Route Educational and Social Science Journal
Volume 2(2), April 2015

DATA ANALYZING AND DATA MINING APPLICATIONS IN INCOMPLETE DATA IN THE BUSINESS DATABASE

EKSİK VERİLİ İŞLETME VERİ TABANLARINDA VERİ ANALİZİ VE VERİ MADENCİLİĞİ UYGULAMASI

***Eyüp AKÇETİN*[1]**
***Hüseyin TURGUT*[2]**

Abstract

Converting to information by analyzing the data is very important for businesses in today's knowledge economy. In this study, data of the Vefa Gıda Pazarlama Company are used. This data consists of buying and selling between September 1, 2012 and February 17, 2015 dates. 8 successful data mining algorithms were tested during the study period. The aim of this study which of the data mining method is most suitable for date analyzing when the structure of the data is missing or insufficient in business database. More than 78 000 data was obtained. The transaction is made by 63958 lines of data. 80% of the data are used as training data and 20% was used as the test data. In this study, the most appropriate model logitboost classification algorithm was determined for data base with missing data.
Keywords: Management information systems, incomplete data, database, data mining, knowledge discovery.

Özet

Günümüz bilgi ekonomilerinde, verilerin analiz edilerek bilgiye dönüştürülmesi işletmeler için çok önemlidir. Bu çalışmada, Vefa Gıda Pazarlama firmasına ait veriler kullanılmıştır. Bu veriler 1 Eylül 2012 ve 17 Şubat 2015 tarihleri arasındaki alış ve satışlardan oluşmaktadır. Çalışma sürecinde 8 başarılı veri madenciliği algoritması test edilmiştir. Bu işlem ile 78 000'den fazla veri elde edilmiştir. 63958 satır veri ile işlem yapılmıştır. Verilerin %80'i eğitim ve %20'si test verisi olarak kullanılmıştır. Bu çalışmanın amacı işletme veri tabanlarında eksik veya yetersiz yapıdaki verilerin analizine en uygun veri madenciliği yönteminin bulunmasıdır. Yapılan çalışmada eksik verili veri tabanları için en uygun modelin logitboost sınıflandırma algoritması olduğu tespit edilmiştir.
Anahtar Kelimeler: Yönetim bilişim sistemileri, eksik veri, veri tabanı, veri madenciliği, bilgi keşfi.

[1] Yrd. Doç. Dr., Balıkesir Üniversitesi, Denizcilik Fakültesi, Deniz İşletmeleri Yönetimi Anabilim Dalı, Balıkesir. e.akcetin@gmail.com
[2] Öğr.Gör., Mehmet Akif Ersoy Üniversitesi, Tefenni MYO, Burdur. hturgut.com@gmail.com

Giriş

Günümüz işletmeleri için bilgi, kritik öneme sahip bir kavramdır. Bilginin analiz edilmesi işletme körlüğünü yok ederek rekabetçi üstünlüğü getirir. Üretim faktörlerine doğrudan etki etmekle birlikte bilgi ekonomisi ile bütünleşmeyi de sağlar. İşletme kaynaklarının tek bir merkezden kolayca yönetilmesine imkân vererek verimlilikte artışı, atıl kapasitede ise azalışı temin eder.

Küresel dünyada bilgi teknolojileri ekonomik büyümenin birincil itici gücü olmuştur. Peter Sondergaard'agöre *"Bilgi 21. Yüzyılın petrolü, bilginin analizi ise içten yanmalı motorudur."* Bilgi teknolojilerine yapılan yatırım, işletme performansını doğrudan etkilemektedir. Bu nedenle bilgi teknolojilerine yatırım her geçen gün artmaktadır (Gartner, 2012). Dünya genelinde 2014 yılında bilgi teknolojilerine 3,7 trilyon dolarlık yatırım yapılmış, 2015 yılında ise 3,8 trilyon dolarlık yatırım yapılacağı tahmin edilmektedir (Gartner, 2015).

Günümüz işletmeleri müşterilerini yakından tanımak için öncelikle etkin veri ambarları oluşturmak zorundadır. 3600 marketi bulunan ve Dünyanın en büyük perakende marketi olan Wal-Mart, haftalık 100 milyon müşteriye satış yapmaktadır. Wal-Mart, müşterilerinin sosyal güvenlik kayıtlarından ehliyet numaralarına kadar bütün bilgilerini (kredi borçları, trafik cezaları vb.) kayıt altında tutmaktadır. Wal-Mart'ın ana bilgisayarlarında toplanan bilgi miktarı 460 terabayttır. Bu bilgi, internetteki tüm bilgilerin yarısı kadardır. Wal-Mart, veri analizini o kadar iyi yapmaktadır ki bir yere market açmadan önce müşterilerin o marketten en çok neyi alacaklarını tahmin edebilmektedir. Wal-Mart, bugün raflarında bulunan 50 milyar dolarlık ürünleri kimlerin satın alacağını bu sayede tahmin edebilmektedir (Hays, 2004).

A.C. Nielsen isimli araştırma şirketi, Wal-Mart'a topladığı bilgileri almak için milyonlarca dolar ödemektedir. A.C. Nielsen, bu bilgileri diğer perakendecilere satmaktadır. Fakat Wal-Mart verilerin ana kaynağı olduğu için rekabetçi üstünlüğü elde ettiğinden Amerikan pazarında birinci sırada yer almaktadır. Wal-Mart 1991 yılında 4 milyar dolar harcayarak tüm ürün akış süreçlerini dijitalleştirmiştir. Wal-Mart, verileri müşterilerinden toplamak için de çeşitli stratejiler geliştirmiştir. Eğer müşteri, kredi kartı ile ödeme yapmak istiyorsa Wal-Mart sosyal güvenlik numarasını talep etmekte, müşteri bu bilgiyi paylaşmak istemezse de peşin ödeme yapması gerektiğini ve ehliyet numarasının yeterli olduğunu belirtmektedir. Ayrıca Wal-Mart online satış yapmakta ve bu sayede edindiği verileri diğer verilerle eşleştirerek müşterilerini daha yakından tanıma fırsatı bulmaktadır. Örneğin; online satışlarda özellikle salgın dönemlerinde soğuk algınlığı ilacı, tavuk suyu çorbası ve portakal suyu birlikte alınmaktadır. Üstelik fiyatların indirimli olması müşteriler tarafından umursanmamaktadır. Çünkü müşteriler, büyük bir alış veriş sepetini %10 ile 20 daha düşük fiyata mal edebildiklerinden, hemen hemen tüm ürünleri düşük fiyata Wal-Mart'tan alacaklarına inanmaktadırlar. Wal-Mart'ın CEO'su Linda M. Dillman'a göre günümüzde *"Bilgi yalnızca güç değil aynı zamanda kârdır."*(Hays, 2004).

Küresel rekabet ortamı ve bilgi ekonomisi işletmeler için hem fırsat hem tehdittir. İşletmeler gerekli alt yapıyı kurarak, işletme içinde yer alan her türlü veriyi kıymetli bilgilere dönüştürmek zorundadır. Müşterilerin verileri şirket içinde yer alırken aynı zamanda sosyal platformlarda da yer alabilmektedir. Önemli olan bu verileri bütünleştirip çeşitli analizler yaparak, elde edilen kıymetli bilgilerle, etkin müşteri ilişkileri geliştirebilmektir. İşletmeler açısından, günümüz küresel bilgi ekonomisinde bu durum çok büyük önem arz etmeye başlamıştır.

