T.C. İSTANBUL TİCARET ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ TEKNİKLERİ İLE TELEKOM SEKTÖRÜNDE AYRILAN MÜŞTERİ ANALİZİ

Özlem ODABAŞ

Danışman
Yrd. Doç. Dr. Mustafa Cem KASAPBAŞI

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
İSTANBUL - 2017

KABUL VE ONAY SAYFASI

Özlem ODABAŞ tarafından hazırlanan **"Veri Madenciliği Teknikleri İle Telekom Sektöründe Ayrılan Müşteri Analizi "** adlı tez çalışması 17/5/2017 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

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

Onay Tarihi : 08/06/2017

Doç. Dr. Necip ŞİMŞEK
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.

17.05.2017

Özlem ODABAŞ

İÇİNDEKİLER

ÖZET

Yüksek Lisans Tezi

VERİ MADENCİLİĞİ TEKNİKLERİ İLE TELEKOM SEKTÖRÜNDE AYRILAN MÜŞTERİ ANALİZİ

Özlem ODABAŞ

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Yrd. Doç. Dr. Mustafa Cem KASAPBAŞI

2017, 49 sayfa

Veri madenciliği, çok büyük veri kümeleri içinden anlamlı bilgi çıkartma sürecidir. Günümüzde de hızla gelişmekte olan bir tekniktir. Bu teknikte; bir ön işlemden sonra veriler arasındaki ilişki kullanılarak bir model oluşturulur. Son aşamada ise oluşturulan model yorumlanır.

Veri madenciliğinin yaygın olarak kullanıldığı alanlardan biri de ayrılma eğilimi gösteren müşterilerin analizidir. Bu tez çalışmasında, telekom sektörüne ait müşterilerden ayrılma eğilimi gösteren müşteriler analiz edilerek; ayrılma eğilimi gösteren müşteriler tahmin edilmiştir. Ayrılan müşteri analizi için sınıflandırma algoritmaları ve nitelik seçimi teknikleri kullanıldı. Karşılaştırmalar sonucunda %94.41 ile en yüksek doğruluk oranına sahip algoritma, ham dataya uygulanan J.48 algoritması olmuştur.

Anahtar Kelimeler: Ayrılan müşteri analizi, sınıflandırma, veri madenciliği

ABSTRACT

M.Sc. Thesis

CHURN CUSTOMER ANALYSES WITH DATA MINING TECHNIQUES FOR TELECOMMUNICATION INDUSTRY

Özlem ODABAŞ

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

Supervisor: Assist. Prof. Dr. Mustafa Cem KASAPBAŞI

2017, 49 pages

Data mining is the process of obtaining meaningful data from vast amount and very large data sets. It is also a rapidly developing technique nowadays. In this technique; after preporcessing a model is created by using the relationship between data. In the last stage; the generated model is interpreted.

One of the widely used field of data mining is the analysis of customer tending to churn. In this thesis study, the customers of telecom sector churn tendency is estimated while analysing them accordingly. Classification algorithms and attribute selection techniques are utilized for churn analysis. The algorithm which has the highest accuracy rate amongs the compared algorithm is determined as J48 with %94.41 accuracy.

Keywords: Churn analysis, classification, data mining

TEŞEKKÜR

Tez çalışmamda bana yardım ve desteğini hiçbir zaman esirgemeyen tez danışmanım, değerli öğretmenim Sayın Yrd. Doç. Dr. Mustafa Cem KASAPBAŞI' na ve doğduğum günden bu yana yanımda olan aileme gösterdikleri maddi-manevi destekleri, anlayışları ve teşvikleri için içtenlikle teşekkür ederim.

Özlem ODABAŞ
İSTANBUL, 2017

ŞEKİLLER

ÇİZELGELER

Sayfa

SİMGELER VE KISALTMALAR

MİY Müşteri İlişkileri Yönetimi

ID.3 Iterative Dichotomiser 3

WEKA Waikato Environment for Knowledge Analysis

SQL Structured Query Language

CLV Customer Lifetime Value

CSR Certificate Signing Request

TP True Positive

FP False Positive

TN True Negative

FN False Negative

MD Maryland

MI Michigan

NJ New Jersey

TX Texas

1. GİRİŞ

Günümüzde teklonojinin gelişmesi ile birlikte firmalar arasındaki müşteri rekabeti artmıştır. Teknoloji, yaşam koşulları gibi nedenlerle müşterilerin satın alma davranışları, eğilimleri, beklentileri büyük ölçüde değişmiştir. Firmalar müşteri edinme, müşteri sadakati kazanma, şirket karlılığını artırma gibi düşüncelerini iyi bir müşteri ilişkileri yönetimi ile elde edebilir.

Firma sahipleri, rekabetin olduğu ortamda müşteri edinmek, onların beklentilerine cevap vermek, müşteri davranışlarına uygun hareket etmek için müşteri ilişkileri yönetimi kavramını iyi bilmeli ve yorumlamalıdır. Bu kavramın, hem müşteri payını hem de pazar payını içerdiği bilinmelidir. Bir müşteriye birden fazla ürün satmanın yolları, karşılıklı iyi ilişkiler kurarak müşteri sadakati kazanmayı ve müşteriyi aktif hale getirmek için çalışılmalıdır. Bu amaç doğrultusunda müşterilerin dikkatini çekecek kampanya ve pazarlama yöntemleri belirlenmelidir. Çünkü firmalar için müşteri önemlidir. Firmayı ayakta tutan, firmanın devamlılığını sağlayan müşterilerdir. Bu nedenle mevcut müşterileri elde tutmalı ve ek müşteri kazanımı yoluna gidilmelidir.

Gelişen teknoloji ile birlikte artık daha büyük verilerin veritabanlarında tutulabilmesi mümkün olmuştur. Büyük veriler içersinden anlamlı veriler elde edilmek istendiğinde veri madenciliği yöntemi işin içine girer. Veri madenciliğinde kullanılan yöntem ve araçlarla veriler, uygun şekilde modellenerek işe yarar bilgiye dönüştürülür.

Çalışmanın ikinci bölümünde veri madenciliği tanım ve süreçleri üzerinde durulmuş, üçüncü bölümünde müşteri ilişkileri yönetiminden ve ayrılma eğiliminden bahsedilmiş, dördüncü bölümünde yapılan uygulama anlatılmış ve son bölümde de sonuçlar belirtilmiştir.

