Fen Bilimleri Dergisi, 25(3) (2013) 100-114
Marmara Üniversitesi

http://dx.doi.org/ 10.7240/201332362

Dinamik Verilere Yönelik Karar-Tahmin Mekanizması Oluşturulması

Özkan ÇELİKTEN[1*], Hacer KARACAN[2]

[1]Türkiye Tarım Kredi Kooperatifleri Merkez Birliği, Krediler Daire Başkanlığı, Bilgi Teknolojileri Müdürlüğü, 06490, Bahçelievler/ANKARA

[2]Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 06570, Maltepe/ANKARA

Özet

Günümüzde, internetinde yaygın kullanımı ve veri depolamadaki manyetik ortamların teknoloji ile gelişmesi sonucu elimizde her sektörde sürekli büyüyen veri yığınları oluşmuştur. Depolanan verileri incelemek, analiz edip bilgi çıkarmak daha karmaşık hale gelmiş, bu da yeni yöntem ve teknolojilerin gelişmesi ihtiyacını ortaya çıkarmıştır. Bu amaçla, günümüzde mevcut problemleri çözmek, kritik kararları almak veya geleceğe yönelik tahminler yapmak için veri madenciliği yöntemi kullanılmaktadır. Bu çalışmada; kredi veren bir kurumun verileri kullanılarak, kredi politikaları oluşturmak için Oracle Data Miner üzerinde karar ağacı uygulaması geliştirilmiş, güvenilirliği yüksek bir karar-tahmin mekanizması oluşturmak için veri hazırlamanın, bunu her an aynı doğrulukta gerçekleştirip karar ağacı modelini sürekli güncel tutmak için ise dinamik veri hazırlamanın önemi ve yöntemleri üzerinde durulmuştur. Dinamik veri setlerinde modelin izlenerek yenilenmesi gerekeceğinden, tüm veri işleme adımlarının otomatik yapılması (dinamik veri hazırlama) sağlanmıştır. Ayrıca, karar ağacı veri setlerinin ve algoritma düğüm ayarlarının model başarısı üzerindeki etkisi değerlendirilmiştir. Son aşamada en iyilenmiş model uygulanmış ve karar ağacı kuralları ile kredi almak için gelenlerin durumlarını değerlendiren karar-tahmin mekanizması oluşturulmuştur.

Anahtar Kelimeler: Veri Optimizasyonu, Sınıflandırma, Karar Ağacı, Dinamik Veri Hazırlama

Constructing a Decision-Prediction Mechanism for Dynamic Data

Abstract

Nowadays, the amount of stored data is constantly growing because of the increasing internet usage and the developments in magnetic medium technology. It has been more complex to analyze this increasing amount of data, therefore new methods and technologies are needed. Today, data mining, is used for solving some of these problems and making future predictions. In this study; a decision tree application is developed for creating a credit policy by using data from a credit company. The application is developed on Oracle Data Miner and dynamical data preprocessing is done for creating the best model, and keeping the current model stable. Dynamical data processing is done automatically with some generated procedures. In addition, the success of data sets in decision tree and tuning of algorithm's nodes over the model is discussed in the study. At the last stage of the study, the best model is applied and a decision-prediction mechanism is constructed for evaluation of the people's case scoring for credit.

Keywords: Data Optimization, Classification, Decision Tree, Dynamical Data Processing

*Sorumlu Yazar: Özkan ÇELİKTEN, ocelikten@tarimkredi.org.tr , 0312 216 41 40

1. Giriş

Veri madenciliği uygulamalarının; ister belirli, ister belirsiz isterse de risk altındaki ortamlarda olsun, özellikle karar verme amacıyla kullanıldığında güvenilir olabilmesi gerekmektedir. Bunun için doğru model kurulması çok önemlidir. Yani, doğru veriyi seçer ve doğru tekniklerle onu işlemeye hazır hale getirirsek doğru kararlar alabilecek otomatize bir sistemden bahsedebiliriz. Burada cevap aranması gereken bazı sorular ortaya çıkmaktadır. Hangi verileri toplamalıyız? Doğru veri nasıl seçilir? Seçilen veri hangi işlemlerden hangi tekniklerle geçirilerek işlenebilir hale getirilmelidir? Doğru model nasıl kurulur? Bu model nasıl uygulanır? Nasıl değerlendirilir? Asıl problem, bunları yapıp bilgiye ulaştıktan sonra verimizin değişiklik göstermesidir. İşte o zaman, tüm bu soruları cevaplamak için gerekli veri işleme adımlarını tekrar tekrar uygulamak gerekecektir.

Bu çalışmada; veri madenciliğinin en önemli amaçlarından olan doğru ve güvenirliği yüksek bir karar-tahmin mekanizması oluşturmak için, aslında en iyi modeli kurmak için veri hazırlamanın, bunu her an aynı doğrulukta gerçekleştirebilmek için ise dinamik veri hazırlamanın önemi ve yöntemleri üzerinde durulmuş, problemin tanımlanmasından sonra başlayıp gerçek modelin kurulmasına kadar geçen veri madenciliği süreçlerinin nasıl otomatize edilebileceği uygulamalı olarak anlatılmıştır. Uygulama sonrası çıktılar değerlendirmek suretiyle bir karar-tahmin mekanizması oluşturulmuştur.

2. Veri Madenciliği

Günümüzde, internetinde yaygın kullanımı ve veri depolamadaki manyetik ortamların teknoloji ile gelişmesi sonucu elimizde her sektörde gitgide büyüyen veri yığınları oluşmuştur. Depolanan verileri incelemek, analiz edip bilgi çıkarmak daha karmaşık hale gelmiş ve yeni yöntem ve teknolojilerin gelişmesi ihtiyacını ortaya çıkarmıştır. Makine öğrenmesi, istatistik, veri tabanı, görselleştirme, bilgisayar bilimi ve diğer disiplinlerin (yapay zekâ, veri analizi vb.) bir araya gelmesi sonucu veri madenciliği kavramı ortaya çıkmıştır [1]. Veri Madenciliği, veri tabanlarında bilginin keşfedilip çıkarılması yani veri arkeolojisi, örüntü analizi, veri eşeleme, bilgi hasadı gibi isimlerle de anılır [2]. Diğer bir tanımı ise; büyük hacimli veri tabanlarında bulunan, mevcut durum veya gelecek tahminlerle ilgili henüz keşfedilmemiş fakat anlamlı ve yararlı olabileceği düşünülen bilgi elde etmek için bilgisayar programlarının yardımıyla kullanılan veri analizi tekniğidir [3]. Veri madenciliği kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli olan bilgileri sağlamaya yarayan bir araçtır. Veri madenciliği; analiste, iş yapma aşamasında oluşan veriler arasındaki şablonları ve ilişkileri bulması konusunda yardım etmektedir [4].