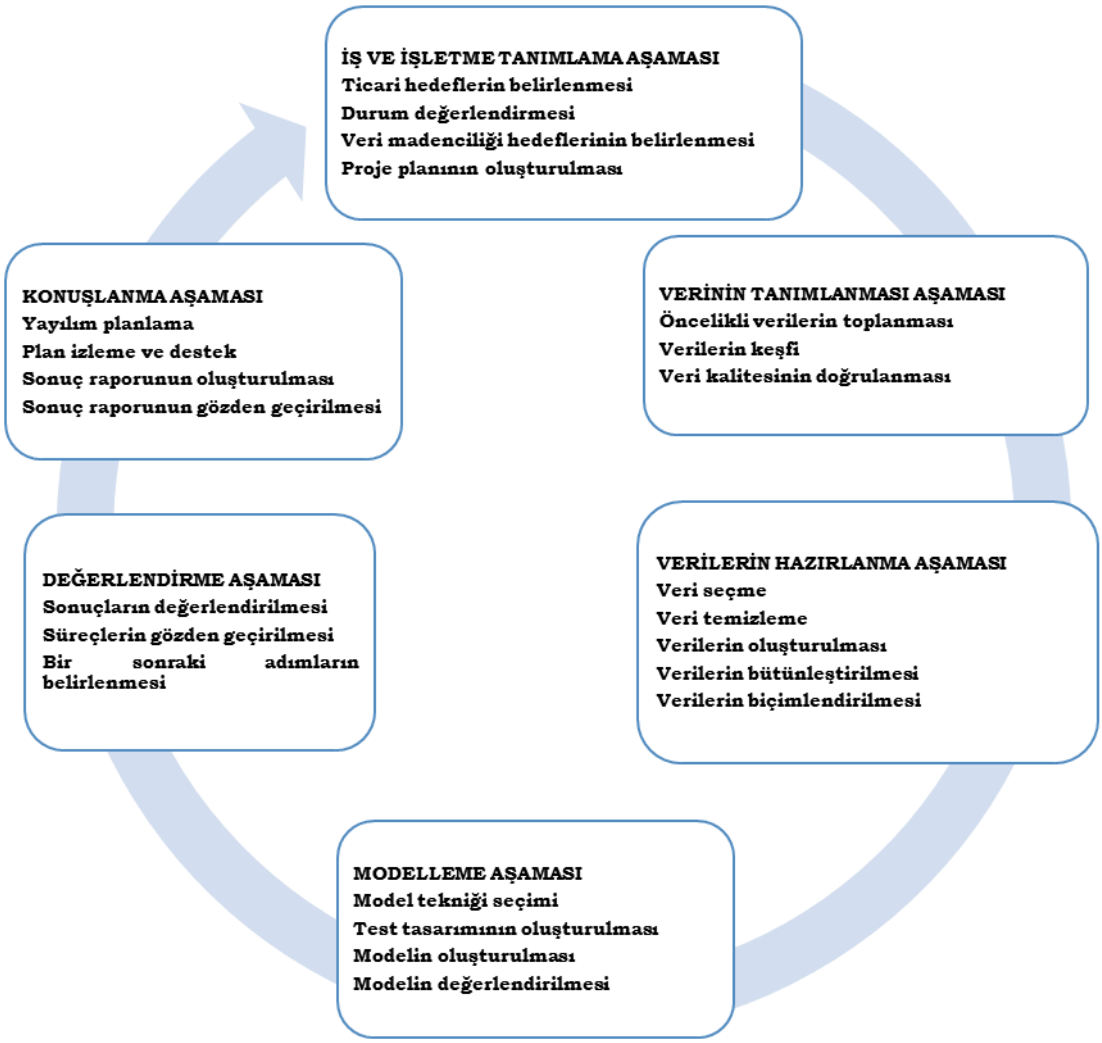
Literatür Taraması

İşletmeler veri madenciliği yöntemlerini, daha iyi karar verebilmek ve iş kararlarında karar vericilere yardımcı olmak amacıyla, veri desenlerini ve bu desenlerin ilişkilerini keşfetmek için kullanmaktadır. Veri madenciliği, satış eğilimlerinin belirlenmesinde, doğru pazarlama kampanyalarının geliştirilmesinde ve müşteri sadakatinin kesinleştirilmesinde işletmelere, yardımcı olabilmektedir (Sumathi & Sivanandam, 2006, s. 16-17). Veri madenciliği, trend analizlerinin yapılmasında da kullanılmaktadır. Böylece önceki tarihlerde yapılan satışlar incelenerek geleceğe yönelik tahminde bulunmak mümkündür (Oz, 2006, s. 355).

Doğrudan pazarlama için doğru müşteriye, doğru ürünle, minimum maliyet ve doğru iletişim kanalı ile ulaşılması, müşteri profillerinin veri madenciliği yöntemleriyle analiz edilmesine bağlıdır. Böylece ürünü talep etmeyecek müşterilerle iletişim kurulmayarak zaman ve maliyet kaybının önüne geçilecektir (Liao, Chen, & Hsieh, 2011, s. 6059–6069). Veri madenciliği yöntemleriyle yapılan bu tür çalışmalar, işletmelere stratejik ve taktiksel güç kazandırarak rekabetçi üstünlük sağlayacaktır (Crone, Lessmann, & Stahlbock, 2006, s. 781-800). Yine yapılan bu tür çalışmalar doğru amaçların ve yeni müşterilerin bulunmasında yardımcı olur. Müşteri özelliklerinin detaylı analiz edilmesini sağlar. Alış veriş örnekleri ile müşterilerin zamanla daha iyi tanınması sağlanarak müşteriye uygun servis ve ürünler müşteriye sunularak müşteri memnun edilerek, müşteri sadakati artırılır. Etkileşimli pazarlamada, veri madenciliği yönteminin kullanılması özellikle internet üzerinden alış veriş yapan tüketici ve/veya müşterilerin ne tür işlemler yaptığı ve hangi ürünlerle hangi servisleri kullandığı analiz edilebilir. Sonrasında müşteri ilişkilerin geliştirilerek müşterilere özel hizmet sunulabilir (Shaw, Subramaniama, Tan, & Welge, 2001, s. 127–137).

İşletme karar verme süreci ve uygulanması; işletme kararı ve uygulaması, iş sorunu, sorunların süreç adımları ve sorunun özellikleri olmak üzere dört bölümde ele alınır. Veri madenciliği algoritmaları da dört bölümde tasvir edilir. Bunlar; veri madenciliği metotları, veri giriş birimleri, veri çıkış birimleri ve algoritmik adımlardır. İşletme kararları ve uygulamaları, veri madenciliği yöntemlerinin özellikleri ile sistemik olarak eşleştirildikten sonra seçim modeli tercihlere uyumlu bir küme oluşturur. Oluşturulan bu kümeler ile işletme sorunlarına etkin çözüm yolları bulunur (Senga & Chen, 2010, s. 8042–8057).

Veri madenciliği iş dünyasında ve ticari kuruluşlarda hızla büyüyen bir uygulama alanıdır (Bose & Mahapatra, 2001, s. 211–225). Çünkü veri madenciliği, karmaşık ve büyük verilerden değerli bilgiyi ayıklamanın tek yoludur (Lausch, Schmidt, & Tischendorf, 2015, s. 5–17).



Şekil 1: İşletmelerde veri madenciliği süreci

Kaynak:Sharma, S.,&Osei-Bryson, K.M. (2009). Framework forformalimplementation of thebusinessunderstandingphase of dataminingprojects. *ExpertSystemswith Applications*, 4114–4124.

Şekil 1'den anlaşılacağı üzere iş ve işletmenin tanımlanması hangi verilerin seçilmesi ve bu verilerin hangi biçimde olması gerektiğini belirler. İşletmenin amaçlarına göre model belirlenir. Bu modelin tekniğine dayalı olarak süreç içindeki adımlar belirlenir ve tüm süreç gözden geçirilir. En son aşamada yayılım planlaması yapılır, geliştirilir ve ayrıca final raporu oluşturularak gözden geçirilir. İşletmelerde veri madenciliği sürecinin belirleyicisi iş ve işletme amaçlarıdır. Bu amaçlar veri madenciliğinin adeta pusulasıdır (Sharma & Osei-Bryson, 2009, s. 4114–4124).

Çalışmanın Önemi ve Amacı

Vodafone'un dünyanın önde gelen yönetim danışmanlığı şirketlerinden Accenture ile yaptığı araştırmaya göre, Türkiye'deki işletmelerin dijitalleşme endeksinin yüzde 48 olduğu tespit edilmiştir. Yapılan bu araştırmaya göre;

- Her 10 işletmeden 6'sı sahada araç kullanmakta fakat bu işletmelerin %76'sı araçlarını takip etmemektedir. Öte yandan araç takip sistemi kullanan işletmeler %15'e varan oranda yakıt tasarrufu sağlayabilmektedir.
- Her 10 işletmeden 6'si sahada ekibi ile iletişim kurmaya ihtiyaç duymasına rağmen, bu işletmelerin %96'sının ekipleri arasında anlık iletişim sağlayacak teknolojileri kullanmamaktadır.
- Her 10 işletmeden 7'si müşterileri için prezantasyon faaliyetlerine ihtiyaç duymaktadır. Lakin bu işletmelerin %55'i bu ihtiyacına yardımcı olacak teknolojileri kullanmamaktadır. Diğer taraftan müşterilerine toplu mesaj ile tanıtım yapan işletmeler satışlarını %28'e varan oranlarda artırabilmektedir.
- Her 10 işletmeden 6'sının işyeri dışında tahsilât yapma ihtiyacı olmasına rağmen yalnızca %43'ü müşterilerine kolay ödeme imkanı sunan Mobil Pos teknolojisini kullanmaktadır.
- Ayrıca her 10 işletmeden 4'ünün işini internet ortamına taşıyarak satış yapmaya ihtiyacı olmasına rağmen bu işletmelerin %63'ü bu ihtiyacına yönelik teknolojik çözümü kullanmamaktadır (eticaretmag, 2014).

