi aşamasıdır. Tüm projenin değerlendirilmesinden çok modelin teknik değerlendirilmesi amaçlanır.

5. Değerlendirme (Evaluation):

Bu aşamaya gelindiğinde kurulmuş bir model vardır. Bu aşama, modelin nihai olarak sunulmasından önce modeli yoğun olarak değerlendirilmesi ve iş hedefleri ile uyumlu olup olmadığının kontrol edilmesini amaçlar. Kapsanmamış (açıkta kalmış) konu olup olmadığı değerlendirilmelidir.

a. Sonuçları Değerlendirmek:

On değerlendirme aşamaları modelin geçerliliği ve uygunluğu konusunda değerlendirme sunarken, bu aşamada modelin iş hedeflerini ne ölçüde karşıladığı değerlendirilir. Eğer zaman



ve bütçe varsa gerçek veriler ile modelin test edilmesi tercih edilir. Modelleme sırasında ortaya çıkan ancak ana hedefler ile ilişkisi olmayan diğer ek faydaların da bu aşamada tanımlanması uygun olur.

b. Sureci Değerlendirmek:

Kalite güvence aşamasıdır. Modelin iş hedeflerini karşılamaya yeterli olduğu kararını aldıktan sonra, modelin doğru kurulup kurulmadığı, sadece eldeki verilerden mi yararlanıldığı, gelecekte kullanılabilecek farklı verilerin neler olabileceği gibi konularda değerlendirmeler yapılmalıdır.

c. Gelecek Adımları Planlamak:

Projenin geldiği noktanın yeterli olup olmadığı, ek çalışma gerekliliğinin değerlendirilmesi aşamasıdır. Atılacak başka adımlar nelerdir, bunu gerçekleştirebilecek bütçemiz var mı, eğer devam etmek mantıklı ise nereden devam etmeliyiz gibi konularda değerlendirmeler yapılmalıdır.

6. Yayma (Deployment):

Modelin tamamlanmış olması projenin nihai sonucu değildir. Modelin amacı veriler hakkında bilinenleri artırmak dahi olsa, elde edilen veri kullanılacak biçimde organize edilmeli ve sunulmalıdır. Genellikle gerçek verilerden örneklerin sunulması şeklinde olur.

a. Yayma Planını Oluşturmak:

Sonuçları değerlendirilerek, yayma stratejisinin oluşturulması aşamasıdır.

b. Takip ve Bakımı Planlamak:

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Uzun süre yanlış veri kullanarak çalışmanın önüne geçmek için bakım çok önemlidir.

c. Final Raporu Hazırlamak:

Yapılan çalışmanın başkaları tarafından da tekrarlanabilirliğini sağlamak ve sonuçlarını karar vericilere aktarabilmek üzere hazırlanan rapordur. Çalışmanın üçüncü taraflarca denetlenebilmesini ve güvenilirliğini sağlamak açısından önem taşır.

d. Projeyi Değerlendirmek:



Yapılan çalışmaya dayandırılan kararların ve sonuçların belli bir zaman sonrasında beklentilerle karşılaştırılması ve gerektiğine çalışmanın yenilenmesi aşamalarını içerir.

Kaynakça

Savaş S., Topaloğlu N., Yılmaz M., (2012), Veri Madenciliği Ve Türkiye’deki Uygulama Örnekleri, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, Yıl:11, Sayı: 21, s. 1-23

Akpınar H., (2000), Veritabanlarında Bilgi Keşfi Ve Veri Madenciliği, İ.Ü. İşletme Fakültesi Dergisi, C: 29, s.1-22

Arguden Y., Erşahin B., (2008), Veri Madenciliği Veriden Bilgiye, Masraftan Değere, Arge Danışmanlık Yayınları, No: 10

<http://www.koyuncugil.org/files/ders/bolum6.pdf> Erişim Tarihi: 12.03.2013

http://yzm5550.files.wordpress.com/2012/02/hafta-6-veri-madencilic49fi_giric59f.pdf Erişim Tarihi: 02.05.2013

<ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf> Erişim Tarihi: 10.05.2013



VERİ MADENCİLİĞİ YÖNTEMLERİ



Cem
YILDIZ

VERİ MADENCİLİĞİ YÖNTEMLERİ

Veri madenciliği, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içerir. Başka bir deyişle, veri madenciliği, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların otomatik olarak keşfedilmesidir.

Temel olarak veri madenciliği, geniş veri setleri arasındaki desenlerin ya da düzenin, verinin analizi ve yazılım tekniklerinin kullanılması ile ilgilidir. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirlemekten bilgisayar sorumludur. Amaç, daha önceden fark edilmemiş veri desenlerini tespit edebilmektir.

Veri madenciliğini istatistiksel bir yöntemler serisi olarak görmek mümkün olabilir. Ancak veri madenciliği, geleneksel istatistikten birkaç yönde farklılık gösterir. Veri madenciliğinde amaç, kolaylıkla mantıksal kurallara ya da görsel sunumlara çevrilebilecek nitel modellerin çıkarılmasıdır. Veri madenciliği sahası, istatistik, yapay zeka, veri tabanları ve veri görselleştirme gibi alanlar ile yakından ilişkilidir (Baykasoğlu, 1).

Veri madenciliğinde kullanılmak üzere birçok yöntem ve algoritma geliştirilmiştir ve geliştirilmeye de devam edilmektedir. Bu yöntemlerden en çok kullanılanlarını aşağıdaki gibi sınıflandırmak mümkündür.

- Denetimli ve Denetimsiz Yöntem Ayrımı
- Klasik ve Yeni Nesil Veri Madenciliği Yöntemleri
- Betimsel ve Keşifsel Veri Madenciliği Yöntemleri

Denetimli ve Denetimsiz Yöntemler

Denetimli (Supervised): İyi tanımlanmış veya kesin bir hedef olduğunda denetimli (supervised) ifadesi kullanılır.

- En yakın k komşuluk (k-Nearest-Neighbor)
- K-ortalamlar kümeleme (K-means clustering)

- 
- Regresyon modelleri (Regression models)
 - Birliktelik kuralları (Association rules)
 - Karar ağaçları (Decision trees)
 - Sinir ağları (Neural networks)

Denetimsiz (Unsupervised): Elde edilmesi istenen sonuç için özel bir tanımlama yapılmamışsa veya belirsizlik söz konusu ise denetimsiz (unsupervised) ifadesi kullanılır.

- Aşamalı kümeleme (Hierarchical clustering)
- Kendi kendini düzenleyen haritalar (Self organized maps)

Klasik ve Yeni Nesil Yöntemler

Klasik yöntemlerin başlıcaları,

- Regresyon
- K - En Yakın Komşuluk
- Kümeleme

Yeni nesil yöntemlerin başlıcaları ise,

- Karar Ağaçları,
- Birliktelik Kuralları,
- Sinir Ağları

Betimsel ve Keşifsel Yöntemler

Veri madenciliği yöntemlerini, kullanım amacı bakımından Betimsel ve Keşifsel olmak üzere iki başlık altında da incelemek mümkündür.

Betimsel veri madenciliği yöntemleri, veri içerisindeki mevcut örüntüleri, gruplaşmaları, ayrışmaları, ilişkileri açığa çıkartmak ve veri hakkında genel bir bilgiye sahip olmak amacıyla kullanılır.

- Hiyerarşik Kümeleme Analizi,
- K-ortalamlar Kümeleme Analizi,
- En yakın k komşuluk,
- Birliktelik Kuralları,
- Kohonen Ağları



Keşifsel veri madenciliği yöntemleri ise mevcut örüntüleri, gruplaşmaları, ayrışmaları, ilişkileri temel alarak geleceğe dönük olarak tahmin ve öngörü de bulunmayı sağlar.

- Doğrusal Regresyon Analizi,
- Lojistik Regresyon Analizi,
- K-ortalamlar Kümeleme Analizi,
- En yakın k komşuluk,
- Birliktelik Kuralları,
- Karar ağaçları,
- Sinir ağları,
- Temel Bileşenler Analizi,
- Diskriminant Analizi,
- Faktör Analizi,
- Genetik Algoritmalar (Koyuncugil, 2010: 1-5).

Veri madenciliği yöntemleri incelendiğinde kullanım amacı bakımından kesin bir ayırım yapmak oldukça güçtür. Çünkü, amaca bağlı olarak, bir durumda betimsel amaçlı kullanılan yöntem, başka bir durumda keşifsel amaçlı kullanılabilir. Buna karşın, betimsel ve keşifsel veri madenciliği yöntemlerini genel anlamda ayırmak mümkündür.

Yukarıda belirtilen sınıflandırmaların haricinde Veri Madenciliği daha genel ve yaygın olarak aşağıda belirtilen şekilde de sınıflandırılabilir.

İstatistiksel Yöntemler:

Veri madenciliği temel olarak ileri bir istatistik uygulamasıdır. Verilen bir örnek kümesine bir kestirici oturtmayı amaçlar. İstatistik literatüründe son elli yılda bu amaç için birçok teknik geliştirilmiştir. Bu tekniklerin en önemlileri çok boyutlu analiz başlığı altında toplanır ve genelde verinin parametrik bir modelden geldiğini varsayar. Bu varsayım altında sınıflandırma, regresyon, öbekleme, boyut azaltma, hipotez testi, varyans analizi, bağıntı kurma için teknikler istatistikte uzun yıllardır kullanılmaktadır.

Örnek Temelli Yöntemler:

Bu yöntem istatistikçiler tarafından 1950'li yıllarda önerilmiş olmasına rağmen o yıllarda gerektirdiği hesaplama ve bellek yüzünden kullanılamamış ama günümüzde bilgisayarların

kapasitelerinin artmasıyla, özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla, kullanılabilir olmuştur. Bu yöneme en iyi örnek en yakın k komşu algoritması ve vaka temelli muhakemedir.

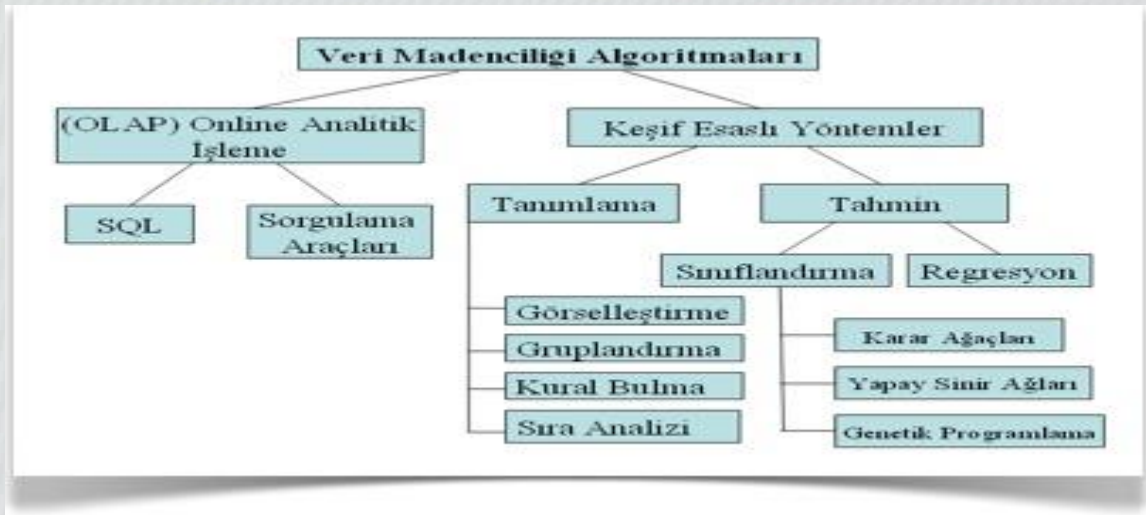
Yapay Zekâ Temelli Yöntemler:

Bu yöntemler veri madenciliğinin beklide en etkili yöntemleri arasında yer alan yapay sinir ağları, genetik algoritmalar, genetik programlama, bulanık mantık gibi modern teknikleri içerir. Yapılan pek çok çalışma bu tekniklerin özellikle doğrusal olmayan, bulanık ve değişken yapıdaki veri uzaylarında klasik istatistik temelli yöntemlere göre çok daha başarılı sonuçlar verdiğini göstermektedir.

Karar Ağaçları:

Karar ağaçları esas itibari ile kural çıkarma algoritmalarıdır. Bu algoritmalar bir veri kümesinden kullanıcıların çok kolay anlayabileceği türündeki kuralları bir ağaç yapısında türetebilirler. Karar ağaçları veri madenciliği uygulamalarında en çok kullanılan ve en fazla yazılımın bulunduğu algoritmalar kümesini oluşturur.

Veri madenciliğinde kullanılan algoritmaları yöntem bağımsız olarak sınıflandırılacak olursa Şekil 1'deki gibi bir sınıflandırma yapılabilir. Aşağıda bu yöntemlerden bazıları açıklanmıştır.



Şekil 1. Veri Madenciliğinde Kullanılan Algoritmalar

Regresyon Analizi

Regresyon analizi, aralarında sebep-sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi belirlemek ve bu ilişkiyi kullanarak o konu ile ilgili tahminler (estimation) ya da kestirimler (prediction) yapabilmek amacıyla yapılır. Doğada birçok olayda sebep sonuç ilişkisine rastlamak mümkündür.

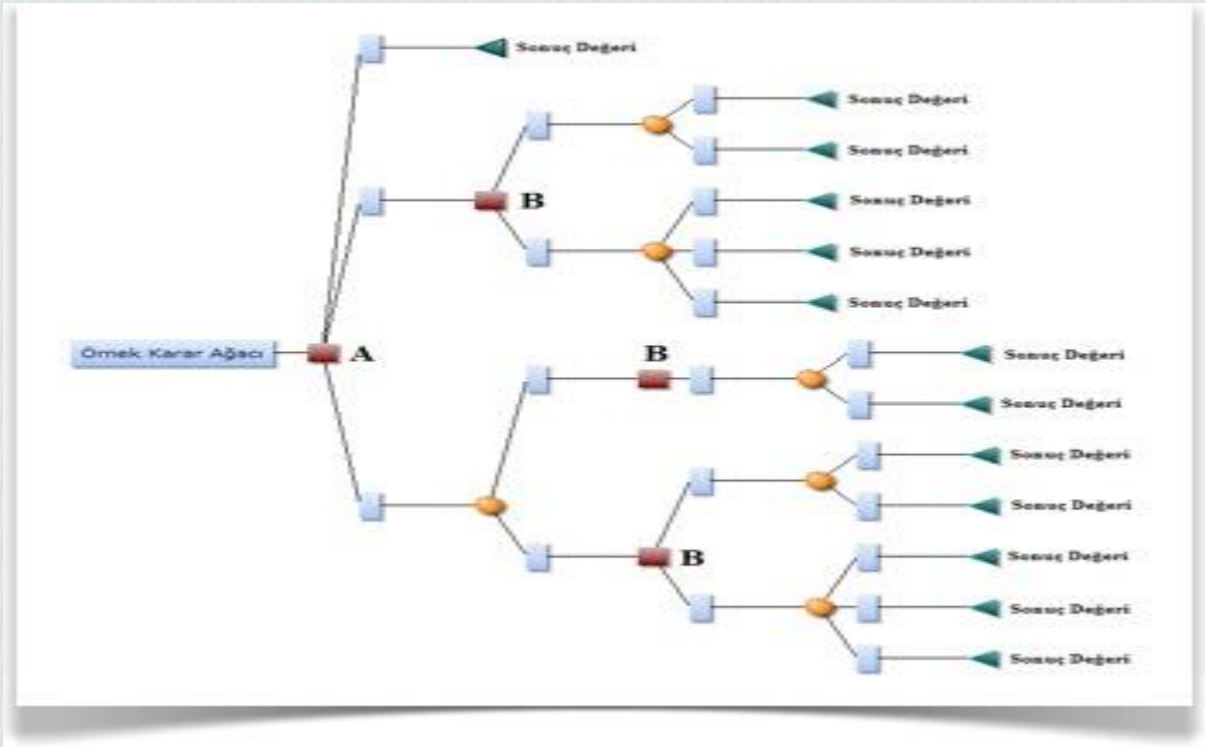
Bu analiz tekniğinde iki (basit regresyon) veya daha fazla değişken (çoklu regresyon) arasındaki ilişki açıklamak için matematiksel bir model kullanılır ve bu model regresyon modeli olarak adlandırılır.

Karar Ağaçları

Olası tüm eylem seçeneklerini, bu eylem seçeneklerine etkisi olabilecek tüm olası faktörleri ve tüm bu faktörlere dayanan her bir olası sonucu, verilere bağlı olarak değerlendiren çizgi, kare, daire gibi geometrik semboller kullanımı yoluyla karar vericiye problemi anlamada kolaylık sağlayan grafiksel bir teknik olarak tanımlanabilir.

Karar problemlerinde söz konusu olan diğer bir özellik problemin simetrik ya da asimetrik yapıda olmasıdır. Karar problemlerinin karar ağacı gösteriminde, başlangıç düğümünden bitiş düğümüne uzanan her bir yol senaryo olarak adlandırılır. Karar ve şans düğümlerinin ardışıklığı tüm senaryolarda aynı olduklarında karar problemi simetriktir. Buna karşın karar ağacında yer alan farklı senaryolarda karar ve şans düğümlerinin ardışıklığı farklı gösteriyorsa problem asimetrik karar problemidir.

Şekil 2’de verilen Karar Ağacı A ve B ile gösterilen iki farklı kararın ardışık olarak alındığı çok aşamalı bir karar problemine ilişkindir. Şekilde, her senaryoda yer alan düğümlerin ardışıklığı aynı olmadıktan asimetrik bir karar problemi olduğu da gözlenebilmektedir. Herhangi bir karar problemine ilişkin karar ağacı gösteriminin tamamlanmasının ardından, her bir bitiş değeri, şans ve karar düğümlerinin beklenen değerlerinin hesaplanması ile çözülebilir (<http://www.istatistikciler.org/dergi/IstDer100204.pdf>).



Şekil 2. Ardışık Asimetrik Karar Problemine İlişkin Karar Ağacı

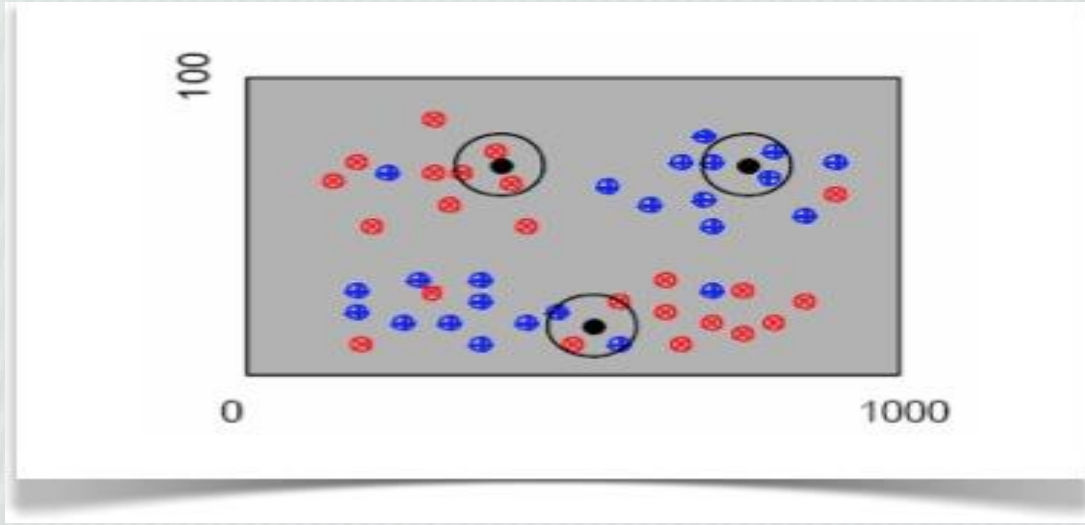
Ayrırma Analizi

Ayrırma analizi, iki veya daha fazla sayıdaki grubun ayırımı ile ilgilenen çok değişkenli ilgi analizidir. Amaçları arasında, analiz öncesi tanımlanmış iki veya daha fazla sayıda grubun ortalama nitelikleri arasında önemli farkların olup olmadığının test edilmesi, gruplar arasındaki farka her bir değişkenin katkısının saptanması ve grup içi değişime oranla gruplar arasındaki ayırımı maksimize eden tahmin değişkenleri kombinasyonunun belirlenmesi sayılabilir.

Örneğin, bira içenleri, bira içmeyenlerden ayırt etmenin bir pazarlama sorunu olduğu kabul edilirse, büyük bir bira üreticisinin yaptığı araştırma ayrırma analizine örnek olarak gösterilebilir. Bu nedenle, tesadüfi olarak seçilen 500 kişilik bir tüketici bölümünü örnek olarak alınmış ve bu kişilerin bira içip içmedikleri, cinsiyetleri ve sporla ilgilenme dereceleri saptanmıştır. Cinsiyet ve sporla ilgilenmenin tahmin değişkenleri olarak kullanılmalarının nedeni, daha önceki çalışmaların bu değişkenlerle bira içme arasında kuvvetli bir ilginin olduğunu göstermiş olmasıdır. Ayrırma analizi sonuçlarının test edilme olanağının bulunması sonuçların geçerliliğini ve güvenilirliğini ve dolayısıyla analizin gücünü artıran önemli bir etmendir.

K – Yakın Komşuluğu

Şekil 3’de K Yakın Komşuluğunun yapısı gösterilmiştir. Kayıtlar, bir veri uzayındaki noktalar olarak düşünülürse, birbirine yakın olan kayıtlar, birbirinin civarında (yakın komşusu) olur. K-yakın komşuluğunda temel düşünce “komşunun yaptığı gibi yap”tır. Eğer belirli bir kişinin davranışı tahmin edilmek isteniyorsa, veri uzayında o kişiye yakın, örneğin on kişinin davranışlarına bakılır. Bu on kişinin davranışlarının ortalaması hesaplanır ve bu ortalama belirlenen kişi için tahmin olur. K-yakın komşuluğunda, K harfi araştırılan komşuların sayısıdır. 5- yakın komşuluğunda, 5 kişiye ve 1-yakın komşuluğunda 1 kişiye bakılır.



Şekil 3. K-Yakın Komşuluğu Yapısı

Birliktelik Kuralları

Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama amaçlı olarak pazar sepeti analizi adı altında veri madenciliğinde yaygın olarak kullanılmaktadır.

Pazarlama araştırmalarındaki hedef, müşteri profillerini açık olarak belirlemektir. Örneğin, spor bir arabaya sahip olan ve evinde köpek besleyen kadınların %90’ı Chanel marka parfüm kullanırlar. Veri madenciliği teknikleriyle bu tür ilişkiler bulunabilir. Bu ilişkiler veri madenciliğinde birliktelik kuralı olarak adlandırılır. Bu örnek için müşterilerin cinsiyeti, sahip oldukları araba türü ve hayvan besleyip beslemedikleri ve satın almaktan hoşlandıkları



ürünlerden oluşan bir veri tabanında oluşturulacak kural: cinsiyeti bayan, arabası spor ve hayvan besleyen kayıtların %90'ı Chanel marka parfüm kullanabilir.

Bir birliktelik algoritması oluşturmada önce kurallar belirlenmelidir. Büyük veri tabanında ilişkileri bulacak algoritmalar geliştirmek çok zor değildir. Fakat, geliştirilen algoritmalar önemli ilişkileri ortaya çıkaracağı gibi önemsiz bir çok ilişkiyi de ortaya çıkarır. Bu yüzden, büyük veri tabanlarında küçük alt kümeler bulunmalıdır. Ayrıca, büyük veri tabanlarında çok sayıda ilişki bulunabileceğinden, birliktelik kuralları sayısı da sınırsız olabilir. Dolayısıyla ilginç ilişkilerle önemsiz ilişkilerin ayrılması gerekir.

Yapay Sinir Ağları (YSA)

İnsan beyninin bilgi işleme uygulamasından esinlenerek geliştirilmiş bir teknolojidir. YSA ile basit biyolojik sinir sisteminin çalışma şekli simüle edilir (benzetilir). Simüle edilen sinir hücreleri [nöronlar](#) içerirler ve bu nöronlar çeşitli şekillerde birbirlerine bağlanarak ağı oluştururlar. Bu ağlar öğrenme, hafızaya alma ve veriler arasındaki ilişkiyi ortaya çıkarma kapasitesine sahiptirler. Diğer bir ifadeyle, YSA'lar, normalde bir insanın düşünme ve gözlemlemeye yönelik doğal yeteneklerini gerektiren problemlere çözüm üretmektedir. Bir insanın, düşünme ve gözleme yeteneklerini gerektiren problemlere yönelik çözümler üretebilmesinin temel sebebi ise insan beyninin ve dolayısıyla insanın sahip olduğu yaşayarak veya deneyerek öğrenme yeteneğidir.

Biyolojik sistemlerde öğrenme, nöronlar arasındaki [sinaptik](#) (synaptic) bağlantıların ayarlanması ile olur. Yani, insanlar doğumlarından itibaren biryaşayarak öğrenme süreci içerisine girerler. Bu süreç içinde beyin sürekli bir gelişme göstermektedir. Yaşayıp tecrübe ettikçe sinaptik bağlantılar ayarlanır ve hatta yeni bağlantılar oluşur. Bu sayede öğrenme gerçekleşir. Bu durum YSA için de geçerlidir. Öğrenme, eğitme yoluyla örnekler kullanarak olur; başka bir deyişle, gerçekleşme girdi/çıktı verilerinin işlenmesiyle, yani eğitme algoritmasının bu verileri kullanarak bağlantı ağırlıklarını bir yakınsama sağlanana kadar, tekrar tekrar ayarlamasıyla olur.

YSA'lar, ağırlıklandırılmış şekilde birbirlerine bağlanmış birçok işlem biriminden (nöronlar) oluşan [matematiksel](#) sistemlerdir. Bir işlem birimi, aslında sık sık transfer fonksiyonu olarak anılan bir denklemdir.



Bu işlem birimi, diğer nöronlardan sinyalleri alır; bunları birleştirir, dönüştürür ve sayısal bir sonuç ortaya çıkartır. Genelde, işlem birimleri kabaca gerçek nöronlara karşılık gelirler ve bir ağ içinde birbirlerine bağlanırlar; bu yapı da sinir ağlarını oluşturmaktadır.

Sinirsel (neural) hesaplamanın merkezinde dağıtılmış, adaptif ve doğrusal olmayan işlem kavramları vardır. YSA'lar, geleneksel işlemcilerden farklı şekilde işlem yapmaktadırlar. Geleneksel işlemcilerde, tek bir merkezi işlem birimi her hareketi sırasıyla gerçekleştirir. YSA'lar ise herbiri büyük bir problemin bir parçası ile ilgilenen, çok sayıda basit işlem birimlerinden oluşmaktadır. En basit şekilde, bir işlem birimi, bir girdiyi bir ağırlık kümesi ile ağırlıklandırır, doğrusal olmayan bir şekilde dönüşümünü sağlar ve bir çıktı değeri oluşturur. İlk bakışta, işlem birimlerinin çalışma şekli yanıltıcı şekilde basittir. Sinirsel hesaplamanın gücü, toplam işlem yükünü paylaşan işlem birimlerinin birbirleri arasındaki yoğun bağlantı yapısından gelmektedir. Bu sistemlerde geri yayılım metoduyla daha sağlıklı öğrenme sağlanmaktadır.

Çoğu YSA'da, benzer karakteristiğe sahip nöronlar tabakalar halinde yapılandırılırlar ve transfer fonksiyonları eş zamanlı olarak çalıştırılırlar. Hemen hemen tüm ağlar, veri alan nöronlara ve çıktı üreten nöronlara sahiptirler.

YSA'nın ana ögesi olan matematiksel fonksiyon, ağın mimarisi tarafından şekillendirilir. Daha açık bir şekilde ifade etmek gerekirse, fonksiyonun temel yapısını ağırlıkların büyüklüğü ve işlem elemanlarının işlem şekli belirler. YSA'ların davranışları, yani girdi veriyi çıktı veriyeye nasıl ilişkilendirdikleri, ilk olarak nöronların transfer fonksiyonlarından, nasıl birbirlerine bağlandıklarından ve bu bağlantıların ağırlıklarından etkilenir.

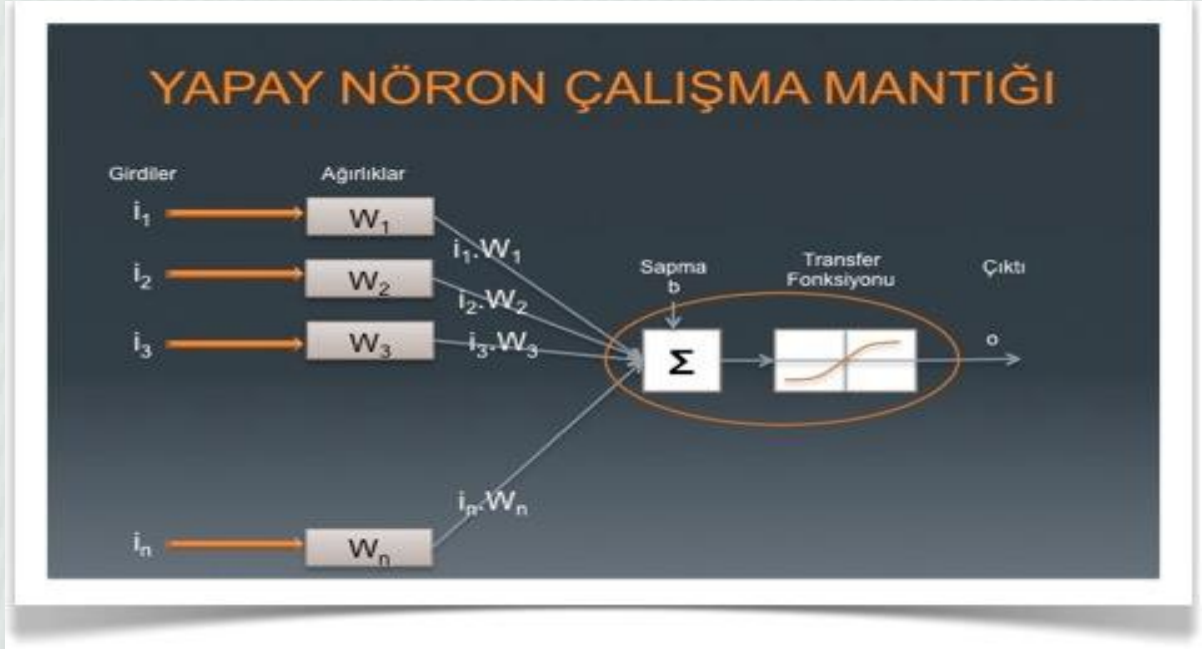
Yapay sinir ağlarının üstünlüklerinin yanı sıra bazı sakıncaları da vardır. Bu sakıncalar şu şekilde listelenebilir (<http://volkansoner.wordpress.com/tag/yapay-sinir-aglari/>).

