YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 3, Sayı 4, Aralık 2016

Veri Madenciliği Araçları (Data Mining Tools)

Merve YILDIZ[1], Şadi Evren ŞEKER[2]

1. *Karabük Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü*
2. *İstanbul Şehir Üniversitesi, Yönetim Bilişim Sistemleri Bölümü*

Özet

Veri madenciliği, günümüzde hızla gelişen ve önemini arttıran bir alandır. Bu alanın gelişmesinde üniversite projeleri ve açık kaynak kodlu projelerin önemli katkısı olmuştur. Günümüzde veri madenciliği ve buna bağlı olarak veri bilimi, makine öğrenmesi ve büyük veri gibi çalışma alanlarında yoğun olarak açık kaynak kodlu projelerden faydalanılmakta ve bu projeler gerek akademik ve bilimsel, gerekse de endüstriyel çoğu uygulama için kritik bir önem taşımaktadır. Bu araştırmada açık kaynak kodlu veri madenciliği araçlarından olan WEKA, Rapid Miner, Knime, Keel, Orange ve MOA incelenmiş, kurulumları yapılarak basit birer proje geliştirilmiş ve hem proje geliştirme süreçleri hem proje geliştirme hızı incelenmiş, yazılımların kurulumundan kullanılmasına kadar olan süreçler analiz edilerek karşılaştırılmıştır. Bu makalenin amacı, bu yazılımları kullanma potansiyeli olan araştırmacılara yazılımlar hakkında genel bilgiler vermek ve hızlı bir karşılaştırma imkanı sunmaktır.

Anahtar Kavramlar: *Veri Madenciliği, Veri Madenciliği Araçları*

Abstract

Data mining is one of the trending topics now a days. University projects and open sourced projects have a great contribution to this area. Contemporary studies in data mining and connected areas, like data science, machine learning and big data, are massively benefitting from open sourced projects and they are critical for both academic, scientific and industrial projects. This study mainly focuses on the data mining tools, WEKA, Rapid Miner, Kime, Keel, Orange and MOA. During the study, all these software are installed, and a simple project is implemented over them. By the way, the installation, project development process and time are compared and analyzed. Aim of this paper is providing an introductory baseline about the software, for researchers interested in the area and a baseline comparison about the tools in the area.

Keywords: *Data Mining, Data Mining Tools*

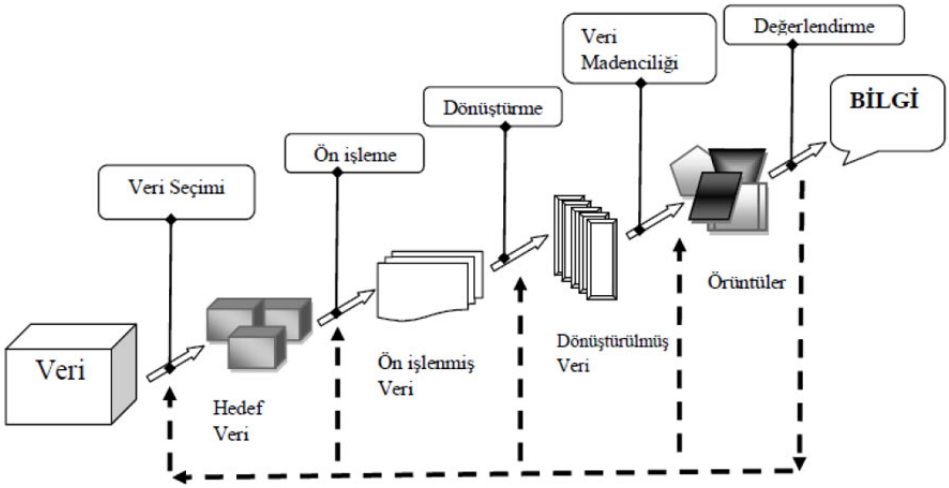
1. Giriş

Veri madenciliği büyük veriler arasından bilgiye ulaşma işidir. Veri madenciliği araçlarını, çok farklı gerçek hayat projelerini, sınıflandırma, bölütleme, doğrulama, gelecek ile ilgili tahminde bulunma veya birliktelik analizi için kullanabiliriz. [1]

Veri madenciliği bilgi bulma işleminin önemli bir adımıdır. Bilgi bulma süreci, literatürde bilgi ve veri keşfi olarak da geçmektedir (knowledge and data discovery, KDD) ve bu bilgi bulma adımları aşağıdaki gibi sıralanabilir.

1. Veri Temizleme (gürültülü, tutarsız verileri silip eksik verileri doldurmak gibi)
2. Veri Bütünleştirme (veri kaynaklarını birleştirmek)
3. Veri Seçme (İşe yarayan verileri seçmek)

4. Veri Dönüşümü (Verilerin veri madenciliği araçlarında daha başarılı kullanabilecek hale dönüştürmek)
5. Veri Madenciliği (Veri örüntülerinin yakalayabilmek için akıllı yöntemler uygulamak)
6. Örüntü Değerlendirme (Yapılan ölçümlere göre elde edilmiş bilgiyi temsil eden ilginç örüntüleri bulmak)
7. Bilgi Sunumu (Elde edilen bilginin kullanıcıya sunumu) [2,3].