Vodafone Türkiye İcra Kurulu Başkan Yardımcısı Engin Aksoy göre; tüm dünyada dijitalleşen şirketler kaynaklarını artırmadan gelirlerini sektörlerinin yüzde 9 üzerinde büyütmekte ve kârlılıkları sektörlerinin %26 üzerinde olmaktadır. Kısaca dijitalleşen işletmeler operasyonel verimlilik kazanarak daha çok müşteriye ulaşmaktadır (Haber Türk, 2015). Doğru müşterilerle doğru ürünleri eşleştirebilen ve müşterisinin ihtiyaçlarını öngörebilen işletmeler, satışlarını arttırıp yüksek kârlar elde etmektedir. İşletmeleri bu hedefe taşıyacak köprü ise veri madenciliğidir. Performansı yüksek, detaylı bir veri tabanı ise veri madenciliği sürecinin başarı seviyesinin yüksek olmasında kritik öneme sahiptir. Kâr marjı yüksek olan büyük işletmelere bakıldığında veri tabanlarının detaylı ve toplanan verilerinin işletme amaçlarına ve analiz hedeflerine yönelik olduğu görülmektedir.

Ürün ve hizmetlerinin pazarlanmasında veri madenciliği, pazarlama tekniklerinin geliştirilmesinde ve hedef müşterilerin tespitinde kullanılmaktadır. Kullanılan bu yöntem ile müşteri bölümlendirme yapılarak daha önce yapılan kampanyalara iştirak eden müşteriler belirlenir. Böylece hangi kampanya hangi tür müşterilerin iştirak edeceği tahmin edilir. Çapraz satış için kârlı müşterilerin profili çıkarılır ve onlara uygun ürün geliştirilebilir. Ayrıca kurumsal yıpranma analizi yapılarak müşteri davranışlarındaki sapmalar tespit edilebilir (Bach, Jukovic, Dumiči, & Šarlija, 2013, s. 32-41).

Aynı ürünleri ve hizmetleri talep eden müşterilerinin ortak özelliklerini veri madenciliği yöntemlerini kullanarak tanımlayabilir. Market sepeti analizi yapılarak hangi ürünlerin birlikte ve ne zaman satın alındığına yönelik detaylı bilgilere ulaşılır. Böylece müşterilere uygun ürün ve servis oluşturulurken, ürün ve servislere uygun müşteriler bulanabilir (Brito, Soares, Almeida, Monte, & Byvoet, 2015, s. 1-8).

Kısaca işletmeler, veri madenciliği yöntemini kullanarak aşağıda sıralanan sorulara yanıt bulmaya çalışırlar.

- Müşterinin sadakati nasıl sağlanır?
- Yüksek riskli tüketiciler ve müşteriler kimlerdir? Belirgin özellikleri nelerdir?
- Sadık müşteriler genelde hangi ürünleri ve/veya servisleri kullanmaktadır?
- Sadık müşteriler ne tür ürünleri ve servisleri talep etmektedir?
- Müşteri gözünde işletmeyi yıpratan unsurlar nelerdir?
- Müşterileriler hangi dönemlerde artmaktadır?
- Mevcut müşterilere daha çok ürün ve hizmet nasıl satılabilir?

- Bir müşteri rakip işletme ürün ve hizmetlerini tercih etmeden önce hangi davranışları sergilemektedir?
- Yeni müşteriler, en düşük maliyetle nasıl edinilir?
- Hangi müşteri profiline hangi mesajla, hangi ürünler önerilmelidir?
- İşletme hedefleri gerçekleştirirken kime, nerede ve hangi kanaldan satış yapılmalıdır?
- Hangi kanala daha çok yatırım yapmalıdır?
- Çalışanlarımızın üretkenliği neye göre değişmektedir?
- Çapraz satış için, müşterilerin detaylı profili nedir?
- Çapraz satış ve diğer ürünler için müşterilerin belirlenmesinde hangi müşteri grubu hangi ürünleri birlikte kullanmaktadır?
- En kârlı ürünümüz hangisidir?
- Hangi operasyonlara ağırlık verilmelidir?
- Potansiyel müşteri adaylarını kalıcı müşteri yapmak için neler yapılmalıdır?

Bu çalışmada veri madenciliği yöntemlerinden Balıkesir ilinde yer alan Vefa Gıda Pazarlama firmasına ait verilere uygun algoritmalar test edilerek bu algoritmaların performanslarının kıyaslanması amaçlanmıştır. Böylece yukarıda sıralanan sorulara işletmeler tarafından yanıt aranırken işletme veri tabanlarında eksik veya yetersiz yapıdaki verilerin analizine en uygun veri madenciliği yöntemi bulunup test edilmiştir.

Materyal

Veri madenciliği uygulamasında kullanılacak olan materyal olarak Balıkesir Vefa Gıda Pazarlama firmasına ait 1 Eylül 2012 ile 17 Şubat 2015 tarihleri arasında yapılmış olan satışlara ait fatura bazlı veriler kullanılmıştır. Veri tabanından elde edilen veriler Tablo 1'deki yapıdadır.

Tablo 1:Satışlara ait veri tabanı yapısı

Tarih	01.09.2012	11.01.2013	11.01.2013	21.01.2013	25.01.2013
İrsaliye					
Fatura	A-215387	A-215387	A-215387	A-215519	A-215527
Konsinye	0	0	0	0	0
Müşteri ID	1	1	3	270	270
Müşteri	Firma Adı1	Firma Adı 1	Firma Adı 3	Firma Adı 270	Firma Adı 270
Personel	Sipariş	Sipariş	Sipariş	Sipariş	Sipariş
Kod	C 9000 1125	C 9000 1038	K 003888	PK 00002	PK 00034
Ürün Adı	Urunadı1	Urunadı1	Urunadı1	Urunadı1	Urunadı1
Miktar	6,00	6,00	3,00	4,00	2,00
Birim	Ad	Ad	Ad	Ad	Ad
Birim Fiyat	4,90	9,00	6,60	4,60	1,11
Ara Toplam	29,40	54,00	19,80	18,40	2,22
İskonto	0,00	0,00	0,00	0,92	0,11
İskonto(%)	0,00	0,00	0,00	5,00	4,95
Açıklama	A	B			

Veri tabanından elde edilmiş olan orijinal veri tablosuna ait 20 farklı başlık ve her başlıktaki veri çeşitliliği ise şu şekildedir:

Şube ID : Satışı yapan şubeye ait kimlik numarasıdır. Çeşitliliği tektir.
Şube : Şube ID'ye bağlı olan şube ismidir.
Tip : Satış tipidir. ("Satış" ve "Alıştan İade")

Tarih : Faturanın kesildiği satışın yapıldığı tarihtir. Gün, ay ve yıl verisini içermektedir.
İrsaliye : İrsaliye bilgisini içermektedir.
Fatura : Fatura numarası bilgisidir.
Konsinye : Konsinye durumunu belirtir.
Müşteri ID : Müşteri firmaya ait eşsiz kimlik verisidir.
Müşteri : Müşteri ID verisini tanımlayan isimdir.
Personel : Sipariş ve Muhasebe olarak 2 farklı tipe sahiptir.
Kod : Ürüne ait eşsiz ürün kodudur.
Ürün Adı : Ürün koduna bağlı ürün adıdır.
Miktar : Satılan birim sayısıdır. Ağırlık veya adet bazlı bir veridir.
Birim : Kg ve adet olarak satılan ürünlerin çeşitliliği için kullanılmaktadır.
Birim Fiyat : Birime ait fiyat bilgisidir.
Ara Toplam : (Miktar) x (Birim Fiyat) değeridir.
İskonto : Yapılan iskonto tutarı değeridir.
İskonto (%) : İskonto oranıdır.

Açıklama : Faturada belirtilemeyen ancak satışı hatırlatıcı nottur.

Çeşitliliği ve özellikleri belirtilmiş olan veriler iş yeri için yeterli seviyede bulunmaktadır. Ancak verilerin analize uygunluğunun belirlenmesi amacıyla veri tabanından elde edilmiş olan verilerle veri madenciliği sürecine aşağıdaki aşamalarla geçilmişti (Witten, Frank, & Hall, 2011, s. 369-423).