2. LİTERATÜR ÖZETİ

Veri madenciliği, gelişen teknoloji ve veri boyutunun büyümesiyle beraber önemli hale gelmiştir. Telekom sektöründeki firmalar daha çok müşteriye hizmet etmeye başlamışlardır. Müşteri, firmalar için ayakta durmalarını sağlayan sermaye olarak görülmeye başlandı. Bu gereklilik ile beraber müşteri ayrılma (churn) analizi çalışmaları önem kazanmıştır. Ayrılma eğilimi, müşterinin mevcut hizmetini bırakıp, rakip firmanın hizmetinden yararlanmaya başlaması olarak tanımlanabilir. Müşteri davranışlarını, alışveriş özelliklerini, hizmet alma eğilimlerini konu alan belli başlı çalışmalar incelenmiştir.

"Veri Madenciliği' nde yapısal olmayan verinin analizi : Metin ve Web Madenciliği" adlı makalede; web madenciliği, yapısal olmayan verileri yapısal şekle dönüştürür ve web ile ilişkili olan verinin analizini içerir. Metin madenciliği, çok büyük belgelerin analizi demektir. Bu yöntemlere dayanarak telekom sektöründe ayrılan müşteri analizi yapılmıştır. Bu yöntem için karar ağacı algoritmalrından C5.0 kullanılmıştır. Bu algoritmaya göre müşterinin ayrılma eğilimi göstermesindeki sebep yurtdışı dolanımı (roaming) 'dir. Sonuç olarak, metin ve web madenciliğinin yapısal olmayan verileri yapısal hale dönüştürüp,analiz edilmesi büyük başarı olarak öngörülmüştür (Dolgun vd., 2009).

"Veri Madenciliği Teknikleri İle Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Bölümlenmesi" adlı makalede; kozmetik firmasına ait müşteriler analiz edilerek ayrılma eğilimi gösteren müşteri grubu için uygun stratejiler geliştirilmiştir. Müşteri analizi için sınıflandırma ve kümeleme algoritmalarından yararlanılmıştır. J48 karar ağacı algoritması doğruluk oranı en yüksek çıkmıştır. Yapılan çalışma ilk defa kozmetik şirketine uygulanmış olup, analiz yöntemleri üretim açısından yol gösterici olacaktır (Aydoğan vd. 2009).

"Veri Madenciliğinde Sepet Analizi İle Tüketici Davranışı Modellemesi" adlı çalışmada, büyük market zincirine ait veriler birliktelik kuralına göre analiz edilmiştir. Müşterilerin hangi ürünle beraber, hangi ürünü de aldıkları tespit edilmiştir. Müşterilerin satınalma davranışlarına göre market kar marjını arttırma

amaçlı; raf düzenlemeleri, ürünlerin birbiriyle ilişkileri, bazı kampanyalar ve indirim çalışmalarına başlar (Timor ve Şimşek, 2008).

"Telekomünikasyon Sektöründe Müşteri Ayrılma Analizi" adlı çalışmada, telekom firmasının ayrılma eğilimi gösteren müşterileri belirlenmiştir. Analiz için çalışmada Lojistik Regresyon ve karar ağacı algoritmalarından yararlanılmıştır. Elde edilen sonuçlara göre firma müşteleri elde tutmak için onlara özel pazarlama stratejileri geliştirmeyi hedeflemektedir (Şimşek ve Umman, 2010).

"Yapay Sinir Ağları ve Sosyal Ağ Analizi Yardımı İle Türk Telekomünikasyon Piyasasında Müşteri Kaybı Analizi" adlı çalışmada, telekomünikasyon sektöründe müşteri kaybı Yapay Sinir Ağları algoritması ile analiz edilmiş ve Sosyal Ağ Analizi ile de müşteri iletişim ağı analizi yapılarak ayrılma eğilimi olan müşterilerin ağdaki konumları ve etkileri incelenmiştir (Gülpınar, 2015).

"Churn in Telecom Dataset" adlı çalışmada, telekom sektörüne ait veriler analiz edilmiştir ve ayrılma eğilimi gösteren müşteriler J.48 karar ağacı algoritmasına göre tahminlenmiştir. Tezimde, bu çalışmada analiz edilen veri kümesinden faydalandım. Bu çalışmada yapılanlardan farklı olarak, veri kümesine nitelik seçimi algoritmaları uygulanmış olup sonuca en çok etki eden nitelikler belirlenmiştir. Ek olarak, veriler deney kümelerine ayrılarak, her deney kümesi için farklı sınıflandırma algoritmaları uygulanmıştır ve bu algoritmalar karşılaştırılmıştır (Vis ve Zwet, 2009).

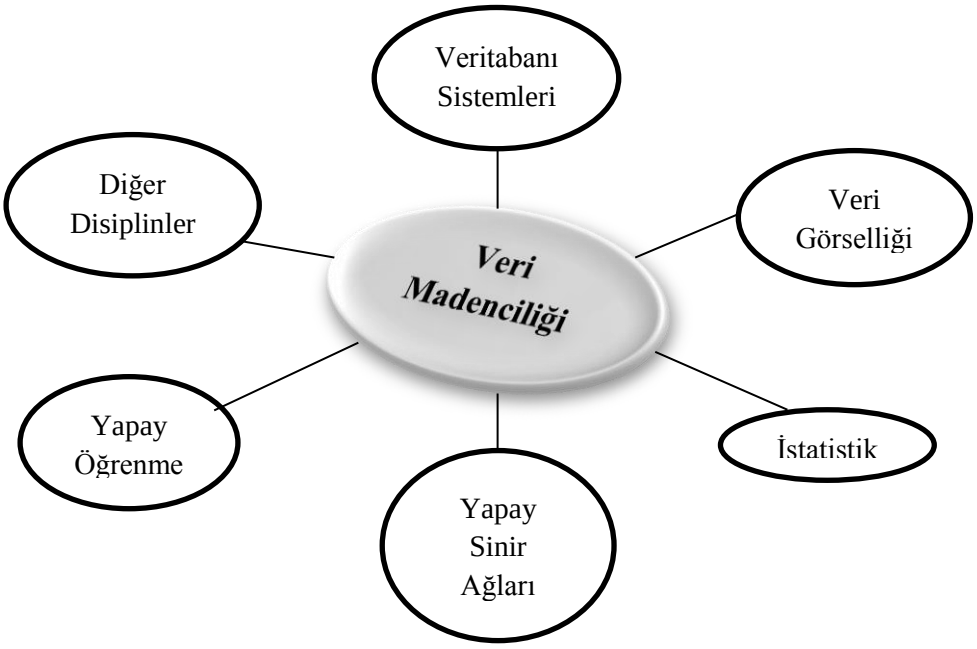
3. VERİ MADENCİLİĞİ TANIMI

Veri madenciliği tanımlayacak olursak, veriden bilgiye giden sürecin bütünüdür. Büyük karmaşık veri yığını içinden, bir sonraki adım veya adımların tahmin edilmesini sağlayan anlamlı kural ve bilgilerin, çeşitli yazılımlarla bulunup analiz edilmesidir.