Sistem içerisinde ne olduğu bilinemez.

Bazı ağlar hariç kararlılık analizleri yapılamaz.

Farklı sistemlere uygulanması zor olabilir.

Başka bir bakış açısıyla Yapay Sinir Ağları teknolojisi insanlığın doğayı araştırma ve taklit etme çabalarının en son ürünlerinden bir tanesi olarak da yorumlanabilir.



Şekil 4. Yapay Sinir Ağları Çalışma Mantığı

Yapay sinir hücreleri, genel özellikleri ile bir yapay hücre modeli (Şekil 4):

Girdiler, ağırlıklar, birleştirme fonksiyonu, aktivasyon (etkinleştirme) fonksiyonu ve çıktılar olmak üzere beş bileşenden meydana gelir. Bu bileşenlerden birincisi, girdiler diğer hücrelerden ya da dış ortamlardan hücreye giren bilgilerdir. Girdiler ($i_1, i_2, \dots, i_n$) kendinden önceki sinirlerden veya dış dünyadan sinir ağına gelebilir. İkincisi, ağırlıklar ($W_1, W_2, \dots, W_n$) yapay sinir tarafından alınan girişlerin sinir üzerindeki etkisini belirleyen uygun katsayılarıdır.

En basit şekilde, bir işlem elemanı, bir girdiyi bir ağırlık kümesi ile ağırlıklandırır, doğrusal olmayan bir şekilde dönüşümünü sağlar ve bir çıktı değeri oluşturur (<http://matlabdersi.com/?p=10>).

Yapay Sinir Ağlarının Özellikleri

YSA'nın hesaplama ve bilgi işleme gücünü, paralel dağılmış yapısından, öğrenebilme ve genelleme yeteneğinden aldığı söylenebilir. Genelleme, eğitim ya da öğrenme sürecinde karşılaşılmayan girişler için de YSA'nın uygun tepkileri üretmesi olarak tanımlanır. Bu üstün özellikleri, YSA'nın karmaşık problemleri çözebilme yeteneğini gösterir. Günümüzde birçok bilim alanında YSA, aşağıdaki özellikleri nedeniyle etkin olmuş ve uygulama yeri bulmuştur .



Doğrusal Olmama : YSA' nın temel işlem elemanı olan hücre doğrusal değildir. Dolayısıyla hücrelerin birleşmesinden meydana gelen YSA da doğrusal değildir ve bu özellik bütün ağı yayılmış durumdadır. Bu özelliği ile YSA, doğrusal olmayan karmaşık problemlerin çözümünde en önemli araç olmuştur.

Öğrenme : YSA' nın arzu edilen davranışı gösterebilmesi için amaca uygun olarak ayarlanması gerekir. Bu, hücreler arasında doğru bağlantıların yapılması ve bağlantıların uygun ağırlıklara sahip olması gerektiğini ifade eder. YSA' nın karmaşık yapısı nedeniyle bağlantılar ve ağırlıklar önceden ayarlı olarak verilemez yada tasarlanamaz. Bu nedenle YSA, istenen davranışı gösterecek şekilde ilgilendiği problemten aldığı eğitim örneklerini kullanarak problemi öğrenmelidir.

Genelleme : YSA, ilgilendiği problemi öğrendikten sonra eğitim sırasında karşılaşmadığı test örnekleri için de arzu edilen tepkiyi üretebilir. Örneğin, karakter tanıma amacıyla eğitilmiş bir YSA, bozuk karakter girişlerinde de doğru karakterleri verebilir yada bir sistemin eğitilmiş YSA modeli, eğitim sürecinde verilmeyen giriş sinyalleri için de sistemle aynı davranışı gösterebilir.

Uyarlanabilirlik : YSA, ilgilendiği problemdeki değişikliklere göre ağırlıklarını ayarlar. Yani, belirli bir problemi çözmek amacıyla eğitilen YSA, problemdeki değişimlere göre tekrar eğitilebilir, değişimler devamlı ise gerçek zamanda da eğitime devam edilebilir. Bu özelliği ile YSA, uyarlamalı örnek tanıma, sinyal işleme, sistem tanılama ve denetim gibi alanlarda etkin olarak kullanılır.

Hata Toleransı : YSA, çok sayıda hücrenin çeşitli şekillerde bağlanmasından oluştuğundan paralel dağılmış bir yapıya sahiptir ve ağına sahip olduğu bilgi, ağıdaki bütün bağlantılar üzerine dağılmış durumdadır. Bu nedenle, eğitilmiş bir YSA'nın bazı bağlantılarının hatta bazı hücrelerinin etkisiz hale gelmesi, ağına doğru bilgi üretmesini önemli ölçüde etkilemez. Bu nedenle, geleneksel yöntemlere göre hatayı tolere etme yetenekleri son derece yüksektir.

Donanım ve Hız : YSA, paralel yapısı nedeniyle büyük ölçekli entegre devre (VLSI) teknolojisi ile gerçekleştirilebilir. Bu özellik, YSA'nın hızlı bilgi işleme yeteneğini artırır ve gerçek zamanlı uygulamalarda arzu edilir.

Analiz ve Tasarım Kolaylığı : YSA' nın temel işlem elemanı olan hücrenin yapısı ve modeli, bütün YSA yapılarında yaklaşık aynıdır. Dolayısıyla, YSA'nın farklı uygulama alanlarındaki yapıları da standart yapıdaki bu hücrelerden oluşacaktır. Bu nedenle, farklı uygulama alanlarında kullanılan YSA'ları benzer öğrenme algoritmalarını ve teorilerini paylaşabilirler. Bu özellik, problemlerin YSA ile çözümünde önemli bir kolaylık getirecektir.

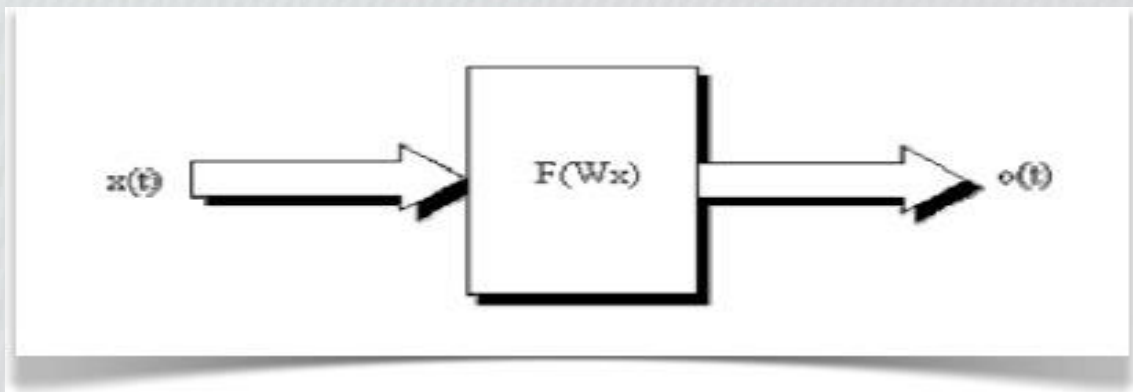
Yapay Sinir Ağlarının Sınıflandırılması

YSA'lar, genel olarak birbirleri ile bağlantılı işlemci birimlerden veya diğer bir ifade ile işlemci elemanlardan (nöron) oluşurlar. Her bir sinir hücresi arasındaki bağlantıların yapısı ağın yapısını belirler. İstenilen hedefe ulaşmak için bağlantıların nasıl değiştirileceği öğrenme algoritması tarafından belirlenir. YSA'lar yapılarına ve öğrenme algoritmalarına göre sınıflandırılırlar.

YSA'ların Yapılarına Göre Sınıflandırılması :

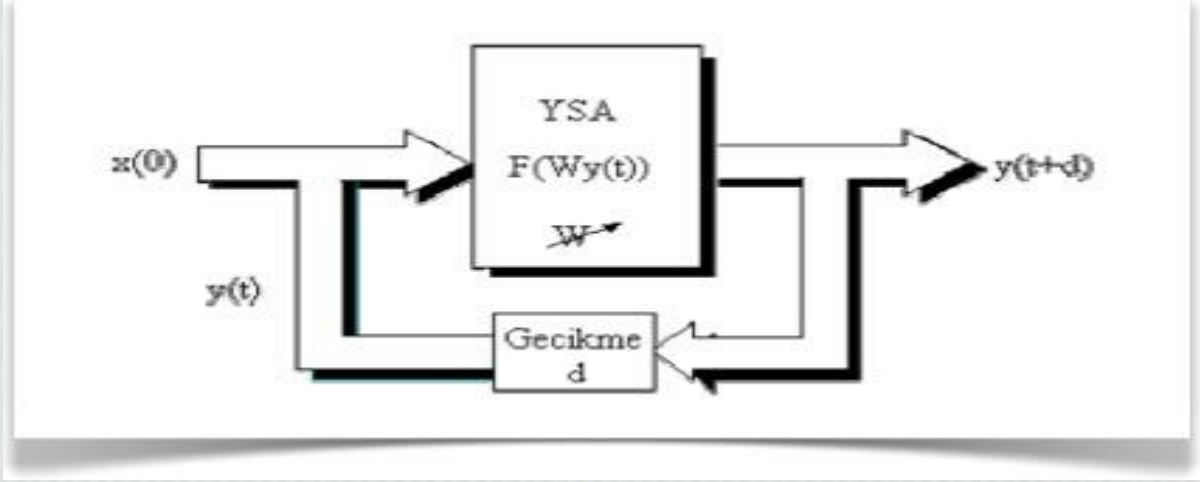
Yapay sinir ağları, yapılarına göre, ileri beslemeli ve geri beslemeli ağlar olmak üzere iki şekilde sınıflandırılırlar.

İleri Beslemeli Ağlar : İleri beslemeli bir ağda işlemci elemanlar (İE) genellikle katmanlara ayrılmışlardır. İşaretler, giriş katmanından çıkış katmanına doğru tek yönlü bağlantılarla iletilir. İE'ler bir katmandan diğer bir katmana bağlantı kurarlarken, aynı katman içerisinde bağlantıları bulunmaz (Şekil 5).



Şekil 5. İleri Beslemeli Ağ İçin Blok Diyagram

Geri Beslemeli Ağlar : Bir geri beslemeli sinir ağı, çıkış ve ara katlardaki çıkışların, giriş birimlerine veya önceki ara katmanlara geri beslendiği bir ağ yapısıdır. Böylece, girişler hem ileri yönde hem de geri yönde aktarılmış olur. Şekil’de bir geri beslemeli ağ görülmektedir. Bu çeşit sinir ağlarının dinamik hafızaları vardır ve bir andaki çıkış hem o andaki hem de önceki girişleri yansıtır. Bundan dolayı, özellikle önceden tahmin uygulamaları için uygundur. Bu ağlar çeşitli tipteki zaman-serilerinin tahmininde oldukça başarı sağlamışlardır.



Şekil 6. Geri Beslemeli Ağ İçin Blok Diyagram

YSA’ların Öğrenme Algoritmalarına Göre Sınıflandırılması

Öğrenme; gözlem, eğitim ve hareketin doğal yapıda meydana getirdiği davranış değişikliği olarak tanımlanmaktadır. O halde, birtakım metot ve kurallar, gözlem ve eğitime göre ağdaki ağırlıkların değiştirilmesi sağlanmalıdır. Bunun için genel olarak üç öğrenme metodundan ve bunların uygulandığı değişik öğrenme kurallarından söz edilebilir. Bu öğrenme kuralları aşağıda açıklanmaktadır.

Danışmanlı Öğrenme : Bu tip öğrenmede, YSA’ya örnek olarak bir doğru çıkış verilir. İstenilen ve gerçek çıktı arasındaki farka göre İE’ler arası bağlantıların ağırlığını en uygun çıkışı elde etmek için sonradan düzenlenebilir.

Danışmansız Öğrenme : Girişe verilen örnekten elde edilen çıkış bilgisine göre ağ sınıflandırma kurallarını kendi kendine geliştirmektedir. Bu öğrenme algoritmalarında, istenilen çıkış değerinin bilinmesine gerek yoktur. Öğrenme süresince sadece giriş bilgileri verilir. Ağ daha sonra bağlantı ağırlıklarını aynı özellikleri gösteren desenler oluşturmak üzere ayarlar.

c) **Takviyeli öğrenme** : Bu öğrenme kuralı danışmanlı öğrenmeye yakın bir metodudur. Denetimsiz öğrenme algoritması, istenilen çıkışın bilinmesine gerek duymaz. (Aynekin.2006)

Yapay Sinir Ağlarının Kullanım Alanları

YSA'lar pek çok sektörde değişik uygulama alanları bulmuştur. Bunlardan bazıları;

Uzay: uçuş simülasyonları, otomatik pilot uygulamaları, komponentlerin hata denetimleri vs.

Otomotiv: otomatik yol izleme, rehber, garanti aktivite analizi, yol koşullarına göre sürüş analizi vs.

Bankacılık: kredi uygulamaları geliştirilmesi, müşteri analizi ve kredi müracaat değerlendirilmesi, bütçe yatırım tahminleri vs.

Savunma: silah yönlendirme, hedef seçme, radar, sensör sonar sistemleri, sinyal işleme, görüntü işleme vs.

Elektronik: kod sırası öngörüsü, çip bozulma analizi, non-lineer modelleme vs.

Eğlence: animasyonlar, özel efektler, pazarlama öngörüsü vs.

Finans: kıymet biçme, pazar performans analizi, bütçe kestirimi, hedef belirleme vs.

Sigortacılık: ürün optimizasyonu, uygulama politikası geliştirme vs.

Üretim: üretim işlem kontrolü, ürün dizaynı, makina yıpranmalarının tespiti, dayanıklılık analizi, kalite kontrolü, iş çizelgeleri hazırlanması vs.

Sağlık: göğüs kanseri erken teşhis ve tedavisi, EEG, ECG, MR, kalite artırımı, ilaç etkileri analizi, kan analizi sınıflandırma, kalp krizi erken teşhis ve tedavisi vs.

Petro kimya: arama, verim analizi vs.

Robotik: yörünge kontrol, forklift robotları, görsel sistemler, uzaktan kumandalı sistemler, optimum rota belirleme vs.

Dil: sözcük tanıma, yazı ve konuşma çevrimi, dil tercüme vs.

Telekomünikasyon: görüntü ve data karşılaştırma, filtreleme, eko ve gürültü sönmülendirilmesi, ses ve görüntü işleme, trafik yoğunluğunun kontrolü ve anahtarlama vs.

Güvenlik: parmak izi tanıma, kredi kartı hileleri saptama, retina tarama, yüz eşleştirme vs.

Kaynakça

Baykasoğlu, A., Veri Madenciliği ve Çimento Sektöründe Bir Uygulama

Koyuncuoglu, S., Veri Madenciliği Ders Notları, Güz 2010



Aynekin, G. ; İnternet İçerik Madenciliğinde Yapay Sinir Ağları ve Bir Uygulama, Uludağ Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2006

<http://matlabdersi.com/?p=10> (E.T. 20.05.2013)

<http://volkansoner.wordpress.com/tag/yapay-sinir-aglari/>(E.T. 19.05.2013)

VERİ MADENCİLİĞİ UYGULAMALARI



Gülcihan
ŞENTÜRK

VERİ MADENCİLİĞİNİN KULLANIM AMAÇLARI

Veri madenciliğinin amacı anlamlı bilgiler elde etmek ve bunu eyleme dönüştürecek kararlar için kullanmaktır. Buradaki temel amaç, değişkenler arasındaki ilişkilerden çok, geleceğe yönelik sağlıklı öngörülerin üretilmesidir. Bu anlamda Veri Madenciliği, öz bilginin keşfedilmesi anlamında bir “kara kutu” bulma yaklaşımı olarak kabul edilmektedir ve bu doğrultuda yalnızca keşifsel veri analizi tekniklerini değil, sinir ağı tekniklerinden hareketle geçerli öngörüler yapmak ve öngörülen değişkenler arasındaki ilişkilerin belirlenmesi mümkün olduğu için aynı zamanda sinir ağı tekniklerini de kullanmaktadır. Yöntemin işletmelerde kullanımı sonucunda sağlanabilecek faydalar aşağıdaki gibi özetlenebilir:

- Bir işletme kendi müşterisiyken rakibine giden müşterilerle ilgili analizler yaparak rakiplerini tercih eden müşterilerinin özelliklerini elde edebilir ve buradan hareketle gelecek dönemlerde kaybetme olasılığı olan müşterilerin kimler olabileceği yolunda tahminlerde bulunarak onları kaybetmemek, kaybettiklerini geri kazanmak için farklı stratejiler geliştirebilir.
- Mevcut müşterilerin işletme tarafından daha iyi tanınmasını sağlayabilir. Özellikle finans sektöründe mevcut müşterilerinin segmentlere ayrılarak çıkarılacak kredi risk davranış modellerinin yeni başvuruda bulunan müşterilere uygulanmasını sağlayarak riski minimize edebilir. Bir anlamda kredi risk skorlamasının altyapısının oluşturulmasında kullanılabilir.
- Mevcut müşterilerin ödeme performansları incelenerek kötü ödeme performansı gösteren müşterilerin ortak özellikleri belirlenerek, benzer özelliklere sahip tüm müşteriler için yeni risk yönetim politikaları oluşturulabilir.
- En karlı mevcut müşteriler belirlenerek, potansiyel müşteriler arasından en karlı olabilecekler belirlenebilir. Karlı Müşteriler tespit edilerek onlara özel kampanyalar uygulanabilir. En masraflı müşteriler daha masrafsız müşteri haline dönüştürülebilir. Örneğin en çok bankacılık işlemi yapanlar ortaya çıkarılıp bunlar şube bankacılığı yerine daha masrafsız internet bankacılığına yönlendirilebilir.
- Mevcut müşteriye tanıyarak işletmelerin müşteri ilişkileri yönetimlerinde düzenleme ve geliştirmeler yapılabilir. Bu sayede firmanın müşterilerini daha iyi tanıyarak müşteri gibi düşünme kapasitelerinin artırılması sağlanabilir. Bunun da işletmelere pazarda avantaj sağlayacağı unutulmamalıdır.
- Geçmiş ve mevcut yapı analiz edilerek geleceğe yönelik tahminlerde bulunulabilir. Özellikle ciro, karlılık, pazar payı gibi analizlerde Veri Madenciliği çok rahat kullanılabilir.

- Mevcut müşteriler üzerinde firma ürünlerinin çapraz satış kapasitesinin artırılması sağlanabilir. Mesela firmanın X ürününü alan müşterilerin çok büyük bir bölümünün Y ürününü de aldıkları biliniyorsa, buna yönelik pazarlama stratejileri geliştirilebilir.
- Piyasada oluşabilecek değişikliklere mevcut müşteri portföyünün vereceği tepkinin firma üzerinde oluşturabileceği etkinin tespitinde kullanılabilir.
- Operasyonel süreçte oluşabilecek olası kayıpların veya suiistimallerin tespitinde kullanılabilir.
- Kurum teknik kaynaklarının en optimal şekilde kullanılmasını sağlamakta kullanılabilir.
- Firmanın finansal yapısının, makroekonomik değişimler karşısındaki duyarlılığı ve oluşabilecek risklerin tespitinde kullanılabilir.
- Günümüzde var olan yoğun rekabet ortamında firmaların hızlı ve kendisi için en doğru kararı almalarını sağlayabilir.

VERİ MADENCİLİĞİNİN KULLANIM ALANLARI

Ülkemizde son yıllarda yeni yeni tanınmaya başlayan Veri Madenciliği kavramının, Avrupa ve Kuzey Amerika ülkelerinde birbirinden çok farklı alanlarda kullanıldığı görülmektedir. Pazarlama ve satış alanında, hedef pazarların tespitinde, müşteri ilişkilerinin yönetiminde, sepet analizinde, çapraz satışlarda, pazar segmentasyonlarında ve müşteri hatırlamada sık sık veri madenciliğinden yararlanılmaktadır. Veri kaynaklarını işlemek için müşteri kartı bilgilerinin kaydedilmesinde, müşteri şikâyetlerinin incelenmesinde, e-ticarete oldukça büyük işlevlere sahiptir. Diğer taraftan satış kampanyalarının, verimlilik analizlerinin yapılması, reklamcılık, indirim kartları ve bonuslandırmaları, karlılığın artırılması gibi daha birçok kullanım alanı bulunmaktadır. Sayılan bu kullanım alanlarının yanında, astronomi, biyoloji, finans, sigorta, tıp gibi birçok başka alanda da uygulanmaktadır. Son 20 yıldır Amerika Birleşik Devletleri'nde çeşitli Veri Madenciliği algoritmalarının gizli dinlemeden, vergi kaçakçılıklarının ortaya çıkarılmasına kadar çeşitli uygulamalarda da kullanıldığı görülmektedir. Özellikle, son yıllarda, risk analizi ve yönetiminde de, doğru ve etkin kredi kararı verebilme, kredi geri ödemesi yapmamaya meyilli müşterileri belirleme, risk derecelendirme, finansal işlemlerde sahtekârlığa yönelik eğilimleri izleme, ekonomik ve finansal yatırımları kararlaştırma, iflas / başarısızlık tahmini gibi alanlarda da yaygın olarak kullanılmaya başlanmıştır. Görüldüğü gibi Veri Madenciliği teknikleri çok çeşitli alanlarda kullanılmaktadır. Bu uygulama alanları ana başlıklar altında Tablo 1'deki gibi özetlenebilir:

Uygulama Alanları	
Pazarlama	• Müşteri segmentasyonunda, • Müşterilerin demografik özellikleri arasındaki bağlantıların kurulmasında, • Çeşitli pazarlama kampanyalarında, • Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında, • Pazar sepeti analizinde, • Çapraz satış analizleri, • Müşteri değerlendirme, • Müşteri ilişkileri yönetiminde, • Çeşitli müşteri analizlerinde, • Satış tahminlerinde,
Bankacılık	• Farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında, • Kredi kartı dolandırıcılıklarının tespitinde, • Müşteri segmentasyonunda, • Kredi taleplerinin değerlendirilmesinde, • Usulsüzlük tespiti, • Risk analizleri, • Risk yönetimi, • Kredi kartı dolandırıcılıklarının tespitinde, • Kredi taleplerinin değerlendirilmesinde, • Usulsüzlük tespiti,
Sigortacılık	• Yeni poliçe talep edecek müşterilerin tahmin edilmesinde, • Sigorta dolandırıcılıklarının tespitinde, • Riskli müşteri tipinin belirlenmesinde,
Perakendecilik	• Satış noktası veri analizleri, • Alış-veriş sepeti analizleri, • Tedarik ve mağaza yerleşim optimizasyonu,
Borsa	• Hisse senedi fiyat tahmini, • Genel piyasa analizleri, • Alım-satım stratejilerinin optimizasyonu.
Telekomünikasyon	• Kalite ve iyileştirme analizlerinde, • Hisse tespitlerinde, • Hatların yoğunluk tahminlerinde, • Kalite ve iyileştirme analizlerinde, • Hisse tespitlerinde, • İletişim desenlerinin belirlenmesi, • Kaynakların daha iyi kullanılması, • Servis kalitesinin artırılması,
Sağlık ve İlaç	• Test sonuçlarının tahmini, • Ürün geliştirme, • Tıbbi teşhis • Tedavi sürecinin belirlenmesinde • Semptomlara göre hastalık tespiti, • Magnetik rezonans verileri ile sinir sistemi bölge ilişkilerinin belirlenmesi.
Bilim ve Mühendislik	• Ampirik veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözülmesi.

Endüstri	• Kalite kontrol analizlerinde • Lojistik, • Üretim süreçlerinin optimizasyonunda
----------	---

Tablo 1. Veri Madenciliği Uygulama Alanları

VERİ MADENCİLİĞİ UYGULAMALARI

1-LİSE TÜRÜ VE LİSE MEZUNİYET BAŞARISININ KAZANILAN FAKÜLTE İLE İLİŞKİSİNİN VERİ MADENCİLİĞİ TEKNİĞİ İLE ANALİZİ

Bu çalışmada Atatürk Üniversitesinden 1976 yılından itibaren mezun olan ve halen okumakta olan öğrenci bilgilerinin bulunduğu veritabanı üzerinde Veri Madenciliği teknikleri uygulanmıştır. Verilerin tanımlanması, anlaşılması ve hazırlanması çalışmanın başında oldukça önemli bir zaman almıştır. Veri hazırlama işlemi, toplama, birleştirme temizleme ve dönüştürme aşamalarından oluşmuştur. Daha sonra modelin kurulması aşamasına geçilmiştir. Bu sırada Veri Madenciliği tekniğinin uygulanabilmesi için, bu amaca uygun Purple Insight MineSet™ 3.2 programı lisanlı olarak satın alınmıştır. Purple Insight MineSet 3.2 programı Windows tabanlıdır ve Veri Madenciliği ve görsel araçların bütünleşmesi uygundur. Ayrıca Veri Madenciliği süreci ve kapsamının hızlı ve doğru gerçekleşmesini sağlamaktadır. Programın kurulumu yapıldıktan sonra, problemin çözümü için en uygun model araştırılmaya başlanmıştır. En uygun modelin bulunabilmesi için Veri Madenciliği tekniğine uygun bütün modeller denenmiştir. Yapılan denemelerden veritabanımızda bulunan verilerin anlaşılabilir en uygun analizinin Sınıflama (Classification) modeli ile yapılabildiği gözlenmiştir. Atatürk Üniversitesi mezunlarına ait veritabanı verilerinin hazırlandıktan sonraki yapısı Tablo 2’de gösterilmiştir.

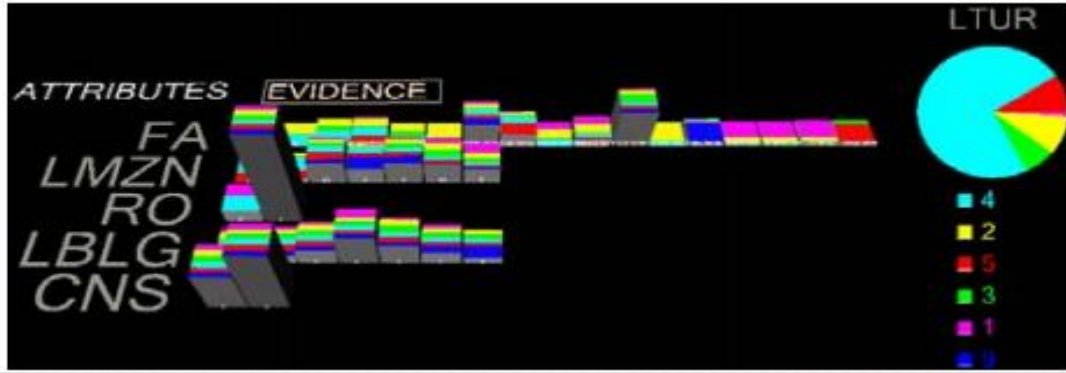
Alan Adı	Alan Kısa Adı	Kod Anlamı
Lise Türü	LTUR	1.Fen 2.Anadolu 3.Anadolu Öğretmen 4.Genel 5.Meslek
Fakülte Adı	FA	İkt. ve İd. Bil.TıpMühendislik Eczacılık vb.
Lise Mez. Bşr.	LMZN	0-2, 2-2.5,2.5-3, 3-3.5, 3.5-4, 4-4.5, 4.5-5
Rsmi-Özel	RO	1Resmi 2Özel
Lise Bölge Kodu	LBLG	1.Marmara 2. Ege 3. Akdeniz 4. İç Anadolu 5.Karadeniz 6.Doğu Anadolu 7.GüneyDoğu Anadolu
Cinsiyet	CNS	1.Erkek 2.Kız

Tablo 2. Veritabanı Alanları ve Kodları

Model uygulanmaya hazır hale gelmiş bu veriler yaklaşık 50000 kayıttan oluşmaktadır. Bu çalışmada bir öğrencinin lisede okuduğu bölümün, yani lise türünün ve lisedeki başarı seviyesinin Üniversiteye girerken kazandığı fakülte ile bir ilişkisinin var olup olmadığının anlaşılabilmesi için Veri Madenciliği tekniği uygulanmıştır. Durum MineSet 3.2 programı yardımı ile değerlendirilmiş ve elde edilen sonuçlar aşağıda verilen birkaç örnek tablo yardımı ile anlatılmıştır.

A. Lise Türünün Kazanılan Fakülte İle İlişkisi

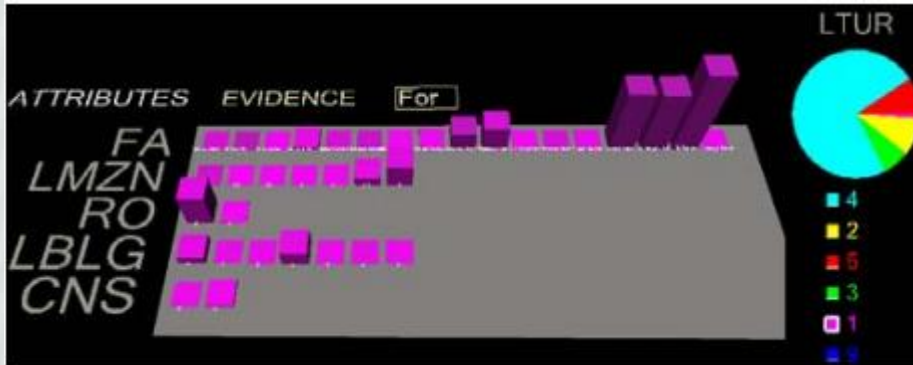
Atatürk Üniversitesini kazanan öğrencilerin genel dağılımları incelendiğinde, öğrencilerin büyük bir kısmının genel liselerden geldiği, diğer lise türlerinden gelen öğrenci sayılarının ise birbirine yakın oldukları görülmektedir. Ayrıca öğrencilerin büyük bir çoğunluğu resmi liselerden mezun olduğu, cinsiyet olarak aralarında çok önemli bir farkın olmadığı, bölgeler itibarıyla da en çok Doğu Anadolu Bölgesi olmak üzere, sırasıyla Karadeniz, Akdeniz ve İç Anadolu bölgesi ve diğer bölge liselerinden mezun öğrencilerin bulunduğu anlaşılmaktadır.