Kısıtlar

Eksik verili veri tabanlarına sahip işletmelerde, veri analizinin yapılabilirliğini ölçmek amacı ile gerçekleştirilen bu çalışmada; Balıkesir ilinde yer alan Vefa Gıda Pazarlama işletmesinden temin edilen veriler ile birçok veri madenciliği algoritması denenmiş ve en başarılı 8 algoritmanın analizi yapılmıştır. Bu algoritmalarda pek çok kısıtlarla karşılaşılmıştır. Bu çalışma kapsamında karşılaşılan en önemli kısıtlar şunlardır: Karar ağacı algoritmalarında nadiren kurulabilen modellerin çok yapraklı ve çok dallı olmasından dolayı sonuca ulaştıracak uygulamanın basit bir yapıya sahip olamayacağı tespit edilmiştir. Sınıflandırma algoritmalarında yaşanan en temel kısıt ise modelin ancak nominal değişkenlerle kurulabilmesidir. Bu durum veri madenciliği sürecinin veri temizleme ve dönüştürme aşamalarında ciddi zaman kaybına neden olmuştur.

Yöntem

Bu çalışmada geçmiş tarihli satışlara ait verilerin, kendi aralarındaki ilişkilerini tanımlayabilmek için veri madenciliği yöntemi kullanılmıştır. Veri madenciliği, verideki eğilimleri, ilişkileri ve profilleri belirlemek için veriyi sınıflandıran bir analitik araç ve bilgisayar yazılım paketi olarak tanımlanabilir. Spesifik veri madenciliği yazılımları; kümeleme, doğrusal regresyon, sinir ağları, bayes ağları, görselleştirme ve ağaç tabanlı modeller gibi pek çok modeli içermektedir. Veri madenciliği uygulamalarında uzun yıllar istatistiksel yöntemler kullanılmıştır. Bununla birlikte, bugünün veri madenciliği teknolojisinde eski yöntemlerin tersine büyük veri kümelerindeki eğilim ve ilişkileri kısa zamanda saptayabilmek için yüksek hızlı bilgisayarlar kullanılmaktadır. Böylece veri madenciliği, bilenmeyen eğilimleri ortaya çıkarmaktadır (Turgut, 2012, s. 16).

Veri madenciliğinde kullanılan yöntemler işlevlerine göre aşağıdaki gibi üç temel grupta sıralanmaktadır (Akbulut, 2006, s. 20-25):

1. Sınıflama (Classification),

2. Kümeleme (Clustering),

3. Birliktelik kuralları ve sıralı örüntüler (Associationrulesandsequentialpatterns).

Gerek tanımlayıcı gerekse tahmin edici modellerde yoğun olarak kullanılan belli başlı istatistiksel yöntemler; sınıflama (classification) ve regresyon (regression), kümeleme (clustering), birliktelik kuralları (associationrules) ve ardışık zamanlı örüntüler (sequentialpatterns), bellek tabanlı yöntemler, yapay sinir ağları ve karar ağaçları olarak gruplandırılabilir. Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir (Dumouchel, 1999, s. 190-196).

Sınıflama ve Regresyon Modeli

Sınıflama, verinin önceden belirlenen çıktılara uygun olarak ayrıştırılmasını sağlayan bir tekniktir. Çıktıların hazırda bilinmesi sebebiyle sınıflama, veri kümesini denetimli olarak öğrenir. Mevcut verilerden hareket ederek geleceğin tahmin edilmesinde faydalanılan, veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan, sınıflama ve regresyon modelleri arasındaki temel özellik, tahmin edilen bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır. Ancak çok terimli lojistik regresyon (multinomiallogisticregression) gibi kategorik değerlerin de tahmin edilmesine olanak sağlayan tekniklerle, her iki model giderek birbirine yaklaşmakta ve sonuçta aynı tekniklerden yararlanabilmektedir (Turgut, 2012, s. 32-41).

Regresyon analizi:

Kullanılan bağımsız değişken sayısına göre:

- Basit regresyon analizi (Tek bağımsız değişken)
- Çoklu regresyon analizi (Birden çok bağımsız değişken)

Fonksiyon tipine göre:

- Doğrusal regresyon analizi
- Doğrusal olmayan (eğrisel) regresyon analizi

Verilerin kaynağına göre:

- Ana kütle verileriyle regresyon analizi
- Örnek verileriyle regresyon analizi olarak sınıflandırılmaktadır.

Ana kütle (evren) için basit doğrusal regresyon denklemi;

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad (2.1)$$

Bu denklemde β_0, $x = 0$ olduğunda regresyon doğrusunun dikey ekseni kestiği noktayı göstermektedir. β_1 doğrusal fonksiyonun eğimi, yani bağımsız değişken x'deki bir birimlik değişmenin bağımlı değişken Y'de (Y cinsinden) ne kadarlık bir değişme meydana getirdiğini gösteren regresyon katsayısıdır. ε ise, rassal (tesadüfi) hata terimidir. Artık veya kalıntı (residual) adı da verilmektedir. $\varepsilon = Y - \hat{Y}$ 'dir. $\hat{Y}$, tahmini bağımlı değişkenin değerini göstermektedir. Gerçek hayat uygulamalarında β_0 ve β_1 değerleri bilinmiyorsa, ana kütleden örnekler alınarak bunların tahmincileri olan b_0 ve b_1 kullanılarak 2.1 nolu denklem;

$$y = b_0 + b_1 x + e \quad (2.2)$$

olarak yazılır. Ana kütle ve örnek için çoklu doğrusal regresyon denklemleri ise sırasıyla;

$$Y = \beta_0 + \beta_1 \ x_1 + \beta_2 x_2 + \ldots + \beta_n \ x_n + \varepsilon \quad (2.3)$$

$$Y = b_0 + b_1 x_1 + b_2 \ x_2 + \ b_n x_n + e \quad (2.4)$$

olarak ifade edilmektedir (Kalaycı, 2008, s. 273-297).

Kümeleme Modeli

Kümeleme, veriyi sınıflara veya kümelere ayırma işlemidir. Aynı kümedeki elemanlar birbirleriyle benzerlik gösterirken başka kümelerin elemanlarından farklıdır. Kümeleme algoritmaları istatistik, biyoloji ve makine öğrenimi gibi pek çok alanda kullanılır. Kümeleme modelinde, sınıflama modelinde olan veri sınıfları yoktur. Sınıflama modelinde, verilerin sınıfları bilinmekte ve yeni bir veri geldiğinde bu verinin hangi sınıftan olduğu tahmin edilebilir. Oysa kümeleme modelinde, sınıfları bulunmayan veriler gruplar halinde kümelere ayrılır. Bazı uygulamalarda kümeleme modeli, sınıflama modelinin bir ön işlemi gibi görev alabilmektedir (Ramkumar & Swami, 1998, s. 9-14) (Baykal, 2006, s. 95-107).

Birliktelik Kuralları ve Ardışık Örüntüler

Birliktelik kuralları, büyük veri kümeleri arasında birliktelik ilişkileri bulur. Toplanan ve depolanan verinin her geçen gün artması sebebiyle, şirketler veri tabanlarındaki birliktelik kurallarını ortaya çıkarmak istemektedir. Büyük miktarda yapılan işlem kayıtlarından ilginç birliktelik ilişkilerini keşfetmek, şirketlerin karar alma işlemlerini daha verimli hale getirmektedir. Birliktelik kurallarının kullanıldığı en tipik örnek market sepeti uygulamasıdır (Han & Fu, 1999, s. 1-12).

Uygulama

İş yerinin bilgi kontrolü ve denetimi amacıyla kullanmış olduğu veri tabanından elde edilen verilerin analize uygunluğunun incelenmesi için veri madenciliği süreci uygulanmıştır. Gürültülü, eksik ve tanımlanamayan veriler, veri madenciliği sürecinde başarısızlığa neden olabildiğinden; veri temizleme, bütünleştirme, seçme ve dönüştürme işlemlerine tabi tutulmuştur (Akbulut, 2006, s. 39-46). Süreç uygulanabilir algoritmaların kullanımı ile tamamlanmıştır. Verilerin hazırlanması aşaması ve kullanılabilir veri madenciliği algoritmaları ile oluşturulan modellerden en başarılı 8 algoritma modeli aşağıda sırası ile incelenmiş ve açıklanmıştır.

Veri Temizleme ve Veri Bütünleştirme

Veri tabanında; işlenmemiş, tamamlanmamış veya gürültülü birçok değer bulunabilmektedir. Bunlar kullanılmayan veya gerekli olmayan, kayıp, gözlemden sapan, tutarsız verilerden kaynaklanıyor olabilir. Veri madenciliğinin kullanılabilmesi için verilerin ön hazırlıktan geçirilmesi gerekmektedir. Ön hazırlıkta verilere temizleme, bütünleştirme, dönüştürme ve indirgeme işlemleri uygulanmaktadır (Oğuzlar, 2003, s. 70-74).