Şekil 1 Veri madenciliği aşamaları gösterilmiştir.

Şekil 1'de bahsi geçen KDD adımları sırasıyla görselleştirilmiştir. Bu adımlardan, veri seçimi problemle ilgili olmalı ve iyi analiz edilmelidir. Seçilen veri diğer aşamalarda kullanılacağı için bu aşama çok önemlidir doğru veri üzerinde çalışılması gerekir. Seçilen verideki tutarsızlıkların giderilmesi veya gürültülü verilerin temizlenmesi ön işleme aşamasında çözülür. Veri madenciliği aşamasında veriler kullanılabilir haldedir ve bu aşamada uygun veri madenciliği yöntemi seçilerek projeye göre ayarlanır. Veri değerlendirme aşamasında ise sonuçlar değerlendirilir ve uygun bulunması durumunda tekrarlanabilir, farklı amaçlarla kullanılabilir ve insanlar için anlamlı bilgiye ulaşılmış olunur.

Literatürde daha önce de yapılan çok sayıdaki çalışmaya bakarak, Weka, Rapid Miner, Knime, Keel, Orange, Moa, ve Elki, paketlerinin kullanılan en yaygın veri madenciliği araçlarından olduğunu söyleyebiliriz [4]. Amacımız bu araçların özelliklerini anlatıp yaptığımız birkaç uygulama örneğini sunarak kurulum süreçleri, basit proje uygulama süreçleri, diskte kapladığı alan, hiç bilmeyen birisinin araca adaptasyon süreci gibi açılardan bu araçları karşılaştırmaktır. Bu makalede, literatürdeki diğer çalışmalardan farklı olarak araçlarda, titanic verisi üzerinde birkaç örnek yapıp görsellerine yer verilmiştir. Örneklerin kaç dakikada yapıldığı ve araçların bilgisayarda ne kadar yer kapladığı da makalede sunulmuştur.

2. Açık Kaynak Kodlu Veri Madenciliği Araçları

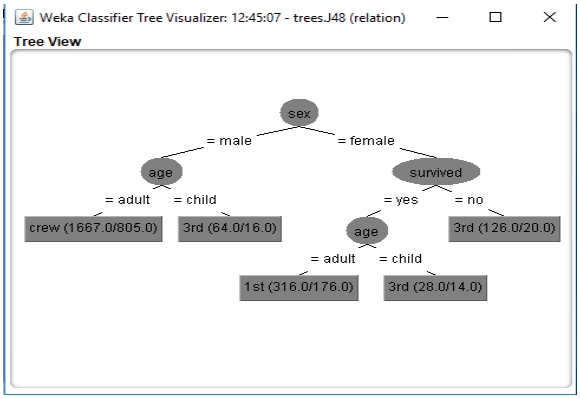
2.1. Weka

Weka(Waikato Environment for Knowledge Analysis) Waikato Üniversitesi(Yeni Zelanda)'nın bir ürünüdür. Açık kaynak kodludur. Java dilinde yazılmıştır ve GPL lisansı ile dağıtılmaktadır. Weka'da algoritmaları doğrudan bir veri kümesine uygulayabilir veya kendi yazdığınız Java kodlarınızdan çağırabilirsiniz. Weka, veri görselleştirme, ön işleme, sınıflandırma, kümeleme, gerileme ve ilişkilendirme araçlarını içerir. [5]



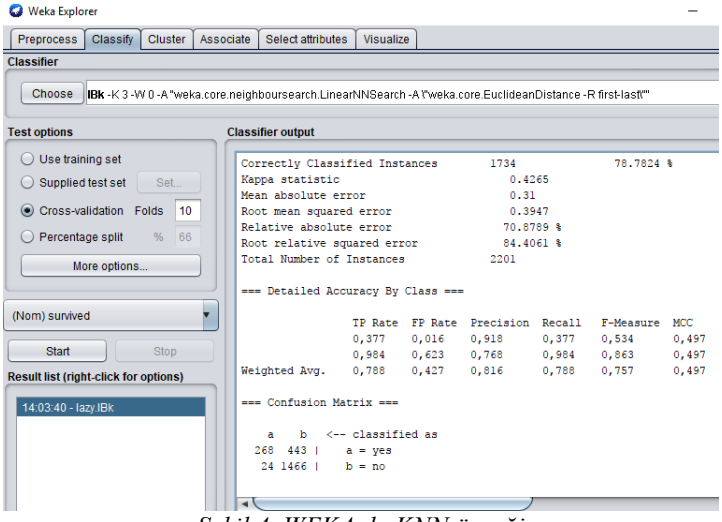
Şekil 2 WEKA Applications Menüsü

Orijinal sürümü tarımsal alanlardan gelen verileri analiz etmek için bir araç olarak tasarlamış olsa da 1997'de Java tabanlı sürümüne geçildikten sonra birçok alanda kullanılmaktadır.



Şekil 3. WEKA'da titanic veri kümesi için üretilen Karar Ağacı

WEKA ile kolaylıkla titanic veri kümesi üzerinde karar ağacı uygulanmış ve ortama tamamen yeni bir kişi tarafından, örnek yaklaşık 3 dakikada yapılmıştır.



