T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

GSM ŞEBEKELERİNDE SAHTEKARLIK YÖNETİMİ İÇİN VERİ MADENCİLİĞİ YÖNTEMLERİNİN UYGULANMASI

Yüksek Lisans Tezi

HÜLYA TAVACI

İSTANBUL, 2012

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİ TEKNOLOJİLERİ (TÜRKÇE TEZLİ)

GSM ŞEBEKELERİNDE SAHTEKARLIKYÖNETİMİ İÇİN VERİ MADENCİLİĞİ YÖNTEMLERİNİN UYGULANMASI

Yüksek Lisans Tezi

HÜLYA TAVACI

Tez Danışmanı: DOÇ. DR. ADEM KARAHOCA

İSTANBUL, 2012

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİ TEKNOLOJİLERİ (TÜRKÇE TEZLİ)

Tezin Başlığı : GSM Şebekelerinde Sahtekarlık Yönetimi İçin Veri Madenciliği Yöntemlerinin Uygulanması

Öğrencinin Adı Soyadı : Hülya Tavacı
Tez Savunma Tarihi : 07.01.2012

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğu Fen Bilimleri Enstitüsü tarafından onaylanmıştır.

Doç. Dr. F. Tunç BOZBURA

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş oldugunu onaylarım.

Y. Doç. Dr. M. Alper Tunga

Bu tez tarafımızca okunmuş, nitelik ve içerik açısından bir Yüksek Lisans tezi olarak yeterli görülmüş ve kabul edilmiştir.

Jüri Üyeleri	İmzalar
Doç. Dr. Adem Karahoca (Tez Danışmanı)	---
Y. Doç. Dr. M. Alper Tunga	---
Y. Doç. Dr. Yalçın Çekiç	---

ÖNSÖZ

Bilişim çağında yaşadığımız şu günlerde, teknolojinin hızlı gelişmesiyle ortaya çıkan yeniliklerin kullanılmaya ve uygulanmaya başlanması, bu sistemlerin başarı ölçümünün değerlendirilmesi, tezin ana fikrini oluşturmaktadır. Bu tez çalışması, GSM şebekelerinde sahtekarlık yönetimi için, veri madenciliği yöntemlerinin uygulamalarını içermektedir.

Bu çalışma sürecinde yardımlarını ve desteğini esirgemeyen tez danışmanım Sayın Doç. Dr. Adem KARAHOCA'ya, tezin içeriğini oluşturmamda yardımcı olan Sayın Tamer Uçar'a, bu uzun vadeli çalışma boyunca bütün sorularımı cevaplayan, desteğini esirgemeyip moral veren Sayın Ertuğrul Özel'e, elinden gelen bütün yardımlarını esirgemeyen sevgili arkadaşım Banu Baklan' a, beni hiçbir zaman yalnız bırakmayan ve her zaman desteğini hissettiğim ve sıkıntıya düştüğüm anlarda hep yanımda olan aileme teşekkürü bir borç bilir, sonsuz sevgilerimi sunarım.

Ocak 2011 Hülya Tavacı

ÖZET

GSM ŞEBEKELERİNDE SAHTEKARLIK YÖNETİMİ İÇİN VERİ MADENCİLİĞİ YÖNTEMLERİNİN UYGULANMASI

Tavacı, Hülya

Fen Bilimleri Enstitüsü IT Türkçe Tezli

Tez Danışmanı: Doç. Dr. Adem Karahoca

Ağustos 2012, 55 sayfa

Teknolojinin ve yerel ağ sistemlerinin hızlı gelişip yayılması, ağ sistemlerine izinsiz girişi de beraberinde getirdi. Bu zararı önlemek için, şirketler sahip olduğu ağlardaki veri akışının güvenliğini sağlamak amacıyla yeni sistemler geliştirmeye başladılar.

Veri madenciliği tüm iş alanlarında uygulanabilen bir yöntem olsada, sıklıkla finans sektöründe, bankacılıkta, GSM sektöründe ve biomedical alanlarda; sahtekarlık belirleme, müşteri tutma, pazarlama ve risk yönetimi gibi amaçlar için kullanılmaktadır.

Sahtekarlık yönetimi (Fraud Management) alanı, veri madenciliği teknikleri uygulanarak, verilerin anlamlı bilgilere dönüştürülebileceği alanlardan biridir. Geçmişte bu konuyla ilgili bir çok çalışma yapılmıştır. Bu çalışmalara, alan yazım çalışması bölümünde örnekler verilmiştir.

Bu çalışmada; GSM sektörlerindeki sahtekarlık yönetimi için, müşteri verilerinin sınıflandırılması problemi üzerinde durulmuş ve sahtekarlık çeşitlerinden olan abone sahtekarlığı (invoice) incelenmiştir. Müşteri bilgileri olarak; yaş, cinsiyet, abonelik yaşı, ortalama aylık fatura tutarı, ortalama aylık kullanılan sms sayısı, geç ödenen fatura sayısı, son borç durumu, sondan 3. fatura ödemesi, sondan 2. fatura ödemesi, son fatura ödemesi ve fraud mu abone bilgilerine sınıflandırma yöntemleri uygulanmıştır.

Sahtekarlığın kesin tanısının konulmasında ise, abonelerin ödeme ve ödememe durumları göz önünde bulundurulmuştur.

Çalışmada uygulanan sınıflandırma yöntemleri; Weka 3.7.1 (Witten & Frank, 2005) veri madenciliği ara yüzü ile; Karar Ağaçları, Çok Katmanlı Algılayıcı, Bayes Kuralı, Bayesian Ağları, Part, Zeror, Oner, Rbf Ağları' dır. MATLAB 7.8.0 (R2009a) (Moler, 2008) Bulanık Mantık aracı kullanılmıştır. Uygulanan sınıflandırma yöntemleri neticesinde; Bulanık Mantık ile diğer sınıflandırma yöntemlerinin performansları kıyaslanmıştır.

Yapılan çalışmalar sonucunda, Bulanık Mantık çalışmasının performansının diğer sınıflandırma yöntemleri olan, Karar Ağaçları, Çok Katmanlı Algılayıcı, Bayes Kuralı, Bayesian Ağları, Part, Zeror, Oner, Rbf Ağları' na göre daha tutarlı ve güvenilir olduğu gözlemlenmiştir.

Anahtar Kelimeler: Sahtekarlık Yönetimi, Veri madenciliği, GSM Ağları, Bulanık Mantık

ABSTRACT

FRAUD MANAGEMENT APPLICATIONS OF DATA MINING METHODS IN GSM NETWORKS

Tavacı, Hülya

The Institute ofr Sciences IT Türkçe Tezli

Supervisor: Doç. Dr. Adem Karahoca

February 2012, 55 pages

Development of the technology and the rapid spread of the local network caused unauthorized access to network systems. To prevent the loss, the companies begin to develop new systems to ensure safe of data flow owned by the network structures. Today, data mining is a data processing technology that is used to solve many problems.

Data mining can be applied to all business areas such as financial industry, banking, telecom and biomedical fields and it is used for fraud detection, customer retention, amrketing and risk management such purposes.

Fraud Detection is one of the methods that converts raw data to rich data using data mining technique. Significant number of studies was done about this matter. These studies mentioned in the literature review section.

This study is focused on customer data classification problem concerning fraud detection in GSM workplaces and examines “member fraud” (invoice).

Classifications are made under customer information according to age, sex, age of membership, average monthly invoice, number of sms sent in a month, number of paid invoice that exceeded deadline, latest debt status, antepenultimate bill payment, penultimate bill payment, last bill payment and fraud information. Customers’ paid and

unpaid bill information is taken into consideration in order to reach definite diagnosis of fraud.

Weka 3.7.1 (Witten, Frank, 2005), with data mining interface; Decision Trees, Multi-Layer Perceptron, Bayes Rule, Bayesian Networks, Part, Zeror, Oner and Rbf Networks are the classification methods that are used in this study. MATLAB 7.8.0 (R2009a) (Moler, 2008) is used for Fuzzy Logic operations.

After many studies, the performance of Fuzzy Logic study is more reliable and more consistent than Decision Trees, Multi-Layer Perceptron, Bayes Rules, Bayesian Networks, Part, Zeror, Oner, Rbf Networks.

Key Words: Fraud Management, Data Mining, GSM Network, ANFIS

İÇİNDEKİLER

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

KISALTMALAR LİSTESİ

BI	:	Bussiness Intelligence (İş Zekası)
BİT	:	Bilişim ve İletişim Teknolojileri
BT	:	Bilişim Teknolojileri
CRM	:	Customer Relationship Management (Müşteri İlişkileri Yönetimi)
DB	:	DataBase (VeriTabanı)
IS	:	Information Systems (Bilişim Sistemleri)
IT	:	Information Technologies (Bilişim Teknolojileri)
OLAP	:	Online Analytical Process (Çevrimiçi Analitik İşleme)
CDR	:	Call Detail Record (Arama Detayları Kaydı)

1. GİRİŞ

1.1 PROBLEM TANIMI

Son yıllarda, verinin bilgiye dönüştürülmesi, oldukça önem taşımakta ve veri madenciliği teknolojisinin kullanımını arttırmaktadır. Farklı süreçlerde üretilen geçmiş verilerin, depolanması, ayrıştırılması ve gruplandırılması sonucunda elde edilen bilgi küpleri, veri ambarlarının oluşturulmasında temel nesneler olmakla birlikte, kurumların farklı iş süreçlerinde geleceğe yönelik kestirim, müşteri analizleri, maliyetler ve doğru hizmet için organizasyonlar, veri madenciliğinden büyük faydalar sağlamaktadırlar. Veri madenciliği kavramı; bankacılık sektörü, üretim sektörü, finans sektörü ve telekomünikasyon olmak üzere birçok sektörlerde uygulama alanları bulmaktadır.

Teknolojinin gelişmesiyle oluşan, diğer yeni bir kavram ise sahtekarlık yönetimidir. Organizasyonlarda yapılan günlük işlemler, elektronik ortamlara taşındıkça bununla ilgili tehditler oluşmaktadır. Ayrıca telekomünikasyon sektöründeki kuruluşların maliyetleri, bu gibi casus yazılımlar nedeniyle, artmaktadır. Şirketlere yapılan bu saldırıların artması nedeniyle, bilgi güvenliği daha da önem kazanmıştır. Kuruluşlara yapılan bu saldırıların engellenmesi için, arama detayı kayıtları (CDR) analiz edilir ve sahtekarlık kriterleri belirlenir.

Bu çalışmanın temel amacı; veri madenciliği modelleri yardımı ile telekomünikasyon sektöründeki dataların analiz edilmesiyle, GSM ağlarına yapılan abone (invoice) saldırılarını tespit etmektir. Abone sahtekarlığı (Subscription Fraud); basit anlamıyla abonelerin ücret ödememe ve operatörlerin sunduğu ek servislerden ücret ödememe niyetiyle yaptığı görüşmelerdir.

Bu tez çalışmasında, abonelerin fatura ödemelerini yapmadıkları durumda oluşan sahtekarlık durumunu analiz edilip, modellemeye yönelik adımlar yürütülmektedir. Modellemeye göre, abonelerin üst üste ikiden fazla faturasını ödememeleri durumu, sahtekar olarak değerlendirilecektir.

Çalışmada; Weka 3.7.1 (Witten & Frank, 2005) veri madenciliği sınıflandırma yöntemleri ve MATLAB 7.8.0 (R2009a) (Moler, 2008) uygulamasında yeralan ANFIS yöntemi ile analizler yapılmıştır. Yapılan çalışmalar neticesinde, sınıflandırma yöntemlerininm performansları ve güvenilirlikleri değendirilecektir.

Yapılan analiz göz önünde bulundurularak; yaş, cinsiyet, il, abonelik yaşı, geç ödenen fatura sayısı, aylık ortalama fatura tutarı, son borç durumu gibi bütün müşteri bilgilerine göre kurallar belirlenir. Bu kurallar belirlenirken analiz aşamasına önem verilmelidir. Yanlış uygulanması durumunda şirket için değerli müşterilerin engellenmesine ve şirketin ciddi zararlara uğramasına yol açabilir. Bu sebeple çalışmalarda uygun veri madenciliği yönteminin kullanılması oldukça önem taşımaktadır. Bu çalışmada bu gibi durumlar göz önünde bulundurularak datalara uygun olan ve performansı yüsek çıkan veri madenciliği sınıflandırma yöntemleri içerisinde yer alan; Bayesian Network, Multiplayer Perceptron, JRIP, Oner, Zeror, Kısmi Karar Ağaçları, Naive Bayes, Part, Rbf Network ve ANFIS kuralları uygulanmış ve bu uygumaların performansları kıyaslanmıştır.

1.2 GSM SAHTEKARLIK TİPLERİ

Teknolojinin gelişmesi ve iletişim teknolojilerinin de içiçe çalışması nedeniyle güvenlik sorunları artmıştır. Özellikle telekomünikasyon, finans, sigorta gibi sektörlerde saldırılar çoğalmaktadır. Bu gibi saldırılar, müşteri yakınlarının saldırıları, kullanıcıların sahte olması gibi nedenler fraud detection yönetiminin oluşmasını gerektirmiştir. Birçok şirket tarafından yöntemler geliştirilmekte ya da satın alınmaktadır. Sahtekerlık Yönetimi veri madenciliği yöntemi ile oluşan saldırıları, tehditleri ve yasa dışı durumların belirlenmesini sağlar. Şirket yapılarına göre farklı çeşitlerde sahtekarlıklar görülmektedir. Bunların içinden en çok rastlanan sahtekarlık tipleri (Taşpınar, 2010);

i. **Abone Sahtekarlığı (Subscription)**; Basit anlamıyla abonelerin ücret ödememe ve operatörlerin sunduğu ek servislerden ücret ödememe niyetiyle yaptığı görüşmelerdir.

ii. **Arama Kartlar(Calling Card);** Basit anlamıyla aboneye ait calling card' ların çalınması ya da kart numarasının illegal olarak ele geçirilmesi ve kullanımına yönelik fraud tipidir.

iii. **Kopyalama(Cloning);** Kullanılan kartların kopyalanması ve yeniden doldurulması ile ilgili yapılan hilelerdir. Bu gün teknolojik olarak smart kartlar kopyalanamamakla birlikte, yarın bunun kopyalanmamasını kimse garanti edemez.

iv. **Tıkınma(Cramming);** Özellikle Fixed (Sabit) Operatörlerde ortaya çıkan bir sorundur. Bazen aboneler ücretsiz veya çok ucuz servis aldıklarını zannederek hizmet alırlar. Ancak, karşılarına çok ciddi rakamlara ulaşmış faturalar çıkar. Cramming sahtekarlık olarak nitelendirilemese bile, Fraud Yönetim içersinde mücadele edilmesi gereken bir sorundur.

v. **Haksız Rekabet (Call Celling);** Call Celling de Cramming gibi, İletişim sektöründe Fraud Management içersinde mücadele edilmesi gereken bir haksız rekabet uygulamasıdır. Operatör olmamasına rağmen, operatör olarak

davranarak, özellikle yurt dışı, konferans ve yönlendirme çağrılarının daha ucuza yaptırılmasıdır.

vi. **İllegal Erişim(Hacking);** Fiziksel Network' e İllegal erişim (PBX veya Voicemail sistemler yoluyla anormal şekilde) yöntemidir.

vii. **İç Sahtekarlık(Internal Fraud);** Şirket içi istismarlar. Santralar den, teknisyenlerin abonelerin hattını illegal erişimler, bu tip yöntemler içine girer.

viii. **Sahte Satıcı Sahtekarlığı (Dealer Fraud);** Telekom adına satıcılık yapanların yaptığı sahtekarlıklar.

ix. **Ön Ödemeli Sahtekarlık(Prepaid Fraud);** Ön ödemeli görüşmelerde karşılaşılan sahtekarlıklar.

