ÜCRETSİZ VERİ MADENCİLİĞİ ARAÇLARI VE TÜRKİYE'DE BİLİNİRLİKLERİ ÜZERİNE BİR ARAŞTIRMA

FREE TOOLS FOR DATA MINING AND A RESEARCH ON THEIR RECOGNITION IN TURKEY

Öğr.Gör.Dr.Onur DOĞAN, Dokuz Eylül Üniversitesi, İzmir Meslek Yüksek Okulu, İktisadi ve İdari Programlar, onur.dogan@deu.edu.tr

Öz

Veri madenciliği, istatistik temelli ve bilgisayar destekli teknikler kullanılarak veriden bilgiye ulaşma süreci olarak tanımlanabilir. Veri analizleri için sıklıkla kullanılan veri madenciliği araçlarının sayısı günden güne artış göstermektedir. Bu çalışmanın amacı ücretsiz veri madenciliği araçlarını derlemek ve tanıtmaktır. Bu amaca uygun olarak, akademik ve ticari araştırmalarda kullanılan ücretsiz veri madenciliği araçlarının önemli bir kısmı belirlenmiştir. Bu yazılımlar hakkında tarihçe bilgisi, kullanım alanı vb. bazı genel bilgiler verilmiştir. Ayrıca, Türkiye'de veri madenciliği konusunda araştırma, proje vb. çalışmalarda bulunan kişilere kartopu örneklem metodu ile ulaşılarak bu veri madenciliği yazılımlarının Türkiye'deki bilinirliği ve kullanım yaygınlıkları belirlenmeye çalışılmıştır. Araştırmacıların, çalışmaya konu olan 38 adet yazılımdan yalnızca 5 tanesi üzerinde yoğunlaştıkları geri kalan yazılımların birçoğunu kullanmadıkları gibi bu yazılımlardan haberdar da olmadıkları görülmüştür. Çalışmanın Türkiye'deki araştırmacılar için veri madenciliği araçları için veri seti oluşturma ve farklı veri madenciliği araçlarını araştırmacılara tanıtma amacına ulaşacağı düşünülmektedir. Veri madenciliği sürecinde girdi sayısı, girdi tipi, kullanılacak veri madenciliği yeteneği gibi vermesi gereken çok sayıda karar vardır. Kullanılacak veri madenciliği aracı da bu kararlardan biridir. Çalışmanın, araştırmacıların çalışmalarında kullandıkları veri madenciliği araçları konusunda tercih şanslarını arttırarak çalışma sürecinin ve sonuçlarının kalitesine katkı yapacağı düşünülmektedir.

Anahtar Kelimeler: *Veri Madenciliği, Ücretsiz Veri Madenciliği Araçları, Yazılım Tercihleri*

Abstract

Data mining can be defined as a process of accessing the knowledge through data by using statistics and computer-based techniques. The numbers of commonly used data mining tools for data analysis are increasing day by day. The aim of this study is to compile data mining tools and introduce them to the users. In this direction, most of the free data mining tools which used in academic and commercial researches have been determined and some general information such as historical background, areas of usage, etc. about the tools, have been given. In addition, a sample which collects people who study on data mining in Turkey have been created by using the snowball sampling method. And awareness and prevalence of usage of this data mining tools has been tried to determine. It has been noticed that, the researchers, frequently use 5 data mining tools within 38 data mining tools. Besides, they are not aware of many of the tools. It has been considered that, the study achieves its goals which are creating a data mining tools set for data scientists in Turkey and introducing different data mining tools to researchers. In the process of data mining, there are many decisions to make such as;

number of inputs, type of inputs, data mining task, etc. One of the decisions to make is which data mining tool will be used. This study will widen researchers' data mining tools preferences and will improve the quality of the mining process and results.

Keywords: *Data Mining, Free Data Mining Tools, Software Preferences*

1. GİRİŞ

Organizasyonlar ve bireyler için topladıkları ve sakladıkları verileri doğru bir biçimde analiz etmek hayati önem taşımaktadır. Veri madenciliği teknikleri, veri analizi konusunda araştırmacıya yardımcı olan başta bilgisayar bilimleri, istatistik gibi farklı alanlardan temellerini alan tekniklerdir. Veri madenciliği ile veri seti içerisinde geleneksel yöntemlerle tespit edilemeyen potansiyel olarak kullanışlı bilgilerin çıkarılması amaçlanır. Bu amaca yönelik olarak veri madenciliği tekniklerinin, sınıflama, kümeleme, örüntü tanıma, değişkenler arası birliktelikleri ortaya çıkarma vb. işlevlerinin yerine getirilmesi için çok sayıda algoritma oluşturulmakta ve bu algoritmalar çeşitli yazılımlar geliştirilerek kullanıcıya sunulmaktadır. Veri madenciliği alanında farklı süreçlere yönelik ücretli, ücretsiz, açık kaynak kodlu olan veya olmayan çok sayıda program geliştirilmiş ve geliştirilmeye devam edilmektedir. Veri madenciliği süreçlerinde, kullanılacak verilerin belirlenmesi, birimlerden toplanan değişkenlerden hangilerinin kullanılacağı, hangi veri madenciliği işlevine yönelik analizler yapılacağı, sonuçlardan hangilerinin karar vericiler için kullanışlı olacağının belirlenmesi gibi problemler söz konusudur. Bu durum veri madenciliği uygulayıcılarını analizler sırasında çok sayıda deneme yapmak durumunda bırakabilir. Bu programların; herhangi bir işlevi daha iyi bir biçimde yerine getirme, daha iyi görselleştirme olanağı sağlama, herhangi bir işletme problemine özgü çözümler sunma, daha hızlı analizler gerçekleştirme, çeşitli dosya formatlarını analiz edebilme gibi çeşitli yönleri ile birbirilerine üstünlükleri vardır. Örneğin bir program sınıflandırma problemine çözümler sunma konusunda özelleşmiş iken bir başka program daha çeşitli ve anlaşılır görselleştirme olanaklarını kullanıcıya sunmaktadır. Neticede, bu kadar farklı sürece ait fazla sayıda probleme cevap verecek biçimde "en iyi yazılım" kavramından bahsetmek mümkün değildir. Bu nedenle program geliştiriciler, genellikle belli ihtiyaçlara, işlevlere, süreçlere cevap verecek biçimde özelleşme yoluna gitmektedirler. Araştırmacıların gerçekleştirmek istedikleri analize uygun programı kullanmaları; zaman, kolaylık, doğruluk gibi açılardan kendilerine yarar sağlayacaktır. Öte yandan, veri madenciliği problemlerine çözümler sunan programların kullanıcıya ücretsiz olarak sunulması doğası gereği kullanıcı için tercih edilir bir durumdur. Özellikle veri madenciliği tekniklerini kar etme amacı gütmeden (örneğin bilimsel amaçlarla) kullanan araştırmacılar için ücretsiz veri madenciliği programlarının daha tercih edilir olacağı açıktır.