Şekil 4. WEKA da KNN örneği

Aynı örnek, en yakın k komşu, KNN (K Nearest Neighbors) algoritması ile çalıştırılmış, k değeri 3 olarak alınmış ve başarı %78,7824 olarak gözlenmiştir. KNN ile aynı örneği yapmak yaklaşık 2 dakika sürmüştür.

Özellikleri:

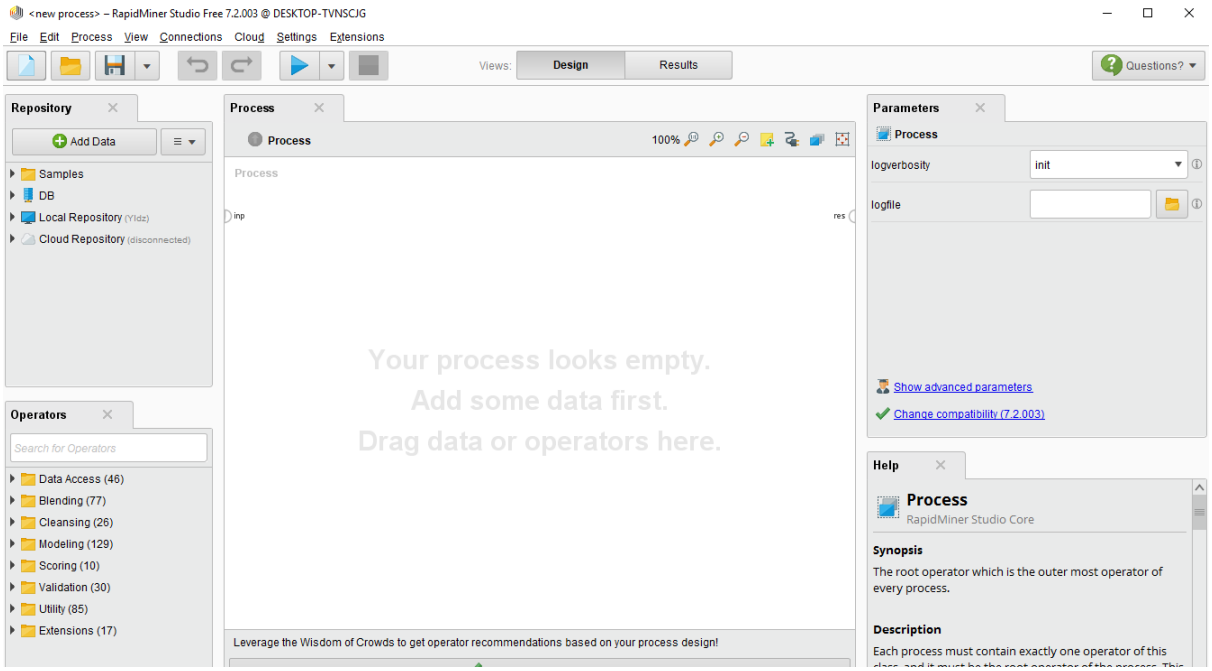
1. Grafiksel kullanıcı ara yüzleri sunar ve sadece fare tıklamaları ile işlemler yapılabilir
2. GNU Genel Kamu Lisansı altında ücretsiz kullanılabilir.
3. WEKA' da bulunmayan algoritmalar, daha sonradan yüklenebilir ve Java ile geliştirme için özel dokümantasyonlar ve yazım kuralları yine aynı araştırma grubu tarafından sunulmaktadır.
4. Weka kütüphaneleri açık kaynak kodludur ve Java projelerinden doğrudan çağrılabilir.
5. Bilgisayarda kapladığı alan 106 MB' dır.
6. Kurulum süresi yaklaşık 12 dakika sürmüştür

2.2. Rapid Miner

Rapid Miner, Dortmund Teknoloji Üniversitesi Yapay Zekâ biriminde geliştirilmiş bir veri madenciliği yazılımıdır. Java ile yazılmıştır. Rapid Miner veri madenciliği, metin madenciliği, makine öğrenmesi, tahmin edici analiz ve iş analizi amaçlarına yönelik olarak geliştirilmiştir [6]. Yazılım Rapid Miner firması tarafından lisanslanmış ve günümüzde ücretli olan profesyonel versiyonunun yanında orijinal olarak başlamış olan topluluk destekli yazılım (community founded software) lisansı ile de daha kısıtlı bir versiyonu dağıtılmaktadır. Yazılım çoğunlukla iş ve ticari amaçlarda kullanılmasının yanı sıra eğitim, araştırma, hızlı prototiplime ve uygulama geliştirme gibi farklı amaçlarda da kullanılır. Ayrıca veri madenciliği sürecinin bütün aşamaları Rapid Miner tarafından desteklenmektedir, bundan dolayı veri hazırlama, sonuçları görselleştirme, doğrulama ve optimizasyon gibi amaçlarla kullanılabilir. [7]