2. İŞ ZEKASI

Günümüzde globalleşme, rekabetin artması, teknolojinin gelişmesi gibi değişikliklerin yaşanması bilgiye verilen önemi arttırmıştır. Bu durum şirketlerdeki bilgi teknolojileri bölümlerini oldukça meşgul etmeye başlamıştır. Ayrıca şirketler, verinin önemimin artmasından dolayı verinin işleme sürecine oldukça yatırım yapmaktadırlar. Verilerin toplanması, saklanması, işlenmesi, bu bilgilerin analiz edilmesi, elde edilen bilgilere göre geleceğe yönelik iş stratejilerinin oluşturulması iş zekası şekllinde adlandırılır. Gartner'a göre; iş zekası gelişmiş ve haberdar edilmiş karar vermeye yol gösteren, veriye ulaşmak, veriyi keşfetmek, veriyi analiz etmek, iç yapıyı geliştirip anlamak gibi işlemleri içeren kullanıcı merkezli bir yapıdır. (Şimşek 2006).

2.1 İŞ ZEKASININ SUNDUĞU ANALİTİK ÇÖZÜMLER

İş zekasında kullanılan araçlar şirketlerin bilgiyi gerektiği şekilde kullanabilmesini ve rekabet ortamından sıyrılmalarını sağlar. İş zekası belirli hedefler doğrultusunda uygulandığında cevap verir ve hedeflenen kullanıcılara hitap edecek şekilde tasarlanmalıdır. Maddeler halinde sıralamak gerekirse iş zekasının sunduğu analitik çözümler aşağıdaki maddelerden oluşmaktadır (Niu&Lu&Zhang, 2009).

i. Entegrasyon, Veri Dönüştürme Uygulamaları

ii. Raporlama, Sorgulama

iii. Veri ambarı, modelleme, tasarım, geliştirme

iv. Dashboard, BI Platformları

v. Veri Madenciliği, Uygulama Geliştirme

vi. Veri Oluşturma

vii. OLAP Teknoloji Platformları üzerinde Uygulama Geliştirme

2.2 İŞ ZEKASI BİLEŞENLERİ

İş zekası bileşenleri, organizasyonlarda iş zekası projelerinin verimli bir şekilde gerçekleştirilmesi için gerekli bilgileri içermektedir. Aşağıda sıralanmış olan bileşenler aynı zamanda, bir iş zekası projesinin döngüsü yada aşamalarını da destekler niteliktedir. Burada belirtilen bileşenlerin tamamı bir iş zekası projesinde yer alabileceği gibi bazı bileşenlere proje içerisinde ihtiyaç duyulmayabilir ya da zaten mevcut yapı içerisinde var olduğundan proje için daha hızlı bir ilerleme söz konusu olabilir (Niu&Lu&Zhang, 2009).

2.2.1 İş Verisi Kaynakları

İş verisi kaynakları; ilişkisel veri tabanları, flat file'lar, xml verileri gibi kaynaklardan oluşabilir. Teknolojik gelişmelere ve gerçekleştirilecek projeye göre bir palm cihazı yada bir buzdolabında ki sensörler de bizim için bir kaynak olarak görülebilir.

2.2.2 Veri Ambarı

Yeni teknolojilerin oluşmasıyla şirketlerin karar alma süreçleri hızlanmış, karmaşık sistemlerden daha çok veri bütünlüğü önem kazanmış; bu durumda veri ambarının oluşmasına yol açmıştır. Veri ambarları veri bütünlüğünü koruyarak bilgiye daha kolay ulaşılmasını sağlar. Ayrıca verinin toplanması, birleştirilmesi, dönüştürülmesi ve yorumlanması süreçlerini içerir. Bu yöntemle şirketler stratejik amaçlarını oluştururlar. Veri ambarları; analizin oluşturulmasını, verilerin depolanmasını, veri modellerinin oluşturulmasını, raporlama yapılmasını, verilerin temizlenmesini, uygulamanın geliştirilmesini sağlamaktadır (Çağıltay, 2010).

2.2.3 Çevrimiçi Analitik İşleme (OLAP)

Verilere daha hızlı erişebilmek ve ilişkisel veritabanlarından faydalı bilgiye ulaşabilmek için oluşturulan bir teknolojidir. Şirketlerin sahip oldukları verileri birçok bakış açısına göre değerlendiren bir teknolojidir. Küp mantığını barındırır. Veri ambarlarındaki verileri işleyip şirketler için iş modellerinin oluşturulmasına katkı sağlayan analiz aracıdır. Olap yapısının bir çok faydaları vardır (Şimşek, 2006).

2.2.3.1 OLAP'ın Faydaları:

- Küp mantığı ile çalışır. Analiz sürecinde çeşitli boyutlarda incelenmek istenen parametreleri raporlama tekniği ile gerçekleştirebilir.
- Veri tabanından sorgu çekebilir.
- Karmaşık hesaplamaları kolayca çözebilir.
- Zaman kavramını göz önünde bulundurarak yapılması istenen analizleri yüzdesel olarak hesaplayabilir. OLAP teknolojisi ile zaman boyutu veri küplerinde kullanılabilir.

2.2.4 İş Zekası Uygulamaları

İş zekası uygulamalarına; şirketlerin departman bazlı yaptıkları raporlamalarda, web tasarım uygulamalarında, veri analizlerinde ve operasyon bölümlerindeki sorgu tasarlamalarda ihtiyaç duyulmaktadır.

İş zekası, veritabanlarındaki dataların işlenmesini ve geçmiş bilgilerden gelecek ile ilgili öngörüde bulunarak şirketlerin karar verme sürecini kolaylaştırır. İş zekası araçları, buradaki bileşenler, kurumlarda gelişmiş bilgi yönetimi mimarilerini, karar vermeyi desteklemek ve düzenlemek amacıyla kullanmaktadır.

Karar destek analizleri, karar vermeyi mevcut bilgi keşfi araçlarıyla kolay hale getirmektedir. Bu kabiliyetler ad-hoc sorgulama sistemleriyle gerekli kullanıcıların analiz yapabilmelerini sağlamaktadır. (Ad-Hoc sorgular sadece bir defaya mahsus yapılan sorgulardır.)

Burada ki araçlar temel olarak; detaya inebilen basit sorgulama araçları (drill-down), veri madenciliği yeteneklerine sahip basit yapay zeka araçlarından oluşmaktadır.

2.2.5 Veri Madenciliği

Globalleşmenin çevre koşullarının etkisi, internetin yaygınlaşması, teknolojinin gelişmesi artık bilginin veriye nazaran daha önemli olduğunu vurgulamaktadır. Veri madenciliği, gerekli bilgiyi saklama, gerektiği zaman hızlı bir şekilde bilgiye ulaşma işidir. Amaç, şirketler için önem taşıyan bilgilere kümeleme yaparak görsel sunumlara çevrilebilecek iş modellerinin oluşturulmasıdır. Büyük ölçüde bulunan ham bilgi gerekli veriye dönüştürülür. Büyük ölçüdeki veri tek bir iş istasyonuna sığmayacak kadar geniş olduğunu ifade etmektedir. Dönüştürülen bilgi, veritabanlarında saklanıp, amaçlar doğrultusunda çağrılarak veri madenciliği için standart formlara çevrilir. Oluşan veri OLAP (Online Analytical Processing) ile analiz edilir. OLAP ile küpler oluşturulur ve ortaya çıkan analiz sonrasında gruplama ve raporlama yapılabilir. Diğer bir deyişle bu yöntemle geçmiş verilerden gelecek ile ilgili tahmin yapma şansını elde edebilmemizi sağlar. Alternatif olarak veri madenciliği aslında bilgi keşfi sürecinin bir parçası şeklinde kabul görmektedir. Bigi keşfi gereksinimleri aşağıdaki süreçleri takip eder (Çağıltay, 2010).

i. Uygulama alanının incelemesi, analizin gerçekleştirilmesi (hedef ve problemlerin belirlenmesi)

ii. Veri temizleme işleminin yapılması (gerekli bilgilerin çağrılması)

iii. Veri kaynaklarının kullanılmasıyla verinin bütünleştirilmesi

iv. Veri kümesinin oluşturulması- Veri seçme (Analiz ile gerekli olan verileri belirlemek)

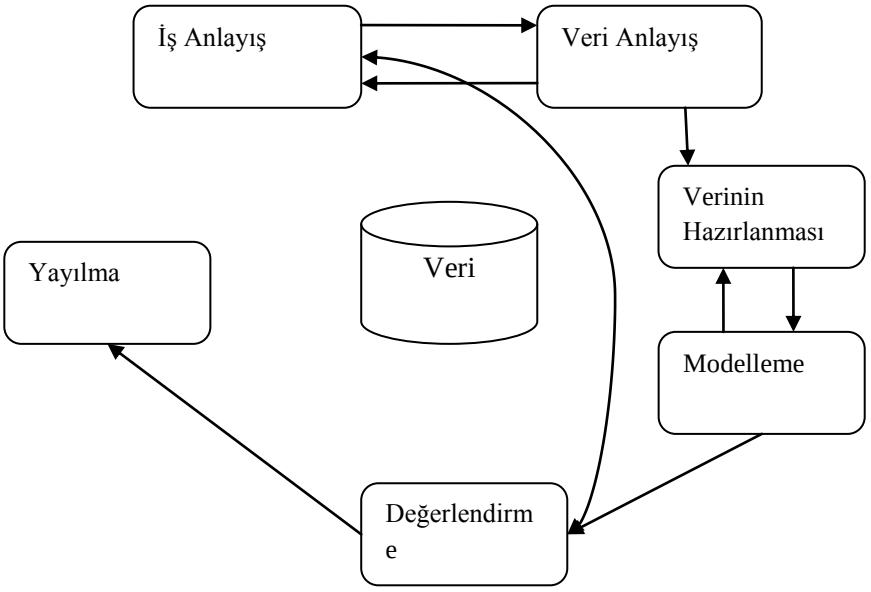
v. Veri Dönüşümü (Toplanan verinin azaltılarak veri madenciliği ile dönüşümünün gerçekleştirilmesi)

vi. Veri madenciliği algoritması ile en uygun modelleme metotların uygulanması

vii. Değerlendirme (Bazı ölçümlere göre belirlenmiş modellerin ve bilgiyi temsil eden ilginç örüntülerin tanımlanması ve hedefler doğrultusunda incelenmesi)

viii. Bilgi sunumunun gerçekleştirilmesi (Bulunan bilginin değerlendirilip yorumlanması, kullanıcıya sunumunun yapılması ve projenin oluşturulması.

Veri madenciliğinde birçok modelleme tekniği kullanılır. Analiz aşamasından sonra uygun modelleme metotlarına karar verilir. Belirlenen hedefler doğrultusunda iş stratejileri oluşturulur ve oluşan modeler ilgili parametrelerle birleştirilir. Dikkate alınması gereken diğer bir durum ise ülkemizde çok fazla kullanılmayan bu yöntemin bir takım sonuçlar doğurmasıdır. Öncelikle maliyetli oluşu üst düzey yöneticileri düşündürmektedir. Yanlış bir adım sonucu telafi edilemeyecek sonuçlar doğurması olasıdır. Bu sebeple üst düzey yöneticilerin tam anlamıyla desteği gerekmekte ve tam ve uygun bir sürecin işlenmesi sağlanmalıdır. Uygulanacak süreçlerin iyi belirlenmesi, planlanması ve hedefler doğrultusunda uygun bir şekilde uygulanması gerekmektedir.



Şekil 2.1: Veri Madenciliği Süreci

3. ALAN YAZIM ÇALIŞMASI

Birçok organizasyonun dinamik iş ortamı, zamana duyarlı iş fırsatlarından yararlanmak için gerçek zamanda kendi süreçlerini izlemek durumundadır. Değişen iş çevresine ayak uydurma ve yorumlama yeteneği BT açısından işlerin hızlı ve etkili yürümesi için önem taşımaktadır. Fakat İş Zekası ve Veri Depolama Teknikleri doğrudan analitik gereksinimleri etkilemez. Bu çalışma bütün süreçleri kapsayan, yorumlayan, tanımlayan, tahmin yürüten, iş kararları için gerekli reaksiyon süresini azaltan gelişmiş İş Zekası mimarisi sunmaktadır. BT teknolojilerini kullanarak mobil telefon sahtekarlık yönetimi alanından, önerilen yaklaşımı örnekleyen prototip seçilmiştir. İş süreçleri ve iş zekası arasındaki operasyonları etkili bir şekilde gerçekleştirip incelemektedir (Nguyen & Schiefer & Tjoa, 2005).

En yaygın dolandırıcılık türlerinden olan abone dolandırıcılığı (subscription fraud) incelenmiştir. Bu çalışmanın amacı bilgi keşfi sürecini inceleyip, veri madenciliği tekniklerini kullanarak abone dolandırıcılığını tespit etmektir. Bu çalışmada, kümeleme ve sınıflandırma aşamaları uygulanmıştır. Kümeleme aşamasında SOM ve K-anlamı, sınıflandırma aşamasında ise karar ağaçları, yapay sinir ağları, destek vektör makineleri incelenmiştir. Bu çalışmada Tahran Telekomünikasyon Şirketi tarafından, önerilen yöntemin doğruluğunu göstermek için bir veri kümesi kullanılmıştır. Bu şirket tarafından geliştirilen veri kümesi, önerilen yöntem üzerine uygulanmıştır ve etkinliği ve istatistik değerleri incelenmiştir. Ortaya çıkan sonuç sınıflandırma yöntemiyle ağaçların performansını arttırdığı yönündedir. Araştırma bulguları ortaya çıkan modelin, yapılan ölçümlerin yüksek doğruluk oranına sahip olduğunu göstermektedir (Farvaresh & Sepehri, 2011).

Mobil Telekom operatörleri arasındaki rekabetin artması iş alanlarının çeşitlenmesine yol açmıştır. Özellikle mobil operatörler geleneksel sesli iletişimden kullanıcı başına ortalama geliri oluşturmak için yeni bir teknoloji olan değer katan servislere dönmeye başlamışlardır. Bunun anlamı da çapraz satışın gelirleri arttırmak ve kar yapmak adına Telekom operatörlerinin ciddi öneme sahip olmasıdır. Bu çalışma mobil Telekom pazarında satışları kolaylaştırmak için müşteri sınıflandırma modeli önermektedir. Modelde iki adım uygulanmıştır ve çeşitli veri madenciliği teknikleri kullanılmıştır. İlk adımda regresyon, yapay sinir ağları ve karar ağaçları gibi çeşitli sınıflandırma teknikleri, tahmini sonuçlar üreten modellere uygulanmıştır. İkinci adımda ise model Genel Algoritma (GA)'yı kullanarak hedef müşteriyi kararlaştırmıştır. Modelin kullanılabilirliğini doğrulamak için model, Kore'deki mobil Telekom şirketine uygulanmıştır. Sonuç olarak model çapraz satış için yüksek kalitede bilgi üretmektedir. Ayrıca ikinci adımda uygulanan GA' nın performansı arttırdığı gözlemlenmiştir (Ahn & Ahn & Oh & Kim, 2011).

Telekomünikasyon sektöründeki sahtekarlıkları yakalamak yüksek seviyeli arama trafiğindeki sahte aramalarda kimlik tespitini gerektirir. Buradan yola çıkarak bu etkili ve verimli algoritmaların tasarımı telekomünikasyon sahtekarlığıyla mücadelede önemli bir araştırmayla meydan okuyor. Bu çalışma, kullanıcı kimlik imzası oluşturmak için LDA' yı destekler. Kullanıcı aktivitesi yüksek arama hatlarında, kullanıcı çağrısını davaya sunmasıyla yükümlü olarak yüksek aramalarda olası dağılım olarak tanımlanır. Bu olası dağılım, LDA den türemiş olan, farklı sınıflara ait dağıtılmış kullanıcı profillerinin tam anlamıyla kombinasyonu olarak tanımlanabilir. Skor aramalar için, aynı çağrıyı üreten bir dolandırıcı ile karşı bir çağrı açan kullanıcının olasılığını karşılaştırır. Bu çalışma sahte aramaların algılanmasında olası dağılım kullanarak sorunun düzeldiğini göstermektedir. Bu yöntem ile saldırılar etkili şekilde hesaplanıp yakalanabilir (Xing & Girolami, 2007).