Bu çalışma kapsamında ücretsiz veri madenciliği araçları incelenecektir. Tespit edilen ücretsiz veri madenciliği araçları hakkında bilgiler verilecek ve araçlar diğer araştırmacılara belirli hatları ile tanıtılmaya çalışılacaktır. Türkiye'de veri madenciliği araçlarının incelendiği, araçlar hakkında bilgi sunulan ve araçlar arasındaki farklılıklar, üstünlükler vb. hakkında bilgi verilen çalışmalar mevcuttur. Dener, Dörterler&Orman (2009)'da Rapidminer, Weka ve R kullanıcıya tanıtılmış, aynı zamanda Weka'da örnek bir uygulama yapılmıştır. Çalışmada ele alınan açık kaynak kodlu veri Madenciliği programlarının farklılıkları üzerinde durulmuş, Weka'nın en çok kullanılan veri madenciliği programı olduğu tespit edilmiştir. Tekerek (2011) çalışmasında Rapidminer, Weka, Knime, Orange, R ve Tanagra hakkında bilgiler vermiştir. Kaya ve Özel (2004)'te ise Keel, Knime, Orange, R, Rapidminer ve Weka tanıtılmıştır. Araştırmacılar, inceledikleri programları kullanıcı dostluğu, desteklediği dosya formatları, içerdikleri algoritmalar ve makine öğrenmesi paketleri gibi birçok açıdan incelemiş; Weka, Rapidminer ve Keel yazılımlarını en kullanışlı yazılımlar olarak tespit

etmişlerdir. Bu üç program arasından ise öğrenim ve kullanım kolaylığı açısından en başarılı programın Weka olduğunu belirtmişlerdir. Literatürde incelenen çalışmalarda incelenen veri madenciliği programı sayıları kısıtlı sayıdadır. Bu çalışmada farklı hedeflerle veri madenciliği çalışmaları yürüten araştırmacılar için veri madenciliği araçlarının derlenmesi amaçlanmıştır. Bu amaçtan hareketle daha fazla sayıda ücretsiz veri madenciliği aracını genel hatları ile tanıtılmıştır. Çalışmada 38 adet ücretsiz veri madenciliği aracı incelenmiştir. İncelenen ücretsiz veri madenciliği yazılımlarının teknik özellikleri derinlemesine incelenmeyip, programların temel yetenekleri hakkında bilgi verilmesi planlamıştır. Ücretsiz yazılımların sunulduğu internet siteleri ve yazılımların tanıtıldığı bilimsel makaleler incelenmiş ve çalışma içerisinde sunulmuştur. Ayrıca, bu araçların Türkiye'den kullanıcılar arasındaki yaygınlığı, haberdarlık durumları ve kullanım durumları gibi bilgiler hazırlanan soru formu vasıtasıyla, veri madenciliği alanında çalışan araştırmacılardan elde edilmeye çalışılmıştır.

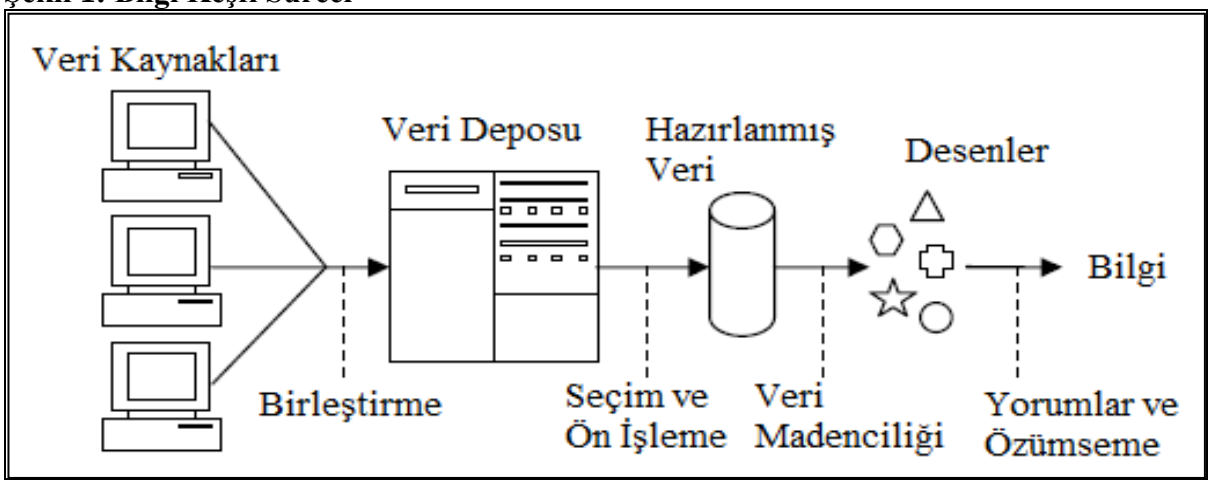
2. VERİ MADENCİLİĞİ

Veri tabanlarından bilgi keşfi (VTBK), verilerden; geçerli, özgün, potansiyel olarak faydalı ve nihayetinde anlaşılabilir yapıların, anlaşılması zor bir süreç ile belirlenmesi olarak tanımlanabilir (Fayyad, Piatetsky-Shapiro ve Smyth 1996: 40-41). Han ve Kamber (2001)'e göre veri tabanlarından bilgi keşfi sürecinin adımları aşağıdaki gibidir;

1. Veri Temizleme
2. Veri Birleştirme
3. Veri Seçimi
4. Veri Dönüşümü
5. Veri Madenciliği
6. Veri Değerlendirme
7. Bilgi Sunumu

VTBK, üstü kapalı, önceden bilinmeyen ve potansiyel olarak kullanışlı bilginin, veriden çıkarılmasıdır. Veri madenciliği ise bilgi keşfi sürecinin yalnızca bir parçası ancak en önemli parçasıdır (Bramer, 2007: 2). Bramer 2007 tarafından yapılan bilgi keşfi süreci modeli Şekil 1'de gösterildiği gibidir. Bilgi keşfinin ilk adımı veri kaynaklarından elde edilen verinin birleştirilmesidir. Veri ardından seçme ve ön işleme tabi tutulur. Hazırlanmış veri, veri madenciliği teknikleri ile analiz edilir. Veri içerisindeki bağlantılar ortaya çıkarılarak, son aşamada yorumlar ve özümseme adımı ile bilgiye dönüştürülür.

Şekil 1: Bilgi Keşfi Süreci



Kaynak: (Bramer, 2007: 2).

Akpınar (2000) de veri madenciliğinin VTBK sürecinin en önemli basamağı olduğunu belirtmiş, bu nedenle VTBK ve veri madenciliği terimlerinin birçok araştırmacı tarafından eş anlamlı olarak kullanıldığına dikkat çekmiştir.
Veri madenciliği kavramının birçok tanımı yapılmıştır. Bunlardan bazıları aşağıda sunulmuştur;

- Veri madenciliği veriden desenler keşfetme süreci olarak tanımlanabilir. Bu süreç, otomatik ya da yarı otomatik olmalıdır. Keşfedilen örüntüler anlamlı olmalıdır ve bazı avantajlar (genellikle ekonomik) getirmelidir (Witten ve Frank, 2005: 5)
- Veri madenciliği; genellikle büyük veri setlerinin, veri sahibi için yararlı ve anlaşılır olacak biçimde, umulmadık ilişkiler yakalamak ve özgün bir biçimde özetlemek için analiz edilmesidir (Hand, Mannila ve Smyth, 2001: 6).
- Veri madenciliği; büyük veri tabanlarından bilgi çıkarımı için kullanılan ve makine öğrenimi, örüntü tanıma, istatistik, veri tabanları, görselleştirme gibi alanlardan teknikleri bir araya getiren disiplinler arası bir alandır (Cabena vd. 1998).