Şekil 1. Lise Türlerine Göre Kazanılan Fakültelerin Dağılımı

1. Fen Lisesi Mezunu Olmanın Kazanılan Fakülte İle İlişkisi

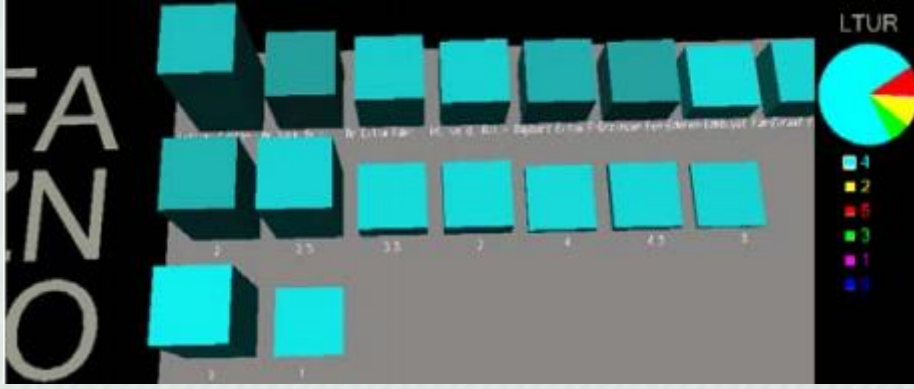
Tıp, Diş ve Eczacılık fakültelerine giren öğrencilerde Fen Lisesi'nin ağırlığının diğer fakültelere göre daha fazla olduğu görülmektedir. Öğrencilerin çoğu özel fen liselerinden mezun olup, bölgeler itibariyle de en kalabalık grubu Doğu Anadolu Bölgesi mezunları oluşturmaktadır.



Şekil 2. Fen Lisesi Mezunlarına Göre Kazanılan Fakültelerin Dağılımı

2. Genel Lise Mezunu Olmanın Kazanılan Fakülte İle İlişkisi

Genel lise mezunlarının Fen-Edebiyat Fakültesi, Eğitim Fakültesi ve İktisadi ve İdari Bilimler Fakültelerinde oranlarının daha yüksek olduğu gözlenmektedir.



Şekil 3. Genel Lise Mezunlarına Göre Kazanılan Fakültelerin Dağılımı

B. Lise Başarısının Kazanılan Fakülte İle İlişkisi

Atatürk Üniversitesine ait fakülteleri kazanan öğrencilerin lise genel başarı ortalamasının 5 üzerinden 3.5 olduğu, daha sonra 4, 4.5 ve 5 ortalamalı öğrencilerin de sırasıyla önemli yer tutukları tespit edilmiştir.



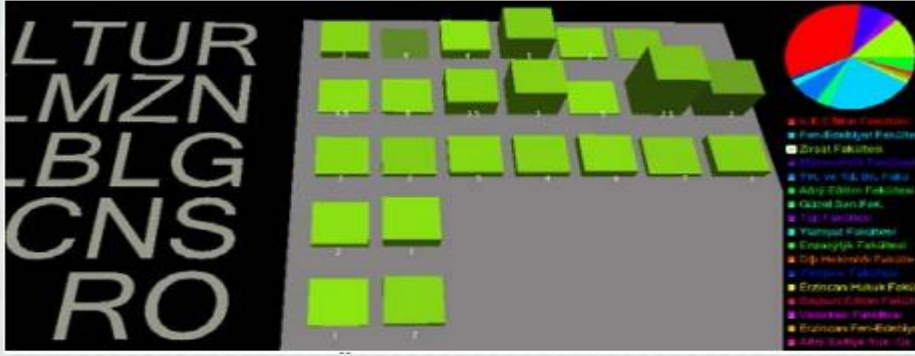
Şekil 4. Lise Başarı Notlarına Göre Kazanılan Fakültelerin Dağılımı

C. Fakültelerde, Lise Türü ve Başarı Notuna Göre Dağılımlar

Fakülte Adı baz alınarak yapılan Veri Madenciliği uygulamasında, Lise Türü ve Lise Mezuniyet Notuna Göre dağılımlarda her fakülte için farklı sonuçların elde edildiği görülmüştür. Örnek olarak Ziraat Fakültesi ile Tıp fakültesinin dağılımları gösterilmiştir.

1. Ziraat Fakültesi Öğrencilerinin Lise Türü ve Lise Başarı Durumları

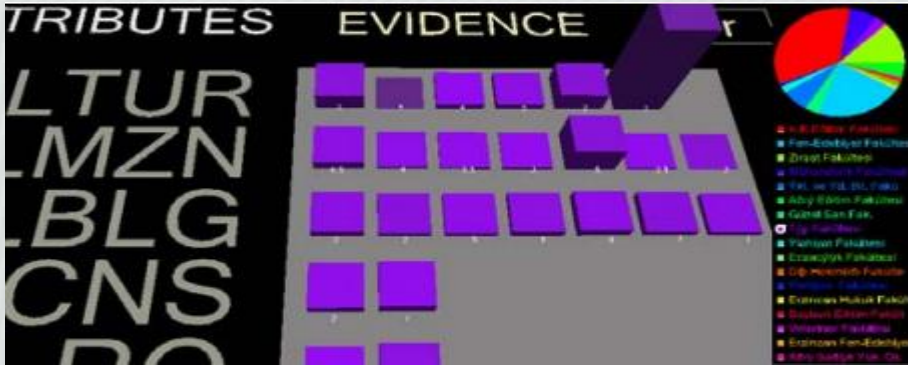
Ziraat Fakültesini kazanan öğrencilerin büyük oranda Meslek Liselerinden geldikleri ve en çok lise mezuniyet notu 2.5 olan öğrencilerden oluştukları gözlenmektedir.



Şekil 5. Ziraat Fakültesi Öğrencilerinin Lise Türü ve Lise Mezuniyet Notuna Göre Dağılımları

2. Tıp Fakültesi Öğrencilerinin Lise Türü ve Lise Başarı Durumları

Tıp Fakültesini kazanan öğrencilerin büyük oranda Fen Liselerinden geldikleri ve en çok lise mezuniyet notu 5 olan öğrencilerden oluştukları tespit edilmiştir.



Şekil 6. Tıp Fakültesi Öğrencilerinin Lise Türü ve Lise Mezuniyet Notuna Göre Dağılımları

Sonuç

Veri miktarındaki baş döndürücü artışla, insan ömrünün bu veri ile mücadelede daha verimli harcanması gerektiği düşüncesi oluşmuştur. Veriler ön elemenden geçirilmeli ve gizli bilgiler mutlaka çıkarılmalıdır. Bu amaçla kullanılan Veri Madenciliği Teknikleri ve Veritabanlarında Bilgi Keşfi süreci son yıllarda oldukça olumlu sonuçlar vermiştir. Veri elde edildikten sonra tanımlanmakta, anlaşılmakta ve faydalı bir hale getirilmesi için hazırlanmaktadır. Daha sonra Veri Madenciliği tekniği uygulanabilmesi için tekrar tekrar denemeler yapılarak veri yapısına uygun bir model belirlenmektedir (Jing Luan, 2002). Model değerlendirilerek kullanıma sunulmakta, böylece yığın haldeki verilerden daha anlamlı bilgiler daha kısa zamanda elde edilebilmektedir. Çalışmada, Atatürk Üniversitesi öğrencilerinin yıllar boyunca birikmiş



karmaşık verileri üzerinde uzun bir hazırlık çalışması yapılarak, Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte üzerindeki öneminin tespit edilebilmesi amaçlanmıştır. Çalışma sonucunda, lise türünün arzu edilen bir fakültenin kazanılmasında çok büyük öneminin olduğu, yine lise başarısının da aynı derecede önemli olduğu tespit edilmiştir.

PERAKENDE İŞLETMESİ MÜŞTERİLERİNİN KOHONEN AĞLARI İLE KÜMELENMESİ

Amaç

Bu çalışmada; sebze, meyve, taze et, süt ürünleri, konserve sebze, konserve et, konserve yemek, balık, bira, şarap, alkolsüz içki ve giyecek satışı yapan bir perakende işletmenin belirli bir dönem boyunca alışveriş yapmış müşterilerinin Kohonen tekniği ile kümelenmesi ele alınmaktadır. Kohonen tekniğinin kümeleme analizi aracı olarak seçilmesinin nedeni, büyük hacimli veriler üzerinde çalışabilme yetisi ve kümeleme analizi için en önemli karar olan küme sayısını, tekniğin kendisinin en uygun olarak belirleyebilmesidir. Kümeleme analizinin amacı; ele alınan işletmenin veritabanında yer alan müşterilerin veritabanında yer alan tüm özelliklere göre kümeler ayrılmasını sağlayarak, işletmenin pazar bölümlendirmesi ve hedef pazar seçimi gibi stratejik pazarlama kararlarını almasında işletmeye gerekli öngörüğü sağlamaktır. Ayrıca, yapılan kümeleme analizi ile işletmenin mevcut pazarının anlamlı ve etkin pazar bölümlerine ayrılabilmesi için kritik müşteri özellikleri ve önem dereceleri de ortaya konmaktadır.

Veri Seti ve Analiz Yöntemi

Bu çalışmada, yapılan analiz için Kohonen ağı tekniği kullanılmakta ve Clementine v8.1 Veri Madenciliği yazılımından yararlanılmaktadır. Clementine programı veri akışına dayanmaktadır. Bu akış içinde düğümler ile gösterilen her bir işlem, birbirine oklar ile bağlanmaktadır. Analizler, 6 aylık bir dönemde satın alınan ürünleri ile tutarlarını ve müşterilere ait bazı özellikleri içeren işletmenin 10000 adet müşteri kaydından oluşan bir müşteri veri tabanı üzerinde gerçekleştirilmektedir. Müşterilerin çoğunluğu bu dönemde birden fazla alışveriş gerçekleştirmiştir. Bununla beraber, aynı müşteriye ait her bir alışveriş kaydı birbirinden farklı olabilmektedir. Örneğin, X müşteri numarasına sahip tüketici, bir alışverişinde “konserve sebze” alırken, başka bir alışverişinde almayabilir. Dolayısıyla da, aynı müşteri numarasına sahip kayıtlardan birinde “konserve sebze”, “evet” olarak kodlanırken, diğerinde “hayır” olarak kodlanmıştır. Öz bir ifade ile her bir müşteri alışveriş veri tabanında ayrı bir kayıt olarak gözükmemektedir. Veri tabanında işletmenin satış verileri müşteriye göre düzenlenmiştir. Müşteri kayıtlarında yer alan değişkenler Tablo 3’de verilmektedir:

Müşteri Numarası	Sebze/Meyve
Aylık SatışTutarı	Taze Et
Ödeme Tipi	Süt Ürünleri
Cinsiyet	Konserve Sebze
Evsahibi	Konserve Et
Aylık Gelir	Konserve Yemek
Medeni Durum	Hane HalkıSayısı
Yaş	Bira
Alkolsüz	İçki Şarap
Balık	Giyecek

Tablo 3. Müşteri Kayıtlarındaki Değişkenler

Analizin Adımları

Analize verilerin programa yerleştirilmesi ile başlanmıştır. Ele alınan işletmeye ait veri tabanı Clementine yazılımındaki “Source” düğümü yardımı ile okutulur. Değişkenlerin tipini belirlemek için ise “Type” düğümü kullanılır. Ek 1’de, değişken tiplerine ve değişkenlerin girdi olarak seçimine bir örnek verilmektedir. Burada, Kohonen analizi için tüm değişkenler girdi olarak seçilmiştir. Böylece tüm özelliklere göre doğal kümeler elde edilmesi amaçlanmaktadır. Kümeleme yapılabilmesi için, “Kohonen” düğümü seçilerek, “Type” düğümüne eklenir. “Kohonen” düğümü ise Ek 2’de 2010 verilmektedir. “Kohonen” düğümünde yer alan “execute” düğmesi tıklandıktan sonra verilerin eğitimi süreci başlamaktadır.

Bulgular

“Kohonen” modeli sonucunda sekiz küme elde edilmiştir. Bu kümeler ve kümeleme analizi sonucunda önemli bulunan nitelikler Tablo 4’te verilmektedir. Elde edilen kümelerden beş tanesinin ayrıntılı özellikleri ise Tablo 5’te verilmektedir. Tablo 5’te yer alan “küme1”, “küme 2”, “küme 3” ve “küme 4” en büyük dört küme olması, “küme 5” ise diğer kümelere göre daha farklı bir alışveriş kalıbının bulunması nedeni ile seçilmiştir.

Sebze/Meyve: Önemsiz	Taze Et: Önemsiz
Aylık SatışTutarı: Önemsiz	Süt Ürünleri: Önemsiz
Ödeme Tipi: Önemsiz	Konserve Sebze: Önemli
Cinsiyet: Marjinal	Konserve Et: Önemli
Ev sahibi: Önemli	Konserve Yemek: Önemli

Aylık Gelir: Önemli	Hane Halkı Sayısı: Önemli
Medeni Durum: Önemli	Bira: Önemli
Yaş: Önemli	Şarap: Önemli
Alkolsüz İçki: Önemli	Giyecek: Önemli
Balık: Önemli	

Tablo 4. Değişkenlerin Önem Dereceleri

Tablo 4’te, kümeleme için kullanılan girdilerden “ev sahibi olma”, “aylık gelir”, “balık”, “konserve sebze”, “konserve et”, “konserve yemek” ve “bira” değişkenlerinin kümeleme için önemli olduğu, cinsiyet değişkeninin ise önemli olmakla beraber, diğer girdilere göre daha az önemli olduğu (programda marjinal olarak adlandırılmaktadır) görülebilir. Belirtilen değişkenler dışında kalanlar ise önemli olarak bulunmuştur. Aşağıda yer alan Tablo 5’te kümeleme analizi sonucunda bulunan en önemli nitelikler ve önem dereceleri verilmektedir.

En Önemli Nitelikler (Önem Dereceleri)	Küme 1 (X = 3, Y = 2)	Küme 2 (X = 3, Y = 0)	Küme 3 (X = 0, Y = 0)	Küme 4 (X = 0, Y = 1)	Küme 5 (X = 0, Y = 2)
Aylık Gelir (1.0)	2000-3000 TL	1000-2000 TL	3000 TL üzeri	2000-3000 TL	1000 TL’den az
Ev Sahibi Olma (1.0)	Evet (%70)	Evet (%60)	Evet (%60)	Hayır (%55)	Evet (%60)
Balık (1.0)	Evet (%80)	Evet (%70)	Evet (%60)	Evet (%60)	Evet (%60)
Konserve Sebze (1.0)	Evet (%90)	Evet (%60)	Evet (%100)	Evet (%60)	Hayır (%70)
Konserve Yemek (1.0)	Evet (%80)	Evet (%60)	Evet (%90)	Evet (%60)	Hayır (%70)
Bira (1.0)	Evet (%90)	Evet (%60)	Evet (%90)	Evet (%60)	Hayır (%70)
Konserve Et (0.99)	Evet (%70)	Evet (%70)	Evet (%80)	Evet (%80)	Evet (%70)
Cinsiyet (0.95)	Erkek (%55) Kadın (%45)	Erkek (%70) Kadın (%30)	Erkek (%60) Kadın (%40)	Erkek (%45) Kadın (%55)	Erkek (%50) Kadın (%50)
Küme Büyüklüğü	(%20)	(%18)	(%16)	(%13)	(%11)

Tablo 5. Kohonen Analizi Sonuçları

Tablo 5’teki Kohonen analizi sonuçlarına göre, söz konusu işletme için pazar bölümlendirmede kullanılabilecek en önemli değişkenlerden birinin tüketicilerin “aylık geliri” olduğu görülebilir.



Birinci küme, “orta gelirli” bir pazar bölümü olarak nitelenebilir. Bu pazar bölümündeki tüketicilerin %70’inin kendi evi vardır. “Balık”, “konserve sebze” ve “bira” tüketim oranları oldukça yüksektir. İkinci küme ise, “orta-alt gelirli” olarak nitelenebilir. %70’i erkek tüketicilerden oluşmaktadır. Üçüncü küme “yüksek gelirli” ve %60’ı kendi evi olan tüketicilerden oluşan bir pazar bölümüdür. Üçüncü kümenin %60’ı erkeklerden oluşmaktadır. Bu pazar bölümünün diğer pazar bölümlerine göre daha fazla “konserve ağırlıklı” bir beslenme biçimine sahip olduğu söylenebilir. Ayrıca bira ve balık da satın alma örüntüleri içinde yer almaktadır. Dördüncü kümeye bakıldığında “orta gelirli” ve %55’i kadın tüketicilerden oluştuğu görülmektedir. Ayrıca, üçüncü küme gibi “konserve yiyecek” satın alma örüntüsüne rastlanmıştır. Beşinci küme ise “düşük gelirli” ve genel olarak “konserve ağırlıklı” beslenme biçimi olmayan tüketicilerden oluşmaktadır. Diğer iki kümeden önemli farkı alkollü içecek satın alma örüntüsüne diğer kümelere göre çok daha az olmasıdır. Kohonen ağırları analizi ile söz edilenler gibi başka bilgiler de elde edilebilir.

Sonuç

Günümüzün ağır rekabet koşulları altında; rekabetçi avantaj elde edebilmek için, işletmelerin potansiyel tüketicilerine ya da müşterilerine ait verilerin düzenli bir şekilde, hassas, doğru ve çok boyutlu olarak analiz edilmesi gerekir. Özellikle pazar bölümlendirme ve hedef pazar seçimi gibi stratejik kararlar çok boyutlu analiz yapabilen modern karar destek araçları yardımı ile alınmalıdır. Bu araçlardan biri olan Veri Madenciliği çeşitli teknikleri sayesinde gizli bazı örüntüleri ortaya çıkararak, işletmelere rekabetçi üstünlük sağlamada çok önemli bazı bilgiler sağlayabilir. Bu çalışmada, bir perakende işletmenin belirli bir dönem boyunca alışveriş yapmış müşterilerinin Kohonen tekniği ile kümelenebilir ele alınmıştır. Yapılan kümeleme analizi ile işletmenin mevcut pazarının anlamlı ve etkin pazar bölümlerine ayrılabilmesi için önceden bilinmeyen kritik müşteri özellikleri ve önem dereceleri de ortaya çıkarılmıştır. Elde edilen bu ve benzeri sonuçlar, işletmenin pazar bölümlendirmesi ve hedef pazar seçimi gibi stratejik pazarlama kararlarını almasında yardımcı olarak kullanılabilir. Çalışmada, kümeleme analizi için Kohonen tekniğinin tercih edilmesinin en önemli iki nedeni ise, uygun küme sayısını doğal olarak hesaplayabilme özelliği ve büyük hacimli veriler üzerinde çalışabilme yetisidir. Bununla birlikte, Kohonen tekniğinin küme sayısını girdi olarak kullanan teknikler ile bileşik olarak kullanılması durumunda daha etkin kümeleme sonuçları elde edilebilir. Benzer konularda yapılması düşünülen çalışmalarda, özellikle iki aşamalı tekniklerin uygulanmasına ağırlık verilebilir.

3-BULANIK VERİ MADENCİLİĞİ UYGULAMASI

Amaç

Çalışmada, İnternet üzerinden ulaşılabilen basın-yayın kaynaklarında yer alan görsel ve metinsel verilerin hızlı ve etkin bir şekilde erişimi ve bu kaynaklardan anlamlı ve önemli bilgilerin çıkarılması hedeflenmektedir. Çalışmalar öncelikli olarak istihbarat açısından önem taşıyan kişi ve örgütlerle ilgili haberler üzerinde yoğunlaşmıştır. Sunucu bilgisayarda İnternet üzerinde yer alan haber kaynaklarından toplanmış ve işlenmiş metinsel belgelerden oluşan veri-tabanı ile bu bilgileri işleyen uygulama yazılımları bulunmaktadır. Bir arayüz ile kullanıcının bu bilgileri sorgulaması sağlanmıştır. (Erkan Ü. ve ark. 2005) Verilerin toplanması sırasında İstihbarat açısından önemli kişi, olay ve tarihlere erişim için İnternet üzerindeki haber kaynakları taranmıştır. Başlangıç noktası olarak, Işık Üniversitesi tarafından hazırlanan Terör Örgütleri ile ilgili sayfadaki bilgiler incelenmiş ve önemli isimler kaydedilmiştir. Daha sonra bu kaynaklarda geçen isimler üzerinde Google kullanılarak otomatik arama yapılmış, fotoğraf ve yazı içeren sonuçlar bulunarak işlenmek üzere veri tabanında saklanmıştır. Kaydedilen belgeler gereksiz bilgileri de içermektedir. Metin analizi yöntemleriyle veriler kısmen işlenerek ve sadece gerekli bilgiler daha sonra kullanılmak üzere veri tabanına kaydedilmiştir. (Erkan Ü. ve ark. 2005) Toplanan verilerde Kişi-kişi, kişi-organizasyon, olay-yer, olay-tarih ilişkileri Apriori algoritması kullanılarak öğrenilmektedir. Öğrenilen ilişkiler sonucu ortaya çıkan yeni isimler üzerinde arama yapılması sağlanarak, yukarıda bahsedilen işlemler tekrarlanarak, veri tabanı sürekli olarak zenginleştirilebilir. (Erkan Ü. ve ark. 2005) Kullanıcıya aranan kişiler hakkında bütün bilgileri içeren bir belge sunulmaktadır. Bu bilgiler içerisinde kişinin ismi, haberde geçen olaylar, yer ve tarihleri, birlikte bulunduğu kişilerin ismi (eğer varsa) yer almaktadır. İlişkiler de kullanıcıya sunulduğu için, kullanıcı yeni bilgiler ışığında yeni sorgulamalar yapabilecektir. (Erkan Ü. ve ark. 2005)

Basın-Yayın Alanında Birliktelik Kural Madenciliği Sistemine Bulanık Mantık Uygulaması

Basın-yayın alanında birliktelik kural madenciliği sisteminde arama yapılacak kelimeler “büyük-küçük harf duyarlı” olacak şekilde ele alınır. Çalışmada uyguladığım bulanık mantık uygulaması ile bu alana girilen isimler “büyük-küçük harf duyarlı” olacak şekilde değil, veri tabanımızdaki isimler içerisinde en uygun olan isim bulunarak bu isim üzerinden arama yapılması düşünülmüştür. Çalışmanın kullanıcıya olacak yararı, arama yapılacak isimler girilirken.

Çalışma Prensipleri

Girilen isim ile veri tabanından karşılaştırılan ismin harf harf karşılaştırılması esas alınmıştır. Girilen isimdeki harfler tek tek veri tabanında karşılaştırılan isim içinde aranır ve en yakın konumdaki harf uzaklığı üzerinden bulanık mantık algoritması uygulanır. Tüm harflerden gelen sonuçlar değerlendirilip girilen isim ile veri tabanındaki isim arasındaki benzerlik oranı 0 – 1 aralığında hesaplanır.

Algoritmanın Açıklaması

Veri tabanımızda bulunan isim alanları algoritmamız için VT anlamına gelmektedir. VT'deki isimler ilk isimden itibaren son isim gelene kadar kullanıcı tarafından girilen kelime ile karşılaştırılır. Tüm karşılaştırma sonucunda elde edilen en yüksek benzerlik oranı olan kelime yazılmak istenilen ancak yanlış yazılan kelime olarak seçilir.

Sisteme Giriş Değerleri:

Sistemin ana girişleri kullanıcı tarafından girilen isim ile veri tabanından gelen, karşılaştırma yapılacak olan isimdir. Ancak bulanık sistemimizin giriş değerleri ele aldığımız karşılaştırma yapılacak olan isimlerin harfleridir. Sisteme kullanıcı tarafından girilen ismin harfleri asıl harflerdir. Veri tabanından gelen ismin harfleri ise içinde arama yapılacak olan kısımdır. Yani sistem kullanıcının girdiği ismi temel almaktadır.

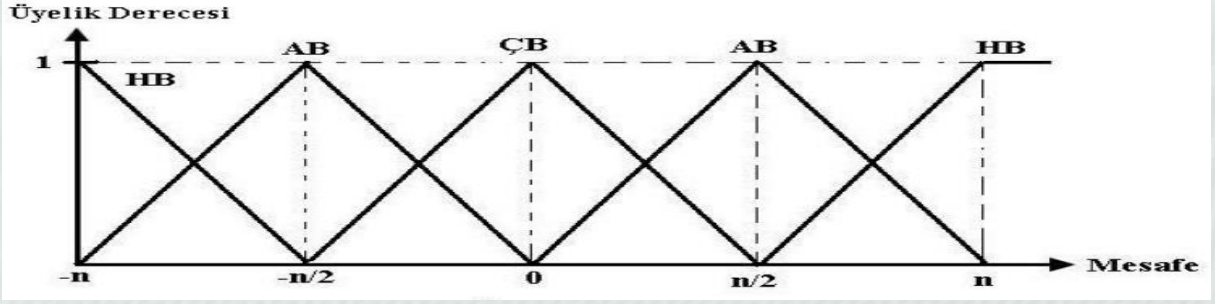
Sistemin Çıkış Değerleri:

Sistemin çıkış değeri, kullanıcının girdiği kelimenin her harfi için elde edilen bir benzerlik değeridir. Her bir benzerlik değeri birbirinden bağımsız ve eş ağırlıklara sahiptir. Birbirinden bağımsız bu değerlerin tümünün ortalaması iki kelime arasındaki benzerlik değerini vermektedir.

Bulanıklaştırma Fonksiyonu

Mesafe: Girilen isimde ki harfin, veri tabanındaki isimdeki bulunduğu konuma uzaklığı

n : Girilen isimdeki harf sayısı



Şekil 7. Üyelik Derecesi

Üyelik Fonksiyonları:

$$\text{HB fonksiyonunun denklemleri: } \begin{cases} -n \leq x \leq -\frac{n}{2}; y = -\frac{2}{n}x - 1 \\ \frac{n}{2} \leq x \leq n; y = \frac{2}{n}x - 1 \end{cases}$$

Bu denklemlerde Y uzaklığı, X üyelik derecesini ifade etmektedir.

Üyelik fonksiyonu olarak Bulanık Mantık Uygulamalarında en sık ve yaygın olarak kullanılan

$$\text{AB fonksiyonunun denklemleri: } \begin{cases} -n \leq x \leq -\frac{n}{2}; y = \frac{2}{n}x + 2 \\ -\frac{n}{2} \leq x \leq 0; y = -\frac{2}{n}x \\ 0 \leq x \leq \frac{n}{2}; y = \frac{2}{n}x \\ \frac{n}{2} \leq x \leq n; y = -\frac{2}{n}x + 2 \end{cases}$$

$$\text{ÇB fonksiyonunun denklemleri: } \begin{cases} -\frac{n}{2} \leq x \leq 0; y = \frac{2}{n}x + 1 \\ 0 \leq x \leq \frac{n}{2}; y = -\frac{2}{n}x + 1 \end{cases}$$

“Örtüşmeli Üçgen” fonksiyonu seçilmiştir. Ayrıca mesafenin negatif ve pozitif değeri için aynı sonucu almamız gerekir. Bu sebepten dolayı seçtiğimiz fonksiyonun simetrik olmalı. Örtüşmeli üçgen fonksiyonu da istediğimiz referans noktasına göre simetrik bir fonksiyondur.

Bulanık Çıkarım Motoru

Bulanık Çıkarım Motoru Fonksiyonları:

Y= üyelik derecesi, X=benzerlik oranı

Çok fonksiyonunun denklemi	: $X^2 + (Y-1)^2 = 1$
Az fonksiyonunun denklemi	: $Y - 2X = 0$
Hiç fonksiyonunun denklemi	: $(X-1)^2 + (Y-1)^2 = 1$

Bu fonksiyonların seçilmesinin sebebi, doğrusal olmayan Çok ve Hiç fonksiyonları benzerliğin az olduğu kelimeleri 0'a yakınsatmak ve benzerliğin fazla olduğu kelimeleri ise 1'e yakınsatmasıdır. Benzerliğin az olduğu Az fonksiyonunda ise yakınsama işlemi sıfır ya da bir'e doğru olabileceğinden, fonksiyonun buna direk etkisini azaltmak için lineer seçilmiştir. Yani bu fonksiyon benzerliği sıfır ya da bir'e doğru itelememektedir. Bu şekilde az benzeyen kelime ile çok benzeyen kelime arasındaki fark lineer fonksiyonlara göre daha çok açılacaktır ve yorum yapmayı kolaylaştıracaktır. Durulaştırma fonksiyonunda hedeflenen program kullanıcısının karar vermesini kolaylaştıracak sonuç vermektir. Yani iki kelimenin benzerlik oranının ortalama değerlerde olması hedeflenmektedir. Ara değerler yukarıya da aşağı itelenerek kullanıcıya daha net sonuçlar verecektir.

Bulanık Kural Tablosu:

Eğer üyelik derecesi ÇB üyelik fonksiyonundan geliyor ise; durulaştırma işlemi için Çok fonksiyonu kullanılır.

Eğer üyelik derecesi AB üyelik fonksiyonundan geliyor ise; durulaştırma işlemi için Az fonksiyonu kullanılır.

Eğer üyelik derecesi HB üyelik fonksiyonundan geliyor ise; durulaştırma işlemi için Hiç fonksiyonu kullanılır.