Öncelikle değerleri tüm satırlarda aynı ve sabit olan veri sütunları gürültülü ve tutarsız verileri yok etmek amacıyla silinmiştir. Bunlar "ŞubeID", "Şube" ve "Konsinye" sütunlarıdır. Birden fazla dosyada bulunan veriler tek bir dosyada birleştirilmiştir. Bu işlem ile 78.000'den fazla veri elde edilmiştir.

Veri Seçme

Veri seçimi aşamasında analiz için kullanılabilir veri başlıkları ile yeni bir veri tablosu oluşturulması planlanmaktadır. İlk olarak sütunlar arası değişiklik göstermeyen ya da biri diğerini tanımlayan veri sütunlarının elenerek devam edilmesi planlanmıştır. "MüşteriID" ile "Müşteri" başlıklarından "MüşteriID"; "Kod" ile "Ürün Adı" başlıklarından "Kod" başlıkları tercih edilmiştir. Bu ikişer değişkenin her biri diğeri ile aynı tanımda kullanılmıştır. Ayrıca tüm verilerde;

- "Açıklama" sütunu dolu olmadığından anlamlandırma yapılamamış ve "Açıklama" sütunları,
- "Birim fiyat" ile "miktar" değerleri "ara toplam" verisini oluşturulabildiği için "Ara Toplam",
- "İskonto" ile "İskonto yüzdesi" değerleri "Ara Toplam" değerinden hesaplanabileceği için "İskonto" sütunları ile
- Adet azlığından dolayı "İrsaliyeli" satış ve "iskontolu satış" sütunları tercih edilmemiştir.

Veri başlıklarının belirlenmesinin ardından veri tablosundan tek sefer veya tek ürün satışları ile ticari kaygısı bulunmayan müşteriye yapılan satışlar ve promosyon olarak verilen faturalı tanıtım ürünlerine ait satışlar silinmiştir. Bir sonraki aşamaya 14 sütunlu ve 63958 satırlı veri tablosu ile devam edilmiştir.

Veri Dönüşümü

Ham veri kitlesinden kullanılabilir bir yapı çıkartmayı hedefleyen veri dönüşümü, veri madenciliği sürecinin en önemli aşamasıdır. Veri dönüşümü, kullanılacak model ve algoritmaların veriyi tanımlaması ve fark edebilmesi için gereklidir (Turgut, 2012). Hedefe ulaşmak amacıyla uygulanan aşamalar aşağıda tanımlanmıştır;

- "Tarih" verisinden; gün, ay, yıl, haftanın günü (Pazartesi, Salı, Çarşamba…) ve yılın kaçıncı haftası (1..52) olmak üzere beş ayrı yeni başlık oluşturulmuştur.
- Satılan ürünlere ait ürün firma adı (marka) başlığı ile yeni bir sütun oluşturulmuştur.
- Müşteri numarasını ve fatura numarasını sıralandırmak için sayısal içerikli bir sütun oluşturulmuştur.
- Müşteriye ilk tarih, son tarih ve bu iki tarih aralığında yapılan satış miktarı ayrıca incelenmiştir. İlk tarih verisi ile müşteri yaşı hesabı yapılmış ve yeni bir sütun oluşturulmuştur. Son tarih ise listenin son fatura tarihine yakınlığı göz önünde tutularak "devam ediyor" veya "biten" olarak değer alan iletişim adıyla tanımlanmış yeni bir sütun daha oluşturulmuştur.
- Tüm veri listesinde yer alan ürünlerden kaçar adet satıldığı ayrıca bir sütun değeri olarak belirlenmiş, 50 ve altı için "az", diğerleri için "cok" değeri kullanılmıştır.

Veri dönüşüm aşamasının tamamlanmasıyla uygulamaya dâhil edilen verilerin adı, tanımı, türü ve çeşitlilik bilgileri Tablo 2'de belirtilmiştir.

Tablo 2: Verilerin öznitelikleri

Değişken Adı	Tanımı	Veri Türü	Çeşitlilik
yıl	Fatura tarihinde yıl	Sayısal	2012 ile 2015
hgun	Haftanın günü	Nominal	Pazartesi-Pazar
Hafta	Yılın kaçıncı haftası	Sayısal	1-52
Musterieski	Müşteriye ilk satış tarihine göre yaş	Nominal	0-1-2
İletisim	Son satışın güncelliği	Nominal	Devam – biten
İslem	Müşterinin satın aldığı ürün sayısı	Nominal	Cok-az
Total	Ürüne ait toplam satış adeti	sayısal	6-2418
Marka	Ürünün markası	Nominal	54 farklı isim
Miktar	Satılan birim adeti	sayısal	0,25 – 1800 arasında
Birim	Birim tipi	Nominal	Kg – ad
birimfiyat	Ürünün birim fiyatı	sayısal	0,15 – 92,59

Uygulamada kullanılan veriler sayısal ve nominal olarak iki farklı yapıdadır. Sayısal verilerde sıralama yapılabilirken, nominal verilerde sıralama yapılamamakta ve bu veriler seçenek olarak değer almaktadır. Kullanılan sayısal değişkenlere ait minimum, maksimum, ortalama ve standart sapma değerleri Tablo 3'de, nominal değişkenlere ait alternatif değerler ve tüm satırlardaki tekrar adetleri Tablo 4'de yer almaktadır.

Tablo 3: Sayısal değişkenler ve analiz değerleri

Değişken	Minimum	Maksimum	Ortalama	Standart S.
YIL	2012	2015	2013,394	0,743
Hafta	1	52	26,994	15,01
Total	6	2418	618.711	592.241
Miktar	0.25	1800	19.381	43.221
birimfiyat	0.15	92.59	6.563	6.785

Tablo 4: Nominal değişkenler ve özellikleri

Değişken	Seçenek	Adet			
Hgun	1_pazatesi	1884		bolci	1261
	2_Sali	4946		lolipops	252
	3_carsamba	2013		snickers	41
	4_persembe	12447		twix	22
	5_Cuma	7310		polo	14
	6_Cumartesi	15286		kutlu	34
	7_Pazar	20072		karaca	174
İletisim	devam	50121		bind	135
	biten	13837		ARKAY	742
İslem	cok	61510		soray	258
	az	2448		budak	100
marka	birlik	128		ipek	882
	elit	23159		tempo	108
	kertil	165		maccun	39
	gunes	357		giba	150
	ulker	262		tac	11
	taha	192		tamkur	415
	ozelif	334		beypa	117
	erenler	3303		oztursan	654
	kent	695		pet	10
	ustundag	147		bindalli	283
	noname	333		AKIS	50
	koala	584		natur	251
	mabel	1385		sukran	80
	istanbul	9908		paloma	9
	vefa	16089		miskos	26
	keskin	59		mefendi	8
				teknopak	9

	usas	71		cennet	17
	lart	246		tatlan	4
	sugabee	43		hopla	10
	pelit	27	Birim	Kg	15126
	melodi	117		Ad	48832
	BALIN	188			

Veri Madenciliği

Veri madenciliği, birimin sahip olduğu veri veya enformasyon kaynaklarında yönetici ve analistin sormayı düşünmediği sorular hakkındaki cevapları arayan (Watson & Gray, 1997, s. 102) veya ham verinin tek başına sunamadığı bilgiyi çıkaran bir veri analizidir (Jacobs, 1999, s. 43-46). Ayrıca insanın asla bulmayı hayal bile edemeyeceği eğilimlerin keşfedilmesi olarak da tanımlanabilir (Bransten, 1999, s. 16-20).

Veri madenciliği yazılımının seçimi aşamasında performans/maliyet değerlendirme kriteri ile WEKA 3-7-8 yazılımı kullanılmıştır (Witten & Frank, 2000, s. 365-368). WEKA kullanımında uygun kaynak dosya biçimi CSV veya ARFF olması sebebiyle dosya ARFF formatına dönüştürülmüştür. ARFF dosyasına ait ilk satırlar Şekil 2'de gösterilmiştir.

```
@relationvericsv-weka.filters.unsupervised.attribute.Remove-R1-2-weka.filters.unsupervised.attribute.Remove-R4-
weka.filters.unsupervised.attribute.Remove-R10-weka.filters.unsupervised.attribute.Remove-R13-
weka.filters.unsupervised.attribute.Remove-R4-weka.filters.unsupervised.attribute.Remove-R4-
weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.attribute.Remove-R7

@attributeyilnumeric
@attributehgun {7_Pazar,6_Cumartesi,4_persembe,5_Cuma,1_pazatesi,2_Sali,3_carsamba}
@attribute hafta numeric
@attributemusterieski {0,1,2}
@attributeiletisim {devam,biten}
@attributeislem {cok,az}
@attribute total numeric
@attribute marka
{birlik,elit,kertil,gunes,ulker,taha,ozelif,erenler,kent,ustundag,noname,koala,mabel,istanbul,vefa,keskin,usas,lart,sugabee,p
elit,melodi,BALIN,bolci,lolipops,snickers,twix,polo,kutlu,karaca,bind,barbaros,ARKAY,soray,budak,ipek,tempo,maccun,giba
,tac,tamkur,beypa,oztursan,pet,bindalli,AKIS,natur,sukran,paloma,miskos,mefendi,teknopak,cennet,tatlan,hopla}
@attribute Miktar numeric
@attribute Birim {Kg,Ad}
@attributeBirimFiyatnumeric

@data
2012,7_Pazar,36,2,devam,cok,51,birlik,5,Kg,17.5
2012,7_Pazar,36,2,devam,cok,431,elit,5,Kg,18
2012,6_Cumartesi,40,2,devam,cok,51,birlik,15,Kg,17.5
2012,7_Pazar,41,2,devam,cok,29,elit,6,Ad,6.94
2012,7_Pazar,41,2,devam,cok,203,elit,12,Ad,8.75
2012,7_Pazar,41,2,devam,cok,431,elit,5,Kg,18
2012,4_persembe,43,2,devam,cok,196,kertil,20,Ad,6.6.
```