Bu çalışma uzun mesafeli taşıyıcılarda, yüksek etkili telekomünikasyonlarda abone dolandırıcılığını önlemek amacıyla bir sistem önermektedir. Sistem sınıflandırma ve tahmin modülü olmak üzere iki kısımdan oluşmaktadır. Sınıflandırma modülü geçmiş davranışları göz önüne alarak 4 gruba ayrılır: Abone sahtekarlığı, diğer sahtekarlık, sahtelik ve normal. Tahmin modülü saldırı sırasındaki müşterileri tanımlamamızı sağlamaktadır. Sınıflandırma modülü bulanık kurallar kullanılarak geliştirilmiştir. Bu çalışma Chile'deki 10.000'den fazla abonenin bulunduğu telekomünikasyon şirketinin veri tabanına uygulanmıştır. Bu veri tabanında yüzde 2.2 sahtekarlığın olduğu belirlenmiştir. Tahmin modülü çok katmanlı algılayıcı sinir ağları şeklinde uygulanmıştır. Bu test, gerçek dolandırıcıların yüzde 56.2' si sadece abonelerin yüzde 3.5' üni tarayarak tanımlamaktadır. Bu çalışma, uygulama zamanı dikkate alınarak müşteri bilgilerinin analiz edilmesiyle telekomünikasyon sektöründeki saldırıların önlenebileceğini göstermektedir (Este´vez & Held & Perez, 2006).

Bu çalışma uzun mesafeli taşıyıcılarda yüksek etkili telekomünikasyonlarda abone dolandırıcılığını önlemek amacıyla bir sistem önermektedir. Sistem sınıflandırma ve tahmin modülü olmak üzere iki kısımdan oluşmaktadır. Sınıflandırma modülü geçmiş davranışları göz önüne alarak 4 gruba ayrılır: Abone sahtekarlığı, diğer sahtekarlık, sahtelik ve normal. Tahmin modülü saldırı sırasındaki müşterileri tanımlamamızı sağlamaktadır. Sınıflandırma modülü bulanık kurallar kullanılarak geliştirilmiştir. Bu çalışma Chile'deki 10.000'den fazla abonenin bulunduğu telekomünikasyon şirketinin veri tabanına uygulanmıştır. Bu veri tabanında yüzde 2.2 sahtekarlığın olduğu belirlenmiştir. Tahmin modülü çok katmanlı algılayıcı sinir ağları şeklinde uygulanmıştır. Bu test, gerçek dolandırıcıların yüzde 56.2' sini, sadece abonelerin yüzde 3.5' ini tarayarak tanımlamaktadır. Bu çalışma, uygulama zamanı dikkate alınarak müşteri bilgilerinin analiz edilmesiyle telekomünikasyon sektöründeki saldırıların önlenebileceğini göstermektedir (Este´vez & Held & Perez, 2006).

4. ARAŞTIRMA YÖNTEMLERİ

4.1 TEZDE KULLANILAN TELEKOM DATALARININ İÇERİĞİ

Bu çalışmadaki datalar bir telekom şirketine ait müşterilerden oluşan 2448 kayıttan oluşmaktadır ve içerisinde 14 farklı değişken mevcuttur. Bu değişlkenler tablo 4.1 de gösterilmektedir.

Tablo 4.1: Değişken Listesi

CİNSİYET	ORTALAMA AYLIK FATURA TUTARI
YAŞ	AYLIK ORTALAMA KULLANILAN SMS SAYISI
İL	SON BORC DURUMU
ABONELIK YASI-AY	SONDAN 3. FATURA ODEMESI
OPERATOR DEGISIKLIGI-ESKI OPERATOR	SONDAN 2. FATURA ODEMESI
KAC OPERATORDE ABONELIGI VAR	SON FATURA ODEMESI
MEVCUT OPERATORU	FRAUD MU
GEC ÖDENEN FATURA SAYISI	

Dataların aldığı değerler ve sayısal değerlere çevrilmiş hali ise tablo 4.2 de gösterilmiştir. Uygulanacak bazı modellerde sayısal datalar kullanılacak, bazılarında ise doğrudan aldığı değerlerle analiz edilecektir.

Tablo 4.2: Ayrık Zamanlı Değişken Listesi

Değişken Adı	Data Tipi	Değerler
CINSIYET	Boolean	ERKEK=0, BAYAN=1
YAS	Integer	18-24=1, 25-32=2, 33-40=3, 41-45=4, 46-51=5, 52-57=6, 58+=7
IL	Integer	ISTANBUL=34, BURSA =16, KONYA =42, ESKISEHIR = 26, MARDIN=47, ADANA =01, BOLU=14, AYDIN =09, SAKARYA=54, KAYSERI=38, GIRESUN=28, TRABZON=61
ABONELIK YASI-AY	Integer	0-6=0, 6-12=1, 12-24=2, 24-36=3, 36-48=4, 48-60=5, 60-72>6
OPERATOR DEGISIKLIGI-ESKI OPERATOR	Integer	AVEA=0, VODAFONE=1, TURKCELL=2
KAC OPERATORDE ABONELIGI VAR	Integer	1=0, 2=1, 3=2
MEVCUT OPERATORU	Integer	AVEA=0, VODAFONE=1,

		TURKCELL=2
GEC ÖDENEN FATURA SAYISI	Integer	1=0, 2=1, 3=2, 4=3, 5=4
ORTALAMA AYLIK FATURA TUTARI	Integer	10-30=0, 30-45=1, 45-60=2, 60-75=3, 75-100=4, 100-125=5, 125-150=6, 150>7
AYLIK ORTALAMA KULLANILAN SMS SAYISI	Integer	10-30=0, 30-45=1, 45-60=2, 60-75=3, 75-100=4, 100-125=5, 125-150=6, 150>7
SON BORC DURUMU	Boolean	10-30=0, 30-45=1, 45-60=2, 60-75=3, 75-100=4, 100-125=5, 125-150=6, 150>7
SONDAN 3. FATURA ODEMESI	Boolean	ODEDI=0, ODEMEDI=1
SONDAN 2. FATURA ODEMESI	Boolean	ODEDI=0, ODEMEDI=1
SON FATURA ODEMESI	Boolean	ODEDI=0, ODEMEDI=1
FRAUD MU	Boolean	EVET=0, HAYIR=1

Tablo 4.3: Sıralama Değerleri

Sıralama Değeri	Değişkenler
0.40122	AYLIK ORTALAMA KULLANILAN SMS SAYISI
0.39765	SON BORC DURUMU
0.33073	YAS
0.32071	ABONELIK YASI-AY
0.24238	ORTALAMA AYLIK FATURA TUTARI
0.21843	SONDAN 2. FATURA ODEMESI
0.18139	SON FATURA ODEMESI
0.11595	IL
0.09198	GEC ÖDENEN FATURA SAYISI
0.01896	FATURA ODEMESI
0.01279	MEVCUT OPERATORU
0.00651	OPERATOR DEGISIKLIGI-ESKI OPERATOR
0.00483	KAÇ OPERATORDE ABONELIGI VAR
0.00111	CINSIYET

Dataların sahtekarlık etkilerini belirlemek için, Weka 3.7.1 (Witten & Frank 2005) platformunda, sıralama (Ranker) yöntemi ve değerlendirme grubundan da InfoGainAttributeEval seçilerek, Tablo 4.3' teki değerler elde edilmiştir. İnfoGaninAttributeEval fonksiyonu, her bir değişkenin sınıflandırmaya etkisinin gücünü ölçmektedir. Aşağıdaki tabloda, değişkenlerin sahtekarlığa etkileri gösterilmektedir. Sıralama değerlerinin büyüklüğü, değişkenlerin sahtekarlık üzerindeki etkilerinin gücünü göstermektedir. Bu tabloya göre, bir abonenin sahtekar olmasını en fazla etkileyen niteliği, aylık ortalama kullanılan sms sayısıdır. Sahtekar olmasına en az etki eden niteliği ise, cinsiyet ve kaç operatörde aboneliği olduğudur.

4.2 UYGULANAN SINIFLANDIRMA YÖNTEMLERİ

4.2.1 Bayesian Ağları

Bayesian ağları, lojistik regresyon modelleri gibi ağ çıkışı olarak olasılık tahminleri üretir. Tahmin sistemin kendisi tarafından üretilmez. Sistemin temel amacı ise, her sınıf değeri için, sınıf modeline uygun olup olmadığına dair örnek olasılık tahmininde bulunmaktır. Olasılık tahminleri ile normal tahminleri karşılaştıracak olursak, olasılık tahminlerinin normal tahminlerden daha kullanışlı ve yararlı olduğunu, normal tahminlerinde, olasılık tahminleri ile sıralama değerlerinin verildiğinden çıkarabiliriz. Bayesian ağlarında, verilen bir sınıflama niteliklerinin koşullu olasılık değerleri başka bir sınıflama nitelikleri içinde tahmin ediliyor (Witten & Frank 2005).

Aşağıdaki formülde A, B, C bilinen olaylarına göre, X değerinin tahmininde bulunuluyor.

$$P(X|A,B,C) = \frac{P(A|X)P(B|X)P(C|X)P(X)}{P(A,B,C)} \quad \textbf{(4.1)}$$

4.2.2 Çok Katmanlı Algılayıcı (Multiplayer Perceptron)

Yapay sinir ağları, matematiksel modellere dayalı bir simülasyon sistemidir. Bu sistemler, çalışma prensiplerini biyolojik sinir ağlarından esinlenerek almıştır. Yapay sinir ağları, temelde doğrusal olmayan istatistiksel veri modelleme araçlarıdır. Genellikle birçok girişi (her biri farklı öneme sahip) ve yalnızca bir çıkışı vardır. Bir sinir ağı, birden fazla katmandan oluşmaktadır. Bu katmanlar çoğunlukla, girdi katmanı, gizli katman ve çıkış katmanından oluşmaktadır. Giriş katmanında, ağ değişkenlerini vektör değerlerinden alır. Gizli katmanda, her bir girdi, kendi değeri ile çarpılır ve sonuçlar toplanarak yeni değerler elde edilir. Sonra, bu değer fonksiyonunun çıktısını besler (Witten, Frank 2005).

Çok katmanlı algılayıcı, ileri beslemeli yapıya sahip olan yapay sinir ağılarıdır. İleri beslemeden kasıt, değerler sadece ağ katmanları üzerinden hareket eder, iç katmanlarda çıktı değerlerinin geri beslemeleri yapılmaz. Çok katmanlı Algılayıcı Ağlarının bir giriş ve bir çıkış katmanına sahip olması gerekmektedir. Fakat gizli katmanların sayısı ağ mimarisi nedeniyle değişebilir.

4.2.3 Ripper Algoritma (JRIP)

Bu algoritmada, sınıflar kendi büyüklüklerine göre düşünülür ve artarak azalan hata kırpma ripper algoritmasının ana prensibidir. Her sınıf için ayrı kural kümeleri oluşturduktan sonra, her bir kural içinde iki model üretilmektedir. Yine, azalan hata kırpma kullanarak diğer sınıflar tarafından kullanılan örneklem ortadan kaldırılır. Bu uygulamadan sonra eğer değişkenlerden biri orjinal kuraldan daha iyi ise kural ile değişken yer değiştirir (Witten, Frank 2005).

4.2.4 Kısmi Karar Ağaçları (Partial Desicion Trees)

Veri madenciliği sınıflandırma yöntemleri arasından en sık ismini duyduğumuz modelleme karar ağaçlarıdır. Sonuca ulaşabilmek için değişkenler arasındaki ilişkilere tahminleme yöntemleri uygulayarak, ağaç yapısı oluşturulmaktadır. Karar ağaçları düğümler ve dallardan oluşan, anlaşılması oldukça kolay olan bir tekniktir.

Karar ağacında bulunan her bir dalın belirli bir olasılığı mevcuttur. Bu sayede son dallardan köke veya istediğimiz yere ulaşana kadar olasılıkları hesaplamamız mümkündür. Çıktı olarak bize sağladığı ağaçlandırma yapısı sayesinde yorumlanması ve anlaşılması oldukça kolaydır. Karar ağaçları eğiticili öğrenme için çok yaygın bir yöntemdir. Algoritmanın işleyişinde öncelikle bir öğrenme kümesi oluşturulur ve daha sonra bu kümedeki verilerin özellikleri ile ağaç yapısındaki düğümler oluşturlur, ardından da çocuk düğümleri ve yaprakları ile ağaç yapısı oluşturularak alt veri kümeleri elde edilir. En son aşamada ise artık ayrılacak bir özellik kalmadığı için ağaç yapısı kendisini sonlandırır. (Koyuncugil 2010):

Ağaç yapısı dallarına ayrılırken hiçbir şekilde veri kaybı yaşanmamaktadır. Karar ağaçlarındaki yapının nasıl oluştuğunu ve nasıl kullanıldığını anlamak oldukça kolaydır ve yapılan tahminler de ağaçlandırma yapılarına dahil edilebilirler (Koyuncugil, 2010).

Karar ağacı oluşturmanın temel adımları; dataları analiz etme, seçenekleri belirleme, karar ağacını yapılandırma, olasılıkları hesaplama, beklenen değer veya beklenen yararların hesaplanması ve şans düğümlerine yazılması ve sonuçları duyarlılık analizi kullanarak test etmedir. Karar Ağaçları olasılıkları hesaplarken; entropi, kazanç ve hata oranlarını ölçerek tahminde bulunur.

Burada entropi, ağacın dallarını belirlemeyi sağlayan bir algortimadır. Entropi, eldeki verilerin birbirinden farklılığına dayalı bir formül olarak düşünülürse aynı değere sahip 2 ifadenin entropi' leri 0 çıkmaktadır. Entropi formülsel olarak şağıdaki gibi gösterilir.

(4.2)

Entropi değişkenin dallanmasına göre farklı değerleri almaktadır. Bu değerlerin farkına kazanım denir ve hangi değişkenin kazanımı fazla ise bu değişkene göre de ağacı dallandırmaktadır. Kazanım formülü aşağıdaki gibidir.

(4.3)

4.2.5 Bayes Kuralı

Naive Bayes algoritmasında her kriterin sonuca olan etkilerinin olasılık olarak hesaplanması temeline dayanmaktadır. Bayes kuralı olasılıksal çıkarıma dayanır. Eldeki verilerin ve hipotezlerin doğru olma olasılığına göre hareket ederek, gelen verilere göre maksimum olasılığa sahip hipotez seçilir.

4.2.6 Oner Kuralı

Oner algoritması adını Tek Kural' (One Rule) ın baş harflerinden almıştır. Basit ve güvenilir bir sınıflandırma modeli olan Oner, verilerdeki her bir tahmin için bir kural oluşturur. Daha sonra bu kurallardan hata oranı en düşük olanı yani, gerçekleşme olasılığı en yüksek olan tahmini seçmiş olur. Böylece elimizde tek kural olmuş olur. Oner algoritmalarında her bir tahmin bir kural oluşturur, sınıflandırmaların hangi sıklıklarla göründüğünü hesaplar ve daha sonra en sık olan sınıflandırmayı seçer. En sonunda da en sıklıkla olan sınıflandırmayı kural olarak seçer. Özetle, her bir tahminin sınıflandırmaya olan etkilerinden toplam hata ölçülür, toplam hata ne kadar küçük çıkarsa, tahmin okadar kuvvetlidir (Witten, Frank, 2005).

4.2.7 Zeror Kuralı

Zeror yaygın olarak görünen sınıfların tahmininde bulunan bir yapıya sahiptir, basit yapılarda pek uygun değildir fakat daha kompleks yapılarda, temel performans değerlendirmeleri için kullanışlıdır (Witten, Frank, 2005).

4.2.8 MATLAB Adaptif Ağ Tabanlı Bulanık Mantık (ANFIS)

MATLAB, mühendislik alanında matematiksel ve teknik çalışmaların analizleri için kullanılan ve matris yapısı ile çalışan bir araçtır. Dalgalar, görüntü ve ses işleme, analog ve sayısal işlemeler, yapay sinir ağları ve bulanık mantık gibi alanlarda sıklıkla kullanılmaktadır. Bu çalışmada MATLAB 7.8.0 (R2000a) (Moler, 2008) Bulanık Mantık aracı kullanılmıştır.

Doğru ve düzgün bir biçimde düşünmenin bilgisi olarak tanımlanabilen mantık, akıl yürütme olarakta ifade edilebilmektedir. Akıl yürütme kavramı ise; var olan bilgi veya bilgileri kullanarak yani bilgilere ulaşmaktır. Akıl yürütmede kullanılan bilgilerin mantığın konusuna girilebilmesi için de bilgiler dilsel olarak ifade edilmeli

ve ifade edilen cümleler de bir yargı şeklinde olmalıdır (Öztürk, Mercan, Toprak, 2003).