Veri Madenciliği uygulamalarını gerçekleştirmek için programlara ihtiyaç duyulur. Bu kapsamda, SPSS Clementine, Excel, SPSS, SAS, Angoss, KXEN, SQL Server, MATLAB ticari ve RapidMiner (YALE), WEKA, R, C4.5, Orange, KNIME açık kaynak olmak üzere birçok program geliştirilmiştir (Dener, Dörterler ve Orman, 2009: 788). Veri madenciliği ücretsiz yazılımlarından ulaşılabilen 38 tanesi (Tablo 1) izleyen bölümde açıklanmıştır. Ücretsiz veri madenciliği yazılımları alfabetik sırada sunulmuş ve yazılımlar hakkında bilgiler verilmiştir.

Tablo 1: İncelenen Veri Madenciliği Yazılımları

Yazılım İsimleri			
ADaM AdamSoft Alpha Miner Apache mahout CMSR Data Miner Databionic ESOM Tools Data Melt Dlib ELKI Fityk	GGobi GNU Octave Jubatus KNIME Keel LIBSVM LIBLINEAR Lattice Miner Mallet Mining Mart	ML-Flex MDP NLTK OpenNN Orange Pandas Pybrain R Rapid Miner (Yale) Rattle GUI	Rosetta SIPINA Shogun Scikit Learn SenticNet API TANAGRA Vowpal Wabbit Weka

ADaM: The Algorithm Development and Mining System (ADaM) Alabama Üniversitesi Bilgi Teknolojileri ve Sistem Merkezi tarafından veri madenciliği teknolojilerini, uzaktan algılama verilerine ve diğer bilimsel verilere uygulamak için geliştirilmiştir (ADaM, 2016). ADaM veri madenciliği yazılımı sınıflama, kümeleme ve birliktelik kuralları gibi diğer veri madenciliği sistemlerinde yaygın olan metotları kullanıcıya sağlamaktadır. Bunlara ek olarak, bilimsel veri madenciliği yapılırken; özellik indirgeme, görüntü işleme, veri temizleme ve ön işleme yetenekleri ile de değer katmaktadır (Rushing vd. 2011: 607).

AdamSoft: ADaMSoft, CASPUR (Interuniversity Consortium for Supercomputing) merkezinde bir istatistikçi ekibi tarafından geliştirilmiştir. İstatistikçiler tarafından geliştirilmesi nedeni ile hem modern analiz metotları içerir hem de veriyi verimli bir şekilde

işlemek yeteneklerine sahiptir. Yazılım, ücretsiz ve açık kaynak kodlu, veri yönetimi, veri ve web madenciliği, istatistiksel analizler ve daha fazlasını yapabilmektedir (Adamsoft, 2016).

Alpha Miner: AlphaMiner açık kaynak kodlu bir veri madenciliği platformudur (java uygulamalı). Hong Kong Üniversitesi E-İş Teknoloji Enstitüsü tarafından geliştirilmiştir (AlphaMiner, 2016).

Apache Mahout: Apache Mahout ölçeklenebilir öz devinimli öğrenim için açık kaynak kodlu bir kütüphanedir ve Apache Maohut projesi öz devinimli öğrenim uygulamaları için bir çevre oluşturulması amacı da taşır (Mahout, 2016). Mahout ayrıca, kullanılan en eski ve yaygın işbirliğine dayalı filtreleme yapılarını sunmaktadır (Schelter ve Owen, 2012).

CMSR Data Miner: Daha önce StarProbe Data Miner olarak adlandırılan CMSR DaTa Miner Suite kullanıcılar için tahmin modelleme, segmentasyon, veri görselliği, istatistiksel veri analizi ve kural tabanlı modelleme çözümleri sağlamaktadır (CMSR Data Miner, 2016).

Databionic ESOM Tools: Databionic ESOM Tools kümeleme, görselleştirme, kendi kendini örgütleyebilen (özörgütlemeli) haritalar yoluyla sınıflandırma gibi veri madenciliği görevlerini yapabilen bir takım programları içermektedir (Databionic ESOM Tools, 2016).

Data Melt: DataMelt nümerik hesaplama, istatistik, sembolik hesaplama, veri analizi ve veri görselleştirme yeteneklerine sahiptir. 2005 yılında DESY laboratuarlarında jHepWork adı altında başlayan proje, 2013 yılı itibariyle SCaVis ismini, 2015 yılında bugünkü adını almıştır. Projeyi başlatan Dr.Sergei Chekanov dünyanın farklı yerlerinden 100 Java geliştiricisinin katkısının bulunduğu programın herkese açık sürümünü yayınlamıştır (DataMelt, 2016).

Dlib: Dlib-ml özellikle bilim adamları ve mühendisleri hedefleyen C++ dilinde yazılmış bir özdevinimli öğrenme yazılımıdır (Dlib, 2016). King (2009), R, Python, Matlab ve Lua dillerinde geliştirilen birçok özdevinimli öğrenme yazılımı var iken C++ dili üzerinde geliştirilen az sayıda yazılım olduğunu belirtmiş, Dlib-ml yazılımının bu açığı kapatmak ve özdevinimli öğrenme yazılımları için C++ dilini kullanan araştırmacıların bir araya geldiği bir ortam oluşturmak amacı taşıdığını belirtmiştir.

ELKI: ELKI Java dilinde yazılmış bir veri madenciliği yazılımıdır. ELKI algoritmaları özellikle, uç değer (sapan değer) bulma ve kümeleme analizlerin güdümsüz (denetimsiz) metotları üzerine yoğunlaşmaktadır (ELKI, 2016).

Fityk: Fityk, veri işleme ve doğrusal olmayan eğri uydurma (non-linear curve fitting) için kullanılan bir programdır (Fityk, 2016).

GGobi: GGobi büyük boyutlu verilerde çıkarımda bulunmaya yarayan, açık kaynak kodlu bir veri görselleştirme programıdır (Ggobi, 2016). GGobi kökleri XGobi programı olan, çoklu çizim yapabilen, çizelge yönetebilen bir veri görselleştirme programıdır. Ggobi programının en büyük avantajları; kolayca genişletilebilmesi, API kullanılarak kolayca kontrol edilebilmesi ve diğer yazımlara gömülerek çalıştırılabilmesidir (Swayne vd. 2003).

GNU Octave: GNU Octave öncelikle nümerik hesaplamalar üzerine yoğunlaşan yüksek düzeyli bir dildir. Program doğrusal ve doğrusal olmayan problemlere çözüm sağlamaktadır. Bunu yanı sıra GNU Octave kullanıcıya, veri görselleştirme ve manipülasyonu için kapsamlı grafikler sağlamaktadır (GNU Octave).

Jubatus: Jubatus büyük veri kümeleri içerisindeki veri akışı üzerinde çevrimiçi dağıtımlı öz devinimli öğrenme yapabilen ilk açık kaynak kodlu platformdur (Jubatus, 2016).

