T.C. İSTANBUL TİCARET
ÜNİVERSİTESİ

E-TİCARET SİTELERİ İÇİN SAHTEKÂRLIK TESPİT SİSTEMLERİ

YASİN KIRELLİ

Danışman
Yrd. Doç. Dr. Mustafa Cem KASAPBAŞI

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

İSTANBUL 2016

KABUL VE ONAY SAYFASI

Yasin KIRELLİ tarafından hazırlanan **"E-Ticaret Siteleri İçin Sahtekârlık Tespit Sistemleri"** adlı tez çalışması 01.06/2016 tarihinde aşağıdaki jüri üyeleri önünde İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak başarı ile savunulmuştur.

Danışman	**Yrd. Doç. Dr. Mustafa Cem Kasapbaşı** İstanbul Ticaret Üniversitesi	
Jüri Üyesi	**Prof. Dr. Selim AKYOKUŞ** Doğuş Üniversitesi	
Jüri Üyesi	**Yrd. Doç. Dr. Metin TURAN** İstanbul Ticaret Üniversitesi	

Onay Tarihi : ..../..../20...

Prof. Dr. Doğan KAYA
Enstitü Müdürü

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

17.06.2016

Tarih

İmza

Yasin KIRELLİ

İÇİNDEKİLER

ÖZET

Yüksek Lisans Tezi

E-TİCARET SİTELERİ İÇİN SAHTEKÂRLIK TESPİT SİSTEMLERİ

YasinKIRELLİ

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Ana Bilim Dalı

Danışman: Yrd. Doç. Dr. Mustafa Cem KASAPBAŞI
2016, 58 Sayfa

İnternet üzerinden yapılan alış verişlerde sahtecilik içeren işlemler, ana kaygılardan biridir. Dolandırıcılık işlemleri sadece müşteriler ve E-Ticaret şirketlerini değil, aynı zamanda bankalar için de mali kayıplara neden olmaktadır. Bu nedenle, sahtecilik olarak ilişkilendirebilecek siparişleri sınıflandırabilmek ve tespit edebilmek E-Ticaret siteleri için büyük önem taşır. Bu türde sahtecilik tespiti, bankacılık sektöründe olduğu gibi müşterileri hakkında bolca bilgi bulunduğundan daha kolaydır ancak, ticari internet sitelerindemüşteri hakkında uygun nitelikleri bulmak daha zordur. Bu çalışmada bir E-Ticaret sitesinin gerçek verileri, yasa dışı kredi kartı kullanımlarınıanaliz etmek için kullanılmıştır. Öncelikle tüm ham veri analiz edilmiş ve eksik değerlerinden filtre edilmiştir. Gainratio algoritmasıyla en uygun değerler seçilmiş, sonrasında veri madenciliği tekniğiyle Navie Bayes, Karar Ağacı (J48) algoritmaları kullanılarak, %95'ten fazla doğru sınıflandırma oranıyla sahtecilik içeren işlemler tespit edilip sınıflandırılmıştır.

Anahtar Kelimeler: Sahtecilik, makine öğrenmesi, e-ticaret sahtekârlık tespiti.

ABSTRACT

M.Sc. Thesis

FRAUD DETECTION SYSTEM FOR E-COMMERCE SITES

Yasin KIRELLİ

İstanbul Commerce University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Assist. Prof. Dr. Mustafa Cem KASAPBAŞI
2016, 58 Pages

Fraudulent transactions are one of the main concerns regarding online shopping. Fraud transactions cause financial losses for not only to customers and E-Commerce merchants but also to the banks. Therefore, it is crucial for E-commerce sites to have capabilities to detect and to classify product orders that can be attributed as fraud. These kinds of fraud detections are easier when there is available abundant information about clients as in Banking but it becomes challenging to find proper attributes in commercial web sites. In this study real transaction data of an E-Commerce site are used to analyze for illegitimate use of credit card transactions. Firstly all raw data analyzed and filtered from missing values. Appropriate attributes are selected using gainratio algorithms, after then Fraudulent transactions have been detected and classified and a true positive rate more than %95 is obtained using data mining techniques namely, Naïve Bayesian, Decision trees (J48).

Keywords: Fraud, machine learning, e-commerce fraud detection.

TEŞEKKÜR

Bu çalışmanın gerçekleşmesinde katkılarından dolayı ve danışmanım olarak tezin yazılmasında yol gösteren Saygıdeğer öğretmenim Yrd. Doç. Dr. Mustafa Cem Kasapbaşı ve verilerin toplanması sürecinde bana yardımını eksik etmeyen Gökhan Uygun'a teşekkür ederim.

Yasin KIRELLİ

İSTANBUL, 2016

ŞEKİLLER

ÇİZELGELER

Sayfa

SİMGELER VE KISALTMALAR

FN	Yanlış tahmin edilen negatif değer (False negative)
FP	Yanlış tahmin edilen pozitif değer (False positive)
IBK	En yakın komşu algoritması (Instance based learning algorithm)
J48	Karar ağacı algoritması
KNN	K en yakın komşu algoritması (K nearest neighbor)
NBTree	Naive Bayes Tree
RBF	Radyal temelli fonksiyon (Radial basis function)
ROC	Alıcı işletim karakteristiği (Receiver operation characteristics)
TN	Doğru tahmin edilen negatif değer (True negative)
TP	Doğru tahmin edilen pozitif değer (True positive)

1. GİRİŞ

Günümüzde Internet üzerinden sipariş yaygınlaşmasıyla birlikte sanal ortamda ticaret hacmi de giderek büyüme göstermektedir. Web ortamında binlerce ürün e-ticaret siteleri sayesinde çok kısa sürede sipariş yapılarak kullanıcıya buluşma olanağı da sağlamıştır. İnsanlar kolay alış veriş için sanal ortamlara bu nedenle yönelmektedirler.

Sahtecilik, e-ticaret siteleri için farklı bir anlam ifade etmektedir. Çalınmış bir kredi kartıyla ya da kredi kartı bilgilerinin kopyalanarak yapılan siparişler sahtecilik(fraud) olarak nitelendirilmektedir. E-ticaret sitelerinin sahtecilik olasılığını tamamen tespit edebilmeleri mümkün değildir, fakat bunu en minimum seviye düşürmeleri mümkündür. E-ticaret site sahibi kurumlar bu risk faktörünü en aza indirmek için ve müşteri memnuniyeti için sahtekârlık yöntemlerine karşı bir kontrol mekanizması kurmaları gerekmektedirler.

Internet üzerinden alışverişlerde herhangi bir siparişin sahtecilik olduğuna dair ipuçlarımız varsa sahteciliği yakalamak için de şansımızın olduğuna anlamına gelmektedir. Bu ipuçlarından başlıcaları, online sipariş üzerinden yapılan bir işleme işlemin yapıldığı bankadan riskli bir cevap kodu dönmesi, siparişte kullanılan kart sayısı, kullanıcının online alışerişte ip'sinden belirlenen konumunu üzerinde yorumlanması, kullanıcının bölgeye göre isim uzunluğu ve kullanım oranı, satın alınan ürünün fiyat bigisi veya adedi, kullanılan elektronik posta adresleri ve bu adreslerin uzantıları gibi bir çok kontrollerden geçirilerek, oluşturulan siparişin sahtekarlık niteliği taşıyıp taşımadığını anlamamız mümkündür.

Belirlenen kurallara bağlı olarak sahtecilik tespit işlemini otomatize edilmesi ve sahtekârlık niteliği taşıyan siparişlerin denetimli sınıflandırılması makine öğrenmesi yönetimiyle sağlanmıştır. Makine öğrenmesinde sınıflandırma yöntemi daha önceden bilinen ve gerçek siparişlerde yer alan sahtecilik olarak işaretli siparişlerinden bilgi çıkarımı ve kümlenmesi işlemidir.

Bu araştırmada veri madenciliğinde veri öğrenmesine ve tahminine dayalı sınıflandırma algoritmaları yardımıyla eldeki veriler ışığında sahtecilik olup olmadığı sınıflandırılmaya çalışılmıştır.

Proje sonucu oluşturulan yöntemle beraber gerçek bir elektronik ticaret sitesinin 1615 siparişinin özellikleri incelenerek doğru sınıfiseçme başarısı %95 olarak belirlenmiştir. Projelerin hazırlanış ve safhaları aşağıda sunulmuştur. Proje genel

hatlarıyla aşağıda belirtilmiştir. Değişkenlerin seçimi, eğitim verilerinin oluşturulması, algoritmanın uygulanması,deneme kümesinin sınanması ve sonuçların değerlendirilmesi olarak ele alınmıştır. Tezimizin son bölümünde sonuçlar değerlendirilmiş ve yorumlanarak gelecek dönemde e-ticaret sitelerinde sahtecilik tespiti konusunda öneriler sunulmuştur.

2. LİTERATÜR ÖZETİ

Veri madenciliği son yıllarda büyük finansal kurumların olmazsa olmaz ihtiyaçları haline gelmiştir. Veri anlamlandırma birçok sektörde aktif olarak kullanılmakta, kurum ve şirketlerin her anlamda stratejik kararlar almasında aktif rol almaktadır. Bu kısımda finansal ve diğer sektörlerde veri madenciliği kullanımı hakkında teknik çalışmalar belirtilmiştir.

Bir diğer çalışmada, Göral (2007) "Kredi Kartı Başvuru Aşamasında Sahtecilik Tespiti İçin Bir Veri Madenciliği Modeli" adlı tezinde, bankacılık sektöründe kredi kartı başvurusu için veri madenciliği modeli kullanarak karar mekanizması geliştirilmiştir. Geçmişten günümüze banka sayısında artışa bağlı olarak bankaların kar oranlarını artırmak için riskli gruptaki müşterilere de kredi kartı vermesinde artış göstermiş ve kredi kartına sahip kullanıcı sayısında ciddi artış gözlenmiştir. Bu çalışmada bir banka için kredi kartı başvurusunda sahtecilik tespiti kontrolü için veri madenciliği teknikleri kullanılarak başvuru aşamasında gerçekleşmesi sağlanmıştır.

Bir diğer çalışmada, Tavacı (2012) "Gsm Şebekelerinde Sahtekârlık Yönetimi İçin Veri Madenciliği Yöntemlerinin Uygulanması" tezinde ele alınan temel konu, gelişen gsm sektöründe abone saldırılarının, müşterilerin verilerinin analizi ile tespiti sağlanmıştır. Genel amaç abonelerin fatura ödememe veya sunulan kampanyalardan yararlanma durumu göz önüne alınarak abone saldırılarının önüne geçmeyi hedeflemektedir.

Bir diğer çalışmada, Özbay (2007) "Veri Madenciliği ile Dolandırıcılık Tespiti" çalışmasında günümüzde internet bankacılığı kullanımının yaygınlaşmasıyla beraber bu alanda güvenlik tehditlerini de yanında getirmiştir. Dolandırıcılık tespitinde Internet bankacılığı kullanan müşterilerde gruplamalar yapılarak veri madenciliği teknikleri kullanılarak sahtecilik işlemlerinin sayısı minimuma indirgenmeye çalışılmıştır.

Bir diğer çalışmada, Aral (2009) "Veri Madenciliği Teknikleri ile Reçete Usulsüzlüklerinin Tespiti" çalışmasında sağlık harcamaları konusunda reçete usulsüzlükleri ele almıştır.Çalışmanın amacı, hâlihazırda uzmanlar tarafından rasgele seçim yoluyla yapılan reçete usulsüzlüğü denetiminin etkin bir otomasyon sistemiyle sağlanması için özelleştirilmis veri madenciliği teknikleri geliştirililmesidir.

Bir diğer çalışmada, Sezen (2011) “Telekomünikasyon Sistemlerinde Sahtekârlık Tespiti ve Yönetimi” telekominikasyon şirketlerinde zarara yol açan sahtekarlık amaçlı kullanımların, arama detay kayıtlarından elde edilen veriler kullanılarak yapay sinir ağları yaklaşımları ile tespit edilmesi gerçekleştirilmiştir.

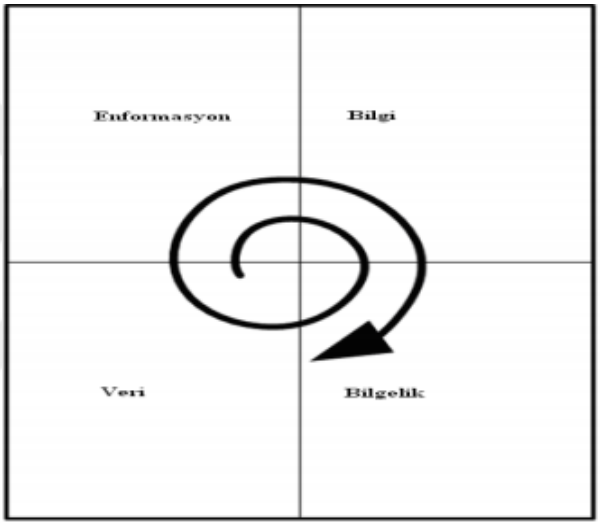
Bir diğer çalışmada, Uğurlu (2007) “Finansal Tablolardaki Hile Riskinin Belirlenmesi” tezinde bankaların temel fonksiyonu olan kredi kullandırım konusunu ele almıştır. kullandırımı olması nedeniyle bankaların karşı karşıya kaldığı risklerin başında kredi riski gelmektedir. Finansal tablolara dayalı olarak gerçekleştirilen ticari kredi kullandırımlarında, hileli finansal tablolar üzerinden kullandırılan kredilerin kısmen veya tamamen geri dönüşümünün sağlanamaması kredi riskine sebebiyet vermekte ve bankalar açısından önemli bir sorun oluşturmaktadır. Finansal tablolardaki hile riskinin öngörülmesi ve değerlendirilmesinde yöntem olarak yapay sinir ağı teknolojisinden yararlanılmıştır.

Bir diğer çalışmada, Aşuk (2010) “Sağlık Sigortası Şirketleri İçin Veri Madenciliği Tabanlı Suistimal Tespit Sistemi” çalışmasında sağlık sektöründeki suiistimalleri ele almıştır. Sağlık suistimalleri, kişilerin veya kurumların haksız kazanç sağlamak adına kasıtlı olarak yaptıkları hile, yanlış beyan ve benzeri sahtekârlıklardır. Sigorta sistemine müşteri olarak giren insanların sayısının çok artması ile tıbbi suistimallerin sağlık sigorta sektörüne verdiği zarar gittikçe büyümektedir. Bu çalışmada sağlık sigorta şirketleri için veri madenciliği tabanlı suistimal tespit sistemi incelenmektedir.

