Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği Araçları

Adem Tekerek
Gazi Üniversitesi, Elektronik-Bilgisayar Eğitimi Bölümü, Ankara
atekerek@gazi.edu.tr

Özet: Veri tabanları veya dosyalarda bulunan verilerin belirli istatistik yönetmeleri 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. Veri Madenciliği işlemlerini gerçekleştirmek için ticari ve açık kaynak olmak üzere birçok araç bulunmaktadır. Bu araştırmada açık kaynak kodlu Veri Madenciliği programlarından olan RapidMiner, WEKA, R, Orange, R, KNIME ve Tanagra anlatılmış ve bu programların özelliklerine göre karşılaştırılmaları yapılmıştır.

Anahtar Kelimler: Veri Madenciliği, Açık Kaynak Veri Madenciliği Araçları

Data Mining Processes and Open Source Data Mining Tools

Abstract: Data mining is the process that managing data in databases or files by using particular statistical methods. Data mining can be expressed as whole of techniques of obtaining information from the data. it is the prosess that disclosuring of confidential information in the data and coverting data to usable information by using statistical analysis techniques, genetic algorithms and artificial intelligence methods. There are many commercial and open source tools to perform data mining operations. In this study, the RapidMiner, Weka, R, Orange, R, and Tanagra and KNIME open source data mining programs are explained and compared according to the characteristics of these programs.

Keywords: Data Mining, Open Source Data Mining Tools

1. Giriş

Veri Madenciliği, büyük ölçekli veriler arasından bilgiye ulaşma, veriyi madenleme işlemidir. Veri tabanlarındaki, veri ambarlarındaki veya dosyalarda bulunan veriler arasında bulunan ilişkiler, örüntüler, sapma ve eğilimler, belirli yapılar gibi bilgilerin ortaya çıkarılması ve keşfi veri madenciliğinin temelini oluşturur. "Veri Tabanlarından Bilgi Keşfi" (Knowledge Discovery in Databases) uygulamaları ile birlikte faaliyet alanına yönelik karar destek mekanizmaları için gerekli ön bilgileri temin etmek için kullanılır. Veri madenciliğinin amacı, toplanmış verilerin bir takım istatistiksel yöntemlerle incelenip ilgili kurum ve yönetim destek dizgelerinde kullanılmak üzere değerlendirilmesidir [1]. 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 [2]. Büyük veritabanlarından gizli kalmış örüntüleri çıkarma sürecine veri madenciliği adı verilmektedir. Geleneksel yöntemler kullanılarak çözülmesi çok zaman olan problemlere veri madenciliği süreci kullanılarak daha hızlı bir şekilde çözüm bulunabilir [3]. Veri madenciliğinin temel amacı elimizde bulunan veriden gizli kalmış örüntüleri çıkarmak, verinin değerini arttırmak ve veriyi bilgiye dönüştürmektir [4].

Günümüzde veri madenciliği; bankacılık, pazarlama, sigortacılık, telekomünikasyon, borsa, sağlık, endüstri, bilim ve mühendislik gibi birçok dalda uygulama alanı bulunmaktadır. Bu alanlardaki uygulamalar aşağıdaki gibi örnekler verilebilir.

- **Bankacılık:** Risk analizleri ve usulsüzlük tespiti.
- **Pazarlama:** Çapraz satış analizleri, müşteri segmentasyonu.
- **Sigortacılık:** Müşteri kaybı sebeplerinin belirlenmesi, usulsüzlüklerin önlenmesi.
- **Telekomünikasyon:** Hile tespiti, hatların yoğunluk tahminleri.
- **Borsa:** Hisse senedi fiyat tahmini, genel piyasa analizleri.
- **Tıp:** Tıbbi teşhis, uygun tedavi sürecinin belirlenmesi.
- **Bilim ve Mühendislik:** Amprik veriler üzerinde modeller kurularak bilimsel ve teknik problemlerin çözümlenmesi.
- **Endüstri:** Kalite kontrol, lojistik.