KNIME: 2004 yılının başlarında Konstanz Üniversitesinde başlayan KNIME projesinin ilk versiyonu 2006 yılında yayınlanmıştır. Başlangıçta ilaç sanayisindeki firmaları hedefleyen ve bu sektörde analizleri hedefleyen proje daha sonra farklı alanlardaki veri analizleri için de kullanılmaya başlanmıştır (KNIME, 2016). 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 (Tekerek, 2011: 166). Kurulum şartı olmadan çalışabilmektedir. Knime yazılımı .txt uzantılı metin dosyalarından veya .arff, .table formatından veri alabilmektedir. Knime, en zengin görselleştirme araçları sunan yazılımlarından biridir (Kaya ve Özel, 2014: 49).

KEEL: KEEL (Knowledge Extraction based on Evolutionary Learning) çok sayıda bilgi keşfi görevi için kullanılabilecek açık kaynak kodlu bir Java yazılımıdır. Yazılım kullanıcılara, sayısal zekâ algoritmaları (özellikle evrimsel algoritmalar) ve farklı veri setleri ile deneyler tasarlanabilecek basit bir kullanıcı ara yüzü sağlamaktadır (KEEL, 2016). Ayrıca, Kaya ve Özel, 2014 KEEL programının veri görselleştirme açısından zayıf olduğunu belirtmişlerdir.

LIBSVM: LIBSVM destek vektör makineleri için 2000 yılından beri geliştirilen bir kütüphanedir. Proje amacı kullanıcıların destek vektör makinelerini kendi uygulamalarında kolaylıkla kullanabilmeleridir (Chang ve Lin, 2011).

LIBLINEAR: LIBLINEAR büyük ölçekli doğrusal sınıflandırma için kullanılan açık kaynak kodlu bir kütüphanedir. Program lojistik regresyon ve doğrusal destek vektör makinelerini desteklemektedir (Fan vd. 2008). LIBLINEAR programı LIBSVM ile aynı üniversitenin (Ulusal Tayvan Üniversitesi) araştırma ekibi tarafından geliştirilmiştir. Proje sahipleri, kullanıcılara veri analizleri için başlangıç aşamasında iseler ve veri setleri çok büyük değil ise LIBSVM programını önermektedirler. Ayrıca, LIBLINEAR 'ın bazı durumlarda yavaş kalsa da doküman sınıflandırma için ön tanımlı çözücülerinin oldukça hızlı sonuç verdiği belirtilmektedir.

Lattice Miner: Lattice Miner veri kümesi içinde örüntüleri yaratan, görselleştiren ve ortaya çıkaran bir veri madenciliği prototipidir. Program kavram analizi ve birliktelik analizlerinin ortaya çıkarılmasına olanak tanır (Lattice Miner, 2016).

Mallet: MALLET istatistiksel doğal dil işleme, belge sınıflama, kümeleme, başlık modelleme, bilgi çıkarımı ve diğer metinlere uygulanan özdevinimli öğrenme uygulamaları için geliştirilmiş Java tabanlı bir paket programıdır (Mallet, 2016).

Mining Mart: Veri madenciliği süreçlerinde en fazla zaman harcanan kısmı veri ön işleme kısmıdır. Pratikteki deneyimlere göre zamanın %50'si ile %80'i arasındaki zaman veri ön işlemeye harcanmaktadır. Bu durum veri ön işlemeyi veri analizlerinin anahtar süreci yapmaktadır. Bu zaman genellikle; hangi öğrenme görevinin seçileceği, örnekleme, özellik belirleme, çıkarım ve seçimi, veri temizleme, model seçimi, değerleme kriterlerinin belirlenmesi gibi basamaklara ayrılmıştır. Mining Mart bu ihtiyaca cevap veren bir yazılımdır (Mining Mart, 2016).

ML-Flex: ML-Flex, bioinformatik alanındaki büyük boyutlu ve heterojen yapılardaki veriler kümelerinde, iki sınıflı ve çoklu sınıflı analizlerin yapılmasına olanak veren bir özdevinimli öğrenme programıdır. ML-Flex, Java dilinde yazılmıştır ancak üçüncü parti birçok

programlama diliyle birlikte çalışabilmektedir ve farklı formatlardaki veri setlerini girdi olarak alabilmektedir (Piccolo ve Frey, 2012: 555).

Modular Toolkit for Data Processing: MDP, Python dilinde yazılmış bir veri işleme yazılımıdır. Kullanıcı yönünden bakıldığında MDP danışmalı ve danışmasız öğrenme algoritmalarının toplandığı bir yazılım olduğu gibi ayrıca diğer ön beslemeli ağ mimarilerine bağlanabilir. MDP sinirbilimindeki teorik araştırmaları kapsamında yazılmış olsa da diğer alanlarda kullanılabilirdir (Zito vd., 2009: 1)

NLTK (Natural Language Toolkit) :İnsan dili verileri ile çalışma yapılabilen NLTK (Natural Language Toolkit), Python programlarının kurulumu için öncü bir platformdur. NLTK; hesaplanabilir dil bilimi ve doğal dil işleme konusunda program modülleri, veri kümeleri, araştırma ve öğretimler için başlangıç dersleri içeren bir ortamdır (Bird, 2006: 69)

OpenNN: OpenNN yapay sinir ağları uygulamaları içeren, başlıca çalışma alanı derin öğrenme (deep learning) olan, C++ dilinde yazılmış açık kaynak kodlu bir kütüphanedir. Özellikle ileri düzeyde C++ bilen ve özdevinimli öğrenme yeteneklerine sahip ileri düzey kullanıcılara yöneliktir (OpenNN, 2016).

Orange: 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 (Slovenya) Bilgisayar Fakültesi ve Bilgi Bilimi'nde geliştirilmiştir (Tekerek, 2011: 166).

Pandas: Pandas; python programlama dili için veri yapıları ve veri analizleri sağlayan, kütüphane desteği sunan, yüksek performanslı, kullanılması kolay, açık kaynak kodlu bir yazılımdır (Pandas, 2016).

Pybrain: PyBrain (Python-Based Reinforcement Learning Artificial Intelligence and Neural Network Library); python programlama dili için çok yönlü bir özdevinimli öğrenme kütüphanesidir. Pybrain; özdevinimli öğrenme işleri için, esnek, kolay kullanılır ancak güçlü algoritmalar içerir (Pybrain, 2016).

R: R istatistiksel hesaplama ve grafikleri için kullanılan bir bilgisayar programıdır. S programlama dili ile birçok konuda benzerlikler taşıyan R, aynı zamanda bir programlama dilidir.

Rapid Miner (Yale): Rapidminer; bazı eklentileri ve özelleştirmeleri ile sektöre yönelik çalışmaya başlasa da, programın çekirdek hali hala açık kaynak kodludur. Rapidminer; .aml, arff, att, bib, clm, cms, cri, csv, dat, ioc, log, mat, mod, obf, par, per, res, sim, thr, wgt, wls, xrff ve a gibi birçok dosya formatıyla kullanılabilmektedir (Kaya ve Özel, 2014: 51).