Belirtilen çalışmalara istinaden veri madenciliği bankacılık ve finans, sigortacılık, haberleşme, sağlık ve eğitim gibi birçok alanda eldeki verilerin anlamlı hale getirilerek önlem alınması konusunda önemli bir araç olmuştur. Günümüzde bilgisayarların ve akıllı teknolojik aletlerin günlük hayatımızda önemli yer tutmasıyla birlikte bankacılık ve e-ticaret konusunda güvenlik diğer sektörlere göre daha önemli hale gelmiştir. Çalışmamızda e-ticaret sektöründe kredi kartı sahtekârlık tespitine yönelik, sınıflandırma algoritmaları kullanılarak önlem alınması sağlanmıştır.

3. BİLGİNİN OLUŞUM SÜRECİ

Veri, enformasyon ve bilgi taşıdıkları anlam seviyelerine göre sıralandığında piramitte en üst noktayı bilgi alacaktır. En alt seviyede ise veri yer alacaktır. Burada konu alan veri en temel nitelikleri ve gerçekleri incelerken enformasyon verilerden beslenerek ortaya veri ilişkilerinden doğan sonuçlar elde edilir. Bilgi ise en üst seviye potansiyeli olan ve geleceğe dair çıkarımlar yapmaya yarayan enformasyonlar olarak adlandırılır (Medeni, 2007). En genel tabirle veri geçmişte olan olaylarla ilgilenirken enformasyon şimdiki zaman dilimiyle ilgilenir. En üst düzeyde bulunan bilgi ise gelecekte olabilecek çıkarımları elde etmemizde önemli rol oynar (Şekil 3.1).



Şekil 3.1. Enformasyonun oluşumu (Medeni, 2007)

Veri bilginin oluşmasında ayrık durumda olan nesnelerden ve gerçeklerden gücünü alır ve inceler. Enformasyon değerlendirilecek olursa elde edilen verilere bağlı olarak bir yorum yapma ve veriyi anlamlandırma yeteneğidir. Bilgide ise insan beyninde oluşan ve şekillenen düşünce ve izlenecek yola kadar geniş bir alana yayılır (Medeni, 2007).

3.1. Veri (Data)

Bilgisayar bilimlerinin günümüz teknolojisiyle hızlı geliştiği gibi veritabanı depolama araçları da yaygınlaşıp tüm organizsyonların kullabileceği kolaylığa gelmiştir. Organizasyonlar süreçlerini, fonksiyonlarını bir hesaplama niteliği olarak kullanma ya da bu fonksiyonel gerçeklerinde manipilasyonlar yapabilmek adına kodlama ve saklaması gelişen veritabanı teknolojisiyle kolay hale gelmiştir. Bu saklama biçimi ikili değer, harf veya sayısal bir karşılığı ifade edebilir (Çerkez, 2003).

Veri tabanı teknolojisiyle bu saklanmış veriler daha yönetsel bir eyleme olanak da sağlamıştır. Örnek olarak bir e-ticaret sitesinde bir müşterinin alışverişinde, o müşteriye ait kullanıcı bilgileri yer alacaktır. Kullanıcının adı, soyadı ya da bir kullanıcı id'si olabilir. Aynı zamanda bu kullanıcının site üzerinden almış olduğu ürünün bilgileri de yer almalıdır. Ürünün ismi, ürünün kodu, ürünün fiyatı ve bu kullanıcının bu ürünü hangi tarihte aldığı gibi süreçleri bir işlemi ifade etmektedir. Aşağıda da bu sürece ait verilerini veritabanında örnek gösterimi yer almaktadır (Çizelge 3.1).

Çizelge 3.1. Kullanıcının bir hareketinin veri örneği

Kategori id	**Fiyat**	**Kullanıcı No**	**Tarih**
Eldiven	15	5	12.5.2010

3.2. Enformasyon

Veriler kendi başlarına bir anlam ifade etmezler. Bunları anlamlandıracak durum bu verilerin işlenmesidir. Veri örneğinden devam edilecek olursa site üzerindeki sipariş verileri çoğaldıkça bu verilerin saklanması, toplam sipariş sayılarının alınması, dönem tabanlı sipariş raporlarının oluşturulması, aylık kaç liralık ürünün satıldığı gibi toplu şekilde verilerin kullanılması gerektiren işlemler enformasyonu oluşturacaktır. Aşağıdaki örnekle ayrı süreçlerde işlenen verilerin bir araya getirdiği enformasyon yer almaktadır (Çizelge 3.2).

Çizelge 3.2. Verilerin bir araya getirdiği enformasyon

Kategori	**Fiyat**	**Kullanıcı No**	**Tarih**
Eldiven	15	5	12.5.2010
Ayakkabı	50	4	14.6.2010
Fener	10	34	11.4.2010
Gözlük	70	77	10.7.2010

Enformasyonu daha etkili kullanmak gerekirse, elde edilen veriden yeni kısıtlamalar veya birleştirme ölçütleriyle daha anlamlı veriler elde etmek mümkün olacaktır (Kılıç, 2010). Örnek üzerinden devam edildiğinde, kategorilerin dönem bazında satış adetleri veya toplam değerleri elde edilen enformasyonun değerini daha da artıracaktır (Çizelge 3.3). Yeni elde edilen enformasyon durumunu özetlemek gerekirse:

Çizelge 3.3. Yeni oluşturulan enformasyon

Kategori	**3.Ay**	**4.Ay**	**5.Ay**	**Miktar**	**Fiyat**	**Toplam Tutar**
Eldiven	36	46	37	119	15	1785
Ayakkabı	120	80	100	300	50	15000
Fener	60	30	20	110	10	1100
Gözlük	50	60	75	185	70	12950

Tabloya enformasyonundan elde edilen sonuca göre eldiven ve ayakkabı satışları dönemden etkilenmeyip fener ve gözlük satışları dönemlere göre artma ve azalması gözlemlenebilmiştir.

3.3. Bilgi

Elde edilen verilere göre göre bilgi, enformasyon ve veriden daha geniş kapsamlı olarak durumlara bakmayı sağlamaktadır. Bilgi bize gelecek hakkında öngörü sunmayı ve tahmine olanak sağlayacak bir kavrama karşılık gelmektedir.

Elde edilen enformasyonun yorumlanması için veri madenciliği teknikleri kullanılarak anlamlı bir bilgi ortaya çıkması mümkündür.

Tablodaki verileri veri madenciliği teknikleri sonucu elde edilen bir pazarlama stratejisi olarak düşünüldüğünde, eldiven alan kişilerden yüzde ellisinin da ayakkabı aldığını öğrenmek kolay düşünülemeyecek bir bilgiyi ifade etmektedir (Çizelge 3.4).

Bu anlamda bakıldığı zaman bilgi kavramı veri ve enformasyondan faklı olarak anlamlı bir çıkarım sunabilmektedir.

Bilginin oluşma aşaması ve teknolojik karşılığı tabloda belirtilmiştir.

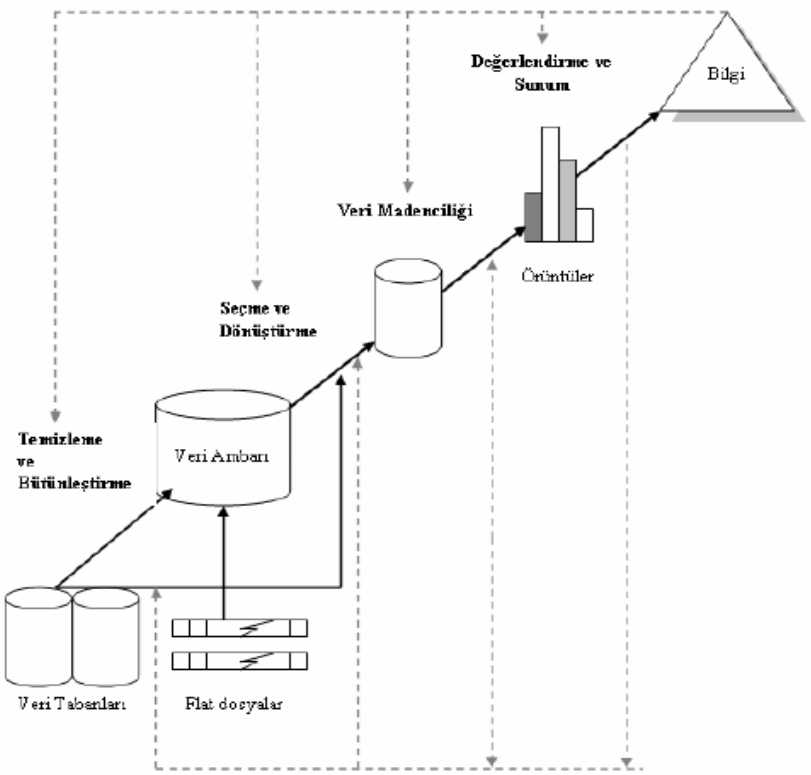
Çizelge 3.4. Bilginin oluşma aşaması

Anlam aşaması	**Teknoloji**
Veri	Veri tabanı sistemleri
Enformasyon	Sorgulama ve raporlama araçları
Bilgi	Veri madenciliği teknikleri

4. VERİ MADENCİLİĞİ

Veri depolama gereksinimiyle beraber veritabanı teknolojisi de gün geçtikçe daha ileri boyutlara gelmiştir. Bilgi depolama teknolojisinin gelişmesiyle birlikte bu depolanan bilgilerden nasıl yararlanılması gerektiği asıl problem olarak ortaya çıkmıştır. Bu sorunlara çözüm olarak klasik raporlama ve sorgulama teknikleriyle bilgiler yararlı bilgiye dönüştürülebilmiştir. Fakat gelecek için çıkarımlar veya bilgideki örüntüleri anlamak için klasik yöntemler yetersiz kalmıştır. Bu durumda depolanmış veriden bilgi keşfi diye nitelendirebileceğimiz "Veri Madenciliği" ortaya çıkmıştır.

Veri Madenciliğinde, veriden anlamlı bilgiler elde etmek için kullanılan verinin bir takım safhalardan geçirilerek, çeşitli veri işleme algoritmalarıyla anlamlı ve yararlı bilgiye ulaşmak mümkün hale gelmiştir (Şekil 4.1).



Şekil 4.1. Yararlı bilgiye ulaşma safhaları (Göral,2007)

Veri Madenciliğinde anlamlı bilgiye ulaşmak için veri işleme safhaları özetle aşağıdaki gibidir (Türkoğlu, 2009):

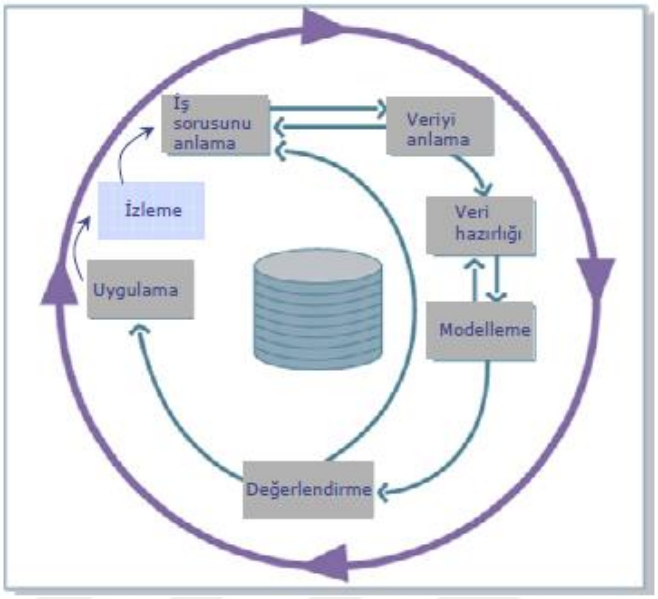
- *Veri Temizleme*: Veri tabanında gereksiz kayıtların temizlenmesi işlemidir. Bu süreç için örnek olarak test amaçlı oluşturulan veriler veya veri kümesi için gereksiz null kayıtlı işlemlerinveri kümesinden temizlenme aşamasıdır.
- *Veri Bütünleştirme:* Bu safhada bütünsel bir veri kümesi oluşturabilmek adına farklı tabloların birleştirmesi işlemidir.
- *Veri seçme:* Bu safhada ise analiz edilecek gerçek verilerin veritabanından çekilmesi işlemidir.
- *Veri Dönüştürme:* Veri analizi için veritabanında bulunan alanların veri madenciliğinde kullanılacak veri tiplerine dönüştürülme aşamasını ifade eder.

Veri madenciliği organizasyonların pazarda rekabetçi bir rol üstlenmesinde önemli bir görev üstlenmektedir. Büyük verinin işlenmesi ve bilgi içerisindeki örüntünün keşfi bu sayede ortaya çıkmıştır.

4.1. Veri Madenciliği Yararlı Bilgiye Ulaşma Süreçleri

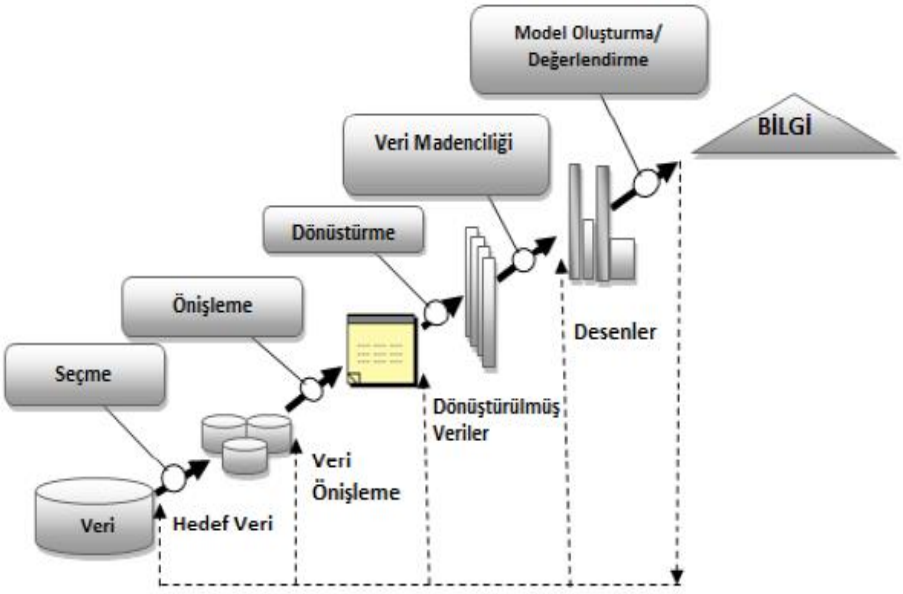
Veri madenciliğinde standardizayon sağlamak amacıyla belirli süreçler kabul görmüştür. Farklı kullanım alanlarına kolay uyum sağlaması, daha kolay gerçekleştirilebilmesi edilebilmesi, daha güvenli bir çözüm sunması adına belirlenmiş süreçler bütününden oluşur. Bu standardizyon için proje olarak CRISPDM (Cross Industry Standart Process For Data Mining) modeli ortaya atılmıştır (Ahi, 2015).