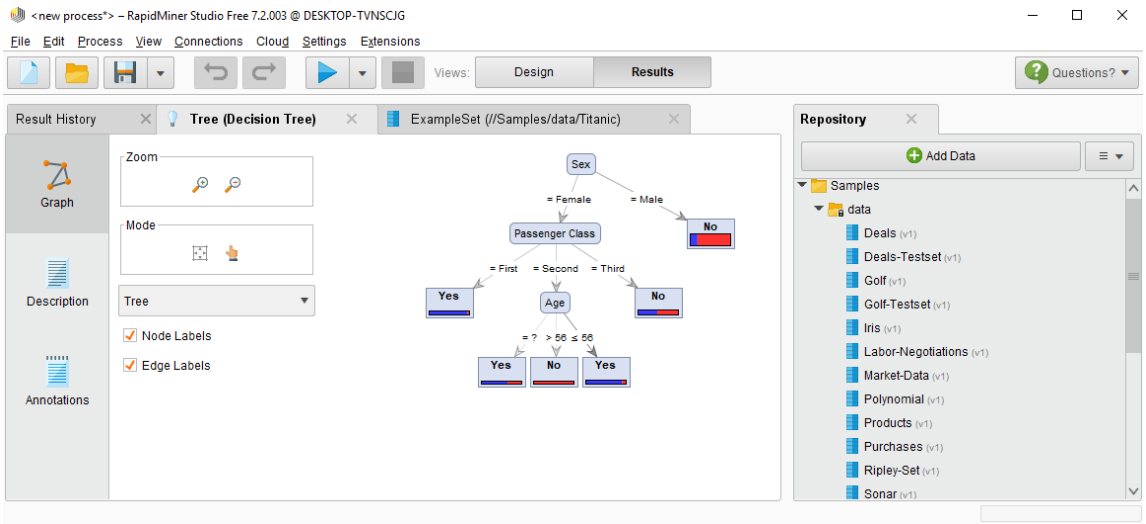
Şekil 5. Rapid Miner Logosu

Şekil 5'te rapid miner'ın ticari logosu ve Şekil 6'da açılışta gelen boş proje ekranı görüntüsüne yer verilmiştir.



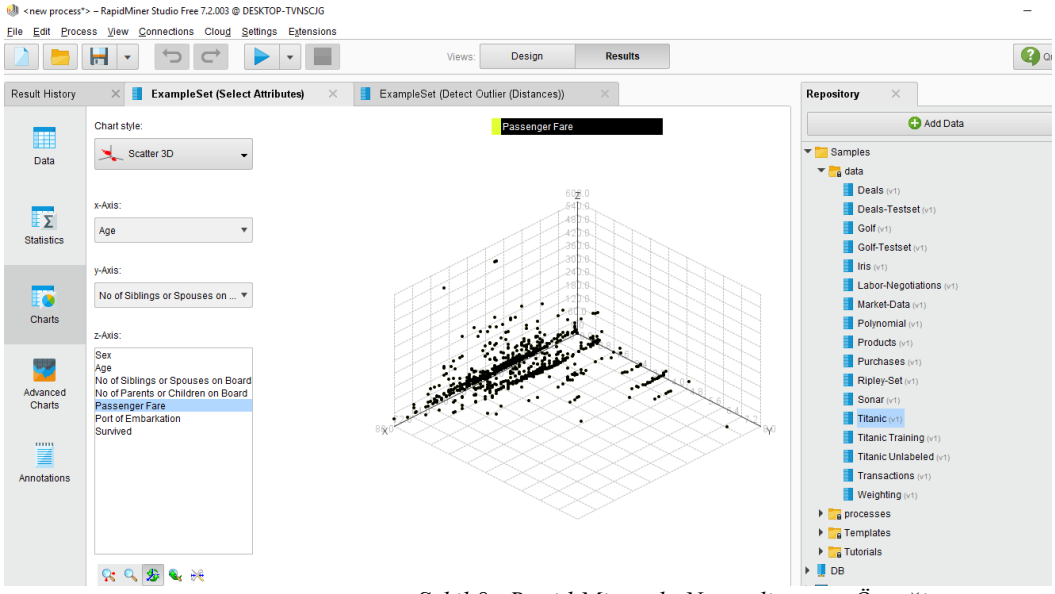
Şekil 6. Rapid Miner Application Menüsü

Bir önceki alt bölümde, Weka ile yapılan titanic veri kümesi üzerindeki karar ağacı projesinin aynısının Rapid Miner ile yapılmış hali ise Şekil 7'de görselleştirilmiştir.



Şekil 7. Rapid Miner'da Titanic Veri Kümesi için çizilen Karar Ağacı

Rapid Miner ile titanic veri kümesi üzerinde bir karar ağacı uygulaması, yazılımı ilk kez kullanan bir kullanıcı tarafından yaklaşık 3 dakikada yapılmıştır.



Şekil 8. Rapid Miner da Normalizasyon Örneği

Rapid Miner, aynı zamanda farklı görselleştirme özellikleri ve veri ön işleme imkanlarını da sunar. Bu özelliklerini denemek için Rapid Miner üzerinde yapılan normalizasyon ve sonrasındaki görselleştirme Şekil 8'de sunulmuştur ve örnek yaklaşık olarak 4 dakikada yapılmıştır.

