YATIRIM TEŞVİK VERİLERİNİN VERİ MADENCİLİĞİ İLE ANALİZİ

Doç. Dr. Mehmet Ali ALAN[1] & Sinan DÜNDAR[2]

Öz

Bu çalışmada, yatırım teşvik verilerinden yararlanılarak veri madenciliği yapılmıştır. Bu verilerle yapılan çalışmada, hem bu verileri en başarılı sınıflandıran algoritma, hem de bu algoritmanın ürettiği sınıflar belirlenmeye çalışılmıştır. Çalışmanın sonucunda BFTree algoritmasının yatırım teşviki verilerini sınıflandırmada en başarılı algoritma olduğu belirlenmiştir. Ayrıca aynı algoriymayla elde edilen sonuçlara göre indirimlerin teşviklerden yararlanmada daha belirleyici olduğu ortaya konmuştur.

Anahtar Kelimeler: Sınıflandırma, Karar Ağaçları, BFTree, Yatırım Teşvik

Analysis of Investment Incentive Data with Data Mining

Abstract

In this study, data mining using datum of investment incentive was conducted. Both the most successful classification algorithm and classes generated via this algorithm were tried to be determined in this study. BFTree algorithm came out as the best classification algorithm in consequence of the study. According to the results derived via the same algorithm, it was presented that discounts are more determinative in case of incentive utilisation.

Keywords: Classification, Decision Trees, BFTree, Investment Incentive

1 Cumhuriyet Üniversitesi, İİBF, Yönetim Bilişim Sistemleri Bölümü. alan@cumhuriyet.edu.tr
2 Doktora Öğrencisi. Cumhuriyet Üniversitesi, Sosyal Bilimler Enstitüsü. sinandundar@hotmail.com

Giriş

Veritabanları, rasyonel karar almayı sağlayacak gizli bilgiler bakımından zengindir. Sınıflandırma ve tahmin, gelecek veri trendlerinin tahmini veya önemli veri sınıflarının açıklanmasında kullanılan iki önemli veri analiz tekniğidir. Bu analizler büyük miktarlardaki verilerin daha iyi anlaşılmasında kullanışlı olabilmektedir (Han and Kamber, 2006:285).

Veritabanlarındaki veriler satandart kullanım amaçları dışında, kurumların işine yarayacak bilgiler ya da ilişkiler barındarıbilir. Bu yararlı bilgileri ortaya çıkaran en önemli disiplinlerin başında veri madenciliği gelmektdir.

Günümüzde kurumlar büyük miktarlarda veri üretmekte, ancak bu veriler içinde anlamlı ve yararlı bilgiyi ortaya çıkarmakta zorluklar yaşamaktadırlar. Geleneksel istatistik yöntemlerle büyük boyuttaki veriyi çözümlemek kolay değildir. Bu nedenle verileri işlemek ve çözümlemek için özel yöntemlere gereksinim duyulmuştur. Veri madenciliği yöntemleri bu gereksinimi karşılamak üzere ortaya çıkmıştır (Özkan, 2008:IV).

Bu çalışmanın amacı, veri madenciliği tekniğini kullanarak, yatırım teşvik verileri yardımıyla sınıflandırma analizi yapmaktır. Bu amaçla mevcut veriler ele alınarak, veri madenciliğinin en yaygın kullanılan tekniklerinden "sınıflandırma" yöntemi kullanılmıştır.

Yatırım teşvik sisteminin amacı, Kalkınma Planları ve Yıllık Programlarda öngörülen hedefler ile uluslararası anlaşmalara uygun olarak, tasarrufları katma değeri yüksek yatırımlara yönlendirmek, üretimi ve istihdamı artırmak, yatırım eğiliminin devamlılığını ve sürdürülebilir kalkınmayı sağlamak, uluslararası rekabet gücünü artıracak teknoloji ve araştırma-geliştirme içeriği yüksek büyük ölçekli yatırımları özendirmek, doğrudan yabancı yatırımları artırmak, bölgesel gelişmişlik farklılıklarını gidermek, çevre korumaya yönelik yatırımlar ile araştırma ve geliştirme faaliyetlerini desteklemektir http://www.ekonomi.gov.tr/portal/content/conn/UCM/uuid/dDocName:EK-107151, 28.02.2017).