Veri Madenciliği bir süreçtir. Bu süreçte ana unsur süreci gerçekleştiren uygulamacıdır. Süreçte bulunan adımlar doğru olarak yerine getirilmediği sürece istenilen sonuca ulaşılması mümkün değildir. Veri madenciliği bilgi keşfi işlemidir, bu bilgi keşfi adımları aşağıdaki gibi sıralanabilir.

1. Veri Temizleme (gürültülü ve tutarsız verileri çıkarmak)
2. Veri Bütünleştirme (birçok veri kaynağını birleştirebilmek)
3. Veri Seçme (Yapılacak olan analiz ile ilgili olan verileri belirlemek)
4. Veri Dönüşümü (Verinin Veri Madenciliği tekniğinden kullanılabilecek hale dönüşümünü gerçekleştirmek)
5. Veri Madenciliği (Veri örüntülerini yakalayabilmek için akıllı metotları uygulamak)
6. Örüntü Değerlendirme (Bazı ölçümlere göre elde edilmiş bilgiyi temsil eden ilginç örüntüleri tanımlamak)
7. Bilgi Sunumu (Madenciliği yapılmış olan elde edilmiş bilginin kullanıcıya sunumunu gerçekleştirmek), [5, 6].

2. Veri Madenciliği Süreçleri

2.1. Veri Toplama

Veri madenciliğinin ilk aşaması veri toplamadır. Veriler birçok farklı ortamda depolanmaktadır. Örneğin; Microsoft'da veriler yüzlerce OLTP veritabanında ve 70'in üzerinde veri ambarında saklanmaktadır. Burada ilk adım veri tabanlarından veya veri ambarlarından yapılacak uygulama için uygun verileri çekmektir Veri toplama işlemi tamamlandıktan sonra, veriler test ve analiz veri seti olarak iki gruba ayrılır. Genellikle yapılan uygulamalarda verilerin %80'i analiz %20'si ise test verisi olarak ayrılır. [7].

2.2 Veri Temizleme ve Dönüştürme

Veri dönüşümünün amacı ise, kaynak veriyi farklı formatlara veya değerlere dönüştürmektir [7]. Örneğin; Veritabanındaki mantıksal (boolean) bir alan integer bir tipe dönüştürülebilir. Bunun sebebi ise kullanılan bazı veri madenciliği algoritmalarının integer veri tipiyle Boolean veri tipine göre daha başarılı sonuçlar üretmesidir. Veri temizleme işleminin amacı, veriler içindeki uygun olmayan veya hatalı girilmiş verileri ayıklamaktır [7]. Bu işlemde eksik veriler uygun değerler ile doldurulur. Eğer eksik veri çok ise bu kaydın silinmesi gerekir.

2.3. Model Kurma

Model kurma veri madenciliğinin çekirdeğidir. Modeli doğru bir şekilde kurabilmek için yapılacak projenin amacı çok iyi bir şekilde kavranmış olmalıdır. Her amaç ile ilgili birden

fazla algoritma mevcuttur. Bu durumda eldeki veriler üzerinde uygun algoritmaların hepsi çalıştırılır ve en doğru sonucu veren algoritma kullanılır.

2.4. Model Değerlendirme

Eldeki veriler üzerinde uygun algoritmalar çalıştırıldıktan sonra en doğru sonucu hangisinin verdiğini bulmak için çeşitli yöntemler mevcuttur. Örneğin, tahmine yönelik sayısal veriler varsa ve kullanılan modelin doğruluğu test edilmek isteniyorsa MAPE (Mean Absolute Percentage Error) yöntemini kullanabilir.

2.5. Raporlama

Raporlama veri madenciliği bulgularını göstermek için önemli bir dağıtım kanalıdır. Birçok veri madenciliği aracı elde edilen modelden kullanıcıların daha önceden tanımladığı raporları göstermek için gerekli araçlara sahiptir.