Veri madenciliği süreçleri 4 ana başlıkta toplanabilir:

1. Problemin Tanımlanması
2. Verilerin Hazırlanması
   a. Veri Ön İşleme (Toplama ve Uyumlaştırma, Değer Biçme ve Seçim, Birleştirme)
   b. Veri İşleme (Temizleme ve Yeniden Yapılandırma)
3. Modelin Kurulması Test Edilmesi ve Uygulanması
4. Modelin Değerlendirilmesi (İzlenmesi, Güncellenmesi ve Tekrar Uygulanması)

Veri madenciliği modellerini genel olarak tanımlayıcı ve tahmin edici modeller olmak üzere iki grupta inceleyebiliriz. *Tanımlayıcı modeller*, veriler arasındaki ilişki, benzerlik ya da sapmaların ortaya konmasını sağlamaktadır. Tanımlayıcı modeller tahmin için değil, mevcut durum analizi için kullanılmaktadır. *Tahmin edici modeller*, henüz sonuçları ya da aralarındaki ilişkileri bilinmeyen veri setleri için, daha önceden analiz edilmiş ve sonuçları

bilinen verilerle bilgi tahminine yaramaktadır. Buradaki temel amaç, var olan verilerden yararlanarak, gelecek verilerin özellik ve ilişkilerini tahmin etmektir. Tahmin edici modeller gelecek durum analizi için kullanılır. Bu çalışmada tahmin edici veri madenciliği araçlarından biri olan ve başarılı sonuçlar elde edilebilen 'Karar Ağacı' uygulanmıştır.

2.1. Karar Ağaçları

Kurulumu entegrasyonu modellemesi ve güvenilirliği bakımından daha kolay ve anlaşılır olduğundan sınıflandırmada en çok başvurulan algoritmadır [5]. Karar ağacı, adında belirtildiği şekilde ağaç görünümünde çıktıya sahip bir algoritmadır. Algoritmaya verilerin belli nitelikleri girdi ve çıktı olarak verilir. Bu çıktı niteliğindeki değerlere ulaşmak için hangi girdi niteliklerinin olması gerektiği ağaç veri yapıları ile oluşturulur [6].
Karar ağacı, veri seti içinden seçilen eğitim seti kullanılarak bir karar ağacı oluşturma esasına dayanır. Karar ağacının kalitesi, gerçekleştirdiği sınıflanma işleminin doğruluğuna ve ağacın boyutuna bağımlıdır. Bu aşamada karar ağacını oluşturan düğümlerin tespiti çok önemlidir. Eğitim verisindeki hangi alanların, hangi sırada kullanılarak ağacın oluşturulacağı bilinmelidir. Bu amaçla en yaygın olarak kullanılan ölçüm Entropy ölçümüdür.

Türkiye de farklı program ve yöntemlerle, karar ağaçları üzerinde bu çalışmaya benzer veri madenciliği uygulamaları yapılmıştır. Bunlardan en fazla benzerlik göstereni bir bankanın geçmişte müşterilerine verdiği ve kontratları sona ermiş olan kredileri inceleyip karar ağacı ve sınıflama kuralları oluşturarak, bu sınıflama kurallarını kullanmak suretiyle, kredi kontratı halen devam etmekte olan müşterilerin, kontrat sonunda kredilerini geri ödeme durumlarını tahmin etmeyi amaçlayan bir çalışmadır [11].

Diğer çalışmalarda ise; müşterilerin satın alma davranışlarını modelleyerek işletmeye pazar stratejisi belirleme [10], MineSet3.2 ile lise türü ve lise mezuniyet notunun, kazanılan fakülte üzerindeki önemini tespit etme [12], kan biyokimya parametreleri ile demir eksikliği anemisi teşhisinde doktorun vereceği kararları belirleme [13], öğrenci ve çalışanların, gıda tüketim desenlerini çıkarma [14], müşterinin sürücü koltuğundan memnuniyetini etkileyen en önemli değişkenleri belirleme [15], telekomünikasyon sektöründe faaliyet gösteren büyük bir firmanın, ayrılma eğilimi gösteren müşterileri belirleyerek; bu müşterilere özel pazarlama stratejileri geliştirme [16], ÖSYM tarafından 2008 ÖSS adayları için resmi internet sitesi üzerinden yapılan anket verileri üzerinde veri madenciliği karar ağacı yönetimi kullanılarak, öğrencilerin başarılarını etkileyen faktörleri belirleme [17], İMKB 100 endeksinde sanayi ve hizmet sektörlerinde faaliyet gösteren 173 işletmenin 2004–2006 yıllarına ait yıllık finansal göstergelerinden yararlanarak, sanayi ve hizmet sektörlerinde faaliyet gösteren firmaları ayıran en önemli değişkenleri saptama [18] gibi konularla karar ağacı uygulanmıştır.

Dünyada karar ağacı ile yapılmış birçok uygulama bulmak mümkündür. Micheline Kamber ve Lara Winstone özellikle büyük ölçekli veri tabanlarındaki karar verme süreçlerinin uzunluğuna, verimlilik ve ölçeklenebilirlik sorunlarını gidermeye yönelik çalışmalarda bulunmuştur [9]. Ankerst, Elsen, Ester ve Kriegel veri madenciliğinde karar ağaçlarının daha kolay (küçük boyutlu) ve anlaşılır olması için karmaşık algoritmalar yerine görsel ve etkileşimli yeni bir sınıflandırma sistemi geliştirmişlerdir [19]. Wenliang Du ve Zhijun Zhan, özel verilerde karar ağacı geliştirme modeli üzerinde çalışmış, iki ayrı kişiden gelen fakat gizliliği korunması gereken veriler üzerinde karar ağacı model uygulaması yapmışlardır [20]. Karar-tahmin amaçlı birçok sınıflandırmada karar ağacı uygulamasına rastlamak mümkündür. Ancak literatürde, Oracle Data Miner kullanımı ile aynı ya da benzer algoritmalarla geliştirilmiş bir veri madenciliği uygulamasına rastlanılamamıştır.

3. Çalışmanın Amacı ve Önemi

Uygulama, Türkiye Tarım Kredi Kooperatifleri verileri kullanılarak gerçekleştirilmiştir. Çiftçilerin tüm tarım girdi ihtiyaçlarını karşılamakla yükümlü olan Türkiye Tarım Kredi Kooperatifleri' nin en büyük gelir kaynağı ve en büyük risk teşkil eden faaliyet alanı çiftçilere yani ortaklarına kredi vermesidir. Burada kredi verirken bazı sorulara cevap bulup, bunlara göre kredi verme işlemini gerçekleştirmek gerekir. Zira daha çok önemli olan verilen krediyi geri tahsil etmektir. Bu sebepledir ki, bu çalışmanın temel amacı; karar vericilere, öncelikle iş risklerini azaltıp sonrasında sektörde hayatta kalmak ve başarıyı yakalamak için gerekli olan gelecek öngörüsünü verebilmektir. Geliştirilen uygulamada, karar vericiler, kredi faaliyeti gerçekleşirken hangi özelliklerin çiftçilerin kredi ödemeleri üzerinde daha önemli olduğu, hangi çiftçilerin hangi miktar kredileri geri ödeme olasılıklarının daha kuvvetli olduğu sorularına cevap bulabilir olduğundan mevcut durumlar için yeni kararlar verebileceklerdir.

Yeni gelen bir çiftçinin kredi limitini daha doğru belirlemek, limit dahilinde de olsa kredi isteyen bir çiftçiye doğru miktarda kredi sağlamak, bölgelerin risk analizlerini çıkararak yatırımlar yapmak, ödeme alışkanlığı en kötü olan gruptaki ortaklar üzerinde eğitici yada önlem alıcı faaliyetlerde bulunmak gibi geleceğe yönelik birçok stratejik karar bu sayede alınabilecektir.