Rattle GUI: Rattle (the R Analytical Tool To Learn Easily) R veri madenciliği programını kullanma için popüler bir grafik kullanıcı ara yüzüdür. Rattle istatistiksel ve görsel olara veri özetlenmesi, veri dönüştürme, danışmalı ve danışmasız modeller kurma, modellerin preformanslarının grafikleştirilmesi ve yeni veri setleri oluşturulması olanaklarını sağlar.

Ayrıca Rattle; RStudio CRAN ’dan günlük yaklaşık 300 indirme sayısına ulaşmıştır (Rattla, 2016)

Rosetta: ROSETTA kaba küme teorisi çerçevesinde veri analiz etmek için bir araç takımıdır. ROSETTA veri madenciliği ve bilgi keşfi süreçlerinin gözden geçirilmesi, veri ön işleme, eğer-ise kuralları ortaya çıkarma, desen tanımlama süreçlerine destek sağlamak için tasarlanmıştır (Rosetta, 2016).

SIPINA: SIPINA, özellikle karar ağaçlarını (sınıflandırma ağaçlarını) hedeflemektedir. SIPINA, çeşitli danışmalı öğrenme paradigmalarını uygulayan bir veri madenciliği yazılımıdır. SIPINA; bütün aktiviteleri ücretsiz olan akademik bir araçtır. SIPINA;1995 yılından beri internet üzerinden dağıtılmaktadır. Temel olarak; ID3, CHAID, C4.5, ASSISTANT-86, vb. sınıflandırma ağaçlarına odaklanmış olsa da, diğer danışmalı metotlarda (örneğin; k-NN, Multilayer Perceptron, Naive Bayes, vb.) programda kullanılabilir durumdadır (Kaur ve Singh, 2013: 50).

Shogun: Shogun, geniş çapta öğrenme ortamı ve özellik türleri için tasarlanmış, büyük ölçekli öğrenim programıdır. SHOGUN kullanıcıya; destek vektör makineleri, saklı markov modelleri, çoklu çekirdek öğrenimi, doğrusal diskriminant analizi, vb birçok makine öğrenimi modeli sunmaktadır. Programın işlemsel biyoloji alanında, 50 milyondan fazla eğitim, 7 milyardan fazla test örneği ile uygulandığı olmuştur. Dünya çapında binden fazla sayıda indirilen SHOGUN C++ programlama dili ile yazılmış olup, MATLAB, R, Octave, Python gibi programlarla entegre çalışabilmektedir (Sonnenburg vd. 2010).

Scikit Learn: Scikit Learn; orta ölçekli danışmalı ve danışmasız problemler için bir Python modülüdür. Sınıflandırma, regresyon, kümeleme, boyut indirgeme, model seçimi, ön işleme gibi amaçlara hizmet eden bir yazılımdır (Scikit Learn, 2016).

SenticNet API: Sentic API, doğrudan ve imalı anlatım desteği sağlayan duygusal analiz yapan bir programdır [55].

TANAGRA: SIPINA programının halefi olan TANAGRA, özellikle görsel ve etkileşimli olarak karar ağaçlarının kurulumuna yoğunlaşan danışmalı öğrenim algoritmaları içeren bir açık kaynak kodlu bir programdır. TANAGRA, danışmalı öğrenim konusunda güçlü algoritmalar içerir ancak kümeleme, faktör analizi, parametrik ve parametrik olmayan istatistiksel analizler, birliktelik kuralları, özellik seçimi ve algoritma kurulumu gibi problemlerin çözümlerine de cevap verir (TANAGRA, 2016).

Vowpal Wabbit: Vowpal Wabbit (VW), başlagıçta *Yahoo! Research* tarafından geliştirilmiş ancak şu an *Microsoft Research* bünyesinde bulunan bir öğrenme sistemi kütüphanesidir (Vowpal Wabbit , 2016)

Weka: Weka projesi fikri 1992 yılına dayanmaktadır ve öğrenme algoritmaları birçok dile uygun, farklı platformlarda kullanılabilir ve çeşitli veri formatlarında işlem yapılabilirdir (Hall vd., 2009). Weka yazılımı ismini The Waikito Environment for Knowledge Analysis kelimelerinin baş harflerinden almıştır. Weka ayrıca, projenin ortaya çıktığı Waikito Üniversitesi’nin bulunduğu Yeni Zelanda’ya özgü bir kuş türünün de ismidir. Wekada hazır algoritmalar bir veri setine direk olarak uygulanabileceği gibi uygulayıcı kendi java kodu ile de algoritma yazabilir.

Arff, Csv, C4.5 formatında bulunan dosyalar Weka'da import edilebilir. Ayrıca Jdbc (Java Database Connectivity) kullanılarak veritabanına bağlanıp burada da işlemler yapılabilir (Dener, Dörterler ve Orman, 2009: 790).

3. ÜCRETSİZ VERİ MADENCİLİĞİ ARAÇLARININ TÜRKİYE'DEKİ YAYGINLIKLARI

Çalışmada 38 adet ücretsiz veri madenciliği yazılımı hakkında araştırmacıların cevaplaması için sorular oluşturulmuştur. Soruların ilk kısmı katılımcıların yazılımlardan haberdarlıklarının belirlenmesine yöneliktir. Rickert 2015, Malcolm Gladwell'in aykırılıklar ölçeğini (outliers scale), R programı öğrenimi için düzenlediği yazısında araştırmacıların R öğreniminde sadece konu ile ilgili bilgiye sahip olma, kullanıcı olma, programcı olma, katılımcı olma ve geliştirici olma seviyelerinde bulunabileceğini belirtmiştir. Buradan hareketle, bu çalışmada; veri madenciliği alanında çalışan bir araştırmacının, bir veri madenciliği yazılımı ile ilişkisi, yazılımı hiç duymamış olması, yalnızca haberdar olması, haberdar olması ve aynı zamanda bir araştırmada kullanmış olması ve araştırmada hazır olarak kullandığı gibi kendi yazdığı bir kod ile de kullanması düzeylerinde değerlendirilerek soru formu oluşturulmuştur. İlk kısımdaki sorular bu düzeyi belirlemeye yönelik olarak oluşturulmuştur. Ücretsiz program listesi sunularak bu düzey ölçülmeye çalışılmıştır. Katılımcılara yöneltilen bir diğer soru, yazılımların hangi amaçla (akademik, ticari, eğitim/öğretim, vb.) kullandıklarının ortaya konmasına yöneliktir. Ayrıca, katılımcıların kendilerini veri madenciliği konusundaki bilgi düzeylerini değerlendirmelerine yönelik bir soru daha yöneltilmiştir. Son olarak katılımcılardan listede belirtilmeyen bildikleri başka bir ücretsiz veri madenciliği aracını bilip bilmedikleri açık uçlu bir soru ile sorulmuştur. Bu soruları içeren soru formu internet üzerinden bir bağlantı ile paylaşılabilecek biçimde oluşturulmuştur.