Şekil 2:ARFF dosyası şablonu

Şekil 2'de de görüldüğü üzere veriler anlamlı bir biçimde dizili ve tanımlıdır. 63.958 satır veri ile işlem yapılmıştır. Aşağıda açıklanan 8 farklı algoritma uygulanmış, verilerin %80'i eğitim ve %20'si test verisi olarak kullanılmıştır.

Logitboost Sınıflandırma Algoritması

Logitboost algoritması Adaboost algoritmasına benzeyen sınıflandırma algoritmasıdır (Friedman, Hastie, & Tibshirani, 2000, s. 337-407). Pratik verilerle çalışabilen bu algoritma, daha iyi bir performans ve düşük azınlık oranı ile hatalı tahmin değerini en aza indirmektedir (Song, Lu, Liu, & Wu, 2011, s. 974 - 977). Oluşturulan algoritma Şekil 3'te gösterilmiştir.

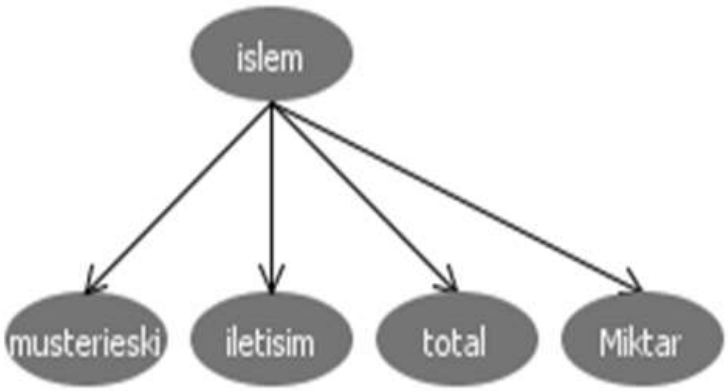
```
Iteration 1 -Class 1 (islem=cok)
iletisim = devam : 1.936713154166916
iletisim != devam : 1.521572595215726
Iteration 2 - Class 1 (islem=cok)
```

```
musterieski<= 0.5 : -0.635354146575275
musterieski> 0.5 : 0.8679939922671618
Iteration 3 - Class 1 (islem=cok)
iletisim = devam : 0.7665965019280738
iletisim != devam : -0.2824772404645931
Iteration 4 -Class 1 (islem=cok)
musterieski<= 1.5 : -0.3325222516156669
musterieski> 1.5 : 0.4032844261589557
Iteration 5- Class 1 (islem=cok)
iletisim = devam : 0.3538807844742449
iletisim != devam : -0.19209043882114804
Iteration 6 - Class 1 (islem=cok)
Miktar <= 2.55 : -0.672081393729815
Miktar > 2.55 : 0.10836339150297196
Iteration 7 - Class 1 (islem=cok)
total<= 54.0 : -0.5954018109414733
total > 54.0 : 0.08628703136165965
Iteration 8 - Class 1 (islem=cok)
Miktar <= 24.5 : 0.06928514050378365
Miktar >24.5 : -0.3418750926946234
Iteration 9 - Class 1 (islem=cok)
musterieski<= 0.5 : -0.3558324436222467
musterieski> 0.5 : 0.05527961126747192
Iteration 10 - Class 1 (islem=cok)
Miktar <= 0.55 : -1.7629763943901071
Miktar > 0.55 : 0.013172949948854925
```

Şekil 3: Logitboost algoritması ile oluşturulan model

BayesNet Sınıflandırma Algoritması

BayesNet algoritması rastgele veriler arasında olasılık hesabına dayalı bir model oluşturmaktadır. Sonuca bağlı değişkenlerin ise aralarında ilişki kurmamakta ve veriler arası karmaşık ilişkiyi azaltmaktadır (Cooper & Herskovits, 1992, s. 309-347). BayesNet algoritması ile elde edilen modelin düğüm yapısı Şekil 4'de gösterilmiştir.



Şekil 4: BayesNet algoritması ile oluşturulan düğümler

NaiveBayes Sınıflandırma Algoritması

Sınıflandırma algoritmalarından olasılıklı yapıya sahip olan NaiveBayes, önermeler üzerine olasılık hesaplaması yapmaktadır. Değerlerin tüm verilerdeki tekrarlanışını belirleyerek aynı zamanda her bağımsız veya bağımlı değişken eşleşmesinin meydana gelme sıklığını da bulmaktadır. Bu yapısı ile kullanım alanı da daralmaktadır (John & Langley, 1995, s. 338-345). Bu algoritmanın çalıştırılması sonucu elde edilen değerler, yapısının büyüklüğü nedeni ile bu çalışmada gösterilememiştir.

SimpleLogistic Sınıflandırma Algoritması

Basit yapıda bir model oluşturan sınıflandırma algoritmasıdır ve dolayısıyla hata payı yüksek olabilir. Veri çeşidi başlığının az olduğu durumlarda kullanılabilir (Sumner, Frank, & Hall, 2005, s. 675-683). Uygulama sonrası oluşturulan model Şekil 5'de gösterilmiştir.

```
Class 0 :
        0.99 +
        [musterieski] * 0.64 +
        [iletisim] * -1.09 +
        [total] * 0    +
        [Miktar] * 0
Class 1 :
        -0.99 +
        [musterieski] * -0.64 +
        [iletisim] * 1.09 +
        [total] * 0    +
        [Miktar] * 0
```

Şekil 5: SimpleLogistic Algoritması ile oluşturulan model

Sequential Minimal Optimization Sınıflandırma Algoritması (SMO)

Ardışıl minimal optimizasyon algoritması olan SMO (sequential minimal optimizationalgorithm), verileri ölçeklendirerek sınıflandıran, bir vektör uygulamasıdır (Platt, 1998, s. 42-65). Oluşturulan model Şekil 6'da gösterilmektedir.

```
Machine linear: showingattributeweights, not supportvectors.
      -0.0001 * (normalized) musterieski
+     -0.0002 * (normalized) iletisim
+      0.0008 * (normalized) total
+     -0.0008 * (normalized) Miktar
-      1.0001
Number of kernelevaluations: 23993115 (47.118% cached)
```

Şekil 6: SMO algoritması ile oluşturulan model

Adaptive Boosting Sınıflandırma Algoritması (AdaBoostM1)

İkili sınıflandırma problemlerini çözmek için tasarlanmış, Boosting algoritmaları toplama yöntemi adı verilen bir makine öğrenme algoritmasıdır. (Freund & Schapire, 1996, s. 148-156).

Toplama yöntemi algoritmaları, zayıf öğrenme algoritmalarını birleştirerek güçlü bir öğrenme algoritması oluşturmaktadır. İşlem sırasında modelin daha önce yanlış sınıflandırdığı örnekleri daha iyi sınıflandırması için her güçsüz sınıflandırıcıya bir ağırlık atanmaktadır. Bu nedenle ağırlıklar farklılık göstermektedir. Oluşturulan model katsayılarla desteklenen karar ağacı modeli olarak da adlandırılabilir. AdaBoost algoritması ise en çok kullanılan boosting algoritmaları arasındadır (Özgür & Erdem, 2012, s. 43).

JRipper Sınıflandırma Algoritması (JRIP)

JRIP algoritması IREP algoritmasının optimizasyonu ile oluşturulmuştur. Tekrarlanan veri budama yöntemi ile modeli optimize ederek hata payını azaltan bir yapıdadır. Jrip algoritması optimizasyonun ilk aşamasında başlangıç kural seti tanımlanır. Rastgele veriler işleme alındıkça kural üzerinde düzenlemeler yapılmaktadır. İlk var edilen kurala dönüştürülerek, üretilen yeni kuralların eklenmesi ile genel bir model oluşturulmaktadır (Cohen, 1995, s. 115-123). Bu algoritma ile üretilen model Şekil 7'de gösterilmiştir.