Özellikleri:

1. Görselleştirme açısından diğer yazılımlara göre avantajlar sunar.
2. Topluluk lisansı altında gelen versiyonu, AGPL açık kaynak lisansı altında ücretsiz kullanılabilir.
3. Profesyonel versiyonu ücretlidir ve ücret karşılığında eğitim veya teknik destekler sunulabilir.
4. Kurulumu yaklaşık 6 dakika sürmüştür.
5. Bilgisayarda 157 MB yer kaplar.

2.3. KNIME

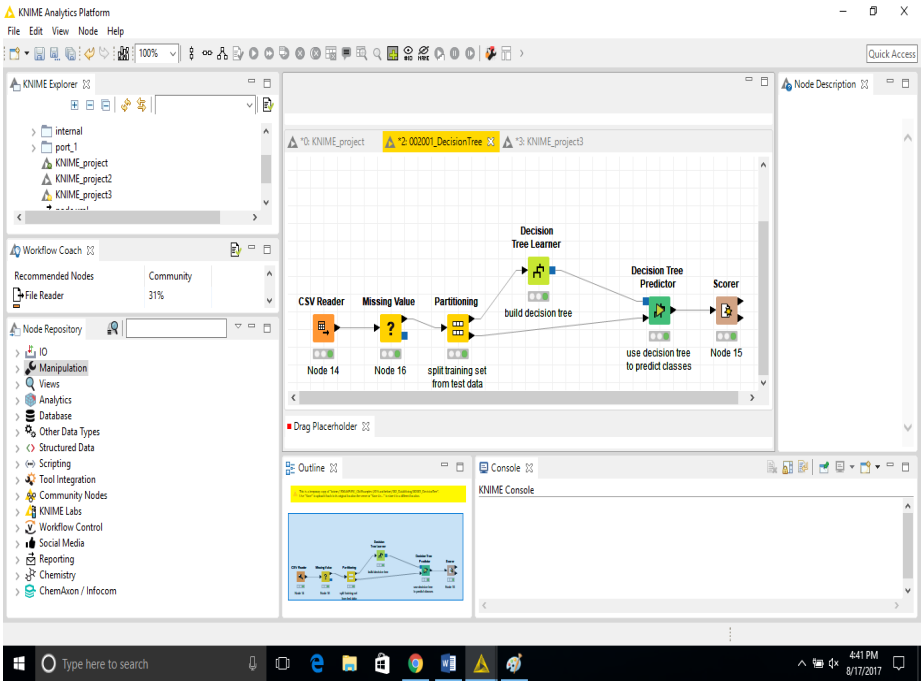
Knime (Konstanz Information Miner) Konstanz Üniversitesi görsel veri madenciliği araştırma grubu tarafından Eclipse Rich Client Platform üzerinde Java ile yazılmış açık kaynak kodlu ücretsiz bir yazılımdır. Açık kaynak kodlu olmasıyla JDK kurarak makine öğrenimi çalışmaları için uygun bir araçtır. Hem kurulabilen hem de yazılımı kurmadan disk üzerinde çalışabilen bir yazılımdır.

KNIME modüler API özelliği ile Eclipse ortamını temel alır ve kolayca genişleyebilir. Bu özelliği Knime'ın öğretim ve araştırma özelliklerinin yanında ticari olarak da kullanılmasını sağlar.[8]



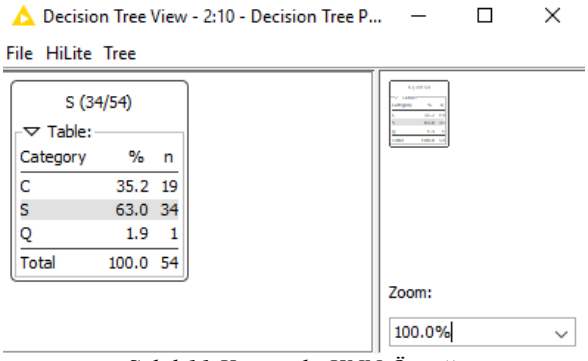
Şekil 9. Knime Logosu

Knime aracının logosu Şekil 9'da sunulmuştur ve diğer araçlarda da çözülen Titanic veri kümesinin Knime ile karar ağacı kullanarak çözülmüş hali ve görselleştirilmesi Şekil 10'da sunulmuştur.



Şekil10. Knime da Karar Ağacı

Örnek yaklaşık 4 dakikada yapılmıştır ve başarı oranlarını gösteren sonuç tablosu Şekil 11'de sunulmuştur.



Şekil 11.Knime da KNN Örneği

Knime'ın genel özellikleri:

1. Görsellestirme açısından en iyi yazılımlardan biridir.
2. Kurulumu 5 dakika sürmüştür.
3. Bilgisayarda 340 MB yer kaplar.