Şekil 4.2'de belirtildiği gibi veri madenciliği süreci yedi aşamadan oluşmaktadır. Bu gerçek bilgiye ulaşmaküzere için bir yaşam döngüsüdür.



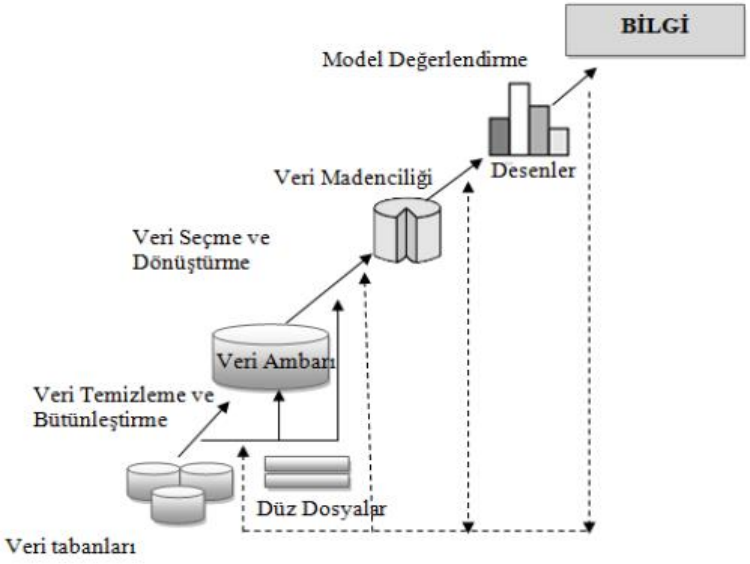
Şekil 4.2. Veri madenciliği süreci (Göker, 2012)

Fayyad'a göre veri madenciliği modeli ise Şekil 4.3'te safhalara göre belirlenmiştir.



Şekil 4.3. Fayyad'a göre veri madenciliği modeli (Fayyad, 1996)

Han'a göre veri madenciliği süreci ise Şekil 4.4'teki safhalardan oluşmaktadır.



Şekil 4.4. Han'a göre veri madenciliği modeli (Han, 2001)

Bütün standartlaşma süreçlerini incelediğimizde hepsinin benzer süreçlerden geçtiğini görebilmektedir. Genel anlamda süreçleri ifade edilecek olursa;

- Problemin tanımı
- Gerçek bilgi için verinin işlenme aşaması
- Verinin temizlenmesi
- Verinin dönüştürülmesi
- Verinin bütünleştirilmesi
- Değişken seçimi
- Model Oluşturma ve Değerlendirme
- Biginin sunum aşaması

4.1.1. Problemin tanımı

Problemin belirlenme aşamasında yapılan projenin hangi alanda kullanılacağı belirlenmelidir. Hedefler ve veri işleme sonucu ortaya çıkan örüntünün ne şekilde kullanılacağı ve değerlendirileceği net bir şekilde belirlenmelidir.

Bu aşamada veri madenciliği projenin ne için yapıldığı, projenin ihtiyaçları, analizleri ve stratejiler için bir yol haritası oluşturulur.

4.1.2. Verinin temizlenmesi

Veri kümesi oluşturma kısmında önemli adımlardan bir tanesidir. Bu aşamada veri kümesindengürültülü verilerden arındırılması aşamasıdır. Gürültülü bir veri kümesi başarılı bir sonuca ulaşmamızı engelleyecektir.

Bunun için aşağıdaki yöntemler uygulanabilir:

- Veri kümesinde eksik bir değere ait olan alanlar silinebilir.
- İşlem kümesinde null kayıtlar için belli bir değer verilebilir ya da veriden silinebilir.
- İşlem kümesiiçin eksik değerler ortalama bir değerle düzeltilebilir. Buna karar vermek için karar ağacı ya da regresyon kullanılabilir.
- Mükerrer kayıt içeren kayıtlar temizlenmelidir.

4.1.3. Verinin bütünleştirilmesi

Veri birleştirme işlemi farklı şemalarda bulanan verilerin birleştirilmesi olayını kapsar. Örneğin bir tabloda “sipariş-no” olarak kaydedilen bir alan farklı bir şemadaki tabloda “siparis-Id” olarak karşılaşılabilir. Bu tip tablo birleştirmelerinde meta anahtarlardan faydalanılarak birleştirme işlemi yapılır.

Farklı bir durum olarak da verinin bir kısmı tablolarda tutuluyorken bir kısmı da dosya ortamında tutuluyor olabilir. Farklı kaynaklardaki verileri kullanarak bir sonuca ulaşmaya çalışmak hem maliyet hem de zaman kaybına yol açmaktadır.

4.1.4. Verinin dönüştürülmesi

Verileri daha net belirleyebilmek adına,verilerin kategorilendirilmesi olarak tanımlanabilir. Bu aşamada verinin min-max değerleri göz önüne alınarak bu aralığın fazla olması diğer değerlerin başarısını etkileyebilir. Değişkenin diğer değişkenlere etkisinin azaltılabilmesi için aralığı geniş olan veriler normalize edilir. Böylece sonucun daha başarılı olması gözlemlenebilir.

Veri değiştirme işlemleri Z-Score kartları veya Min-Max Normalleştirme gibi algoritmalarla yapılır.

Z Score,değişken verilerinde oransal dağılım gösterdiği durumlarda en çok kullanılan algoritmalardan bir tanesidir.

$$z_i = \frac{Xi - \vec{X}}{S} \tag{4.1}$$

Formule göre veriler Z Score'larına dönüştürülür.

Xise belirli değişkendeki değerlerin aritmetik ortalamasını ifade eder.

S değeri verideki standart sapmadır.

Min-Max algoritmasında ise verideki en büyük değer Xmax olarak ifade edilir ve verinin en küçük değeri de Xmin'dir. Xmax-Xmin değeri ise veri genişliğini(R) ifade eder.

$$R = Xmax - Xmin \tag{4.2}$$

$$X_i = \frac{Xi - Xmin}{R} \tag{4.3}$$

4.1.5. Değişken seçimi

Veri madenciliğinde en önemli kararlardan biri de değişken seçimidir. Analiz ve zaman maliyeti göz önüne alındığında en uygun değişken seçimi yapılmalıdır. Analiz ve modelleme sürecinde karmaşıklığı azaltmak için veri kümesi üzerinde değişken arası uyum göz önüne alınır.Veri kümesindeki uyumsuz değişkenler başarı oranını da

olumsuz etkileyecektir, bu yüzden değişken seçiminde birbirleri arasında bir korelasyonun olması sonucun başarısını etkileyecektir.

Değişken seçimi algortimada boyutu indirgeyebilmek için uygulanır. Değişken sayısının en az değeri sınıf sayısı kadar olmalıdır. Değişken ayırt edici ise seçilmelidir.Veri madenciliği uygulamalarında yaygın kullanıma sahip olan Weka uygulamasında değişken seçiminde CfsSubSetEval, GainRatioAttributeEval algoritmaları kullanılabilir.

4.1.6. Model oluşturma ve değerlendirme

Son durumda işlem görmüş ve modelleme safhasına gelmiş veri kümesi en iyi sonuç veren algoritma kullanılarak uygulamaya geçilir. Algoritma seçiminde veri kümesi üzerinde uygulanan çeşitli algoritmalar içerisinde en iyi sonuç vereni temel alınır. Algoritma sonucu oluşan örüntüye bağlı olarak yorumlamalar yapılır.

Modelleme aşaması denetimli ve denetimsiz öğrenme öğrenme modeline göre farklılık gösterir.

Denetimli öğrenmede sınıf bilgisi önceden bellidir. Verilerin algoritma sonucu olarak hangi sınıfa ait olduğu tahminde bulunur.

Denetimsiz öğrenme sınıf bilgisi bilinmeyen veri kümeleri için kullanılır.Veriler arasında kümeleme yapılmaya çalışılır. Veriler arasındaki uyum ve bağımlılık incelenir.

Sınıflandırma algoritmaları denetimli, kümeleme algoritmaları ise denetimsiz algoritmalara örnektir.

Denetimli öğrenmede veri kümesi için en uygun sınıflandırma algoritması seçildikten sonra veriler eğitim ve sınama verisi olarak ikiye ayrılır. Eğitim verisinde model eğitilir ve test verisiyle sınıflandırmalar tahmin edilmeye çalışılır.Tahmin edilen sınıflandırmaların başarı oranı belli yöntemlerle belirlenir. Bu yöntmeler aşağıdaki gibidir:

*Yalın Doğrulama (Simple Validation)**:*** Sonuç değerlendirmede oluşturulan en basit yöntemlerden bir tanesidir. Geçerlilik yönetiminde verilerin %5-33 arasında bir test kısmı oluşturulur. Verinin geri kalan kısmıyla modelin öğrenmesi sağlanır. Bu

yöntemde başarı oranı yanlış sınıflandırılmış veri sayısının tüm veriye oranı hata oranı verir. Doğru sınıflandırılmış veri sayısının tüm veriye oranı ise doğruluk oranını verir.

Doğruluk Oranı + Hata Oranı = 1

*Çapraz Doğrulama (Cross Validation)**:*** Veri kümesinin sınırlı olduğu durumlarda kullanılır. Veri kümesi rastgele bir oranla ikiye bölünür. Bölünen veri kümesinin ilk bölümü model ve ikinci bölümü sınamakümesi olarak sınanır ve bir hata oranı elde edilir. Sonrasında tam tersi işlem uygulanarak ilk bölüm test ve ikinci bölümü model olarak belirlenip yeni hata oranı bulur. Çıkan hata oranlarının ortalaması başarı oranını belirler(Wikipedia, 2016a).

*K-Kat Çapraz Doğrulama (K-Fold Cross Validation)**:*** Bu algoritmada veriler n parçaya bölünür. N parçadan bir tanesi test için diğer parçaları öğrenim olarak değerlendirilir ve bir hata oranı elde edilir. Bu durum her bir parça için ayrı ayrı sınamakümesi olacak şekilde sınandıktan sonra elde edilen hata oranlarının ortalaması başarı oranını belirler.

*Ön Yükleme (BootStrapping)**:*** Genelde veri kümesinin az olduğu durumlarda yararlanılan bir yöntemdir. Bu algoritmanın temel mantığı her defasında bir veri dışarıda bırakılarak bir hata oranı ölçülür ve her bir işlem için belirlenen hata oranının ortalamasıyla bir başarı oranını belirler.

Uzatma (Holdout): Veri dağılımında bir denge var ise verideki sınıf değerleri eğitim ve test verisinde eşit dağılım gösteriyor ise basit bir geçerlilik yöntemi uygulanabilir. Ama holdout yönteminde belli bir kısım test için kullanılır kalan kısım eğitimde kullanılır. Ortaya çıkan hata oranından başarı oranı elde edilir.

Seçilen algoritmanın ve kurulan modelin başarı ölçümünde karışıklık matrisi kullanılır. Örneğin iki sınıf için kullanılan bir karışıklık matrisi Şekil 4.5.'teki gibi değerlendirilebilir (Göker 2012).

		Tahmin Edilen Sınıf	
		Sınıf=1	Sınıf=0
Gerçek Sınıf	Sınıf=1	TP	FP
	Sınıf=0	FN	TN

Şekil 4.5. Karışıklık matrisi (Göker 2012)

Sonuç olarak elde edilen:

TP (True Positive)ve FP (False Positive) doğru tahmin edilen değerler,

FN (False Negative) ve TN (True Negative) yanlış tahmin edilen değerleri ifade eder.

Seçilen algoritmaların performansının değerlendirilmesinde başarı ölçütlerindekesinlik oranı, duyarlılık ve f oranları başarı oranı tespitinde kullanılır.

$$\text{Doğruluk} = \frac{TP+TN}{TP+TN+FN+FP} \quad (4.4)$$

$$\text{Duyarlılık} = \frac{TP}{TP+FP} \quad (4.5)$$

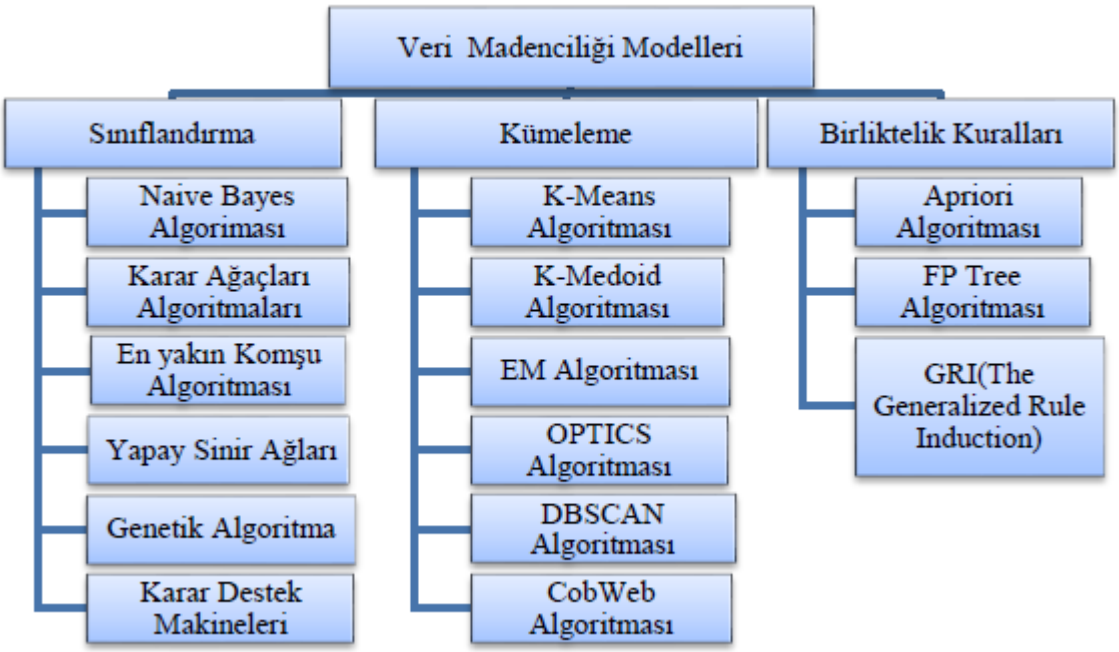
$$\text{Kesinlik} = \frac{TP}{TP+FN} \quad (4.6)$$

$$\text{F Ölçütü} = \frac{2 * \text{duyarlılık} * \text{kesinlik}}{\text{duyarlılık}+\text{kesinlik}} \quad (4.7)$$

Başarı ölçtünde ROC eğrisi de True Positive ve False Positive değerlerinin kullanılarak hazırlandığı bir grafiktir. ROC değerinin 1'e yakınsaması model başarısının yüksek olduğunu 0'a yakınsaması modelin başarısız sonuçlandığını belirtir.

