Şirket İflaslarının Tahmin Edilmesinde Karar Ağacı Algoritmalarının Karşılaştırmalı Başarım Analizi

Aytuğ ONAN

Celal Bayar Üniversitesi, Bilgisayar Mühendisliği Bölümü
aytugonan@gmail.com
(Geliş/Received: 24.08.2014; Kabul/Accepted: 30.12.2014)
DOI: 10.17671/btd.36087

***Özet*—** Bu çalışmada, önemli bir ekonomik problem olan, şirket iflaslarının tahmin edilmesi ele alınmıştır. Bunun için iki yüz kırk farklı şirkete ilişkin finansal özellikleri içeren bir veri seti kullanılmıştır. Ele alınan veri seti, sınıflandırma ve tahmin etmede kullanılan önemli yöntemlerden biri olan karar ağacı yöntemine ilişkin yedi farklı algoritma uygulanarak, doğru sınıflandırma yüzdesi, ortalama mutlak hata, ortalama karesel hatanın karekökü, kesinlik, geri çağırma, F-ölçütü gibi ölçütler bakımından değerlendirilmiştir. Deneysel sonuçlar incelendiğinde, karar ağacı algoritmalarının şirket iflaslarının tahmin edilmesi için uygun bir yöntem olduğu ve kısmen başarılı doğru sınıflandırma yüzdesi elde ettiği gözlemlenmiştir.

***Anahtar Kelimeler*—** tahmin etme, karar ağacı algoritmaları, sınıflandırma, şirket iflasları

Comparative Performance Analysis of Decision Tree Algorithms in the Corporate Bankruptcy Prediction

***Abstract*—** In this study, corporate bankruptcy prediction, a crucial economic problem is tackled. To do this, a data set of 240 distinct companies with financial features is used. This data set is applied to one of the most important classification and forecasting methods, i.e. decision tree method. Seven different decision tree algorithms are evaluated in terms of accuracy percentage, mean absolute error, root mean squared error, precision, recall, F-measure. According to experimental results, decision tree algorithms are appropriate methods for corporate bankruptcy prediction with relatively successful accuracy rates.

***Keywords*—** forecasting, decision tree algorithms, classification, corporate bankruptcy.

1. GİRİŞ (INTRODUCTION)

Şirket iflaslarının tahmin edilmesi, başta bankalar, sigorta şirketleri ve yatırımcılar olmak üzere birçok farklı paydaşı ilgilendiren önemli bir ekonomik problemdir. Tahmin modellerinin oluşturulabilmesi için ilgili alanda yeterli verinin toplanması ve bu verilerin uygun algoritma ya da yöntemler aracılığı ile modelin oluşturulmasında kullanılması gerekmektedir. Günümüzde bilgi ve iletişim teknolojilerindeki ilerlemeler, depolanan ve işlenen veri miktarının önemli ölçüde artmasını olanaklı kılmaktadır. Buna paralel olarak, büyük miktarda verinin etkin bir biçimde ele alınabilmesi ve bu verilerden anlamlı ve kullanışlı bilgiler elde edilebilmesi gereksinimi, veri madenciliği yöntemlerini önemli bir konuma yerleştirmektedir. Veri madenciliği, büyük veri setlerinden, açık olarak bilinmeyen ve yararlı olması olası bilgilerin elde edilmesini amaçlayan, istatistik, matematik, bilgisayar bilimleri gibi birçok farklı alandan yöntem ve algoritmaların bir araya getirildiği disiplinler arası bir çalışma alanıdır. Veri madenciliği yöntemleri, birçok farklı alanda uygulama alanı bulmaktadır. Bu alanların başında işletme ve ekonomi alanları gelmektedir.

Şirketlerin ekonomik ve stratejik başarıları için veri madenciliği teknikleri stratejik bir öneme sahiptir. Veri madenciliği teknolojisi ile sağlanabilecek fırsatlar arasında, eğilim ve davranışların otomatik olarak tahmin edilmesi bulunmaktadır. Eğilim ve davranışların otomatik olarak tahmin edilmesi ile gerçekleştirilen uygulamalar arasında daha önceki kampanyalardan hareketle yatırım getirisini maksimize edecek hedef kitlenin belirlenmesi ve

şirket iflaslarının tahmin edilmesi yer almaktadır [1]. Bunun yanı sıra, satışlardan birbirleriyle ilgisiz görünen ancak birlikte satışı gerçekleştirilen kayıtlara ilişkin örüntülerin belirlenmesi ve kredi kartı hileciliğinin saptanması gibi uygulamalar da sıklıkla gerçekleştirilmektedir.

Veri madenciliği teknikleri temel olarak, veri madenciliğinin amacına dayalı olarak tahmin edici ve tanımlayıcı yöntemler olmak üzere iki sınıf altında incelenebilir [2]. Tahmin edici yöntemler, mevcut değişkenleri kullanarak, gelecekteki bilinmeyen değerleri sınıflandırma ya da regresyon gibi yaklaşımlar aracılığıyla belirlemektedir. Tanımlayıcı yöntemler ise veri setindeki desenleri ortaya çıkararak, verinin kullanıcı tarafından kolayca yorumlanabilmesini olanaklı kılmaktadır. Tanımlayıcı yöntemler içerisinde kümeleme analizi, birliktelik kuralları madenciliği, ardışıl örüntü madenciliği gibi yöntemler yer almaktadır [2].

Karar ağacı yöntemi, basit ve kolay anlaşılır yapısı sayesinde, sınıflandırma ve tahmin etme problemi için uygun bir yöntem olarak kullanılmaktadır. Bu çalışmanın temel amacı, başlıca karar ağacı algoritmalarının şirket iflaslarının tahmin edilmesinde performanslarının belirlenmesidir. Bu doğrultuda, karşılaştırmalı analizler başlıca karar ağacı algoritmalarından C4.5 [3], Decision Stump [4], Hoeffding Tree [5], LMT (Logistic Model Trees) [6], Random Forest [7], Random Tree [8] ve RepTree [9] yöntemleri kullanılarak gerçekleştirilmiştir. Çalışmanın ikinci bölümünde, şirket iflaslarının tahmin edilmesi için gerçekleştirilen başlıca akademik çalışmalar yer almaktadır. Üçüncü bölümde, karar ağacı yönteminin genel ilkeleri ve çalışmada kullanılan algoritmaların kısa açıklamalarına, dördüncü bölümde veri seti ve bulgulara, sonuç bölümünde ise çalışmanın genel değerlendirmesine yer verilmektedir.

2. LİTERATÜR (LITERATURE)

Şirket iflaslarının tahmin edilmesi, doğru stratejik kararlar alınabilmesini olanaklı kılması bakımından işletme, finans ve ekonomi alanlarında önemli bir konuma sahiptir. Şirket iflaslarının yüksek doğruluk başarısı ile sınıflandırılması, hissedarlar, kredi veren kurum ve kuruluşlar, politika yapıcılar, yöneticiler gibi önemli ekonomik paydaşlar için çok kritik bir önem taşımaktadır [10].