Veri madenciliğinin tanımında en temel süreçleri şu şekilde sıralayabiliriz:

1- Her çeşit veri yığınını kullanarak anlamlı ve yararlı bilgiler elde edilebilir.
2- Karmaşık ve büyük verilerle çalışır.
3- Birçok endüstri alanında kullanılmaktadır.
4- Problemlerin türüne göre değişen çözüm yöntemleri vardır.
5- Veri analizlerinde otomatik veya yarı otomatik olarak çalışan çözüm araçları kullanır.
6- Daha önceden bilinmeyen fakat; doğrulanabilir, etkinleştirilebilir bilgi arar.

Veri analizi tekniğinin sahip olduğu uygulama alanları geniştir. Bu alanlara; Veri Tabanı Sistemleri, Veri Görselliği, Yapay Sinir Ağları, İstatistik, Yapay Öğrenme, vb. gibi disiplinleri sıralayabiliriz (Şekil 3.1).



Şekil 3.1. Veri Madenciliği Uygulama Alanı

Bu zamana kadar veri madenliği tanımı ile bir çok literatürde farklı tanımlara rastlanmıştır. Veri madenciliği ile ilgili bu tanımlardan bazıları şu şekildedir;

- Veri madenciliği, çok büyük miktardaki verilerin içindeki ilişkileri inceleyerek aralarındaki bağlantıyı bulmaya yardımcı olan ve veri tabanı sistemleri içerisinde gizli kalmış bilgilerin çekilmesini sağlayan veri analizi tekniğidir (Kalikov, 2006).

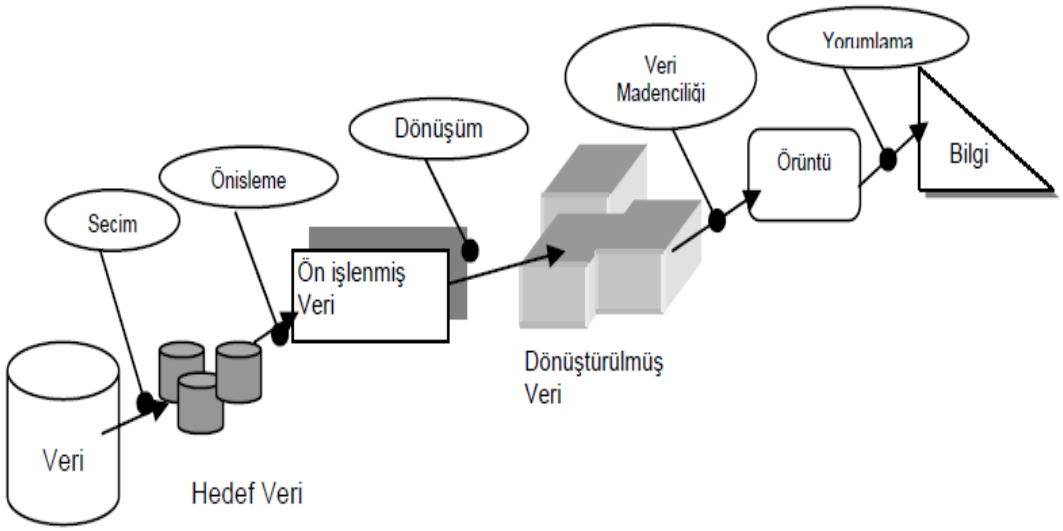
- Veri madenciliği, öncelikle bilinmeyen desenlerin ortaya konması amacıyla bilimsel ve teknik veri araştıran, veri tabanındaki bilgi keşfi süreçlerinden biridir (Rokach ve Maimon, 2005:2).

- Veri madenciliği, veri ambarlarındaki tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş verileri ortaya çıkarmak, bunları karar vermek ve gerçekleştirmek için kullanma sürecidir (Swift, 2001).

Tanımlardan da anlaşılacağı gibi; veri madenciliği, çok büyük boyutlu verilerin depolandığı veri tabanlarından, problemi çözüme kavuşturacak, değerlendirme ve gelecek ile ilgili tahminler yapmamızı sağlayacak yararlı, anlamlı doğru bilgilere ulaşmak ve bu bilgileri kullanmaktır.

3.1. Bilginin Keşfi

Veri yığınlarından bilgi keşfi, sürekli tekrar eden bir süreçtir. Veri madenciliği, bir nevi bilgi keşfinin bir parçasıdır. Veri yığınları arasında anlamlı veriyi ortaya çıkarmanın yanı sıra, dönüştürülmüş veriden örüntüleri temizleyip yorumlamak ve verileri bir üst adıma hazırlamak da bilgi keşfi sürecinin bir parçasıdır (Çelik ve Akçetin, 2015). Bilginin keşif adımları Şekil 3.2'de gösterilmiştir.



Şekil 3.2. Bilginin Keşfi

Bilgi keşfi aşağıdaki adımlardan oluşmaktadır:

1. Uygulama alanının incelenmesi: Analiz edilecek veri ve analiz amacının belirlenmesi.

2. Amaca uygun veri kümesi yaratma: Verinin hangi veritabanında yapılacağını belirterek, veri seçmek, alt veri oluşturma.

3. Veri ayıklama ve önişlem: Gürültülü ve tutarsız, anlamsız verileri temizleme.

4. Veri azaltma ve veri dönüşümü: Analize etki edecek özellikleri belirleme, özellikler arasında ilişki kurma ve veri dönüştürme adımı ile veriyi azaltarak anlamlı hale getirme.

5. Veri madenciliği tekniği seçme: Doğruluk oranı yüksek olan algoritma seçilir ve seçilen algoritma ile uygun modeli oluşturulur.

6. Veri madenciliği algoritmasını seçme: Hangi tekniğin uygun olduğu belirlendikten sonra veri modeline en uygun algoritma seçilir. Veriler arasında ilişki, sınıflandırma bu şekilde sağlanır.

7. Veri Madenciliği: Dönüştürülmüş veriden, örüntüleri yakalayabilmek için akıllı metotlar uygulanır.

8. Örüntü Değerlendirme: Bu adımda elde edilen örüntüler görselleştirilir, fazla ve gerekli ilişi kurulmayan veriler elenir ve yararlı olan veriler kullanıcılar tarafından anlaşılacak ifadelere dönüştürülür.