Kuralları sözel olarak açıklayacak olursak; Eğer harflerimiz birbirine çok yakın ise, bu durumda benzerlik fazla olacaktır ve üyelik fonksiyonlarından ÇB'ye üye olacaktır. Bu durumda durulaştırma fonksiyonu olarak Çok fonksiyonu kullanılacaktır. Eğer harflerin yakınlığı çok yakın değil ama uzak ta değil ise, bu durumda benzerlik az olacaktır ve AB üyelik fonksiyonunda bulunacaktır. Bu durumda durulaştırma fonksiyonu olarak Az kullanılacaktır. İki harfin uzaklığı çok ise, HB üyelik fonksiyonundan gelen değer dikkate alınır ve bu durumda Hiç fonksiyonu durulaştırma işlemi için kullanılır.

Durulařtırıcı

Bulanık mantık algoritmasının durulařtırıcı adımında, durulařtırma metotlarından ‘‘Ađırlıklı Ortalama Yöntemi’’ kullanılmıřtır. Bu yöntem en çok kabul gören yöntemlerden biridir. Ayrıca konumuz isimlerin benzerlik oranını bulmak olduđu için ve algoritmada isimlerin harflerinden birisinin diđerinden daha önemli olması gibi bir tutum olmadığı için yani kelimelerdeki tüm harfler eşit ađırlıklara sahip oldukları için durulařtırma işleminde ađırlık ortalama yöntemi kullanılması en mantıklı ve sonucu en sađlıklı durulařtırma işlemi olarak görölmüřtür.

(Z*= Durulařtırılmıř Deđer, Kelimenin Benzerlik Oranı, ü=her harfin benzerlik oranı, ü(n)=harf sayısı)

Algoritmanın Bir Örnekle Açıklaması

Veri tabanımızda bulunan isimlerden bit tanesi ‘‘TALABANİ’’ olsun. Kullanıcı bu ismi yazarken ‘‘ATBALANE’’ olarak yanlış yazmıř kabul edelim. Algoritmamızı bu iki kelimenin benzerliğini bulamak üzere uygular isek;

n = 8 olacaktır.

n’nin 8 deđerı için üyelik fonksiyonumuz;

‘‘ATBALANE’’ Kelimesinin her harfi için algoritmamız ayrı ayrı uygulanır.

‘‘A’’ için; KW deki ilk ‘‘A’’nın VT’deki kelimemiz içindeki en yakın ‘‘A’’ ya olan uzaklığı-1 birimdir.

u = 1 – 2 = -1;

1 2 3 4 5 6 7 8

T A L A B A N İ

A T B A L A N E

1 2 3 4 5 6 7 8

1-2 = -1

Üyelik fonksiyonlarımızdan mesafe(x)’in -1 deđerı için elde edeceđimiz üyelik dereceleri;

AB fonksiyonu için; ü(A) = 0,25

ÇB fonksiyonu için; ü(A) = 0,75

Bulanık kural tabanımızdaki;

- Eğer bulanık giriş deđerı AB fonksiyonundan geliyor ise çıkış deđerı

Az fonksiyonu... ve

- Eğer bulanık giriş deđerı ÇB fonksiyonundan geliyor ise çıkış deđerı

Çok fonksiyonu... kurallarını uygularsak;

A harfi için; AB fonksiyonundaki üyelik derecesi = 0,25



ÇB fonksiyonundaki üyelik derecesi = 0,75

Az fonksiyonunun Y (üyelik derecesi) = 0,25 için benzerlik oranı= 0,125.

Çok fonksiyonunun Y (üyelik derecesi) = 0,75 için benzerlik oranı= 0,968

Buradan $\bar{u}(A) = 0,968$ bulunur.

T harfi için; AB fonksiyonundaki üyelik derecesi = 0,25

ÇB fonksiyonundaki üyelik derecesi = 0,75

Az fonksiyonunun Y (üyelik derecesi) = 0,25 için benzerlik oranı= 0,125.

Çok fonksiyonunun Y (üyelik derecesi) = 0,75 için benzerlik oranı= 0,968

Buradan $\bar{u}(T) = 0,968$ bulunur

B harfi için; AB fonksiyonundaki üyelik derecesi = 0,5

ÇB fonksiyonundaki üyelik derecesi = 0,5

Az fonksiyonunun Y (üyelik derecesi) = 0,5 için benzerlik oranı= 0,225.

Çok fonksiyonunun Y (üyelik derecesi) = 0,5 için benzerlik oranı= 0,866

Buradan $\bar{u}(B) = 0,866$

A harfi için; AB fonksiyonundaki üyelik derecesi = 0

ÇB fonksiyonundaki üyelik derecesi = 1

Az fonksiyonunun Y (üyelik derecesi) = 0 için benzerlik oranı= 0.

Çok fonksiyonunun Y (üyelik derecesi) = 1 için benzerlik oranı = 1

Buradan $\bar{u}(A) = 1$

L harfi için; AB fonksiyonundaki üyelik derecesi = 0,5

ÇB fonksiyonundaki üyelik derecesi = 0,5

Az fonksiyonunun Y (üyelik derecesi) = 0,5 için benzerlik oranı= 0,225.

Çok fonksiyonunun Y (üyelik derecesi) = 0,5 için benzerlik oranı= 0,866

Buradan $\bar{u}(L) = 0,866$

A harfi için; AB fonksiyonundaki üyelik derecesi = 0

ÇB fonksiyonundaki üyelik derecesi = 1

Az fonksiyonunun Y (üyelik derecesi) = 0 için benzerlik oranı= 0.

Çok fonksiyonunun Y (üyelik derecesi) = 0 için benzerlik oranı = 1

Buradan $\bar{u}(A) = 1$

N harfi için; AB fonksiyonundaki üyelik derecesi = 0

ÇB fonksiyonundaki üyelik derecesi = 1

Az fonksiyonunun Y (üyelik derecesi) = 0 için benzerlik oranı= 0.

Çok fonksiyonunun Y (üyelik derecesi) = 1 için benzerlik oranı = 1

Buradan $\bar{u}(N) = 1$

E harfi için; AB fonksiyonundaki üyelik derecesi = 0

ÇB fonksiyonundaki üyelik derecesi = 0

Az fonksiyonunun Y (üyelik derecesi) = 0 için benzerlik oranı= 0.

Çok fonksiyonunun Y (üyelik derecesi) = 0 için benzerlik oranı = 0

Buradan $\bar{u}(E) = 0$

Bulunacaktır.

Durulaştırma fonksiyonumuzu kullanırsak;

$Z^* = (0,968 + 0,968 + 0,866 + 1 + 0,866 + 1 + 1 + 0) / 8 = 0,8335$ bulunur.

Sonuç olarak VT’ deki “TALABANİ” isminin, kullanıcı tarafından girilen KW=’ATBALANE” kelimesine olan benzerliği 0,8335’dir denilebilir.

Algoritmanın Özel Durumları

Kullanıcı tarafından girilen kelimenin veri tabanından karşılaştırılan kelimedenden uzun olması durumu Örneğin kullanıcı girdi olarak “ERDOĞAN” girmiş olsun, bu kelime veri tabanından gelen “ERCAN” kelimesi ile karşılaştırılıyor olsun. Girilen “ERDOĞAN” kelimesi 7 karakterlidir. Karşılaştırılan “ERCAN” ise 5 karakterlidir. “ERDOĞAN” kelimesindeki ‘N’ harfi 7. sıradadır, “ERCAN” da ise ‘N’ harfi 5. sıradadır. Ancak bu harf iki kelime de son sıradadır. Bu durumda bu harfin iki kelime üzerindeki yeri aslında aynıdır.

Bu durum oluştuğunda sistemin esas aldığı kelime kullanıcı tarafından girilen kelime olduğundan veri tabanından gelen karşılaştırma kelimesi kullanıcının girdiği kelimenin boyutuna göre sündürülür. Verdiğimiz örneğe uyarlayacak olursak, “ERCAN” kelimesi “ERDOĞAN” kelimesinin boyutuna göre sündürülür. 5 harfin konumları 7 harfle eşleştirilebilecek hale getirilir. 7 harfli “ERDOĞAN” kelimesinde birbiri ardına gelen harflerin uzaklığı 1 birim olarak hesaplanırken, 5 harfli “ERCAN” kelimesinde birbiri ardına gelen harflerin uzaklığı 1,5 birim olarak alınmaktadır. Bu hesaplama şu şekilde yapılmaktadır:

Kullanıcı tarafından girilen kelimenin uzunluğu L1 olsun.

Veri tabanından gelen karşılaştırılacak olan kelimenin uzunluğu ise L2 olsun.

Veri tabanından gelen kelimenin harfleri arasındaki uzaklığı ise X birim olsun.

Bu durumda X sıfırdan büyük bir değer çıkacaktır...

Algoritmanın Performans Analizi

Algoritma tek başına, birliktelik kural madenciliği içinde kullanılmadan performans testi yapılmıştır. Test için algoritma Turbo C++ editöründe yazılmıştır. Kullanılan bilgisayar; Microsoft Windows XP Professional with Service Pack 2 işletim istemi yüklü olan Intel



Pentium III 863Mhz işlemcili ve 512 MB belleği olan bir makinedir. Testler sırasında sistemde Turbo C++ programından başka hiçbir program çalışmamıştır.

Test kütüklerinin özellikleri:

Test için kullanılan dosyalar diske 100 Bin, 200 Bin, 400 Bin, 600 Bin ve 800 Bin kayıt içerecek şekilde kaydedilmiştir. Bu boyutların üzerindeki testlerde ise bu dosyalar birbiri ardına eklenerek kullanılmıştır. Bu kütüklerde geçen kelimeler;

- o Saddam
- o laden
- o erdogan
- o mahmut
- o talabani
- o ahmet
- o mehmet
- o mukaddes
- o nurullah
- o serdar

Yukarıdaki 10 isim birbiri ardına tekrarlanarak kütüklere yazılmıştır. Bu değişik harf sayısına sahip 10 adet ismin ortalama harf miktarı;

$(6+5+7+6+8+5+6+8+8+6)/10=6,5$ dir.

Test isimlerinin özellikleri:

Yukarıda açıklanan test kütüklerinde denenmek üzere 5 adet isim seçilmiştir. Bu isimler değişik karakter uzunlarına sahiptir. En kısa isim 3 harfli “ALİ” dir. Diğer isimler, 5 harfli “AHMET”, 7 harfli “ERDOGAN”, 9 harfli “SANLIURFA” ve 11 harfli “ARSLANGIRAY” dır.

Yukarıda belirtilen kıstaslarda algoritma, ortalama karakter uzunluğu 6,5 karakter olan 100bin, 200bin, 400bin, 800bin, 1milyon, 1,5milyon, 2milyon, 2,5milyon, 3milyon, 3,5milyon, 4milyon, 4,5milyon, 5milyon, 7,5milyon, 10milyon ve 15milyon kayıt için, 3, 5, 7, 9 ve 11 karakter uzunluğundaki isimler girilerek test edilmiştir. Test sonuçları bir tablo şekline getirilerek

Sonuçlar

Çalışmanın Bulanık Mantık Kısımının Sonuçları

Apriori algoritması kullanılarak yapılan “Birliktelik Kural Madenciliği” çalışmasının kişi-kişi birlikteliğinin bulunmasında yapılan yazım hatalarını elemek amaçlı uygulanan bulanık mantık çalışması son derece önemlidir. Bu çalışmanın sonuçlarına göre “büyük-küçük harf” duyarlı bir

karşılaştırma yapan klasik karşılaştırma algoritmasına göre “bulanık mantık” algoritması kullanılarak yapılan karşılaştırma algoritması çok önemli bir üstünlük sağlamıştır. Ayrıca “bulanık mantık” algoritması uygulanarak gerçekleştirilen algoritma, lineer olarak gerçekleştirilen benzerlik algoritmalarına da üstünlük sağlamıştır. Lineer bir algoritma ile bulanık mantık ile yapılan algoritmanın birkaç ismin karşılaştırmasına göre sonuçları aşağıdaki Tablo 6’da verilmiştir.

V.T. deki isim	Girilen İsim	Lineer Algoritma	Bulanık Mantık
SADDAM	TASDAM	0.76	0.79
MEHMET	METHEM	0.80	0.86
TALABANİ	TABALANİ	0.92	0.95
ERDOĞAN	EDROGEN	0.83	0.92

Tablo 6. Bulanık Mantık ile Lineer Algoritmanın Karşılaştırma Sonuçları

Çalışmanın Veri Madenciliği Kısımının Sonuçları

Elde edilen görsel ve metinsel verinin saklanması, işlenmesi ve Sorgulanmasını sağlayacak bir saklama mekanizmasının gereği hissedilmektedir. Bu nedenle, hem daha güvenli ve verimli olması için hem de gerekli sorguların mümkün olan en hızlı bir şekilde yapılması için bir veri tabanının tasarlanması ve oluşturulmasına karar verilmiştir. Bu amaçla MySQL Veri Tabanı Yöneticisi Sistemi kullanılarak metinsel veri tabanı oluşturulmaktadır. Bu veri tabanı sayesinde Hypertext Preprocessor (PHP) script dilini kullanarak geliştirilmiş yazılımla web üzerinden de bir arayüzle sorgulama yapılması sağlanmaktadır. İstihbarat açısından önemli kişiler hakkındaki bilgilerin toplanması için, Işık Üniversitesi Öğretim Üyesi ve Milli İstihbarat TeşkilatıEski MüsteşarıSönmez Köksal'ın hazırlamış olduğu Terör Örgütleri ve bunların ilişkilerini anlatan Veri Tabanı öncelikli kaynak olarak seçilmiştir. Bu kaynak, örgütlerin kurucuları ve önemli elemanları ile ilgili ilk bilgilerin elde edilmesi için kullanılmıştır. Bir sonraki aşama verilerin internet kullanılarak zenginleştirilmesidir. Işık Üniversitesi Terör Örgütleri Veri Tabanından elde edilen önemli kişi isimleri, Google Arama Motoru (<http://www.google.com>) aracılığı ile İnternet üzerindeki haber içerikli metinsel (haber metni, resimler altındaki açıklayıcı bilgiler, vs.) ve görsel (haber ile ilgili resimler) verilere ulaşmak için kullanılmıştır. Bu işlem Google Arama Motorunun 'grafikler' seçeneğini kullanarak kişi isimleri üzerine arama yapma ve bulunan sayfaların tümünün (resim, yazı, vs.) bilgisayarın sabit diskine Kopyalanması aşamalarını içermektedir. Bu amaçla sed, wget gibi hazır

programların yanı sıra, elde edilen web sayfalarını ayrıştıran, C Programlama diliyle yazılmış geliştirilen fonksiyon ve rutinler kullanılmıştır.

Metinsel bilgi haberin çok önemli bir parçasıdır. Bu sebeple metin üzerinde detaylı bir inceleme yapılması ve metin içinde geçen kişi adlarının, yer ve zaman öğelerinin çıkarılması ve diğer kelimelerden ayrıştırılması büyük önem taşımaktadır. Bu bilgiler şimdiye kadar kısmen ve elle ayrıştırılmıştır. Daha sonra ayrıştırılmış olan bu anahtar kelimeler item olarak kabul edilerek her bir haberin frequent itemset' leri oluşturulmuş ve Apriori algoritmasının first candidate itemsets olarak algoritmaya sunulmuştur. Apriori algoritması 4-frequent itemset'e kadar çalışmaktadır.

Tablo 7'de güven ve destek değerleri ile kişi-yer, kişi-kişi, kişi-olay sorgulamalarından bir bölüm görülmektedir. Burada gerçek kişi adları yerine semboller kullanılmıştır.

A	B	A - B	B – A	Bulunan Diğer Nesneler
Abdullah	Irak	0.432	0.102	Bülent Ecevit, Türkiye, 2002, Yorum, Terör, Tayyip Erdoğan, ...
Kemal	Türkiye	0.743	0.698	Terör, Örgüt, Şehit, 2001, 1999, 1985, 1995, Şemdin sakık, Talabani, ...
Öcalan	Barzani	0.328	0.508	Saddam Hüseyin, Usame Bin Laden, Talabani, Irak, Türkiye, Tayyip Erdoğan, . .
laden	bush	0.281	0.859	A.B.D., Bombalama, Irak, Saldırı, Arama, Elebaşı, İkiz Kule, suçlama, terör,

Tablo 7. Destek ve Güven Değeri İle kişi-yer, kişi-kişi, kişi-olay Birliktelik Kurallarından Bazıları

4- TIBBİ VERİLERLE VERİ MADENCİLİĞİ UYGULAMASI

Amaç

İnsanların deneyimlerden sonuç çıkartma yeteneği, geçmişten uygun örneklerin tanınması yeteneğine bağlıdır. Hastalıklara teşhis koyan bir doktor, öncelikle deneyimlerinden benzer vakaları tanımlar ve ardından bu vakaların bilgilerini eldeki probleme uygular. Bilinen vaka kayıtlarının tutulduğu veritabanı, sınıflandırılmış kayıtlar içinden yeni vakaya benzeyenleri bulmak için taranır. Mevcut hasta için en etkili tedavi, muhtemelen benzer hastaların sonuçlarından elde edilen bilgilerle yapılan tedavidir.

Bu tezde tıp alanında geçmiş kayıtları kullanmanın önemi dikkate alınarak gırtlak kanseri ameliyat verileri için bir analiz aracı geliştirilmiştir. Bu çalışmanın amacı, veri madenciliğinde



bir kümeleme tekniği olan k-means algoritmasını incelemek ve bu algoritmayı kullanarak geliştirilen bir yazılım aracılığı ile gırtlak kanseri ameliyat verilerinin analizini yapmaktır. Uygulamanın tıp doktorlarının kullanımına uygun şekilde verileri çeşitli açılardan analiz etmesi hedeflenmiştir. Ancak klasik hastane bilgi sistemleri daha çok idari ve finansal konulara ağırlık verecek şekilde tasarlanmaktadır. Hastaların ayrıntılı teşhis ve tedavi bilgileri ancak gereksinim duyulması halinde ve kimi zaman doktorların bireysel çabaları ile toplanıp kaydedilmektedir. Bu çalışmada kullanmak amacıyla tıbbi veri araştırılırken gırtlak kanseri ameliyat verileri bulunmuştur. Veriler çalışma için özel bir seçim değildir. Bu bölümde, gırtlak kanseri ameliyatları ile ilgili verilerin tutulduğu veritabanı kullanılarak geliştirilen yazılım uygulaması anlatılmaktadır. Yazılım, veritabanından okunan verileri 4 farklı açıdan analiz etmektedir. Dördüncü bölümde açıklanan k-means kümeleme algoritması analizde kullanılmış ve algoritmanın gösterdiği davranışlar ile performansı incelenmiştir.

Hastalık Hakkında Genel Bilgiler

Kanser, anormal hücrelerin kontrolsüz çoğalması ve yayılması olarak bilinen bir grup hastalığa verilen isimdir. Gırtlak kanseri ise, Kulak Burun Boğaz Hekimliğinde en sık görülen kanser türlerinden biridir. Gırtlak (larenks), boğazın hemen altında ses tellerinin bulunduğu bir organdır ve gıda alımı sırasında besinlerin nefes borusuna kaçmasını engeller. Gırtlak kanseri gırtlakın herhangi bir kısmında gelişebilir ve çoğu zaman ses kısıklığı ile erken bulgu verir. Genellikle 50-60 yaş grubundaki erkeklerde sık görülür. Sigara bu kanser türü için en önemli risk etkenidir. Yoğun alkol kullanımı da riski arttırır. Tedavi tümörün türü, yeri ve evreye göre belirlenir. En önemli tedavi şekli cerrahidir ve bunun yanında ışın tedavisi kullanılabilir. TNM sınıflaması, malignant(kötü huylu) tümörler için standart bir kanser evrelendirme sistemidir. Birçok tümörün kendi TNM sınıflaması bulunur. Sınıflandırmanın genel şekli şöyledir: Zorunlu parametreler T, N ve M'dir. T: Tümör büyüklüğünü temsil eder ve 0 ile 4 arası değer alır. N: Node değeri, tümörün lenflere yayılımını ifade eder ve 0 ile 3 arası değer alır. M: Metastas, 0 veya 1 değeri alır ve diğer organlara yayılımı ifade eder. Gırtlak kanseri vakalarında dikkate alınmadığından metastas parametresi veritabanında tutulmamaktadır. Tedavide patolojik ve diğer görüntü verilerine dayanarak tümör hakkında tahmini bilgi edinilir ve kanserin evresi öngörülür. Bu bilgiler ışığında yapılacak ameliyat şekline karar verilir. Hastanın yaşı da göz önünde tutulan bir etkidir. Ameliyat sırasında nadiren de olsa öngörülen bilgilerden farklı bir görüntüyle karşılaşılabilir. Tümör büyüklüğü öngörülenden farklı olabilir ve kanser evresi tahmin edilenden daha yüksek olabilir. Bu durum ameliyat sırasında uygulanan tekniğin değiştirilmesini gerektirir. Ameliyatla alınan tümör daha sonradan nüks edebilir. Hastaların



durumu daha sonraki yıllarda da izlenir ve tümör 5 yıl içinde nüks etmez ise hastanın sorundan kurtulduğu kabul edilir.

Veritabanının Çalışma İçin Hazırlanması

Bu çalışmada kullanılan gırtlak kanseri ameliyat verileri Kocaeli Üniversitesi Tıp Fakültesi Hastanesi, Kulak, Burun ve Boğaz Bölümünden alınmıştır. Veritabanı 1995 yılından bu yana toplanan 400 kayıt içermektedir. Diğer birçok veritabanında olduğu gibi tıp veritabanlarının da program aracılığı ile kullanılması için bozuk verilerden temizlenmesi ve düzenlenmesi gerekir. Hatalı girilmiş bilgilerin olması sık rastlanır bir durumdur. Çalışmada kullanılan veritabanında rastlanan hatalar şunlardır; Boş bırakılmış alanlar, birkaç bilginin birleştirilerek aynı alana yazılması, bir durumun değişik kayıtlarda birkaç farklı isimle yer alması, benzer şekilde aynı anlama gelen bilgilerin farklı formatlarda yazılması, yanlış yazılmış tıbbi terimler. Bu çalışmada kullanılan gırtlak kanseri ameliyat bilgilerinin tutulduğu veritabanı öncelikle bozuk verilerden temizlenmiş ve program aracılığı ile kullanılabilir şekilde düzenlenmiştir. Veriler Excel tablosu şeklinde alınmış ve Access veritabanına aktarılmıştır. Düzenlemeler için Access sorgu nesnesi kullanılmıştır. Veri tabanında aşağıdaki işlemler yapılmıştır: Hasta ismi ve adresi gibi analiz için gerekli olmayan alanlar çıkartılmıştır. Boş (null) değer içeren sayısal alanlara 0 değeri atanmıştır. K-means algoritmasında karakter alanlar kullanılmadığından, patoloji, survive ve operasyon alanları sayısal karşılıklarına çevrilmiştir. Patoloji sonucu yassı epitel hücre (tabloda “Yeh Ca” olarak kısaltılmıştır) ise bu alanın değeri 1, diğer sonuçlar için 0 yapılarak veritabanına yeni eklenen patoloji_kodu alanına yerleştirilmiştir. Nüks bilgisi, ölüm nedeni ve tarihini içeren survive alanından nüks ve hayatta bilgileri alınarak veritabanına yeni eklenen aynı isimli alanlara yerleştirilmişlerdir. Nüks varsa nüks isimli alanın değeri 1, yoksa 0 yapılmıştır. Benzer şekilde, hasta hayatta ise hayatta isimli alanın değeri 1, aksi durumda 0 yapılmıştır.

Çalışmada kullanılmak amacıyla düzenlenen veritabanı Tablo 8’de yer almaktadır.

Yas	Patoloji	preop_T	preop_N	preop_evre	postop_T	postop_N	postop_evre	Operasyon	Operasyon_tarihi	nüks	hayatta	patoloji_kodu
45	Epidermoic	T4	N1	4				TL	05.10.2000	0	1	0
40	YEH Ca	T2	N0	2				SCL	02.11.2000	0	1	1
51	YEH Ca	T3	N0	3	T4	N2	4	TL	21.11.2000	1	0	1
70	YEH Ca	T4	N2	4				TL	28.11.2000	0	0	1
43	YEH Ca	T1	N1	3	T2	N2	4	SGL	08.12.2000	0	1	1
39	YEH Ca	T4	N2	4				TL	04.01.2001	1	0	1
60	YEH Ca	T2	N0	2				TL	11.01.2001	0	0	1
53	YEH Ca	T2	N3	4				SGL	01.01.2016	1	0	1
50	YEH Ca	T2	N2	4				SGL	01.02.2001	0	1	1
52	Insitu Ca	T1	N0	1				FLL	08.02.2001	0	1	0

Tablo 8. Çalışmada Kullanılan Gırtlak Kanseri Ameliyat Bilgileri Veritabanı

Veritabanındaki Alanların Açıklamaları:

Yaş	:Hastanın yaşı.
Patoloji	:Ameliyat öncesinde tespit edilen patolojik tetkik sonucu. Yeh yassı epitel hücre anlamına gelmektedir.
Preop_T	:Teşhis sırasında tespit edilen tümör büyüklüğü.
Preop_N	: Teşhis sırasında tespit edilen node (lenf) büyüklüğü (Bkz. 5.2).
Preop_evre	: Preop_T ve Preop_N değerlerinden hesaplanan kanser evresi.
Postop_T	:Ameliyat sırasında görülen tümör büyüklüğü.
Postop_N	: Ameliyat sırasında görülen node (lenf) büyüklüğü (Bkz. 5.2).
Postop_evre	:Postop_T ve Postop_N değerlerinden hesaplanan kanser evresi.
Operasyon	:Yapılan ameliyatın ismi. Veritabanında yedi çeşit ameliyat bulunmaktadır.
Operasyon_tarihi	: Ameliyatın yapıldığı tarih.
Nüks	:Ameliyat sonrası tümör nüks ederse bu alanın değeri 1 olur. Aksi durumda 0'dır.
Hayatta	:Ameliyat sonrası hasta hayatta ise bu alanın değeri 1 olur. Aksi durumda 0'dır.
Patoloji_kodu	:Ameliyat öncesinde tespit edilen patolojik tetkik sonucu "Yeh Ca" ise bu alanın değeri 1 olur. Diğer bütün sonuçlar için 0'dır.

K-means Algoritmasının Tercih Nedenleri

Bu çalışmada kullanılan k-means algoritması, aşağıdaki özellikleri nedeniyle tercih edilmiştir:

1. Küme sayısı olan k değeri parametre olarak algoritmaya dışardan verilmektedir. Verilerin kaç kümeye ayrılacağını net olarak bilinmediği durumlarda farklı değerler

vererek sonuçları izlemek mümkün olmaktadır. Bu olanak uygulamanın analizini esnek hale getirmektedir.

2. Algoritmanın uygulanması kolaydır ve hızlı çalışmaktadır. Veriler matristen okunarak algoritmaya verilmektedir. İterasyon sayısı, veri sayına oranla oldukça azdır ve bu nedenle algoritma hızlı çalışır.
3. Değişik dağılımlarda başarılı sonuçlar alınabilmektedir. Veriler birbirine çok yakın veya çok uzak değilse, kümeleme işlemleri başarıyla gerçekleştirilmektedir.
4. Kategorik verilerle çalışacak şekilde adapte edilebilmektedir. Algoritma rakamsal verilerle çalışmaktadır. Ancak kategorik veriler uygun rakamsal karşılıklarına çevrilerek algoritma tarafından işlenebilir.
5. Kümeleme sonuçları hem grafik olarak hem de yazı ve rakamlarla kolayca ifade edilebilmektedir. Algoritma sonucunda kümelenen veriler matrise yerleştirilmektedir. Buradan grafiğe kolayca aktarılabilen ve her küme farklı renkle gösterilebilmektedir. Kümelerin eleman sayıları ve diğer verilere olan oranları kolayca hesaplanıp, yazılı olarak ifade edilebilmektedir. Algoritmanın bazı dezavantajları da bulunmaktadır. Çok boyutlu verilerde kötü sonuçlar verebilir ve kümeleme yüksek iterasyon gerektirebilir. Ancak gırtlak kanseri ameliyatlarının verileriyle çalışılırken bu dezavantajlar ortaya çıkmamıştır. Çünkü veriler çok boyutlu değildir ve noktalar düzgün dağıldığından kümeleme yüksek iterasyon gerektirmemiştir.

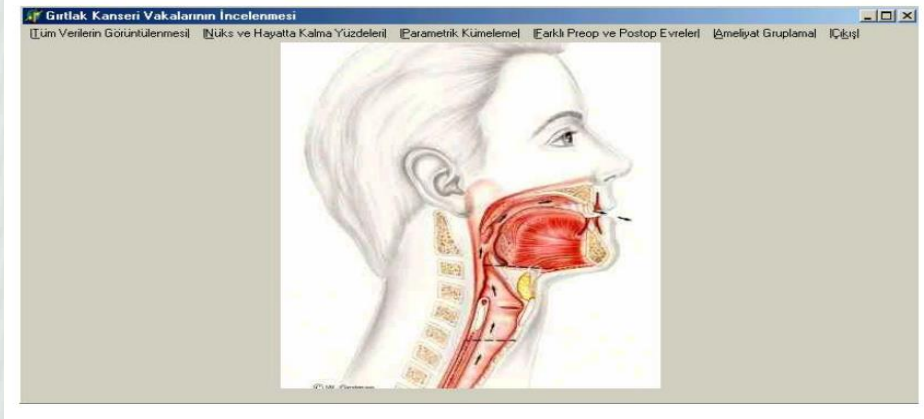
K-means algoritmasının Delphi ile yazılan kodu Ek-a'da yer almaktadır.