Türkiye'de veri madenciliği üzerine çalışan kişilerin sayısı konusunda bilgiye sahip olunmadığı için, örneklemi oluşturacak birimlere kartopu örneklem metodu ile ulaşılma kararı verilmiştir. Kartopu örnekleme yöntemi araştırmacının, araştırma yapılacak evrenin sınırları hakkında yeterli bilgiye sahip olmadığı durumlarda tercih edilen örnekleme yöntemlerinden biridir. Kartopu örnekleme yönteminde, öncelikle araştırma evreni içerisinde yer alan ve araştırmacının ulaşabileceği ilk birim belirlenir. Bu birim üzerinden elde edilecek veriler ışığında sonraki birime ve daha sonra bunu zincirleme olarak takip eden diğer birimlere ulaşılarak evreni temsil edebileceği düşünülen örneklemin oluşturulması, böylelikle başlangıçta tek birimden oluşan örneklemin hacminin kartopu gibi büyütülerek oluşturulması amaçlanır (Ural ve Kılıç, 2006: 46). Buradan hareketle, veri madenciliği konusunda çalışma yaptığı bilinen kişilerden bir liste oluşturulmuş, soru formu bu kişilere gönderilmiştir. Katılımcılardan soru formunu veri madenciliği konusunda çalıştığını bildikleri kişilere ulaştırmaları istenmiştir ve 40 günlük bir zaman dilimi veri toplama süresi olarak belirlenmiştir. Bu süre sonunda veri toplama işlemi sonlandırılmıştır. Süre sonunda örneklem büyüklüğü 63 kişi olarak belirlenmiştir.

Tablo 2: Araçların Bilinirlikleri ve Kullanım İstatistikleri

	Haberdar Olunanlar		Kullanılanlar		Bir Kod ile Kullanılanlar	
	N	%	N	%	N	%
ADaM	3	4,76	0	0,00	0	0,00
AdamSoft	3	4,76	0	0,00	0	0,00
Alpha Miner	8	12,70	0	0,00	0	0,00
Apache mahout	9	14,29	3	4,76	1	1,59
CMSR Data Miner	2	3,17	0	0,00	0	0,00
Databionic ESOM Tools	0	0,00	0	0,00	0	0,00
Data Melt	1	1,59	1	1,59	1	1,59
Dlib	3	4,76	0	0,00	0	0,00
ELKI	1	1,59	0	0,00	0	0,00
Fityk	0	0,00	0	0,00	0	0,00
GGobi	0	0,00	0	0,00	0	0,00
GNU Octave	6	9,52	0	0,00	0	0,00
Jubatus	0	0,00	0	0,00	0	0,00
KNIME	30	47,62	17	26,98	6	9,52
Keel	0	0,00	0	0,00	0	0,00
LIBSVM	7	11,11	3	4,76	1	1,59
LIBLINEAR	1	1,59	0	0,00	0	0,00
Lattice Miner	0	0,00	0	0,00	0	0,00
Mallet	2	3,17	0	0,00	0	0,00
Mining Mart	0	0,00	0	0,00	0	0,00
ML-Flex	1	1,59	0	0,00	0	0,00
MDP	0	0,00	0	0,00	0	0,00
NLTK	3	4,76	1	1,59	0	0,00
OpenNN	3	4,76	0	0,00	0	0,00
Orange	28	44,44	15	23,81	5	7,94
Pandas	2	3,17	1	1,59	0	0,00
Pybrain	1	1,59	0	0,00	0	0,00
R	43	68,25	37	58,73	18	28,57
Rapid Miner (Yale)	40	63,49	29	46,03	11	17,46
Rattle GUI	5	7,94	2	3,17	0	0,00
Rosetta	5	7,94	1	1,59	1	1,59
SIPINA	2	3,17	0	0,00	0	0,00
Shogun	1	1,59	0	0,00	0	0,00
Scikit Learn	5	7,94	1	1,59	1	1,59
SenticNet API	0	0,00	0	0,00	0	0,00
TANAGRA	8	12,70	3	4,76	1	1,59
Vowpal Wabbit	1	1,59	1	1,59	1	1,59
Weka	51	80,95	41	65,08	15	23,81

Tablo 2 incelendiğinde; örneklemdeki kullanıcıların bazı yazılımlardan hiç haberdar olmadıkları görülmektedir. Örneklemdeki kullanıcılar tarafından en çok haberdar olunan yazılımın Weka olduğu tespit edilmiştir. Ücretsiz veri madenciliği araçlarından haberdarlık söz konusu olduğunda, Weka (%80,95), R(%68,25), Rapid Miner (%63,49) KNIME (%47,62), ve Orange (%44,44) isimli programlar dışındaki yazılımların çok az kişi tarafından bilindiği ortaya çıkmıştır.

Yazılımların herhangi bir araştırmada kullanılma istatistikleri incelendiğinde ise, kullanıcıların listede yer alan birçok yazılımı hiç kullanmadıkları görülmektedir. Örneklemde yer alan kullanıcıların en çok Weka'yı (%65,08) tercih ettikleri görülmektedir. Weka'yı sırasıyla; R (%58,73), Rapid Miner (%46,03), KNIME (%26,98), Orange (%23,81) izlemektedir. Haberdar olma ve kullanma istatistiklerinde ilk sırada Weka var ise de; araştırmacıların kendi yazdıkları ya da yazdırdıkları bir kod ile beraber kullandıkları program istatistiklerinin en başında ise R (%28,57) programı bulunmaktadır.

Araştırmada araştırmacıların yazılımları hangi amaçla kullandıklarının ortaya konulmasına yönelik olan sorunun cevaplarına ilişkin istatistikler Tablo 2'deki gibidir.

Tablo 3: Araçların Kullanım Amaçları

Kullanım Amacı	Kişi Sayısı
Akademik	39
Ticari	23
Öğretim-Öğrenim	19

Tablo 4:Kişilerin Kendilerini Değerlendirmelerine Yönelik İstatistikler

Değerlendirme Derecesi	Kişi Sayısı	Yüzde
1	12	%19
2	7	%11
3	12	%19
4	15	%24
5	17	%27

Katılımcıların birden fazla seçeneği işaretleyebileceği su soruda, 39 kullanıcı akademik amaçla, 23 kullanıcı ticari amaçla ve 19 katılımcı öğretim-öğrenim amacı ile herhangi bir veri madenciliği yazılımı kullandıklarını belirtmişlerdir.

Tablo 3 katılımcıların kendilerinin veri madenciliği bilgi düzeyini (1: Başlangıç Düzeyi - 5: İleri Düzey ölçeğinde) değerlendirmelerine yönelik soruya ilişkin cevapları göstermektedir.

Son olarak katılımcılardan listede belirtilenler dışında bildikleri veya kullandıkları ücretsiz bir yazılım var ise belirtmeleri istenen açık uçlu bir soruya ilişkin değerlendirme yapılmıştır. Bu soruya katılımcıların büyük çoğunluğu, listedekiler dışında bir yazılım bilmediklerini belirtmişlerdir. Bazı katılımcılar ise ücretli bazı veri madenciliği yeteneğine sahip program isimleri vermişlerdir. Ayrıca bir katılımcı; "Tamamen algoritmayı öğrendikten sonra kendim kodluyorum. Daha sonraki zamanlarda açık kaynak olarak internet ortamına aktaracağım." şeklinde bir cevap vermiştir.