İflas tahmin edilmesinde kullanılan temel yöntemlerin başında, yapay sinir ağları, lojistik regresyon, karar ağacı sınıflandırıcıları gibi yöntemler gelmektedir. Bu bölümde, literatürde iflas tahmin etme alanında gerçekleştirilen başlıca akademik çalışmalara değinilmiştir.

Lam [11], geri yayılım algoritması kullanarak yapay sinir ağı ile finansal performansın tahmin edilmesi üzerinde çalışmıştır. Deneysel sonuçlar, yapay sinir ağlarının finansal performans tahmin etmede yüksek başarım oranı elde ettiğini göstermektedir. Yapay sinir ağına daha önceki yıllara ilişkin finansal veriler girdi vektörü olarak sunulduğunda, başarımın önemli ölçüde arttığı görülmüştür.

Aoki ve Hosonuma [12] çalışmalarında, CHAID karar ağacı algoritmasını kullanarak, 73'ü iflas etmiş ve 73'ü iflas etmemiş olan 146 Japon şirkete ilişkin bir sınıflandırma modeli oluşturmuştur. CHAID karar ağacı algoritması kullanıldığında, %91,2 doğru sınıflandırma oranı elde edildiği ve Japon şirketlerinin iflasına ilişkin veri seti için en önemli özniteliğin faiz karşılama gücü olduğu görülmüştür.

Santos vd. [13] tarafından yapılan diğer bir çalışmada, şirket iflaslarının tahmin edilmesi problemi için yapay sinir ağları ve karar ağaçları gibi on altı farklı veri madenciliği modelini değerlendiren bir çatı sunulmuş, modelde sunulan algoritmalar, eğitimde kullanılan stratejiler, öznitelik seçimleri gibi ölçütler bakımından farklılaştırılmış ve Portekiz'deki çeşitli şirketlerden toplanan gerçek veri seti aracılığıyla değerlendirilmiştir. Deneysel çalışmalar ile yöntemlerden %86-%99 aralığında doğru sınıflandırma oranlarının elde edilmesi, veri madenciliği yöntemlerinin, finansal başarım tahmin edilmesinde uygun bir araç olarak kullanılabileceğini vurgulamaktadır.

Neves ve Vieria [14], şirket iflaslarının tahmin edilmesinde, gizli katman öğrenme vektör nicelendirmesi algoritmasını kullanmıştır. Çok katmanlı algılayıcının çıktıları düzeltilmiş ve yöntem, şirket iflaslarının tahmin edilmesi için uygulanmıştır. Geliştirilen yöntemin, diskriminant analizi ve geleneksel yapay sinir ağı uygulamalarından daha iyi sonuçlar verdiği görülmüştür.

Alfaro vd. [15] çalışmalarında, kurumsal başarısızlık tahmini için yapay sinir ağları ve AdaBoost öğrenme algoritmasını uygulamış ve her iki yöntemin tahmin etmede elde ettikleri doğruluk oranları Avrupa'daki çeşitli şirketlere ilişkin veriler üzerinde karşılaştırılmıştır. Çalışma kapsamında elde edilen modelin, genelleştirme hatasını önemli ölçüde azalttığı gözlenmiştir.

Nachev [16] çalışmasında, şirket iflaslarının tahmin edilmesinde bulanık ARTMAP sinir ağlarını kullanmış, bu yöntemin güçlü ve zayıf yönlerini deneysel olarak incelemiştir. Çalışmada kullanılan yöntemin, hızlı öğrenen, ağ yapısını belirleme yetisine sahip ve yüksek tahmin etme doğruluk oranına sahip bir yöntem olduğu gözlenmiştir.

Cho vd. [17] çalışmalarında, iflas tahmin edilmesi için, karar ağacı ve olay tabanlı çıkarsama yaklaşımlarına dayalı melez bir yöntem geliştirmiştir. Yöntemde, karar ağaçları, değişken seçimi için kullanılmıştır. Çalışma mevcut Öklit uzaklığına dayalı olay tabanlı çıkarsama yaklaşımlarından farklı olarak, en yakın komşuların

konumlandırılmasında Mahalanobis uzaklığını kullanmaktadır. Finansal veri setlerinin çok miktarda finansal öznitelik içermesi nedeniyle uygun öznitelik alt kümesinin seçilmesi için geliştirilen yöntemde öznitelik seçimi de uygulanmıştır. Deneysel sonuçlar, geliştirilen melez yöntemin, mevcut yöntemlerden daha yüksek başarım elde ettiğini göstermiştir.

Olson vd. [10] tarafından gerçekleştirilen diğer bir çalışmada, temel bazı veri madenciliği yöntemleri şirket iflaslarını içeren veri setinde uygulanarak, yöntemler sınıflandırma doğruluk oranı ve oluşturdukları kural sayıları bakımından karşılaştırılmıştır. Deneysel çalışmalarda, karar ağaçlarının şirket iflasları için yapay sinir ağı ve destek vektör makineleri yöntemlerine kıyasla kısmen yüksek başarımlı sonuçlar elde ettiği ancak daha çok kural oluşturduğu görülmüştür.

Doolatabadi vd. [18] çalışmalarında, şirketlerin iflaslarının tahmin edilmesi için faktör analizi, lojistik regresyon ve CHAID karar ağacı algoritmalarının etkinliklerini incelemiştir. Deneysel çalışmada, Tahran menkul kıymetler borsasından 2006 ile 2011 yılları arasında elde edilen veriler kullanılmıştır. Regresyon yöntemi ile %84,5, CHAID karar ağacı yöntemi ile ise %77,6 doğru sınıflandırma oranı elde edildiği görülmüştür.

Zibanezhad vd. [19] çalışmalarında, Tahran menkul kıymetler borsasından 1996 ile 2009 yılları arasında elde edilen veriler kullanarak, C&R karar ağacı algoritmasının şirket iflaslarının tahmin edilmesinde başarımını incelemiştir. Deneysel sonuçlar algoritma ile %94,5 gibi yüksek bir doğru sınıflandırma oranı elde edildiğini göstermektedir.

Mohan [20] tarafından gerçekleştirilen çalışmada, karar ağacı algoritması için geleneksel öğrenme ve evrimsel algoritmaya dayalı iki farklı yöntemin etkinliklerini araştırmıştır. Geleneksel öğrenme yönteminde, bilgi kazancı ve kazanç oranı ölçütleri, evrimsel algoritmaya (genetik algoritma) dayalı yöntemde ise uygunluğa dayalı ve sıraya dayalı seçim stratejileri uygulanmıştır. Geliştirilen yöntemlerin aşırı uygunluk problemini ortadan kaldırmak için budama işlemi uygulanmıştır. Deneysel sonuçlar, geleneksel öğrenme yöntemine dayalı karar ağacı algoritmasının daha yüksek başarım elde ettiğini ve genetik algoritmaya dayalı öğrenmenin daha uzun eğitim süresi gerektirdiğini göstermektedir.