Geliştirilen Uygulama ile Verilerin Analizi

Bu çalışmada geliştirilen uygulama, Borland Delphi 6,0 kullanılarak Windows XP işletim sistemi üzerinde geliştirilmiştir. Veriler Access veritabanında tutulmuştur. Kmeans algoritması da Delphi ile kodlanmıştır. Uygulama aracılığı ile veriler dört farklı açıdan analiz edilmiştir. Her bir analiz ayrı bir arayüz ile görüntülenmiştir. Arayüz ekranlarına Şekil 8'de görülen bir anamenü üzerinden erişilmektedir. Oluşturulan arayüzler şöyledir:

1. Nüks ve Hayatta Kalma Yüzdeleri: Gelecek vakaların tahminlerinde kullanılmak amacıyla, geçmiş vakaların tümör nüks etme yüzdeleri ve hayatta kalma yüzdeleri incelenir. Sadece bu arayüzde k-means algoritması kullanılmamıştır.
2. Parametrik Kümeleme: Veri tablosunun seçilen iki alanı arasındaki ilişim izlenir.
3. Farklı Preop ve Postop Evreler: Doğru öngörülen ve öngörülemeyen ameliyat öncesi evreler incelenerek ameliyat öncesi tahmin başarısı değerlendirilir.

4. Ameliyat Grublama: Başarılı ameliyatlara izlenerek, gelecek ameliyat tercihlerine karar verilir.



Şekil 8. Geliştirilen Programın Anamenüsü

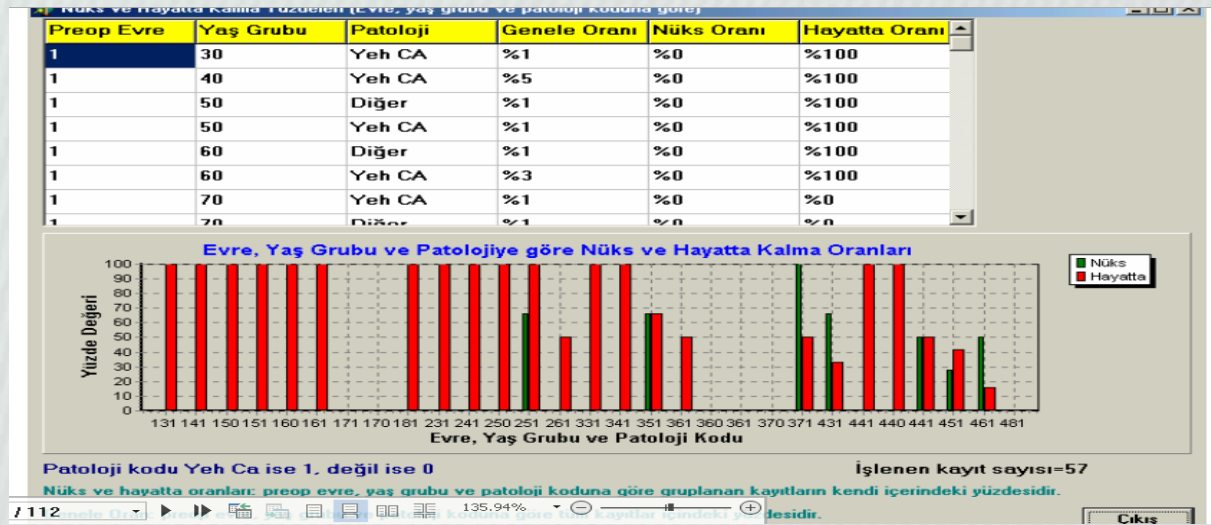
Analiz yapılan arayüzlere ilave olarak, bütün veritabanının filtre edilmeden görüntülenmesi için Tüm Verilerin Görüntülenmesi seçeneği altında bir liste ekranı hazırlanmıştır.

Uygulama Arayüzleri

Analiz yapılan arayüzler benzer görüntülerde tasarlanmıştır. Ortak özellikler şöyledir; arayüz ekranının üst kısmında analizi yapılan veriler listelenmektedir. Parametre girişi yapılmayan arayüzlerde, seçenek çalıştırıldığında, veriler hesaplanarak görüntülenmektedir. Kümelenmiş veriler ve içlerindeki yoğunlaşmalar orta kısımda grafik üzerinde gösterilmiştir. Arayüzlerde k-means algoritması ile oluşturulacak küme sayısı, kullanıcı tarafından 2 ile 9 arasında bir değer ile belirlenebilmektedir. Sınama sonuçlarına göre, 9 kümeden daha fazlası verimli olmamaktadır. Kullanıcı “K-means çalıştır” tuşunu tıklayarak kümeleme sonucunu grafik üzerinde görüntüleyebilmektedir. Her bir küme ayrı bir renk ile gösterilerek birbirinden ayrıştırılır. Bu şekilde kümelenmiş veriler, kolay analiz olanağı sağlamaktadır. Kümelenmiş sonuçların kısa açıklaması arayüz ekranlarının altındaki bir pencereden görüntülenir. Açıklama, her bir kümedeki kayıt sayısını ve bu sayının işlenen tüm kayıtlara oranını içerir. Her bir açıklama satırı, grafikteki kümesi ile aynı renktedir. Bu kısımda ayrıca kümeleme sonucunun yazılı ifadeleri ile arayüze ilişkin mesajlar ve açıklamalar bulunmaktadır. Her arayüzün çalıştırılması sırasında hesaplanan bilgiler ayrıca bir metin bir dosyasına yazdırılmaktadır. Bu dosyada ekranda görüntülenen rakamsal bilgiler ve bunlarla ilişkili diğer ayrıntı bilgileri yer almaktadır. Dosyalar grafik içermemektedir. Uygulamanın arayüzleri aşağıda ayrıntılarıyla açıklanmıştır.

Nüks ve Hayatta Kalma Yüzdeleri

Bu arayüzde diğer üç arayüz ekranından farklı olarak veri kümeleme yerine sınıflama yapılmıştır. Analizi yapılacak kayıtların algoritmaya uygulanamaması nedeniyle, burada k-means algoritması kullanılmamıştır. Bu arayüzün amacı, mevcut ve gelecek vakalar için, ameliyat sonrasında tümörün nüks etme olasılığının ve hastanın hayatta kalma olasılığının tahmin edilmesine destek sağlamaktır. Bunun için geçmiş vaka verileri sınıflandırılarak, tümör nüks etme yüzdeleri ve hayatta kalma yüzdeleri görüntülenir. Veriler, hasta yaş grubu, ameliyat öncesi evresi ve patolojik bilgiye göre sınıflanmıştır. Yüzdeler her sınıf için ayrı ayrı hesaplanmıştır.



Şekil 9. Nüks ve Hayatta Kalma Yüzdeleri Arayüzü

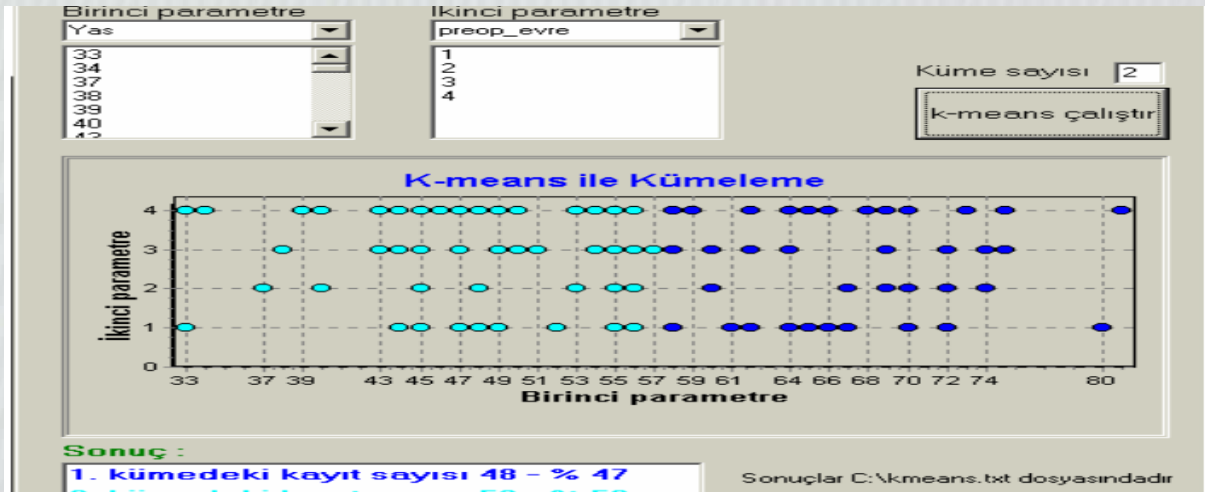
Arayüz görüntülediğinde hesaplanan sonuçlar listelenir. Her bir sınıftaki kayıt sayısının bütün kayıtlara oranı hesaplanarak “Genele Oranı” adı altında ayrı bir sütunda listelenmiştir. Grafikte çubuk (bar) seri tipi kullanılmıştır. Şekil 9’da görüldüğü gibi her bir çubuk bir sınıfı temsil etmektedir. Grafikte görüntülenene yeşil renkli seri, her bir sınıfın tümör nüks yüzdesini, kırmızı renkli seri ise hayatta kalma yüzdesini göstermektedir. Sınıflar hasta yaş grubu, ameliyat öncesi evresi ve patolojik bilgilerin birleştirilmesinden oluşturulmuştur. Yaş grubunda onlar basamağı aynı olan yaşlar gruplanmıştır. Örneğin 50-59 arası 50 grubu, 60-69 arası 60 grubu gibi. Birleştirilen alanların bilgileri sayısal karşılıklarına dönüştürülerek, her bir sınıf üç basamaklı bir sayı ile temsil edilmiştir. Sayının yüzler basamağını evre kodu, onlar basamağını yaş grubu bilgisinin onlar basamağı, birler basamağını değeri 0 veya 1 olan patoloji kodu oluşturmaktadır. Program kodunda sayının hesaplanmasında izlenen yol şöyledir: (evre kodu)*100+(yaş grubunun onlar basamağı)*10+ (patoloji kodu). Örneğin evre kodu 1, yaş

grubu 70 ve patolojik tetkik sonucu “Yeh Ca” ise bu sınıfın kodu 171’dir. Evre kodu 2, yaş grubu 60 ve patolojik tetkik sonucu “Diğer” ise bu sınıfın kodu 260’dır.

Örneğin Şekil 9’da görülen ilk satırda, birinci evrede olan ve patolojik tetkik sonucu “YehCa” olan 30 yaş grubundaki vakalar yer almaktadır. Bu gruptaki vakalar sayıca bütün kayıtların yüzde birini oluşturmaktadır. Vakaların hiç birinde tümör nüksü kaydedilmemiş ve vakaların tamamı hayattadır. Bu grup 131 kodu ile grafikte yer almıştır. Nüks oranı 0 olduğundan yeşil seri görüntülenmemektedir. Hayatta oranı 100, kırmızı seri ile gösterilmiştir. Yedinci satırda, birinci evrede olan ve patolojik tetkik sonucu “YehCa” olan 70 yaş grubundaki vakaların nüks ve hayatta bilgileri veritabanında yer almadığı için oranları 0’dır.

Parametrik kümeleme

Bu arayüzde geriye dönük inceleme kolaylığı ve ileriye dönük karar verme desteği sağlanması amaçlanmıştır. Doktor değişken parametreler kullanarak değerlendirme yapabilir, geçmiş verileri analiz edebilir, mevcut ve gelecek vakalar için tahminde bulunabilir. Veritabanından iki alan seçilerek, bunların değerleri arasındaki etkileşim izlenmektedir. Örneğin yaş ile tümör boyutu veya yaş ile uygulanmış ameliyatlar gibi. Seçilen veriler k-means algoritması ile kümelenecek, veri içindeki yoğunlaşmalar kullanıcıya görüntülenebilmektedir. Geliştirilen uygulama içinde, kullanıcının alanlar seçerek işlem yapabildiği tek arayüzdür. Eğer kullanıcı



birbiriyle ilişkisiz alanlar seçerse, analiz sonucu anlamsız olabilir.

Şekil 10. Parametrik Kümeleme Arayüzü

Arayüz görüntülendiğinde Şekil 10’da görüldüğü gibi, veritabanındaki bütün alanların isimleri üstte yer alan iki açılan kutu (combo box) içinde listelenir. Kullanıcı, alan isimlerinin üstünü tıklayarak herhangi iki tanesini seçer. Seçim sonrası alanların değerleri, her bir değerden bir



tane olacak şekilde liste kutusunda (list box) listelenir. Seçilen alanların grafik üzerinde kümelenmiş görüntüsünün oluşturulması için, kullanıcı küme sayısını girerek “K-means çalıştır” tuşunu tıklar. Grafik üzerindeki her bir seri seçilen bir alanı temsil eder.

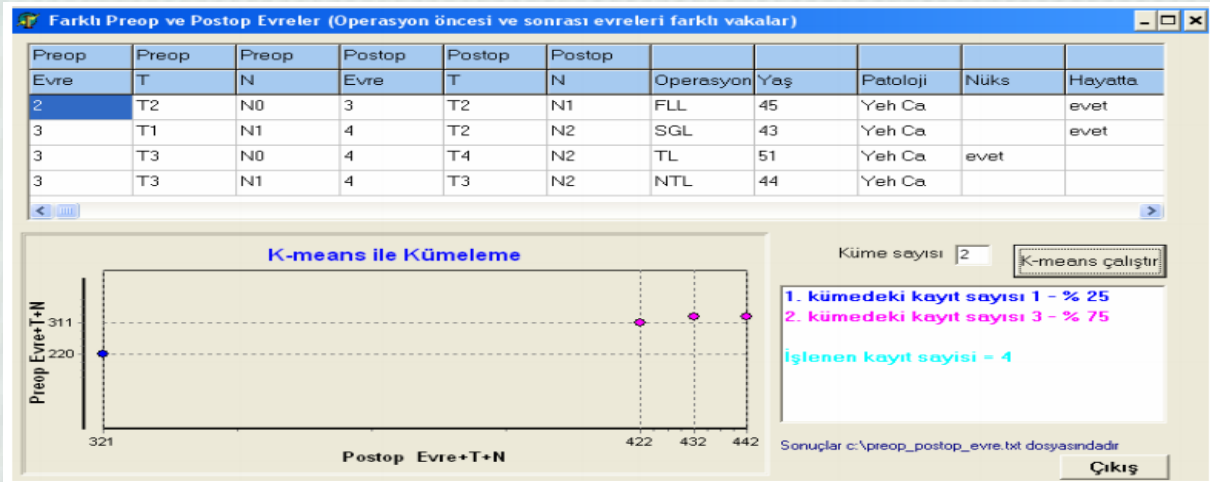
Şekil 10'daki örnek görüntüde, yaş ve ameliyat öncesi evre alanları seçilmiş, ilgili veriler k-means algoritması ile iki kümeye ayrılmıştır. Bu seçimle, ameliyat öncesi öngörülen evrelerin yaşlarla ilişkisi incelenmektedir. Kayıt sayıları birbirine yakın iki kümeden birincisi 33 ile 57 yaş arası, ikincisi ise 57 ile 83 yaş arasındaki vakaları içermektedir. Birinci kümede, 43 yaş sonrası dördüncü evredeki vakaların çokluğu dikkat çekmektedir. Dördüncü evreyi, yoğunluk açısından üçüncü ve birinci evreler izlemektedir. İkinci kümede ise 57 yaş ile 70 yaş arası dördüncü evrede, 61 ve 67 yaş arasında birinci evrede yoğunluk izlenmektedir. Program kodunda, ameliyat ismi ve patoloji kodu gibi bütün karakter alanlar kmeans algoritmasıyla işlenebilmek için sayısal karşılıklarına dönüştürülmüştür.

Farklı preop ve postop Evreler

Bu arayüzde doğru öngörülen ve öngörülemeyen ameliyat öncesi evrelerin görüntülenerek incelenmesi ve bu şekilde ameliyat öncesi tahmin başarısının değerlendirilmesi amaçlanmıştır. Kanseri ameliyatlarında, nadiren de olsa ameliyat öncesinde tahmin edilen evre, ameliyat sırasında görülen evreden farklı olmaktadır. Bu da ameliyat öncesi evreye göre yapılan hazırlığın, ameliyat sırasında değiştirilmesi ve yeni kararlar alınması anlamına gelmektedir. Gerçek hayatta az rastlanır bir durum olduğundan, veritabanında bu inceleme için listelenen kayıt sayısı azdır. Şekil 10'daki örnekte sadece 4 kayıt görüntülenmektedir. Burada evre değerleri, ameliyat öncesinde ve sonrasında birbirinden farklı kayıtlar listelenmektedir. Koşulu sağlayan her bir vaka için listelenen bilgiler şunlardır:

Ameliyat öncesi ve sonrasına ait evre değeri, tümör ve node kodları, uygulanan ameliyatın kısa adı, hastanın yaşı, patoloji kodu, nüks ve hayatta bilgileri.

Grafikte ameliyat öncesi ve sonrasına ait evre, tümör ve node değerleri karşılaştırılmaktadır.



Şekil 11. Farklı preop ve postop Evreler Arayüzü

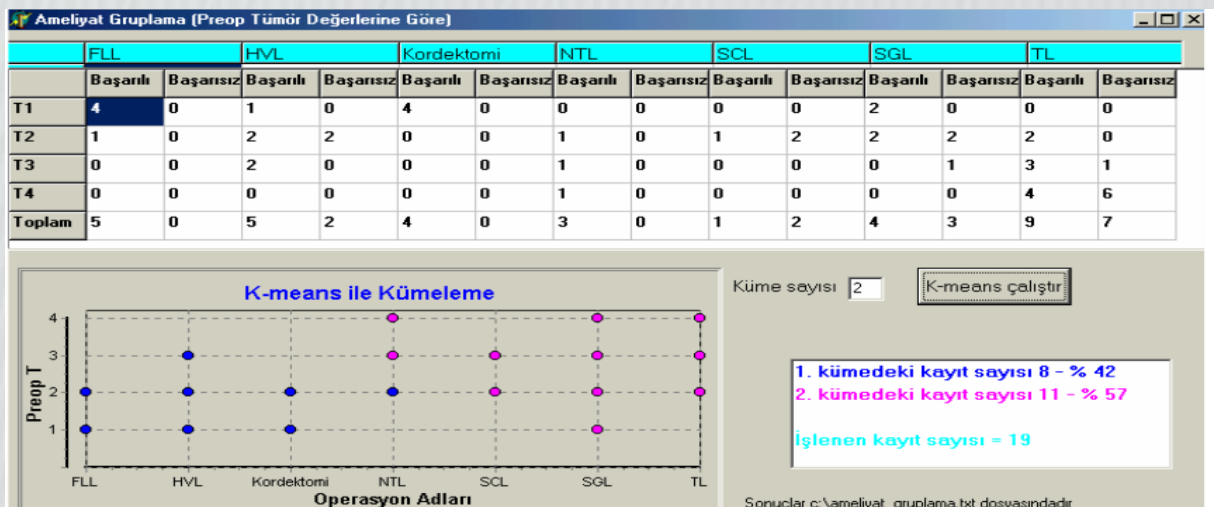
Program kodunda, evre, tümör ve node değerlerinin k-means algoritmasında kullanılabilmesi için alanlar birleştirilerek tek bir sayı haline getirilmiştir. Birleştirmede izlenen yol şöyledir: $(\text{evre kodu}) \times 100 + (\text{tümör değeri}) \times 10 + (\text{node değeri})$. Ameliyat öncesi ve sonrası için bu şekilde oluşturulan sayılar k-means algoritmasının her bir boyutunu oluşturur. Şekil 11’de görülen örnekte, ilk kaydın preop evresi 2, tümör değeri 2 ve node değeri 0’dır. Bu değerler grafikte 220 sayısına karşılık gelmektedir. Aynı şekilde postop evresi 3, tümör değeri 2 ve node değeri 1’dir. Bu da grafikte 321 sayısına karşılık gelmektedir. Şekil 11’de listelenen 4 kayıt iki kümeye ayrılmıştır. Postop evresi 4 olan kayıtlarla preop evresi 3 olan kayıtlar üç elemanlı büyük kümeyi oluşturmuşlardır. Grafikten görüldüğü gibi, ameliyat öncesinde üçüncü evrede olduğu öngörülen vakalar, ameliyat sırasında dördüncü evrede bulunmuştur. Öngörü farkı, en çok üçüncü evrede, patolojisi “yehca” olan 40-50 yaş arası vakalarda ortaya çıkmıştır.

Ameliyat Gruplama

Bu arayüzde başarılı ameliyatların izlenerek, gelecek ameliyat tercihlerine destek sağlanması amaçlanmıştır. Burada gırtlak kanseri ameliyatları, vakalara göre kümelenmiştir. Vakalar ameliyat öncesi tümör değerleri ile temsil edilmektedir. Kümeleme sırasında aynı tümör değerine sahip vakaların sayıları belirlenmiştir. Aynı tümör değerine sahip her vaka için başarılı veya başarısız ameliyatların sayıları hesaplanmıştır. Hesaplama yöntemi şöyledir; eğer ameliyat sonrasında tümör nüks etmiş ise ameliyat başarısız kabul edilir, eğer tümör nüks etmemiş ve hasta hayatta ise ameliyat başarılı kabul edilir. Böylece arayüzde, gerçekleştirilen ameliyatlar görüntülenerek, hangi tümör değerlerine hangi ameliyatların ne çoklukta yapıldığı ve başarılı/başarısız sayıları izlenir. Başarılı ve başarısız ameliyatların sayıları grafiğe yansıtılmamıştır.

Arayüz vasıtasıyla kümelenmiş ameliyatlar listelenir. Birçok ameliyatın ismi uzun olduğu için isimler yerine kısaltmalar kullanılmıştır. Şekil 12’de görülen tabloda ameliyat öncesi tümör değerlerine karşılık olarak, başarılı ve başarısızlık hesaplaması her bir ameliyat için ayrı ayrı listelenmektedir. Şekildeki örnekte FLL kodlu ameliyat, T1 için 4 kez, T2 için 1 kez başarıyla yapılmıştır. Bu ameliyatta hiç başarısız vaka kaydedilmemiştir. Son satırda bütün ameliyatların başarılı ve başarısız sayıları toplanmıştır. Örnekte, FLL kodlu ameliyatın başarı sayısı 5, başarısız sayısı 0’dır. Bu tabloda, veritabanında nüks ve hayatta alanları dolu olan vakalar hesaplamaya dahil edilmiştir. Her kayıt için bu alanlar dolu olmadığından, burada listelenen kayıt sayısı, veritabanındaki toplam kayıt sayısından daha azdır.

Şekil 12’deki grafikte ameliyatlar ile ameliyat öncesi tümör değerleri karşılaştırılmaktadır. Burada tümör tiplerine uygulanan ameliyat çeşitleri kümelenmiştir. Kümeleme için veritabanından okunan vakaların nüks ve hayatta alanları dikkate alınmamıştır. Grafikte gösterilmediğinden tümör tiplerine uygulanan ameliyat sayıları hesaplanmamıştır. Bu nedenle tabloda listelenen ile grafikte ele alınan kayıt sayıları farklıdır. Tabloda 45 kayıt için listeleme yapılırken, grafikte 19 kayıt kümelenmiştir. Şekildeki örnekte, ameliyatlar iki kümeye ayrılmıştır. Sayıca az olan ilk kümede, daha çok 1 ve 2. derece tümörlere uygulanan üç ameliyat yer almıştır. Büyük kümede ise, 2. derecenin üstündeki tümörlere uygulanan ameliyatlar bulunmaktadır. 2. derece tümörlere bütün ameliyat çeşitlerinin uygulandığı, SGL kodlu ameliyatın bütün tümör türlerine uygulanabildiği görülmektedir. Diğer taraftan 2., 3. ve 4. derecelere uygulanan NTL ve TL kodlu ameliyatlar, başarı rakamlarına göre incelendiğinde, TL’nin daha riskli bir teknik olduğu göze çarpmaktadır. Buna rağmen TL, sayıca NTL’den fazla gerçekleştirilmiştir.



Şekil 12. Ameliyat Gruplama Arayüzü

Program kodunda ameliyatlar k-means algoritmasıyla işlenebilmek için sayısal karşılıklarına dönüştürülmüştür. Bunun için ameliyat isimlerine alfabetik sırada birden yediye kadar sıra numarası verilmiştir.

Tüm Verilerin Görüntülenmesi

Bu arayüz gırtlak kanseri ameliyatlarına ilişkin tüm verilerin görüntülenmesi amacıyla hazırlanmıştır. Verilerin yer aldığı Access veritabanındaki bütün alanlar listelenmiştir. Şekil 13’de görüldüğü gibi veriler üzerinde herhangi bir işlem yapılmamıştır.

Yas	Patoloji	preop_T	preop_N	preop_evre	postop_T	postop_N	postop_evre	Operasyon	Operasyon_tarihi	nuks	hayatta	patoloji_kodu
65	YEH Ca	T1	N0	1				FLL	03.01.1996	0	0	1
56	YEH Ca	T4	N2	4				TL	13.12.1995	1	0	1
67	YEH Ca	T2	N0	2				SGL	28.03.1996	0	0	1
45	YEH Ca	T2	N0	2	T2	N1	3	FLL	14.02.1996	0	1	1
58	YEH Ca	T4	N2	4				NTL	20.05.1997	0	1	1
49	YEH Ca	T4	N2	4				TL	03.06.1997	0	1	1
45	YEH Ca	T3	N0	3				TL	03.03.1997	0	1	1
33	epidermoid Ca	T4	N2	4				TL	02.09.1997	0	0	0
48	YEH Ca	T1	N0	1				SGL	17.09.1997	0	1	1
67	YEH Ca	T1	N0	1				Kordektomi	03.11.1997	0	0	1
74	YEH Ca	T2	N0	2				HVL	19.01.1998	0	0	1
62	Ağır displazi	T1	N0	1				Kordektomi	11.03.1997	0	1	0
44	YEH Ca	T3	N1	3	T3	N2	4	NTL	23.02.1998	0	0	1
60	YEH Ca	T3	N0	3				TL	18.12.1997	0	1	1
61	Ağır displazi	T1	N0	1				Kordektomi	10.05.1996	0	0	0

Şekil 13. Tüm Verilerin Görüntüldüğü Arayüz

Sonuçlar ve Öneriler

Bu bölümde, tez çalışması hakkında bilgi verilerek ve gerçekleştirilen uygulama ile elde edilen sonuçlar özetlenmiştir. Tıp alanında veri madenciliği uygulamalarının ilerlemesinin önündeki engeller anlatılmış ve bölümün sonunda çalışmanın geliştirilebilmesi için ileride yapılabilecek öneriler sıralanmıştır. Günümüzde başta iş dünyası olmak üzere birçok farklı alanda kullanılan Veri Madenciliği MIT, (Massachusetts Institute of Technology) tarafından 2001 yılında yayınlanan bildirgeye göre dünyayı değiştirecek 10 teknoloji arasında gösterilmiştir. Gelecekte daha çok önem kazanacak olan veri madenciliği üzerinde yapılan çalışmalara her geçen gün yenileri eklenmektedir. Veri madenciliğinde yeni gelişen teknolojilerin birçoğu henüz tıp alanında kullanılan yazılımlara dahil edilmemiştir. Hastalıklara teşhis koyan bir doktor, öncelikle deneyimlerden benzer vakaları tanımlar ve ardından bu vakaların bilgilerini eldeki probleme uygular. Bilinen vaka kayıtlarının tutulduğu veritabanı, sınıflandırılmış kayıtlar içinden yeni vakaya benzeyenleri bulmak için taranır. Mevcut hasta için en etkili tedavi, muhtemelen benzer hastaların sonuçlarından elde edilen bilgilerle yapılan tedavidir. Bu tezde



tıp alanında geçmiş kayıtları kullanmanın önemi dikkate alınarak gırtlak kanseri ameliyat verileri için bir analiz aracı geliştirilmiştir. Çalışmanın amacı, veri madenciliğinde bir kümeleme tekniği olan k-means algoritmasını incelemek ve bu algoritmayı kullanarak gırtlak kanseri ameliyat verileri üzerine geliştirilen bir yazılım aracılığıyla, verilerin analizini yapmaktır. Uygulamanın tıp doktorlarının kullanımına uygun şekilde verileri çeşitli açılardan analiz etmesi hedeflenmiştir. Çalışmada kullanılan gırtlak kanseri ameliyat verileri, Kocaeli Üniversitesi Tıp Fakültesi Hastanesi, Kulak, Burun ve Boğaz Bölümünden alınmıştır. Veriler dört farklı açıdan analiz edilmiş ve her bir analiz ayrı bir ekranda görüntülenmiştir. Bu çalışmada kullanılan k-means algoritması aracılığı ile veriler kümelenmiştir. Küme sayısı, kullanıcı tarafından 2 ile 9 arasında bir değer olarak belirlenebilmektedir. Sınama sonuçlarına göre, 9 kümeden daha fazlası verimli olmamaktadır. Kümeleme sonucu grafik üzerinde gösterilerek veri içindeki yoğunlaşmaların ortaya çıkması sağlanmıştır. K-means algoritması aşağıdaki özellikleri nedeniyle tercih edilmiştir:

1. Küme sayısının okunan bir parametre olması analizi esnek hale getirmektedir.
2. Algoritmanın uygulanması kolaydır ve hızlı çalışmaktadır.
3. Değişik dağılımlarda başarılı sonuçlar alınabilmektedir.
4. Kategorik verilerle çalışacak şekilde adapte edilebilmektedir.
5. Kümeleme sonuçları hem grafik olarak hem de yazı ve rakamlarla kolayca ifade edilebilmektedir.

Algoritmanın bazı dezavantajları da bulunmaktadır. Çok boyutlu verilerde kötü sonuçlar verebilir ve kümeleme yüksek iterasyon gerektirebilir. Ancak gırtlak kanseri ameliyatlarının verileriyle çalışılırken bu dezavantajlar ortaya çıkmamıştır. Çünkü veriler çok boyutlu değildir ve noktalar düzgün dağıldığından kümeleme yüksek iterasyon gerektirmemiştir.

Yazılım aracılığı ile yapılan analizler sonucunda aşağıdaki yararlar elde edilebilir;

1. Geçmiş verileri analiz ederken değişken parametreler kullanılarak değerlendirme yapılabilir, vakalar için tahminde bulunulabilir,
2. Mevcut ve gelecek vakalar için ameliyat sonrasında tümörün nüks etme olasılığı ve hastanın hayatta kalma olasılığı değerlendirilebilir,
3. Doğru öngörülen ve öngörülemeyen ameliyat öncesi evreler görüntülenerek incelenebilir ve bu şekilde ameliyat öncesi tahmin başarısı değerlendirilebilir,
4. Başarılı ameliyat bilgileri izlenerek, gelecek ameliyat tercihlerinde fikir alınabilir.