4.2. Veri Madenciliği Modelleri

Temel olarak veri madenciliğinde modellemeyi üç grupta toplayabiliriz. Veri madenciliğinde kullanılan uygulama modellleri Şekil 4.6'daki gibi gruplamak mümkündür.Sınıflandırma, kümeleme, birliktelik kuralı bu modeller arasında sıklıkla kullanıllan algoritmalar belirtilmiştir.



Şekil 4.6. Veri madenciliği modelleri

4.2.1. Sınıflandırma modeli

Sınıflandırma modeli veri madenciliğinde sıkça kullanılan bir tekniktir. Verilerin belli bir kategoride sınıflandırılabilmesi öncelikle veritabanında bulunan veriler analiz edilerek gelecek verilerin hangi kategoriye ait olacağı tespitidir.

Veri kümesinde bulunan bilgilerin hangi kategoriye ait olduğu belirlenir ve yeni verilerin hangi kategoriye ait olduğu sınıflandırılır. Bu süreçte eldeki veriler sınıflandırma algoritmalarıyla eğitime tabi tutulur. Sınıflandırma modelinde her sınıflandırma algoritması aynı oranda başarılı sonuç vermeyecektir. Veri kümesine göre seçilen algoritma başarı sonucu değişkenlik gösterir. Yeni gelen veriler sınama kümesi olarak değerlendirilip hangi kategoriye ait olduğu tespit edilir.

Sınıflandırma modelinde her algoritma seçimi farklı sonuçlar elde edilebilir, sınıflandırmada en başarılı sonuca ulaştıran algoritma seçimi yapılır. En sık kullanılan sınıflandırma algoritmaları aşağıdaki gibidir.

- Karar ağaçaları
- Bayes algoritmaları
- En yakın komşu algoritması
- Yapay Sinir Ağları
- Genetik Algoritmalar
- Karar Destek Makinaları

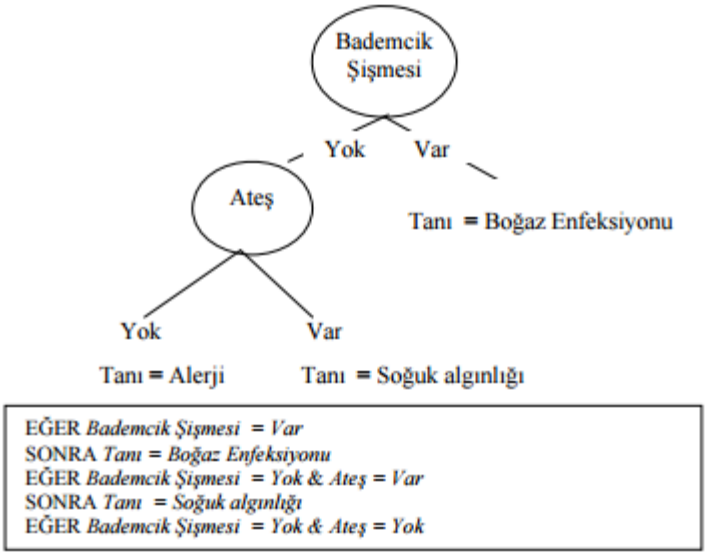
4.2.1.1. Karar ağaçları

Sınıflandırma kullanılan modelleme algoritmalarından karar ağacı yöntemi sıklıkla kullanılmaktadır. Karar ağaçlarının sık kullanılmasında en büyük etkenler;

- Maliyetinin düşük olması
- Yorumlanma kolaylığı ve veritabanına kolaylıkla entegre edilebilmesi
- Güvenilirliğinin iyi olması

Karar ağacı oluşturmak için eğitim kümesinde örnekleri en iyi ifade eden değişken bulunur. Seçilen dalın üstündeki örneklerden yeni bir değişken bulunur böylece yeni dallar oluşur. Örnekleri ayrıt edecek başka bir ayırtaç kalmamışsa dallanma bitmiş olur.

Ayırt edici olacak ağac kökü gibi hesaplamasıyla elde edilir. Değişkenlerden hangi nitelik en fazla dala sahipse kök olarak o nitelik seçilir. Entropi bilgi kazancı ölçümü için kullanılır. Entropi ileride beklenilmeyen bir olayın meydana gelme olasılığını bulur. Örneğin hastalık sınıflandırması tespitinde karar ağacı tespiti Şekil 4.7'deki gibi belirlenebilir (Akçetin, 2014).



Şekil 4.7. Karar ağacı tespiti (Taşkın, 2005)

İncelenen bir veri kümesinde bilgikazancı hesabı yapıldığında S'nin veri kümesi içerisinde Ci sınıfına bağlı olarak Si tane kaydı varsa sınıflandırmada aşağıdaki denklem kullanılır.

$$Bilgi(s_1, s_2, \dots, s_n) = -\sum_{i=1}^{n} \frac{s_i}{s} \left(\log_2 \left(\frac{s_i}{s}\right)\right) \tag{4.8.}$$

Değişken olarak kabul edilen A'nın değerleri (an) entropisinni tespiti için aşağdaki denklem kullanılır.

$$Entropi(A) = \sum_{j=1}^{v} \left(\frac{s_{1j} + \dots + s_{nj}}{s}\right) \cdot Bilgi(s_1, s_2, \dots, s_n) \tag{4.9.}$$

Karar ağaçları için sıklıklar kullanılan algoritmalar ID3, CART ve CHAID önemli örneklerdir.

4.2.1.2. Naive bayes algoritması

Sınıflandırma algoritmalarından Bayes Sınıflandırıcısı Şekil 4.9'daki gibi açıklanabilir. Navie Bayes modeli, Bayes modeline karar teorisine bağlı bir sınıflandırma yöntemidir. Niteliklerin birbirlerinden bağımsız ve eşit etki düzeyinde değerlendirilir. Bu yüzden anlaşılması kolay bir sınıflandırma algoritmadır.

Normal yaşamda nitelikler birbirleriyle bağlantılı olduğundan Navie Bayes eksik yanlarından biri olarak değerlendirilebilir.

X(x1,x2,x3…xn) oluşan bir veri örneği olsun. X'i sınıfı belli olmayan bir bir veri kümesi olarak kabul edebiliriz. Bu veri kümesinde n tane sınıfın var olduğunu kabul edelim. Bu şekilde C1, C2, C3, Cn şeklinde n tane sınıf olduğunu varsayalım.
Veri kümesinde bir işlemin sınıf değerini belirlemek için el alınan örnek:

$$P(C_i \mid X) = \frac{P(X|C_i)\,P(C_i)}{P(X)} \qquad (4.10.)$$

Örneğin olasılık değerleri hesaplanır. Xi proseslerinin ayrı ayrı olasılık değerlerinden aşağıdaki bağıntı kurulabilir. Böylelikle hesaplamada kolaylık sağlancaktır.

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k \mid C_i) \qquad (4.11.)$$

X'i sınıflandırmak adına P(Ci|X) için çıkan değerlerde paydalar eşit olduğunda sadece pay kısımıları karşılaştırılarak büyük olan değere göre sınıf seçimi yapılır.

$$\arg\max_{Ci} \{P(X \mid C_i)\,P(C_i)\} \qquad (4.12.)$$

Yukarıdaki çıkarımlara göre sonrasal olasılıklar incelenerek max olasılığa sahip sınıf seçildiğinden aşağıdaki durumla sonuca ulaşılabilir.

$$C_{MAP} = \underset{c}{\operatorname{argmax}} \prod_{k=1}^{n} P(X_k \mid C_i) \qquad (4.13.)$$

Navie Bayes uygulaması tez uygulamamızda yer aldığından örneklendirmek gerekirse;

Eğitim veri kümesi aşağaki gibi olan bir model ele alındığında Şekil 4.8'deki gibi bir sonuç tablosu elde edilir.

		KABUL			
		EVET		HAYIR	
Nitelikleri	Değeri	Sayısı	Olasılık	Sayısı	Olasılık
EĞİTİM	İLK	1	1/5	2	2/3
	ORTA	3	3/5	0	0
	YÜKSEK	1	1/5	1	1/3
YAŞ	GENÇ	0	0	2	2/3
	ORTA	3	3/5	1	1/3
	YAŞLI	2	2/5	0	0
CİNSİYET	ERKEK	3	3/5	1	1/3
	KADIN	2	2/5	2	2/3

Şekil 4.8. Bayes eğitim kümesi (Akçetin 2014)

Rastgele seçilen bir örneğin Navie Bayes Sınıflandırıcısı ile sınıfını belirlenirse:

X1: Eğitim = Yüksek

X2: Yaş = Orta

X3: Cinsiyet = Kadın

C Kabul = ?

Navie Bayes olasılıkları formülüzasyonu için verikümesifrekanslarına göre tekrar düzenlenirse Şekil 4.9'daki sonuç tablosu elde edilir.

		KABUL			
		EVET		HAYIR	
Nitelikleri	Değeri	Sayısı	Olasılık	Sayısı	Olasılık
EĞİTİM	İLK	1	1/5	2	2/3
	ORTA	3	3/5	0	0
	YÜKSEK	1	1/5	1	1/3
YAŞ	GENÇ	0	0	2	2/3
	ORTA	3	3/5	1	1/3
	YAŞLI	2	2/5	0	0
CİNSİYET	ERKEK	3	3/5	1	1/3
	KADIN	2	2/5	2	2/3

Şekil 4.9. Frekansa göre düzenlenmiş eğitim kümesi (Akçetin 2014)

Sınıf değerleri bulunmasında olasılık değerleri hesaplanmaya devam edilirse:

C1: Kabul = Evet ve C2 = Kabul = Hayır olarak iki sınıf değeri verikümesinden anlaşılmaktadır.

P(X|C1)P(C1), P(X|C1)P(C1) olasılıkları hesaplanıp max değere sahip olan ifade örneğin bize sınıfını gösterecektir.

P(X|C1)P(C1)değerinin hesaplanması:

P(X1|C1) = P(Eğitim = Yüksek | Kabul = Evet) = 1/5

P(X2|C1) = P(Yaş = Orta | Kabul = Evet) = 3/5

P(X3|C1) = P(Cinsiyet= Kadın| Kabul = Evet) = 2/5

Bu durumda;

P(X|C1) = P(X|Kabul = Evet) = (1/5) (3/5) (2/5) = 6/125 olarak bulunur.

P(C1) = P(Kabul = Evet) = 5/8

Formulde son durum yerine koyulursa:

P(X|C1) P(C1) = P(X|Kabul = Evet) P(Kabul = Evet) = (6/125) (5/8) = 0,03 olarak olasılık değeri bulunur.

P(X|C2)P(C2)değerinin hesaplanması:

P(X1|C2) = P(Eğitim = Yüksek | Kabul = Hayır) = 1/3

P(X2|C2) = P(Yaş = Orta | Kabul = Hayır) = 1/3

P(X3|C2) = P(Cinsiyet= Kadın| Kabul = Hayır) = 2/3

Bu durumda;
P(X|C2) = P(X|Kabul = Hayır) = (1/3) (1/3) (2/3) = 2/27 olarak bulunur.
P(C2) = P(Kabul = Hayır) = 3/8

Formulde son durum yerine koyulursa:
P(X|C2) P(C2) = P(X|Kabul = Hayır) P(Kabul = Hayır) = (2/27) (3/8) = 0,027 olarak olasılık deperi bulunur.

P(X|C1)P(C1) ve P(X|C2)P(C2) değerlerine göre sonuç:
argmax ci {P(X|Ci)P(Ci) } = max { 0,03, 0,027 } = 0,03

En büyük olasılık değeri 0,03 ün ve sınıfı "Evet" olduğunda rastgele seçilen:

X1: Eğitim = Yüksek
X2: Yaş = Orta
X3: Cinsiyet = Kadın

C Kabul = Evet

Sınıfı "Evet" olarak sınıfı belirlenmiş olur.

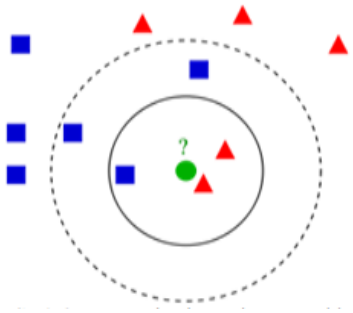
4.2.1.3. K en yakın komşu algoritması

Denetimli sınıflandırma algoritmalarından K en yakın komşu algoritmasında sınıflandırma işleminde kaç adet k değerine bakılarak sınıflandırma yapılacağına karar verilir. Yeni gelen değerin sınıflandrılımasında k değerine bakılarak o değere en yakın k tane değer karşılaştırılır ve en çok sınıfı olan değer değerin sınıfı olmuş olur (Özalp, 2013).

K en yakın nokta hesaplanırken çeşitli uzaklık hesaplama fonksiyonlarından yararlanılır. Aşağıdaki uzaklık hesaplama fonksiyonları başlıcalarıdır.

- Manhatten Hesaplaması
- Minkowski Hesaplaması
- Öklid Hesaplaması

Ağırlık belirleme yöntemi ile W = 1/(d*d) ile her bir uzaklığın ağırlığı bulunur ve seçilen k değerine göre sınıf karşılaştırması yapılır (Şekil 4.10).



Şekil 4.10. KNN model grafiği (Wikipedia, 2016b)

Örneklemek gerekirse, 10 tane örnek içeren bir veri kümesi Çizelge 4.1'deki gibidir.