4. Materyal ve Yöntem

Uygulama Oracle 11gR2 veri tabanı üzerinde kurulu 11.1. versiyonlu Oracle Data Miner [7] programı aracılığı ile gerçekleştirilmiştir. Problemin ve amacımızın özelliğinden kaynaklı veri madenciliği yöntemlerinden karar ağacı ile sınıflandırma yöntemi seçilmiş ve karar ağacı oluşturma algoritması olarak Entropy ve Gini kullanmıştır.

4.1. Problemin Tanımlanması

En önemli faaliyetlerinden biri kredi vermek ve verdiği kredileri toplamak olan Tarım Kredi Kooperatifleri' nin kredi vermedeki asıl problemi: 'BORCUNU HANGİ ÖZELLİKTEKİ ÇİFTÇİ, HANGİ OLASILIKLA ÖDER / ÖDEMEZ?' olarak özetlenebilir. Öncelikle çiftçi ortakların özelliklerinin belirlenmesi ve bu özelliklerin borç ödeme alışkanlığına ne denli etki ettiğinin hesaplanması gerekecektir. Böylelikle geçmiş verilerden bir sonuca vararak, gelecek için tahmin mekanizması oluşturabilir.

4.2. Verilerin Hazırlanması

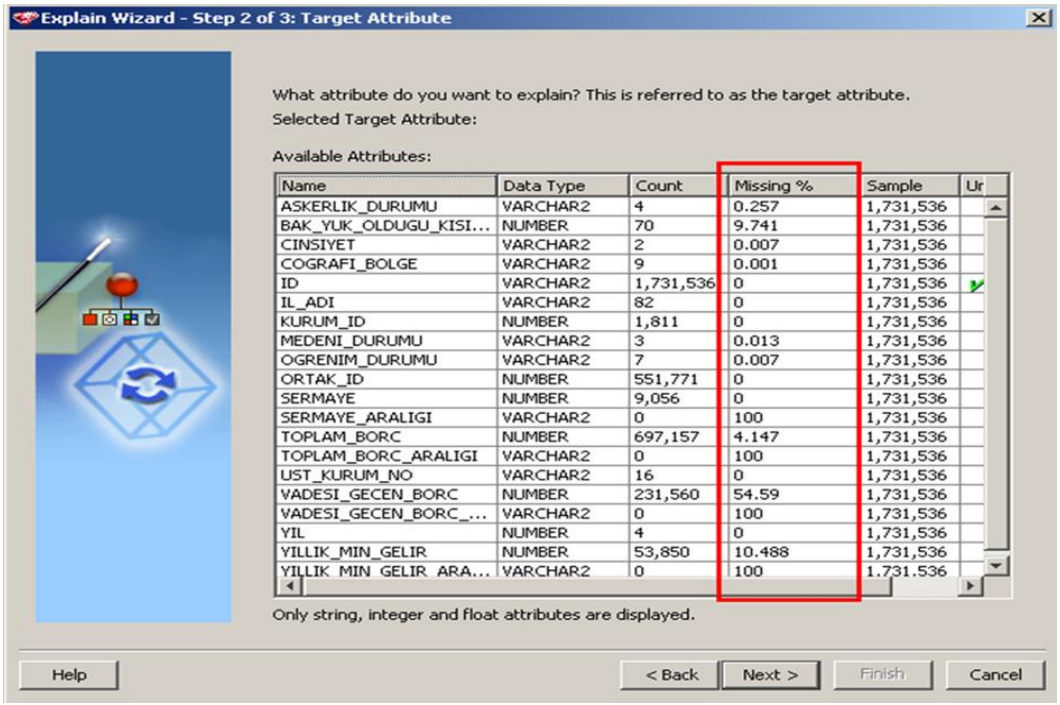
Problemimiz doğrultusunda ortağın hangi özelliklerinin, borcunu ödeme üzerinde diğerlerinden daha değerli olduğunun belirlenmesi gerekmektedir. Bunun için verilerin tek bir platformda toplanarak birleştirilmesi ve seçilmesi ile kirli verilerden temizlenmesi gerekmektedir.

Sürekli değişken yapıdaki veri setlerine sahip, özellikle de büyük ölçekli veri tabanlarında karmaşık veri madenciliği model yapısından kaçınmak gerekir. Bu kapsamda, farklı veri tabanları ve farklı tablolardan alınan çiftçi bilgilerinin yerel bir veri tabanında toplanması (oracle data pump tekniği ile) sağlanmıştır.

Problemimize daha fazla etki edebilecek veri setlerini seçmek önemlidir. Hedefimizin çiftçi ve onun borçları olduğunu hatırlamak hangi verileri seçeceğimiz konusunda yol gösterici

olmuştur. Toplanan verilerin her defasında farklı veri tabanlarından ve farklı tablolardan çekilmesi yerine; gerekirse view gibi veri tabanı nesneleri ile ya da yeni tablolar yapmak suretiyle birleştirmeye tabi tutulmuş, gerekli bilgiler farklı platformlarda da olsa, hepsine tek bir yerden ulaşılması sağlanmıştır.

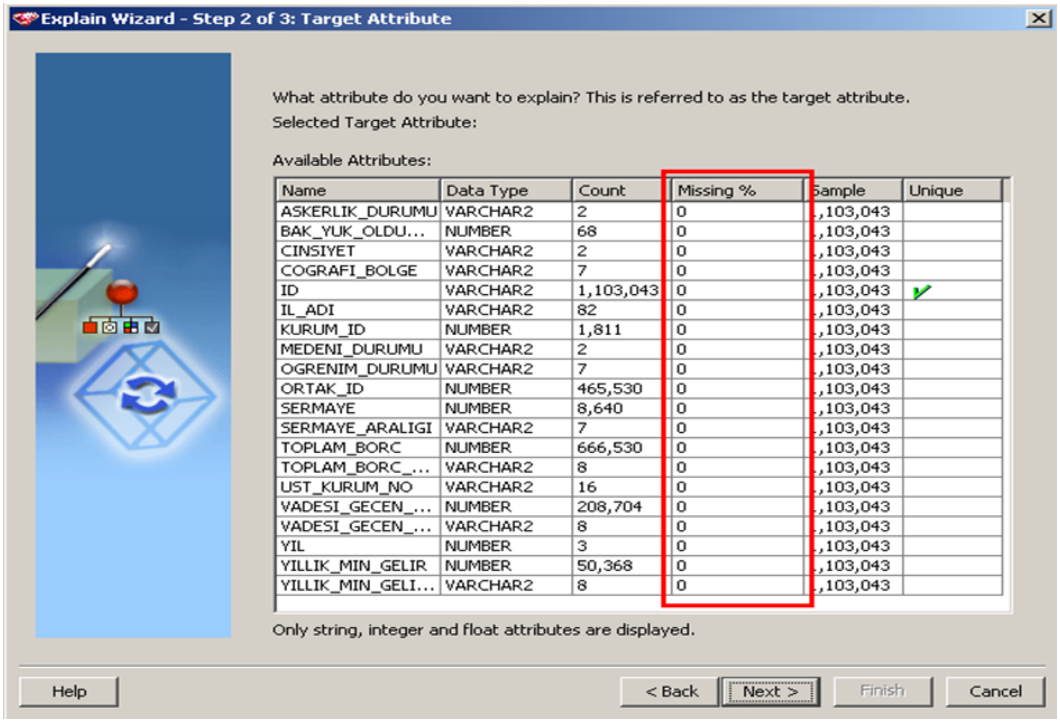
Öncelikle işimize yaramayacak, modelimizin güvenilirliğini düşürüp doğru çıktı üretmesini engelleyecek ya da performansımızı etkileyecek verilerden kurtulmamız gerekmektedir. Gürültülü, boş, eksik, artık, tutarsız diye sınıflandırılan bu kirli verilerin Oracle Data Miner aracılığı ile (Data – Transform – Missing Value) görsel olarak belirlenmesi mümkündür.