Geliştirilen yazılım, tıp doktorlarının geçmiş kayıtları analiz ederek, ileriye dönük tahminde bulunabilmelerini kolaylaştırmaktadır. Karar almada yardımcı olabilecek bir araçtır. Ayrıca araştırma, denetim ve eğitim etkinliklerinde de kullanılabilir. Mevcut durumda, hastanede geçmiş kayıtları analiz etmek için hasta dosyaları tek tek taranarak araştırma yapılmaktadır.

Tez için yapılan literatür taramasında, gırtlak kanseri ameliyat verileri üzerinde, k-means algoritması kullanarak yapılan başka bir Veri Madenciliği çalışmasına rastlanmamıştır. Veri madenciliği çalışmalarında genellikle veriler SPSS ve MATLAB gibi paket programlar aracılığı ile analiz edilmektedir. Ancak paket programlar çeşitli kısıtlar içermektedir. Kullanıcı açısından bu çalışmada geliştirilen yazılımın öğrenme süresi çok kısadır ve kullanılması kolaydır. Tıp alanında veri madenciliği uygulamalarının ilerlemesinin önünde bazı engeller bulunmaktadır. Öncelikle hasta teşhis ve tedavi bilgileri çoğunlukla klasik hastane bilgi sistemlerinde tutulmadığından, çalışmalarda kullanılacak verilerin bulunmasında zorluk yaşanmaktadır. İstenen verilere bulmak zor olduğundan, bireysel çabalarla ve belirli konular için toplanmış verilere ulaşmak gerekmektedir. Verilerin elde edilmesi, çalışmanın başlaması için her durumda yeterli olmamaktadır. Tıp alanındaki verilerde belli bir standardın olmayışı verilerin işlenmesini zorlaştırmaktadır. Çoğu kez, çalışmanın yönlendirilmesi için konuyla ilgili bir tıp doktorunun bulunması ve destek alınması gerekebilir. İlgili tıp doktorunun bilgisayar uygulamaları konusunda bilgi sahibi olmaması çalışmayı olumsuz yönde etkilemektedir. Bu zorlukların en aza indirilmesi, tıp alanında Veri Madenciliği uygulamalarının artmasını sağlayacaktır. Geliştirilen yazılımın gerçek veritabanı üzerinde kullanılması, uygulamanın etkinliğini görmeyi sağlamıştır. Yazılıma eklemeler yaparak, Kocaeli Üniversitesi Tıp Fakültesi Hastanesi, Kulak, Burun ve Boğaz Bölümünde kullanılması planlanmaktadır. Tıbbi veritabanları üzerinde yazılım uygulaması geliştiren çalışmaların azlığı dikkat çekmektedir. Bu tez çalışması, yakın gelecekte tıp alanında geliştirilecek ve Veri Madenciliği algoritmalarını içerecek yazılımlar için bilgilendirici bir kaynak olabilir. Bu çalışma ile veri madenciliğindeki kümeleme algoritmalarının çeşitli tıbbi verilerin değerlendirilmesi ve analizine katkısı gözlenebilir. Bu tip yazılımların başka klinikler için de geliştirilmesinin tıbbi çalışmalarda yararlı olabileceği düşünülmektedir. Analizlerde k-means dışında başka teknikler kullanılarak, sonuçlar karşılaştırılabilir.

VERİ MADENCİLİĞİNDE YENİ YAKLAŞIMLAR

Yapay Bağışıklık Sistemi Bağışıklık sistemindeki etkileşimleri daha iyi anlayabilmek için bağışıklık sisteminin bir modelini oluşturmak ve sistemdeki olayları hesapsal araç olarak



kullanabilmek amacıyla ortaya atılmıştır Bilgisayar bilimcileri, mühendisler, matematikçiler, filozoflar ve diğer araştırmacılar, karmaşıklığı beyne benzer olan bu sistemin özellikle yetenekleri üzerine ilgi duymaktadırlar. Yapay bağışıklık sistemi ile VERİ MADENCİLİĞİ tekniklerinden sınıflama ve kümeleme için yeni yeni birkaç çalışma yapılmıştır. Birliktelik kurallarının keşfi için de tek bir çalışma vardır. Ancak diğer teknikler için de çalışmaların olacağı açıktır. Özellikle sistemin sahip olduğu özelliğiyle dağıtık ve paralel VERİ MADENCİLİĞİ'nde etkili olarak kullanılabilecektir.

Karınca Koloni Optimizasyonu

Karıncalar, yuvalarından bir gıda kaynağına giden en kısa yolu, herhangi görsel ipucu kullanmadan bulma yetisine sahiptirler. Koloni halinde yaşayan karıncalar yiyecek bulmak için ilk olarak öncü karıncaları tek başına gönderirler. Bu öncüler çevreyi araştırarak uygun yiyecek kaynağını bulmaya çalışır. Öncüler yiyecek bulursa, koloninin olduğu yere geri dönerken arkalarında özel bir koku izi bırakarak ilerler. Bu iz sayesinde diğer karıncalar da bu yiyecek kaynağını bulabilirler. Araştırmacılar bu arkalarında iz bırakarak ilerleyen öncü karıncaların uyguladığı yöntemi "sanal karıncalar" oluşturarak bilgisayarlarla simüle ettiklerinde çoğu problemin daha kolay çözülebileceğini gösterdiler. Karınca koloni optimizasyon algoritması VERİ MADENCİLİĞİ'nde sınıflama tekniğinde yeni yeni kullanılmaya başlanmıştır. Elde edilen sonuçlar tatmin edicidir. Daha optimize parametrelerle dağıtık ya da paralel gerçeklemelerle çok daha iyi sonuçların alınacağı kesindir. Diğer tekniklerde de bu yaklaşım kullanılabilir.

Destek Vektör Makineleri

Destek vektör makineleri yeni bir öğrenme metodudur. Çekirdek tabanlı doğrusal olmayan sınıflandırıcıların sinyal işleme, yapay öğrenme ve VERİ MADENCİLİĞİ alanındaki pratik problemlerde iyi sonuçlar verdiği bulunmuştur. V. Vapnik tarafından önerilen destek vektör makineleri (DVERİ MADENCİLİĞİ) ileri yönde beslemeli yeni bir ağ kategorisidir. İstatistiksel öğrenme teorisinde iyi şekilde kurulmuş bir teoriye sahiptir ve sınıflandırma problemlerine yaklaşım için uygundur. Özellikle iki sınıf sınıflandırma probleminde, DVERİ MADENCİLİĞİ iki sınıf arasındaki sınırı büyükleen optimal ayırt etme yüzeyini belirlemekte, yani eğitim kümesi ile ayırt etme yüzeyine en yakın noktaların arasındaki mesafeyi en büyüklemektedir. Kısaca DVERİ MADENCİLİĞİ, doğrusal olmayan bir şekilde ayrılabilen öbekler için optimal hiper düzlemi bulmaya çalışır. Bu yüzden DVERİ MADENCİLİĞİ'nin



VERİ MADENCİLİĞİ’deki uygulamaları özellikle sınıflama tekniğinde ortaya çıkmıştır. Elde edilen sonuçlar bu yöntemin sınıflama tekniğinde oldukça başarılı olduğunu göstermiştir.

Kaos

Kaos kelimesi insanda pek de hoş olmayan çağrışımlar yapar. Karmaşıklık, belirsizlik ve hatta anarşi. Bilimde ise kaos kelimesi belirlenemezlik olarak kabul edilir. Yani günlük yaşamda kullanımı ile bilimde kullanımı oldukça farklıdır. Kaos teorisi engin uygulama alanına sahip olan bir yaklaşımdır. Her türlü alanda uygulanabilme yeteneğinden dolayı, kaos teorisinin bilim dallarını birbirinden soyutlayan engelleri aştığı söylenebilir. Çok küçük görünen bir nedenin kendisinden çok daha büyük sonuçlara yol açabileceği mantığından hareket eden kaos kuramı, düzensizlik ve karmaşadan çok, bu düzensizlik içerisinde belli bir düzeni, düzenli düzensizliği anlamaya yöneliktir. Endüstriyel alanlarda çoğu işlemlerin doğrusal olmamasından dolayı kaotik işlemlerin tahmini yapılmaya çalışılmaktadır. VERİ MADENCİLİĞİ ’de kaos kuramı, kümelemede ve zaman verisinin de kullanıldığı durumlarda birliktelik kural keşfinde kullanılmıştır.

KAYNAKLAR

Tezler

Sıramkaya, E., Veri Madenciliğinde Bulanık Mantık Uygulaması, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2005

Dinçer, E., Veri Madenciliğinde K-Means Algoritması ve Tıp Alanında Uygulanması, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2006

Tiryaki, S., LLojistik Alanında Bir Veri Madenciliği Uygulanması, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2006

Makaleler

Taşkın Ç., Emel, G. G., Veri Madenciliğinde Kümeleme Yaklaşımları ve Kohonen Ağları ile Perakendecilik Sektöründe Bir Uygulama, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 2010

İnternet

<http://www.ferdikucuk.com.tr/programlama/veri-madenciligi-ve-apriori-algoritmasi.html>

(E.T. 05.06.2013)



http://tr.wikipedia.org/wiki/Bulan%C4%B1k_mant%C4%B1k (E.T. 05.06.2013)

Ayık, Y. Z., Özdemir, A., Yavuz, U., Lise Türü ve Lise Mezuniyet Başarısının Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi, Atatürk Üniversitesi, (E.T. 28.05.2013)

Koyuncu A. , Veri Madenciliği Ders Notları, Başkent Üniversitesi , Yönetim Bilişim Sistemleri Bölümü Güz 2010 (E.T. 05.06.2013)

**VERİ
MADENCİLİĞİ
UYGULAMA
PROGRAMLARI**



**M u s t a f a
Y U R T S E V E R**

GİRİŞ

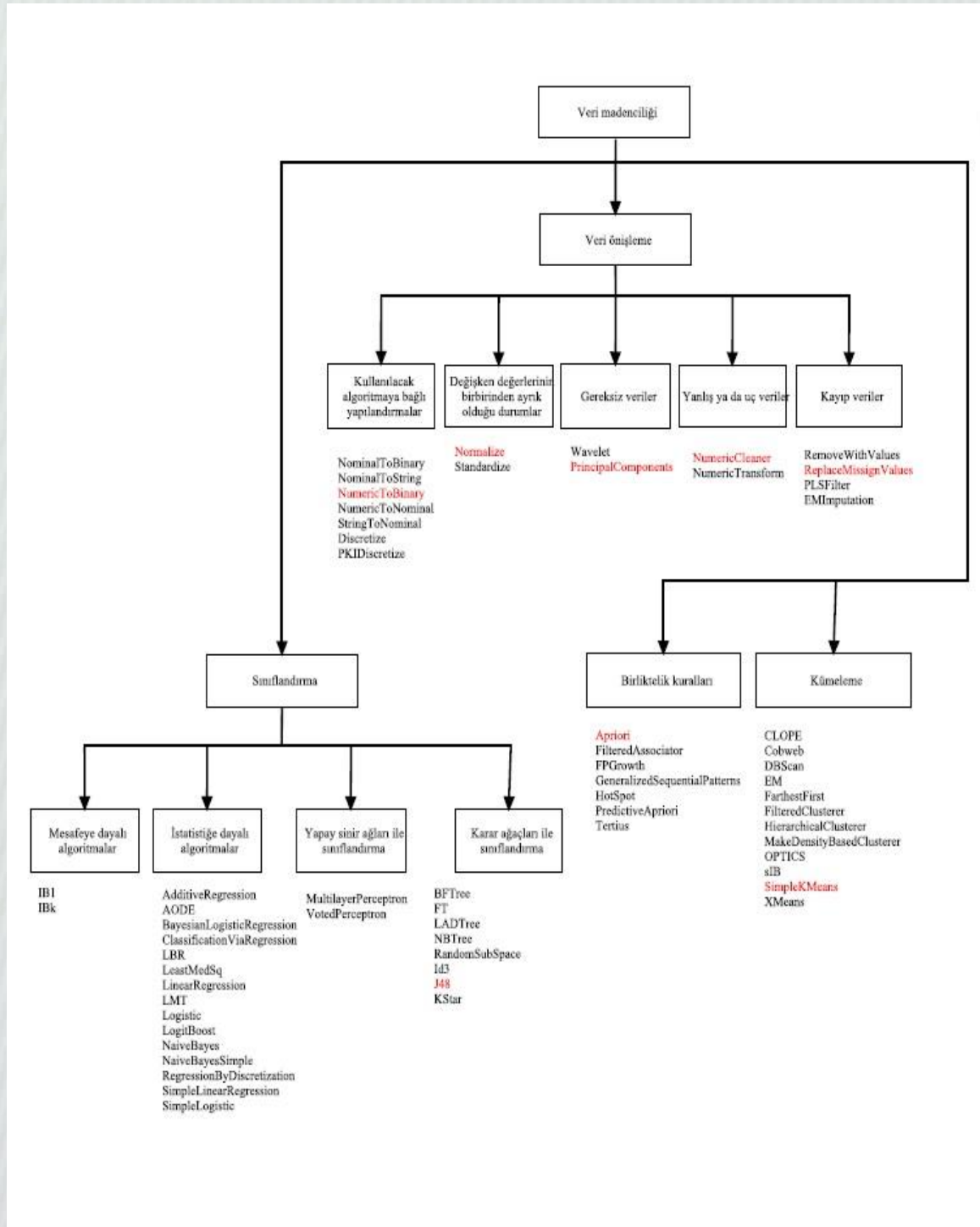
Veri Madenciliği, büyük ölçekli veriler arasından bilgiye ulaşma, veriyi madenleme işlemidir. Veri tabanları veya dosyalarda bulunan verilerin belirli istatistik yöntemleri kullanılarak kullanılabilir hale getirilmesi işlemine veri madenciliği denir. Veri madenciliği veriden bilgi elde etmek için kullanılan tekniklerin bütünü olarak da ifade edilebilir. İstatistiksel analiz tekniklerinin, genetik algoritma yöntemlerinin ve yapay zekâ algoritmalarının bir arada kullanılarak veri içerisindeki gizli bilgilerin açığa çıkarılması ve verinin kullanılabilir bilgiye dönüştürülmesi sürecidir. Görsel veri madenciliği ise görselleştirmeyi insan ile bilgisayar arasında bir iletişim kanalı olarak kullanarak yeni ve yorumlanabilir örüntüler ortaya çıkarma sürecidir.

Veri toplama kaynaklarının ve bilişim teknolojisinin hızlı gelişimi neticesinde kurumlar üretmiş oldukları büyük boyutlardaki verilerden anlamlı ve yararlı bilgiler ortaya çıkarmakta zorluklar yaşamaktadırlar. Diğer taraftan geleneksel istatistiksel yöntemler, bu tür verileri çözümlemekte yetersiz kalmaktadır

Veri Madenciliği işlemlerini gerçekleştirmek için ticari ve açık kaynak olmak üzere birçok araç bulunmaktadır. Bu çalışmada açık kaynak kodlu Veri Madenciliği programlarından olan RapidMiner, WEKA, Orange, KNIME anlatılmış ve bu programlar hakkında genel bilgi verilmiştir.

Veri madenciliği yöntemleri ve programlarının amacı büyük miktarlardaki verileri etkin ve verimli hale getirmektedir. Bilgi ve tecrübeyi birleştirmek için Veri Madenciliği konusunda geliştirilmiş yazılımların kullanılması gerekmektedir. Bu kapsamda, pek çok ticari ve açık kaynak program geliştirilmiştir. Ticari programların başlıcaları SPSS Clementine, Excel, SPSS, SAS, Angoss, KXEN, SQL Server, MATLAB'dır. Açık kaynak programlardan başlıcaları ise Orange, RapidMiner, WEKA, Scrip-tella ETL, jHepWork, KNIME, ELKI' dir.

Veri madenciliğinin temel amacı elimizde bulunan veriden gizli kalmış örüntüleri çıkarmak, verinin değerini arttırmak ve veriyi bilgiye dönüştürmektir. Günümüzde veri madenciliği; bankacılık, pazarlama, sigortacılık, telekomünikasyon, borsa, sağlık, endüstri, bilim ve mühendislik gibi birçok dalda uygulama alanı bulunmaktadır. Şekil 1'de Veri madenciliğinde kullanılan programlar özetlenmiştir.



Şekil 1. Veri Madenciliğinde Kullanılan Programlar



Veri madenciliğini sistematik bir şekilde yürütebilmek için veri madenciliği uygulamacıları tarafında geliştirilen bazı süreçler mevcuttur. Avrupalı şirketlerin oluşturduğu bir konsorsiyum tarafından geliştirilen CRISP-VM ve SAS enstitüsü tarafından geliştirilmiş olan SEMMA metodolojileri bunlara örnek olarak verilebilir. Bu süreçlere ait ana başlıklar aşağıdaki gibidir.

CRISP-VM Süreci:

1. İşi Anlama
2. Veriyi Anlama
3. Veri Hazırlama
4. Model Oluşturma
5. Test ve Değerlendirme
6. Uygulama

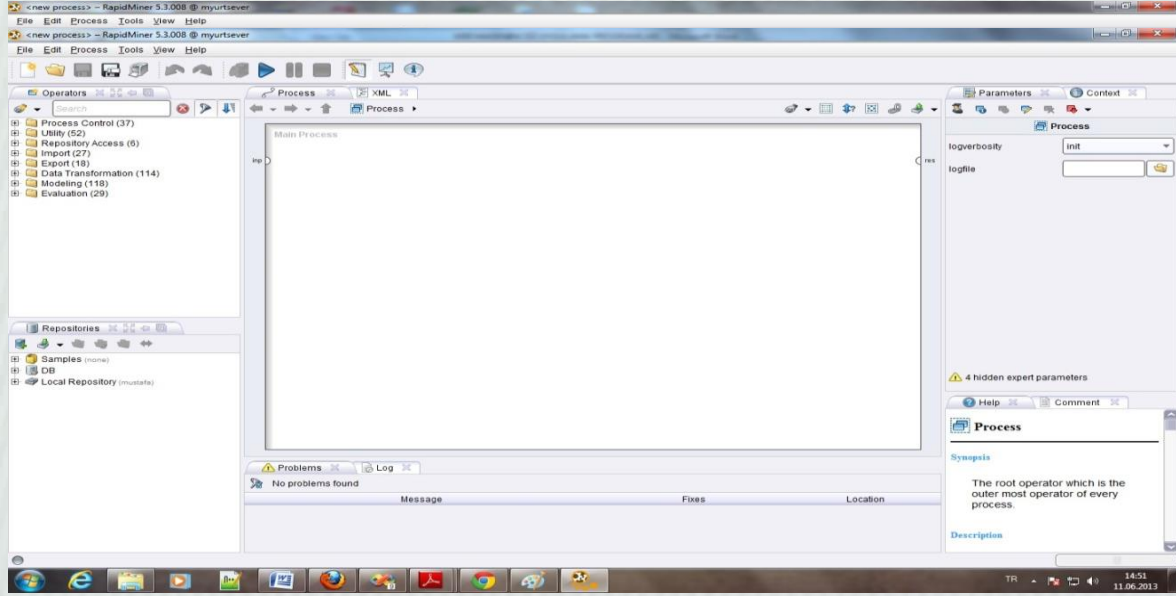
SEMMA Süreci:

1. Sample (Veriyi temsil edecek bir “Örneklem Üretme ya da Alma”)
2. Explore (Verinin görselleştirilmesi ve temel tanımlamasına ilişkin “Keşif”)
3. Modify (Değişkenlerin seçimi ve dönüşümü içeren “Değiştirme”)
4. Model (Çeşitli istatistiksel ve makine öğrenmesi teknikleri ile “Modelleme”)
5. Assess (Modellerin doğruluğu ve kullanılabilirliğini “Değerlendirme”)

AÇIK KAYNAK KODLU VERİ MADENCİLİĞİ PROGRAMLARI

RAPIDMINER (YALE)

Amerika’da bulunan YALE üniversitesi bilim adamları tarafından Java dili kullanılarak geliştirilmiştir. RapidMiner’da çok sayıda veri işlenerek, bunlar üzerinden anlamlı bilgiler çıkarılabilir. Aml, arff, att, bib, clm, cms, cri, csv, dat, ioc, log, mat, mod, obf, bar, per, res, sim, thr, wgt, wls, xrff uzantılı dosyaları desteklemektedir. Diğer programlar gibi birkaç tane format desteklememesi YALE’nin artılarından. Makine öğrenme algoritmaları olarak destek vektör makinelerini içeren büyük sayıdaki öğrenme modelleri için sınıflandırma ve regresyon, Karar Ağaçları, Bayesian, Mantıksal Kümeler, İlişkilendirme Kuralları ve Kümeleme için birçok algoritma (k-means, k-medoids, dbscan), WEKA’da olan her şey, veri ön işleme için ayırma, normalleştirme, filtreleme gibi özellikler, genetik algoritma, yapay sinir ağları, 3D ile verileri analiz etme gibi birçok özelliği bulunmaktadır. 400’den fazla algoritmaya sahiptir. Oracle, Microsoft SQL Server, PostgreSQL veya MySQL veritabanlarından veriler YALE’ye aktarılabilir.



Şekil 2. YALE Arayüzleri

WEKA

WEKA bir proje olarak başlayıp bugün dünya üzerinde birçok insan tarafından kullanılmaya başlanan bir Veri Madenciliği uygulaması geliştirme programıdır. WEKA Java platformu üzerinde geliştirilmiş açık kodlu bir programdır. Weka veri işleme, sınıflandırma, regresyon, kümeleme, ilişki kuralları ve görüntüleme araçları içerir. Weka Java Database Connectivity kullanarak SQL veritabanlarına erişim sağlar ve bir veritabanı sorgusundan dönen sonucu işleyebilir. Çoklu ilişkisel veri madenciliği yapamaz ama Weka kullanılarak işlemek için bir koleksiyon bağlı veritabanı tablosunu tek tabloya dönüştürebilen ayrı bir yazılım vardır.

WEKA aşağıdaki özelliklere sahiptir:

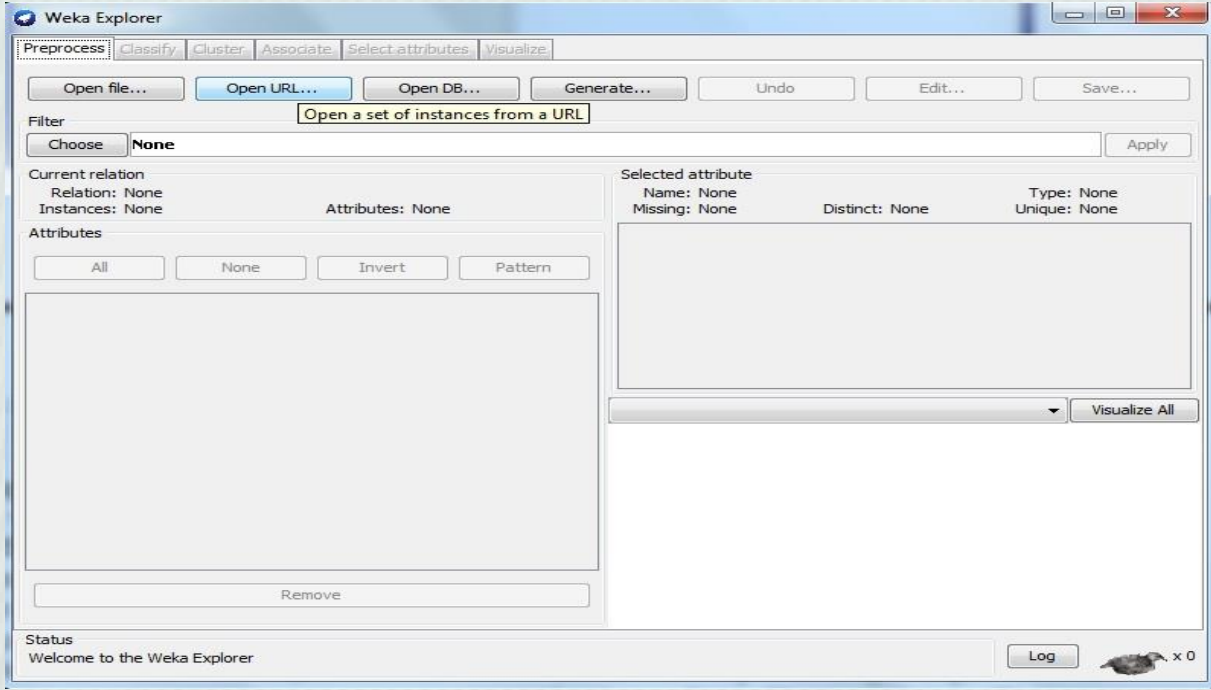
- Veritabanındaki analiz ve ön- işleme özelliklerinin ve verinin doğruluğunu değerlendirme.
- Örnek setlerin uygun sınıflara bölünüp sınıf niteliklerinin tanımlanması
- Sınıflandırma için kullanılacak muhtemel özelliklerin çıkarılması
- Öğrenme işleminde kullanılması için özelliklerin bir alt set olarak seçilmesi
- Seçilen veri seti için mümkün sapmaların araştırılması ve etkisinin nasıl önlenebileceği.
- Örnek alt setin seçilmesi, örneğin makine öğrenme baz alınarak yapılan kayıtlar.
- Öğrenme işlemi için sınıflandırma algoritması programı
- Seçilen algoritmanın performansını tahmin etmek için bir test yöntemine karar verilmesi



Şekil 3. WEKA Application Penceresi

WEKA çalıştırıldıktan Şekil 3’de görüldüğü gibi, Application menüsünde çalışılabilecek modlar listelenmektedir. Bunlar komut modunda çalışmayı sağlayan Simple CLI, projeyi adım adım görsel ortamda gerçekleştirmeyi sağlayan Explorer ve projeyi sürükleyip bırak yöntemiyle gerçekleştirmeyi sağlayan KnowledgeFlow seçenekleridir.

Explorer seçeneği seçildikten sonra üzerinde çalışılacak verilerin seçilmesi, bu veriler üzerinde temizleme ve dönüştürme işlemlerinin gerçekleştirilebilmesini sağlayan yukarıdaki ekran ile karşılaşılmaktadır. WEKA’da import edilebilir herhangi bir text dosyasındaki verileri WEKA ile işlemek olanaksızdır, Arff, Csv, C4.5 formatında bulunan dosyalar, ayrıca Jdbc kullanılarak veritabanına bağlanıp burada da işlemler yapılabilir.

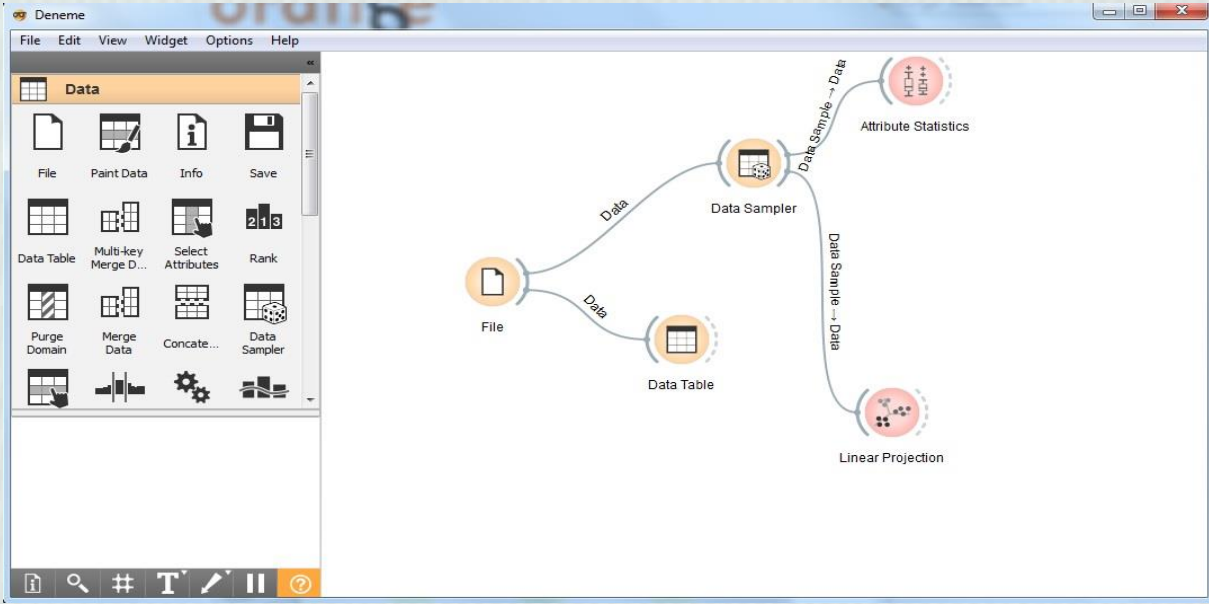


Şekil 4. WEKA Veri Giriş Ekranı

Orange

Orange kullanıcıya veri hazırlama, keşifsel veri analizi, modelleme gibi imkânlar sağlayan bir açık kaynak veri madenciliği programıdır. Orange C ++ tabanlı olup; makine öğrenimi ve veri madenciliği ile ilgili çok fazla algoritma içermektedir. Programda veri madenciliği işi görsel programlama ya da Python scripting ile yapılabilir. Orange kullanıcı dostu güçlü ve esnek görsel programlama, arama amaçlı veri analizi ve görüntüleme ve Python bağlama ve kodlama için kütüphaneler içeren bileşen tabanlı bir veri madenciliği ve makine öğrenmesi yazılım takımıdır. Veri ön işleme, özellik skorlama ve filtreleme, modelleme, model değerlendirme ve keşif teknikleri gibi geniş kapsamlı bileşen seti içerir.

Orange GPL (Genel Kamu Lisansı) altında ücretsiz olarak dağıtılmaktadır. Ljubljana Üniversitesi (Slovenia) Bilgisayar Fakültesi ve Bilgi Bilimi'nde geliştirilmiştir. Orange, Linux, Apple's Mac OS X, ve Microsoft Windows'un çeşitli versiyonlarını destekler. Program yalnızca metin dosyadan (TXT, TAB) veri import işlemine imkan sağlamaktadır.