2.6. Değerlendirme (Scoring)

Veri madenciliği projesinde, örüntüleri bulmak çalışmanın yarısını oluşturur. Esas amaç, değerlendirme için modeli kullanmaktır. Değerlendirme veri madenciliği terminolojisinde scoring olarak da adlandırılır. Değerlendirme yapabilmek için eğitilen model ve yeni durumları içeren veri setinin olması gerekir. Böylece, eğitilen model kullanılarak yeni durumlar için tahminde bulunulabilir.

2.7. Uygulama Entegrasyonu

Bu aşamada kurulan veri madenciliği modeli gerçek zamanlı olarak çalıştırmak üzere geliştirilen uygulama içerisine gömülür.

2.8. Model Yönetimi

Her bir veri madenciliği modeli bir yaşam döngüsüne sahiptir. Bazı uygulamalarda işler, özellikler durağandır ve modelin yeniden eğitilmesine gerek yoktur. Fakat birçok iş özellikleri sık sık değişir. Yeni veriler geldikçe modelin yeniden eğitilmesine gerek vardır. Yani bir model kurulduktan sonra eğer çok sık olarak veri setinde değişiklik yapılıyorsa model sık sık güncellenmelidir [7].

3. Açık Kaynak Kodlu Veri Madenciliği Programları

3.1. 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. [8]. Diğer programlar gibi birkaç tane format desteklememesi YALE'nin artılarındandır.

RapidMiner ve eklentileri Veri Madenciliği'nin tüm yönleri için 400 den fazla operatör sunar. Meta operatörler deneysel tasarımları otomatik olarak optimize eder ve kullanıcıların tekil adımları ya da parametreleri ayarlamaları gerekmez [9].

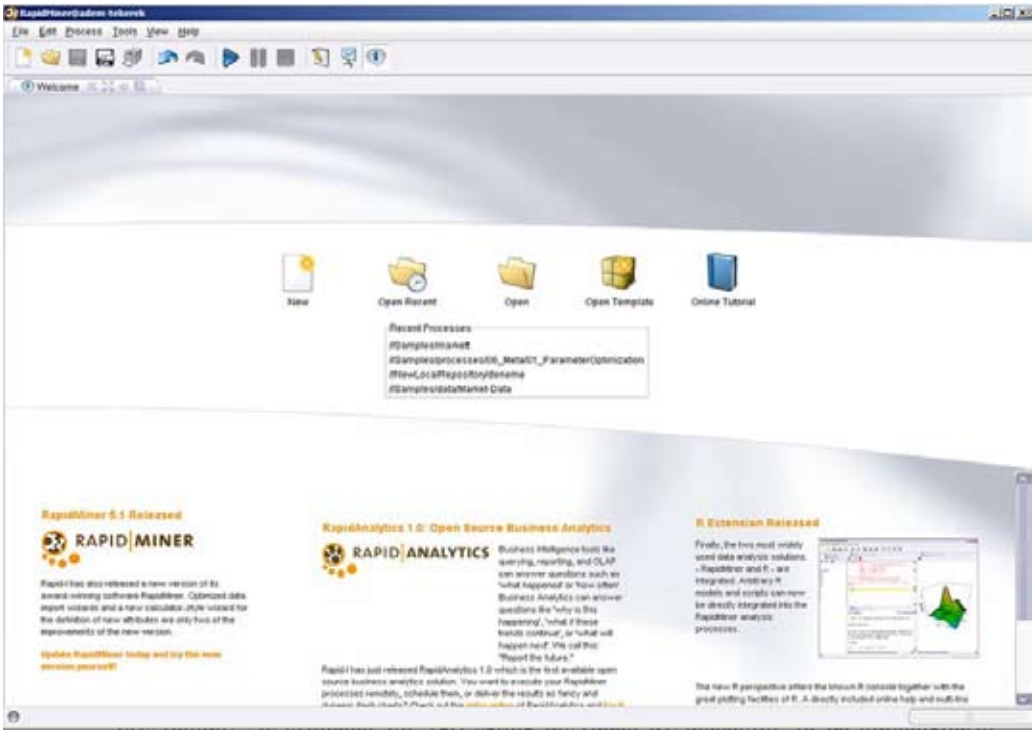
Makine öğrenme algoritmaları olarak destek vektör makinelerini içeren büyük sayıdaki öğrenme modelleri için sınıflandırma ve regresyon, Karar Ağaçları, Bayesian, Mantıksal Kümeler, İlişkilendirme Kuralları ve Kümeleme için birçok algoritma (k-means, k-medoids, dbscan), WEKA'da olan her şey, veri önişleme 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 [2].