Çalışma üç bölümden oluşmaktadır. Birinci bölümde veri madenciliği, sınıflandırma, karar ağaçları ve BFTree algoritmasıyla ilgili tanımlayıcı açıklamalara yer verilmiştir. İkinci bölümde konuyla ilgili yapılmış çalışmalar literatür bazlı değerlendirilmiş ve nihayet üçüncü bölümde ise yatırım teşvik verileri üzerinde veri madenciliği yapılmıştır.

Veri Madenciliği, Sınıflandırma, Karar Ağaçları ve BFTree Algoritması

Veri madenciliği, hem yararlı hem de anlaşılabilir verilerle, alışılmamış yollarla, verileri özetleyen ve gizli ilişkileri ortaya koyan bir analiz yöntemidir (Larose, 2006). Bu yöntem, öncelikle bilinmeyen desenlerin ortaya konması amacıyla bilimsel ve tek-

nik veri araştıran, veritabanındaki bilgi keşfi süreçlerinden biridir (Rokach ve Maimon, 2005:2).

Disiplinler arası nitelik taşıyan veri madenciliğini en yaygın kullanan bilim dalları; veritabanı sistemleri, istatistik, matematik, makine öğrenmesi, görselleme ve bilişim bilimleridir (Han ve Kamber, 2006:29). Veri madenciliği, verinin bütününü kullanması bakımından diğer istatistiksel verilerden ayrılmaktadır. Bu yöntemle, geleneksel yollarla elde edilmiş küçük verilerle çalışma yerine daha kolay değerlendirme yapabilecek, yeni bağımsız veriler tercih edilebilmektedir (Weiss ve Zhang, 2003:426).

Veri madenciliği, reklamcılık, biyoinformatik, veritabanı pazarlama, dolandırıcılık tespiti, e-ticaret, sağlık, güvenlik gibi alanların yanın da farklı alanlarda da uygulanabilen, değişik bakış açısı ve çalışması ile veri analizinden bilgi keşfetme süreci olarak bilinir(Jain, 2011).

Birliktelik kuralları, "kümeleme, karar ağaçları, diskriminant analizi, yapay sinir ağları, genetik algoritmalar" vb. birçok veri madenciliği algoritmasını içerir. Bu algoritmalar sıradan bilgiyi bulup çıkarmak ve bir yöneticinin kararlarını yönlendirebilen özel bilgiye ulaşmak için çeşitli alanlardan elde edilen verileri işlemek amacıyla kullanılır (Wu ve Li, 2003).

En yaygın veri madenciliği algoritmaları ve modelleri içinde karar ağaçları, sınıflandırma ağaçları olarak ta adlandırılır (Bramer, 2007:6); birliktelik kuralları, kümeleme, sınıflandırma, çoklu lineer regresyon, sıralı örüntüler ve zaman serileri tahmini, örüntü tanıma ve özelliklerinin belirlenmesi sayılabilir. Sınıflandırma, regresyon ve zaman serisi analizleri gizli örüntülerin ortaya çıkarılmasında ve şekillendirilmesinde uygun iken, birliktelik kuralları, kümeleme ve sırasal keşif yaklaşımları, hava tahmini ve şiddeti araştırmak ve tanımlamak için yararlı araçlar olabilir (Tadesse 2009).

Sınıflandırma, günlük yaşamda çok sıklıkla başvurulan bir işlemdir. Sınıflandırma ile nesneler bölünerek ayrıştırılır, yani karşılıklı olarak özel ya da genel kategorilerden her biri bir sınıf olarak atanabilir. Pek çok pratik karar verme işlemi, bir sınıflandırma problemi olarak formüle edilebilir. Örneğin kişiler ya da nesneler birçok kategoriden biri olabilir (Bramer, 2007:23).