Çizelge 4.1. KNN örnek veri kümesi

x1	x2	y
2	4	kötü
3	6	iyi
3	4	iyi
4	10	kötü
5	8	kötü
6	3	iyi
7	9	iyi
9	7	kötü
11	7	kötü
10	2	kötü

Sınıflandırılmak istenen değer: X1 = 9 ve X2 = 4 olarak incelersek, k en yakın algoritmasıyla k yı 4 olarak sınıflandırmaya çalışırsak öncelikle değerlerin öklid uzaklık değerleri Çizelge 4.2'deki gibi olacaktır.

Değerlerin hesaplanmış değerleriyle veri kümesi tekrar düzenlenirse:

Çizelge 4.2. Uzaklık değerleri hesaplanmış veri kümesi

X1	X2	Uzaklık	y
2	4	6.00	kötü
3	6	5.39	iyi
3	4	5.00	iyi
4	10	7.21	kötü
5	8	5.00	kötü
6	3	2.24	iyi
7	9	5.10	iyi
9	7	3.16	kötü
11	7	4.24	kötü
10	2	2.83	kötü

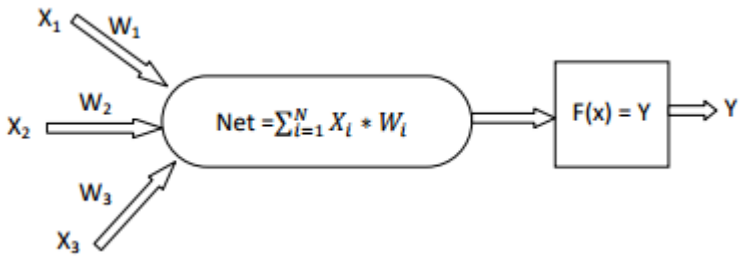
Veri kümesinde en yakın k=4 uzaklık değeri sınıfları kontrol edilirse: sınıf değerleri bir adet “iyi” ve üç tane ”kötü” sınıflandırma değeri olduğunda fazla olan sınıflandırma değeri “kötü” olarak örneğin sınıfı belirlenmiş olur.

4.2.1.4. Yapay sinir ağları

Yapay sinir ağlarıyla sınıflandırma yönetiminde insan beyni model alınarak oluşturulan bir tahminleme ve sınıflandırma algoritmasıdır. İnsan beyninde de Şekil 4.11’de olduğu gibi sinir hücreleri nöronlar yardımıyla belirli sinayalleri alır bir yorumlayıcı toplayıcı işlemden geçirilerek bir sonuç elde edilir (Cayiroğlu, 2016).

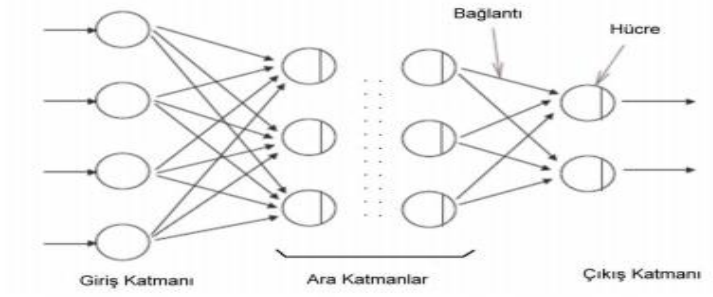
Yapay sinir hücresi beş bölümden oluşur:

Girdiler: İletilmek üzere gelen veriler olarak nitelendirilebilir.
Ağırlıklar: Girdiler üzerinden gelen değerlerin ağırlıkarı çıktılar üzerinde etkiyi belirler.
*Birleştirme:*Hücreye gelen verilerin ağırlıklarıyla çarğılarak oluşan toplam veriyi ifade eder.
Aktivasyon: Hücreye gelen net bilgi toplamını işleyerek bir çıktı işlemi belirlenir.



Şekil 4.11. Yapay sinir ağları (Cayiroğlu, 2016)

Sinir hücrelerinin birbirine bağlanmasıyla oluşan yapay sinir ağları üç katmandan oluşur (Şekil 4.12).



Şekil 4.12 Yapay sinir ağları katmanları (Cayiroğlu, 2016)

Giriş katmanı: Herhangi bir işleme tabi tutulmadan orta katmana iletilen sinir hücrelerinden gelen bilgiler bu katmanda yer alırlar.

Ara katman: Ara katmanda bağlantı sayısı giriş ve çıkış katmanındaki veri sayısından bağımsızdır. Bu katmanda artan veri fazlalığı uygulama performansını, çıktı süresini azaltacaktır. Giriş katmanından çıkan veriler direkt bu katmana iletilir. Ara katmanının sayısı değişkenlik gösterebilir.

Çıkış katmanı: Ara katmandan işlenerek gelen verilerin sunulduğu katmandır. Bu katmanda yeni ağlarda kullanılmak üzere ağırlık değerleri hesaplanır.

4.2.1.5. Genetik algoritmalar

Genetik algoritmalar veri madenciliğinde doğal seleksiyon yönetmini temel alarak kullanılan bir yöntemdir. Bu algoritmayla başlangıçta yer alan bir nesil, mutasyon ve çaprazlama yöntemlerinden geçirilerek doğal seçilim sağlanır (Kılıçaraslan, 2013).

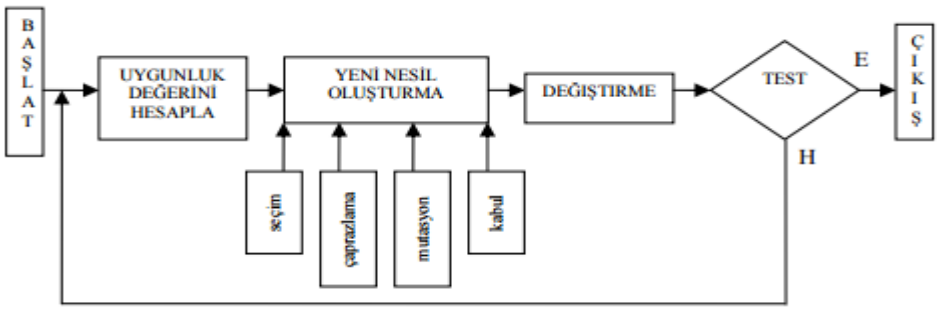
Genetik algoritmalar genel prensipleriyle Şekil 4.13.'te gösterilmiştir. Probleme göre rastgele seçilen n kromozoma sahip bir populasyon ele alınır ve başlangıç fonksyionuna sokulur. Sonrasında seçilen her bir kromozom uygunluk fonksiyonundan geçirilerek yeni populasyon elde edilir.

Seleksiyon: Rastgele seçilen iki kromozom karşılaştırılarak baskın karakterler göz önünde tutularak yeni bir seçim yapılır.

Çaprazlama: Yeni bireyler oluşturmak amacıyla iki kromozom genleri çaprazlanır. Böylece ebeveynlerden farklı bireyler oluşur.

*Mutasyon:*Genlerde belli oranlarda manipulasyon yapılır. Böylece daha farklı kromozomlar elde edilir.

Algoritmaya göre yeni populasyona ulaştıktan sonra, yeni populasyon ile yer değiştirilir.



Şekil 4.13. Genetik algorimaları (Bolat 2004)

Genetik algoritmalarda kromozom sayısı artışı algoritmanın hızını azaltır ve maliyetinin azaltır. Tersi durumda ise kromozom çeşitliliğinde azalmaya yol açar. Kromozomların çeşitliliğinin azaldığı durumlarda mutasyona başvurmak yararlı olacaktır.

Genetik algoritmalarda güçlü bir algoritma omasında arama uzayının geniş olmasında etkisi büyüktür. Çok yönlü ve amaçlı problemlerin çözümünde bir tercih sebebidir ve kısa sürede başarılı sonuçlar verirler.

Son kullanıcının oluşturulan genetik algoritma modelini anlmasının zor olması dezavantajlarındandır. Uygunluk fonksiyonu belirlemek, mutasyon ve çaprazlama fonskyionlarını belirlemekmaliyetlidir (Bolat 2004).

5. SAHTECİLİK TANIMI VE SAHTECİLİĞİ ÖNLEME YÖNTEMLERİ

Fraud kelime anlamı itibariyle sahtecilik anlamına gelmektedir. Elektronik ticaret sitelerinde ise sahtecilik yöntemi kredi kartları ile oluşmaktadır. Çalıntı kredi kartlarıyla ya da kredi kartı bilgilerinin kopyalanmasıyla Internet üzerinden alış veriş sitelerinde sahtecilik riski oluşturmaktadır. Bu dolandırıcılık işlemleri sonucu ile kart sahibi, iş yeri ve banka ciddi zararlar görebilmektedir.

5.1. Sanal ve Gerçek Ortamda Sahtecilik

Gerçek fiziki ortamlarda gerçekleştirilen sahtecilik olaylarında banka kartının banka kartı şeridinde yer alan bilginin kopyalanması veya kart çip bilgilerinin kopyalanması ile mümkün olmaktadır. Yapılan sahtecilik işlemleri genellikle temassız işlemlerde veya manyettik şeritle yapılan işlemlerde sıkça rastlanmaktadır.

Fiziksel ortamlarda gerçekleştirilen sahtecilik uygulamalarına genellikle ATM'ler, mağazalar, kuyumcular veya telefon bayileri gibi hassas noktalar hedef alınmaktadır. Bu şekilde yapılan sahtecilik içeren işlemlerden korunabilmek için kart çipi okutularak ve şifre girişi yapılarak yapılan ödeme işlemlerinde sahtecilik riski en aza indirmek mümkündür.

Sanal ortamlarda sahtecilik için en sık hedef olan ortamlar Internet üzerindensatış siteleri içerisinde uçak veya otobus bilet satışı için kullanılan siteler, online uygulama indirme siteleri (iTunes, Google Play), online telefon fatura ödeme işlemi yapan siteler ve online alışveriş yapılan her türlü ödeme sistemi içinsahtecilik mevcuttur.

Herhangi bir işyeri için sahtecilik puanı aşağıdaki şekilde hesaplanır:

Toplam (sahtecilik tutarı / toplam ciro) * 10000

Bu orana base point denir. Örnek olarak bir şirketin aylık sahte sipariş tutarı 100 tl ve cirosu 100000 tl olarak ele alınırsa, sahtecilikpuanı:

(100/100000)*10000 = 0,001 olarak bulunur.

Bir işyeri için maksimum sahtecilik puan standardı ise 0,004'tür, bu oran avrupa bölgesinde ortalama 0,008 ve Türkiye'de ise ortalama sahtecilik puanı 0,002 dir.

2012 yılı verileri incelendiğinde Kanada'da elektronik alış veriş sitelerinde sahtecilik işlemleri nedeniyle üç buçuk milyar dolar zarar meydana gelmiştir ve bu işyerleri için sahtecilik kontrolünün ne denli önemli olduğunu göstermektedir (Webrazzi, 2013).

5.2. Sanal Ticarette Sahtecilik Önleme Yöntemleri

Sahtecilik kontrolleriyle biralış veriş sitesinde sahtecilik işlemlerini azaltmak mümkündür. Bu kontroller site üzerinden verilen bir siparişin karakteristik özellikleri incelenerek verilen karar mekanızmalarıdır.

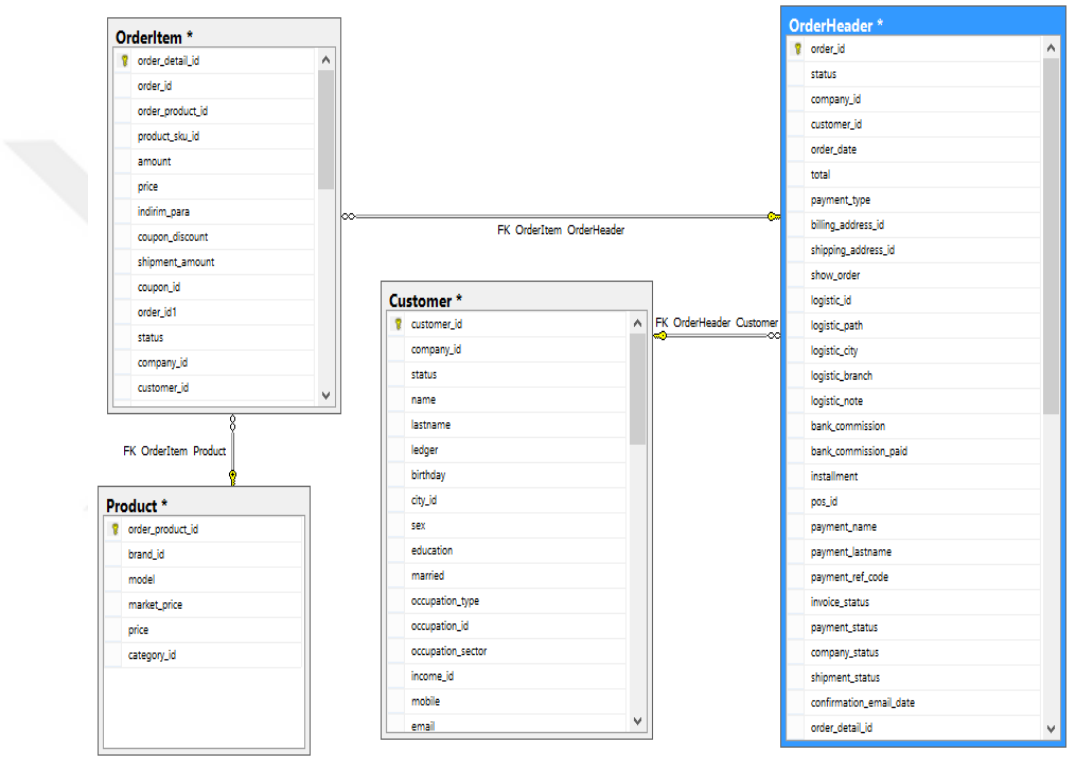
Sahtecilikkontrolleri, bir siparişin incelenmesi ve önlenmesinde belli başlı yöntemlerdir (Fraudandchargeback, 2010).

- Bir e-ticaret sitesinden ilk kez alışveriş yapan bir müşterinin siparişi kontrol için önem taşır. Bu müşteriler için 3d ödeme kontrolü bir önlem olarak alınması.
- Site için müşterilerin yaptığı ortalama sipariş fiyatı üzeri yapılan siparişler kontrole alınması.
- Bir ürünün birden fazla adet sipariş verilmesi olağan durum dışında değerlendirilir, bu yüzden bu kontrol de ele alınır.
- Hızlı teslimat içeren ürünler sahtekârlık unsuruna açık olacağından dağıtım öncesi kullanıcıyla görüşmek riski aza indirgeyecektir.
- Anlamsız veya olağan dışı email adresleri sahtekârlık niteliği taşıdığından kontrole alınması gereklidir.
- Aynı kart ile en ucuz fiyatlı ürünler alınarak kart çekimi test edildiği çıkarımı yapılabilir.