İnsanın elde ettiği bilginin türü ne olursa olsun, insan var olan bilgilerden yeni bilgiler elde etmede mantığını kullanmaktadır. Bilim adamı da yaptığı çalışmalarda ve incelemelerde teoriler kurmakta ve bu teorileride deneyler vasıtasıyla gerçekleştirmeye çalışmaktadır. Bilim akılla gerçek arasında mutlaka birer uygunluk olduğunu kabul etmekte, aklın nasıl çalıştığı ise Aristo tarafından kesin olarak prensiplerle ortaya konulmaktadır. Aristo mantığında kesin bilgi anlayışı sorgulamasına karşın iki bin yıldır bilimin üzerine inşaa edildiği bu mantık sistemine karşı alternatif mantık sistemleride geliştirilmiştir (Clear, Yuan, 1995).

Yapay zeka araçlarından biri olan bulanık mantık kavramı, her gün kullandığımız ve davranışlarımızı yorumladığımız bir sisteme ulaşmamızı sağlayan matematiksel bir disiplin olarak karşımıza çıkmaktadır. Bulanık mantık değişik biçimlerde ortaya çıkan karmaşık ve belirsizlik gibi tam ve kesin olmayan bilgi kaynakları olarak düşünülebilir. Bulanık mantıkta herşey 0-1 aralığında belirli bir derece ile gösterilmektedir ve matematiksel modeli karmaşık ve zor olan sistemler için çok kullanışlıdır (Clear, Yuan, 1995).

Bulanık küme teorisi ilk bilgilerin Zadeh tarafından literatüre mal edilmesine karşın aslında bulanık mantığı oluşturacak ilk temel düşünceyi Plato oluşturmuş daha sonra 1900’ lerde Polonyalı mantıkçı Jan Lukasiewicz ilk kez Aristo' nun iki değerli mantığına sistematik bir alternatif geliştirerek çok değerli yada bulanık mantık küme sistemlerini geliştiren Black ise bulanık küme üyelik fonksiyonlarından bahseden ilk kişi olmuştur (Clear, Yuan, 1995).

Matematik ve mantık kavramlarının esaslarını teşkil eden kümeler insan düşüncesinin en temel öğelerini meydana getirmektedir. Düşünce sisteminde mantığın ve matematiğin kullanılmasıyla küme kavramı gündeme gelmektedir. Küme kavramı klasik kümeler ve bulanık kümeler olarak karşımıza çıkmaktadır. Bulanık mantık insan düşüncesinin getirdiği sözel bilgileri işleyebilmekte ve bulanık küme teorisi ile açıklayabilmektedir (Timothy, 1995)

Klasik küme kavramında bir X kümesindeki A alt kümesi kendisine ait karakteristik fonksiyon olan Xa ile ifade edilmektedir. Buradaki karakteristik fonksiyon X in elemanlarını {0,1} kümesine dönüştürmektedir. Klasik bir A kümesini karakteristik

ifadesi yardımıyla aşağıdaki şekilde ifade etmek mümkündür. Aşağıdaki formülasyonda da göründüğü gibi A kümesine ait elemanlar 1 değerini alırken ait olmayanlar 0 değerini alıyor.

(4.4)

X boş olmayan bir küme olmak üzere X deki bir A kümesi

olmak üzere;

: Anfis kümesine karşılık gelen fonksiyondur. A kümesi elemanlarından beklenen niteliklerin ne kadar sağladığı bilgisi olarak düşünülebilir.

Anfis kümeleri kesikli ve sürekli bulanık kümeler olmak üzere iki grupta incelenmektedir.

olmak üzere,

Anfis Kesikli A kümesi, **(4.5)**

Anfis Sürekli A kümesi olarak ifade edilebililir.

4.2.9 İstatistik Doğruluk Ölçümleri (Statistical Accurancy Metrics)

İstatistik doğruluk ölçütleri, deneysel sonuçlar ölçmek için kullanılır. Yaygın olarak kullanılan ortak ölçümleri; mutlak hata (mean absolute error), ortalama kare hata (mean square error) ve ortalama hata kareleri kökü toplamıdır (root mean squared error). Bu tez çalışmasında, yöntemleri kıyaslamak için tercih edilen ölçüm ortalama hata kareleri kökü toplamıdır (McClish, 1987).

4.2.10 Ortalama Hata Kareleri Kökü Toplamı (Root Mean Squaed Error)

Ortalama hata karelerinin kökü toplamı, yapay sinir ağlarında ortamın performansını ölçmek için kullanılan bir indistir. Aynı zamanda RMSE olarak adlandırılan küçük hatalardan daha büyük hataları tartmak için kullanılır. RMSE aşağıdaki formül ile hesaplanır. Formüldeki tahmin edilen değerler, bilinen değerleri N ise toplam değerleri gösterir (McClish, 1987).

(4.6)

4.2.11 Tanı Testi Performansı (Receiver Operating Characteristic)

Tanı Testi performanslarının değerlendirilmesi ve kıyaslanması için en yaygın kullanıma sahip olan yöntemdir.

Bu Çalışmada ROC alan grafiğinin kullanılma sebebi, uygulanan sınıflandırma yöntemlerinin performansını ve başarısını ROC grafiğine göre değerlendirme ve uygulanan yöntemlerin birbirleri ile kıyaslanmasının yapılmasıdır.

ROC grafiğinin x koordinatında yanlış positif değeri, y koordinatında ise gerçek positif değeri bulunmaktadır. ROC eğrisinin altında kalan alan 1‘ e ne kadar yakın ise performans o kadar yüksek çıkmaktadır. Bu durumda FPR değerinin çok düşük, TPR değerinin ise 1 ‘ e çok yakın bir değer olması beklenmektedir. FPR ve TPR değerlerinin dışında grafiğin anlamlı çıkması için başlangıç noktası olarak (0,0) değeri, bitiş değeri olarakta (1,1) değerleri verilmelidir (McClish, 1987).

Tablo 4.4 Karışık Matris Yapısı

		Gerçek Değerler		
		P	N	
Tahmin Edilen Değerler	p’	Gerçek Pozitif	Yanlış Pozitif	P’
	n’	Yanlış Negatif	Gerçek Negatif	N’
	Toplam	P	N	

Duyarlılık: Gerçek fraud aboneler içinden fraud aboneleri ayırma yeteneğidir.

(4.7)

Gerçek fraud abonelere konan tanılar açısıdan; Gerçek tanı sonucuna uygun olarak testinde fraud değil dediği gerçek negatif olgulardır. Gerçekte fraud olmadıkları halde testin hatalı olarak fraud dediği yanlış pozitif olgulardır.

Özgüllük: Gerçek fraud olmayanlar içinden fraud olmayanları ayırma yeteneğidir.

(4.8)

Yanlış Negatif Oranı: Gerçek fraud aboneler içinden testin hatalı olarak fraud değil dediği durumlardır.

(4.10)

Yanlış Positif Oranı: Gerçek fraud olmayan aboneler içinden testin hatalı olarak fraud dediği durumlardır.

(4.11)

Testin, fraud dediği zaman doğruyu bildirmesinin yanılmasına oranı Fraud Tanısı Koymanın Doğruluk Oranı' dır. Bu değer ne kadar yüksek olursa, gerçek fraudlar o derecede iyi çıkarılabilmektedir.

(4.12)

Negatif test sonucu olasılık oranı ise; Fraud olmama tanısının doğruluk oranıdır. Bu değer ne kadar küçük olursa, gerçek fraudlar o kadar iyi çıkarılabilmektedir.

(4.13)

Doğruluk: Gerçekte testin fraud olan ve fraud olmayan olarak toplam doğru tanı oranına denir.

(4.14)

4.2.12 Kappa İstatistik Katsayısı

Kappa istatistik katsayısı, aynı nesneyi derecelendiren iki derleyici arasındaki uyumu test etmek amacıyla kullanılır. Tanı testi performansı ölçmek için kullanılan bir katsayıdır ve beklenen uyum ile tahmini uyum arasındaki bağıntı olarak düşünülebilir. Aşağıdaki formülde; Pr(a) beklenen uyum, Pr(e) ise tahmini uyumdur. Kappa katsayısı 1 'e ne kadar yakın ise performans okadar iyidir (Bluman, 2004).

(4.15)

4.2.13 Ortalama Mutlak Hata (MAE)

Ortalama mutlak hata zaman serisinde, istatistik ölçümlerdeki hataları oranlarını gösteren, yüzdesel bir değerdir. Gerçek değer ile tahmin edilen değer arasındaki farkın, mutlak değerinin toplamının, yine tahmin sayısına bölünmesidir. Aşağıdaki formüldeki; At, gerçek değerler, Ft, tahmin edilen değerler ve n ise tahmin sayısıdır. M değerinin 0' a yakın çıkması, testin performansının başarılı çıkması yönünde önemlidir.

(4.16)

5.BULGULAR

Abone sahtekarlığı kestirimi ve tahminlemesi için, veri madenciliği sınıflandırma yöntemlerinden sırası ile aşağıdaki modeller uygulanmıştır. .

5.1 WEKA BAYES AĞLARI UYGULAMASI

Weka 3.7.1 sınıflandırma methodu Bayes Ağları uygulamasından aşağıdaki çıktılar elde edilmiştir. Tablo 5.1'de de görüldüğü gibi 2448 müşteri bilgilerinden 2211 (yüzde 90.3186) kayıt, doğru sınıflandırılmış örneklem, 237 (yüzde 9.6814) kayıt ise yanlış sınıflandırılmış örneklemdir. RMSE değerinin 0' a yakın çıkması uygulamanın başarılı olduğunun bir göstergesidir.

Tablo 5.1: BayesNet İstatistik Değerleri

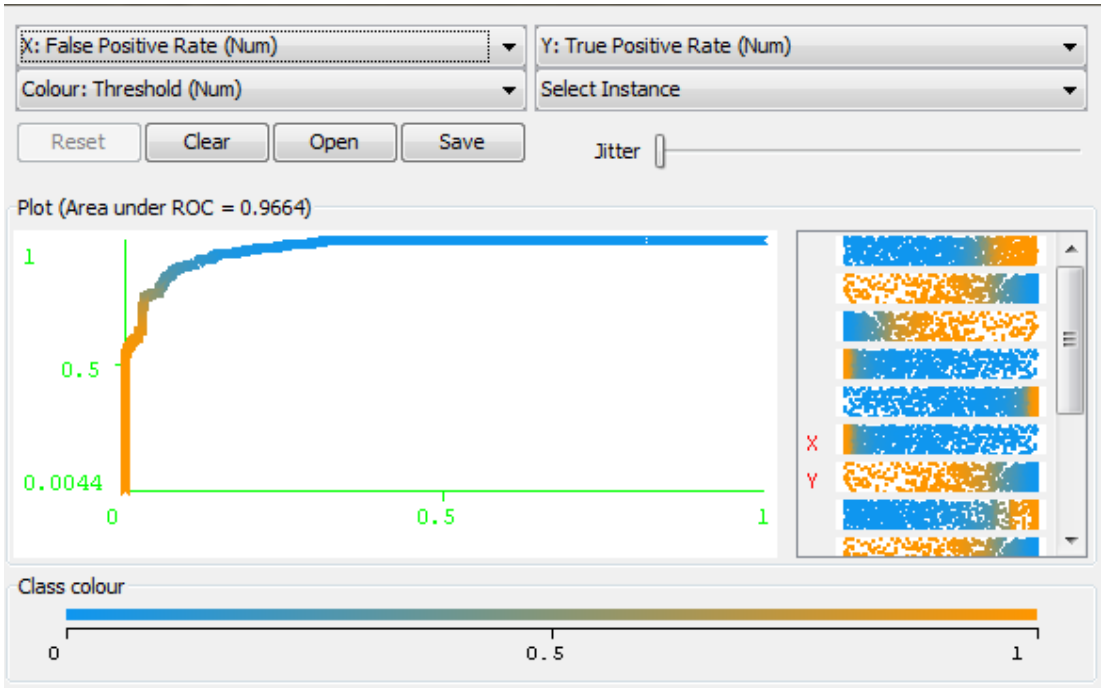
Correctly Classified Instances	2211	90.3186 %
Incorrectly Classified Instances	237	9.6814 %
Kappa statistic	0.7535	
Mean absolute error	0.1058	
Root mean squared error	0.265	
Relative absolute error	26.2475 %	
Root relative squared error	59.0337 %	
Coverage of cases (0.95 level)	97.3448 %	
Mean rel. region size (0.95 level)	60.5801 %	
Total Number of Instances	2448	

Weka 3.7.1 sınıflandırma methodu Bayes Ağları uygulamasından aşağıdaki çıktılar elde edilmiştir. Tablo 5.1'de de görüldüğü gibi 2448 müşteri bilgilerinden 2211 (yüzde 90.3186) kayıt, doğru sınıflandırılmış örneklem, 237 (yüzde 9.6814) kayıt ise

yanlış sınıflandırılmış örneklemdir. RMSE değerinin 0’ a yakın çıkması uygulamanın başarılı olduğunun bir göstergesidir.

Tablo 5.2: Bayes Ağları Doğruluk Değerleri

TP Rate		FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.785		0.051	0.857	0.785	0.819	0.966	EVET
0.949		0.215	0.919	0.949	0.934	0.966	HAYIR
Avg.	0.903	0.169	0.902	0.903	0.902	0.966	-



Şekil 5.1: Bayes Ağları ROC Eğrisi

5.2 WEKA NAIVE BAYES UYGULAMASI

Tablo 5.3 ’de Naive Bayes uygulamasından elde elilen değerler bulunmaktadır. Naive Bayes uygulasında yüzde 85 oranında bir başarı elde edilmiştir. Kappa istatistik değerinden ve RMSE değerlerinden de anlaşılacağı gibi, bu uygulamadan çok iyi sonuçlar elde edilememiştir .Naive Bayes uygulamasındaki tutarlılığın BayesNet ‘e göre daha düşük olduğu gözlemlenmiştir.

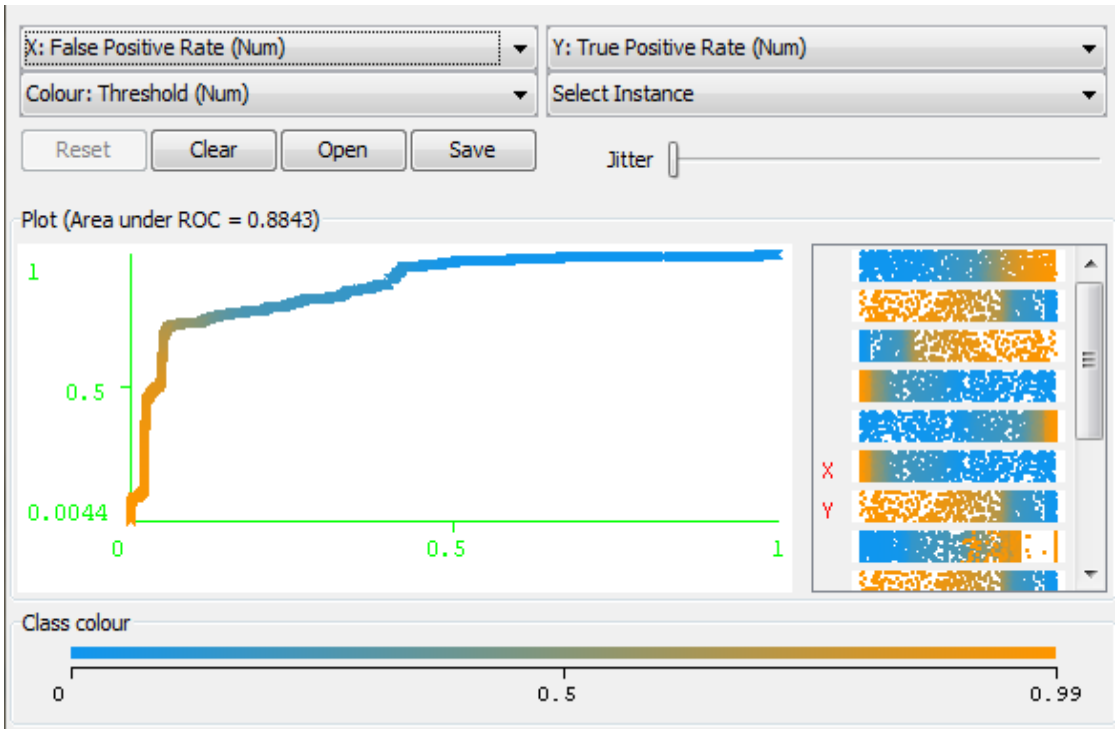
Tablo 5.3: Naive Bayes İstatistik Değerleri

Correctly Classified Instances	2098	85.7026 %
Incorrectly Classified Instances	350	14.2974 %
Kappa statistic	0.6446	
Mean absolute error	0.1946	
Root mean squared error	0.3351	
Relative absolute error	48.2686 %	
Root relative squared error	74.6562 %	
Coverage of cases (0.95 level)	97.1405 %	
Mean rel. region size (0.95 level)	75.674 %	
Total Number of Instances	2448	

Tablo 5.4’ de ROC değerleri yer almaktadır. Şekil 5.2 de görüldüğü gibi eğrinin altında kalan alan 0.884 ‘ dür. FP değerinin BayesNet ‘ e göre arttığı ve TP değerininde azalışından dolayı, hata oranı BayesNet ’ e göre daha yüksektir.