YALE'de veri kümesi XML olarak ifade edilir. Aşağıda örnek veri kümesi verilmiştir.

```
<attributeset default source="golf.
dat"> <attribute name ="Outlook"
sourcecol ="1" valuetype - 'nominal"
blocktype
="single value" classes ="rain
overcast sunny"/>
<attribute name ="Temperature"
sourcecol ="2" valuetype ="integer"
```

```
blocktype =”single value”/>
<attribute name - ‘Humidity”
sourcecol =”3” valuetype =”integer”
blocktype =”single value”/>
<attribute name =”Wind” sourcecol
=”4” valuetype =”nominal” blocktype
=”single value” classes =”true false
“/> <labelname =”Play” sourcecol
=”5” valuetype =”nominal” blocktype
=”single value” classes =”yes no”/>
</attributeset>
```

İçerisinde yüzlerce özellik barındırdığı gibi kullanıcıya yakınlığı açısından da diğer programlardan oldukça üstündür. YALE ilk çalıştırıldığında, New diyerek yeni bir uygulama oluşturulabilir, Open diyerek varolan uygulamalar açılabilir. Program bünyesinde her bir algoritma için örnek bulunmaktadır. Şekil 1’de YALE’de açılış ekranı örneği verilmiştir.



Şekil 1. RapidMiner (Yale) Açılış Ekranı

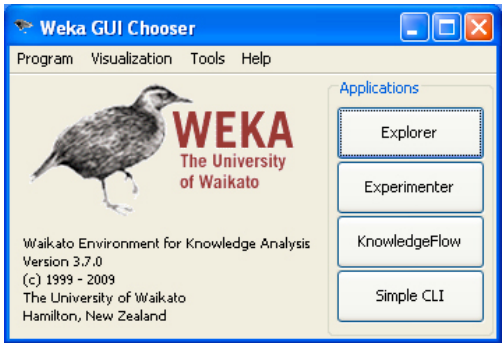
3.2. WEKA

Weka veri madenciliği için makine öğrenmesi algoritmalarının olduğu bir programdır. Algoritmalar bir veri setine doğrudan uygulanabilir ya da kullanıcıların kendi Java kodu içerisinden çağırılabilir. Weka veri işleme, sınıflandırma, regresyon, kümeleme, ilişki kuralları ve görüntüleme araçları içerir. Ayrıca yeni makine öğrenmesi şemaları geliştirmek için uygun yapıdadır [10]. Waikato Üniveristesi tarafından java paltformu üzerinde açık kaynak kodlu olarak geliştirilen ve devamlı güncellenen WEKA’dır. 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 dönüştürebi-