3. KARAR AĞACI YÖNTEMİ (THE DECISION TREE METHOD)

Karar ağacı yöntemi, sınıflandırma ve tahmin etmede kullanılan önemli veri madenciliği teknikleri arasında yer almaktadır. Karar ağacı, girdisi olmayan bir kök düğüm ve her biri birer girdi alan iç düğümlerden oluşan yönlü bir ağaçtır. Çıktıları bir başka düğüm tarafından girdi olarak alınan düğümler iç ya da test düğümü, çıktıları bir başka düğüme girdi olmayan düğümler ise yaprak düğümler olarak adlandırılmaktadır. Karar ağacında her bir iç düğüm, örnek uzayını girdi öznitelik değerlerinin belirli bir fonksiyona tabi tutulmasına dayalı olarak iki ya da daha fazla parçaya ayırmaktadır [21]. Karar ağacının iç düğümleri öznitelikler üzerinde gerçekleştirilen testleri, dallar test sonuçlarını ve her bir yaprak düğüm sınıf etiketini temsil etmektedir.

Karar ağaçlarının sınıflandırmada kullanılmasında, karar ağaçlarının basit yapısı sayesinde, oluşturulan sınıflandırma modelinin kolay anlaşılabilir olması, karar ağaçlarının parametrik olmaması sayesinde, bilgi keşfi için uygun bir yapı sunması, diğer sınıflandırma yöntemlerine kıyasla kısmen daha hızlı bir biçimde oluşturulması gibi özellikler rol oynamaktadır [22]. Bunun yanı sıra, karar ağaçlarından kuralların elde edilmesi de oldukça kolay bir biçimde gerçekleştirilebilmektedir. Karar ağaçları hem kategorik hem de nümerik verilerin sınıflandırılmasında kullanılabilmektedir. Karar ağaçları, değinilen üstün özelliklerine karşın, birden fazla öznitelik içeren çıktıları olanaklı kılmamaları, kısmen değişken sonuçlar üretmeleri, test verisindeki küçük değişikliklere karşı bile duyarlı olmaları, nümerik veri setleri için karmaşık bir ağaç yapısı oluşturmaları gibi problemler ile karşı karşıya kalmaktadır [9].

Sınıflandırma ve tahmin etmede karar ağaçlarının kullanılması, eğitim verisinden karar ağacı modelinin oluşturulması, bu modelin, test verisi kullanılarak uygun sınama ölçütleri aracılığıyla değerlendirilmesi ve ilgili modelin gelecekteki değerleri tahmin edilmesinde kullanılması şeklinde işlemektedir [23-24].

Veri setlerinden otomatik olarak karar ağacı yapısını oluşturmak amacıyla geliştirilmiş birçok karar ağacı algoritması bulunmaktadır. Karar ağacı algoritmaları genellikle genelleştirme hatasını en aza indirgeyen en uygun karar ağacı yapısını oluşturmayı hedeflemek ile birlikte, düğüm sayısı, ortalama derinlik ya da başka amaç fonksiyonlarını en aza indirgemeyi hedeflemek de mümkündür [21]. Karar ağacı algoritmalarının küçük boyutlu ve az derinlikli ağaçlar oluşturması amaçlanmaktadır. Karar ağacı algoritmaları sonucu oluşturulan büyük ve karmaşık karar ağaçları, düşük genelleştirme başarımına sahiptir. Bu nedenle, küçük boyutlu karar ağaçları oluşturmak için birçok yaklaşım geliştirilmiştir. Karar ağacı oluşturmada kullanılan yaklaşımlardan biri düğüm ayırmada ölçütlerin kullanılmasıdır. Bilgi kazancı, ki-kare istatistiği, GINI indeksi gibi ölçütler, başlıca kullanılan düğüm ayırma ölçütleri arasındadır [25].

Karar ağacından elde edilen genelleştirme performansını artırmak için kullanılan yaklaşımlar arasında budama yöntemi de yer almaktadır. Budama yöntemi, ağacın düşük istatistiksel geçerliliğe sahip alt ağaçlarını ortadan kaldırarak, daha küçük boyutlu bir ağaç elde edilmesini ve böylelikle genelleştirme doğruluk oranının iyileştirilmesini sağlamaktadır. Budama yöntemleri, düğümlerin yukarıdan aşağıya ya da aşağıdan yukarıya doğru taranması ile gerçekleştirilmektedir. Budama ile bir ölçütü iyileştiren düğümler budanmaktadır [26]. Maliyet-karmaşıklık budama yöntemi, iki aşamalı olarak gerçekleştirilmektedir. Birinci aşamada, T_0, başlangıçtaki ağacı, T_K, kök ağacı temsil etmek üzere eğitim verisi üzerinde T_0, T_1, ... T_K şeklinde bir ağaç dizisi oluşturulmaktadır. İkinci aşamada, bu ağaçlar içerisinden genelleme hatası tahminine dayalı olarak bir tanesi budanmış ağaç olarak seçilmektedir. Budama yöntemlerinden bir diğeri, azaltılmış hata budamasıdır. Bu yöntem, karar ağacının iç düğümlerinde aşağıdan yukarıya doğru gezinerek, her bir iç düğümün, en sık görülen sınıf ile yer değiştirilmesinin ağacın doğruluğunu azaltıp azaltmadığını kontrol etmekte ve bu kontrole dayalı olarak düğümleri budamaktadır. Bunun yanı sıra, en düşük hata budaması, karamsar budama, hata tabanlı budama, en uygun budama ve minimum tanım uzunluğu budama gibi çeşitli budama yaklaşımları bulunmaktadır [21]. Budama yöntemleri değerlendirildiğinde, maliyet-karmaşıklık budama yöntemi, azaltılmış hata budaması yöntemi gibi yaklaşımların gereğinden fazla budama yaparak, küçük ancak düşük doğruluk oranına sahip karar ağaçları oluşturduğu, en düşük hata budaması, karamsar budama ve hata tabanlı budama gibi yaklaşımların gereğinden az budama yaptıkları gözlenmektedir [26]. Her koşulda en uygun sonucu veren standart bir budama yönteminin bulunmadığı gözlenmektedir.

3.1. Karar Ağacı Algoritmaları (Decision Tree Algorithms)

Birçok farklı alanda başarı ile uygulanmış karar ağacı algoritmaları uygulamaları bulunmaktadır. Bu bölümde, çalışma kapsamında kullanılan karar ağacı algoritmaları kısaca tanıtılmıştır.

C4.5 algoritması [3], en çok bilinen karar ağacı algoritmalarından biridir. C4.5 algoritmasında, test öznitelik seçim ölçütü olarak bilgi kazancı oranı kullanılmakta ve her bir set için, en yüksek bilgi kazancı oranına sahip öznitelik seçilmektedir. C4.5 algoritması, ID3 algoritmasına [27] dayanan ve bu algoritmanın bazı kısıtlarını ortadan kaldıran bir yöntemdir. C4.5 algoritması hem sürekli hem ayrık öznitelikler ile çalışabilmektedir. Buna ek olarak, eksik öznitelik değerleri içeren eğitim veri setleri ile çalışabilmektedir. Bunun yanı sıra, karar ağacı oluşturma sırasında ya da sonrasında bazı düğümlerin ya da alt ağaçların silinmesi ile aşırı uygunluk problemini ortadan kaldırmakta, eğitim setindeki istisnai ve gürültülü değerlerin çıkarılmasını sağlamaktadır [28].