9. Keşfedilen bilginin kullanılması: Elde edilen bilgi, kirli verilerden arındırılmış ve analiz tamamlanmıştır. Elde edilen anlamlı veri ilgili kişi veya kurumlara sunulur.

3.2. Veri Madenciliğinin Tarihi

İnsanlar her zaman geçmişten günümüze bütün verileri harmanlayıp bunlardan yaralı bilgiler elde etmeye çalışmıştır. Elde edilen bu bilgilerin taşınması için çeşitli donanımlar oluşturulmuştur. Zamanla teknolojinin de gelişmesiyle her alanda veriler toplanmaya başlanmıştır. Veri madenciliğinin gelişimi kronolojik olarak aşağıdaki gibi aktarılmış ve son olarak Çizelge 3.1'de özetlenmiştir.

1950 lerde matematikçiler mantık ve bilgisayar bilimleri üzerinde çalışarak yapay zeka ve makine öğrenme tekniklerini bulmuşlardır.

1960 lı yıllarda regresyon analizi, sinir ağları vb. yeni algotmalar keşfedilmiştir. Bu metotlar veri madenciliğinin temelini oluşturmuştur. Ayrıca, döküman ve bilgilerin saklanması için veritabanı sistemleri geliştirilmiştir.

1970 – 1990 lı yıllarda, bilgisayar teknikleri geliştirilmiş, yeni algoritmalar geliştirilmiştir.

1990 yıllarında bilgi keşfinin ilk adımları atılmış, büyük veriler için veri ambarları oluşturulmuştur. Gelişen teknoloji ile beraber veri madenciliği yaygın olarak kullanılmaya başlanmıştır.

Çizelge 3.1. Veri Madenciliği Gelişim Süreci (Çıngı,2006)

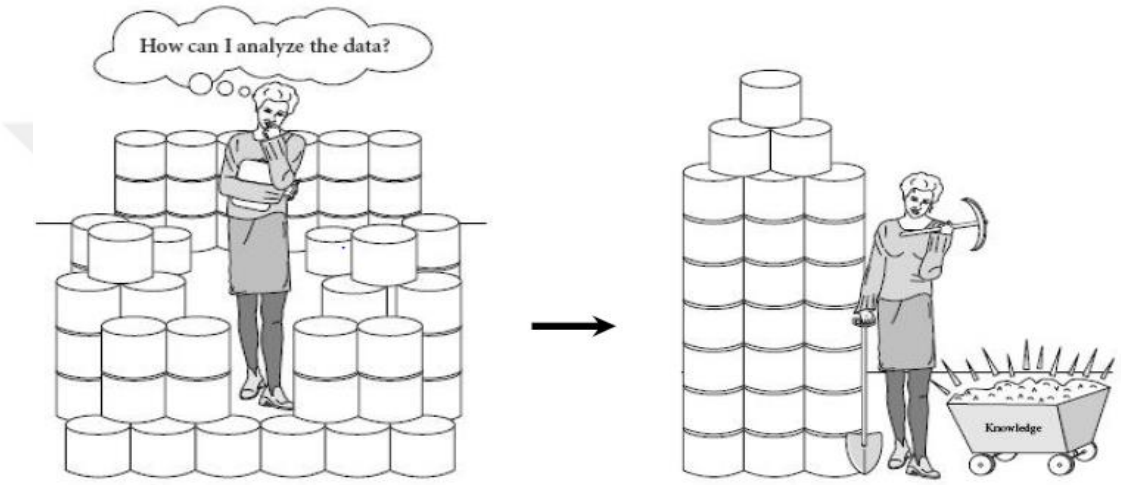
Gelişim Adımları	Cevaplanan Karar Problemi	Kullanılan Teknolojiler	Ürün Sağlayıcıları	Karakteristikler
Veri Toplama (1960' lar)	"Benim toplam karım geçen 5 yılda ne kadardı?"	Bilgisayarlar, Teypler, Diskler	IBM, CDC	Geriye dönük, statik veri dağıtımı
Veri Erişimi (1980' ler)	"Türkiye' de geçen Mart ayında birim satışları ne kadardı?"	İlişkisel Veritabanları, SQL, ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Kayıt düzeyinde geriye dönük, dinamik veri dağılımı
Veri Ambarlama ve Karar Destek Sistemleri (1990'lar)	"Türkiye'de geçen Mart ayında birim satışları ne kadardı?	OLAP, Çok Boyutlu Veritabanı Sistemleri, Veri Ambarları	Pilot, Comshare, Arbor, Cognos, Microstrategy	Çoklu düzeylerde, geriye dönük dinamik veri dağıtımı
Veri Madenciliği (Bugün)	"Gelecek ay Boston' daki birim satışlar muhtemelen ne olabilir, niçin?"	İleri düzeyde algoritmalar, çok işlemcili biligisayarlar, büyük veritabanları	Pilot, Lockheed, IBM, SGI, SPSS, SAS, Microsoft vs.	Geleceğe dönük, proaktif enformasyon dağıtımı

3.3. Veri Madenciliğinin Gelişmesinin Nedenleri

Veri madenciliğinin günümüzde gelişmesinin nedenleri aşağıda belirtilmiştir.

- Bilgisayarların kullanımı teknolojinin gelişmesiyle beraber daha da arttı ve gerek duyulan verilerin miktarında da artış oldu. Veritabanlarında tutulan yüksek boyutlu verilerin saklanma isteği veri madenciliğinin önemini arttırdı.
- Veri toplama yöntemlerinin (veritabanları, disk) gelişmesi ile daha çok verilerin elde olması
- İş dünyasının, bilim dünyasının ve toplumsal uygulamaların büyük veri kaynaklarına ihtiyaç duyması.

- Kişiselleştirilmiş ürünler, CSR yönetimi gibi alanlarda ticari rekabetin ve baskının artması
- Gelişen veri madenciliği teknoljileri birçok endüstrilerde kullanılmaya başlandı. Algoritmalar ile daha büyük ve daha karmaşık veriler daha doğru bir şekilde analiz edilmektedir.
- Veri madenciliği arayüzleri standartlaştığı için yazılım geliştiriciler daha iyi uygulama geliştirebilmektedirler.



Şekil 3.3. Neden Veri Madenciliği

Veri madenciliğinin amacı; büyük veriler arasında arama yapmak değil, aradığımız veriyi anlamlı bilgi şeklinde sonuçlandırmaktır (Şekil 3.3).