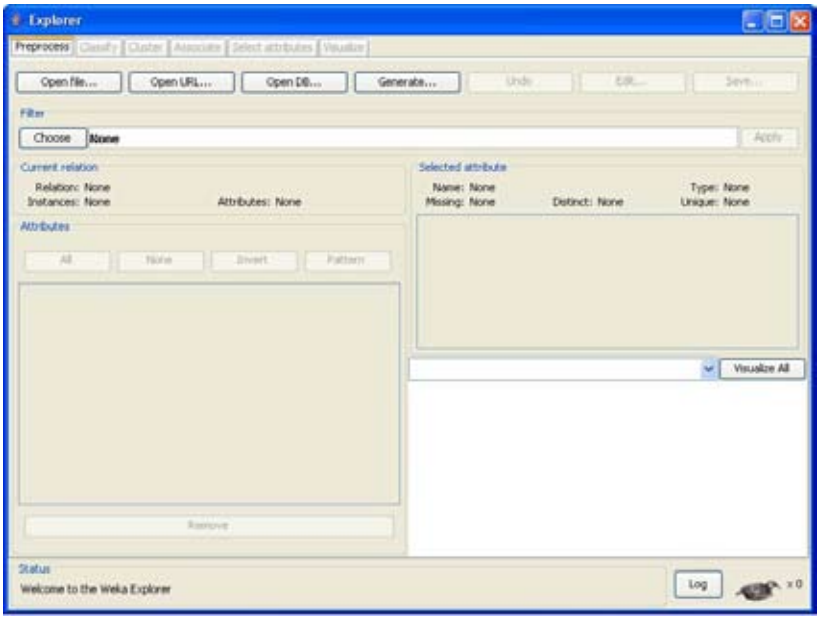
len ayrı bir yazılım vardır [11]. WEKA veri sınıflandırma, dönüştürme, kümelenmesi, ilişki kuralı oluşturma ve görüntüleme özelliklerine sahiptir. Java dilinde geliştirilmiş ve GNU genel kamu lisansı altında bildirilmiş açık kaynak kodlu veri madenciliği aracıdır [12]. WEKA aşağıdaki özelliklere sahiptir. :[13]

- Veritabanındaki analiz ve ön- işleme özelliklerinin ve verinin doğruluğunu değerlendirme.
- Örnek setlerin uygun sınıflara bölünüp sınıf niteliklerinin tanımlaması
- 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 2. WEKA'da Applications Menüsü

WEKA çalıştırıldıktan sonra Şekil 2'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 sürükle bırak yöntemiyle gerçekleştirmeyi sağlayan KnowledgeFlow seçenekleridir. Explorer seçeneği seçildikten sonra üzerinde çalışılacak verilerin seçilmesi, bu veriler üzerinde temizleme ve dönüştürme işlemlerinin gerçekleştirilebilmesini sağlayan Şekil 3'deki ekran ile karşılaşılmaktadır.



Şekil 3. WEKA'da Veri Seçimi

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. WEKA'nın içerisinde Veri İşleme, Veri Sınıflandırma, Veri Kümeleme, Veri İlişkilendirme özellikleri mevcuttur.

Bu adımdan sonra yapılacak olan projenin amacına göre açılan sayfadaki uygun tabdaki (Sınıflandırma, Kümeleme, İlişkilendirme) uygun algoritma veya algoritmalar seçilerek veriler üzerine uygulanmakta ve en doğru sonucu veren algoritma seçilebilmektedir [2].

3.3. Orange

Orange kullanıcıya veri hazırlama, keşifsel veri analizi, modelleme gibi imkânlar sağlayan bir açık kaynak veri madenciliği programıdır. Programda veri madenciliği işi görsel programlama ya da Python scripting ile yapılabilir. Makine öğrenmesi için bileşenleri vardır. Bioinformatik ve metin madenciliği için eklentileri vardır. Veri analiz özellikleri ile paketlenmiştir. [14].

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 önişleme, özellik skorlama ve filtreleme, modelleme, model değerlendirme ve keşif teknikleri gibi geniş kapsamlı bileşen seti içerir. C++ (hız) ve Python (esneklik) 'a uygulanmıştır. Grafik kullanıcı arayüzü çapraz-platform üzerine inşa eder. 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.

- 1996, Ljubljana Üniversitesi ve Jozef Stefan Enstitüsü ML geliştirmeye başladı, C++ ile bir makine öğrenmesi framework.
- 1997, ML için Python bağlayıcılar geliştirildi, gelişmekte olan Python modülleri ile birlikte Orange denen framework birleştirildi
- Geçen yıllarda C++ yada Python modülleri ile en önemli veri madenciliği ve makine öğrenmesi algoritmaları geliştirildi.
- 2002, Pmw Python megawidgets kullanılarak esnek grafik arayüzü için ilk prototipler oluşturuldu.
- 2003, Qt framework için PyQt Python bindings kullanılarak grafik kullanıcı arayüzü tekrar dizayn edildi ve geliştirildi.
- Görsel programlama çatısı tanımlandı, veri analiz hattı için grafik bileşenlerinin geliştirilmesine başlandı.
- 2005, bioinformatik için data analizi eklentileri oluşturuldu.
- 2008, Mac OS X DMG ve Fink- tabanlı kurulum paketleri geliştirildi.
- 2009 dan itibaren, Orange 2.0 beta sürümünde ve web sitesi günlük derleme döngüsü ile kurulum paketleri sunuyor.