4. SONUÇLAR ve ÖNERİLER

Organizasyonların ve bireylerin karar süreçleri açısından mevcut verilerin analizi büyük önem taşımaktadır. Bu analizleri veri madenciliği teknikleri ile gerçekleştirme işine duyulan ilgi Türkiye'de ve dünyada artış göstermektedir. Bu nedenle birçok veri madenciliği tekniği geliştirilmiş ve bu teknikleri kullanarak çözüme ulaşmayı kolaylaştırması açısından çok sayıda yazılım geliştirilmiştir. Bu yazılımların bazıları ücretli iken bazıları ise ücretsiz olarak kullanıcıların hizmetine sunulmaktadır. Ücretsiz yazılımlar, yazılımı çok sık kullanmayıp yalnızca özellikli bir araştırma için kullanan, yazılıma para ödemek istemeyen araştırmacılar tarafından tercih edilmektedir. Yazılımlara internet üzerinden kolaylıkla ulaşılması da bu yazılımların tercih edilir olması hususunda bir başka etmendir.

Bu çalışmada internet üzerinden indirilerek kullanılabilir olan 38 adet veri madenciliği yazılımı incelenmiştir. Bu yazılımların bir kısmının (KNIME, Weka, R, Rapid Miner, Orange) hali hazırda Türkiye 'den kullanıcılar tarafından bilinir olduğu ancak birçok veri madenciliği yazılımının da kullanıcılar tarafından bilinmediği tespit edilmiştir. Araştırmada kullanıcıların neden bu programlara tercih ettikleri sorulmamıştır. Ancak çalışma kapsamında edinilen bilgiler derlenecek olursa kullanıcıların bu öne çıkan programları tercih etmeleri izleyen şekilde sıralanabilir. Weka programının oluşturulma sürecinin başlangıcı 1992 yıllarına dayanmaktadır. Bu alandaki ilk programlardan biri olması nedeniyle kullanıcılar tarafından daha çok tercih edildiği söylenebilir. KNIME programının görsel veri akışı sağlaması, veri ve modelden sonuçlarını interaktif olarak sağlaması vb. avantajları kullanıcılar tarafından tercih edilmesinin nedenleri arasında sayılabilir. R programa dilinin geniş bir alanda kullanım bulması ve çoğu araştırmacının analizlere özel geliştirdikleri R paketlerini diğer araştırmacılara internette sunmaları R programını daha tercih edilebilir kılmıştır yorumu yapılabilir. Rapid Miner programının ise önemli avantajlarından biri çok sayıda veri formatı ile analiz yapılmasına olanak sağlamasıdır. Orange programının ise veri önişleme, özellik skorlama ve filtreleme, modelleme, model değerlendirme ve keşif teknikleri gibi geniş kapsamlı bileşen seti içeriyor olması programı öne çıkaran etmen olarak değerlendirilebilir. Çalışmada programları neden tercih edildiğinin araştırmacılara sorulması ve programların kullanıcılar için farklı yönden avantajları, dezavantajları, içerdikleri algoritmalar vb.leri ortaya konulabilecek bir biçimde daha kapsamlı bir ölçek haline getirilmesi bundan sonraki çalışmalarda ele alınabilecek bir konudur. Bunlara ek olarak yazılımlar arası farklılıklar ve benzerliklerde analiz edilebilecek konular arasındadır. Yazılımların aynı veri kümeleri için sonuç performansları, analiz hızı performansları gibi karşılaştırmalı analizler de bundan sonraki çalışmalarda ele alınabilecek konular arasındadır.

Elbette herhangi veri madenciliği aracı araştırmacın ihtiyacına cevap veriyor ise araştırmacının başka bir araca ihtiyaç duymayacağı düşünülebilir. Ancak, araştırmacının kullanacağı araçlarda tercih şansını arttırmanın önemli olduğu düşünülmektedir. Bu nedenle tespit edilen tüm ücretsiz veri madenciliği yazılımlarına değinilmeye çalışılmıştır. Veriden bilgiye giden yolda; verinin yapısı, kullanılacak kriterlerin tespiti, kullanılacak veri madenciliği tekniklerinin belirlenmesi, elde edilen sonuçların elden geçirilmesi gibi süreçlerin tamamı araştırmacı için bir karar problemidir. Bu süreçte spesifik bir konuda araştırmacıya hali hazırda bildiği ve kullandığı yazılımdan daha çok yardımcı olabilecek bir yazılım bulunabilir. Söz gelimi incelenen veri madenciliği yazılımlarından Mining Mart, veri ön işleme sürecindeki eylemlere yardımcı olmak için tasarlanmıştır ya da SIPINA yalnızca karar ağaçları konusunda çözüm sunan bir yazılımdır. Bazı durumlarda, kullanıcının daha çok yeteneği hedefleyen bir yazılımı kullanmak yerine bu tür spesifik bir araca hizmet eden bir yazılımım tercih etmesi daha uygun olabilir.

Çalışmada örnekleme dâhil olan kullanıcılara listelenen yazılımlar dışında bir yazılım bilip bilmedikleri sorulmuş ve listelenenler dışında ek bir yazılım tespit edilememiştir. Ancak, bu alan günden güne hızlı bir şekilde gelişmeye devam eden bir alandır. Elbette gözden kaçan başka ücretsiz yazılımlar bulunabilir. Son olarak, bu haliyle çalışmanın; Türkiye'de veri madenciliği alanında çalışan araştırmacılar için kullanışlı olabilecek özellikteki ücretsiz yazılımları, temel özellikleri ile açıklaması, tanıtması ve bir arada sunması açısından önemli olduğu ve Türkiye'de bu alana katkı yaptığı düşünülmektedir.

KAYNAKÇA

ADaM, (2016)
http://projects.itsc.uah.edu/datamining/adam/index.html (Erişim Tarihi: 16.04.2016)

Adamsoft,(2016)
http://adamsoft.sourceforge.net/index.html (Erişim Tarihi: 16.04.2016)

Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İstanbul Üniversitesi İşletme Fakültesi Dergisi, 29 (1): 1-22

AlphaMiner, (2016)
http://www.datamining.gr/en/links/20--open-source.html (Erişim Tarihi: 18.04.2016)

Bird, S. (2006). NLTK: The Natural Language Toolkit, *Proceedings of the COLING/ACL 2006*, Interactive Presentation Sessions, Sydney, ss.69–72.

Bramer, M. (2007). *Principles of Data Mining*, Springer-Verlag London.

Cabena P., Hadjinian P., Stadler R., Verhees J., ve Zanasi A. (1998), *Discovering Data Mining: From Concept to Implementation*, Prentice Hall, Upper Saddle River, NJ.

Chang, C. C. ve Lin, C.J. (2011). LIBSVM: A Library for Support Vector Machines, *ACM Trans. Intell. Syst.Technol*. 2 (3): 1-27

CMSR Data Miner, (2016)
http://www.roselladb.com/starprobe.htm (Erişim Tarihi: 21.04.2016)

Databionic ESOM Tools, (2016)
http://databionic-esom.sourceforge.net/index.html (Erişim Tarihi: 10.04.2016)

DataMelt, (2016)
http://jwork.org/dmelt/ (Erişim Tarihi: 19.03.2016)

Dener, M., Dörterler, M. ve Orman, A. (2009). Açık Kaynak Kodlu Veri Madenciliği Programları: WEKA'da Örnek Uygulama Akademik Bilişim'09 - XI. Akademik Bilişim Konferansı Bildirileri, Harran Üniversitesi, Şanlıurfa, ss.787-796.