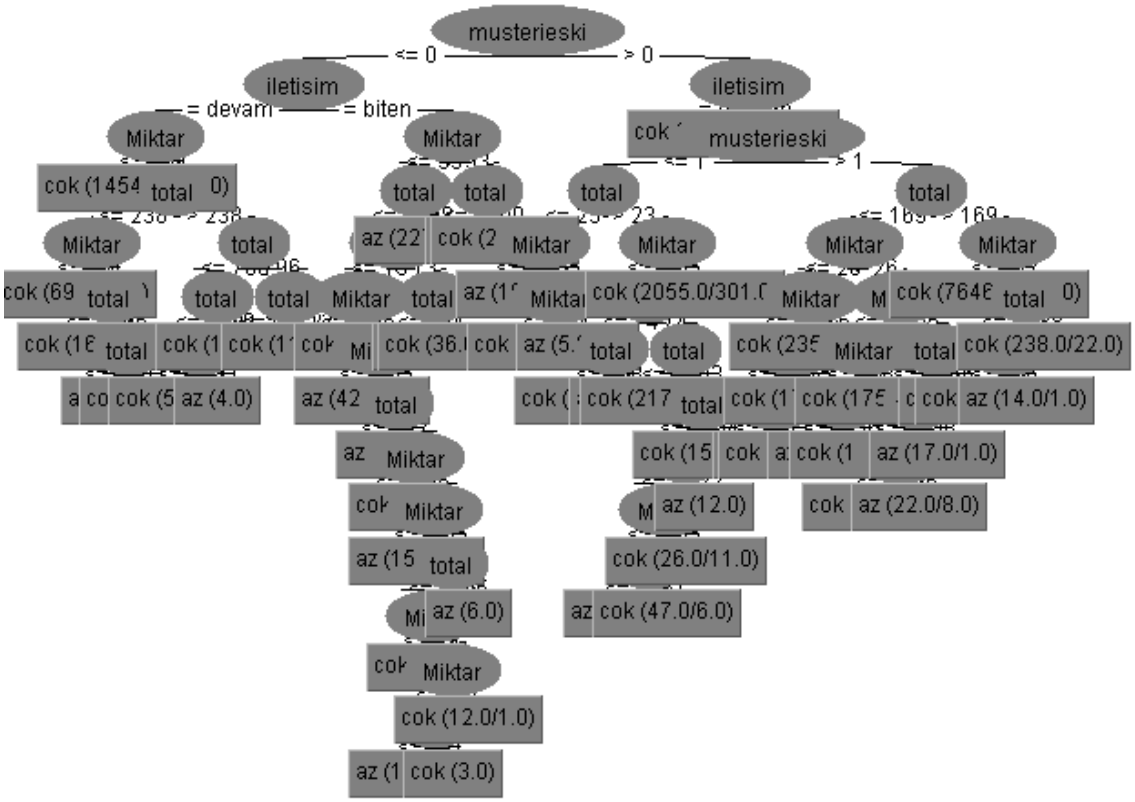
```
(iletisim = biten) and (musterieski=0) =>islem=az (415.0/133.0)
(iletisim = biten) and (musterieski<= 1) and (Miktar >= 25) and  (Miktar <= 27) =>islem=az (19.0/9.0)
(iletisim = biten) and (musterieski<= 1) and (Miktar <= 3) and (total <= 73) =>islem=az (70.0/28.0)
```

(iletisim = biten) and (musterieski<= 1) and (total >= 400) and (total <= 440) and (Miktar >= 3) and (Miktar <= 8) =>islem=az (34.0/10.0)
(iletisim = biten) and (total <= 169) and (Miktar >= 13) and (total >= 145) and (total <= 145) and (Miktar >= 43) =>islem=az (18.0/1.0)
(iletisim = biten) and (Miktar >= 19.5) and (total <= 229) and (total >= 215) and (Miktar >= 80) =>islem=az (17.0/6.0)
=>islem=cok (63385.0/2062.0)

Şekil 7: JRIP algoritması ile oluşturulan model

J48 Karar Ağacı Algoritması

J48 karar ağacı algoritması, temel olarak c4.5 karar ağacı algoritmasını kullanan, modeli en yüksek bilgi kazancına sahip nitelik üzerinden, verilerin bölünmesiyle karara ulaştıran düğümlerin tamamıdır. J48 tüm nitelikler için metrik bilgi kazancını hesaplayarak en yüksek kazanç seviyesinde düğüm belirler. Bu durum ağacın sonuna kadar ardışık devam eder (Quinlan, 1995, s. 235-240).



Şekil 8:J48 algoritması ile oluşturulan model

Şekil 8'deki oluşturulan karar ağacı modeli 105 dal ve 53 yapraklıdır. Veriler üzerinde yapılan çalışmada müşterinin geçmişi J48 karar ağacı algoritması için en büyük düğüm olarak kararlaştırılmıştır.

Bulgular

Uygulama ile kurulabilen modellere ait algoritmalardan başarı değerleri en yüksek sekiz algoritmanın isimleri ve ölçüm değerleri Tablo 5'de belirtilmiştir.

Algoritma	Süre (sn)	Toplam Doğru Tahmin (%)	"Cok" doğru tahmin (%)	"az" doğru tahmin (%)	Doğru tahmin	Yanlış tahmin	Hassasiyet	Kesinlik	ROC Puanı
Meta.Logitboost	**1,38**	**96,404**	**99,8**	**0,2**	**12332**	**460**	**0,956**	**0,952**	**0,841**
BayesNet	**0,5**	96,2398	99,7	**0,3**	12311	481	0,951	0,949	**0,861**
NaiveBayes	**0,14**	95,7161	99	0,1	12244	548	0,943	0,948	**0,829**
SimpleLogistic	19,27	96,3336	**99,8**	**0,2**	12323	469	0,954	0,950	0,834
SMO	9,64	96,1069	100	0	12294	498	0,924	0,942	0,500
AdaBoostM1	1,49	96,153	100	0	12300	492	0,950	0,944	0,828
JRIP	7,14	**96,5056**	**99,8**	**0,2**	**12345**	**447**	**0,959**	**0,954**	0,574
J48	1,54	**96,5134**	**99,9**	0,1	**12346**	**446**	**0,960**	**0,953**	0,812

Tablo 5: Uygulama sonunda en yüksek doğruluk değerine sahip sekiz algoritma

Tablo 5'de WEKA yazılımı ile üretilen 8 farklı modele ait ölçüm kriterinde başarı değeri kendi aralarında yüksek olanlar ayrıca işaretlenmiştir. Bu işaretlemede modelin oluşturulması süresi en düşük, toplam doğru, "cok" ve "az" değeri doğru tahmin yüzdesi, hassasiyet ve kesinlik değerleri ortalamanın üzerinde olan algoritmalar belirtilmiştir. Tablo 5'deki algoritmalar kendi aralarında değerlendirilmiştir. İsimleri işaretlenmemiş algoritmalar tablodan silinerek yeni bir değerlendirme düzeneği oluşturulmuş ve bu düzenek Tablo 6'da gösterilmiştir.

Algoritma	Süre (sn)	Toplam Doğru Tahmin (%)	"Cok" doğru tahmin (%)	"az" doğru tahmin (%)	Doğru tahmin	Hassasiyet	Kesinlik	ROC Puanı
Meta.Logitboost	1,38	96,404	99,8	0,2	12332	0,956	0,952	0,841
BayesNet	0,5							0,861
NaiveBayes	0,14							0,829
JRIP		96,5056	99,8	0,2	12345	0,959	0,954	
J48		96,5134	99,9		12346	0,960	0,953	

Tablo 6: Değerlendirme ölçeklerine göre baskın olan algoritmalar

Tablo 6'da görüldüğü üzere listedeki beş algoritmadan her biri farklı kriterlere göre kullanılabilir. Toplam Doğruluk ve hassasiyet değerlerine göre J48, ROC puanı (Kelly, O'Malley, & Laura, 2007, s. 665-666) değerine göre BayesNet, kesinlik değerine göre ise JRIP algoritması baskın görünmektedir. Tablo 6'daki kriterlere göre algoritmaların almış olduğu değerler birbirlerine çok yakındır. Ancak tüm kriterlerde en başarılı olan algoritma, Logitboost algoritmasıdır.