3.4. Konstanz Information Miner (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 üniversitede aynı zamanda öğretim ve araştırma için de kullanılmaktadır. Üniversitede geliştirilen bir çok veri analiz yöntemi programa entegre edilmiştir.

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. Wekanın veri madenciliğinde kullanılan en iyi bilinen bütün analiz modülleri ve R-scripts'in eklentilerinin çalıştırılmasına, istatistik kütüphanelerinin kullanılmasını sağlar.

KNIME modüler API özelliği ile Eclipse platformunu temel alır ve kolayca genişleyebilir. bu modülerlik ve genişleyebilirlik özelliği KNIME'nin öğretim ve araştırma özelliklerinin yanında ticari olarakta kullanılmasını sağlar. KNIME çok fazla fonksiyonelliklere sahiptir.

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

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

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

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

3.5. R

R programı grafikler, istatistiksel hesaplamalar, veri analizleri için geliştirilmiş bir programdır. S diline benzer bir GNU (GNU Genel kamu lisansı) projesidir. Yeni Zelanda'da bulunan Auckland Üniversitesi İstatistik bölümü tarafından geliştirilmiştir. R & R olarak ta bilinir. R, farklı uygulamalar ile S diline üstünlük sağlamaktadır. Lineer ve lineer olmayan modelleme, klasik istatistiksel testler, zaman serileri analizi, sınıflandırma, kümeleme gibi özellikleri bünyesinde bulundurmaktadır. R, Windows, MacOS X ve Linux sistemleri üzerinde çalışabilmektedir [15].

R yaygın olarak pencereli sistemlerde kullanılır. R'nin X Window sistemi üzerinde kullanılması tavsiye edilmektedir. Açık sistemlerin kullanıcıya sunduğu en büyük özelliklerinden biri olan X Window, Linux'un doğduğu andan itibaren destek görmeye başlamıştır. İnternet üzerinde bedava dağıtılmasıyla Linux dağıtımı altında bir standart olarak kendine yer edinmiştir. X Window, istemci sunucu modeline göre çalışır. Ana makina üzerinde çalışan X sunucusu, grafik donanımı üzerindeki tüm giriş-çıkış yetkilere sahiptir. Bir X istemcisi, sunucuya bağlanarak istediği işlemleri sunucuya yaptırır. İstemcinin görevi emir vermek, sunucunun ise verilen emri görünür hale getirmektir [16]. Windows veya MacOS üzerinde R'yi çalıştırmak için uzman yardımına ihtiyaç vardır. Kullanıcılar, R'yi çoğunlukla Unix makineler üzerinde çalıştırırlar. R'yi Unix makinelerde çalıştırabilmek için aşağıdaki adımlar izlenir. Problemi çözümü için gereken veri dosyaları barındırmak için dizin oluşturulur.

```
$ mkdir work
$ cd work
```

R programının çalıştırılması için aşağıdaki komut yazılır.

```
$ R
```

R programından çıkmak için aşağıdaki komut yazılır.

```
q()
```

Fonksiyonların özelliklerini öğrenmek için aşağıdaki komutlar yazılabilir.

```
help(solve)
?solve
```

Verilerin işleniş şekli de aşağıda verilmektedir.

```
incomes <- c(60, 49, 40, 61, 64, 60,
59, 54, 62, 69, 70, 42, 56, 61, 61,
61, 58, 51, 48, 65, 49, 49, 41, 48,
52, 46, 59, 46, 58, 43)
```

3.6. Tanagra

Tanagra akademik ve araştırmalar için kullanıma sunulan ücretsiz veri madenciliği programıdır. Veri analizi, istatistiksel ve makine öğrenme gibi bir çok veri madenciliği metotlarını sağlar. Bu proje, çeşitli denetlenmiş öğrenme algoritmaları, özellikle interaktif ve görsel karar ağaçlarının yapımında uygulanan SIPINA'nın ardından geliştirilmiştir. Tanagra, SIPINA'dan daha güçlüdür. Tanagra, kümeleme, faktöriyel analiz, parametrik ve nanparametrik istatistik, birliktelik kuralı, özellik seçimi ve yapı algoritması gibi bazı denetlenmiş öğrenme, ayrıca diğer paradigmaları içerir. [17].