Sınıflandırma, farklı sınıflardaki, değişik öğeleri ayırma sürecidir. Bu sınıflar, iş kuralları, sınıf sınırları veya bazı matematiksel fonksiyonlar olabilir. Sınıflandırma işlemi, sınıflandırılmış olan ögenin, bilinen bir sınıf değeri ile özellikleri arasındaki bir ilişki üzerine bina edilebilir. Bu sınıflandırma tipi, "denetimli öğrenme" olarak isimlendirilir. Eğer bir sınıfın bilinen örnekleri yoksa bu sınıflandırma denetimsizdir. En yaygın denetimsiz sınıflandırma yaklaşımı "kümeleme"dir. Kümeleme teknolojisinin en yaygın uygulamaları, perakende ürünlerde birliktelik analizi (market sepet analizi) ve dolandırıcılık tespitidir (Nisbet, vd., 2009: 235).

Veri madenciliğinde denetimli öğrenme kavramı, bir sınıflandırma ile bilinen veriler temelinde bir sınıflandırma fonksiyonu öğretmek ya da bir sınıflandırma modeli inşa etmektir. Bu fonksiyon ya da model, veri tabanındaki verileri hedef niteliklere dönüştürür, dolayısıyla yeni veriler sınıf tahmininde kullanılabilir (Dong-Peng, vd.,2008:36).

Karar ağaçları, sınıflandırma ve tahmin açısından güçlü ve popüler araçlardır. Bu yöntemin çekici tarafı, yapay sinir ağlarının aksine, karar ağaçlarının kuralları temsil etmesidir. Başka bir deyişle bunları yorumlamak daha kolaydır (Nisbet, vd., 2009:465).

Veri madenciliğinde bir karar ağacı, hem sınıflandırıcıları, hem de regresyon modellerini temsil edecek şekilde kullanılabilir. Diğer yandan operasyonel araştırmalarda karar ağaçları, hiyerarşik kararlar modeli ve onların sonuçları ile ilişkili bilgiler ortaya koyar. Karar vericiler olması en fazla muhtemel hedefine ulaşacak şekilde strateji geliştirmek için karar ağaçlarını kullanır. Bir karar ağacı sınıflandırma amacıyla kullanıldığı zaman, genellikle "sınıflandırma ağacı", regresyon amacıyla kullanıldığında ise "regresyon ağacı" olarak adlandırılır (Rokach ve Maimon:2008:5).

Karar ağaçları hâlihazırdaki kullanımıyla en popüler tüme varım metodudur. Karar ağaçları genellikle iki aşamada oluşturulurlar. Büyüme olduğu zaman bu algoritma her bir düğümde sınıflar arasındaki en iyi özellik ayırt ediciyi (veri alt seti) ortaya çıkarır ve daha sonra o özelliğe dayalı olarak bu verileri iki yeni düğüm halinde bölümlere ayırır. Bu, her bir tabaka için bir sınıf tahsis edilinceye kadar ortaya çıkan veri alt setine tekrar tekrar uygulanır. Budamanın ikinci aşaması en iyi dengeye erişebilmek için ağacın en az yararlı dallarını kesmek suretiyle işletilir. Daha basit bir model genellikle daha sağlamdır. Yani yeni veriler hakkında daha doğru sonuçlar ortaya çıkarır. Nihai ağaç etiketlenmiş birkaç bölgedeki bu özel alanı bölümlere ayırır (Nisbet, vd., 2009:300).

Karar ağacı yöntemlerinde parametrik istatistiksel varsayımlar yapılmaz. Öngörüler terminal düğümlerde birkaç mantıksal "if-then" şartıyla sunulabilirler. Normal bir veri dağılımında veya değişkenler ve tepki değişkeni arasındaki lineer ilişkilerde örtülü varsayımlar yoktur. Karar ağacı yöntemleri, değişkenlerin öngörü sağlayabildikleri zamanın ötesini analistlerin bilmedikleri yerlerde veri madenciliği için oldukça uygundurlar. Bu nedenle karar ağacı yöntemleri ilişkileri açığa çıkarabilir ve onları daha fazla bilişimsel yoğun yöntemlerin gözden kaçırdığı birkaç karar kuralı halinde ifadelendirebilirler (Nisbet, vd.,. 2009:278-279).