6. E-TİCARET SİTESİ SİPARİŞ VERİLERİ ANALİZİ

6.1. Verilerin Analiz Süreci

E-ticaret veri kümesihazırlanmadan önce tabloların ilişkisel diagramı ortaya çıkarılmıştır. Tablo verileri Ms Sql Server veritabanı kullanılarak incelemeye alınmış ve oluşturulan diagramdaki tablolarınilişkisel durumu ele alınmıştır (Şekil 6.1).



Şekil 6.1. Site veritabanı yapısı

OrderHeader: Siparişlerin genel bilgilerinin tutulduğu tablodur.

OrderItem: Siparişlerin detaylarının tutulduğu tablodur.

Product: Sitede ürünlerin ve kategorilerinin tutulduğu tablodur.

Customer: Site üzerinde müşteri bilgilerinin tutulduğu tablodur.

6.2. Değişkenlerin Sınıflandırma Üzerinde Etkisinin Belirlenmesi

Oluşturulan modeldeki verilerin sahtecilik sınıflandırmasında etkisinin bulunabilmesi için Weka platformunda yer alan algoritmalardan yararlanılmıştır. Sırasıyla GainRatio, ChiSquared ve InfoGain algoritmaları kullanılarak değişkenlerin sınıflandırma üzerine etkileri sınanmıştır (Çizelge 6.1).

Çizelge 6.1. Değişkenlerin sınıflandırma üzerine etkileri

Sınıflamayı etkileyen değişken isimleri	Anlam	Sipariş verilerinin sınıflandırma üzerine etki sonuçları		
		GainRatio	ChiSquared	InfoGain
Total	Sipariş toplam tutarı	2	1	1
Payment_ref_code	Banka cevap kodu	3	5	4
Amount	Ürün fiyatı	9	12	13
OrderHour	Sipariş Saati 1-24 sınıflandırma	13	9	9
OrderDayOfWeek	Sipariş günü 1-7 sınıflandırma	14	10	11
NameSurnameLen	Müşteri ad soyad uzunluğu	1	2	2
Discount_money	İndirim tutarı	15	15	15
Coupon_Discount	Kupon indirim tutarı	16	16	16
Shipped_Amount	Kargo ücreti	8	11	10
CouponID	Kupon Id	17	17	17
EmailConfirmTime	E-mail doğrulama 1-24 sınıflama	11	14	14
CustomerCityID	Şehir Kodu	5	4	6
CustomerEmailFormat	E-formatı, uzantısına gore sınıflama	7	7	7
OrderBrandID	Ürün marka id	6	6	5
CategoryID	Ürün kategori id	4	3	3
CustomerAge	Müşteri yaş	10	8	8
Gender	Müşteri cinsiyet	12	12	12
IsFraud	Sahtecilik kontrolü 1-0 sınıflandırma	Class Attribute	Class Attribute	Class Attribute

Bu çalışmada kullanılan ilgili e-ticaret sitesinden 1615 adet sipariş incelenmiştir. Modelleme için sahtecilik niteliği oluşturulan değişken listesi Çizelge 6.2'de listelenmiş ve sınıflandırma üzerinde etkisi Çizelge 6.1'de belirtilmiştir.

Çizelge 6.2. Değişken listesi

1	Sipariş Toplam Tutarı	10	E-mail Doğrulma Zamanı
2	Banka Cevap Kodu	11	Müşteri İl Id
3	Sipariş Tutarı	12	Müşteri Email Format Tipi
4	Sipariş Saati	13	Sipariş Marka Id
5	Siparişin Hangi Gün Verildiği	14	Sipariş Kategori Id
6	Müşteri Ad Soyad Uzunluğu	15	Kargo ücreti
7	Sipariş İndirim Tutarı	16	Müşteri Yaşı
8	Sipariş Kupon İndirimi	17	Müşteri Cinsiyeti
9	Kupon Id	18	Sahtecilik kontrol Id

Veri madenciliğinde anlamlı veriye ulaşmak için kullanılan adımlar takip edilerek en performanslı ve sonuca ulaştırabilecek veri kümesi oluşturma aşağıdaki adımlar takip edilmiştir.

- *Veri Temizleme*
- *Veri Bütünleştirme*
- *Veri seçme*
- *Veri Dönüştürme*

Sahtecilik oluşturabilecek veri seçiminin ardından modellemede daha iyi performans alabilmek ve verileri sınıflamada geniş bir sınıflandırma kümesiyle uğraşmak yerine normalize edilmiş bir değer aralığıyla algoritmaları test etmek daha iyi sonuca ulaşmamızda etkili olacaktır. Bu nedenle seçilen değişkenlere göre normalize edilmiş değer karşılık tablosu aşağıda belirtimiştir.

Müşteri email uzantıları Çizelge 6.3'teki gibi karşılık değerleriyle normalize edilmiştir.

Çizelge 6.3. E-Posta normalizasyon tablosu

customerEmailExtension	typeNum	customerEmailExtension	typeNum	customerEmailExtension	typeNum
aerodeon.com	1	link.com.tr	21	turksat.com.tr	41
anadolu.edu.tr	2	mac.com	22	uekae.tubitak.gov.tr	42
baxter.com	3	msn.com	23	vakifbank.com.tr	43
besiktas.ws	4	my.net.com	24	vodafone.com	44
cimri.com	5	mynet.com	25	windowslive.com	45
colpal.com	6	otmail.com	26	yahoo.co.uk	46
creareklam.com	7	redoksas.com	27	yahoo.com	47
cvsship.com	8	renault.com	28	yahoo.com.tr	48
dhl.com	9	rocketmail.com	29	yandex.ru	49
ekolay.net	10	sefine.com.tr	30	yapikredi.com.tr	50
ersell.com	11	sisecam.com	31	zorlu.com	51
garanti.com.tr	12	somacons.com	32		
garantiemeklilik.com.tr	13	spk.gov.tr	33		
gechit.com	14	superonline.com	34		
gmail.com	15	superposta.com	35		
groupama.com.tr	16	teb.com.tr	36		
hotmail.com	17	temasltd.com	37		
hotmail.com.tr	18	thy.com	38		
isbank.net.tr	19	tr.petronas.com	39		
istanbul.com	20	ttef.net	40		

Müşterilerin sipariş günleri Çizelge 6.4'te normalize edilmiş değerleriye belirtilmiştir.

Çizelge 6.4. Sipariş günü normalizasyon tablosu

orderDayOfWeek	typeNum
Pazar	1
Pazartesi	2
Salı	3
Çarşamba	4
Perşembe	5
Cuma	6
Cumartesi	7

Müşterilerin e-mail adreslerini doğrulama yapıp yapmaması da Çizelge 6.5'te gruplandırılmıştır.

Çizelge 6.5. E-Posta doğrulama normalizasyon tablosu

emailConfirmTimeType	typeNum
E-Posta Doğrulanmamış	0
E-Posta Doğrulanmış	1

Müşterilerin sipariş saatleri 1'den 24'e kadar saat aralığında değerlendirmeye alınmıştır. Aynı şekilde müşterilerin email doğrulama saatleri de normalize edilerek modele eklenmiştir.

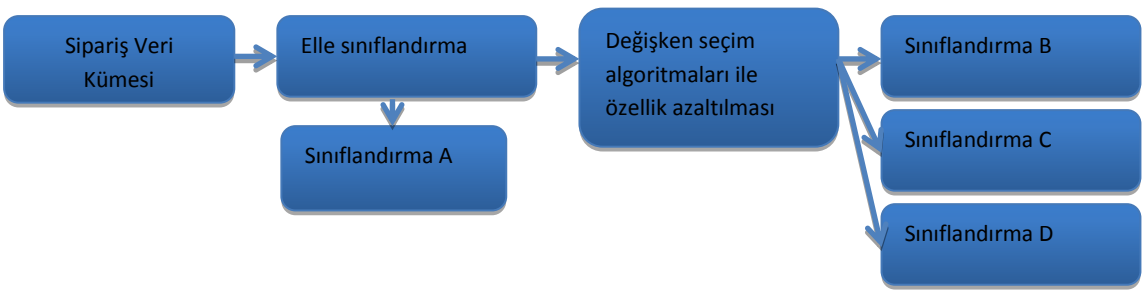
E-ticaret firaması tarafında bankadan dönen cevap kodlarına göre ve elle uygulanansüreçlere göre karar verilen "sahtecilik" kolonu da sahtecilik veya sahtecilik değil olarak normalize edilmiştir (Çizelge 6.6).

Çizelge 6.6. Sipariş sahtecilik normalizyon tablosu

isFraud	typeNum
Sahtecilik içerir	1
Sahtecilik içermez	0

6.3. Model Üzerinde Sınıflandırma Metodlarının Uygulanması

E-ticaret sitesi üzerinde sahtecilik belirleme işleminde öncelikle firmanın operasyon tarafında bankayla ve müşteriyle iletişimi sonucu elde edilmiş sahtecilikbilgileri elde edilmiştir. Veri madencliği için sipariş verileri daha önce anlatılan veri temizleme, veri bütünleştirme, değişken seçimi ve veri normalizasyonu aşamalarından sonra sınıflandırma algoritmalarına tabi tutulmuştur (Şekil 6.2).



Şekil 6.2. Uygulama modeli aşamaları

7. WEKA'DA ALGORİTMALARIN UYGULANMASI

7.1. Navie Bayes Uygulanması

Navie Bayes uygulamasında sırasıyla bütün değişkenler, Weka CfsSubSetEval, Weka ConsistensySubsetEval veWeka FilteredSubsetEval algoritmalarına uygulanmış en başarılı değişkenler algoritmada sınanım sonuçlarında değerlendirilmiştir.

- *Bütün Değişkenler Kullanılarak Algoritmada Uygulanması*

Veri kümesindeki tüm değişkenler kullanılarak ulaşılan algoritma sonuçlarıdır (Çizelge 7.1, 7.2, 7.3).

Çizelge 7.1. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 18 (All attributes)	

Çizelge 7.2. Navie Bayes uygulaması istatistik sonuçları

Correctly Classified Instances	446	92,1488%
Incorrectly Classified Instances	38	7,8512%
Kappa statistic	0.5351	-
Mean absolute error	0.0851	-
Root mean squared error	0.2416	-
Relative absolute error	549802%	-
Root relative squared error	877314%	-
Total Number of Instances	484	-

Çizelge 7.3. Navie Bayes uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.946	0.35	0.968	0.946	0.957	0.956	0
0.65	0.054	0.52	0.65	0.578	0.956	1
Weighted Avg.	0.921	0.326	0.931	0.921	0.925	0.956

- *Weka CfsSubSetEvalDeğişken Seçim Algoritmasının Uygulanması*

Veri kümesindeki değişkenlerin CfsSubSetEval algoritması sonucu çıkan en başarılı değişken seçim sonuçlarıdır (Çizelge 7.4, 7.5, 7.6).

Çizelge 7.4. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 4 (All attributes)	

Çizelge 7.5. Navie Bayes uygulaması istatistik sonuçları:

Correctly Classified Instances	459	94.8347 %
Incorrectly Classified Instances	25	5.1653 %
Kappa statistic	0.5524	
Mean absolute error	0.087	
Root mean squared error	0.1944	
Relative absolute error	56.1547%	
Root relative squared error	70.5833%	
Total Number of Instances	484	

Çizelge 7.6. Navie Bayes uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.995	0.575	0.951	0.995	0.972	0.963	0
0.425	0.005	0.895	0.425	0.576	0.963	1
Weighted Avg.	0.948	0.528	0.946	0.948	0.94	0.963

- *Weka ConsistensySubsetEval Değişken Seçim Algoritmasının Uygulanması*

Veri kümesindeki değişkenlerin ConsistensySubsetEval algoritması sonucu çıkan en başarılı değişken seçim sonuçlarıdır (Çizelge 7.7, 7.8, 7.9).

Çizelge 7.7. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 4 (All attributes)	

Çizelge 7.8. Navie Bayes uygulaması istatistik sonuçları

Correctly Classified Instances	445	91.9421 %
Incorrectly Classified Instances	39	8.0579 %
Kappa statistic	0.3593	
Mean absolute error	0.1028	
Root mean squared error	0.2397	
Relative absolute error	66.3973 %	
Root relative squared error	87.0523 %	
Total Number of Instances	484	

Çizelge 7.9. Navie Bayes uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.973	0.675	0.941	0.973	0.957	0.926	0
0.325	0.027	0.52	0.325	0.4	0.926	1
Weighted Avg.	0.919	0.621	0.906	0.919	0.911	0.926

- *Weka FilteredSubsetEval Değişken Seçim Algoritmasının Uygulanması*

Veri kümesindeki değişkenlerin FilteredSubsetEval algoritması sonucu çıkan en başarılı değişken seçim sonuçlarıdır (Çizelge 7.10, 7.11, 7.12).

Çizelge 7.10. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 3 (All attributes)	

Çizelge 7.11. Navie Bayes uygulaması istatistik sonuçları

Correctly Classified Instances	457	94.4215 %
Incorrectly Classified Instances	27	5.5785 %
Kappa statistic	0.5018	-
Mean absolute error	0.0894	-
Root mean squared error	0.1998	-
Relative absolute error	57.7142 %	-
Root relative squared error	72.5646 %	-
Total Number of Instances	484	-

Çizelge 7.12. Navie Bayes uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.995	0.625	0.946	0.995	0.97	0.961	0
0.375	0.005	0.882	0.375	0.526	0.961	1
Weighted Avg.	0.944	0.574	0.941	0.944	0.934	0.961

7.2. RBF Network Uygulanması

RBF Networkuygulamasında sırasıyla bütün değişkenler, Weka CfsSubSetEval, Weka ConsistensySubsetEval ve Weka FilteredSubsetEval algoritmalarına uygulanmış en başarılı değişkenler algoritmada sınanım sonuçlarında değerlendirilmiştir.

- *Bütün Değişkenler Kullanılarak Algoritmada Uygulanması*

Veri kümesindeki tüm değişkenlerin algoritması sonucu çıkan en başarılı değişken seçim sonuçlarıdır (Çizelge 7.13, 7.14, 7.15).