Şekil 5. Orange Arayüzü

KNIME

KNIME, kullanıcıya görsel veri akışı sağlayan, analiz adımlarının tamamını veya bir kısmı üzerinde seçim yapılarak yürütülmesini sağlayan ve veri ve modelden sonuçlarını interaktif olarak sağlayan modüler bir veri keşif platformudur. KNIME biyoenformatik ve bilgi madenciliği bölümü tarafından Almanya'daki Konztanz üniversitesinde geliştirilmiştir.

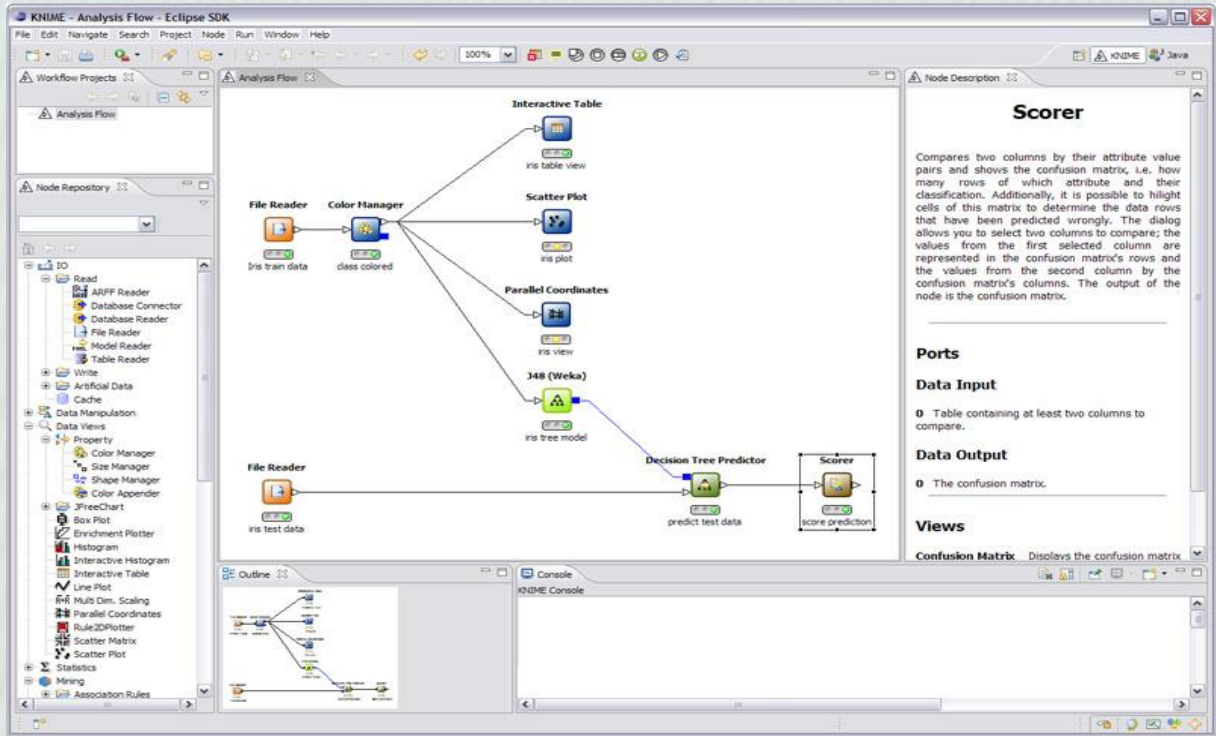
KNIME temel versiyonu, veri ön işleme ve temizleme, analizler ve veri madenciliğin de dahil scatter plots, parallel coordinates ve bir çok interaktif görüntüleme gibi 100'den fazla iş parçasını G/Ç bilgisi olarak birleştirir. KNIME çok fazla fonksiyonelliklere sahiptir.

G/Ç : Dosyalardan yada veritabanlarından veri alışı verişi yapar.

Veri Manipülasyon: Filtreleyerek veri ön işleme, gruplama, pivot, kovalama, normalleştirme, toplama, karıştırma, örnekleme, bölümleme gibi.

Görüntüleme: Veriyi ve sonuçlarını birçok görüntüleme aracı ile verinin keşfinin sağlanmasını kolaylaştırır.

Madencilik: Demetleme, karar ağaçları, kural oluşturma, ilişki kuralları, sinir ağları, destek vektör makineleri.



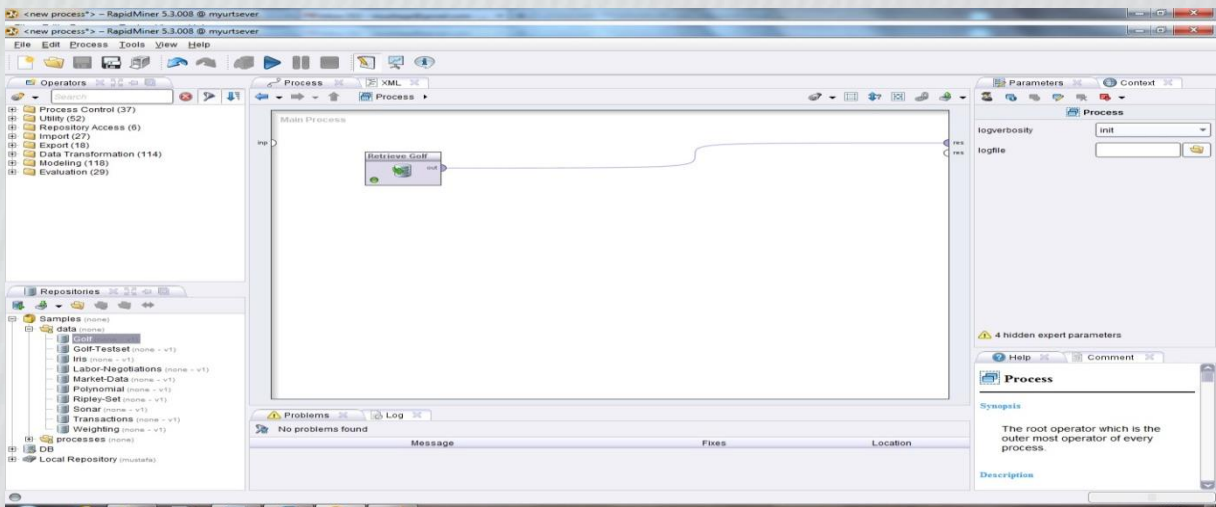
Şekil 6. KNIME Karar Ağacı Menüsü

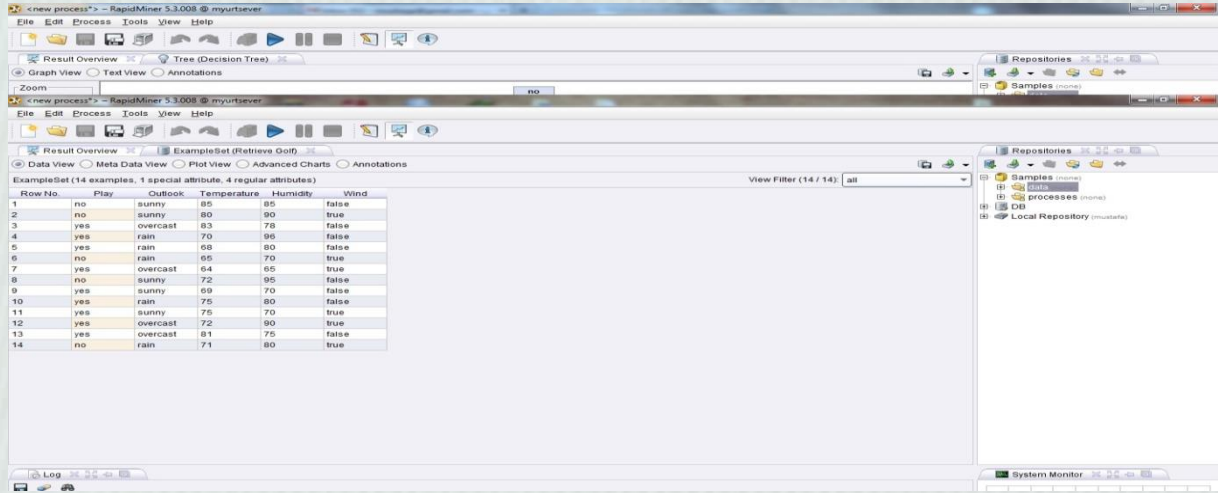
ÖRNEK UYGULAMALAR

Rapid Miner Karar Ağacı Örneği

Şekil 7’de görüldüğü gibi sol taraftaki Samples menüsünde bulunan data kısmından Golf isimli hazırlanan data seti sürüklenerek ekrana bırakılır ve ardından Run butonuna basarak program çalıştırılır. Bu durumda Şekil 8’de görüldüğü gibi ekrana data setinin geldiği görülebilir.

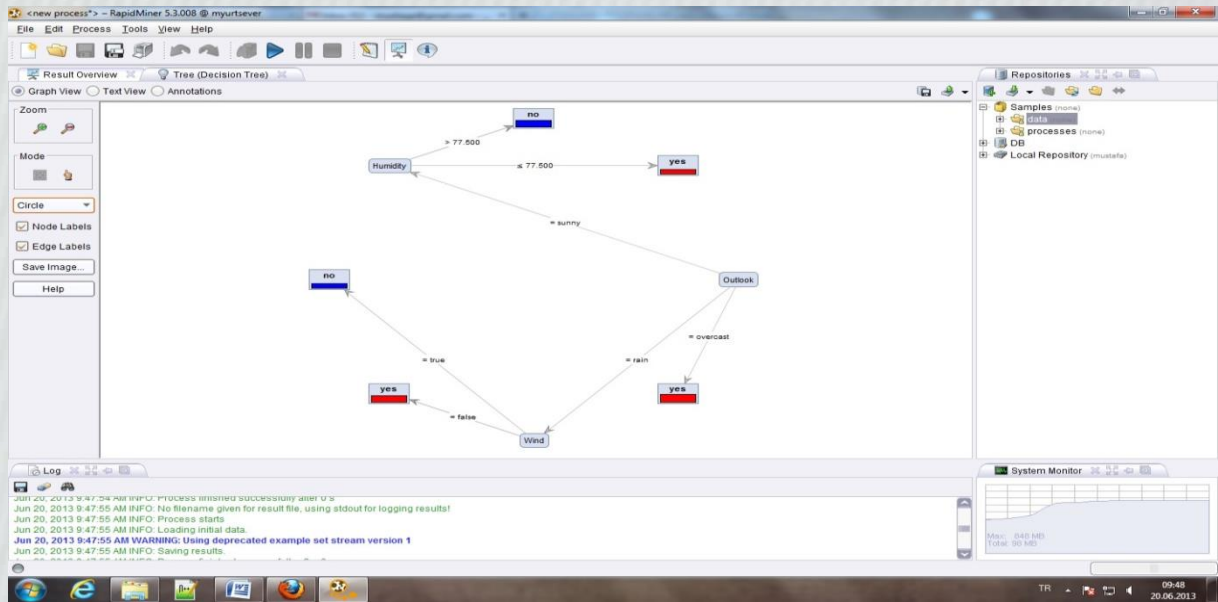
Şekil 7. Rapid Miner Karar Ağacı Başlangıç Menüsü





Şekil 8. Data Seti Ekranı

Daha sonra design perspective butonuna basılarak tasarım ekranına dönülür. Sol tarafta bulunan Operators menusunun altında bulunan Modeling menüsünden Tree Induction-Decision Tree seçeneğini sürükleyip bırak yöntemiyle ekrandaki data seti ile aynı çizgi üzerine yerleştirilir ve Run butonuna tıklanır.

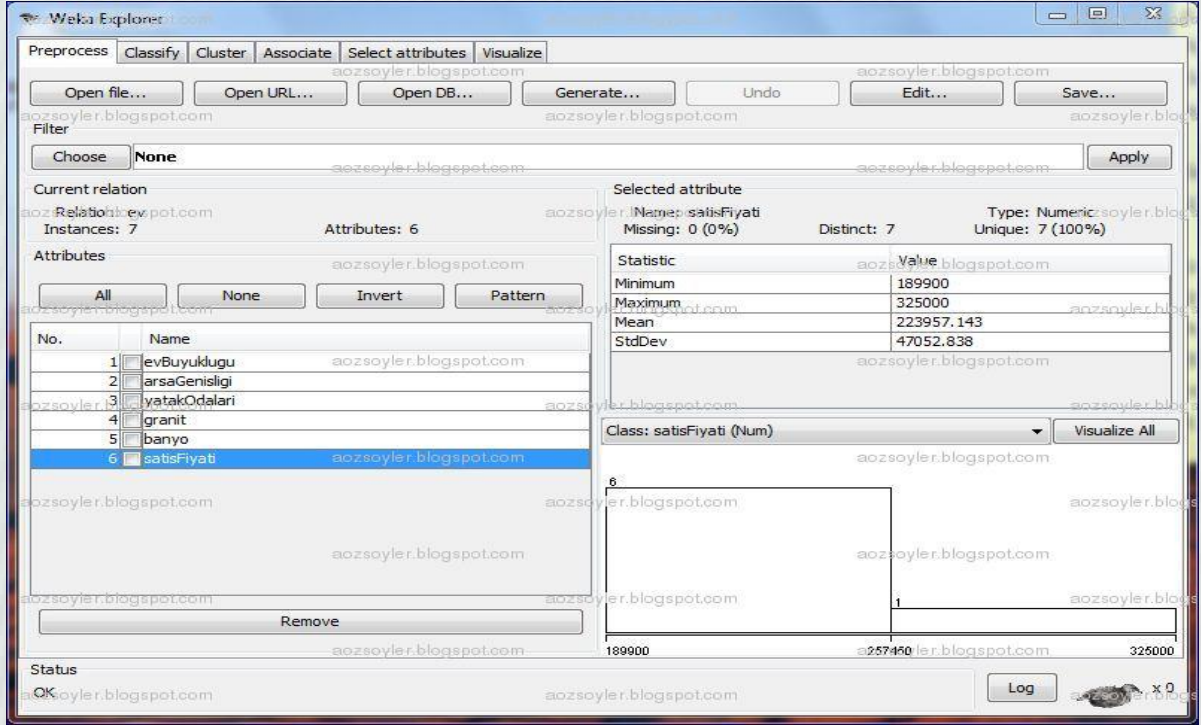


Şekil 9. Sonuç Ekranı

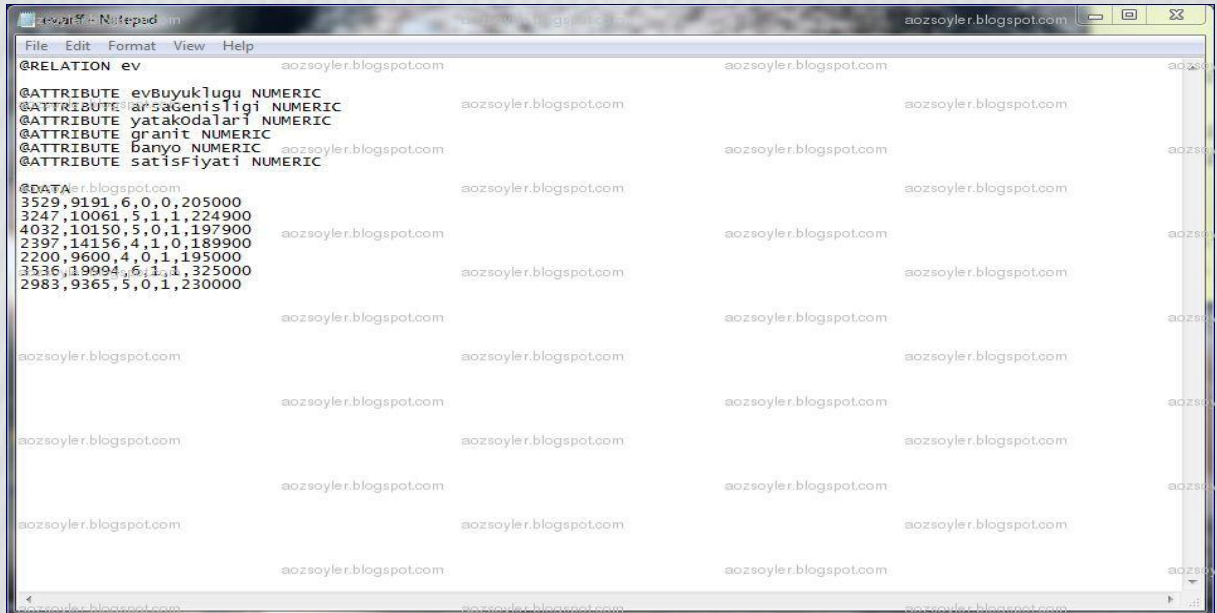
WEKA Daire fiyatı hesaplama örneği (Lineer Regresyon Analizi)

Şekil 10'daki ekran görüntüsünde, daha önceden hazırlanan ev.aff uzantılı (Şekil 11) veri dosyasındaki veriler WEKA yazılımına aktarılır.

Bu örnekte WEKA yazılımı ile bu aff uzantılı dosya ile satışFiyatı niteliğine ait tahmin formülü bulunmaya çalışılmıştır.

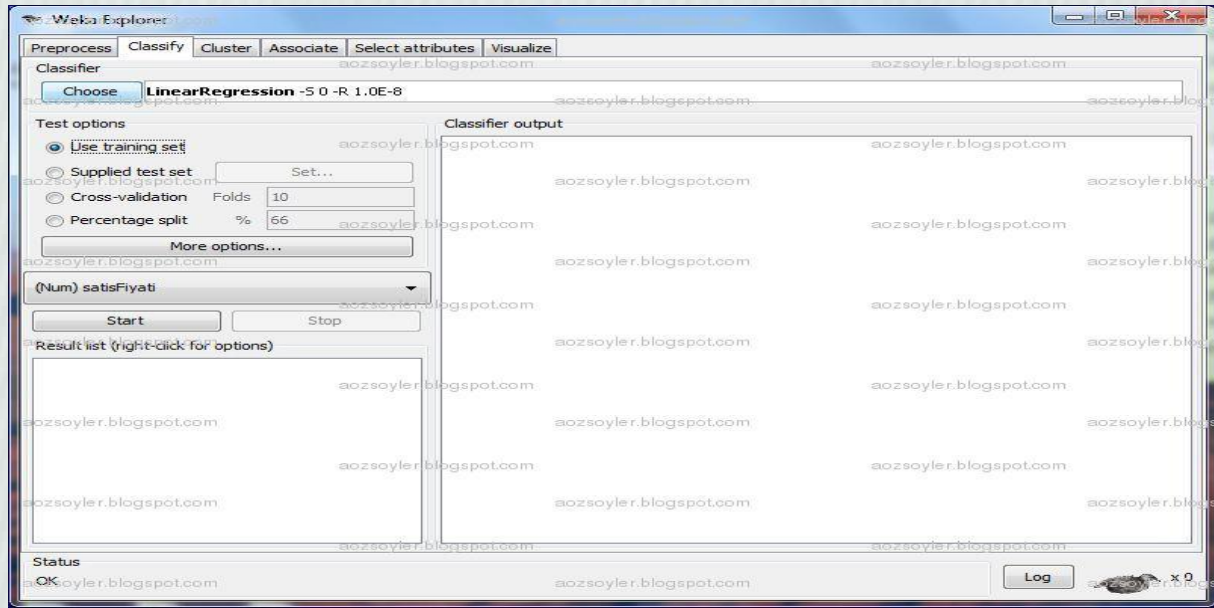


Şekil 10. Veri Seti Ekran Görüntü



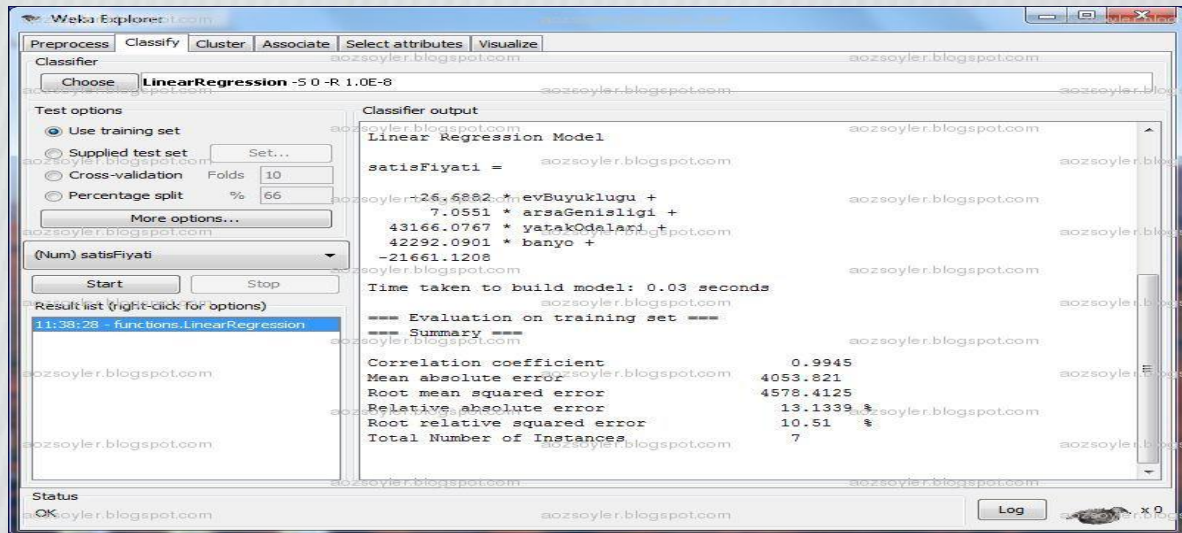
Şekil 11. aff Uzantılı Veri Dosyası

Örnekte Classify sekmesi altından sınıflandırma yöntemlerinden biri olan Lineer Regresyon Analizi kullanılmıştır. Öncelikle Choose çubuğundan ilgili analiz seçilir (Şekil 12).



Şekil 12. Yöntem Seçimi

Sınıflandırma işlemini başlamak için üst-sol kısımdan “Use training set” seçeneğini işaretledikten sonra Start tuşuna basmak yeterli olacaktır (Şekil13).



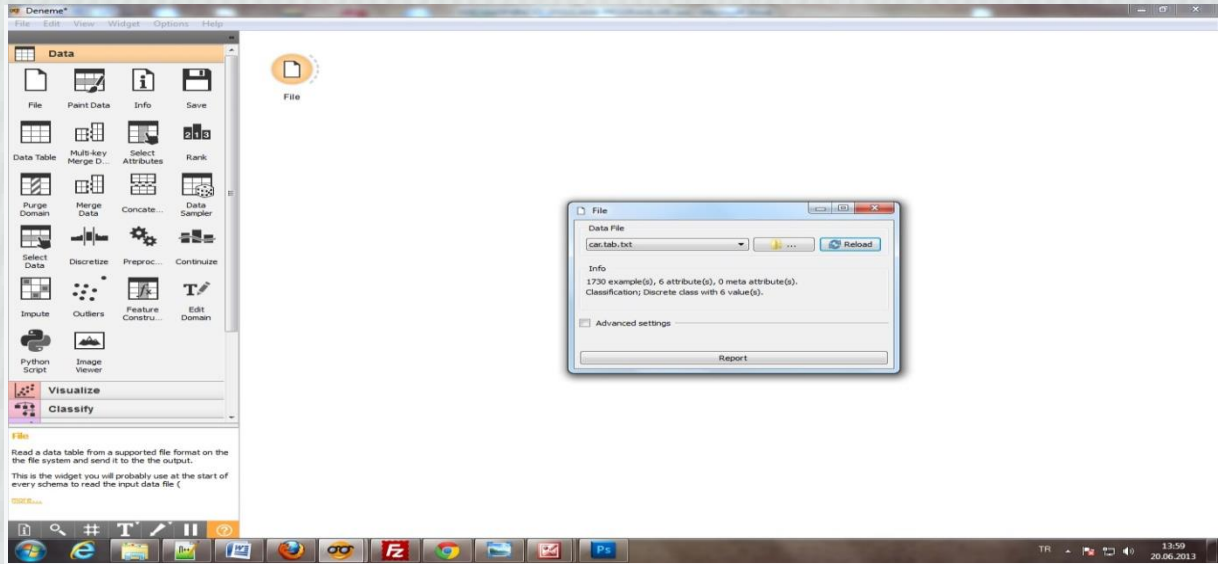
Şekil 13. Analiz Sonuçları

Orange

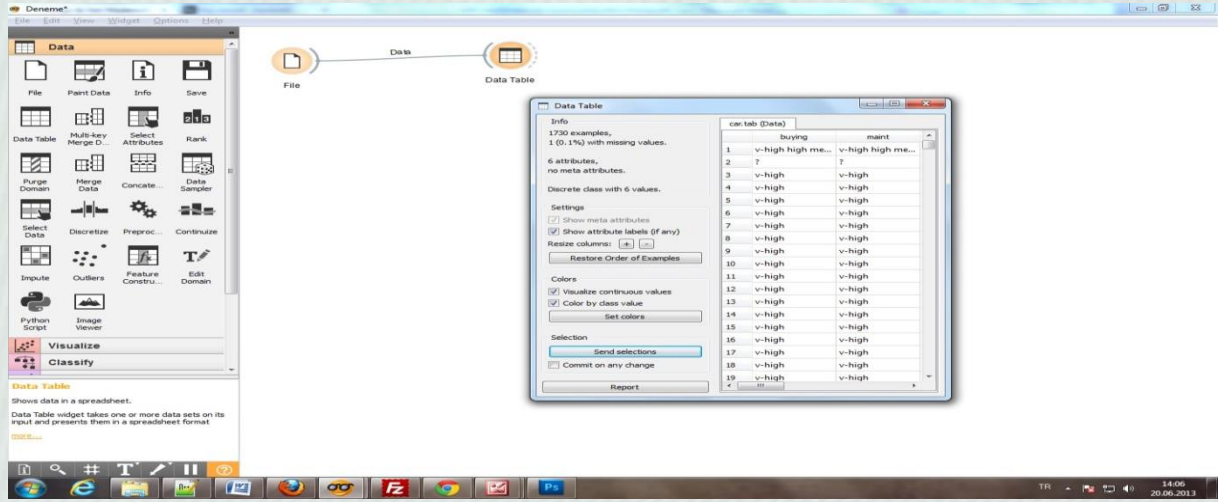
File menüsü dosyalardan gelen verileri almak için kullanılır. Veriler genellikle “.tab” formatındadır. Şekil 14’de görüldüğü gibi “Data” tabının en başında yer almaktadır.

File alana sürüklenir ve File’a çift tıklayarak Şekil 15’deki ekran görüntüsü elde edilir. Yazılım Car.tab formatındaki veri dosyasını okur. Örnek uygulamada 1728 örnek data ve 5 tane öznitelik bulunmaktadır. Car.tab dosyasındaki bazı veriler değiştirilirse Reload düğmesine basıldığında veriler otomatik olarak yenilenecektir.

Dosyadan gelen verileri bir tabloya atarak nasıl veriler olduğunu görmek için Şekil 15’deki gibi Data Table sürüklenip bırakılır. Aralarındaki bağlantıyı kurduktan sonra üzerine çift tıkladığında 2 widget arasındaki çizgiler birleşir.

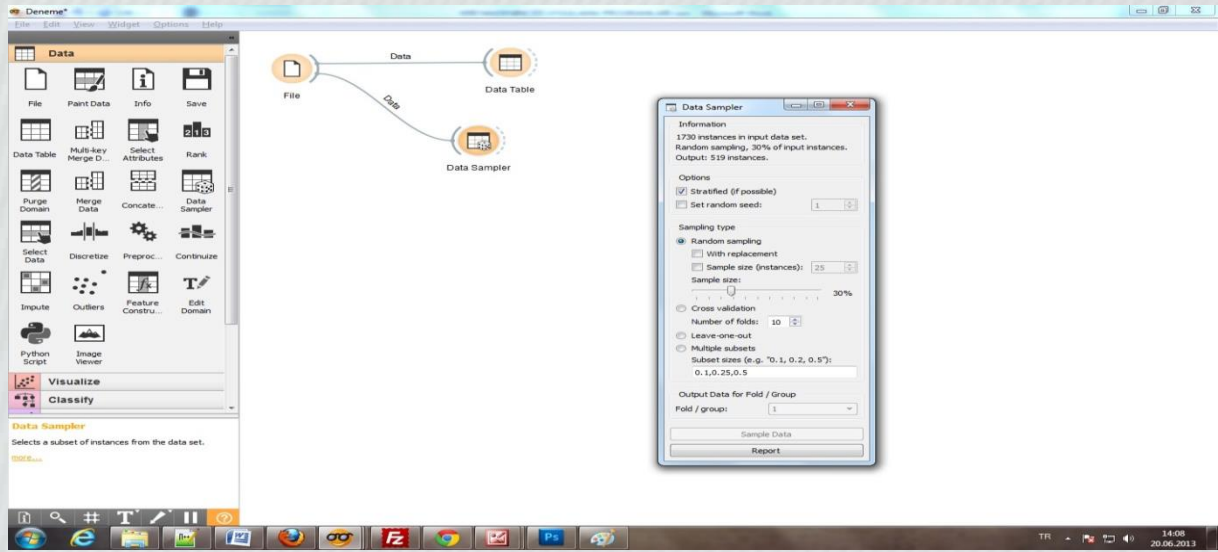


Şekil 14. Orange Başlangıç Ekranı



Şekil 15. Verilerin Aktarılması

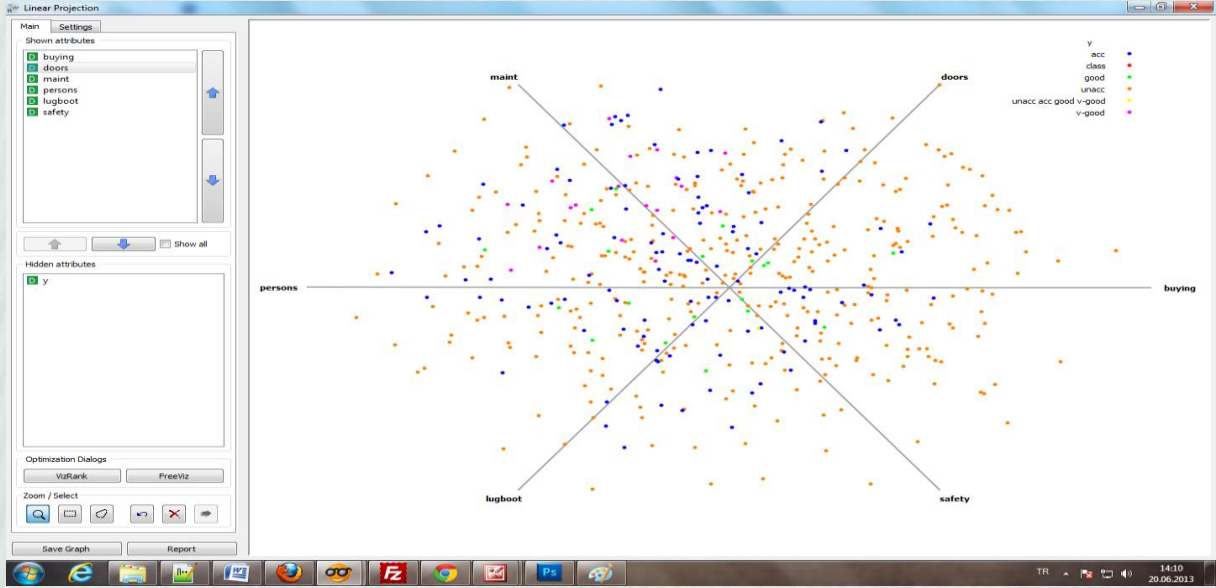
Eğer bir kümeleme ve analiz çıkarımında bulunulacaksa Data Sampler kullanmak zorunluluktur. Şekil 16'da görüldüğü gibi Data Sampler ile File ile birleştirilir. Örnekte Sampling Type ile mevcut verinin %30 unu random olarak alınmıştır. İsteğe göre veri seti arttırılabilir veya azaltılabilir.