Tablo 5.4: Naive Bayes Doğruluk Değerleri

TP Rate		FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.742		0.098	0.746	0.742	0.744	0.884	EVET
0.902		0.258	0.9	0.902	0.901	0.884	HAYIR
Avg.	0.857	0.214	0.857	0.857	0.857	0.884	-



Şekil 5.2: Naive Bayes ROC Eğrisi

5.3 WEKA LOGISTIC UYGULAMASI

Logistik Uygulamasından elde edilen sonuçlar, Naive Bayes ve BayesNet uygulamalarına göre daha tutarlıdır. RMSE değerinin 0' a yakın ve doğru sınıflandırılan örneklemin yüzde 94 çıkması, uygulamanın diğer iki yötemden daha başarılı olduğunun bir göstergesidir.

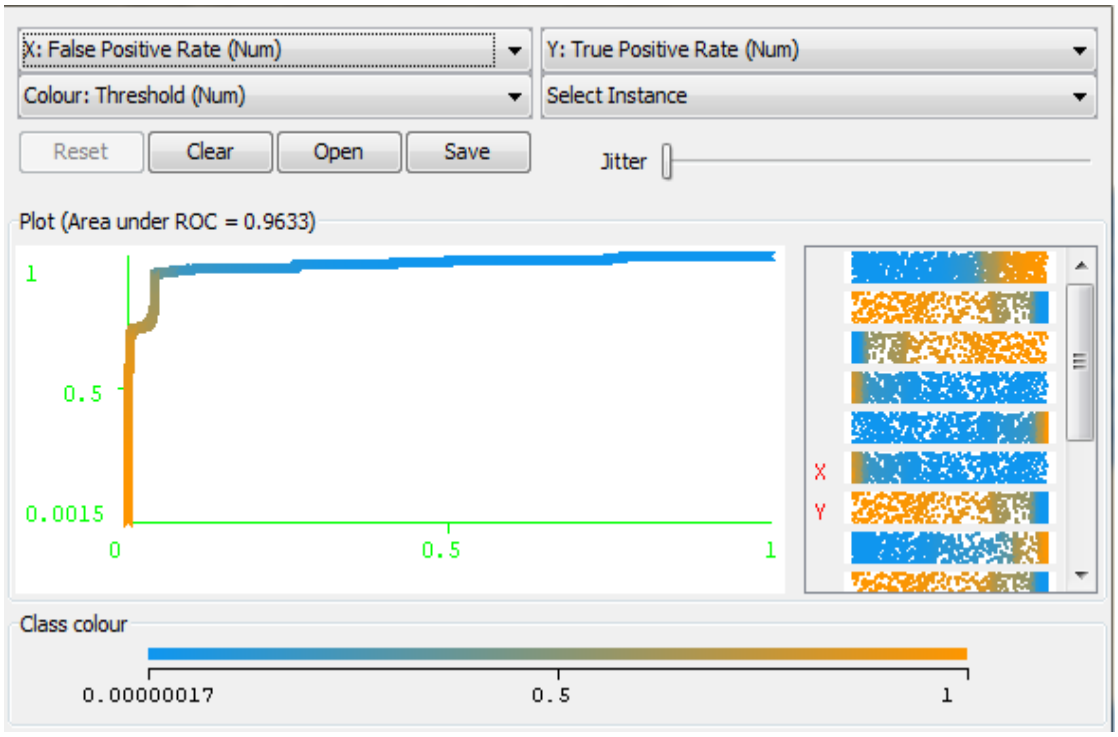
Tablo 5.5: Logistic İstatistik Değerleri

Correctly Classified Instances	2320	94.7712 %
Incorrectly Classified Instances	128	5.2288 %
Kappa statistic	0.8718	
Mean absolute error	0.105	
Root mean squared error	0.2205	
Relative absolute error	26.038 %	
Root relative squared error	49.1119 %	
Coverage of cases (0.95 level)	98.8562 %	
Mean rel. region size (0.95 level)	68.3415 %	
Total Number of Instances	2448	

Tablo 5.6 da ROC değerleri yer almaktadır. FP değeri 0,044 ve TP değeride 0,926 çıkmıştır. ROC eğrisindeki hata oranı, FP değerindeki düşüş ve TP değerinde artış ile birlikte azalmıştır. Şekil 5.3 de görüldüğü gibi eğrinin altında kalan başarılı alan 0.963 ‘ dür.

Tablo 5.6: Logistic Doğruluk Değerleri

TP Rate		FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.926		0.044	0.892	0.926	0.908	0.963	EVET
0.956		0.074	0.971	0.956	0.963	0.963	HAYIR
Avg.	0.948	0.066	0.949	0.948	0.948	0.963	-



Şekil 5.3: Logistic ROC Eğrisi

5.4 WEKA ÇOK KATMANLI ALGILAYICI UYGULAMASI

Tablo 5.7'de de görüldüğü gibi 2448 müşteri bilgilerinden 2431 (yüzde 99.3056) kayıt doğru sınıflandırılmış örneklem,17 (yüzde 0.6944) kayıt ise yanlış sınıflandırılmış örneklemdir. Mean Absolute Error ve Root Mean Square Error ün 0' a çok yakın olması hatanın neredeyse hiç olmadığı anlamına gelmektedir. BayesNet uygulamasından daha başarılı sonuç elde edilmiştir.

Tablo 5.7: Multiplayer Perceptron İstatistik Değerleri

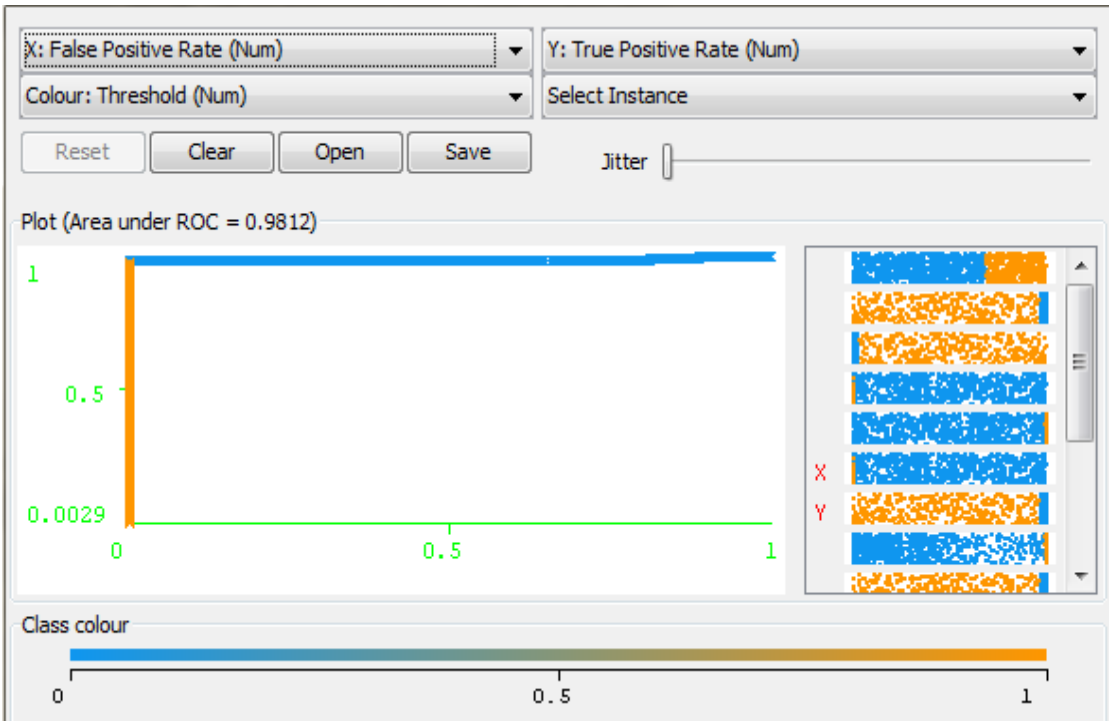
Correctly Classified Instances	2431	99.3056 %
Incorrectly Classified Instances	17	0.6944 %
Kappa statistic	0.9827	
Mean absolute error	0.0081	
Root mean squared error	0.0819	
Relative absolute error	2.0029 %	
Root relative squared error	18.2371 %	
Coverage of cases (0.95 level)	99.3873 %	
Mean rel. region size (0.95 level)	50.0817 %	
Total Number of Instances	2448	

Tablo 5.8 de ROC değerleri yer almaktadır. Şekil 5.4 de görüldüğü, FP ve TP değerlerinin verdiği ROC eğrisinin altında kalan alan 0.981 ' dür. Sistem yüzde 99 başarılı çıkmıştır.

Tablo 5.8: Multiplayer Perceptron Doğruluk Değerleri

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.977	0.001	0.999	0.977	0.987	0.981	EVET

0.999		0.023	0.991	0.999	0.995	0.981	HAYIR
Avg.	0.993	0.017	0.993	0.993	0.993	0.981	-



Şekil 5.4: Multiplayer Perceptron ROC Eğrisi

5.5 WEKA KISMI KARAR AGACLARI UYGULAMASI

Weka 3.7.1 sınıflandırma methodu Desicion Table uygulamasından aşağıdaki çıktılar elde edilmiştir. Tablo 5.9'da de görüldüğü gibi 2448 müşteri bilgilerinden 2438 (yüzde 99.5915) kayıt doğru sınıflandırılmış örneklem, 10 (yüzde 0.4085) kayıt ise yanlış sınıflandırılmış örneklemdir. Kuadratik Ortalamanın (Root Mean Squared Error) 0 'a yakın çıkması uygulamanın başarılı olduğunun bir göstergesidir. BayesNet ve Multiplayer uygulamalarına göre Desicion table daha başarılı olmuştur.

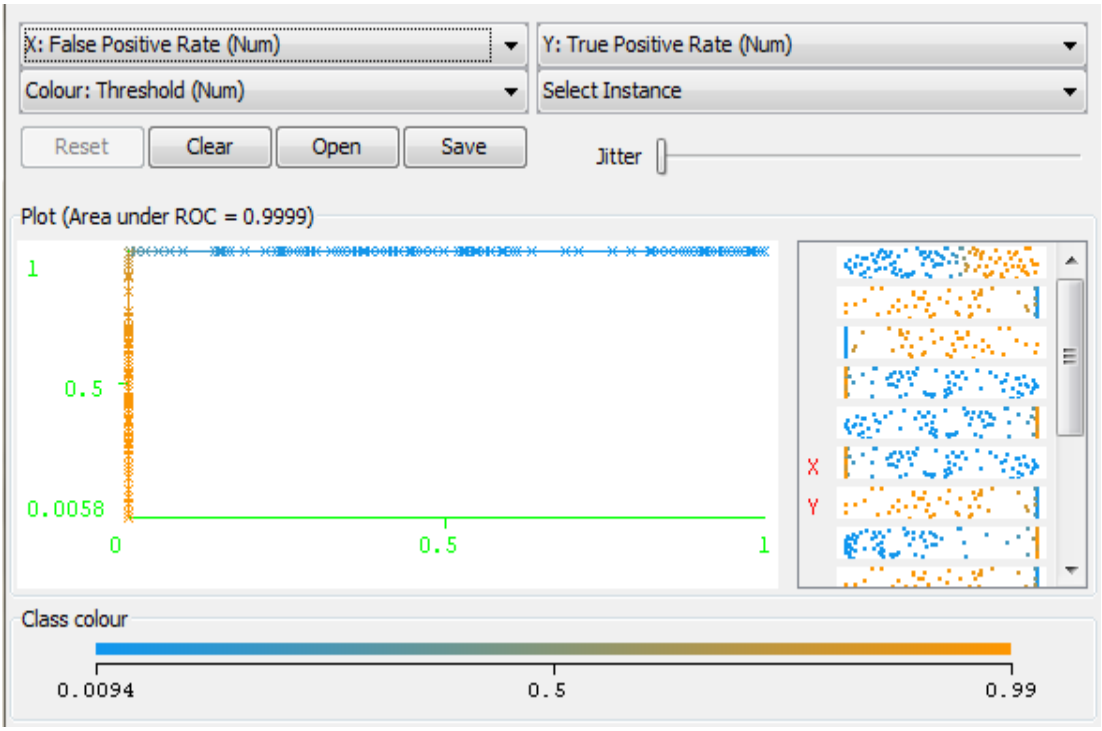
Tablo 5.9: Karar Tablosu İstatistik Değerleri

Correctly Classified Instances	2438	99.5915 %
Incorrectly Classified Instances	10	0.4085 %
Kappa statistic	0.9899	
Mean absolute error	0.0547	
Root mean squared error	0.0911	
Relative absolute error	13.5655 %	
Root relative squared error	20.2904 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	63.6029 %	
Total Number of Instances	2448	

Tablo 5.10 da ROC değerleri yer almaktadır. Şekil 5.5 de görüldüğü gibi eğrinin altında kalan alan 1 ' dir. Müşteri bilgilerinden yanlızca 10 kaydın tutarsız çıkmasından dolayı ROC değeri 1 çıkmıştır.

Tablo 5.10: Karar Tablosu Doğruluk Değerleri

TP Rate		FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.997		0.005	0.988	0.997	0.993	1	EVET
0.995		0.003	0.999	0.995	0.997	1	HAYIR
Avg.	0.996	0.003	0.996	0.996	0.996	1	-



Şekil 5.5: Karar Tablosu ROC Eğrisi

5.6 WEKA JRIP UYGULAMASI

Tablo 5.11'deki JRIP yönteminin sonuçlarına göre; 2448 müşteri bilgilerinden 2439 (yüzde 99.6324) kayıt doğru sınıflandırılmış örneklem, 9 (yüzde 0.3676) kayıt ise yanlış sınıflandırılmış örneklemdir. Kuadratik Ortalamanın (Root Mean Squared Error) 0' a yakın çıkması uygulamanın başarılı olduğunun bir göstergesidir. Diğer 5 sınıflandırma yöntemine göre datalar daha tutarlıdır .