Name	Data Type	Count	Missing %	Sample
ASKERLIK_DURUMU	VARCHAR2	4	0.257	1,731,536
BAK_YUK_OLDUGU_KISI...	NUMBER	70	9.741	1,731,536
CINSIYET	VARCHAR2	2	0.007	1,731,536
COGRAFI_BOLGE	VARCHAR2	9	0.001	1,731,536
ID	VARCHAR2	1,731,536	0	1,731,536
IL_ADI	VARCHAR2	82	0	1,731,536
KURUM_ID	NUMBER	1,811	0	1,731,536
MEDENI_DURUMU	VARCHAR2	3	0.013	1,731,536
OGRENIM_DURUMU	VARCHAR2	7	0.007	1,731,536
ORTAK_ID	NUMBER	551,771	0	1,731,536
SERMAYE	NUMBER	9,056	0	1,731,536
SERMAYE_ARALIGI	VARCHAR2	0	100	1,731,536
TOPLAM_BORC	NUMBER	697,157	4.147	1,731,536
TOPLAM_BORC_ARALIGI	VARCHAR2	0	100	1,731,536
UST_KURUM_NO	VARCHAR2	16	0	1,731,536
VADESI_GECEN_BORC	NUMBER	231,560	54.59	1,731,536
VADESI_GECEN_BORC_...	VARCHAR2	0	100	1,731,536
YIL	NUMBER	4	0	1,731,536
YILLIK_MIN_GELIR	NUMBER	53,850	10.488	1,731,536
YILLIK_MIN_GELIR_ARA...	VARCHAR2	0	100	1.731.536

Şekil 1. Alınan Verilerdeki Kirli-Kayıp Değerler

Verilerimizin işlenmesi; 'Data Tipinin Transformasyonu', 'Sürekli Kolonların Transformasyonu', 'Gruplama', 'Kayıp Verilerin İşlenmesi' ve 'Uç Verilerin Ortadan Kaldırılması' yöntemleri ile olmaktadır. Örneğin, il bilgisi olmayan çiftçilerin kayıtlarının silinmesi, coğrafi bölge bilgisi boş yada yanlış olan kayıtların en çok kullanılan kayıtla güncellenmesi, sermayesi olmayan yada negatif olan ortakların silinmesi, çiftçinin borç miktarı boşsa '0' (sıfır) ile güncellenmesi, tekrarlı kayıtların silinmesi, çiftçinin bakmakla yükümlü olduğu kişi sayısı boşsa tüm tablonun ortalamasıyla doldurulması (3 basamaklı olan bakmakla yükümlü kişi sayısı gibi uç değerlerden temizlenerek), cinsiyeti boş yada 'E' veya 'K' olmayan kayıtların ise 'E' (erkek) olarak güncellenmesi, öğrenim durumu boş yada yanlış olan kayıtların da en çok kullanılan kayıtla güncellenmesi gibi çeşitli işlemler yine prosedür yazılarak otomatize hale getirilerek veriler işlenmiştir.



Şekil 2. Veri İşleme Sonrası Kirli-Kayıp Değerler

Sınıflandırma tekniklerinde verilerin kategorik olması gerekmektedir. Özellikle parasal miktarların (sürekli değerlerin) çok fazla çeşitlilik göstermesinden kaynaklı, karar ağaçlarında düğüm sayısı binleri bulan sonuçlarla (sınıflarla) karşılaşmamız olası olacaktır. O yüzden kesikli değerler 'Sürekli Kolonların Transformasyonu' metodu uygulamak üzere pl/sql diliyle yazılan bir prosedür ile kategorize edilerek şu kolonlar oluşturulmuş ve doldurulmuştur:

- ✓ SERMAYE_ARALIGI,
- ✓ VADESI_GECEN_BORC_ARALIGI,
- ✓ TOPLAM_BORC_ARALIGI
- ✓ YILLIK_MIN_GELIR_ARALIGI

Vadesi geçen borç miktarını belirli aralıklar için 'B0,B1,…,BX' şeklinde kesikli değerlere çevrilmiştir. Vadesi geçen borç miktarının; 0 TL olması 'B0', 0-1000 TL arasında olması 'B1', 1000-1500 TL arasında olması 'B2', 1500-2000 TL arasında olması 'B3', 2000-3000 TL arasında olması 'B4', 3000-5000 TL arasında olması 'B5', 5000-10000 TL arasında olması 'B6' ve 10000 TL'den büyük olması ise 'B7' ile ifade edilmiştir. Burada istenen durum bir çiftçinin borcunu zamanında ödemesi yani vadesi geçen borcunun olmaması ('B0' olması) durumudur.

4.3. Modelin Kurulması, Test Edilmesi ve Uygulanması

Karar ağacına girdi olacak verilerin tamamını modelde girdi olarak göstermek, ters etki eden kolonlarda modelin güvenilirliğini düşüreceğinden, hangi girdinin çıktımız üzerinde daha etkili olacağını belirlemek için "Explain" analiz sonuçlarından faydalanılmıştır. Çıktı değişkeni ise, çiftçi borcunun ne kadarını zamanında ödememişi gösteren VADESİ GEÇEN BORÇ kolonumuzdur. Amacımız hangi nitelikteki çiftçilerin (ki bu nitelikler modelimizin girdileridir) borcunu zamanında öde(me)me olasılığını belirlemektir.

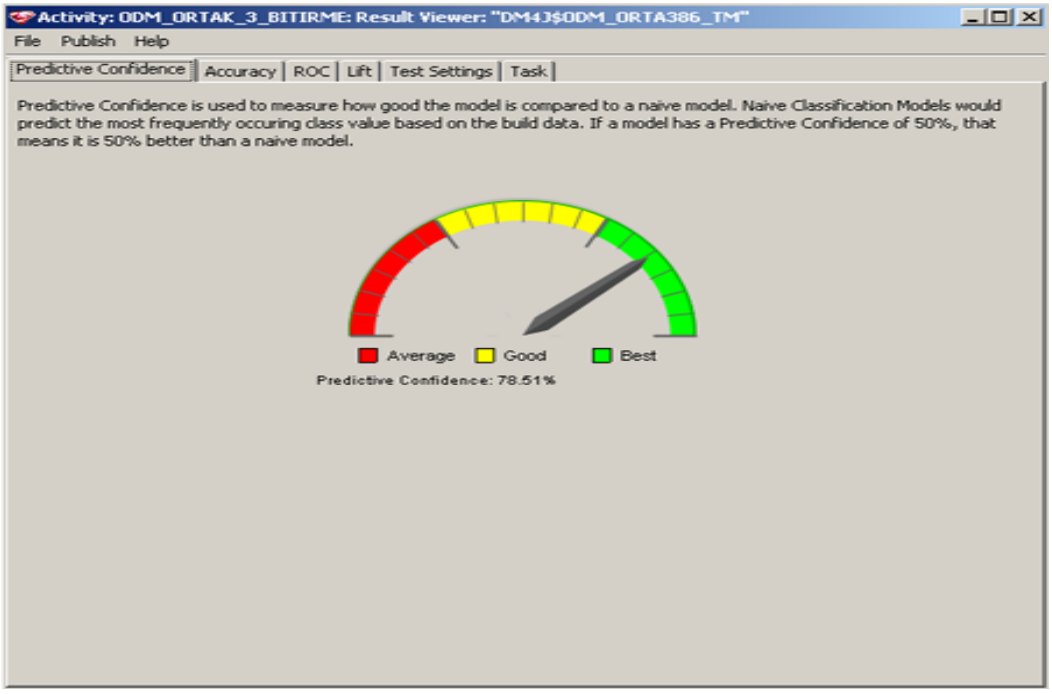
Tahmin edici modeller genellikle denetimli öğrenme modelleridir. Denetimli öğrenme modellerinde öncelikle model, bir miktar veri (öğrenme verisi) üzerinde çalıştırılır. Bu aşama modelin eğitilmesi olarak da adlandırılır. Daha sonra model verinin kalan kısmında test edilerek doğrulanır. Eğitim ve test döngüsü tamamlandıktan sonra model oluşturulur ve test kümesi ile modelin doğruluk derecesi belirlenir. Bir modelin doğruluğunun test edilmesinde çeşitli geçerlilik yöntemleri kullanılır[8]. Günümüzdeki veri madenciliği programları modelin güvenirliğini ROC, LIFT Analizi ve Risk Matrisleri gibi çeşitli yöntemlerle kullanıcıya hesaplayarak döndürmektedir.