Decision Stump algoritması, tek seviyeli bir karar ağacı oluşturan bir yöntemdir. Bu yöntem ile oluşturulan ağaçta kök düğüm, yaprak düğümlere doğrudan bağlıdır. Decision Stump, sınıflandırma işlemini doğrudan tek bir girdi öznitelik değerine dayalı olarak gerçekleştirmektedir. Decision Stump algoritması genellikle boosting yöntemleri ile birlikte kullanılır [29].

Hoeffding Tree algoritması, her bir örneği en çok bir kez okuyarak ve uygun bir zaman aralığında işleyerek, büyük veri setlerinde etkin bir biçimde çalışan bir karar ağacı sınıflandırıcısıdır. Bunun yanı sıra, Hoeffding Tree algoritması, ID3, C4.5 ve SLIQ gibi geleneksel karar ağacı algoritmalarının depolama problemlerini ortadan kaldırmakta, oldukça karmaşık karar ağaçlarının bile kabul edilebilir bir hesaplama maliyeti ile oluşturulmasını olanaklı kılmaktadır. Algoritma, karar ağacının her bir düğümünde, düğümün nasıl parçalanacağına ilişkin kararın verilmesinde Hoeffding sınırı adı verilen istatistiksel değeri kullanmaktadır. Hoeffding Tree algoritmasının önemli özelliklerinden biri algoritma sonucu oluşturulan karar ağacının, her bir düğümün test edilmesi amacıyla tüm örnekleri kullanan sınıflandırıcılar ile hemen hemen aynı olmasıdır [5].

LMT (Logistic Model Trees) algoritması, karar ağacı indüksiyonu ve lojistik regresyon modellerini bir araya getiren bir yöntemdir [6]. Bu algoritmada ağaç yapısı C4.5 algoritmasına benzer şekilde genişletilir. Her bir parçalamada, ebeveyn düğümün lojistik regresyonları alt düğümlere geçirilir. Böylelikle, yaprak düğümlerin tüm ebeveyn düğümlere ilişkin bilgi içermesi ve her bir sınıf için olasılık tahminleri oluşturması sağlanır. Algoritma sonucu oluşturulan ağaç yapısına budama işlemi uygulanarak model sadeleştirilir ve genelleştirme performansı artırılır [30].

Random Forest algoritması [7], eğitim verisindeki örneklerin rastgele olarak seçilmesi ile oluşturulan budanmamış sınıflandırma ve regresyon ağaçlarından oluşan bir modeldir. Bu modelde, sınıflandırıcıların genelleştirme hatası, tüm ağaçların bireysel gücüne ve bu ağaçlar arasındaki bağıntıya dayalıdır. Her bir düğümün parçalanmasında kullanılacak özelliklerin rastgele olarak seçilmesi, algoritmanın Adaboost ile yarışacak sonuçlar vermesine ve gürültülü değerlere karşı daha dayanıklı olmasına neden olmaktadır.

Random Tree algoritması sonucu oluşturulan ağaç, olası ağaç kümesi içerisinden rastgele olarak seçilir. Burada, ağaç kümesi içerisindeki her bir ağaç eşit örnek olarak denenme şansına sahiptir. Ağaçların dağılımı uniform dağılış gösterir. Rastgele ağaçlar, etkin bir biçimde oluşturulabilmekte ve birçok rastgele ağacın oluşturduğu modeller genellikle yüksek doğruluk oranına sahip olmaktadır [8].

REPTree algoritması, hızlı karar ağacı sınıflandırma algoritmalarından biridir. Algoritma, karar ya da regresyon ağacının oluşturulmasında bilgi kazancı ölçütünü kullanmakta ve oluşan ağacı, azaltılmış hata budaması yöntemine dayalı olarak budama işlemine tabi tutmaktadır. REPTree algoritmasında, yalnızca nümerik özniteliklerin sıralanması söz konusudur. Eksik değerler için ise C4.5 algoritmasının örnekleri karşılık gelen parçalara ayırma yaklaşımı uygulanmaktadır [9].

4. VERİ SETİ VE BULGULAR (DATA SET AND FINDINGS)

Çalışmada kullanılan veri seti, Polonyalı şirketlere ilişkin finansal veriler taranarak oluşturulmuştur [31]. Veri seti, 112'si iflas eden şirketlere, 128'i iflas etmemiş şirketlere ait olmak üzere, toplam 240 şirkete ilişkin bilgi içermektedir. Veri seti, şirket iflaslarının gerçekleşmesinden 2-5 yıl önceki kayıtları içermektedir. Veri setinde biri sınıf etiketi olmak üzere toplam 33 öznitelik bulunmaktadır. Veri setinde şirketlerinin finansal yapıları, nakit/kısa vadeli borçlar, nakit/toplam aktifler, dönen varlıklar/kısa vadeli borçlar, dönen varlıklar/toplam aktifler, çalışma sermayesi/toplam aktifler, çalışma sermayesi/satış, satış/stok, satış/alacaklar, net kar/toplam aktifler, net kar/dönen varlıklar, net kar/alacaklar, brüt kar/satış, net kar/yükümlülükler, net kar/özkaynak, net kar/(özkaynak + uzun vadeli borçlar), satış/alacaklar, satış/toplam aktifler, satış/dönen varlıklar gibi toplam 33 öznitelik kullanılarak modellenmiştir. Veri setinde yer alan öznitelikler ve bu özniteliklere ilişkin tanımlayıcı istatistiksel veriler Tablo 1'de sunulmaktadır.

Tablo 1. Özniteliklere İlişkin Tanımlayıcı İstatistikler
(Descriptive Statistics for Attributes)

Öznitelik Adı	En az	En çok	Ortalama	Standart Sapma
Şirket No	1	130	66.108	38.872
Yıl	1997	2001	1998.425	1.084
X1	0	2.075	0.219	0.417
X2	0	0.529	0.061	0.087
X3	0.185	6.806	1.539	1.187
X4	0.048	0.999	0.602	0.237
X5	-1.016	0.707	0.055	0.282
X6	-85.01	0.559	-0.354	5.501
X7	0	76339	538.377	5272.293
X8	0	551.21	12.884	38.52
X9	-0.622	0.632	0.022	0.155
X10	-1.802	0.866	0.013	0.332
X11	-17.34	7.908	-0.087	1.863
X12	-17.34	8.983	0.034	1.945
X13	-0.747	4.885	0.148	0.557
X14	-54.079	2.637	-0.547	4.508
X15	-54.079	2.637	-0.209	3.541
X16	0	551.21	12.884	38.52
X17	0	13.499	2.524	2.137
X18	0	26.585	4.282	3.219
X19	0.662	4768.5	87.749	328.267
X20	0	13.499	2.524	2.137
X21	0.018	1722.3	8.221	111.313
X22	0.018	1722.3	7.806	111.24
X23	0.012	2.957	0.693	0.548
X24	-0.481	0.572	0.001	0.099
X25	0.034	1.899	0.607	0.3
X26	-61.313	1383.9	12.238	91.594
X27	-43.775	755.76	3.592	49.163
X28	-17.539	628.14	8.646	44.333
X29	-0.592	0.872	0.071	0.184
X30	0.038	78.299	0.697	5.079