Tanagra her araştırmacının kaynak koduna erişebildiği ve yazılım dağıtım lisansını onayladığı sürece kendi algoritmalarını ekleyebildiği bir açık kaynak yazılım projesidir. Tanagra projesinin esas amacı araştırmacılara ve öğrencilere kullanımı kolay veri madenciliği yazılımı sunmak, yazılım gelişiminde güncel standartları sağlamak ve reel veya yapay veri ile analiz yapmaya izin vermektir. [17]. Tanagra'nın diğer bir amacı araştırmacılara kendi veri madenciliği metotlarını ekleyebildikleri, performanslarını karşılaştırabildikleri bir mimari sunmaktır. Tanagra kullanıcılara işlerinin gerekli kısmını yapmalarına imkan sağlayan, bu tür programlardaki istenmeyen kısım olan veri yönetimini muaf tutan daha çok deneysel platform olarak çalışır. Üçüncü ve son amacı, acemi geliştiricilerin yönlendirilmesinde, bu tür bir yazılımın yapımında mümkün bir metodoloji yayılmasını meydana getirmektir. Geliştiriciler kaynak koda ücretsiz erişimin, yazılımın adımlarının nasıl oluşturulduğunun, kaçınılması gereken problemlerden, projenin ana adımlarından, hangi araçların ve kod kütüphanelerinin kullanılabildiğinden faydalanırlar. Tanagra öğrenme algoritması teknikleri için pedagojik bir araç olarak düşünülebilir. Tanagra ticari yazılımların sunduğu, geniş bir set veri kaynağı, veri ambarı ve veritabanı, veri temizleme, interaktif kullanım özellikleri gibi özelliklerin tamamına sahip değildir.

4. Programların Karşılaştırılması

YALE, WEKA ve R dâhil olmak üzere açık kaynak kodlu Veri Madenciliği programları arasında liderdir. Hem kullanım kolaylığı hem de içerisinde yüzlerce özelliği barındırması YALE'yi WEKA'dan üstün kılmaktadır. YALE'de 3D görsellerin fazlalığı kullanıcıya oldukça yardımcı olmaktadır. WEKA'nın kullanımı da kolaydır fakat desteklediği algoritmaların sayısı YALE'ye göre daha azdır. YALE 22'ye yakın dosya formatını desteklerken, WEKA'nın desteklediği dosya formatı sayısı 4 ile sınırlıdır. Ancak çoğu Veri Madenciliği uygulamasını geliştirmede WEKA yeterli olmaktadır. R ise hem kullanım kolaylığı hem de desteklediği algoritmalar ile YALE ve WEKA'nın altında bulunmaktadır. R, Unix makinelerde yaygın olarak kullanılmaktadır. R'yi Windows sistemi üzerinde kullanabilmek uzman yardımı istemektedir. Bundan dolayı R, YALE ve WEKA'ya göre fazla tercih edilmemektedir.

2010 yılında yapılan anket sonucunda Şekil 4'de verilen bilgiler elde edilmiştir. Bu anketin yapıldığı web sitesi veri madenciliği uzmanlarının ziyaret ettiği bir sitedir. Araştırmada gerçek projelerde kullanılan veri madenciliği programlarının kullanım oranları sunulmuştur. Şekil 4'te açık kaynak kodlu programların kullanım ağırlıkları verilmiştir.