3.4. Veri Madenciliğinin Yararları

Veri madenciliğinin yararları aşağıdaki gibi sıralanabilir:

- Müşterilerin eğilimlerinin, karakteristik özelliklerinin daha iyi analiz edilmesini sağlayabilir.
- Müşteri bölümlendirilmesi yapılıp her grup müşteri için farklı davranış modelleri oluşturulur. Bu şekilde olası riskler minimize edilerek müşteri bilgilendirilir.

- Veri madenciliği tahminleme yöntemi ile firmanın / kurumun stok fiyatı, kar oranı vs. tahmin edilebilir.
- Müşteri kitlesi göz önüne alınarak çeşitli kampanyalar düzenlenebilir, firmanın kar oranı artırılabilir. Müşteri davranışları, kampanya şartlarını şekillendirir.
- Mevcut müşteriler üzerinde çapraz satış yöntemi uygulanarak satış oranı arttırılabilir.
- Veri madenciliği sistemi ile firma müşterilerini daha iyi tanımak için müşteri ilişkileri yönetiminde düzenleme ve yenilikler yapabilir. Böylelikle, firma müşterilerinin ihtiyaçlarını bilerek onların ihtiyaçlarına daha kolay cevap verebilir. Hem müşteri kazanımı olur hem de firmanın karlılık oranının artmasıyla prestiji de artar.
- Rekabet ortamında firmaların doğru ve hızlı karar vermesini sağlayabilir.
- Firmalar, veri analizi ile müşterilerini kampanya ve hizmetler hakkında bilgilendirebilir.
- Veri madenciliği yöntemleri ile firmalar satış politikalarını geliştirebilir, düzenleyebilir.
- Veri analizi sonucunda, anlamsız verilerden anlamlı veriler elde edilip yorumlanabilir.
- Sağlık sektöründe, test sonuçlarından elde edilen veriler ile kanserlerin ön tanısı belirlenip erken tedavi süreci kazanılabilinir, kalp krizi verileri kullanılarak kalp krizi geçirme riski belirlenerek risk taşıyanlara acil servislerde uygun tedavi ve risklerin önlenmesi sağlanabilir.
- Öğrenim alanında, öğrencilerin bilgileri analiz edilerek başarı ve başarısızlık nedenleri belirlenebilir ve başarının arttırımı için neler yapılması gerektiği belirlenebilir.
- Online e-ticaret sitelerinden kullanıcıların analizi yapılabilir.
- Haber, e-posta vb. dökümanlar arasında elle çalışma gerekmeden, aralarındaki benzerlikler hesaplanabilir.
- Firma mevcut kaynaklarını en verimli şekilde kullanır.
- Geçmiş veriler ile mevcut veriler analiz edilerek geleceğe yönelik tahminlerde bulunur.

- Risk Analizi ile kalite kontrolü, rekabet analizi, sahtekarlıkların saptanması yapılabilir.

3.5. Veri Madenciliği Sürecinde Karşılaşılan Sorunlar

Anlamlı veriye erişmek için kullanılan veri madenciliği sürecinde birtakım sorunlar yaşanmaktadır. Bunlar;

- Güvenlik ve Sosyal Haklar
- Kullanıcı Arabirimi
- Veri MadenciliğindeKullanılanYöntem
- Başarım ve Ölçeklenebilirlik
- Veri Kaynağı

Günümüzde güvenlik en önemli unsur sayılmaktadır. Güvenlik ve sosyal haklar ile kişilere ait elde edilen bilgilerin toparlanıp, kişilere ait bilgilerin onlardan habersiz ve izinsiz olarak kullanılması, çeşitli veri madenciliği yöntemleri ile elde edilen sonuçların izinsiz dağıtılması, açıklanması, gizlilik vb. sorunlar hala çözümlenmemiştir.

Veri madenciliği uygulamaya yönelik çözümler ürettiğinden, yaygın bir kullanıcı arabirimine sahip değildir.

Kullanılan yöntem de çok önemlidir. Kullanılan yönteme göre sonuçlar arasında farklılılar olabilir. Geçerli yöntemin ne olduğuna karar vermek için uzman kişilerin görüş ve açıklamaları değerlendirilmelidir. Sonuçlar uzman kişi tarafından verilmelidir.

Başarım ve ölçeklenebilirlik konusu kişiden kişiye değişiklik gösterebilir. Kuralların geçerliliği konusunda net bir söylem yoktur. Bir uygulama için %90 başarı iyi bir sonuç sayılabilirken, başka bir uygulama alanında kötü bir sonuç sayılabilir.
Veri kaynağıdan elde edilen ham bilginin güvenirliliği konusunda, bilgi ihlali gibi nedenler dolayı bir doğrulama yapılamıyor.

3.6. Veri Madenciliğinde Karşılaşılan Problemler

Veri madenciliğinde ham veri, veritabanlarından elde edilir. Buradaki verilerin eksiksiz, net, anlamlı veri olmaması durumunda sorunlar olur. Büyük hacimli verilerin olduğu ortamlarda büyük sorunlar ortaya çıkar. Veri madenciliğinde sistemlerinde eksik, gürültülü, boş, anlamsız, aykırı verilerin olduğu ortamda yanlış sonuçlar elde edilebilir. Bu sebepten dolayı, model seçiminden önce anlamlı veri bütünlüğü sağlanmalıdır. Veri bütünlüğü sağlanırken karşılaşılan sorunların çözümlenmesi gerekmektedir. Başlıca karşılaşılan sorunları şu şekilde sıralayabiliriz:

- Sınırlı Bilgi
- Artık Veri
- Boş Veri
- Dinamik Veri
- Belirsizlik
- Gürültü
- Eksik Veri
- Ebat, güncellemeler ve konu dışı sahalar
- Veri tabanı boyutu

3.7. Veri Madenciliğinin Çözüm Ürettiği İş Problemleri

3.7.1. Ayrılma analizi

Hangi müşteriler rakiplere gitmeye daha çok eğilim gösteriyor, bu soruya cevap aranır. Birçok sektör bu risk ile karşı karşıya kalmaktadır. Ayrılma analizi şirketlere, müşterilerinin neden rakiplere yöneldiğini anlamalarını sağlar. Bu şekilde şirket, müşteri ilişkilerini yeniden düzene sokar ve müşteri bağlılığını arttırır (Uslu, 2011).

3.7.2. Çapraz satış

Müşterilerimiz daha çok hangi ürünleri alıyor ya da daha çok hangi hizmetten yararlanıyor, sorularına cevap aranır. Özellikle e-ticaret, web üzerinden yapılan satışlarda bu yöntem çok etkilidir. Müşteri bir ürün veya hizmet aldığında ek olarak şunu da alır şeklinde çapraz satış yöntemi ile firma karlılığını arttırır (Uslu, 2011).