Test verimizin uygulanması sonucu modelimizin %22 lik bir güvenilirliğe sahip olduğu görülmüştür. Tüm değişkenleri model üzerinde girdi olarak tanımlamak, görüldüğü üzere modelin güvenilirliğini düşürdüğü gibi modelin uygulanması ve güncellenmesini de imkansız hale getirecektir.

4.4. Modelin Değerlendirilmesi (İzlenmesi ve Güncellenmesi)

Model sistem içerisine gömülerek probleme karşı verdiği cevapların değerlendirilmesiyle kararlar verilir. Fakat unutulmamalıdır ki oluşturulan model dinamik verilerde ya da dinamik sistemlerde sürekli en iyi performansı vermeyecektir ve o modelin sürekli takip edilerek değişen koşullara uyum sağlaması için tekrar test edilmesi ve tekrar eğitilmesi hatta gerekiyorsa tekrar oluşturulması gerekebilir.

Daha başarılı bir model elde etmek için örnekleme, öğrenme ve test verilerimizin ayarları ile girdi değişkenlerimizi sürekli değiştirip yeni çıktılar karşılaştırılmıştır. Daha güvenilir sonuca ulaşmada, çıktı değişkenine (“Vadesi Geçen Borç”) en çok etki eden girdi değişkenlerini belirlemek için “Explain” analizden faydalanılmıştır. En iyilenmiş modelimizin girdileri ‘Bakmakla Yükümlü Olduğu Kişi Sayısı’, ‘İl’, ‘Sermaye’, ‘Toplam Borç’, ‘Üst Kurum No’ ve ‘Yıllık Minimum Gelir’ kolonları olup, toplam 1.731.536 adet verinin işlenmesinden sonra kalan 1.103.043 adet veriden 200.000 örnekleme alınarak ve %90 öğrenme, %10 test verilerine ayırmak suretiyle, Entropy dağılımı 15 seviye olarak belirlenmiştir. Sonuçta oluşan en iyilenmiş modelimizin güvenilirliği %78,51 olarak elde edilmiştir.



Şekil 3. En İyilenmiş Modelin Güvenilirliği

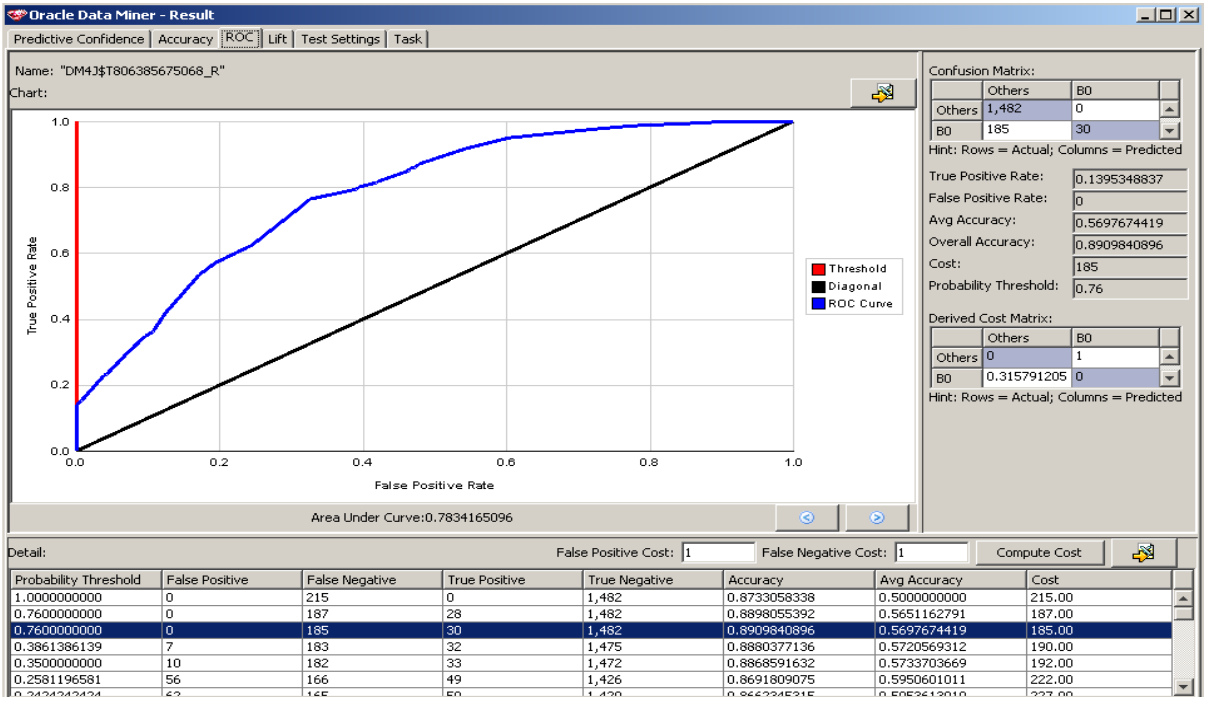
Risk Matrisinde tahmin edilen sınıf değeri kolonda gösterilirken, gerçek değer satırda gösterilir. Matrisin köşegeni doğru tahmin edilen sınıf sayısını göstermektedir. Sonuçta test verisi içerisinden her bir hedef değer grubu için ortalama eşit sayıda örnekleme alınarak model test edildiğinde, kurduğumuz model bize 1.697 verinin 1.375 tanesini doğru tahmin etmiştir. Modelimizin doğruluk oranının %81 olduğu söylenebilir. Modelimizin test verileri asıl hedef kolonumuz olan 'B0' değerini tahmin etmede %100 başarılı gözükmektedir.