Bir karar ağacı, öz nitelikler değeri üzerine bölünme olarak bilinen bir işlem tarafından oluşturulur yani outlook gibi bir öznitelik testi ve ardından olası değerlerin her biri için bir dal oluşturmadır. Sürekli öznitelikli test durumunda normal olarak değer "daha az ya da eşit" ya da "daha büyük" bölünmüş değer olarak bilinen bir değer verir. Her bir dal, sadece bir sınıflandırma ile etiketli oluncaya kadar bölünme işlemi devam eder (Bramer, 2007:43-44).

Çekici bir sınıflandırma yöntemi olan karar ağacı, kök düğümden aşağı doğru yaprak düğümlerinde sonlanıncaya kadar uzayan, dallar tarafından bağlantıları sağlanmış, karar düğümlerinin bir koleksiyonunu içerir. Karar ağacı diyagramının en üstüne yerleştirilmiş olan kök düğümden başlayarak, dalın her bir olası sonucu ile karar düğümleri test edilir. Her bir dal daha sonra diğer bir karar düğmesi ya da sonlandırma yaprak düğmesine yol açar (Larose, 2005:107). Karar ağaçlarının popüler agoritmalarından birisi de BFTree (En iyi - ilk Karar Ağacı- Best-first Decision Tree) algoritmasıdır. BFTree algoritması, böl ve yönet mantığına dayanır. Öncelikle bir nitelik kök olarak alınır ve bu nitelik üzerinden bazı kriterlere göre gruplara ayrılır. Daha sonra kök düğümden her grup genişletilerek eğitim verileri alt gruplara bölünür. Sadece karşılayan verileri kullanarak seçilen her grup için bu işlem tekrarlanır. Her işlemde, genişlemeye en uygun iyi alt grup seçilir. Bu süreç belirli bir genişleme katsayısına göre tüm düğümler net olana kadar devam eder (Akçetin ve Çelik, 2014).

Literatür Özeti

Konuyla ilgili literatürde farklı veri setleri üzerinden yapılmış çok sayıda çalışma bulunmaktadır. Bunlardan Dota, vd., (2015), topraktan elde edilmiş kirli su verilerini kullanarak sınıflandırma yapmış ve mevcut verileri en iyi sınıflandıran algoritmanın BFTree algoritması olduğunu tespit etmiştir. Kumar vd., (2013), toprak ve arazi verilerini kullanarak BFTree algoritmasıyla sınıflandırma yapmışlar ve uygun arazi kullanımı, toprak ve su koruma uygulamalarında karar ağaçlarının önemli bir model olduğunu ortaya koymuşlardır. Aksu ve Güzeller (2016), Türkiye'deki Uluslararası Öğrenci Değerlendirme Programına (PISA) katılan öğrencilere ait verileri, WEKA programı kullanılarak karar ağaçlarıyla sınıflandırmışlar ve bulunan sonuçlar başka yöntemlerle kıyaslanarak, buluntuların başarısına vurgu yapmışlardır. Hota ve Dewangan (2016), UCI depo sisteminden indirdikleri kalp verilerini kullanarak, çeşitli makine öğrenme teknikleriyle sınıflandırma yapmışlar ve en iyi sonucu veren algoritmanın CART olduğunu belirlemişlerdir. Sulaiman vd., (2015), öğrenci veri setindeki çeşitli ders verileri kullanılarak, bir dersin başarısına diğer derslerin etki edip etmediğini araştırmak için, karar ağaçlarına ait J48, Cart ve BFTree algoritmalarıyla sınıflandırma analizleri yapmışlar ve en iyi performans gösteren algoritmanın J48 algoritması olduğunu ortaya koymuşlardır. Thepade vd., (2015), multimedya verilerini, WEKA programınca desteklenen farklı ailelere ait 12 algoritma ile test etmiş ve basit lojistik sınıflayıcının en iyi performanslı sınıflayıcı olduğunu tespit etmişlerdir. Ma vd., (2008), internet ağ trafiği verilerini kullanarak çeşitli algoritmalarla sınıflandırmalar yapmışlar ve C4.5 algoritmasını ağ trafiğini en iyi tanımlayan algoritma olarak bulmuşlardır. Ayık vd. (2007), Atatürk Üniversitesi öğrencilerine ait veritabanındaki tüm verileri kullanarak sınıflandırma analizi yapmışlardır. Kaya vd., (2012), Epileptik EEG işaretlerini karar ağaçları ve karar kurallarını kullanarak sınıflandırmış ve tanı performanslarının oldukça yüksek olduğunu tespit etmişlerdir.