Şekil 16. Data Sampler ile File İlişkisi

Visualize tabına tıklayarak Linear Projection eklenir ve daha sonra Data Sampler ile birleştirilir. Çift tıklayarak verinin nasıl görüldüğüne bakılır. Örnekte arabaların satış miktarını ölçmek için veriyi analiz etmek istenmiştir. Şekil 17'de görüldüğü gibi kapı-kişi sayısı-araç boyutu ve güvenlik seviyesine göre araba satış oranları kümelanmiş halde gösterilmektedir. Kapı sayısının

ve kiři sayısının fazla olduđu arabalar çok az satılırken araç boyutu ve güvenliğin değeri arttıkça kümelemedeki nokta sayısı artmaktadır.



Şekil 17. Analiz Sonuçları

SONUÇ

Her geçen gün katlanarak artan veri miktarından dolayı bilgiye ulaşmak zorlaşmıştır, Veri madenciliği, bilginin her geçen gün daha da hızla çoğaldığı dünyada en güncel araştırma alanlarından biridir. Veri madenciliği büyük miktardaki veriden kullanılabilir bilgiyi üretmek için kullanılır. Veri Madenciliği uygulamaları yapmak için bilgisayar programlarına ihtiyaç vardır. Bu programlar içerisinde veri kümeleme, karar ağaçları, bayes sınıflandırıcılar, apriori yöntemi gibi birçok algoritma mevcuttur. Algoritmalar sayesinde işlenen verilerden, bilgi çıkarımı yapılabilmektedir. Veri akış diyagramlarını temel alan veri madenciliği yazılımları kullanıcıları karmaşık kod kümeleri arasında boğulmaktan kurtararak bilgi keşfi sürecini kısaltmayı hedeflemektedir.

Kaynakça

Bilgin, T.T, Veri Akışı Diyagramları Tabanlı Veri Madenciliği Araçları Ve Yazılım Geliştirme Ortamları, Maltepe Üniversitesi

Dener, M., Dörterler, M., & Orman, A. (2009) Açık Kaynak Kodlu Veri Madenciliği Programları: Weka'da Örnek Uygulama, XI. Akademik Bilişim Konferansı Bildirileri: 11-13 Şubat 2009-Şanlıurfa

Tapkan, P., Özbakır, L., & Baykasoğlu, A. (2011). Weka ile Veri Madenciliği Süreci ve Örnek Uygulama. Endüstri Mühendisliği Yazılımları ve Uygulama Kongresi

Tekerek, A. (2011) Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği Araçları. , XIII. Akademik Bilişim Konferansı Bildirileri: 2-4 Şubat 2011-Malatya

Orange ile Veri Madenciliği. (2012, Nisan 11). Haziran 12, 2013 tarihinde <http://www.datawarehouse.gen.tr/Makale.aspx?ID=516> adresinden alındı

Orange indirme linki : <http://orange.biolab.si/>

RapidMiner indirme linki : <http://rapid-i.com/content/view/26/201/>

Weka indirme linki : <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

Weka örnek uygulama: <http://aozsoyler.blogspot.com/2011/05/weka-ile-daire-fiyatn-tahmin-etme.html>



ULAŞIM TÜRÜ SEÇİMİNİ ETKİLEYEN FAKTÖRLERİN BELİRLENMESİNDE VERİ MADENCİLİĞİ YÖNTEMİNİN KULLANILMASI



Özge
ORAL

Giriş

Ulaşım modelleme çalışmaları içerisindeki adımlardan biri olan ulaşım tür seçimi aşaması özellikle ülkemizde yapılan çalışmalarda ayrıntılı olarak üzerinde durulmayan bir aşamadır. Literatür incelendiğinde ulaşım tür seçimini etkileyen birçok faktör olduğu görülmüştür. Bu çalışmada İzmir ili Buca ilçesindeki kişilerin bireysel özellikleri ve tercihlerinden hareketle tür seçiminde etkili olan faktörler belirlenmeye çalışılmıştır. Faktörlerin belirlenmesi aşamasında veri madenciliği yöntemlerinden faydalanılmıştır. Çalışmada kullanılan veriler İzmir Ulaşım Planı çalışmaları kapsamında gerçekleştirilen 2007 Hanehalkı anket çalışmasından elde edilmiştir.

Ulaşım Tür Seçimini Etkileyen Faktörler

Ulaşım planlaması, seyahat üretimi, seyahat dağılımı, ulaşım türünün *seçimi* (türel dağılım) ve trafik *atama* olmak üzere dört *aşamalı* bir planlama sürecidir. Özellikle ülkemizde tür seçimi aşaması gözardı edilir. Tür seçimi aşamasında tür seçiminin belirlenmesinde literatürde belirtilen çeşitli faktörler bulunmaktadır. Bu faktörler; yolculuk yapan kişinin özellikleri, yolculuğun özellikleri, ulaşım türünün özellikleri olarak ayırmak mümkündür.

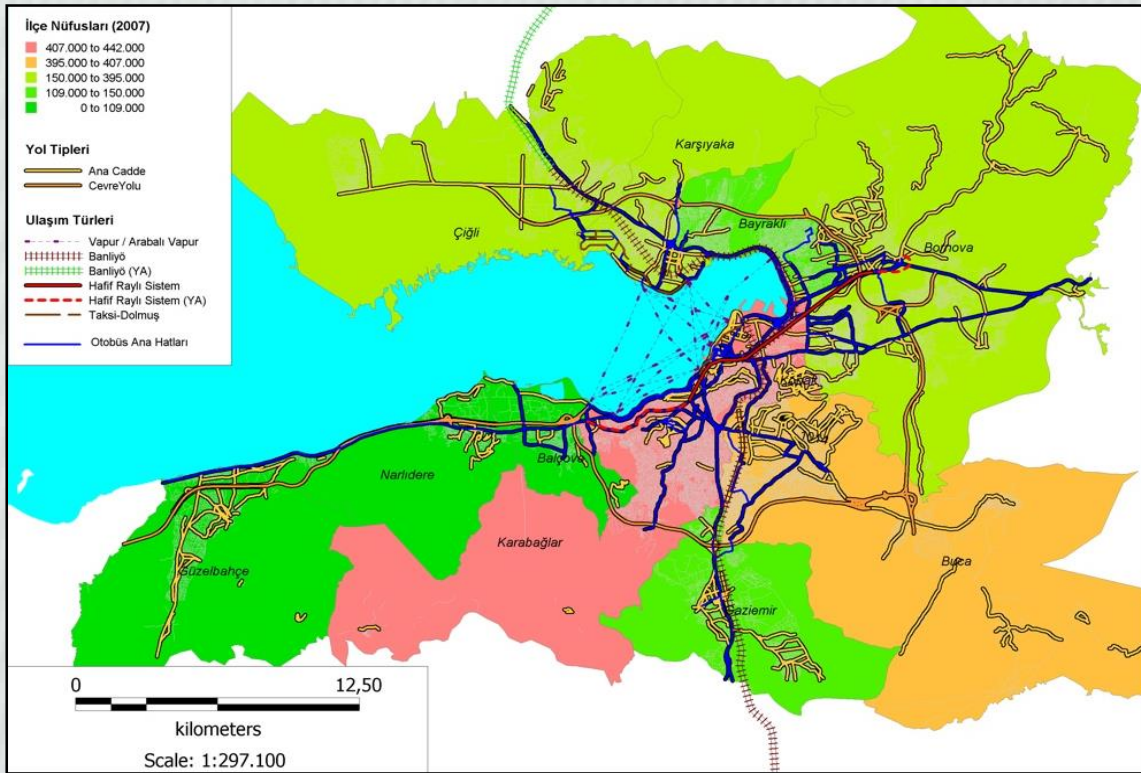
Yolculuk Yapan Kişinin Özellikleri	Yolculuğun Özellikler	Ulaşım Türünün Özellikleri
Araba Sahipliği/ İmkanı Sürücü Belgesi Sahipliği Hanehalkı Yapısı Hanehalkı geliri Gereklikler Konut Yoğunluğu	Seyahatin Amacı Seyahatin Başladığı Zaman Başlangıç ve Bitiş Noktası Seyahatin Uzunluğu	Niceliksel - Harcanan Süre - Katlanılan Maliyetler - Park Uygunluğu / Ücreti -Kapasitesi - Hız
		Niteliksel - Rahatlık ve Kolaylık - Güvenirlilik ve Düzenlilik - Koruma ve Güvenlilik

Tablo 1. Ulaşım Tür Seçimini Etkileyen Faktörler

Ulaşım tür seçimini etkileyen faktörlerin, ulaşım planlaması yapılan çalışma alanına göre önem dereceleri değişmektedir. Dolayısıyla çalışma alanı belirlendikten sonra çalışma alanından elde edilmiş verilerle bu faktörlerin belirlenmesi gerekir.

Çalışma Alanı ve Veri Seti

İzmir'deki Merkez ilçelere (Bayraklı, Bornova, Buca, Çiğli, Gaziemir, Güzelbahçe, Karabağlar, Karşıyaka, Konak, Narlıdere) ait Mevcut Toplu Ulaşım Ağı ve Ana caddeleri Şekil 1'de görüldüğü gibidir. Ayrıca bu şekilde 2007 yılında ait nüfus bilgilerinin ilçelere göre dağılımı gösterilmiştir.



Şekil 1. İzmir Merkez İlçelerine Ait Toplu Ulaşım Ağı ve 2007 Nüfus Dağılımı

İzmir Büyükşehir Belediyesi tarafından İzmir Ulaşım Ana Planı çalışmaları kapsamında 2007 yılında gerçekleştirilmiş Hanehalkı Konut Anketi 4 sayfadan ve yaklaşık 55 sorudan oluşan bir anket çalışmasıdır. 2007 yılı hanehalkı anket çalışması kapsamında elde edilen yolculuk sayılarının ilçelere dağılımı Şekil 2'de gösterildiği gibidir.

Bu anket çalışması kapsamında 20000 haneden 99258 kişiye ait adres bilgisi, konut bilgisi, araç sahipliği, taşıt bilgisi, hanedeki kişi bilgisi, yolculukta kullanılan araç sayısı ve yolculuk bilgisi toplanmıştır



Şekil 2. 2007 Hanehalkı Anket Çalışmasına Göre İzmir Merkez İlçelerindeki Yolculuk Sayıları

Grup Adı	Değişken adı
Konut Bilgisi (her hane için)	Konut türü, Mülkiyet Türü, İkinci Ev sahipliği, Kişi Sayısı, Araç Sayısı, Özel Araç Sayısı
Hanedeki kişi bilgisi (hanedeki her bir kişi için)	Kişi Sıra no, Hane reisine yakınlık, Yaş, Yaş grubu, Cinsiyet, Öğrencilik, Eğitim, Çalışma durumu, Süreklilik, Ehliyet, Kentkart
Yolculuk bilgisi (hanedeki her bir kişi için)	Yolculuk no, Yolculuk nedeni, Başlangıç mahallesi, Baş. Arazi kullanımı, Baş. Mahalle kodu, Başlangıç zamanı, Ulaşım aracı, Bitiş mahallesi, Bitiş mahallesi kodu, Bitiş arazi kullanımı, Bitiş zamanı, Yolculuk süresi, Sorunlar

Tablo 2. Çalışma Kapsamında Dikkate Alınan Değişkenler

Tür seçimini etkileyen faktörlerin belirlenmesi düşünüldüğünde, tüm sorulardan elde edilen değişkenlerin kullanılması yerine sadece literatürde belirtilmiş tür seçimini etkileyen faktörler dikkate alınmıştır. Dolayısıyla veri seti içerisinde Tablo 2’de gösterildiği gibi sadece konut bilgisi, hanehalkı özellikleri ve yolculuk bilgisiyle ilgili sorular dikkate alınmıştır.

Ele alınan değişkenler içerisinde tüm yolculuk bilgileri dikkate alındığında veri seti içerisinde boşluklar olduğu ortaya çıkmıştır. Yolculuğunda bir ulaşım türü seçen her bir bireyin konut bilgisi ve hanehalkı özellikleri dikkate alınarak tür seçimini etkileyen faktörler belirlenmek

istendiği için, Tablo 3 üzerinde gri renkte gösterilen bu boşlukların doldurulması gerekmektedir. Her bir satırın bir bireyi göstermesinden dolayı yeni Yolculuk Sıra No sütunu eklenmesi ihtiyacı doğmuştur. Bu veri düzenleme işlemi Microsoft Excel üzerinde yapılmıştır.

Yolculuk Sıra No	Konut Bilgisi		Hanehalkı Özellikleri		Yolculuk Bilgisi	
	Anket No	X _{1...m}	Kişi No	Y _{1...m}	Seyahat No	Z _{1...k}
1	1	---	1	---	1	---
2	2	---	1	---	1	---
3	2	---	2	---	1	---
4	3	---	1	---	1	---
5	3	---	2	---	1	---
6	3	---	2	---	2	---
7	3	---	2	---	3	---
8	3	---	3	---	1	---
9	3	---	3	---	2	---

Tablo 3. Düzenlenmiş Veri Seti

İzmir ili tüm merkez ilçelerine ait bu veri seti içerisinde ilk aşamada Buca ilçesi üzerinden tür seçimini etkileyen faktörler belirlenmeye çalışılmıştır. Bu amaçla adres bilgisi Buca ilçesi olarak tanımlanmış yolculuk bilgileri sadece dikkate alınmıştır. Buca ilçesindeki 2579 hane içerisinde 14135 kişiye ait yolculuk bilgileri dikkate alınmıştır. Çalışmanın bir sonraki aşamasında ise İzmir merkez ilçelerinin hepsi dikkate alınarak yöntem uygulanacaktır.

Yöntem

Veri seti düzenlendikten sonra kullanılacak veri madenciliği yönteminin belirlenmesi aşamasına gelinmiştir. Bu amaçla öncelikle literatürdeki yöntemler incelenmiştir (Han & Kamber, 2006; Hensher, Et Al. ,2004; Lu & Kawamura, 2010; Ozkan, 2008). Bu çalışmada kişilerin ulaşım tür seçiminde hangi ortak özelliklere sahip olduğu ortaya konmak istendiğinden, veri madenciliği yöntemlerinden *Sınıflandırma* yöntemi seçilmiştir. Sınıflandırma yöntemleri içerisinde makine öğrenmesi olarak da tanımlanan *Karar Ağacı* yöntemi seçilmiştir. Karar ağacı ile karar noktalarındaki farklı değerlere göre tür seçiminin nasıl ayrıştığı ortaya konacaktır. Sınıfları belli olmayan ve ikiden fazla bölünmeyle karar ağaçları oluşturulacağı için karar ağacı yöntemlerinden *Entropi Tabanlı Algoritmalar* seçilmiştir. Veri seti içerisinde hem kategorik hem de sayısal nitelikli veriler olduğu için entropi tabanlı algoritmalar içerisinde *C4.5 algoritması* seçilmiştir.



Şekil 3. Veri Madenciliği Yöntemi Seçim Şeması

Çalışmada kullanılacak veri madenciliği yöntemi belirlendikten sonra modellemenin yapılabilmesi için kullanılacak yazılım belirlenmelidir. Bu çalışma öncesi yapılan araştırmada birçok veri madenciliği yazılımına ulaşılmıştır. Bu yazılımlardan açık kaynak kodlu RapidMiner yazılımı tercih edilmiştir.



Şekil 4. Veri Madenciliği Yazılımları

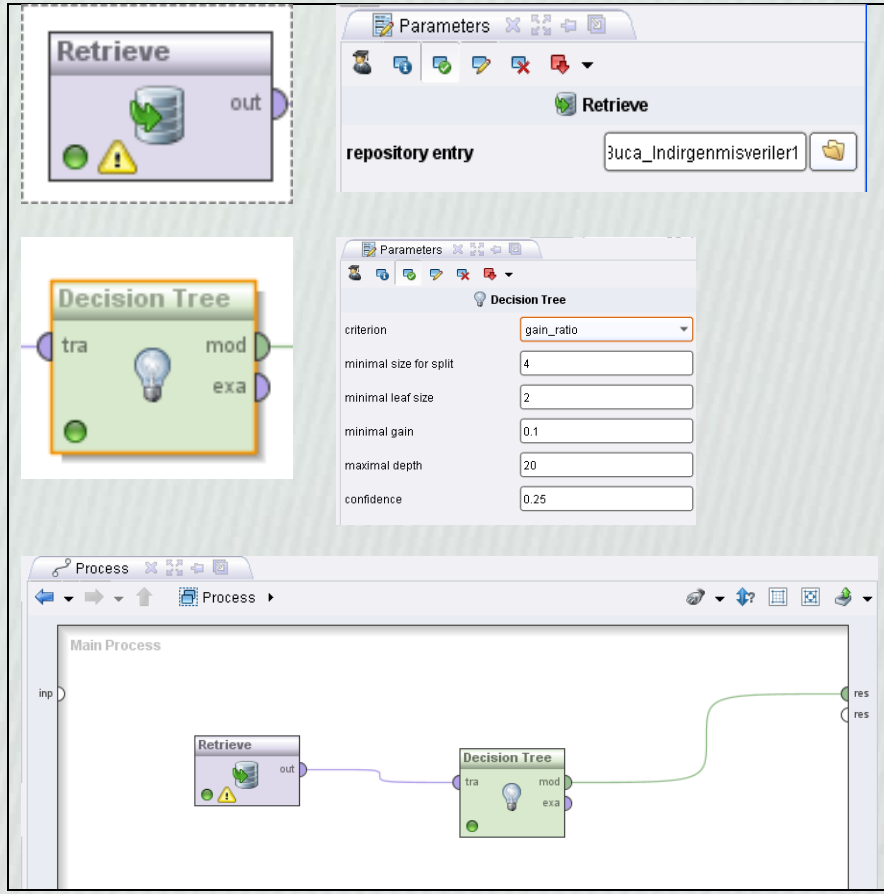
Rapidminer programında, *Retrieve* ve *Decision Tree* operatörleri kullanılarak sonuçlar elde edilmiştir. Veri seti program içerisinde aktarılırken, veri setindeki verilerin türleri ve karar ağacında belirlenmesi gereken hedef değer tanımlanmıştır (Tablo 4).



Değişken adı	Veri türü	Değerler
Konut türü	binomial	müstakil, apartman...
Mülkiyet türü	binomial	kend, ,kira,lojman...
İkinci ev sahipliği	binomial	var,yok
Kişi Sayısı	integer	1,2,...n
Araç Sayısı	integer	1,2,...n
Özel Araç sayısı	integer	1,2,...n
Hane reisine yakınlık	polynomial	eşi,çocuk...
Yaş grubu	polynomial	0-6,7-12,...,55 ve üzeri
Cinsiyet	binomial	kadın, erkek
Öğrencilik	binomial	var, yok
Eğitim	polynomial	ilkokul, ortaokul, lise...
Çalışma durumu	polynomial	özel, kamu
Süreklilik	polynomial	çalışan, iş sahibi...
Ehliyet	binomial	var, yok
Kentkart sahipliği	binomial	var, yok
Yolculuk sayısı	polynomial	1,2,...n
Yolculuk nedeni	polynomial	iş, ev, alışveriş, okul....
Başlangıç mahallesi	polynomial	izmirin mahalleleri
Baş. Arazi kullanımı	polynomial	ev, okul, hastane....
Başlangıç zamanı	time	24 saat
Ulaşım aracı*	polynomial	araba, otobüs, tren...
Bitiş mahallesi	polynomial	izmirin mahalleleri
Bitiş arazi kullanımı	polynomial	ev, okul, hastane....
Bitiş zamanı	time	24 saat
Yolculuk süresi	integer	dakikalar

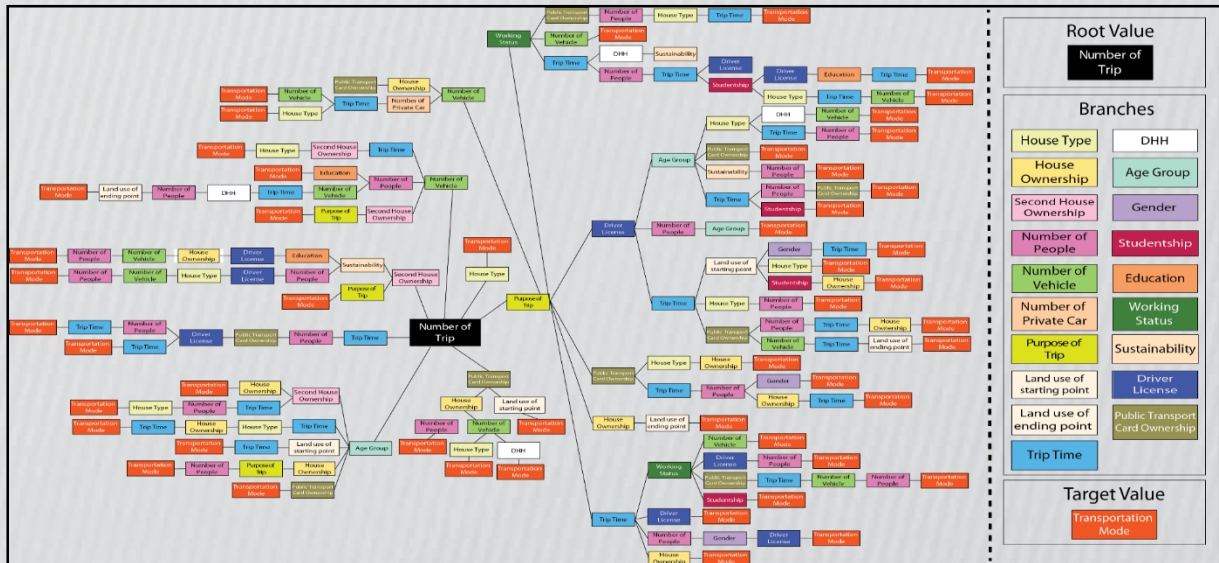
Tablo 4. Rapidminer Değişken Tanımlaması

Retrieve operatöründe veri seti tanımlanmıştır. Daha sonra *Decision Tree* operatöründe kriterler belirlenmiştir (Şekil 5). Sonuçta Buca ilçesine ait hanehalkının verileri dikkate alınarak karar ağacı oluşturulmuştur.



Şekil 5. Rapidminer Programındaki Arayüzler

Oluşturulan karar ağacındaki basitleştirilmiş yapı Şekil 6 da gösterildiği gibidir. Kök değeri “Yolculuk sayısı” olarak oluşmuştur. Buca ilçesi bazında oluşturulan karar ağacından sonra İzmir ili tüm merkez ilçeleri dikkate alınarakta karar ağacı oluşturulmuştur.



Şekil 6. Buca İlçesi İçin Karar Ağacı Yapısı

Sonuç

Ulaşım planlama aşamalarından biri olan ulaşım tür seçimi aşamasında dikkate alınan tür seçimini en çok etkileyen faktörler, hanehalkı anket verisinden faydalanarak belirlenmeye çalışılmıştır. İzmir ili merkez ilçelerinde yapılan ulaşım planı çalışması kapsamında 2007 yılında gerçekleştirilen hanehalkı anketi çalışmasından elde edilen konut bilgisi, hanehalkı bilgisi ve yolculuk bilgisi grupları altındaki değişkenlerden hareketle öncelikle Buca ilçesi için karar ağacı oluşturulmuştur. Bir sonraki aşamada ise İzmir ili tüm merkez ilçeleri için karar ağacı oluşturulmuştur. Bu oluşturulan karar ağacı sonucunda hem Buca ilçesinde hem de İzmir ili merkez ilçeleri için ulaşım tür seçimini en çok etkileyen faktörler ortaya konmuştur.

	Buca	İzmir (Merkez İlçeler)
Anket (hane)	2579	20000
Yolculuk Bilgisi	14135	99258
Faktörler	Ehliyet Durumu Yaş Grupları Kişi Sayısı Seyahat Süresi Hanedeki Kişi Sayısı Çalışma Durumu Mülkiyet Türü Araç Sayısı KentKart Sahipliği Konut Sahipliği Seyahatin Amacı	Yolculuk Sayısı Yolculuk Süresi Yaş Grubu Konut Türü Hane reisine yakınlık Başlangıç Arazi Kullanımı Hanedeki Kişi Sayısı Araç Sayısı Mülkiyet Türü Süreklilik İkinci Evsahipliği

Tablo 5. Buca İlçesi ve İzmir Merkez İlçeleri İçin Tür Seçimini Etkileyen Faktörler

Bu sonuçlara göre Buca ilçesi ve İzmir merkez ilçelerinde hanehalkı anket verisinden hareketle ulaşım tür seçimini en çok etkileyen faktörler Tablo 5’te gösterilmiştir. İleri ki çalışmalarda hanehalkı anket verisinin yanında nüfus vb. veri setleri dikkate alınarak tür seçimini etkileyen faktörlerin belirlenmesinde daha etkili sonuçlara ulaşılabilecektir.



Bu çalışma, L.Özge ORAL'ın Prof. Dr. Vahap TECİM danışmanlığında Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri programında yürütmekte olduğu doktora çalışmasının bir kısmını oluşturmaktadır.

Kaynakça

Bonnel, P. 2003. The Estimation Of Aggregate Modal Split Models. European Transport Conference. Strasbourg, France.

Chaudhury, P.D. 2005. Modal Split Between Rail And Road Modes Of Transport In India. Vikalpa: The Journal Of Decision Makers 30(1):17-33.

Cios, K.J. Et Al. 2007. Data Mining-A Knowledge Discovery Approach Springer.

Haluzova, P. 2008. Effective Data Mining For A Transportation Information Systems. Acta: Polytechnica 48(1): 24-29.

Han, J. & Kamber, M. 2006. Data Mining Concepts And Techniques. San Francisco: Morgan Kaufmann
Hensher, D.A Et Al. 2004. Handbooks In Transport Volume 5: Handbook Of Transport Geography And Spatial Systems. Oxford: Elsevier.

Izmir Büyükşehir Belediyesi. 2009. İzmir Ulaşım Master Planı. Ulaşım Daire Başkanlığı, Ulaşım Koordinasyon Şube Müdürlüğü, İzmir.

Kurgan, L.A & Musilek, P. 2006. A Survey Of Knowledge Discovery And Data Mining Process Models. Knowledge Engineering Review. 21(1):1-24.

Liang, Z. & Ling-Xiang, Z. 2012. Public Transportation Dynamic Guidance Model And Algorithm. Advances In Civil Engineering And Architecture Innovation, Pts 1-6 Book Series: 368-373:3113-3116.

Lu, Y. & Kawamura, K. 2010. Data-Mining Approach To Work Trip Mode Choice Analysis In Chicago, Illinois, Area, Transportation Research Record 2156: 73-80.

Murty, A. & Ashtakala, B. 1987. Modal Split Analysis Using Logit models. Journal Of Transportation Engineering-ASCE. 113(5): 502-519.

Nassar, K. 2007. Application Of Data-Mining To State Transportation Agencies' Projects Databases, Itcon. 12:139-149.

Ortuzar, J. D. & Willumsen, L.G. 2006. Modelling Transport. Chichester: John & Wiley.



Ozkan, Y. 2008. Veri Madenciliği Yöntemleri, İstanbul:Papatya.

Phyu, T. N. 2009. Survey Of Classification Techniques İn Data Mining. International Multi-Conference Of Engineers And Computer Scientists, Vols I And II Book Series: Lecture Notes İn Engineering And Computer Science:727-731.

Tye, W.B. Et Al. 1982. Application Of Disaggregate Travel Demand Models. Transportation Research Board. Washington D.C: Transportation Research Board.

Xie, C. Et Al. 2003. Work Travel Mode Choice Modeling Using Data Mining: Decision Trees And Neural Networks. Traveler Behavior And Values 2003: Planning And Administration, Transportation Research Record. 1854: 50-61

Yao, X. 2007. Where Are Public Transit Needed-Examining Potential Demand For Public Transit For Commuting Trips. Computers, Environment And Urban Systems. 31: 535-550.

Oral,L.Ö. & Tecim,V. , (2013), Using Decision Trees For Estimating Mode Choice Of Trips In Buca-Izmir, International Archives Of The Photogrammetry, Remote Sensing And Spatial Information Sciences,Volume XI-4/W1, 29th Urban Data Management Symposium, 29 – 31 May, 2013, London, United Kingdom. 139:145.