Confusion Matrix: Rows = Actual; Columns = Predicted ☑ Show Total and Cost

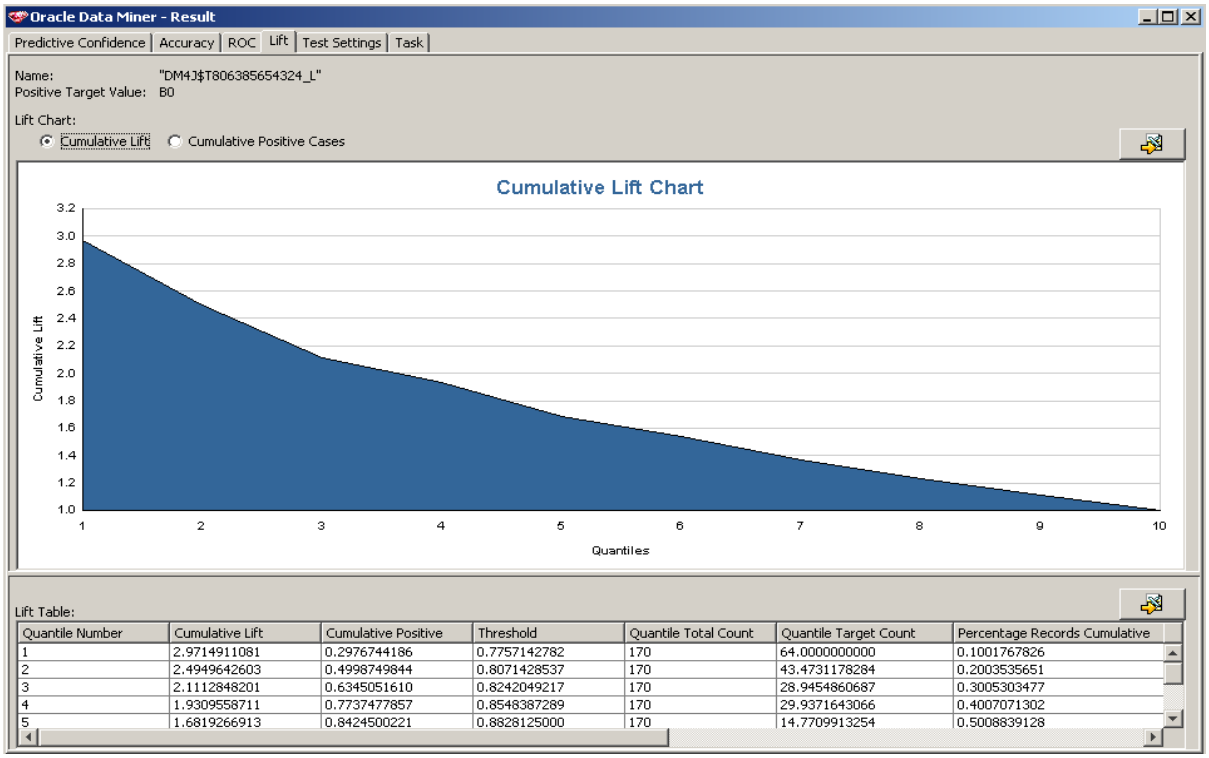
	B0	B1	B2	B3	B4	B5	B6	B7	Total	Corre...	Cost
B0	30	18	13	15	21	44	44	30	215	13.95	185
B1	0	165	3	3	6	8	11	9	205	80.49	40
B2	0	0	198	5	6	6	3	9	227	87.22	29
B3	0	0	0	210	6	2	5	4	227	92.51	17
B4	0	0	0	0	179	9	7	4	199	89.95	20
B5	0	0	0	0	0	195	15	9	219	89.04	24
B6	0	0	0	0	0	0	189	7	196	96.43	7
B7	0	0	0	0	0	0	0	209	209	100	0
Total	30	183	214	233	218	264	274	281	1,697		
Correct %	100	90.16	92.52	90.13	82.11	73.86	68.98	74.38			
Cost	0	18	16	23	39	69	85	72			

Şekil 4. En İyilenmiş Modelin Risk Matrisi

Modelin sağladığı Faydanın doğruluğunu gösteren Lift grafiği ile yine doğruluğu ölçmede kullanılan ROC grafiği çıktıları aşağıdaki şekillerde görülmektedir. ROC grafiğinde, tüm durumların mükemmel ve tüm durumların hatalı (tüm durumların pozitif ya da negatif) tahmin edilmesi olasılıkları incelenmiştir.



Şekil 5. Modelin ROC Analizi



Şekil 6. Modelin Lift Analizi

4.5. (En İyilenmiş) Modelin Uygulanması

Modelin uygulama adımında, yeni bir çiftçi ortağın kredi çekmek için kuruma geldiğini varsayalım.
Gelen çiftçiye ait bilgiler şunlardır;

- ✓ Bakmakla Yükümlü Olduğu Kişi Sayısı: 15
- ✓ Çiftçilik Yaptığı İl Adı: DENİZLİ
- ✓ Sermayesi: 300 TL
- ✓ Kredi Çekmek İstediği Tutar (Toplam Tüm Borçları): 10.000 TL
- ✓ Geldiği Kooperatifin Bağlı Olduğu Üst Kurum: B04 (İZMİR BÖLGE BİRLİĞİ)
- ✓ Beyannamesine Göre Sistem Tarafından Hesaplanan Yıllık Minimum Gelir: 14.000 TL

Bu verileri modelimizi uygulayacağımız "KREDI_ICIN_GELEN_ORTAK" adlı yeni bir tabloya kaydedelim. Sonuç olarak kredi çekmek için gelen çiftçinin vadesini geçireceği borç miktarı B0'a (sıfır) doğru olması gerekirken, olasılığı en yüksek çıkan sonuç B7 (10.000 TL ve daha fazla vadesi geçen) olmuştur. (Bkz. Tablo 1)

Tablo 1. Model Uygulama Sonucu Olasılık Dağılımları

	VADESİ GEÇEN BORÇ								Düğüm
	B0	**B1**	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**	
Olasılık	0,0023	-	0,0011	-	-	-	-	0,9964	67

Sonuçta ortağın hiç borcunun olmadığını varsayar ve 10.000 TL kredi çekmek için geldiğini düşünürsek, bu ortak çekeceği 10.000 TL'lik kredinin tamamını (B7 durumu 10.000 ve fazlası içindi) ödememe olasılığı %99,64 dür. Uç bir örnek olması bakımından bu veriler seçilmiş olup, zira Tarım Kredi Kooperatifleri' n de nakdi ya da ayni toplam 10.000 TL'lik bir kredi verilebilmesi için ortağın en az 500 TL sermayesi olması gerekmektedir (kredi limitinin en az %5'i). Nitekim bu örnekte, ortağın borcunu çok büyük bir olasılıkla ödemeyeceği de bu sebeple çıkmıştır. Bu sonuç, diğer verilerden hareketle elde ettiğimiz modelimizin doğruluğunu da ispatlar niteliktedir. Peki bu ortağa kredi vermek istersek hangi koşullarda ve ne kadar kredi vermeliyiz? Ortağın sermayesini artırması, çiftçilik yaptığı ili değiştirmesi, daha az kredi çekmesi, ya da beyannamesini değiştirerek yıllık gelirini artırması seçeneklerinde hangi olasılıklarda borcunu öder / ödemez? Aşağıda ortağın olası durumlarının listesi verilmiştir. Bu durumlar yapılabilirlik doğrultusunda istenildiği gibi değiştirilebilir.