Sonuç

İşletme için önemli verilerin tıpkı petrol gibi işlenmesi, analizinin yapılması işletme için kıymetli bilgilere dönüştürülmesi küresel rekabet ortamında kritik bir öneme sahiptir. Her şeyin dijitalleştiği günümüzde her geçen gün bu sürecin önemi daha da artacaktır. Dijitalleşme işletmelerin operasyonel süreçlerini kolaylaştırmaktadır. Daha da önemlisi işletmeler dijitalleşme ile rekabetçi olma yolunda büyük aşama kaydederek kârlılığını, verimliliğini ve rekabetçiliğini artırmaktadır.

İşlemenin dijitalleşmesi, kendi ihtiyaçlarına uygun veri tabanlarının oluşturmasına ve depolanan bu verileri bilgiye dönüştürmesine bağlıdır. Yapılan bu çalışma ile işletmelerin veri tabanı oluştururken bu veri tabanlarına girilecek verilerin, kesilen faturalara ve yapılacak tahsilatlara yönelik tasarlandığı bir nevi muhasebe kayıtlarını aşamayan bir yapıya dönüştüğü tespit edilmiştir. İşletmelerin bu kısır döngüyü kırarak işletme amaçlarına uygun veri tabanları kurması ve veri tabanlarında tüm verilerini önemli oranda detaylandırması gereklidir. Çünkü günümüz dünyasında karlılık verilerin ayrıntısında gizlidir.

Yapılan bu çalışmada oluşturulan tüm modellerin birkaç değişkene bağlı olduğu gözlenmiştir. Başarı skorları kıyaslanarak en başarılı algoritma Logitboost algoritması

olarak tespit edilmiştir. Logitboost algoritmasına göre; satışların, müşteriye uğrama sıklığı ve müşteriye yapılan ilk satış tarihinin önceliği ile orantılı olduğu saptanmıştır. Ancak bu algoritmanın diğerlerine önemli ölçüde baskın olmadığı da tespit edilmiştir.

Sonuç olarak işletmelerin dijitalleşmesi kurulan modelin detaylı ve doğruluk değerinin %100'e ulaşması, işletmelerin veri tabanlarında müşteri verilerini daha düzenli, detaylı ve amaca yönelik olarak depolamaları ile doğru orantılıdır. İşletmelerde oluşturulan veri tabanlarının fatura, müşteri ve ürün kategorilerinde detaylı bir sınıflandırmaya ve/veya içeriğe sahip olmasının, bu verilerle daha başarılı analiz yapılmasına olanak tanıyacağı gibi doğru analizlere ve daha isabetli öngörülere ulaşılmasına da yardımcı olacaktır.

Kaynakça

Akbulut, S. (2006). Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu. Ankara: Gazi Üniv. Fen Bilimleri Ens.

Bach, M. P., Juković, S., Dumiči, K., & Šarlija, N. (2013). Business Client Segmentation in Banking Using Self-Organizing Maps. *South East European Journal of Economics and Business* , 32-41.

Baykal, A. (2006). Veri Madenciliği Uygulama Alanları. *Dicle Üniversitesi Ziya* , 95-107.

Bose, I., & Mahapatra, R. K. (2001). Business data mining — a machine learning perspective. *Information & Management* , 211–225.

Bransten, L. (1999). Technology – Power Tools – Looking For Patterns: Data Mining Enables Companies To Better Manage The Ream of Statistics They Collect; The Goal: Spot The Unexpected. *Wall Street Journal* , 16-20.

Brito, P. Q., Soares, C., Almeida, S., Monte, A., & Byvoet, M. (2015). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer-Integrated Manufacturing* , 1-8.

Cohen, W. W. (1995). Fast Effective Rule Induction. *Twelfth International Conference on Machine Learning* , 115-123.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning* , 309-347.

Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research* , 781-800.

Dumouchel, W. (1999). Bayesian Data Mining in Large Frequency Tables, With An. *American Statistician* , 190-196.

eticaretmag. (2014, 07 22). *Vodafone: 'Şirketlerin Dijitalleşme Endeksi Yüzde 48'*. 04 06, 2015 tarihinde eticaretmag: http://eticaretmag.com/vodafone-sirketlerin-dijitallesme-endeksi-h-infografik/ adresinden alındı

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Thirteenth International Conference on Machine Learning* , 148-156.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics* , 337-407.

Gartner. (2012, 10 11). *Gartner Says Worldwide Enterprise IT Spending to Reach $2.7 Trillion in 2012*. Gartner: http://www.gartner.com/newsroom/id/1824919 adresinden alınmıştır

Gartner. (2015, 01 12). *Gartner Says Worldwide IT Spending on Pace to Grow 2.4 Percent in 2015*. Gartner: http://www.gartner.com/newsroom/id/2643919 adresinden alınmıştır

Haber Türk. (2015, 02 26). *'Dijitalleşen işletmeler kârını, verimliliğini ve rekabetçiliğini artırıyor!'*. 04 06, 2015 tarihinde Haber Türk: http://www.haberturk.com/ekonomi/teknoloji/haber/1047285-dijitallesen-isletmeler-krini-verimliligini-ve-rekabetciligini-artiriyor- adresinden alındı

Han, J., & Fu, Y. (1999). Mining Multiple-Level Association Rules in Large Databases. *EEE Transactions on Knowledge & Data Engineering* , 1-12.

Hays, C. L. (2004, 11 04). *What Wal-Mart Knows About Customers' Habits*. The New York Times: http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html?_r=0 adresinden alınmıştır

Jacobs, P. (1999). Data Mining: What General Managers Need to Know. *Harvard Management* , 42-48.

Javaheri, S. H., Sepehri, M. M., & Teimourpour, B. (2014). Response Modeling in Direct Marketing: A Data Mining-Based Approach for Target Selection. Y. Zhao, & Y. Cen içinde, *Data Mining Applications with R* (s. 153-180). Waltham: Elsevier Inc.

John, G., & Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. . *In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo* , 338-345.

Kalaycı, Ş. (2008). *SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri.* Ankara: Asil Yayın Dağıtım.

Kelly, H. Z., O'Malley, A. J., & Laura, M. (2007). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation* , 654-657.

Lausch, A., Schmidt, A., & Tischendorf, L. (2015). Data mining and linked open data – New perspectives for data analysis in environmental research. *Ecological Modelling* , 5–17.

Liao, S.-h., Chen, Y.-j., & Hsieh, H.-h. (2011). Mining customer knowledge for direct selling and marketing. *Expert Systems with Applications* , 6059–6069.

Oğuzlar, A. (2003). Veri Önişleme. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi* , 67-76.

Oz, E. (2006). *Management Information Systems.* Canada: Thomson Course Technology.

Özgür, A., & Erdem, H. (2012). Saldırı Tespit Sistemlerinde Kullanılan Kolay Erişilen Makine Öğrenme Algoritmalarının Karşılaştırılması. *Bilişim Teknolojileri Dergisi* , 41-48.

Platt, J. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Advances in Kernel Methods - Support Vector Learning* .

Quinlan, R. (1995). *C4.5: Programs for Machine Learning.* San Mateo, CA.: Morgan Kaufmann Publishers.

Ramkumar, G., & Swami, A. (1998). Clustering Data Without Distance Functions. *IEEE Bulletin of The Technical Committee on Data Engineering,* , 9-14.

Senga, J.-L., & Chen, T. (2010). An analytic approach to select data mining for business decision. *Expert Systems with Applications* , 8042–8057.

Sharma, S., & Osei-Bryson, K.-M. (2009). Framework for formal implementation of the business understanding phase of data mining projects. *Expert Systems with Applications* , 4114–4124.

Shaw, M. J., Subramaniama, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems* , 127–137.

Song, J., Lu, X., Liu, M., & Wu, X. (2011). A new LogitBoost algorithm for multiclass unbalanced data classification. *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on* , 974 - 977.

Sumathi, S., & Sivanandam, S. (2006). *Introduction to Data Mining and Its Applications.* Tamil Nadu: Springer Science & Business Media.

Sumner, M., Frank, E., & Hall, M. (2005). Speeding Up Logistic Model Tree Induction. *Knowledge Discovery in Databases* , 675-683.

Turgut, H. (2012). Veri madenciliği süreci kullanılarak alzheimer hastalığı teşhisine yönelik bir uygulama. *Yüksek Lisans Tezi* . Isparta: SDÜ Fen Bilimleri Enstitüsü.

Watson, H. J., & Gray, P. (1997). *Decision Support in The Data Warehouse.* USA: Prentice Hall Professional Technical Reference.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques (3rd edition).* San Francisco: Morgan Kaufmann.

Witten, I., & Frank, E. (2000). *E. Frank. Data mining: Practical machine learning tools and techniques with Java implementations.* San Francisco: Morgan Kaufmann.