Çizelge 7.13. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 18 (All attributes)	

Çizelge 7.14. RBF Network uygulaması istatistik sonuçları

Correctly Classified Instances	449	92.7686 %
Incorrectly Classified Instances	35	7.2314 %
Kappa statistic	0.4943	-
Mean absolute error	0.087	-
Root mean squared error	0.2433	-
Relative absolute error	56.1677 %	-
Root relative squared error	88.3558 %	-
Total Number of Instances	484	-

Çizelge 7.15. RBF Network uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.966	0.5	0.955	0.966	0.961	0.926	0
0.5	0.034	0.571	0.5	0.533	0.926	1
Weighted Avg.	0.928	0.461	0.924	0.928	0.925	0.926

- *Weka CfsSubSetEval Değişken Seçim Algoritmasının Uygulanması*

Veri kümesindeki değişkenlerin CfsSubSetEval algoritması sonucu çıkan en başarılı değişken seçim sonuçlarıdır (Çizelge 7.16, 7.17, 7.18).

Çizelge 7.16. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 4	

Çizelge 7.17. RBF Network uygulaması istatistik sonuçları

Correctly Classified Instances	459	94.8347 %
Incorrectly Classified Instances	25	5.1653 %
Kappa statistic	0.6473	-
Mean absolute error	0.0676	-
Root mean squared error	0.2087	-
Relative absolute error	43.6668 %	-
Root relative squared error	75.8057 %	-
Total Number of Instances	484	-

Çizelge 7.18. RBF Network uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.975	0.35	0.969	0.975	0.972	0.954	0
0.65	0.025	0.703	0.65	0.675	0.954	1
Weighted Avg.	. 0.948	0.323	0.947	0.948	0.947	0.954

- *Weka ConsistensySubsetEval Değişken Seçim Algoritmasının Uygulanması*

ConsistensySubsetEval sonucu başarılı seçim sonuçlarıdır (Çizelge 7.19, 7.20, 7.21).

Çizelge 7.19. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 4	

Çizelge 7.20. RBF Network uygulaması istatistik sonuçları

Correctly Classified Instances	452	93.3884 %
Incorrectly Classified Instances	32	6.6116 %
Kappa statistic	0.5432	-
Mean absolute error	0.0808	-
Root mean squared error	0.2259	-
Relative absolute error	52.1645 %	-
Root relative squared error	82.0471 %	-
Total Number of Instances	484	-

Çizelge 7.21. RBF Network uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.968	0.45	0.96	0.968	0.964	0.932	0
0.55	0.032	0.611	0.55	0.579	0.932	1
Weighted Avg.	0.934	0.415	0.931	0.934	0.932	0.932

- *FilteredSubsetEval Değişken Seçim Algoritmasının Uygulanması*

Veri kümesindeki değişkenlerin FilteredSubsetEval algoritması sonucu çıkan en başarılı değişken seçim sonuçlarıdır (Çizelge 7.22, 7.23, 7.24).

Çizelge 7.22. Veri kümesi özellikleri

Instances: 1615	Test mode: split 70.0% train, remainder test
Attributes: 3	

Çizelge 7.23. RBF Network uygulaması istatistik sonuçları

Correctly Classified Instances	457	94.4215 %
Incorrectly Classified Instances	27	5.5785 %
Kappa statistic	0.6099	-
Mean absolute error	0.0706	-
Root mean squared error	0.2139	-
Relative absolute error	45.6029 %	-
Root relative squared error	77.6701 %	-
Total Number of Instances	484	-

Çizelge 7.24. RBF Network uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.975	0.4	0.964	0.975	0.97	0.952	0
0.6	0.025	0.686	0.6	0.64	0.952	1
Weighted Avg.	0.944	0.369	0.941	0.944	0.943	0.952

7.3. IBK Uygulanması

IBK uygulamasında sırasıyla bütün değişkenler, Weka CfsSubSetEval, Weka ConsistensySubsetEval ve Weka FilteredSubsetEval algoritmalarına uygulanmış en başarılı değişkenler algoritmada sınanım sonuçlarında değerlendirilmiştir.

Veri kümesindeki tüm değişkenlerin kullanımı sonucu çıkan en başarılı değişken seçim sonuçlarıdır (Çizelge 7.25, 7.26, 7.27).

Çizelge 7.25. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 18	

Çizelge 7.26. IBK uygulaması istatistik sonuçları

Correctly Classified Instances	455	94.0083 %
Incorrectly Classified Instances	29	5.9917 %
Kappa statistic	0.5909	-
Mean absolute error	0.064	-
Root mean squared error	0.2293	-
Relative absolute error	41.3355 %	-
Root relative squared error	83.2751 %	-
Total Number of Instances	484	-

Çizelge 7.27. IBK uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.971	0.4	0.964	0.971	0.967	0.841	0
0.6	0.029	0.649	0.6	0.623	0.841	1
Weighted Avg.	0.94	0.369	0.938	0.94	0.939	0.841

- *Weka CfsSubSetEval Değişken Seçim Algoritmasının Uygulanması*

CfsSubSetEval algoritması sonucu seçim sonuçlarıdır (Çizelge 7.28, 7.29, 7.30).

Çizelge 7.28. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 4	

Çizelge 7.29. IBK uygulaması istatistik sonuçları

Correctly Classified Instances	460	95.0413 %
Incorrectly Classified Instances	24	4.9587 %
Kappa statistic	0.6002	-
Mean absolute error	0.0688	-
Root mean squared error	0.214	-
Relative absolute error	44.4338 %	-
Root relative squared error	77.727 %	-
Total Number of Instances	484	-

Çizelge 7.30. IBK uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.991	0.5	0.957	0.991	0.973	0.912	0
0.5	0.009	0.833	0.5	0.625	0.912	1
Weighted Avg.	0.95	0.459	0.946	0.95	0.945	0.912

- *Weka ConsistensySubsetEval Değişken Seçim Algoritmasının Uygulanması*

CfsSubSetEval algoritması sonucu seçim sonuçlarıdır (Çizelge 7.31, 7.32, 7.33).

Çizelge 7.31. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 4	

Çizelge 7.32. IBK uygulaması istatistik sonuçları

Correctly Classified Instances	464	95.8678 %
Incorrectly Classified Instances	20	4.1322 %
Kappa statistic	0.6759	-
Mean absolute error	0.0682	-
Root mean squared error	0.1995	-
Relative absolute error	44.0454 %	-
Root relative squared error	72.4562 %	-
Total Number of Instances	484	-

Çizelge 7.33. IBK uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.993	0.425	0.963	0.993	0.978	0.943	0
0.575	0.007	0.885	0.575	0.697	0.943	1
Weighted Avg.	0.959	0.39	0.956	0.959	0.955	0.943

- *FilteredSubsetEval Değişken Seçim Algoritmasının Uygulanması*

CfsSubSetEval algoritması sonucu seçim sonuçlarıdır (Çizelge 7.34, 7.35, 7.36).

Çizelge 7.34. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 3	

Çizelge 7.35. IBK uygulaması istatistik sonuçları

Correctly Classified Instances	460	95.0413 %
Incorrectly Classified Instances	24	4.9587 %
Kappa statistic	0.6002	-
Mean absolute error	0.0688	-
Root mean squared error	0.214	-
Relative absolute error	44.4338 %	-
Root relative squared error	77.727 %	-
Total Number of Instances	484	-

Çizelge 7.36. IBK uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.991	0.5	0.957	0.991	0.973	0.912	0
0.5	0.009	0.833	0.5	0.625	0.912	1
Weighted Avg.	0.95	0.459	0.946	0.95	0.945	0.912

7.4. NBTree Uygulanması

NBTree uygulamasında sırasıyla bütün değişkenler, Weka CfsSubSetEval, Weka ConsistensySubsetEval veWeka FilteredSubsetEval algoritmalarına uygulanmış en başarılı değişkenler algoritmada sınanım sonuçlarında değerlendirilmiştir.

- *Bütün Değişkenler Kullanılarak Algoritmada Uygulanması*

Tüm değişkenleri kullanımına bağlı sonuçlardır (Çizelge 7.37, 7.38, 7.39).

Çizelge 7.37. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 4	

Çizelge 7.38. NBTree uygulaması istatistik sonuçları

Correctly Classified Instances	459	94.8347 %
Incorrectly Classified Instances	25	5.1653 %
Kappa statistic	0.5387	-
Mean absolute error	0.0544	-
Root mean squared error	0.2269	-
Relative absolute error	35.1089 %	-
Root relative squared error	82.3806 %	-
Total Number of Instances	484	-

Çizelge 7.39. NBTree uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.998	0.6	0.949	0.998	0.978	0.828	0
0.4	0.002	0.941	0.4	0.697	0.828	1
Weighted Avg.	0.948	0.551	0.948	0.959	0.939	0.828

Veri kümesindeki değişkenlerin CfsSubSetEval algoritması sonucu çıkan en başarılı değişken seçim sonuçlarıdır (Çizelge 7.40, 7.41, 7.42).

Çizelge 7.40. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 4	

Çizelge 7.41. NBTree uygulaması istatistik sonuçları

Correctly Classified Instances	459	94.8347 %
Incorrectly Classified Instances	25	5.1653 %
Kappa statistic	0.5524	-
Mean absolute error	0.087	-
Root mean squared error	0.1944	-
Relative absolute error	56.1547 %	-
Root relative squared error	70.5833 %	-
Total Number of Instances	484	-

Çizelge 7.42. NBTree uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.995	0.575	0.951	0.995	0.972	0.963	0
0.425	0.005	0.895	0.425	0.576	0.963	1
Weighted Avg.	0.948	0.528	0.946	0.948	0.94	0.963

ConsistensySubsetEval algoritması başarım sonuçlarıdır (Çizelge 7.43, 7.44, 7.45).

Çizelge 7.43. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 4	

Çizelge 7.44. NBTree uygulaması istatistik sonuçları

Correctly Classified Instances	457	94.4215 %
Incorrectly Classified Instances	27	5.5785 %
Kappa statistic	0.469	-
Mean absolute error	0.0632	-
Root mean squared error	0.2334	-
Relative absolute error	40.8044 %	-
Root relative squared error	84.7524 %	-
Total Number of Instances	484	-

Çizelge 7.45. NBTree uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.675	0.943	0.97	0.794	0.963	0
0.325	0	1	0.491	0.794	0.963	1
Weighted Avg.	0.944	0.619	0.944	0.948	0.931	0.794

- *FilteredSubsetEval Değişken Seçim Algoritmasının Uygulanması*

FilteredSubsetEval algoritması sonucu seçim sonuçlarıdır (Çizelge 7.46, 7.47, 7.48).

Çizelge 7.46. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 3	

Çizelge 7.47. NBTree uygulaması istatistik sonuçları

Correctly Classified Instances	457	94.4215 %
Incorrectly Classified Instances	27	5.5785 %
Kappa statistic	0.5018	-
Mean absolute error	0.0894	-
Root mean squared error	0.1998	-
Relative absolute error	57.7142 %	-
Root relative squared error	72.5646 %	-
Total Number of Instances	484	-

Çizelge 7.48. NBTree uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.995	0.625	0.946	0.995	0.97	0.961	0
0.375	0.005	0.882	0.375	0.526	0.961	1
Weighted Avg.	0.944	0.574	0.941	0.944	0.934	0.961

7.5. J48 Uygulanması

J48 uygulamasında sırasıyla bütün değişkenler, Weka CfsSubSetEval, Weka ConsistensySubsetEval veWeka FilteredSubsetEval algoritmalarına uygulanmış en başarılı değişkenler algoritmada sınanım sonuçlarında değerlendirilmiştir.

- *Bütün Değişkenler Kullanılarak Algoritmada Uygulanması*

Tüm değişkenleri kullanımına bağlı sonuçlardır (Çizelge 7.49, 7.50, 7.51).

Çizelge 7.49. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 3	

Çizelge 7.50. J48 uygulaması istatistik sonuçları

Correctly Classified Instances	457	94.4215 %
Incorrectly Classified Instances	27	5.5785 %
Kappa statistic	0.469	-
Mean absolute error	0.1026	-
Root mean squared error	0.2268	-
Relative absolute error	66.2416 %	-
Root relative squared error	82.3448 %	-
Total Number of Instances	484	-

Çizelge 7.51. J48 uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.675	0.943	1	0.97	0.755	0
0.325	0	1	0.325	0.491	0.755	1
Weighted Avg.	0.944	0.619	0.947	0.944	0.931	0.755

- *Weka CfsSubSetEval Değişken Seçim Algoritmasının Uygulanması*

CfsSubSetEval algoritması sonucu seçim sonuçlarıdır (Çizelge 7.52, 7.53, 7.54).

Çizelge 7.52. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 4	

Çizelge 7.53. J48 uygulaması istatistik sonuçları

Correctly Classified Instances	444	91.7355 %
Incorrectly Classified Instances	40	8.2645 %
Kappa statistic	0	-
Mean absolute error	0.1542	-
Root mean squared error	0.2754	-
Relative absolute error	99.6058 %	-
Root relative squared error	99.9966 %	-
Total Number of Instances	484	-

Çizelge 7.54. J48 uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	1	0.917	1	0.957	0.5	0
0	0	0	0	0	0.5	1
Weighted Avg.	0.917	0.917	0.842	0.917	0.878	0.5

- *Weka ConsistensySubsetEval Değişken Seçim Algoritmasının Uygulanması*

ConsistensySubsetEval algoritması sonucu seçim sonuçlarıdır (Çizelge 7.55, 7.56, 7.57).

Çizelge 7.55. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 4	

Çizelge 7.56. J48 uygulaması istatistik sonuçları

Correctly Classified Instances	444	91.7355 %
Incorrectly Classified Instances	40	8.2645 %
Kappa statistic	0	-
Mean absolute error	0.1542	-
Root mean squared error	0.2754	-
Relative absolute error	99.6058 %	-
Root relative squared error	99.9966 %	-
Total Number of Instances	484	-

Çizelge 7.57. J48 uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	1	0.917	1	0.957	0.5	0
0	0	0	0	0	0.5	1
Weighted Avg.	0.917	0.917	0.842	0.917	0.878	0.5

- *Weka FilteredSubsetEval Değişken Seçim Algoritmasının Uygulanması*

FilteredSubsetEval algoritması sonucu seçim sonuçlarıdır (Çizelge 7.58, 7.59, 7.60).