3.7.3. Fraud algılama

Talepte bulunan müşteri acaba sahtekar mıdır, sorusuna cevap aranır. Günlük yoğun olarak birçok talebi Fraud Algılama işleme alan banka ve şirketler, her talebin gerçekliğini ayrıştıramaz. Bu ayrıştırmayı yapmak için veri madenciliği yöntemleri faydalı olabilir (Uslu, 2011).

3.7.4. Risk yönetimi

Müşterinin bu talebini onaylamalı mıyım, sorusuna cevap aranır. Bankacılık ve sigortacılık sektöründe sık yaşanan sorunlardan birisidir. Veri madenciliği yöntemleri ile müşterinin güvenirliliği analiz edilerek talebi işlemi alınır. Böylelikle, doğru kararlar verilerek risk azaltılmış olur (Uslu, 2011).

3.7.5. Müşteri kümeleme

Müşterim kimler, sorusuna cevap aranır. Müşteri kümeleme yöntemi ile müşteriler kategorilendirilir, müşteri profilleri belirlenir ve her kategorideki müşterilere uygun stratejiler geliştirilir.

3.7.6. Hedeflenen reklamlar

Müşteriye hangi reklamları göstermeliyim, sorusuna cevap aranır. Web üzerinden satış yapan şirketler, müşterilerin en çok ziyaret ettiği sayfaları ve içerikleri veri madenciliği yöntemleri ile analiz ederek, müşteriye uygun ürünlerin veya hizmetlerin reklamını gösterir.

3.7.7. Satış tahmini

Gelecek ay kazancım ne kadar olacak, stok miktarım ne olacak, kaç adet ürün satacağım gibi sorulara cevap aranır. Geçmiş ve mevcut durum analizi yapılarak gelecek ayki durumun tahmini yapılabilir.

3.8. Veri Madenciliği Kullanım Alanları

Teknolojinin gelişmesine bağlı olarak ve çağın gereksinimden kaynaklı, veri boyutunda hızlı artış söz konusudur. Veri madenciliği günümüzde karar verme sürecini barındıran pek çok alanda uygulanmaktadır. Veri madenciliğinin kullanıldığı alanlar; özellikle pazarlama, bankacılık, sigortacılık, tıp, telekominikasyon vb.

sektörlerinde yaygın olarak kullanılmaktadır (Baykal, 2006). Kullanım yerleri maddeler halinde gösterilmiştir:

Pazarlama

- Satın alma yöntemlerinin belirlenmesi
- Yeni müşterilerin kazanılması, mevcut müşterilerin elde tutulması
- Müşteri ilişkileri yönetimi
- Müşteri profili değerlendirmesi
- Satış tahminin yapılması

Bankacılık

- Dolandırıcılıkların tespiti
- Harcamalarına göre müşteri sınıflarının belirlenmesi
- Kredi taleplerinin doğru değerlendirilmesi
- Müşteri İlişkileri Yönetimi
- Müşteri Kaybı Analizi

Parakendecilik

- Çapraz satış yöntemi ile eldeki ürünlerin benzerliklerini tahmin etmek
- Müşteriler için birçok ürün bulunması ile ürün satışları arasındaki bağlantıyı tespit etmek

Finans

- Nakit para akışının incelenmesi ve kestirimi
- Talep incelenimi ve kestirimi
- Zaman serilerinin incelenmesi

Elektronik Ticaret

- WEB sayfalarına yapılan ziyaretlerinin çözümlenmesi
- Kullanıcı davranışlarına göre web sitesinin yenilenmesi

Telekomünikasyon

- İletişim ağlarında sorun yaşanan bölgelerin tespiti
- Kaçak hat kullanımı varsa bunların belirlenmesi
- Müşteri profiline göre yeni hizmetlerin sunulması

Tıp

- DNA içersindeki genlerin sıralarının belirlenmesi
- Hastalığa ait yol haritasının hazırlanması
- Hastalık teşhişleri

Bu maddelerin dışında, veri madenciliğinin kullanıldığı alanlardan bazıları şunlardır:

- Nakliyat ve Ulaşım
- Turizm ve otelcilik sektörü
- Eğitim sektörü
- Bilim ve mühendislik
- Meteraloji
- Sistem Yönetimi ve Yardım Masası
- Borsa
- Üretim ve Planlama
- Taşımacılık – Ulaşım

3.9. Veri Madenciliğinde Kullanılan Yöntemler ve Algoritmalar

Veri madenciliğinin belirli adımlarında birtakım yöntemler kullanılır. En uygun modelin belirlenmesi için eldeki anlamlı veri kümesi çeşitli algoritmalara sokulur. Tez çalışmasında kullanılan algortitmaları inceleyelim:

3.9.1. Naive bayes algoritması

Bayes teorimini esas alır. Rassal değişkenler ile olasılıklar arasındaki ilişkiyi ifade eder. Her zaman için en yüksek olasılıklı durum hedef sınıf seçilir. $P(C_i \mid X)$ olasılığı en büyük olmalıdır.

- Sınıflandırma yapılırken en yüksek olasılıklı durum hedef sınıf olarak seçilir.
- Naive bayes formülü, Bayes teoreminden türetilebilir.

$$P(X \mid C_i) = \frac{P(X \mid C_i)\, P(C_i)}{P(X)} \quad (3.1)$$

3.9.2. Lojistik regresyon

Lojistik regresyonda ortaya çıkan model, ortaya çıkacak riski 0 ile 1 arasında tahmin etmeyi amaçlar. Bu nedenle Lojistik fonksiyonu, her zaman 0 ile 1 arasında değişim aralığına sahip olması tercih edilir (Akyol, vd, 2016). Formülü aşağıdaki gibidir :

$$\boldsymbol{\Pi(x) = \frac{e^{\beta 0+\beta 1X1+\cdots\ldots\ldots+\beta pXp}}{1+e^{\beta 0+\beta 1X1+\cdots\ldots\ldots+\beta pXp}}} \quad (3.2.)$$

Lojistik regresyonda en az değişken kullanılarak değişkenler arasındaki ilişkiyi en iyi şekilde tanımlamak ve model kurmak ilk amaçtır.

İncelenen bir olaya ait olasılığın, diğer olayların olasılığına oranına Odds Değeri denir. Lojistik regresyon analizi, odds değerini temel alır.

Odds Değeri = $\boldsymbol{p/(1-p)}$ (3.3.)

Odds oranı (OR), x = 1'in odds değerinin x = 0 odds değerine oranıdır ve $\boldsymbol{e^{\beta 1}}$ 'e eşittir.