Çalışmada başlıca karar ağacı algoritmalarının karşılaştırmalı analizlerinin gerçekleştirilmesi için WEKA aracında bulunan C4.5(J48), Decision Stump, Hoeffding Tree, LMT, Random Forest, Random Tree ve RepTree algoritmaları kullanılmıştır. Deneysel çalışmalarda 10-kat çapraz geçerleme yöntemi kullanılarak orijinal veri seti, rastgele olarak on eşit parçaya ayrılmıştır. Ardından, bu parçalardan bir tanesi modelin test edilmesi için geçerleme verisi olarak tutulurken, geriye kalan dokuz parça eğitim verisi olarak kullanılmıştır. Çapraz geçerleme süreci 10 kez gerçekleştirilerek, 10 parçanın her birinin birer kez geçerleme verisi olarak kullanılması sağlanmıştır.

Bu çalışma kapsamında, karar ağacı algoritmalarından karşılaştırmalı sonuçların elde edilmesinde veri setinde yer alan özniteliklere herhangi bir öznitelik seçim yöntemi uygulanmamış; sınıflandırma modelinin oluşturulmasında tüm öznitelikler dikkate alınmıştır. Veri seti herhangi bir eksik değer içermemektedir. Bu nedenle, veri seti herhangi bir ön işleme yöntemine tabi tutulmamıştır.

Çalışmada kullanılan algoritmaların değerlendirilmesinde, doğru sınıflandırma yüzdesi, ortalama mutlak hata, ortalama karesel hatanın karekökü, kesinlik, geri çağırma ve F-ölçütü kullanılmıştır.

Doğru sınıflandırma yüzdesi, belirli bir ikili sınıflandırma yönteminin bir şartı hangi oranda doğru olarak saptadığını belirleyen istatistiksel bir ölçüttür. Doğruluk, doğru pozitifler ve doğru negatiflerin toplamının, doğru pozitif,

yanlış pozitif, yanlış negatif ve doğru negatifler toplamına oranlanması ile hesaplanmaktadır.

Ortalama mutlak hata (OMH), tahminlerin nihai sonuçlar ile ne kadar yakın olduğunu ölçümlemek için kullanılan bir ölçüttür. Ortalama mutlak hata ölçütü aşağıda belirtilen formüle göre hesaplanmaktadır [32]:

$$OMH = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (1)$$

Burada, f_i, tahmin etme değerini, y_i ise gerçek değeri belirtmektedir.

Ortalama karesel hatanın karekökü (OKHK), tahminler ile gerçek sonuçlar arasındaki farkın ölçümlenmesi için kullanılan diğer bir ölçüttür. Ortalama karesel hatanın karekökü aşağıda belirtilen formüle göre hesaplanmaktadır [33]:

$$OKHK = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}} \quad (2)$$

Burada, i zamanında gözlemlenen değer x_{obs} ile gerçek değer x_{model} ile temsil edilmektedir.

Kesinlik oranı, doğru pozitifin, doğru pozitif ve yanlış pozitif toplamına oranı ile geri çağırma ise, doğru pozitifin, doğru pozitif ve yanlış negatif toplamına oranı ile hesaplanmaktadır [34].

F-ölçütü ise kesinlik ve geri çağırma değerlerini bir araya getiren bir ölçüttür. F-ölçütü değeri, aşağıda belirtilen formüle göre belirlenmektedir [34]:

$$F\ \text{ölçütü} = \frac{(1+\beta)*\text{Kesinlik}*\text{Geri Çağırma}}{\beta*Kesinlik+Geri\ Çağırma} \quad (3)$$

Burada, β değeri genellikle $\beta=1$ olarak alınarak, kesinlik ve geri çağırma değerlerinin eşit ağırlıklı olarak hesaplanması sağlanmaktadır. F-ölçütü, 0-1 aralığında değerler almaktadır ve yüksek başarımlı bir sınıflandırmada F-ölçütünün bire yakın bir değer alması beklenmektedir.

Tablo 2'de karar ağacı algoritmalarının işletilmesi sonucu elde edilen yaprak sayıları, ağaç boyutları ve ilgili modelleri oluşturmak için gereken süre (saniye cinsinde) sunulmuştur. Yaprak sayısı, belirlenemeyen algoritmalar için ilgili satır boş bırakılmıştır.

Tablo 2. Ağaç Yapıları ve Çalışma Süreleri (Structures of Trees and Running Times)

	Yaprak Sayısı	Ağacın Boyutu	Modeli Oluşturmak İçin Geçen Süre (saniye)
C4.5 (J48)	21	41	0.03
Decision Stump	-	Tek seviyeli	0
Hoeffding Tree	-	1	0.12
LMT	1	1	1.73
Random Forest	-	10 Ağaç	0.05
Random Tree	-	49	0
RepTree	4	7	0.02

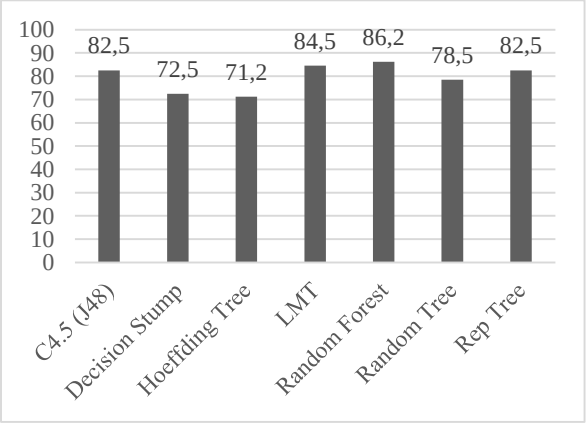
Tablo 3'te karar ağacı algoritmalarının her bir algoritma sonucunda elde edilen doğru sınıflandırılan örnek sayısı, yanlış sınıflandırılan örnek sayısı, doğru sınıflandırma yüzdesi, ortalama mutlak hata ve ortalama karesel hatanın karekökü değerleri sunulmuştur. Bu sonuçlar incelendiğinde, en yüksek doğru sınıflandırılan örnek sayısının Random Forest algoritması ile elde edildiği gözlenmektedir. İkinci en yüksek doğru sınıflandırılan örnek sayısı değerine ise LMT algoritması ile ulaşılmaktadır. Karar ağacı algoritmaları sonucu elde edilen doğru sınıflandırma yüzdeleri birbirlerine yakın değerler olmak ile birlikte, en düşük doğru sınıflandırma yüzdesi, Hoeffding Tree algoritması ile en yüksek doğru sınıflandırma yüzdesi ise Random Forest algoritması ile elde edilmektedir. Karar ağacı algoritmalarının şirket iflaslarının tahmin edilmesi problemine uygulanması ile elde edilen ortalama doğru sınıflandırma yüzdesi ise %79,7'dir. Algoritmalar sonucu elde edilen ortalama mutlak hata değerlerinin sıfıra yakın değerler olduğu gözlenmektedir. Bu nedenle, algoritmalar sonucu elde edilen tahmin değerleri ile gerçek değerlerin birbirlerine yakın değerler olduğu söylenebilir. Ortalama mutlak hata değerleri incelendiğinde, en iyi (en düşük) ortalama mutlak hata değerinin C4.5(J48) algoritması ile en kötü (en yüksek) ortalama mutlak hata değerinin ise Decision Stump algoritması ile elde edildiği görülmektedir. Algoritmalar sonucu elde edilen ortalama karesel hatanın karekökü değerlerinin birbirine yakın olduğu ve en iyi değerin Random Forest algoritması ile en kötü değerin ise Hoeffding Tree algoritması ile elde edildiği görülmektedir.