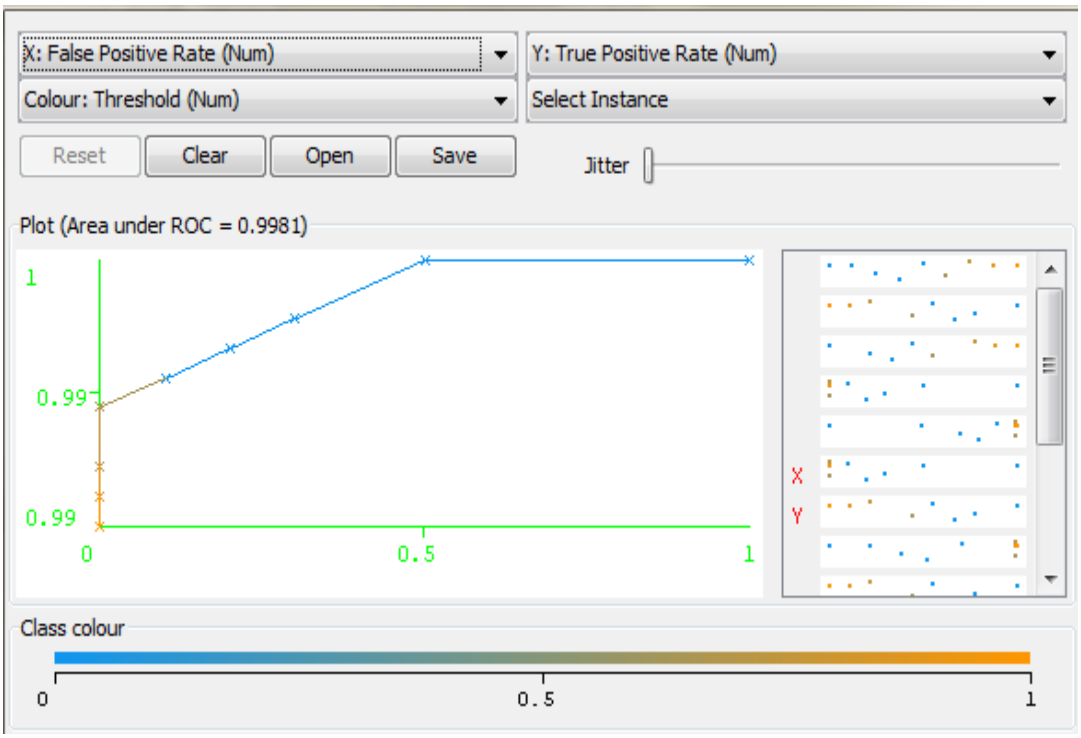
Tablo 5.11: Jrip İstatistik Değerleri

Correctly Classified Instances	2439	99.6324 %
Incorrectly Classified Instances	9	0.3676 %
Kappa statistic	0.9909	
Mean absolute error	0.004	
Root mean squared error	0.0565	
Relative absolute error	1.0022 %	
Root relative squared error	12.5963 %	
Coverage of cases (0.95 level)	99.7958 %	
Mean rel. region size (0.95 level)	50.1634 %	
Total Number of Instances	2448	

Tablo 5.12 de ROC değerleri yer almaktadır. Şekil 5.6' da da görüldüğü gibi eğrinin altında kalan alan 0.998 ' dir.

Tablo 5.12: Jrip Doğruluk Değerleri

TP Rate		FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.993		0.002	0.994	0.993	0.993	0.998	EVET
0.998		0.007	0.997	0.998	0.997	0.998	HAYIR
Avg.	0.996	0.006	0.996	0.996	0.996	0.998	-



Şekil 5.6: Jrip ROC Eğrisi

5.7 WEKA PART UYGULAMASI

Weka 3.7.1 sınıflandırma methodu Part uygulamasından aşağıdaki çıktılar elde edilmiştir. Tablo 5.13'de de görüldüğü gibi 2448 müşteri bilgilerinden 2441 (yüzde 99.7141) kayıt doğru sınıflandırılmış örneklem, 7 (yüzde 0.2859) kayıt ise yanlış sınıflandırılmış örneklemdir. Kuadratik Ortalamanın (Root Mean Squared Error) oldukça düşük çıkması uygulamanın başarılı olduğunun bir göstergesidir. Diğer modellerde olduğu gibi part modelinden de başarılı çıktılar elde edilmiştir.

Tablo 5.13: Part İstatistik Değerleri

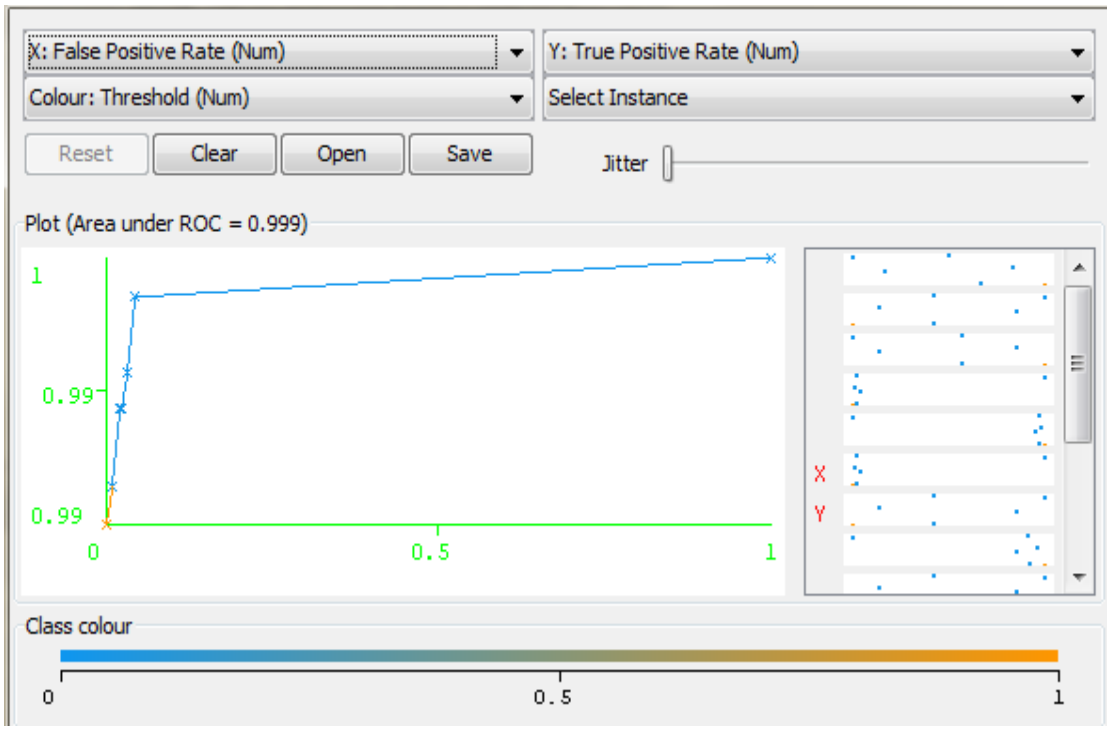
Correctly Classified Instances	2441	99.7141 %
Incorrectly Classified Instances	7	0.2859 %
Kappa statistic	0.9929	
Mean absolute error	0.0035	
Root mean squared error	0.0527	

Relative absolute error	0.8638 %	
Root relative squared error	11.7433 %	
Coverage of cases (0.95 level)	99.7141 %	
Mean rel. region size (0.95 level)	50 %	
Total Number of Instances	2448	

Tablo 5.14 de ROC değerleri yer almaktadır. Şekil 5.7 de görüldüğü gibi eğrinin altında kalan alan 0.999 ‘ dur. Part modelindeki çıkan değerler bu modelinde uygulanabilirliliğin göstergesidir. Part uygulamasındaki başarı yüzde 99 oranındadır. FP değerinin 0, TP değerininde 1 değerlerini almasıda bu başarının göstergesidir.

Tablo 5.14: Part Doğruluk Değerleri

TP Rate		FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.99		0	1	0.99	0.995	0.999	EVET
1		0.01	0.996	1	0.998	0.999	HAYIR
Avg.	0.997	0.007	0.997	0.997	0.997	0.999	-



Şekil 5.7: Part ROC Eğrisi

5.8 WEKA ONER UYGULAMASI

Weka 3.7.1 sınıflandırma methodu Oner uygulamasından aşağıdaki çıktılar elde edilmiştir. Tablo 5.15'de de görüldüğü gibi 2448 müşteri bilgilerinden 2097 (yüzde 85.6618) kayıt doğru sınıflandırılmış örneklem, 351 (yüzde 14.3382) kayıt ise yanlış sınıflandırılmış örneklemdir.

.

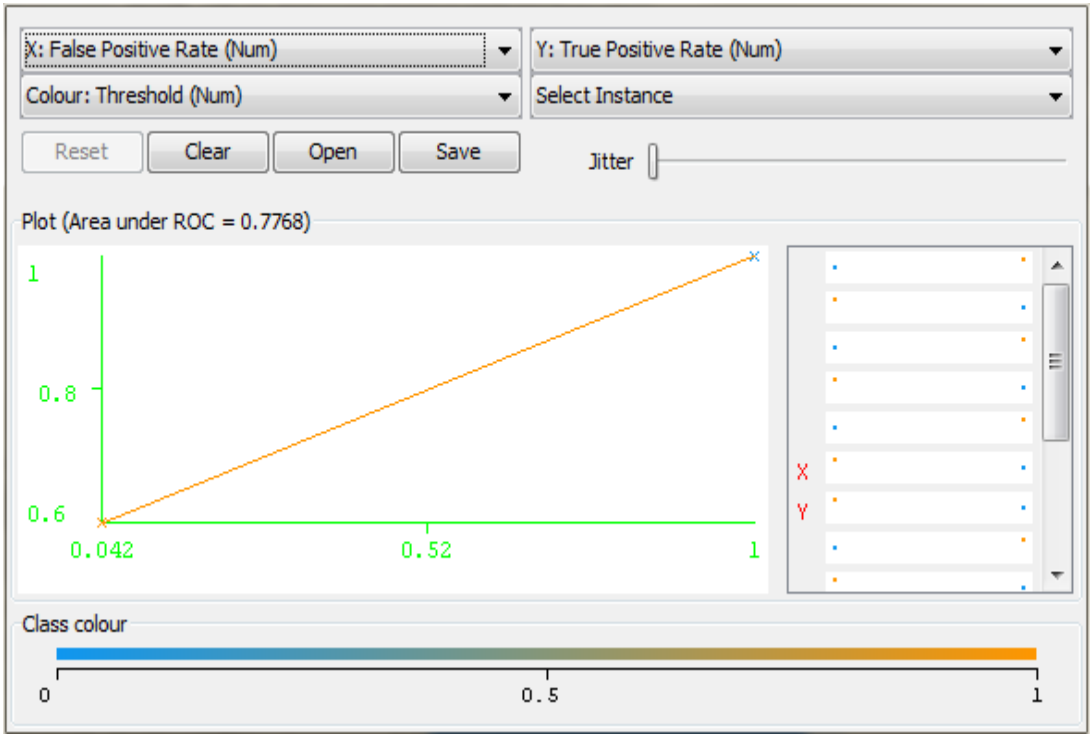
Tablo 5.16: Oner İstatistik Değerleri

Correctly Classified Instances	2097	85.6618 %
Incorrectly Classified Instances	351	14.3382 %
Kappa statistic	0.6088	
Mean absolute error	0.1434	
Root mean squared error	0.3787	
Relative absolute error	35.5672 %	
Root relative squared error	84.3504 %	
Coverage of cases (0.95 level)	85.6618 %	
Mean rel. region size (0.95 level)	50 %	
Total Number of Instances	2448	

Tablo 5.17' de ROC değerleri yer almaktadır. Şekil 5.8 de görüldüğü gibi eğrinin altında kalan alan 0.777' dir. Tabloda çıkan FP ve TP değerlerine baktığımızda, Oner uygumasının hata oranının, diğer yöntemlere kıyasla, daha yüksek çıktığı söylenilebilir.

Tablo 5.17: Oner Doğruluk Değerleri

TP Rate		FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.596		0.042	0.846	0.596	0.699	0.777	EVET
0.958		0.404	0.859	0.958	0.906	0.777	HAYIR
Avg.	0.857	0.303	0.856	0.857	0.848	0.777	-



Şekil 5.8: Oner ROC Eğrisi

5.9 WEKA ZEROR UYGULAMASI

Weka 3.7.1 sınıflandırma methodu Zeror uygulamasından aşağıdaki çıktılar elde edilmiştir. Tablo 5.17'de de görüldüğü gibi 2448 müşteri bilgilerinden 1763 (yüzde 72.018) kayıt doğru sınıflandırılmış örneklem, 685 (yüzde 27.982) kayıt ise yanlış sınıflandırılmış örneklemdir. Kuadratik Ortalamanın 0 a çok yakın çıkması ve Kappa Statistic değerinin 0 çıkması Zeror modelinin bu çalışma için uygun olmadığının göstergesidir.

Tablo 5.17: Zeror İstatistik Değerleri

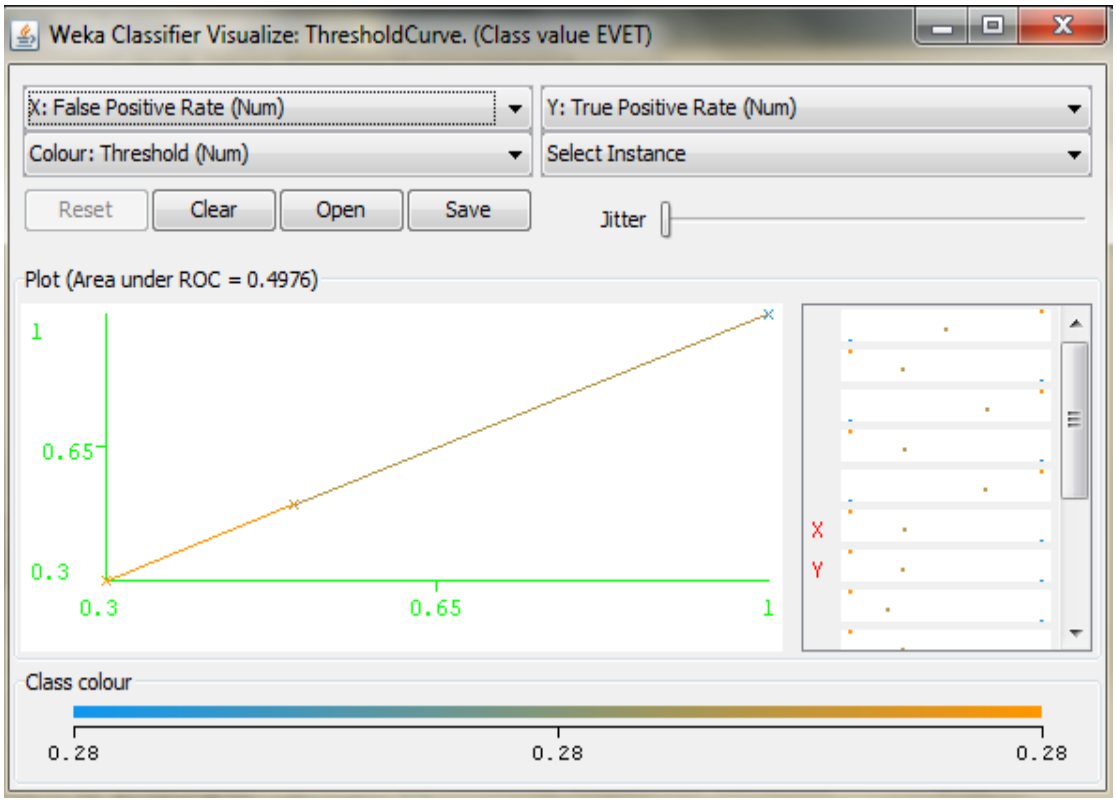
Correctly Classified Instances	1763	72.018 %
Incorrectly Classified Instances	685	27.982 %
Kappa statistic	0	
Mean absolute error	0.4031	
Root mean squared error	0.4489	

Relative absolute error	100 %	
Root relative squared error	100 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	100 %	
Total Number of Instances	2448	

Tablo 5.18 de ROC değerleri yer almaktadır. Şekil 5.9 de görüldüğü gibi eğrinin altında kalan alan 0.498 ‘ dür. Zeror uygulamasından, FP (0,1) ve TP (0,1) değerlerine bakıldığında, diğer modellerdeki gibi tutarlı çıktılar elde edilememiştir.

Tablo 5.18: Zeror Doğruluk Değerleri

TP Rate		FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0		0	0	0	0	0.498	EVET
1		1	0.72	1	0.837	0.498	HAYIR
Avg.	0.72	0.72	0.519	0.72	0.603	0.498	-



Şekil 5.9: Zeror ROC Eğrisi

5.10. WEKA RBF AĞLARI UYGULAMASI

Weka 3.7.1 sınıflandırma methodu Rbf Network uygulamasının diğer uygulamalardan farkı, bu modeldeki datalar tablo 4.2' deki gibi sayısal dataya çevrilmiştir. Bu uygulamadan aşağıdaki çıktılar elde edilmiştir. Tablo 5.19'da da görüldüğü gibi 2448 müşteri bilgilerinden 2201 (yüzde 89.9101) kayıt doğru sınıflandırılmış örneklem, 247 (yüzde 10.0899) kayıt ise yanlış sınıflandırılmış örneklemdir. Kuadratik Ortalamanın pozitif çıkması uygulamanın başarılı olduğunun bir göstergesidir.