Veri Seti ve Yöntem

Bu çalışmada Yatırım Teşviklerinde bölgesel teşvik uygulamalarına ilişkin 2009-2016 yılları arası 32.023 işletmeye ait veriler kullanılmıştır. Analizde uygulamaya temel teşkil edecek, sınıflarla birlikte toplam 9 sütun başlığı tanımlanmıştır. Excel makroları kullanılarak veri madenciliğinde kullanılan standart algoritmaların yanı sıra bulanık mantık ve genetik algoritma analizleri yapılacak şekilde veri ambarı hazırlanmıştır. Gerekli dönüşümler yapıldıktan sonra veriler "veriset.arff" adlı metin dosyasına yazdırılmıştır.

Verilerde; sermaye türünde yerli için 1, yabancı için 2, sektörde; imalat için 1, tarım için 2, hizmet için 3, madencilik için 4 ve diğer sektör için 5 değeri, yatırım cinsinde tevsi için 1, yeni yatırım için 2 ve modernizasyon için 3 değeri atanmıştır. Gelir vergisi stopaj desteğinde muaf olanlar için 1, muaf olmayanlar için 2 değeri girilmiştir. Sigorta Primi İşveren Hissesi Desteği; var olanlar için 1, olmayanlar için 2, vergi indiriminde; var değeri için 1, yok değeri için 2, faiz desteğinde de var değeri için 1, yok değeri için 2 değeri atanmıştır. İstihdamda işletmede çalışan personele göre, büyük için 1, küçük için 2, mikro için 3 ve orta için 4 değeri tanımlanmıştır. Sınıflar ise uygun için 1 ve uygun değil için 2 değerini almıştır.

Uygulama

Yapılan çalışmada Waikato Üniversitesince geliştirilmiş olan WEKA Programının (Waikato Environment for Knowledge Analysis) 3.7.2 sürümü kullanılmıştır (Wekafull paket). WEKA Programı, açık kaynak kodlu bir yazılımdır. Bu program pek çok sınıflandırma, kümeleme ve birliktelik kurallarına ait algoritmayı desteklemektedir. WEKA, metin tabanlı pek çok dosya tiplerinin yanı sıra, veritabanlarını ve verilerin olduğu URL adreslerini de desteklemektedir. WEKA programının bu sürümünde, bulanık mantık ve genetik algoritma desteği de bulunmaktadır.

Mevcut veri seti ile yapılan 10 kat çapraz doğrulama ve tam eğitimli set kullanılarak yapılan uygulama sonucunda izleyen tablodaki sonuçlar algoritmaların başarım derecelerine göre sıralanarak sunulmuştur:

Tablo 1

WEKA ile Elde Edilen Analiz Sonuçları

Algoritmalar	Doğru sınıflandırılan Örnek	Kappa İstatistiği	Ortalama Mutlak Hata	Ortalama Hata Karekök	Göreli Mutlak Hata %	Göreli Hata Karekök %	TP Oranı	FP Oranı	F-Ölçütü
BFTree	25519	0.3074	0.2813	0.3789	80.617	90.7028	0.797	0.539	0.774
J48	25511	0.3058	0.2872	0.3801	82.294	90.9894	0.797	0.54	0.773
SimpleCART	25495	0.3043	0.2846	0.379	81.5516	90.7438	0.796	0.541	0.773
JRip	25480	0.3324	0.3063	0.392	87.7636	93.8513	0.796	0.504	0.779
RandomForest	25480	0.314	0.2791	0.379	79.9862	90.7385	0.796	0.528	0.775
Fuzzy.FURIA	25446	0.3411	0.2059	0.4369	59.0049	104.599	0.795	0.49	0.781
Kstar	25377	0.181	0.2884	0.378	82.6534	90.4974	0.792	0.661	0.738
Ridor	25345	0.2356	0.2085	0.4566	59.7513	109.319	0.791	0.607	0.754
ADTree	25335	0.2763	0.367	0.3984	105.1658	95.3886	0.791	0.563	0.764
NaiveBayes	25202	0.3295	0.2898	0.3836	83.0387	91.8296	0.787	0.487	0.775
NBTree	25202	0.3295	0.2898	0.3836	83.0387	91.8296	0.787	0.487	0.775
Fuzzy.OWENN	25200	0.2985	0.2799	0.3965	80.2005	94.9282	0.787	0.527	0.769
LADTree	25200	0.2036	0.2883	0.3794	82.6243	90.8264	0.787	0.631	0.744
Fuzzy.VQNN	25176	0.3364	0.2339	0.4043	67.0298	96.7846	0.786	0.476	0.776
Genetic Programmaning	24890	0.1808	0.2227	0.4719	63.823	112.982	0.777	0.636	0.736
OneR	24775	0.0593	0.2263	0.4757	64.8521	113.889	0.774	0.733	0.698