Çalışma kapsamında incelenen veri seti, iki sınıf içermektedir. Tablo 4'te birinci sınıf için karar ağacı algoritmalarının doğru pozitif oranı, yanlış pozitif oranı, kesinlik, geri çağırma, F-ölçütü gibi doğruluk değerleri bakımından karşılaştırması verilmiştir. İkinci sınıf için aynı oranlara ilişkin değerler ise Tablo 5'te sunulmuştur. Tablo 6'da ise her iki sınıf için ağırlıklı ortalamalı doğruluk oranı değerleri sunulmaktadır. Tablo 4, 5 ve 6'da sunulan doğruluk oranlarına ilişkin değerler incelendiğinde, sınıflar için ağırlıklı ortalamalı doğru pozitif oranı değerlerinin 0,8 civarında olduğu görülmektedir. Sınıflar için ağırlıklı ortalamalı değerler için en yüksek doğru pozitif oranı Random Forest algoritması ile en düşük doğru pozitif oranı ise Hoeffding Tree algoritması ile elde edilmektedir. Benzer şekilde, en düşük (en iyi) yanlış pozitif oranının Random Forest algoritması ile alındığı görülmektedir. Kesinlik, geri

çağırma ve F-ölçütü bakımından da en iyi sonuçlara değinilen algoritma ile ulaşılmaktadır.

Tablo 3. Sınıflandırma Ölçüt Değerleri (Values for Classification Measures)

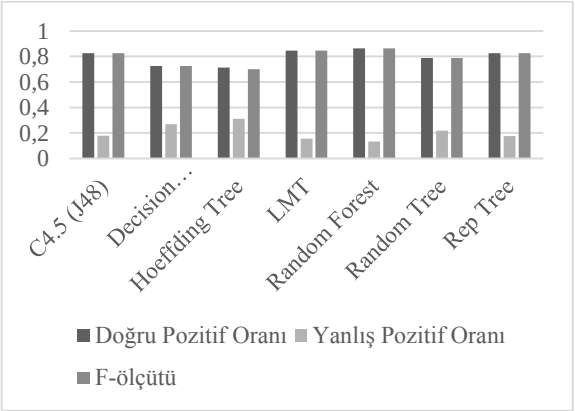
	Doğru Sınıflı Örnek Sayısı	Yanlış Sınıflı Örnek Sayısı	Doğru Sınıflan-dırma Yüzdesi	OMH	OKHK
C4.5 (J48)	198	42	82.5	**0.1949**	0.4008
Decision Stump	174	66	72.5	0.3686	0.4488
Hoeffding Tree	171	69	71.2	0.2884	0.5117
LMT	203	37	84.5	0.2131	0.3506
Random Forest	**207**	**33**	**86.2**	0.2363	**0.3414**
Random Tree	189	51	78.5	0.2125	0.4610
RepTree	198	42	82.5	0.2482	0.3851



Şekil 1. Karar Ağacı Algoritmalarını doğru sınıflandırma oranları (Accuracy Rates for Decision Tree Algorithms)

Bunun yanı sıra, karar ağacı algoritmaları ile elde edilen ortalama doğru sınıflandırma oranlarına ilişkin özet gösterim Şekil 1'de, algoritmaların doğru pozitif oranı, yanlış pozitif oranı ve F-ölçütü bakımından karşılaştırılması ise Şekil 2'de sunulmaktadır.

Şekil 3, 4 ve 5'te ise C4.5, RepTree ve Random Tree karar ağacı algoritmalarının çalıştırılması sonucu elde edilen karar ağaçları verilmiştir.



Şekil 2. Karar Ağacı Algoritmalarını ölçütlerden elde edilen değerleri (Measure Values for Decision Tree Algorithms)

5. SONUÇLAR (CONCLUSIONS)

Şirket iflaslarının önceden tahmin edilmesi ekonomide ele alınması gereken önemli problemlerden biridir. Şirket iflaslarının tahmin edilmesi, şirketlerin finansal yapılarını yansıtan önemli parametrelerin belirlenmesini ve bu parametrelerin tahmin etmede kullanılacak modelin oluşturulmasında kullanılmasını gerektirmektedir. Bu çalışma kapsamında, makine öğrenmesi ve veri madenciliği alanlarının önemli sınıflandırma yöntemlerinden biri olan karar ağacı yöntemi şirket iflaslarının tahmin edilmesi alanına uygulanmıştır. Çalışma kapsamında, başlıca karar ağacı algoritmalarından C4.5(J48), Decision Stump, Hoeffding Tree, LMT, Random Forest ve RepTree yöntemleri kullanılmıştır. Bu algoritmaların çalıştırılması sonucunda elde edilen ağaç yapıları ve özellikleri, çalışma süreleri, doğru sınıflandırma yüzdesi, ortalama mutlak hata, ortalama karesel hatanın karekökü, kesinlik, geri çağırma ve F-ölçütü karşılaştırılmalı olarak sunulmuştur. Karar ağacı algoritmalarının şirket iflaslarının tahmin edilmesinde kısmen başarılı (ortalama %79,7) doğru sınıflandırma yüzdesi elde ettiği görülmektedir.

Tablo 4. Birinci Sınıf İçin Doğruluk Oranı Değerleri (Accuracy Measure Values for First Class)

	Doğru Pozitif Oranı	Yanlış Pozitif Oranı	Kesinlik	Geri Çağırma	F-ölçütü
C4.5 (J48)	0.795	0.148	0.824	0.795	0.809
Decision Stump	0.777	0.320	0.680	0.777	0.725