Dlib, (2016)
http://dlib.net/intro.html (Erişim Tarihi: 19.03.2016)

ELKI, (2016)
http://elki.dbs.ifi.lmu.de/ (Erişim Tarihi: 20.03.2016)

Fityk, (2016)
http://fityk.nieto.pl/ (Erişim Tarihi: 10.03.2016)

Fan, R.E., Chang, K.W., Hsieh, C. J., Wang, X.R. ve Lin, C.J. (2008). LIBLINEAR: A Library for Large Linear Classification, *Journal of Machine Learning Research 9*: 1871-1874.

Fayyad, U. M., Piatetsky-Shapiro, G. ve Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. PiatetskyShapiro, P. Smyth, and R. Uthurusamy, 1–30. Menlo Park, Calif.: AAAI Press, ss.37-54.

Ggobi, (2016)
http://www.ggobi.org/ (Erişim Tarihi: 15.04.2016)

GNU Octave, (2016)
https://www.gnu.org/software/octave/ (Erişim Tarihi: 10.03.2016)

Hall, M., Frank E., Holmes, G., Pfahringer B., Reutemann, P. ve Witten, I.H. (2009). The WEKA Data Mining Software: An Update, *ACM SIGKDD Explorations Newsletter*, 11 (1):10-18

Han, J. & Kamber, M. (2001). *Data Mining, Concepts and Techniques*. Morgan Kaufmann Publishers.

Hand D., Mannila, H., ve Smyth, P., (2001). *Principles of Data Mining*, MIT Press, Cambridge, MA.

Jubatus, (2016)
http://jubat.us/en/overview.html (Erişim Tarihi: 16.03.2016)

Kaur, A. ve Singh, S. (2013). Classification and Selection of Best Saving Service for Potential Investors using Decision Tree – Data Mining Algorithms. *International Journal of Engineering and Advanced Technology (IJEAT)* ISSN: 2249 – 8958, 2 (4): 80-82

Kaya, M. ve Özel, S.A. (2014). Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması. 16. Akademik Bilişim Konferansı, Mersin Üniversitesi, Mersin, Mersin Üniversitesi, Mersin, ss.47-53.

KEEL, (2016)
http://www.keel.es/ (Erişim Tarihi: 09.03.2016)

King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10: 1755-1758

KNIME, (2016)
https://www.knime.org/ (Erişim Tarihi: 08.04.2016)

Lattice Miner, (2016)
http://sourceforge.net/projects/lattice-miner/ (Erişim Tarihi: 08.04.2016)

LIBLINEAR, (2016)
http://www.csie.ntu.edu.tw/~cjlin/liblinear/ (Erişim Tarihi: 08.04.2016)

LIBSVM, (2016)
https://www.csie.ntu.edu.tw/~cjlin/libsvm/ (Erişim Tarihi: 08.04.2016)

Mahout, (2016)

http://mahout.apache.org/ (Erişim Tarihi: 11.03.2016)

Mallet, (2016)
http://mallet.cs.umass.edu/ (Erişim Tarihi: 10.03.2016)

MDP, 2016 http://mdp-toolkit.sourceforge.net/ ve https://pypi.python.org/pypi/MDP/2.4 (Erişim Tarihi: 18.04.2016)

Mining Mart, (2016)
http://mmart.cs.uni-dortmund.de/ (Erişim Tarihi: 18.04.2016)

Mlflex, (2016)
http://mlflex.sourceforge.net/ (Erişim Tarihi: 23.04.2016)

NTLK, (2016)
http://www.nltk.org/ (Erişim Tarihi: 23.04.2016)

OpenNN, (2016)
http://www.artelnics.com/opennn/ (Erişim Tarihi: 12.04.2016)

Orange, (2016)
http://orange.biolab.si/ (Erişim Tarihi: 18.04.2016)

Pandas, (2016)
http://pandas.pydata.org/ (Erişim Tarihi: 18.04.2016)

Piccolo, S. R. ve Frey, L. J. (2012). ML-Flex: A Flexible Toolbox for Performing Classification Analyses In Parallel, *Journal of Machine Learning Research* 13: 555-559.

Pybrain, (2016)
http://pybrain.org/ (Erişim Tarihi: 17.04.2016)

R, (2016) https://www.r-project.org/

Rapidminer, (2016)
https://rapidminer.com/ (Erişim Tarihi: 10.02.2016)

Rattle, (2016)
http://rattle.togaware.com/ (Erişim Tarihi: 17.04.2016)

Rickert, J. (2015). Learning R: Index of Online R Courses. http://blog.revolutionanalytics.com/2015/10/learning-r-oct-2015.html?utm_content=bufferc5df8&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer (Erişim Tarihi: 12.12.2016)

Rosetta, (2016)
http://www.lcb.uu.se/tools/rosetta/ (Erişim Tarihi: 15.04.2016)

Rushing, J., Ramachandran, R., Nair, U., Graves, S., Welch, R. ve Lin H. (2005). ADaM: A Data Mining Toolkit For Scientists And Engineers. *Computers & Geosciences*, 31: 607–618.

Scikit Learn, (2016)
http://scikit-learn.org/stable/ (Erişim Tarihi: 19.04.2016)

Schelter, S. ve Owen S. (2012). Collaborative Filtering with Apache Mahout Recommender, Systems Challenge 2012 in Conjunction with the ACM Conference on Recommender Systems 2012.

SenticNet API, (2016)
http://sentic.net/api/ (Erişim Tarihi: 20.04.2016)

Shogun, (2016)
http://www.shogun-toolbox.org/page/contact/irclog/2013-10-13/ (Erişim Tarihi: 20.04.2016)

Sipina, (2016)
http://eric.univ-lyon2.fr/~ricco/sipina.html (Erişim Tarihi: 20.04.2016)

Sonnenburg, S., Ratsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., Bona, F., Binder, A., Gehl, C. ve Franc, V. (2010). The SHOGUN Machine Learning Toolbox. *Journal of Machine Learning Research,* 11: 1799-1802

Swayne, D. F., Lang D. T., Buja, A. ve Cook, D. (2003). GGobi: Evolving from XGobi In to an Extensible Framework for İnteractive Data Visualization. *Computational Statistics & Data Analysis,* 43: 423-444.

Tanagra, (2016)
http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html (Erişim Tarihi: 26.02.2016)

Tekerek, A. (2011). Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği Araçları. Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri, İnönü Üniversitesi, Malatya, ss.161-169.

Ural, A. ve Kılıç, İ. (2006). *Bilimsel Araştırma Süreci ve SPSS ile Veri Analizi*, 2. Baskı, Detay Yayıncılık, Ankara.

Witten, H. I. ve Frank E. (2005). *Data Mining Pratical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers.

Vowpal Wabbit, (2016)
http://hunch.net/~vw/ ve https://github.com/JohnLangford/vowpal_wabbit/wiki (Erişim Tarihi: 26.02.2016)

Zito, T., Wilbert, N., Wiskott, L. ve Berkes, P. (2009). Modular Toolkit for Data Processing (MDP): A Python Data Processing Framework, *Frontiers in Neuroinformatics,* 2, Article 8: 1-7