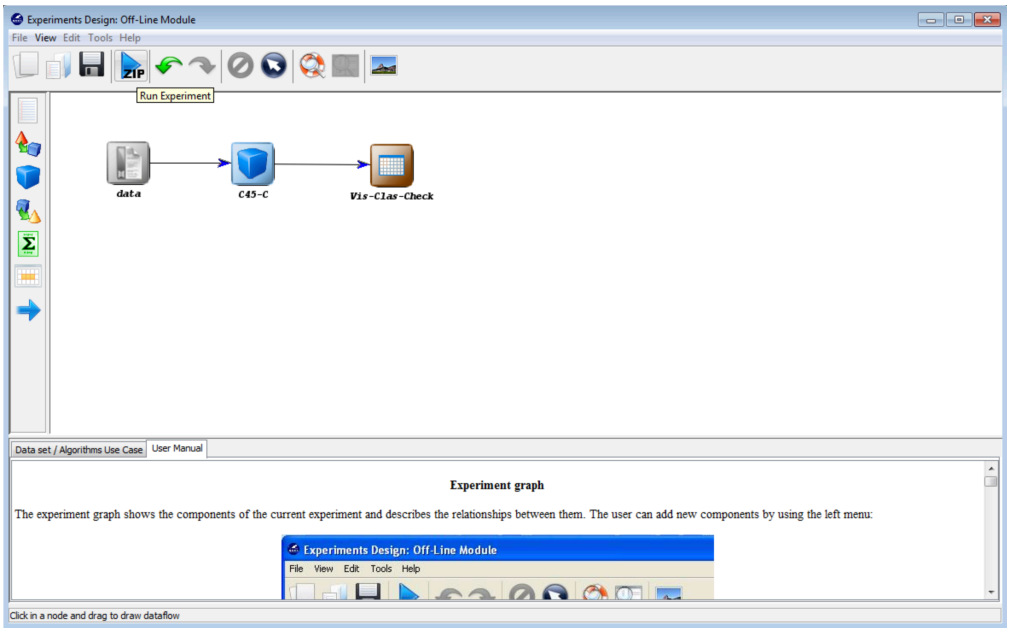
2.4. KEEL

KEEL İspanya Ulusal Bilim Projeleri Kurumunun desteği ile Granada Üniversitesi tarafından geliştirilmiştir. Java dilinde kodlanmıştır.



Şekil 12. Keel Applications Menüsü

Keel uygulamasının çalıştırılınca çıkan ilk ekranı Şekil 12'de ve Keel uygulamasında karar ağacı çalıştırılması için kurulan akış düzeni Şekil 13'te sunulmuştur.



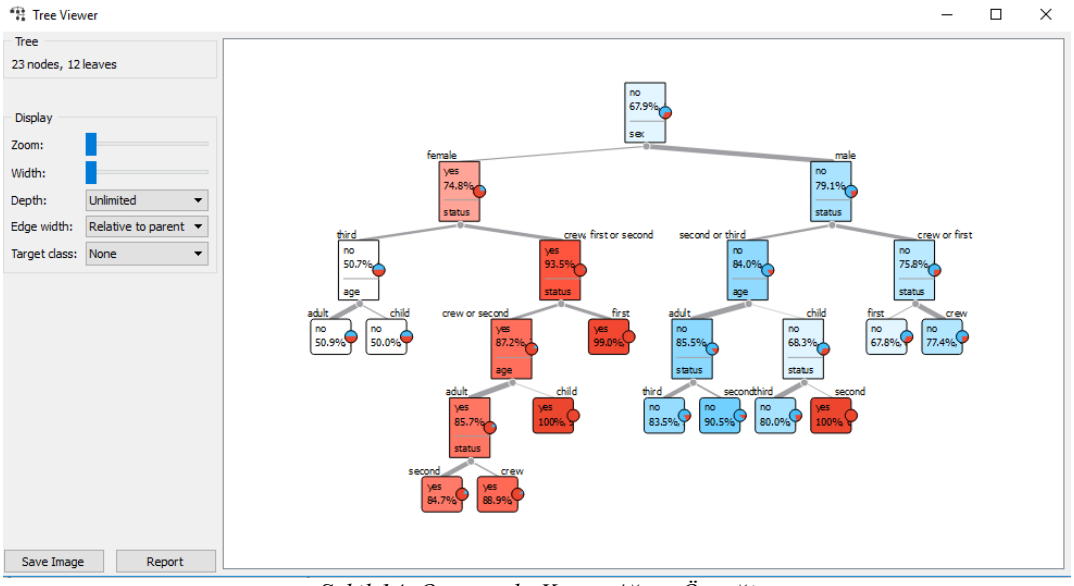
Şekil 13. Keel ile Karar ağacı uygulaması

Özellikleri:

1. Yapay zekâ tabanlı sınıflandırma ve Kural tabanlı kümeleme algoritmalarının birçok çeşidini içermektedir.[9]
2. Kurulum şartı olmadan çalışabilmektedir.
3. Kurulumu yaklaşık 2 dakika sürmüştür.
4. Bilgisayarda 130 MB yer kaplar.