Tablo 2. Yeni Uygulama İçin Olası Durumlar

	BAKTIĞI KİŞİ	**İLİ**	**SERMAYESİ**	**BORCU (KREDİ)**	**UST KURUMU**	**GELİRİ**	**DURUM**
1	15	DENİZLİ	300	10000	İZMİR	14000	Ortağın mevcut durumu
2	15	DENİZLİ	500	10000	İZMİR	14000	Ortağın sermayesinin artırılması
3	15	DENİZLİ	500	6000	İZMİR	16000	Ortağın çekmek istediği kredi miktarının indirilmesi ve beyannamesini değiştirerek yıllık gelirinin yükseltilmesi
4	15	DENİZLİ	1000	2500	İZMİR	16000	Ortağın sermayesinin artırılması ve çekmek istediği kredi miktarının indirilmesi
5	2	İZMİR	1000	2500	İZMİR	16000	Ortağın bakmakla yükümlü kişi sayısının azalması ve çiftçiliğini Denizli de değil de İzmir de yapması

Modelimizi KREDI_ICIN_GELEN_ORTAK tablosu üzerinde tekrar çalıştırdığımızda, yukarıdaki her durum için Tablo 3'teki sonuçlar elde edilmiştir.

Tablo 3. Olası Durumlar İçin Olasılık Dağılımları

	VADESİ GEÇEN BORÇ								DÜĞÜM
	B0	**B1**	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**	
1	0,0023	-	0,0011	-	-	-	-	0,9964	67
2	0,2048	0,0878	0,0878	0,0829	0,0341	0,0829	0,1560	0,2634	78
3	0,2242	0,0942	0,0671	0,0714	0,0599	0,0899	0,3928	-	63
4	0,1989	0,0749	0,0620	0,0956	0,5684	-	-	-	51
5	0,1989	0,0749	0,0620	0,0956	0,5684	-	-	-	51

Ortağa istediği 10.000 TL'lik kredinin hemen verilmesi durumunda zamanında geri ödeme alınması neredeyse mümkün değildir. Ancak, sermaye artırımı ve beyanname yenileme ile 6.000 TL kredi verilirse %22,4 olasılıkla zamanından önce geri ödeme alınacaktır. (Durum:3)

Ortağın bakmakla yükümlü kişi sayısının azaltılması ve çiftçiliğini Denizli de değil de İzmir ilinde yapmasının (Durum:5) bu ortak için önemli olmadığı, borcunu zamanında ödemesi için en iyi durumun ise 3. durum olduğu görülmektedir. 3. durum, karar ağacımızın 63. Düğümüne karşılık gelmektedir. Karar ağacımızın 63. Düğümü (B6 durumu) için oluşturduğu kural yapısını gösteren örnek kod bloğu aşağıda verilmiştir.

```
IF

UST_KURUM_NO is in ( B020000 B030000 B040000 B060000 B160000 ) AND
IL_ADI is in ( ADANA ADIYAMAN AFYON AFYONKARAHİSAR AKSARAY ANKARA ANTALYA AYDIN
AĞRI BALIKESİR BATMAN BAYBURT BOLU BURDUR BURSA BİLECİK DENİZLİ EDİRNE ELAZIĞ
ERZURUM ESKİŞEHİR GAZİANTEP HATAY ISPARTA KAHRAMANMARAŞ KASTAMONU KAYSERİ KIRIKKALE
KIRKLARELİ KIRŞEHİR KOCAELİ KONYA MALATYA MANİSA MUĞLA OSMANİYE SİVAS TEKİRDAĞ
TOKAT YALOVA YOZGAT ÇANAKKALE ÇANKIRI ÇORUM İSTANBUL İZMİR İÇEL ) AND
UST_KURUM_NO is in ( B010000 B020000 B030000 B040000 B060000 B160000 ) AND
SERMAYE > 344.565 AND
TOPLAM_BORC <= 9999.17 AND
TOPLAM_BORC > 5001.385 AND
TOPLAM_BORC > 2999.905 AND
TOPLAM_BORC > 1999.935 AND
TOPLAM_BORC > 1500.19

THEN
VADESI_GECEN_BORC_ARALIGI equal B6

Confidence (%)=39.29
Support (%)=4.66
```

Şekil 7. 63.Düğüm İçin Dallanma Yapısını Oluşturan Kural Kod Bloğu

Sonuçlara bakılarak, bu ortağa kredi şu koşullarda şu risklerle verilir;

Tablo 4. Olası Durumlar İçin Toplam Risk Dağılımları

DURUM	RİSK %	KREDİ KOŞULU
1	99,75	10.000 TL KREDİ HEMEN VER
2	79,49	SERMAYESİNİ 500 TL' ye ÇIKAR 10.000 TL KREDİ VER
3	77,53	SERMAYESİNİ 500 TL' ye ÇIKAR ve YILLIK GELİRİNİ 2.000 TL ARTIR AMA 6.000 TL KREDİ VER
4	80,09	SERMAYESİNİ 1000 TL'ye ÇIKAR ve YILLIK GELİRİNİ 2.000 TL ARTIR AMA 2.500 TL KREDİ VER
5	80,09	SERMAYESİNİ 1000 TL'ye ÇIKAR ve YILLIK GELİRİNİ 2.000 TL ARTIR, BAKTIĞI KİŞİ SAYISINI AZALT ve İZMİR' deki ARAZİSİNDE ÇİFTÇİLİK YAPTIR AMA 2.500 TL KREDİ VER

Kurduğumuz modelin girdisi olan verilerimizin, zamanla özelliklerinde ve boyutunda değişiklikler meydana geleceği için, sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerekecektir. Uygulamamızda girdi değişkenlerimizin dinamik olması, hazırladığımız prosedürlerin tekrar çalıştırılmasını gerektirir.

Fakat veri setlerimizdeki bu değişkenlik (özellikle yeni kolonların eklenip çıkartılmasından sonra) işlenmiş verilerimizden rastgele alacağımız eğitim ve test verilerimizin ilişkilerini değiştirebilecek ya da model güvenilirliği azaltacak nitelikte olabilir. Böyle durumlarda Explain analizle ilişkilerin derecesine tekrar bakılmalıdır. Eğer bir fark görülürse model farklı girdiler ile tekrar kurmalı ve analiz sonuçlarına bakarak yeni model tekrar değerlendirilerek uygulanmalıdır.

5. Bulgular ve Tartışma

Firmaların rekabet ortamında var olabilmek için sadece veri elde etmekten çok, detaylı, anlamlı ve işe yarar veri elde etmeleri son derece önemlidir. Ama firmayı rekabet ortamında var olmaktan öte ayakta durmasını sağlayacak ve hatta onu güçlendirecek olan ise, bu verilerden güvenilir ve yararlı bilgiler elde etmesi ve onu uygulayabilmesidir. Bu kapsamda veri madenciliği önem kazanmaktadır. Firmalar tahminlerinin doğru çıkması oranında başarıyı yakalarlar. Buda ancak doğru girdiler ve doğru modellemelerle, doğru süreçleri takip ederek yapılacak bir veri madenciliği tahminleme yöntemiyle sağlanabilir.

Tahminleme yöntemlerinden biri olan karar ağacı uygulamamızda, başarı için model oluşturmanın önemi üzerinde durulmuştur. En iyilenmiş model için girdi verileri özenle işlenmiştir. Dinamik veri setlerinde modelin izlenerek yenilenmesi gerekeceğinden, tüm veri işleme adımlarının tekrar uygulanması yerine, yazılan prosedürler ile otomatik yapılması sağlanmıştır. Hangi girdinin çıktı üzerinde daha etkili olduğu ise tahmini modellemelerle Explain Analizi doğrultusunda belirlenmiştir. Model en iyilenirken; Roc, Lift ve Risk Matrisi Analizleri ile test edilmiştir. Öğrenme ve test sınıflarının ayrılması ayarları, dallandırma

ölçüm algoritması seçilmesi ve düğüm ayarlarının yapılması ile test verilerinin hedef değişkenler üzerinde seçim yöntemleri de modelin başarısı için özenle yapılmıştır.