OR = $\boldsymbol{e^{\beta 1}}$ (3.4.)

Odds oranı, özellikle epidemiyolojide yaygın olarak kullanılır.

3.9.3. K- Star algoritması

K – star algoritması, örnek tabanlı bir sınıflandırıcıdır. İki nokta arasındaki uzaklığın veya benzerliğin belirlenmesinde kullanılmaktadır. Entopik uzaklık ölçüsü kullanılır. K-star algoritması iki özelliği birbirine bağlayan en kısa uzaklık olarak Kolmogorov mesafesini dikkate almaktadır (Cihan, Kalıpsız, 2015). K-star fonksiyonu aşağıdaki gibi ifade edilir (Çölkesen, Kavzoğlu, 2011).

$$\boldsymbol{K} * \left(\frac{\boldsymbol{b}}{\boldsymbol{a}}\right) = -\mathbf{log}(\mathbf{2})\, \boldsymbol{P} * (\frac{\boldsymbol{b}}{\boldsymbol{a}}) \qquad (3.5.)$$

3.9.4. Ardışık minimal optimizasyon algoritması (SMO)

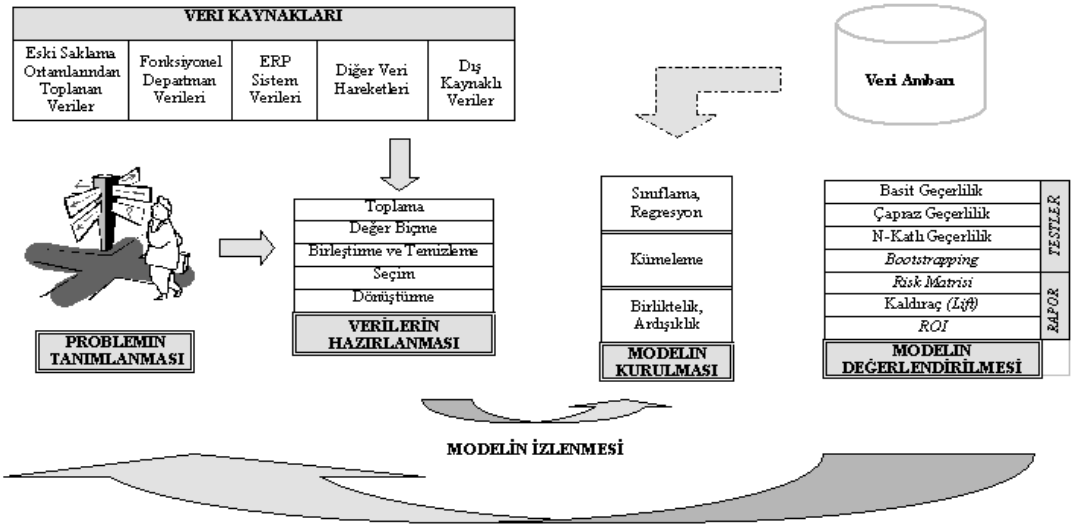
Destek vektörü kullanan bir algoritmadır. Veriyi, birden fazla kümeye bölmeye yarar. Destek vektörü, sınıflandırıcıyı eğitmek için SMO algoritmasını uygular. Bu şekilde kayıp değerler arındırılır, yeni veri ile değiştirilir ve öznitelikler ikili özniteliklere dönüştürülür. Ayrıca öznitelikler, elde tanımlanmış olan değerler ile normalize edilir.

3.9.5. J.48 algoritması

C4.5 algoritmasının bir üst versiyonudur. Karar ağacı algoritmaları, durumlar ve örnek kümeleri ile başlar. Olayların sınıflandırılmasında ağaç yapısı oluşturulur. Ağaçlar, düğüm ve yapraklardan oluşur. Düğümden sonra dallanmalar başlar. Her bir durumun olasılığı dallanmada belirtilir. Karar ağacının her bir düğümü test verisi içerir. Test sonuçlarına göre sonraki dallanmanın hangisi olacağına karar verilir. Yapraklar da sınıf bilgisi içerir (Cihan ve Kalıpsız, 2015). Bu algoritma, sınıf bilgilerini kullanarak sınıflandırma yapar.

3.10. Veri Madenciliği Süreçleri

Veri madenciliği tekniğinin gelişmesi ile bu yöntemin kullanımı artmıştır. Bu yöntemin kullanılmasından anlamlı bir sonuca ulaşılmasına kadar birtakım süreçler izlenmelidir. İzlenecek adımlar Şekil 3.4'te belirtilmiştir.



Şekil 3.4. Veri Madenciliği Süreci

3.10.1. Sorunun tanımlanması

Sorunun tanımlanması veri analizi çalışmalarının en önemli adımıdır. Uygulamanın hangi kurum/kuruluş amacı için yapılacağı kesin bir ifade ile tanımlanmalıdır. Sorun üzerine odaklanmalı, sorun ne ise kesin tanımlarla ifade eidlmeli, sonuçların başarı düzeyini ne kadar ne oranla etkileyeceği net olarak belirlenmelidir. Amaç, problem ile tam örtüşmezse, o çalışmadan sonuç alınmaz ve bu yanlış aynı zamanda başka bir yanlışı da doğurabilir. Aynı zamanda yanlış tanımlanmış amaç, yanlış kararlar alınmasını sağlayarak maliyeti arttırır. Doğru tanımlanan amaç bizi doğru kararlar almaya yönlendirir. Bu şekilde verilmiş doğru kararlar, kazanılmış faydaları belirtir. Sonuç olarak, amaç ve faydalar bu adımda belirlenmelidir.

3.10.2. Verilerin hazırlanması

Veri hazırlığı aşamasında veriler önce toplanır, birbiri ile uyumlaştırılır, birleştirilir, gereksiz verilerden temzilenir ve son olarak uygun veriler seçilerek veri madenciliği çalışmasına devam edilir.

Bu adım verilerin saflaştırlmasındaki en önemli adımdır. Bu adımda yapılacak en ufak hata sürekli bu adımı tekrarlamamıza, verilerin yeniden düzenlenmesine ve buna bağlı olarak zaman kaybına yol açacaktır. Verilerin tekrar hazırlanması ve modelin kurulması aşamaları, karar vericinin karara ulaşmasında; enerji ve zamanının %50 si ve fazlasını harcamasına neden olur.

3.10.3. Verilerin toplanması

Ortadaki problem için işimize yarayacak verilerin ve bu verileri hangi kaynaktan toplayacağımızı belirlediğimiz bir adımdır. Verilerin hangi veri kaynağından alınacağı önemli bir karar aşamasıdır. Verileri eksik veri kaynağından alırsak, çalışmamız da eksik olmuş olur; fazla veri kaynağı daveri kirliliği oluşturacağından veri madenciliği sürecinin uzamasına neden olur.