Yapılan uygulama çalışmasında WEKA programınca desteklenen pek çok algoritma denenmiş, başarım dereceleri Tablo 1'de verilmiştir. Bu çalışmada karar ağaçları, karar kuralları, Bayes, bulanık mantık ve genetik algoritma gibi çeşitli sınıflara ait algoritmalar kullanılarak modeller oluşturulmuş ve oluşturulan modellerin başarım dereceleri karşılaştırılmıştır.

BFTree algoritması, 25519 doğru sınıflandırılmış örnek derecesiyle en başarılı algoritma olmuştur. Bu algoritmanın, sınıflar arası uyumu veren kappa istatistiği 0.3074, birinci sınıftaki doğru olarak sınıflandırılmış kayıtların sayını veren TP (True Positive) oranı 0.797, birinci sınıfta sınıflandırılmış, ikinci sınıftaki kayıtların sayısını veren FP (False Positive) oranı 0.539 olarak elde edilmiştir. Kesinlik ve duyarlılığın harmonik ortalaması olan F-ölçütü (Coşkun ve Baykal, 2011) ise 0.774 olarak bulunmuştur. BFTree algoritmasından sonra en başarılı algoritma ise 25511 doğru örnek sınıflandırmasıyla J48 algoritması olmuştur. Bu algoritmaların kappa istatistiği 0.3058, TP oranı 0.795, FP oranı, 0.54 ve F-Ölçütü 0.773 olarak bulunmuştur. Bulanık mantık sınıflandırma grubundan en başarılı algoritma 25446 doğru sınıflandırma oranıyla FURIA (Fuzzy Unordered Rule Induction Algorithm) algoritmasıdır. Bu algoritmanın kappa istatistiği 0.3411, TP oranı 0.795, FP oranı 0.49 ve F-Ölçütü 0.781 olarak hesaplanmıştır. Genetik Programlama ile doğru sınıflandırılan örnek sayısı 24890'dır. Bu sınıflandırıcının kappa istatistiği 0.1808, TP oranı 0.777, FP oranı 0.636 ve F-ölçütü 0.736'dır. Diğer algoritmaların başarım dereceleri ise tablodaki gibidir. F-ölçütü aşağıdaki formülle elde edilmektedir:

$$\text{F - Ölçütü} = \frac{2\text{xDuyarlılıkxKesinlik}}{\text{Duyarlılık} + \text{Kesinlik}}$$

Formülde kullanılan kesinlik ve duyarlık ise aşağıdaki şekilde hesaplanmaktadır:

$$\text{Kesinlik} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Duyarlılık} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Sınıflandırma başarısı en yüksek olarak bulunan BFTree algoritmasıyla üretilen ağacın büyüklüğü (Size of the Tree) 929 ve yaprak düğüm sayısı (Number of Leaf Nodes) 465'tir. Bu büyüklükteki bir ağacı grafik olarak sunmak pek mümkün olmadığından ancak ağacın baş kısmı izleyen şekilde sunulmuştur:

```
SermayeT=(1)
|  Sektoru=(1)
|  |  istihdam=(1)|(3)
|  |  |  Cinsi=(2): 2(2180.0/55.0)
|  |  |  Cinsi!=(2)
|  |  |  |  Cinsi=(1)|(2)
|  |  |  |  |  GVMuafiyet=(2): 2(10.0/0.0)
|  |  |  |  |  GVMuafiyet!=(2)
|  |  |  |  |  |  istihdam=(1): 2(2.0/0.0)
|  |  |  |  |  |  istihdam!=(1): 2(51.0/12.0)
|  |  |  |  Cinsi!=(1)|(2)
|  |  |  |  |  vergiindirimi=(1): 2(3.0/0.0)
|  |  |  |  |  vergiindirimi!=(1)
|  |  |  |  |  |  GVMuafiyet=(1): 2(26.0/12.0)
|  |  |  |  |  |  GVMuafiyet!=(1): 2(7.0/4.0)
|  |  istihdam!=(1)|(3)
|  |  |  Cinsi=(1)|(2)
|  |  |  |  Cinsi=(1)
|  |  |  |  |  istihdam=(4): 2(13.0/0.0)
|  |  |  |  |  istihdam!=(4)
|  |  |  |  |  |  SPDestegi=(1): 2(2.0/0.0)
|  |  |  |  |  |  SPDestegi!=(1): 2(26.0/4.0)
|  |  |  |  Cinsi!=(1)
```

Şekil 1. BFTree Algoritmasının Ürettiği Bazı Sınıflar

Ayrıca Tablo 2'de yer aldığı biçimiyle veriler iki ayrı gruba ayrılarak yatırım teşvikinde yararlanmada hangi grubun daha belirleyici olduğu araştırılmıştır. Yapılan analize göre gelir vergisi stopajı desteği, faiz indirimi, sigorta primi desteği ve vergi indiriminin yer aldığı grupta doğru sınıflandırılan örnek, kappa istatistiği, TP, FP

oranları ve F-ölçütünün daha yüksek, sermaye türü, sektörü, yatırım cinsi ve istihdamın yer aldığı ikinci grupta ise daha düşük olduğu anlaşılmaktadır.

Tablo 2

Gruplara göre Analiz Sonuçları

	Doğru sınıflandırılan Örnek	Kappa İstatistiği	TP Oranı	FP Oranı	F Ölçütü
Gelir Vergisi Stopajı Desteği, Faiz İndirimi, Sigorta Primi Desteği ve Vergi İndirimi	27130	0.6768	0.847	0.086	0.853
Sermaye Türü, Sektörü, Yatırım Cinsi ve İstihdam	19024	0.2331	0.594	0.385	0.523

Sonuç

Bu çalışmada, veri madenciliği ile yatırım teşvik verileri üzerinde analizler yapılmıştır. Veri madenciliği, gizli, önemli, önceden bilinmeyen, yararlı örüntüleri ortaya çıkaran bir analiz tekniğidir. Veri madenciliğinde standart algoritmaların yanı sıra yapay sinir ağları, bulanık mantık ve genetik algoritmalar gibi yapay zekâ yöntemleri de kullanılabilmektedir. Bu çalışmada da WEKA programınca desteklenen pek çok algoritma denenmiş ve yatırım teşvik verilerini en başarılı sınıflandıran algoritma olarak BFTree algoritması tespit edilmiştir. Ayrıca bu çalışmada yatırım teşviklerinden yararlanmada gelir vergisi stopajı desteği, faiz indirimi, sigorta primi desteği ve vergi indiriminden oluşan indirim grubunun, sermaye türü, sektörü, cinsi ve istihdam oluşan gruba göre daha belirleyici olduğu anlaşılmaktadır.