Tarım Kredi Kooperatifleri için kurulan bu model, kredi başvurusu yapan çiftçi ortaklar için ve genel olarak belirlenmesi gereken kredi politikaları üzerine kurulmuştur. Amaç; gelen ortağın özeliklerine göre, "biz bu krediyi verirsek hangi olasılıkla zamanında geri öder?" ya da "ne yapalım ki ortağın borcunu zamanında ödeme olasılığı artsın?" gibi gelecek için tahminlemeler yapmak ve mevcut durum için ortakların ödeme alışkanlıklarını ve risklerini belirleyerek kredi politikaları ortaya koymaktır. Ayrıca, karar ağacımızdan genel olarak: sermayesi 495 TL den küçük ve kredi borcu 10.000 TL den fazla olan ve belli illerde çiftçilik yapan ortakların kredi geri ödeme alışkanlıklarının çok kötü olduğu, çiftçinin bakmakla yükümlü olduğu kişi sayısının bir tek 8 ve üzerine çıkması durumlarında riskin arttığı, çiftçinin yıllık minimum gelirinde kırılma üst noktasının 15.900 TL ve alt noktasının 4.030 TL olduğu vb. birçok yorum çıkarmak mümkün olmuştur. Böylelikle, bu yorumlardan yola çıkarak, gelecek için yeni politikalar belirlenmesi sağlanabilecektir.

Uygulamamızın diğer yapılan benzer çalışmalardan en önemli farklılığı, farklı bir platformda geliştirilmesi ve model oluşturmaya kadar olan veri madenciliği süreçlerinin farklı yöntemlerle uygulanmasıdır. Zira, sonuçlar üzerine büyük etki edip model başarısını değiştirebilecek kadar sürekli bir değişkenlik gösteren dinamik ve büyük ölçekli verilere sahip farklı veri tabanlarındaki verilerin, modele girdi olana kadar olan tüm işlemleri hazırlanan prosedürler ile sağlanmış ve modelin güncel tutulabilmesi için veri hazırlama işlemi otomatize edilmiştir.

6. Sonuç

Kredi faaliyetlerinin, kurum yatırımlarının, ülke şartlarının, devlet tarım politikalarının, mevsim durumlarının vb. her yıl hatta her ay sürekli değişkenlik göstermesi, geleceğe dair tahmin yürütmek için kurulan veri madenciliği karar ağacı modelinin değişken sayısının değişmesine ve veri setlerinin boyut ve değerlerinde yadsınamayacak değişikliklerin oluşmasına neden olmaktadır. Bu yüzden en hızlı şekilde güncellenebilir bir model oluşturmak gerekmektedir. Verilerin tekrar işlenmesi, değişkenler arası ilişkilerin belirlenmesi, öğrenme ve test verilerinin ayrılması, karar ağacı düğümlerinin ayarlanması ve modelin test edilerek doğrulanması işlemleri tekrarlanmak zorunda kalacaktır. Yapılan çalışmada modele girdi olacak verilerin hazırlanması işlemi prosedürlerle otomatize edilmiştir. Ancak, gelecekte daha verimli sonuçlar elde edilmesi için tüm bu işlemlerin sonuçlarına göre en iyi modeli zamanı gelince kurup güncelleyebilecek uzman sistemlerin geliştirilmesi faydalı olacaktır.

KAYNAKLAR

1. Han, J., Kamber, M.(2006). Data Mining: Concepts and Techniques (second edition), University of Illinois at Urbana-Champaign, Elsevier

2. Kaya, H., Köymen, K. (2008). Veri Madenciliği Kavramı ve Uygulama Alanları, Doğu Anadolu Bölgesi Araştırmalar, Kocaeli-İstanbul

3. Argüden, Y., Erşahin, B. (2008). Veri Madenciliği, Veriden Bilgiye, Masraftan Değere 1st Ed., İstanbul

4. Akgöbek, Ö., Çakır, F. (2009). Veri Madenciliğinde Bir Uzman Sistem Tasarımı, Akademik Bilişim'09, Harran Üniversitesi, Şanlıurfa

5. Data Mining Concepts 11g, Oracle Corporation, http://download.oracle.com/docs/ cd/B28359_01/ datamine.111/b28129/toc.htm, (May 2008).

6. Kayaalp, K. (2007). Asenkron Motorlarda Veri Madenciliği İle Hata Tespiti, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü, Isparta

7. Data Mining Administrator's Guide, Oracle Corporation, http://download.oracle.com/ docs/cd/B28359_01/ datamine.111/b28130/ toc.htm, (August 2008).

8. Tan, P., Steinbach, M., Kumar, V. (2006). Introductionto Data Mining, University of Minnesota:Addisson-Wesley

9. Kamber, M., Winstone, L., Gong, W., Cheng, S., Han, J. (1997). Generalization and Decision Tree Induction: Efficient Classification in Data Mining, Simon Fraser University, B.C., Canada, V5A 1S6

10. Emel, G., Taşkın, Ç. (2000). Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması, Osmangazi Üniversitesi Sosyal Bilimler Dergisi, Cilt 6 Sayı 2, Eskişehir

11. Özekes, S., Çamurcu, Y. (2002). Veri Madenciliğinde Sınıflama ve Kestirim Uygulaması, Marmara Üniversitesi Fen Bilimleri Dergisi, 18, İstanbul

12. Ayık, Z., Özdemir, A., Yavuz, U., Liste Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi

13. Doğan, Ş., Türkoğlu, İ. (2008). Iron-Deficiency Anemia Detection From Hematology Parameters By Using Decision Trees, International Journal of Science&Technology, Cilt 3, No 1, 85-92

14. Bozkır, A.S., Sezer, E. (2009). Usage of Data Mining Techniques in Discovering The Food Consumption Patterns of Students and Employees of University, Balkan-Kafkas

ve Türk Devletleri Uluslararası Mühendislik Sempozyumu, 22-24 October, Isparta, 104-109

15. Güntürkün, F. (2007). A Comprehensive Review Of Data Mining Applications In Quality İmprovement And A Case Study, Yüksek Lisans Tezi, Middle East Technical University, Statistics

16. Gürsoy, Ş., Umman, T. (2010). Customer Churn Analysis in Telecommunication Sector, İstanbul University Journal of the School of Business Administration, Cilt 39, No 1, 35-49

17. Bozkır, A.S., Sezer, E., Gök, B. (2009).Öğrenci Seçme Sınavında (ÖSS) Öğrenci Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti, 5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09), 13-15 Mayıs, Karabük Üniversitesi, Karabük, 37-43

18. Albayrak, A.S., Yılmaz, Ş.K. (2009). Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama, S.D.Ü. İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt 14, No 1, 31-52

19. Ankerst, M., Ester, M., Kriegel, H. P. (1999).Visual Classification: An Interactive Approach to Decision Tree Construction, 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, 392-396

20. Du, W., Zhan, Z. (2002). Building Decision Tree Classier on Private Data, Center for Systems Assurance Department of Electrical Engineering and Computer Science Syracuse University, Syracuse, NY 13244