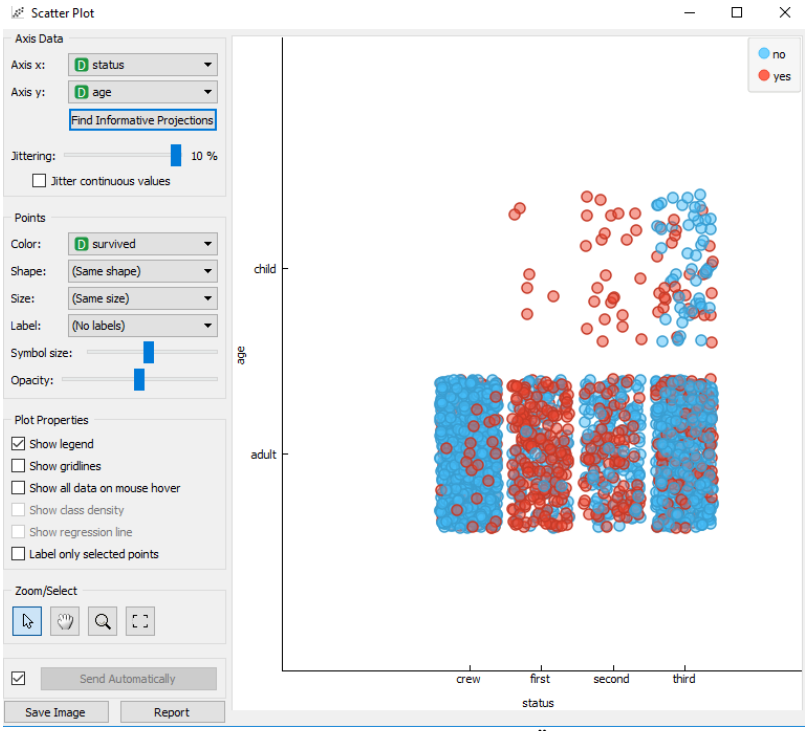
2.5. ORANGE

Orange, Slovenya Ljubljana Üniversitesi Bilgisayar ve Enformatik Bilimleri bölümü yapay zekâ araştırmaları ekibi tarafından geliştirilmiş bir açık kaynak kodlu yazılımdır. Orange yazılımı C++ dili ile geliştirilmiştir. Yazılımın ara yüzleri ve grafik ortamı ise Qt3 kütüphanesi ve Python kullanılarak geliştirilmiştir.[8] Orange kullanıcıya veri hazırlama, keşifsel veri analizi, modelleme gibi imkânlar sağlar. Yazılımda veri madenciliği işi görsel programlama veya Python scripting ile yapılabilir. Makine öğrenmesi için bileşenleri vardır. Bioinformatics ve metin madenciliği için eklentileri vardır. [10]



Şekil 14. Orange da Karar Ağacı Örneği

Diğer araçlarda çalıştırılan Titanic veri kümesi için karar ağacı örneğ yaklaşık 3 dakikada yapılmıştır ve kara ağacının çıktısı Şekil 14'te sunulmuştur. Ayrıca Orange, görselleştirme imkanı da sunmaktadır ve örnek bir görselleştirme, Şekil 15'te sunulmuştur.



Şekil 15. Orange da K-Means Örneği

Örnek yaklaşık 3 dakikada yapılmıştır.

Orange yazılımının özellikleri:

1. Python diliyle yazılmasından dolayı yazılım içerisinde script yazılabilmektedir.
2. Kurulumu yaklaşık 8 dakika sürmüştür.
3. Bilgisayarda 109 MB yer kaplar.

2.6. Massive Online Analysis

Massive Online Analysis (MOA), Yeni Zelanda'nın Waikato Üniversite geliştirilmiş ve Java diliyle yazılmış açık kaynak kodlu yazılımdır [11]. Yazılım daha çok zamana bağlı [12] akan veri (streaming data) [13] üzerinde madencilik ve özel olarak dağıtılmış işleme ve büyük veri problemleri için geliştirilmiştir. Yazılım, Weka'yı geliştiren ekip tarafından geliştirilmiş olmasına karşın, bu açılardan Weka'dan ayrılır.



Şekil 15. MOA Logosu

Özellikleri:

1. MOA, en sık bulunan veri jeneratörlerini içerir. Şu jeneratörler mevcut: Random Tree Generator, Sea Concepts Generator, Stagger Kavramları Jeneratör, Dönen Hiperplane, Rastgele RBF Jeneratör, Led Jeneratör, Dalga Formu Jeneratörü ve İşlev Jeneratörü.[14]
2. Kurulumu 3 dakika sürmüştür.
3. Bilgisayarda 30,3 MB yer kaplar.

2.7. ELKI

Elkı(for Environment for DeveLoping KDD-Applications Supported by Index-Structures), Münih Ludwig Maximilian Üniversitesi'nden Profesör Hans ve Peter Kriegel'in veri tabanı sistemleri araştırma biriminde geliştirilmiş Java ile yazılmış açık kaynak kodlu yazılımdır. [15] Elkı veri madenciliği algoritmalarının geliştirilmesi ve değerlendirilmesine ve bunların veritabanı dizin yapılarıyla etkileşimini sağlar.