Kaynakça

Akçetin, E. ve Çelik, U. (2014). İstenmeyen Elektronik Posta (Spam) Tespitinde Karar Ağacı Algoritmalarının Performans Kıyaslaması, Internet Uygulamaları ve Yönetimi Dergisi, 214/5(2), s.43-56

Aksu, G., Güzeller, C. O. (2016). PISA 2012 Matematik Okuryazarlığı Puanlarının Karar Ağacı Yöntemiyle Sınıflandırılması: Türkiye Örneklemi, Eğitim ve Bilim, Cilt 41, Sayı: 185, s. 101-122

Ayık, Y. Z., Özdemir, A. ve Yavuz U. (2007). Lise Türü Ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi,. Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, Cilt 10, Sayı 2, s.441–454

Baykal, A. ve Coşkun C. (2011). Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması, http://ab.org.tr/ab11/bildiri/67.pdf, 18.01.2017

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

Dong-Peng, Y., Li, J., Lun, R. and Chao, Z. (2008). Applications of Data Mining Methods in the Evaluation of Client Credibility, *Applications of Data Mining in E-Business and Finance C.* Soares et al. (Eds.), IOS Press, Amsterdam, p.35-43

Dota, M. A, Cugnasca, C, E. and Domingos, S. B. (2015). Comparative analysis of decision tree algorithms on quality of water contaminated with soil, Ciência Rural, Santa Maria, v.45, n.2, p.267-273

Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques,

Hota, H.S. and Dewangan, S. (2016). Classification of Health Care Data Using Machine Learning Technique, International Journal of Engineering Science Invention, Volume 5, Issue 9,September 2016, p. 17-20 Inc., Publication, New Jersey

Jain, Y. K., Yadav, V. K. and Panday, G. S., (2011), "An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining", *International Journal On Computer Science And Engineering*, Vol. 3 No. 7, p. 2792-2798. Jersey.

Kaya, Y., Ertuğrul, Ö. F. ve Tekin R. (2012). Batman University International participated Science and Culture Symposium, Batman University Journal of Life Sciences, Volume 1, Number 2, p.403-413

Kumar, N., Reddy, G. P. O. and Chatterji, S. (2013). Evaluation of Best First Decision Tree on Categorical Soil Survey Data for Land Capability Classification, International Journal of Computer Applications (0975 – 8887), Volume 72– No.4, June 2013,p. 9-12

Larose, D. T. (2005). Discovering Knowledge In Data, Wiley Publication, New

Larose, D. T. (2006). Data Mining Methods and Models, A John Wiley & Sons,

Ma, Y., Qian, Z., Shou, G. and Hu, Y. (2008). Study on Preliminary Performance of Algorithms for Network Traffic Identification, 2008 International Conference on Computer Science and Software Engineering, p.629-633

Nisbet, R., Elder, J., and Miner, G. (2009). Handbook of Statistical Analysis and Data Mining Applications, Elsevier Inc, Burlington.

Özkan, Y. (2008). Veri Madenciliği Yöntemleri, Papatya Yayınları, İstanbul

Rokach, L. and Maimon, O.(2008), Data Mining with Decision Trees, World Scientific, New Jersey

Second Edition, Morgan Kaufmann Publications, San Francisco

Sulaiman ,A., Akinbowale, B., Moshood, H. A. and Ronke, B. (2015). Comparative Analysis of Decision Tree Algorithms for Predicting Undergraduate Students' Performance in Computer Programming, A Multidisciplinary Journal Publication of the Faculty of Science, Adeleke University, Ede, Nigeria, 2015 Edition Vol 2., p.79-92

Tadesse, T., Wardlow, B. And Hayes, M.J. (2009). The Application of Data Mining for Drought Monitoring and Prediction, Data Mining Applications for Empowering Knowledge Societies, Edited by Hakikur Rahman, Information Science Reference, New York, p.280-291

Thepade, S. D. and Kalbhor, M. M. (2015). Image Cataloging using Bayes, Function, Lazy, Rule, Tree classifier Families with Row mean of Fourier Transformed Image Content, 2015 International Conference on Information Processing (ICIP) Vishwakarma Institute of Technology., p.680-684

Weiss, S. M. And Zhang, T. (2003). Performance Analysis and Evaluation, The Handbook of Data Mining, Edited by. Nong Ye, Lawrence Erlbaum Associates Publishers. London, pp.436-439

Wu, T. and Li, X. (2003). Data Storage and Management, The Handbook of Data Mining, *Edited by.* Nong Ye, Lawrence Erlbaum Associates Publishers. London, p.393-407

http://www.ekonomi.gov.tr/portal/content/conn/UCM/uuid/dDocName:EK-107151, 28.02.2017).