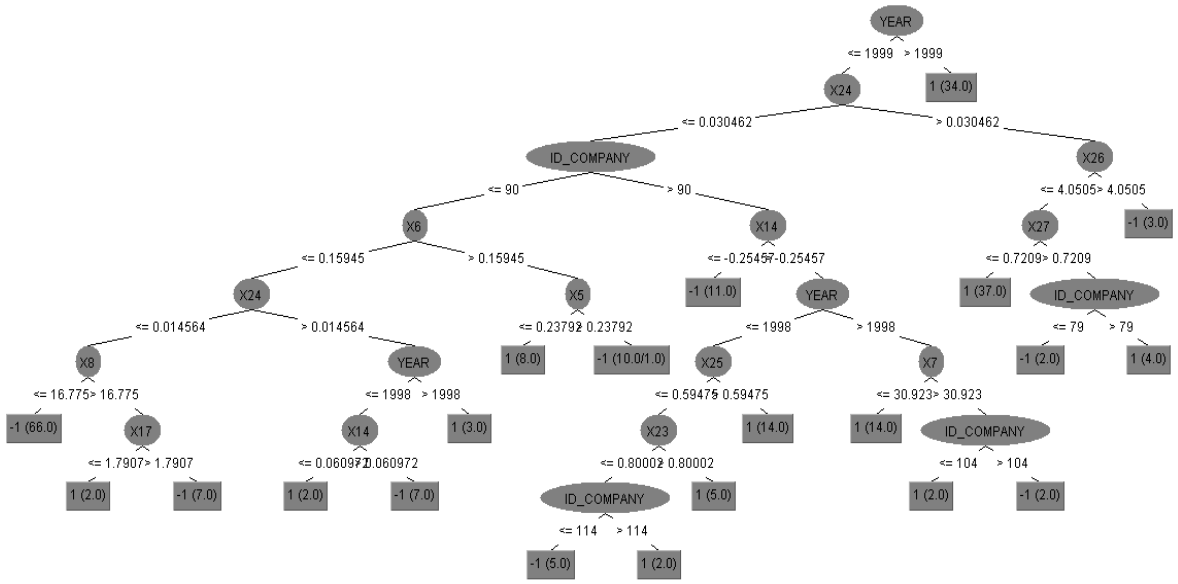
Hoeffding Tree	0.518	**0.117**	0.795	0.518	0.627
LMT	0.839	0.148	**0.832**	0.839	0.836
Random Forest	**0.893**	0.164	0.826	**0.893**	**0.858**
Random Tree	0.741	0.172	0.790	0.741	0.765
RepTree	0.821	0.172	0.807	0.821	0.814

Tablo 5. İkinci Sınıf İçin Doğruluk Oranı Değerleri (Accuracy Measure Values for Second Class)

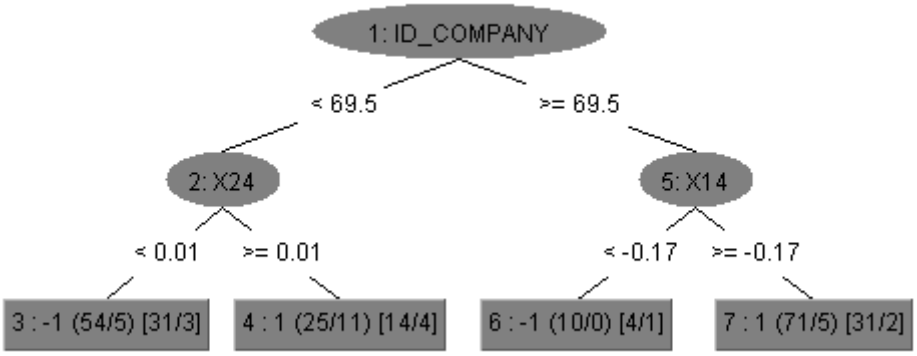
	Doğru Pozitif Oranı	Yanlış Pozitif Oranı	Kesinlik	Geri Çağırma	F-ölçütü
C4.5 (J48)	0.852	0.205	0.826	0.852	0.838
Decision Stump	0.680	0.223	0.777	0.680	0.725
Hoeffding Tree	**0.883**	0.482	0.677	**0.883**	0.766
LMT	0.852	0.161	0.858	0.852	0.855
Random Forest	0.836	**0.107**	**0.899**	0.836	**0.866**
Random Tree	0.828	0.259	0.785	0.828	0.806
RepTree	0.828	0.179	0.841	0.828	0.835

Tablo 6. Karar Ağacı Algoritmalarının Sınıflar için Ağırlıklı Ortalamalı Karşılaştırması (Averaged Comparative Values of Decision Tree Algorithms for Two Classes)

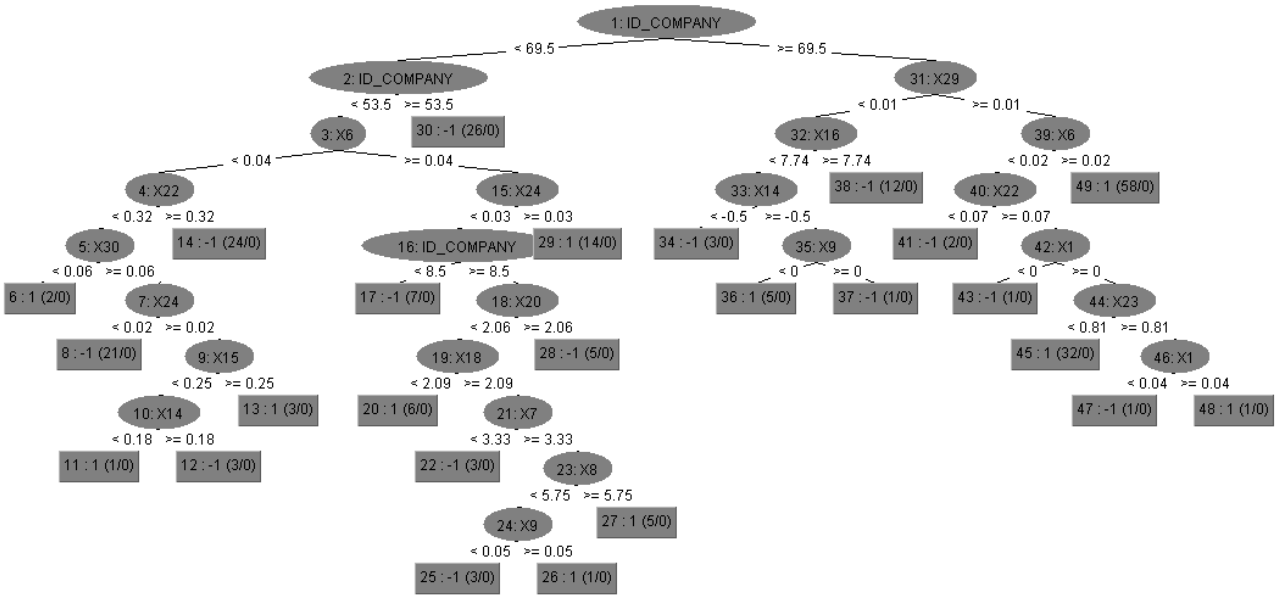
	Doğru Pozitif Oranı	Yanlış Pozitif Oranı	Kesinlik	Geri Çağırma	F-ölçütü
C4.5 (J48)	0.825	0.179	0.825	0.825	0.825
Decision Stump	0.725	0.269	0.731	0.725	0.725
Hoeffding Tree	0.713	0.312	0.732	0.713	0.701
LMT	0.846	0.155	0.846	0.846	0.846
Random Forest	**0.863**	**0.134**	**0.865**	**0.863**	**0.863**
Random Tree	0.788	0.218	0.788	0.788	0.787
RepTree	0.825	0.175	0.825	0.825	0.825



Şekil 3. C4.5 Algoritması ile Elde Edilen Karar Ağacı (Decision Tree Obtained By C4.5 Algorithm)



Şekil 4. RepTree Algoritması ile Elde Edilen Karar Ağacı (Decision Tree Obtained By RepTree Algorithm)



Şekil 5. Random Tree Algoritması ile Elde Edilen Karar Ağacı (Decision Tree Obtained By Random Tree Algorithm)

KAYNAKLAR (REFERENCES)

[1] S. Sumathi, S.N. Sivanandam, **Introduction to Data Mining and Its Applications**, Springer-Verlag, Berlin, 2006.