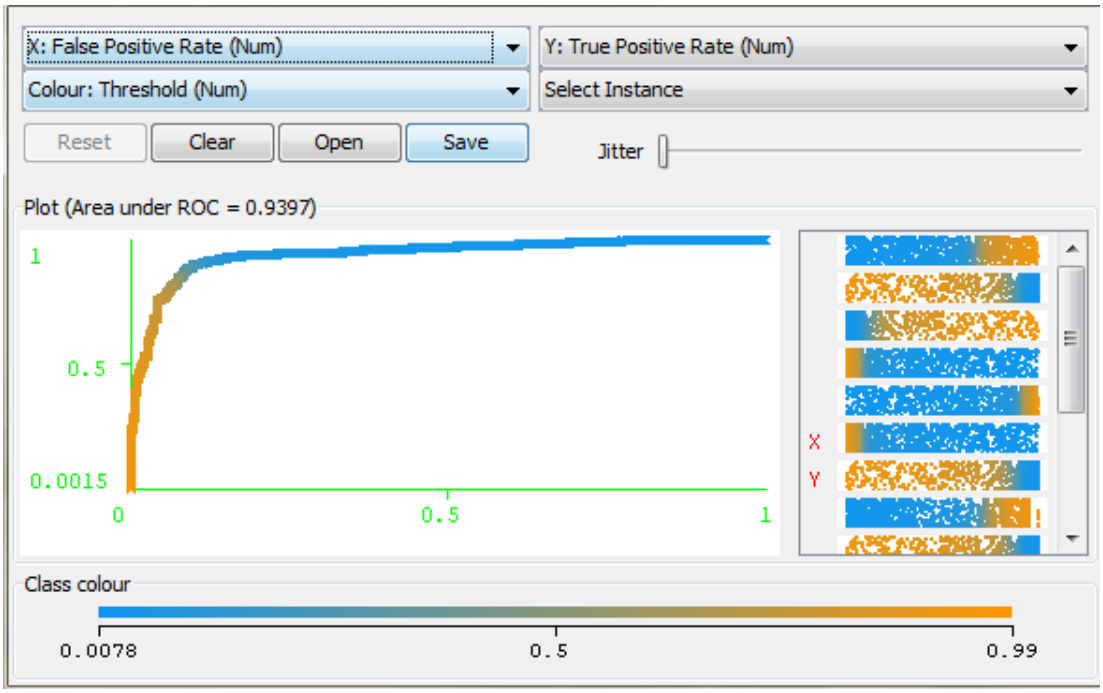
Tablo 5.19: Rbf Ağları İstatistik Değerleri

Correctly Classified Instances	2201	89.9101 %
Incorrectly Classified Instances	247	10.0899 %
Kappa statistic	0.7533	
Mean absolute error	0.1461	
Root mean squared error	0.2761	
Relative absolute error	36.2447 %	
Root relative squared error	61.5103 %	
Coverage of cases (0.95 level)	98.4477 %	
Mean rel. region size (0.95 level)	74.0809 %	
Total Number of Instances	2448	

Tablo 5.20 de ROC değerleri yer almaktadır. Şekil 5.10’ da da görüldüğü gibi eğrinin altında kalan alan 0.94 ‘ dür ve sitem performansı ve çıktılar başarılıdır.

Tablo 5.20: Rbf Ağları Doğruluk Değerleri

TP Rate		FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.844		0.079	0.805	0.844	0.824	0.94	EVET
0.921		0.156	0.938	0.921	0.929	0.94	HAYIR
Avg.	0.899	0.135	0.901	0.899	0.9	0.94	-

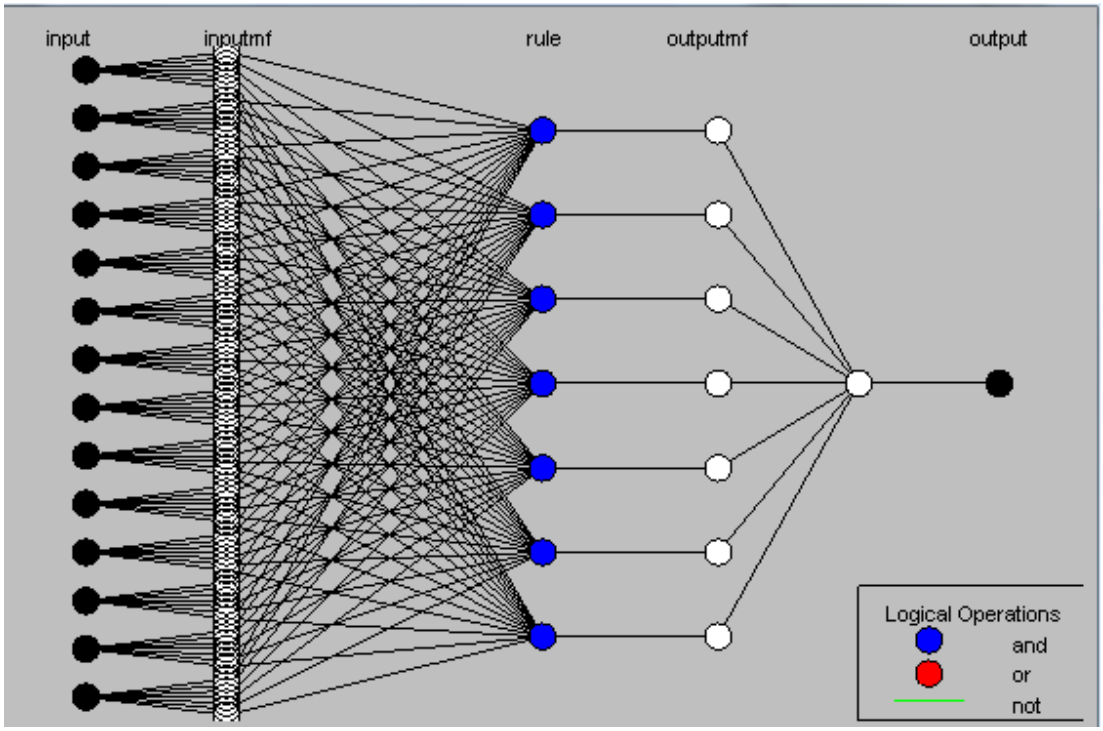


Şekil 5.10: Rbf Ağları ROC Eğrisi

5.11 ADAPTİF AĞ TABANLI BULANIK MANTIK (ANFIS)

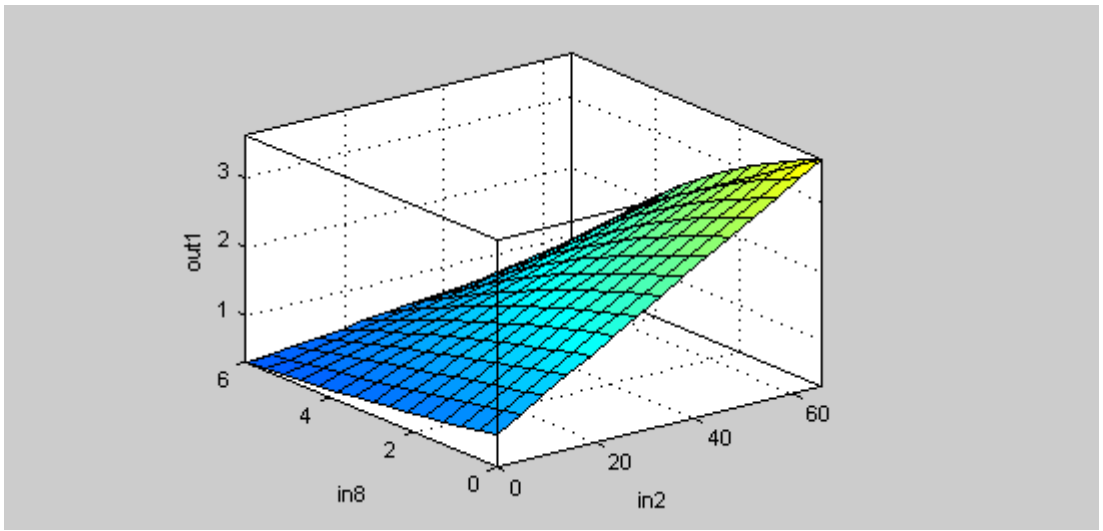
MATLAB, mühendislik alanında matematiksel ve teknik çalışmaların analizleri için kullanılan ve matris yapısı ile çalışan bir araçtır. Dalgalar, görüntü ve ses işleme, analog ve sayısal işlemeler, yapay sinir ağları ve bulanık mantık gibi alanlarda sıklıkla kullanılmaktadır. Bu çalışmada MATLAB 7.8.0 (R2000a) (Moler, 2008) Bulanık Mantık aracı kullanılmıştır.

Adaftif Ağ tabanlı Bulanık Mantık Uygulamasından 14 girişli, tek çıkışlı ve 7 kuraldan oluşan bir ANFIS mimarisi elde edilmiştir. Şekil 5.11 de gösterilmektedir.

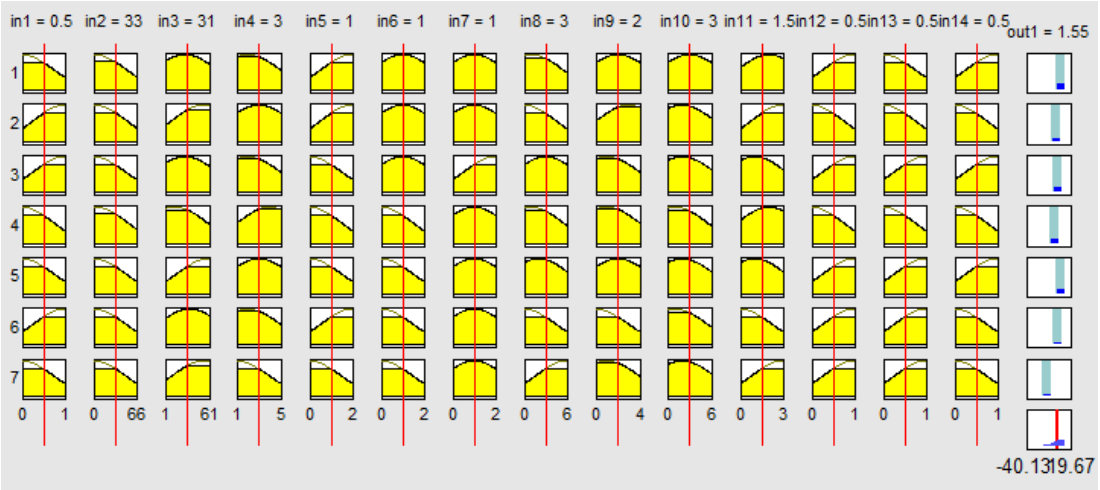


Şekil 5.11 Adaptif Ağ Tabanlı Bulanık Mantık Çıkarım Sistemi

Burada input katmanındaki her bir düğüm, giriş sinyallerinin diğer katmanlara aktarıldığı düğümüdür. İnputmf katmandaki her bir düğüm ise bulanık kümeleri ifade etmektedir. Rule katmanındaki bir düğümler Sugeno bulanık mantık çıkarım sistemine göre oluşturulan kuralları ve sayısını ifade etmektedir. Outputmf katmanı ise arındırma katmanıdır. Rule katmanından bu katmana gelen datalar normalleştirilmiş olarak gelir. Son olarak output katmanında ise her bir düğümün çıkış değerleri toplanarak tek bir düğüm elde edilmiştir.



Şekil 5.12 Abone Yaşının Geç Ödenen Fatura Sayısına Göre Yüzey Çıkarımı



Şekil 5.13 Adaptif Ağ Tabanlı Bulanık Mantık Kural Çıkarımı

Aşağıdaki değerler, Matlab Anfis modeli uygulamasından elde edilen karışık matris değerleridir. Bu değerlerden yola çıkarak doğru pozitif, yanlış positif, duyarlılık ve özgürlük değerleri hesaplanılarak ROC grafiği çizilecektir.

Karışık Matris		Tahmin	
		Negative	Positive
Gerçek	Negative		
	Positive		

tp=245/251=0,976 tn=44/46=0,956

fp=2/46=0,0434 fn=6/251=0.0239

Tahmin =245/247=0.991

Doğru Sınıflandırılmış Örneklem = 289/297 = 0.973 = 97.3%

Yanlış Sınıflandırılmış Örneklem = 1-0.973 = 0.027 = 2.7%

		Gerçek Değerler		
		P	N	
Tahmin Edilen Değerler	p’	0,976	0,0434	108
	n’	0,0239	0,956	313
		0,999	0,994	1,932

Özgürlük= tp / (tp + fn) = 0,976/(0,976+0,0239)= 0.9565=95,65%

Duyarlılık = fp / (fp + tn) = 0,0434/(0,0434+0,956)=0.976=96.76%

Anfis uygumasından elde edilen yukarıdaki değerlerle, 7 farklı kural elde edilmiştir. Bu kurallar aşağıdaki şekilde ifade edilmektedir.

Rule 1: [0 2 34 1 1 1 0 3 4 1 4 0 1 1] [0]

Eğer cinsiyet= erkek ve yaşı= 25-32 arasında ve il= İstanbul ve abonelik yaşı= 6-12 ay ve operatör değişikliği-eski operatörü= vodafone ve kaç operatörde aboneliği var= 2 ve mevcut operatörü= Avea ve geç ödenen fatura sayısı=4 ve ortalama aylık fatura tutarı=75-100 arası ve aylık ortalama kullanılan sms sayısı= 30-45 arası ve son borç durumu= 75-100 arası ve sondan 3. fatura ödemesi=ödedi, sondan 2. Fatura ödemesi= ödemedi ve sondan fatura ödemesi=ödemedi ise fraud mu=evet

Rule 2: [1 1 34 1 0 2 1 4 4 6 5 1 1 1] [0]

Eğer cinsiyet= bayan ve yaşı=18-24 arası ve il= İstanbul ve abonelik yaşı= 6-12 ay arası ve operatör değişikliği-eski operatörü= avea ve kaç operatörde aboneliği var= 3 ve mevcut operatörü=vodafone ve geç ödenen fatura sayısı= 5 ve ortalama aylık fatura tutarı=75-100 arası ve aylık ortalama kullanılan sms sayısı= 125-150 arası ve son borç durumu= 100-125 arası ve sondan 3. fatura ödemesi=ödemedi, sondan 2. Fatura ödemesi=ödemedi ve sondan fatura ödemesi=ödemedi ise fraud mu=evet

Rule 3: [0 1 26 2 1 1 0 2 5 4 4 1 1 0] [1]

Eğer cinsiyet= erkek ve yaşı= 18-24 arası ve il= Eskişehir ve abonelik yaşı= 12-24 ay arası ve operatör değişikliği-eski operatörü= vodafone ve kaç operatörde aboneliği var= 2 ve mevcut operatörü= avea ve geç ödenen fatura sayısı= 3 ve ortalama aylık fatura tutarı= 100-125 arası ve aylık ortalama kullanılan sms sayısı= 75-100 ve son borç durumu= 75-100 ve sondan 3. fatura ödemesi=ödemedi, sondan 2. Fatura ödemesi=ödemesi ve sondan fatura ödemesi= ödedi ise fraud mu=hayır

Rule 4: [1 3 34 3 2 0 2 0 1 1 1 0 0 1] [1]

Eğer cinsiyet= bayan ve yaşı= 33-40 ve il= İstanbul ve abonelik yaşı= 24-36 ay arası ve operatör değişikliği-eski operatörü= turkcell ve kaç operatörde aboneliği var=1 ve

mevcut operatörü= turkcell ve geç ödenen fatura sayısı= 1 ve ortalama aylık fatura tutarı= 30-45 arası ve aylık ortalama kullanılan sms sayısı= 30-45 ve son borç durumu= 30-45 arası ve sondan 3. fatura ödemesi= ödedi, sondan 2. Fatura ödemesi= ödedi ve sondan fatura ödemesi=ödemedi ise fraud mu= hayır