Özellikleri:

1. Kurulumu 4 dakika sürmüştür.
2. Bilgisayarda 78,2 MB yer kaplar.

3. Araçların Karşılaştırılması ve Sonuç

Rapid Miner, Knime ve WEKA veri madenciliği açık kaynak kodlu araçlar arasında en çok tercih edilenlerdir. Rapid Miner 22'ye yakın dosya formatını desteklerken, WEKA'nın desteklediği dosya formatı sayısı 4 ile sınırlıdır. Çoğu Veri Madenciliği uygulamasını geliştirmede bahsi geçen Rapid Miner, Knime veya Weka tek başına yeterli olmaktadır. Metin Madenciliği işlemlerini Keel, Orange, RapidMiner ve WEKA kendi başlarına yapabilirlerken; Knime bir modül sayesinde yapabilmektedir. Yapılan denemelerde, Knime ve Rapid Miner veri görselleştirme açısından diğer yazılımlardan daha gelişmiş bir ara yüz sunmuştur. Weka, doğrudan .jar uzantılı dosyaları kütüphane olarak sunması veya komut satırı arayüzü ile işlemlere izin vermesi açısından diğer yazılımlardan farklılaşır. Orange, Python diliyle yazılmasından dolayı yazılım içerisinde script yazmaya imkan vermektedir ve benzer özellikler Knime veya Rapid Miner içinde Java veya Python dillerinde snippet yazılarak sunulmaktadır. Knime ve Orange doğrudan Excel dosyalarıyla çalışamamaktadır. Karşılaştırılan yazılımlar arasında kurulum şartı olmadan çalışabilen tek yazılım Keel'dir. Diğer yazılımların öncelikle bilgisayara kurulması gerekmektedir[10]. Yazılımların ortak iletişim dili olarak kullandıkları ve makine öğrenmesi algoritmalarını iletmeye yarayan, tahminci model işaretleme dili (predictive model markup language, PMML) [16] , kurulan ve denemesi yapılan araçların tamamı tarafından desteklenmektedir.

Tablo I. Açık Kaynaklı Veri Madenciliği Yazılımlarının Karşılaştırılması

	WEKA	**Rapid Miner**	**KNIME**	**KEEL**	**ORANGE**	**MOA**	**ELKI**
PMML	Var	Var	Var	Var	Var	Var	Var
Veri Görselleştirme	Var	Var	Var	Var	Var	Var	Var
Veri Önişleme	Var	Var	Var	Var	Var	Var	Var
Algoritma Sayısı	76	129	102	393	13	-	147
Veri Analizi	Var	Var	Var	Var	Var	Var	Var
Excel Desteği	Evet	Evet	Hayır	Evet	Hayır	-	Evet
Yazıldığı Dil	Java	Java	Java	Java	Python, C++	Java	Java

Kullanılan versiyonlar ve araçların web siteleri:

1. Weka : Weka-3-8-1jre-x64.exe , http://www.cs.waikato.ac.nz/ml/weka/
2. Rapid Miner: Rapid Miner Studio 7.6 , https://rapidminer.com/
3. Knime: Knime 3.4.0 , https://www.knime.com/
4. Keel: Keel version(2015-03-23), http://www.keel.es/
5. Orange: Orange 3.4 , https://orange.biolab.si/
6. Moa: Moa Release 2016.04 , https://moa.cms.waikato.ac.nz/
7. Elki: Elki 0.7.1(2016.02.11) , https://elki-project.github.io/

Kaynakça:

[1] İnternet:"Veri Madenciliği" https://tr.wikipedia.org/wiki/Veri_madencili%C4%9Fi (2017)
[2] Han, J. ve Kamber M., "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2001.(2017)
[3] 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(2017)
[4] Kaya, Mümine, and Selma Ayşe Özel. "Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması." Akademik Bilişim (2017)
[5] İnternet:"WEKA" http://www.cs.wakikato.ac.nz/ml/weka/ (2017)
[6] İnternet:"Rapid Miner" https://tr.wikipedia.org/wiki/RapidMiner (2017)
[7] Markus Hofmann, Ralf Klinkenberg, "RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)(2017)
[8] İnternet:"Knime" https://en.wikipedia.org/wiki/KNIME (2017)
[9] Bilgin, T.T., "Veri Akışı Diyagramları Tabanlı Veri Madenciliği Araçları ve Yazılım Geliştirme Ortamları", Akademik Bilişim'09 - XI. Akademik Bilişim Konferansı Bildirileri, Şanlıurfa, 807-814 (2009)(2017)
[10] İnternet:"ORANGE", http://orange.biolab.si/(2017)
[11] Bifet, Albert; Holmes, Geoff; Kirkby, Richard; Pfahringer, Bernhard (2010). "MOA: Massive online analysis". The Journal of Machine Learning Research. 99: 1601–1604.(2017)
[12] Şeker, Şadi Evren. "Temporal logic extension for self-referring, nonexistence, multiple recurrence, and anterior past events." Turkish Journal of Electrical Engineering & Computer Sciences 23.1 (2015): 212-230.
[13] Seker, Sadi Evren, Khaled Al-Naami, and Latifur Khan. "Author attribution on streaming data." Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on. IEEE, 2013.
[14] Bifet, Albert, et al. "Moa: Massive online analysis." Journal of Machine Learning Research 11.May (2010): 1601-1604.(2017)
[15] İnternet:"ELKI", https://en.wikipedia.org/wiki/ELKI(2017)
[16] Guazzelli, Alex, et al. "PMML: An open standard for sharing models." The R Journal 1.1 (2009): 60-65.