Son 12 ayda gerçek projelerde kullanılan veri madenciliği veya veri analizi yazılımları arasındaki bazı açık kaynak kodlu yazılımların ağırlığı	
RapidMiner	37.8 %
R	29.8 %
KNIME	19.2 %
Weka	14.3 %
Orange (25)	2.7 %

Şekil 4. Açık Kaynak Kodlu Veri Madenciliği Programları Kullanım Oranı [18].

5. Sonuç

Her geçen gün katlanarak artan veri miktarından dolayı bilgiye ulaşmak zorlaşmıştır, bilgiye ulaşmak için geliştirilen kavram veri madenciliği olarak nitelendirilmektedir. Veri madenciliği

büyük miktardaki veriden kullanılabilir bilgiyi üretmek için kullanılır. Veri Madenciliği uygulamaları yapmak için bilgisayar programlarına ihtiyaç vardır. Bu programlar veri kümeleme, karar ağaçları, bayes sınıflandırıcılar, apriori yöntemi gibi birçok algoritmayı içerir. Algoritmalar sayesinde işlenen verilerden, bilgi çıkarımı yapılabilmektedir. Bu çalışmada açık kaynak kodlu bazı veri madenciliği programları incelenmiştir.

6. Kaynaklar

[1] İnternet : "Veritabanı, Veri Madenciliği, Veri Ambarı, Veri Pazarı", http://mail.baskent.edu.tr/~2039467 6/0302/bil483/HW2.pdf , (2010).

[2] Dener, M., Dörterler, M., Orman, A., "Açık Kaynak Kodlu Veri Madenciliği Programları: Weka'da Örnek Uygulama", Akademik Bilişim'09 - XI. Akademik Bilişim Konferansı Bildirileri, 11-13 Şubat 2009 Harran Üniversitesi, Şanlıurfa.

[3] Hung, S., Yen, D., C., Wang, H., 'Applying data mining to telecom churn management', Expert Systems with Applications, October 2005, pp. 1-10.

[4] İnternet: "Article Detail" http://www.sqlnedir.com/Members/ ArticleDetail.aspx?Id=81 (2010).

[5] Han, J. ve Kamber M., "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2001.

[6] Delen, D., Walker, G., Kadam, A., 'Predicting breast cancer survivability: a comparison of three data mining methods', Artificial Intelligence in Medicine, vol 34, June 2005, pp113-127.

[7] Tang, Z., MacLennan, J. ,"Data Mining with Sql Server 2005", Wiley, 2005.

[8] İnternet: "RAPİDMİNER", http://www.aktueryabilimleri.com/ index.php?option=comcontent&vi ew=category&id=97:rapidminer&It emid=252&layout=default , (2010).

[9] İnternet: "Rapid - I - Operator Overview", http://rapid-i.com/content/view/12/69/ (2010).

[10] İnternet: "Weka 3 - Data Mining with Open Source Machine Learning Software in Java", http://www.cs.waikato.ac.nz/ml/we ka/ , (2010).

[11] İnternet : "Weka Machine Learning", http://en.wikipedia.org/wiki/Weka_ %28machine learning%29 (2010).

[12] Ian H.Witten and Elbe Frank, "Datamining Practical Machine Learning Tools and Techniques," Second Edition, Morgan Kaufmann, San Fransisco, 2005.

[13] B. M. Patil, Durga Toshniwal, R. C. Joshi, "Predicting Burn Patient Survivability Using Decision Tree In WEKA Environment", 2009 IEEE International Advance Computing Conference (IACC 2009), Patiala, India, 6-7 March 2009.

[14] İnternet: " Orange - Data Mining Fruitful Fun", http://www.ailab.si/orange/ (2010).

[15] İnternet : "R-Intro", http://cran.r- proj ect. org/doc/manual s/R- intro.pdf, (2010) Hania Gajewska, Mark S. Manasse and Joel McCormack, Why X Is Not Our Ideal Window System , Software — Practice & Experience vol 20, issue S2 (October 1990).

[17] İnternet: "TANAGRA", http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra. html (2010).

[18] İnternet: "Data Mining / Analytic Tools Used Poll" http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html (2010).