Rule 5: [1 6 34 6 2 1 2 1 3 0 0 0 1 0] [1]

Eğer cinsiyet= bayan ve yaşı= 52-57 ve il= İstanbul ve abonelik yaşı= 60-72 ay arası ve operatör değişikliği-eski operatörü= turkcell ve kaç operatörde aboneliği var= 2 ve mevcut operatörü= turkcell ve geç ödenen fatura sayısı=2 ve ortalama aylık fatura tutarı= 60-75 arası ve aylık ortalama kullanılan sms sayısı= 10-30 arası ve son borç durumu= 10-30 ve sondan 3. fatura ödemesi= ödedi, sondan 2. Fatura ödemesi=ödemedi ve sondan fatura ödemesi=ödedi ise fraud mu=hayır

Rule 6: [0 5 16 4 2 1 2 0 0 0 1 0 0 0] [1]

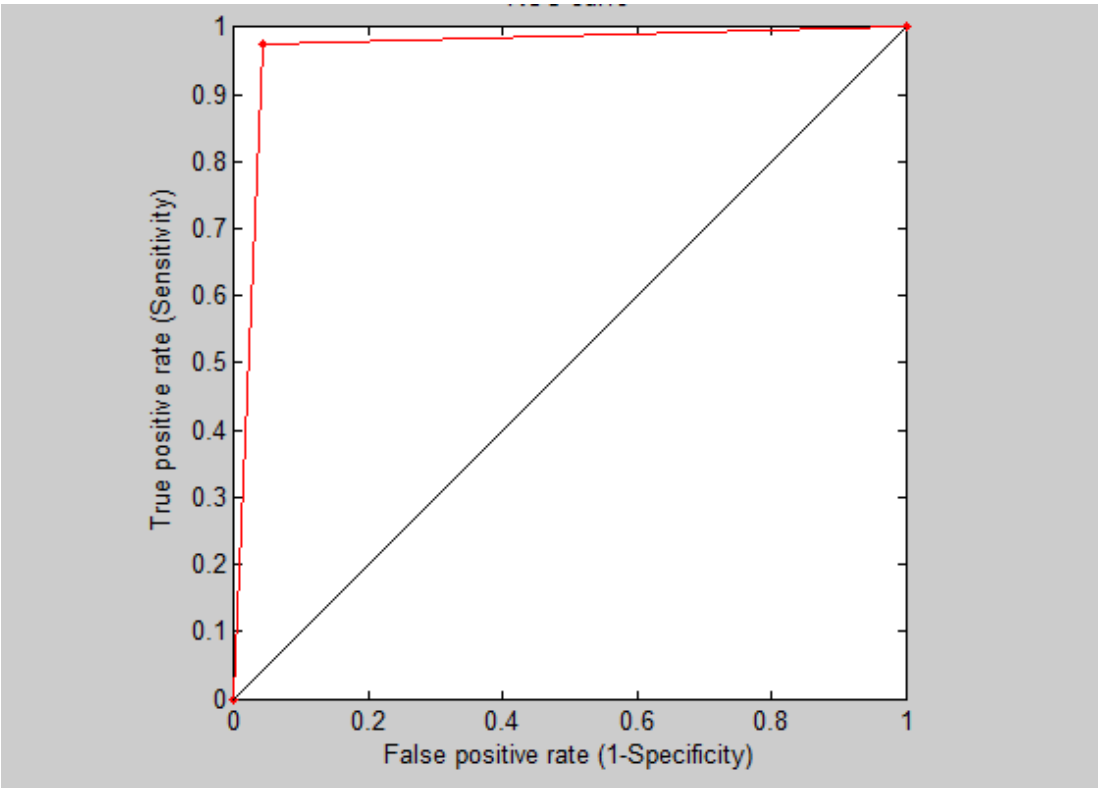
Eğer cinsiyet=erkek ve yaşı= 46-51 arası ve il=Bursa ve abonelik yaşı= 36-48 ay arası ve operatör değişikliği-eski operatörü= turkcell ve kaç operatörde aboneliği var= 2 ve mevcut operatörü= turkcell ve geç ödenen fatura sayısı= 1 ve ortalama aylık fatura tutarı= 10-30 arası ve aylık ortalama kullanılan sms sayısı= 10-30 arası ve son borç durumu= 30-45 ve sondan 3. fatura ödemesi= ödedi, sondan 2. Fatura ödemesi= ödedi ve sondan fatura ödemesi= ödedi ise fraud mu= hayır

Rule 7: [0 4 61 1 1 1 0 3 4 1 4 1 1 0] [1]

Eğer cinsiyet= erkek ve yaşı= 41-45 arası ve il=Trabzon ve abonelik yaşı= 6-12 ay arası ve operatör değişikliği-eski operatörü= vodafone ve kaç operatörde aboneliği var= 2 ve mevcut operatörü= avea ve geç ödenen fatura sayısı= 4 ve ortalama aylık fatura tutarı= 75-100 arası ve aylık ortalama kullanılan sms sayısı= 30-45 arası ve son borç durumu= 75-100 arası ve sondan 3. fatura ödemesi= ödemedi, sondan 2. Fatura ödemesi= ödemedi ve sondan fatura ödemesi=ödedi ise fraud mu=hayır

Anfis modelinin training uygumasıdan elde edilen yukarıdaki kurallara göre, her bir niteliğin, sıralama fonksiyonuna göre, sınıflandırmaya olan etkileri farklı boyutlardadır. Örneğin; aylık ortalama sms sayısı yüzde 40, son borç durumu yüzde 39, yaşı yüzde 33, son operatörde bulunduğu süre yüzde 32, ortalama aylık fatura tutarı yüzde 24, sondan 2. Fatura ödemesi yüzde 21, son fatura ödemesi yüzde 18, il

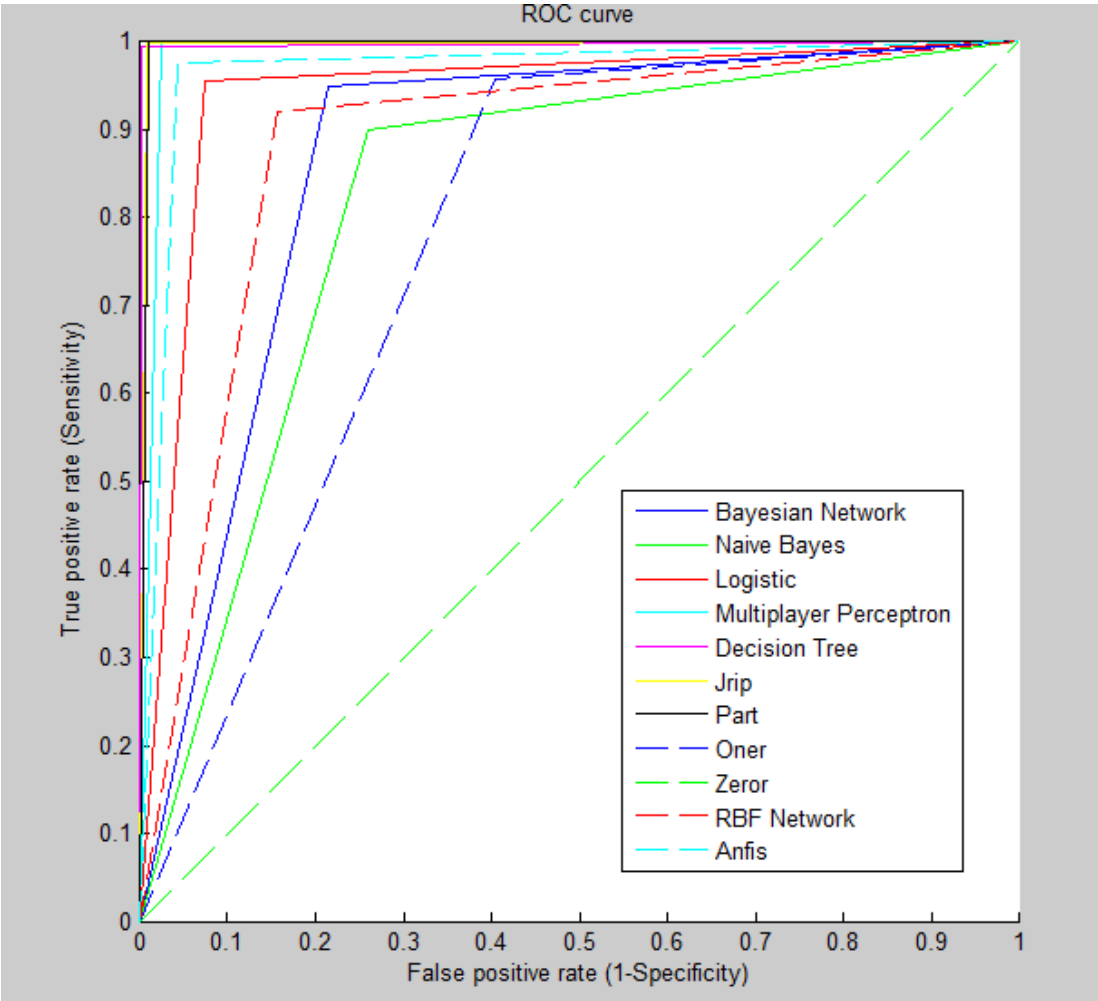
yüzde 11 ve geç ödenen fatura sayısıda yüzde 9 oranında anfis çıktısını etkilemektedir.



Şekil 5.14 ANFIS ROC Eğrisi

5.12 BULGU DEĞERLERİ UYGULAMA ÖZETLERİ

Bütün sınıflandırma yöntemlerinin ROC eğrileri şekil 5.15 te gösterilmiştir. ROC eğrilerinden yola çıkarak, Anfis, Karar Ağaçları, Part ve Çok Katmanlı Algılayıcı yöntemlerinin, tanı testi performans eğrilerinden, uygumamız için tutarlı datalar elde edilmiştir. Zeror, Oner ve Naive Bayes uygulamarındaki tutarsız datalar ise, diğer yöntemlere göre artış göstermiştir.



Şekil 5.15 Sınıflandırma Yöntemleri ROC Eğrileri

Bu eğrilere göre performansı en iyi çıkan yötemler; Kısmi Karar Ağaçları, Çok Katmanlı Alıgılayıcı, Part ve Anfis uygulamalarıdır. Performansı en düşük çıkan uygulamalar ise; Zeror, Naive Bayes ve Oner uygulamalarıdır.

Tablo 5.21: Sınıflandırma Yöntemleri Özet Çıktıları

Model	**Duyarlılık(TPR)**	**Özgüllük (1-FPR)**	**RMSE**
BAYES AGLARI	0.949	0.785	0.265
NAIVE BAYES	0.901	0.741	0.3351
Lojistik	0.956	0.925	0.2205
Çok Katmanlı Algılayıcı	0.999	0.976	0.0819
Kısmi Karar Ağaçları	0.995	0.997	0.0911
Jrip	0.998	0.992	0.0565
Part	1	0.989	0.0527
Oner	0.958	0.595	0.3787
Zeror	1	0	0.4489
Rbf Ağları	0.920	0.843	0.2761
ANFIS	0.976	0.9565	0.038

Tablo 5.21‘ e göre hata oranı en düşük olan yötem ANFIS olarak görünmektedir. Hata oranının en yüksek çıktığı uygulama ise Zeror uygulamasıdır.

6. SONUÇ

Bu çalışmada, veri madenciliği sınıflandırma yöntemlerinin hemen hemen bütün modellemeri datalar üzerinde uygulanılmıştır. Her bir modellemede bir birine çok yakın fakat farklı değerler elde edilmiştir.

Burada kullanılan datalar, gerçek bir telekom şirketinin datalarıdır ve 1256 erkek ve 1192 bayandan oluşan toplam 2448 datadan oluşmaktadır. Çalışma için her bir müşterinin 15 özelliği üzerinden değerlendirme yapılmıştır. Weka programının sıralama fonksiyonu kullanılarak (InfoGainAttributeEval with Ranker), müşterilerin en çok hangi niteliklerinin sahtekarlığa eğilimlerini artırdığını görebiliriz. Buna göre; aylık ortalama kullanılan sms sayısı , son borç durumu, yaş, abonelik yaşı, ortalama aylık fatura ödemesi, son fatura ödemesi, il ve geç ödenen fatura sayısı en fazla etkileyen faktörler olarak sıralanabilir.

Modellemelerin en iyi performans gösterini, RMSE değerinden ve ROC grafiğinden de anlaşılacağı gibi ANFIS ve onu takiben desicion table uygulamasıdır. ROC eğrisinin altında kalan alan 1'i vermektedir. Buradaki doğru sınıflandırılmış örneklem, 2448 kaydın içerisinden yalnızca 9 tanesidir.

Modellemelerin içerisinden performansı en düşük olan uygulama ise ZEROR uygulamasıdır. 2448 müşteri kaydından 635 tutarsız data çıkmıştır. ROC eğrisinden de anlaşılacağı gibi eğrinin altında kalan alan gerçek positif değerleri, üstünde kalan alan ise yanlış positif değerleri, yani başarısız çıkan tutarsız sonuçları göstermektedir. Burada ki eğride hemen hemen bir eşitlik söz konusudur.

Sonuç olarak; ANFIS, Karar Ağaçları, Çok Katmanlı Algılayıcı modellerinin performanslarının başarılı çıktığını, tam tersine ZEROR modellemesinin performansının ise başarız çıktığını söyleyebiliriz.

KAYNAKÇA

Kitaplar

Çağıltay, N. E., 2010. *İş Zekası ve Veri Ambarı Sistemleri*. Ankara: ODTU Geliştirme Vakfı Yayıncılık.

Silahtaroğlu, 2008. *Kavram ve Algoritmaları ile Temel Veri Madenciliği***.** İstanbul: Papatya Yayıncılık Eğitim.

Niu, L., Lu, J. ve Zhang G., 2009. *Cognition-Driven Decision Support for Business Intelligence*: *Models, Techniques, Systems and Application.* Springer.

McClish, D.K., 1987. *Comparing The Areas Under More Than Two Independent ROC curves*. Med Decis Making, pg 148-156.

Karahoca, D. & Karahoca, A., 1998. *İşletmeciler, Mühendisler ve Yöneticiler İçin Yönetim Bilişim Sistemleri ve Uygulamaları*. İstanbul: Beta Yayınları.

Klir, J.G. & Yuan, B., 1995. *Fuzzy Sets and Fuzzy Logic Theory and Applications*. New Jersey: Prentice Hall.

Moler, C., 2008. *Experiments with MATLAB, The MathWorks*. Ebook: www.mathworks.com.

Stinnet, B. , 2008. *Müşterin Gibi Düşün*. Ankara: ODTÜ yayıncılık, ss.15-45

Timothy, J.R., 1995. *Fuzzy Logic with Engineering Applications*. Newyork : Mc Graw-Hill.

Şimşek, U. T., 2006. *Veri Madenciliği ve Müşteri İlişkileri Yönetiminde Bir Uygulama*. İstanbul

Witten, I.H. & Frank, E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers*. San Fransisco.

Diğer Yayınlar

Tanrıkulu, H. & Dr. Sazlı M. H., *Saldırı Tespit Sistemlerinde Yapay Sinir Ağlarının Kullanılması*. Anlara: Ankara Üniversitesi.

Helberg, C. *Data Mining with Confidence*. SPSS sunumu, 2002

Farvaresh, H. & Sepehri M. M., *A datamining framework for detecting subscription fraud in telecommunication*

Ahn, H., Ahn, J. J., Oh, K. J. and Kim, D. H., *Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques*

Öztürk, C. A. & Mercan, D.E., Toprak, F., Kişi, Ö. ve Şahin, U., 2003. *Bulanık Mantık Kurs Notları*, İstanbul: İTU Bulanık Mantık ve Teknoloji Kulübü.

Philip K. C, *Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection.* Columbia University.

Nguyen, T. M., Schiefer, J., Tjoa, A. M., *Vienna University of Technology, An approach towards a real-time business intelligence solution and its use for a fraud detection application*