[2] F. Gorunescu, **Data Mining: Concepts, Models and Techniques**, Springer-Verlag, Berlin, 2011.

[3] J. R. Quinlan, **C4.5: Programs for Machine Learning**, Morgan Kauffman, San Francisco, CA, USA, 1993.

[4] I. Wayne, P. Langley, "Induction of One-Level Decision Trees", **Proceedings of the Ningth International Conference on Machine Learning**, 233-240, 1992.

[5] P. Domingos, P., G. Hulten, "Mining High-speed Data Streams", **Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 71-80, 2000.

[6] N. Landwehr, M. Hall, E. Frank, "Logistic Model Trees", *Machine Learning*, 59(1-2), 161-205, 2005.

[7] L. Breiman, "Random Forests", *Machine Learning*, 45(1), 5-32, 2001.

[8] W. Fan, H. Wang, P. S. Yu, S. Ma, "Is Random Model Better? On Its Accuracy and Efficiency", **The Third IEEE International Conference on Data Mining**, 51-58, 2003.

[9] Y. Zhao, Y. Zhang, "Comparison of Decision Tree Methods for Finding Active Objects", *Advances in Space Research*, 41(12), 1955-1959, 2008.

[10] D. L. Olson, D. Delen, Y. Meng, "Comparative Analysis of Data Mining Methods for Bankruptcy Prediction", *Decision Support Systems*, 52(2), 464-473, 2012.

[11] M. Lam, "Neural Network Techniques For Financial Performance Prediction: Integrating Fundamental And Technical Analysis", *Decision Support Systems*, 37(4), 567-581, 2004.

[12] S. Aoki, Y. Hosonuma, "Bankruptcy Prediction Using Decision Tree", **Proceedings of the Second Nikkei Econophysics Symposium**, 299-302, 2004.

[13] M. F. Santos, P. Cortez, J. Pereira, H. Quintela, "Corporate bankruptcy prediction using data mining techniques", *WIT Transactions on Information and Communication Technologies*, 37, 349-357, 2006.

[14] J. Neves, A. Vieria, "Improving Bankruptcy Prediction With Hidden Layer Learning: Vector Quantization", The *European Accounting Review*, 15(2), 253-271, 2006.

[15] E. Alfaro, N. Garcia, M. Gamez, D. Elizondo, "Bankruptcy Forecasting: An Empirical Comparision of

AdaBoost and Neural Networks", *Decision Support Systems*, 45(1), 110-122, 2008.

[16] A. Nachev, "Fuzzy ARTMAP Neural Network for Classifying the Financial Health of a Firm", *Lecture Notes in Computer Science*, 5027(2008), 82-91, 2008.

[17] S. Cho, H. Hong, B-C. Ha, "A Hybrid Approach Based On The Combination of Variable Selection Using Decision Trees And Case-Based Reasoning Using the Mahalanobis Distance: For Bankruptcy Prediction", *Expert Systems with Applications*, 37(4), 3482-3488, 2010.

[18] H. R. Doolatabadi, S. M. Hoseini, R. Tahmasebi, "Using Decision Tree Model and Logistic Regression to Predict Companies Financial Bankruptcy in Tehran Stock Exchanges", *International Journal of Emerging Research in Management &Technology*, 2(9), 7-16, 2013.

[19] E. Zibanezhad, B. Mobarake, D. Foroghi, "Applying decision tree to predict bankruptcy", **IEEE International Conference on Computer Science and Automation Engineering (CSAE),** 165-169, 2011.

[20] Internet: V. Mohan, Decision Trees: A comparison of various algorithms for building decision trees, http://cs.jhu.edu/~vmohan3/document/ai_dt.pdf, 08.12.2014.

[21] O., Maimon, L. Rokach, "Classification Trees". **Data Mining and Knowledge Discovery Handbook**, Editör: O., Maimon, L. Rokach, Springer, New York, A.B.D., 149-175, 2010.

[22] J. Gehrke, "Decision Trees", **The Handbook of Data Mining**, Editör: Nong Ye, Lawrence Erlbaum Associates Publishers, London, 149-175, 2003.

[23] P.J. García-Laencina, J.L. Sancho-Gómez, A.R. Figueiras-Vidal, "Pattern Classification with Missing Data: A Review", *Neural Comput. Appl.*, 19(2), 263-282, 2010.

[24] D. Birant, "Comparison of Decision Tree Algorithms for Predicting Potential Air Pollutant Emissions with Data Mining Models", *Journal of Environmental Informatics*, 17(1), 46-53, 2011.

[25] R. Kothari, M. Dong, "Decision Trees for Classification: A Review and Some New Results", **Pattern Recognition: From Classical to Modern Approaches**, Editör: S. K. Pal, A. Pal, World Scientific, New Jersey, 2001.

[26] S. B. Kotsiantis, "Decision Trees: A Recent Overview", *Artificial Intelligence Review*, 39(4), 261-283, 2013.

[27] J. Quinlan, "Induction of Decision Trees", *Machine Learning*, 1(1), 81-106, 1986.

[28] X. Niuniu, L. Yuxun, "Review of Decision Trees", **The Third IEEE International Conference on Computer Science and Information Technology**, 105-109, 2010.

[29] I. H. Witten, E. Frank, **Data Mining: Practical Machine Learning Tools and Techniques**, 2. Baskı, San Elsevier, Francisco, 2005.

[30] P. Doetsch, C. Buck, P. Golik, N. Hoppe, M. Kramp, J. Laudenberg, C. Oberdörfer, P. Steingrube, J. Forster, A. Mauser, "Logistic Model Trees with AUCsplit Criterion for the KDD Cup 2009 Small Challenge", **JMLR: Workshop and Conference Proceedings**, 77-88, 2009.

[31] W. Pietruszkiewicz, "Dynamical Systems and Nonlinear Kalman Filtering Applied in Classification", **Proceedings of 2008 7th IEEE International Conference on Cybernetic Intelligent Systems**, 263-268, 2008.

[32] Internet: Mean Absolute Error, http://en.wikipedia.org/wiki/Mean_absolute_error, 08.12.2014.

[33] K. Essig, H. Ritter, O. Strogan, T. Schack, "Influence of Movement Expertise on Visual Perception of Objects, Events and Motor Action: A Modeling Approach", **Developing and Applying Biologically-Inspired Vision Systems,** Editor: Pomplun, M., Suzuki, J., IGI Global, Hershey, A.B.D., 1-30, 2013.

[34] Internet: Precision and Recall, http://en.wikipedia.org/wiki/Precision_and_recal, 08.12.2014.