Çizelge 7.58. Veri kümesi özellikleri

Instances: 1615	Test mode:split 70.0% train, remainder test
Attributes: 3	

Çizelge 7.59. J48 uygulaması istatistik sonuçları

Correctly Classified Instances	444	91.7355 %
Incorrectly Classified Instances	40	8.2645 %
Kappa statistic	0	-
Mean absolute error	0.1542	-
Root mean squared error	0.2754	-
Relative absolute error	99.6058 %	-
Root relative squared error	99.9966 %	-
Total Number of Instances	484	-

Çizelge 7.60. J48 uygulaması doğruluk sonuçları

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	1	0.917	1	0.957	0.5	0
0	0	0	0	0	0.5	1
Weighted Avg.	0.917	0.917	0.842	0.917	0.878	0.5

8. ARAŞTIRMA BULGULARI VE TARTIŞMA

Çıkan sonuçlar değerlendirildiğinde 4 farklı değişken seçim deneyi ile 6 farklı sınıflandırma algoritmasının başarı oranları ölçümlenmiştir. İlk deneyde (Seçim A) bütün değişkenler kullanılarak oluşturulmuş ve sınıflandırma algoritma sonuçlarına ulaşılmıştır. İkinci deneyde seçim "B" sınıflandırma için Weka ConsistencySubsetEval değişken seçim algoritması kullanılarak en iyi 3 değişken kümesi elde edilerek aynı sınıfladırma algoritmaları başarısı ölçülmüş ve aynı şekilde üçüncü seçim "C" ve dördüncü seçim "D" deneylerde sırasıyla Weka CfsSubsetEval ve Weka FilteredSubsetEval değişken seçim algoritmalarına tabi tutulup her bir değişken kümesi sınıflandırma başarı sonuçları Şekil 8.1.'deözetlenmiştir. Tabloda A, B, C ve D özellik seçim algoritmaları sonucu işleme alacağımız özellikler belirtilmiştir. Kullandığımız seçim algoritmalarına bağlı olarak en iyi değişken kümeleri, "sahtecilik" alanını doğru belirlemede en iyi değişkenler olduğundan tabloda her bir sınıflandırıcı için "sahtecilik" değişkeni ayrıca belirtilmemiştir.

Şekil 8.1.'e göre consistencySubsetEval değişken seçim algoritmasıyla elde edilmiş IBK (K en yakın komşu) algoritması %95,86 oranında bir başarı oranıyla en başarılı sonucu elde edilmiştir.

	Best First Attributes	Navie Bayes Ratio	RBF Network	IBK Ratio	NBTree Ratio	J48 Ratio
All Attributes	total					
	payment_ref_code					
	amount					
	orderHour					
	orderDayOfWeekType					
	NameSurnameLen					
	indirim_para					
	coupon_discount	92.1488 %	92.7686 %	94.0083 %	94.8347 %	94.4215 %
	shipment_amount					
	coupon_id					
	emailConfirmTimeType					
	customer_city_id					
	customerEmailFormatType					
	order_brandid					
	category_id					
	Customer_age					
	Sex					
Evaluator: weka.attributeSelection.ConsistencySubsetEval	total					
Search:weka.attributeSelection.BestFirst -D 1 -N 5	payment_ref_code					
	NameSurnameLen	91.9421 %	93.3884 %	95.8678 %	94.4215 %	91.7355 %
Evaluator: weka.attributeSelection.CfsSubsetEval	total					
Search:weka.attributeSelection.BestFirst -D 1 -N 5	NameSurnameLen	94.8347 %	94.8347 %	95.0413 %	94.8347 %	91.7355 %
	coupon_discount					
Evaluator: weka.attributeSelection.FilteredSubsetEval	total	94.4215 %	94.4215 %	95.0413 %	94.4215 %	91.7355 %
Search:weka.attributeSelection.BestFirst -D 1 -N 5	NameSurnameLen					

Şekil 8.1. Sonuç değerleri

Çizelge 8.1. Özellik seçim algoritmalarına bağlı başarı sonuçları

	Navie Bayes	RBF Network	KNN	NBTree	J48
A Durumu Özellik Seçimi	92.1488 %	92.7686 %	94.0083 %	94.8347 %	94.4215 %
B Durumu Özellik Seçimi	91.9421 %	93.3884 %	95.8678 %	94.4215 %	91.7355 %
C Durumu Özellik Seçimi	94.8347 %	94.8347 %	95.0413 %	94.8347 %	91.7355 %
D Durumu Özellik Seçimi	94.4215 %	94.4215 %	95.0413 %	94.4215 %	91.7355 %

Çizelge 8.2.'ye göre A, B, C ve D değişken seçim koşullarına göre, TP (true positive) oranı, FP (false positive) oranı, duyarlılık, keskinlik ve F-ölçütü değerleri belirtilmiştir.

Çizelge 8.2. Özellik seçim algoritmalarına bağlı başarı sonuçları

Sınıflandırıcı	Özellik Seçim Durumu	TP Oranı	FP Oranı	Keskinlik	Duyarlılık	F-Ölçütü	ROC Oranı
	Sınıflandırma A	0.921	0.326	0.931	0.921	0.925	0.956
Navie Bayes	*SınıflandırmaB*	0.919	0.621	0.906	0.919	0.911	0.926
	SınıflandırmaC	0.948	0.528	0.946	0.948	0.94	0.963
	Sınıflandırma A	0.928	0.461	0.924	0.928	0.925	0.926
RBF Network	*SınıflandırmaB*	0.934	0.415	0.931	0.934	0.932	0.932
	SınıflandırmaC	0.948	0.323	0.947	0.948	0.947	0.954
	Sınıflandırma A	0.94	0.369	0.938	0.94	0.939	0.841
KNN	***Sınıflandırma B***	**0.959**	**0.39**	**0.956**	**0.959**	**0.955**	**0.943**
	SınıflandırmaC	0.95	0.459	0.946	0.95	0.945	0.912
	Sınıflandırma A	0.948	0.551	0.948	0.959	0.939	0.828
NBTree	*SınıflandırmaB*	0.944	0.619	0.944	0.948	0.931	0.794
	SınıflandırmaC	0.948	0.528	0.946	0.948	0.94	0.963
	Sınıflandırma A	0.944	0.619	0.947	0.944	0.931	0.755
J48	*SınıflandırmaB*	0.917	0.917	0.842	0.917	0.878	0.5
	SınıflandırmaC	0.917	0.917	0.842	0.917	0.878	0.5

Çizelge 8.2.'ye göre consistencySubsetEval değişken seçim algoritmasıyla elde edilmiş IBK (K en yakın komşu) en başarılı algoritmasonucuna Çizelge 8.3'teki gibi elde edilmiştir.

Çizelge 8.3. Özellik seçim algoritmalarına bağlı en başarılı sonuç oranları

Sınıflandırıcı	Özellik Seçim Durumu	TP Oranı	FP Oranı	Keskinlik	Duyarlılık	F-Ölçütü	ROC Oranı
KNN (K en yakın komşu)	*Sınıflandırma B*	**0.959**	**0.39**	**0.956**	**0.959**	**0.955**	**0.943**

Çizelge 8.2.'ye göre en başarılı oranı belirlemekte kullandığımız TP (True Positive) oranının 1'e yaklaştığı ve FP (False Positive) oranının 0'a yaklaştığı görülmektedir. TP (True Positive) oranının 1'e yakınsaması gerçekte olan "sahtecilik" olarak nitelendirilen kayıtları doğru tahmin ettiği yönünde değerlendiriyor oluruz.

F-Ölçütü, keskinlik ve duyarlılık oranlarının harmonik ortalama değeri olduğundan ortalama bir başarı ölçütü olarak ele alındığında 0.955 oranla diğer sınıflandırma algoritmalarına göre en iyi sonuç elde edilmiştir.

Şekil 8.1.'e göre değişken seçim algoritmaları sonucu kullanılan değişkenlerin başarı oranının, bütün değişkenler kullanılarak alınan başarı oranından daha büyük olduğu gözlenmektedir. Değişkenler arası korelasyonun büyük olması durumunda daha iyi sınıflandırma yapılabildiği sonucunu varabiliyoruz.

9. SONUÇ VE ÖNERİLER

Bu tez çalışmasında e-ticaret sitelerinde üzerinde kurumu finansal ve itibar anlamında yıpratacak ve site üzerinden alışveriş yapan kişileri de doğrudan etkileyecek kredi kartı sahtekârlıkdavranışlarını tespit eden bir sistem geliştirilmiştir.

Çalışmada kullanılan veriler gerçek bir e-ticaret sitesinden alınarak analiz edilmiştir. 1615 adet e-ticaret sipariş verisi üzerinden sınıflandırma algoritmaları deneyi yapılmış ve elle tespit edilmiş sahtecilik içeren sipariş verileri üzerinde %95 oranında başarı yüzdesine ulaşılmıştır.Oluşturulan model“www.indirimplus.com” adlı e-ticaret sitesi sipariş verileri temel alınarak hazırlanmıştır.

E-ticaret siteleri üzerinden yapılan kredi kartı sahtekârlık yöntemleri tespiti yapılan bu çalışmasıyla kurum içi tehditler en aza indirgenmeyi amaçlamıştır. Farklı sektörlerde de kişi ve kurumları yıpratacak sahtecilik girişim tespitinde çalışmamız uyarlama yöntemleriyle kullanılabilir.

KAYNAKÇA

Ahi, L., (2015). Veri Madenciliği Yöntemleri İle Ana Harcama Gruplarının Paylarının Tahmini.Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü.Yüksek Lisans Tezi, Ankara.

Akçetin, E., (2014). İstenmeyen Elektronik Posta(Spam) Tespitinde Karar Ağacı Algoritmalarının Performans Kıyaslaması. İnternet Uygulamaları ve Yönetimi, 5(2), 46- 48

Aral, K., (2009). Veri madenciliği teknikleri ile reçete usulsüzlüklerinin tespiti: Bir yöntem önerisi, İhsan Doğramacı Bilkent Üniversitesi, Mühendislik ve Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi, Ankara.

Aşuk, C., (2010). Sağlık Sigortası Şirketleri İçin Veri Madenciliği Tabanlı Suistimal Tespit Sistemi, Marmara Üniversitesi, Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi, İstanbul.

Bolat, B., (2004). Mühendislik Uygulamalarında Genetik Algoritmalar ve Operatörlerin İşlevleri, Mühendislik ve Fen Bilimleri Dergisi, 4(2004), 266

Cayiroğlu, I., (2016). Yapay Sinir Ağlarış.Erişim Tarihi: 01.06.2016. http://www.ibrahimcayiroglu.com/Dokumanlar/IleriAlgoritmaAnalizi/IleriAlgoritmaAnalizi-5.Hafta-YapaySinirAglari.pdf

Çerkez, S., (2003). Müşteri İlişkileri Yönetiminde İş Zekası ve Veri Madenciliği Yöntemleri. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi, İstanbul.

Fayyad, U., (1996). From Data Mining to Knowledge Discover in Databases. Al Magazine, 37-54

Fraudandchargeback, (2010). Fraud kontrol noktalari. Erişim Tarihi: 01.06.2016. http://www.fraudandchargeback.com/e-ticaret-siteleri-icin-fraud-kontrol-noktalari-neler-olabilir

Göker, H., (2012). Üniversite Giriş Sınavında Öğrencilerin Başarılarının Veri Madenciliği Yöntemleri İle Tahmin Edilmesi. Gazi Üniversitesi Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi, Ankara.

Göral, A., (2007). Kredi Kartı Başvuru Aşamasında Sahtecilik Tespiti İçin Bir Veri Madenciliği Modeli. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi, İstanbul.

Han, J., (2001). Data Mining: Concept and Tecniques.Morgan Kaufmann Publications, USA.

Kılıç, R., (2010). Kredi Kartı Sahteciliği. Marmara Üniversitesi Bankacılık ve Sigortacılık Enstitüsü. Yüksek Lisans Tezi.

Kılıçarslan, H., (2013). Türkiye Gsm Sektöründe Veri Madenciliği Yöntemi ile Sahtekârlık Tespiti ve Bir Uygulama. Beykent Üniversitesi Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi, Ankara.

Medeni, T., (2007). Bilgi Biliminin Mühendislik Gereksinimi ve Bilgi Mühendisliği. Ankara.

Özalp, N., (2013). Yazar Tespiti için İyileştirilmiş Naive Bayesian Algoritması. Conference Paper. 3-4

Özbay, E., (2007). Finans Sektöründe Veri Madenciliği ile Dolandırıcılık Tespiti, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi, Konya.

Sezen, C., (2011). Telekomünikasyon Sistemlerinde Sahtekârlık Tespiti ve Yönetimi, Ege Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İzmir.

Taşkın, Ç., (2005). Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi uygulaması. Sosyal Bilimler Dergisi, 6(2), 8-9

Tavacı, H., (2012). Gsm Şebekelerinde Sahtekârlık Yönetimi İçin Veri Madenciliği Yöntemlerinin Ugulanması. Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi, İstanbul.

Uğurlu, M., (2007). Yapay Sinir Ağı Modeliyle Bir Bankada Uygulama, Dumlupınar Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, Kütahya.

Türkoğlu, İ., (2009). Web Tabanlı Öğretim Materyallerinin Web Kullanım Madenciliği İle Analiz Edilmesi. Fırat Üniv. Mühendislik Bilimleri Dergisi 22(1), 111-122

Wikipedia, (2016a). Cross-validation. Erişim Tarihi: 01.06.2016. https://en.wikipedia.org/wiki/Cross-validation

Wikipedia, (2016b). K Nearest Neigbours. Erişim Tarihi: 01.06.2016. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Webrazzi, (2013). Fraud nedir nasil hesaplanir nasil engellenir. Erişim Tarihi: 01.06.2016. http://webrazzi.com/2013/05/28/fraud-nedir-nasil-hesaplanir-nasil-engellenir

ÖZGEÇMİŞ

Adı Soyadı : Yasin KIRELLİ

Doğum Yeri ve Yılı : KONYA, 21/08/1989

Medeni Hali : (Bekar)

Yabancı Dili : İngilizce

E-posta : yasinkirelli@gmail.com

Eğitim Durumu

Lise : Ilgın Anadolu Lisesi, 2007

Lisans : Kocaeli Üniversitesi, 2012
Mühendislik Fakültesi, Bilgisayar Mühendisliği

Yüksek Lisans : İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği

Mesleki Deneyim

Bright E-Ventures,
Yazılım Mühendisi 2013-2015

Intertech Aş,
Yazılım Mühendisi 2015-...(devam ediyor)