Veri Madenciliğinde verilerin farklı kaynaklardan toplanması uyumsuzluk yaratacaktır. Bu uyumsuzluklar arasında format hataları, güncelleme hataları, bir alanın birden fazla formatta olması, farklı ölçü birimleri gösterilebilir. Ayrıca verilerin nereden toplandığı da büyük bir önem teşkil etmektedir. Güvenli olmayan kaynaklardan veri toplamışsak, sonraki aşamalarda çalışma için büyük bir sorun yaşatır. Bu nedenle verileri güvenli kaynaktan toplamaya özen göstermeli ve veri kaynağını iyi belirlemeliyiz.

Başarılı çalışma ancak ve ancak doğru ve güvenilir verilerle elde edilir. Ayrıca toplanan verielrin birbiri ile uyumu bu aşamada belirlenir.

3.10.4. Verilerin birleştirilmesi ve temizlenmesi

Farklı kaynaklardan toplanan veriler, uyumsuzluklardan arındırıldıktan sonra, temiz veriler tek bir veritabanında toplanır. Veriler birleştirilirken gereksiz olan verilerin ayıklanmasına özen gösterilmelidir. Çünkü, sonraki adımlarda işimize yaramayan gereksiz verilerle uğraşmış oluruz. Bu durum da çalışmamızı engeller.

3.10.5. Verilerin seçimi

Bu adımda, kurulacak olan modele bağlı olarak veri seçimi yapılmalıdır. Seçilen verilerin anlamlı olması gerekir. Anlamlı olmayan değişkenlerin model oluşturma adımına girmemesi gerekir. Çünkü bu değişkenler model kurulmasına etki etmeyecek boş, etkisiz verilerdir. Bu yöntem ile hem gereksiz veri hacmi olmaz hem de sonuca daha hızlı ulaşırız.

3.10.6. Uygun modelin kurulması ve değerlendirilmesi

Modelin kurulma ve değerlendirme aşaması, ileride yaşabilecek hatalar oluşturmamak açısından önemli bir adımdır. Problem çözümüne en uygun modelin

bulunması için birçok model denenir. En başarılı, en doğru sonucu hangi modelde elde ediyorsak o model ile sürece devam edilir.

Uygun modelin tesipi için modele testler uygulanır. Kullanılan en basit yöntem geçerlilik testidir. Kullanılcak diğer bir yöntem de, çapraz geçerlilik testidir. Bu testlerde hata düzeyinin tespiti yapılır. Bu test, tüm verilerle yapılır.

3.10.7. Modelin kullanılması

Kurulan model doğrudan hazır bir uygulama da olabilir ya da diğer uygulamaların alt uygulması şeklinde kullanılabilir. Kurulan modeller müşteri analizi, kaçak arama tespiti, ayrılan müşterilerin analizi gibi pek çok alanda kullanılır veya başka şirketlerin mevcut uygulamalarında entegre edilebilir (Gündüz, 2016).

3.10.8. Kurulan modelin izlenmesi

Model kullanılmaya başlandıktan sonra izlenmelidir. Zaman içerisinde sistemin özelliği veya verilerde ortaya çıkabilecek değişiklikler, modelin hatalı sonuçlar üretmesine neden olacağı için model tekrardan düzenlenebilir.

3.11. Nitelik Seçimi

Veri madenciliği sürecinden olan model kurulması ve izlenmesi aşamasında veriler birtakım algortimalara tutulur. Başarım oranı yüksek olan algoritma ile model kurulur. Bazı durumlarda veriden daha iyi sonuç elde etmek için verisetinden algoritmaya en çok etki eden nitelikler seçilir. Böylelikle, doğruluk oranı arttırımı sağlanmış olur.

3.11.1. Bilgi kazancı (Information Gain)

Karar ağacı yöntemlerinde, en ayırt edici niteliği belirlemek için her bir nitelik için bilgi kazancı ölçülür. Bu ölçümde Entropy temel alınır. Entropy belirsizliği, rastgeleliği ve beklenmeyen durumun ortaya çıkma olasılığını gösterir (Şeker, 2012).

$$Entropy = -\sum_{i=1}^{m} pi \ log_2 \ (p_i) \tag{3.6}$$

- Eğer tüm örnekler aynı sınıfa aitse belirsizlik yoktur dolayısıyla entropy 0 dır.

- Eğer uniform dağıldıysa her sınıf eşit olasılıkla mümkündür ve entropy 1dir.
- Diğer durumlarda 0< entropy < 1 aralığındadır.

Bilgi kazancında amaç, entropiyi en aza indirgemektir. Bu nedenle, bilgi kazanımı en yüksek olan nitelikler seçilir. Bilgi kazanımı aşağıdaki formülle ifade edilir.

$$Gain(A) = Info(D) - Info_A(D) \quad (3.7)$$

Elimizdeki veri kümesi için ayrılma durumunun Entropy sini hesaplayacak olursak;

3333 adet müşteriden 483 tanesi ayrılmış olmuş, 2850 tanesi mevcut hizmetini kullanmaya devam etmiştir.

Ayrılma / ayrılmama durumuna göre olasılıkları yazacak olursak;

$p_1 = 483/3333 = 0.15$

$p_2 = 2850/3333 = 0.85$

Entropi hesabı için formülümüz:

$$H_{(S)} = -(p_1 \log_2(p_1) + p_2 \log_2(p_2)) \quad (3.8)$$

$H_{(S)} = -(0.15. \log_2(0.15) + 0.85. \log_2(0.85)) = 0.60$

Ayrılma durumu için Entropiyi 0.60 olarak bulduk.

3.11.2. Kazanım oranı (Gain Ratio)

Bu metot, çok çeşitli değerlere sahip nitelikleri seçme eğilimdedir. Bu problemin çözümünde C4.5 (ID3 ten geliştirilmiş) kazanım oranı kullanılır.

$$\text{SplitInfo}_A(D) \sum_{i=1}^{m} \frac{|D_j|}{|D|} \text{ x } \log 2 \frac{|D_j|}{|D|} \quad (3.9)$$

$$\text{GainRatio(A)} = \text{Gain(A)}/\text{SplitInfo(A)} \quad (3.10)$$

En yüksek kazanım oranına sahip nitelik seçilir.

3.12. Karışıklık Matrisi

Model başarımını ölçmek için kullanılan ölçüler, karışıklık matrisine göre hesaplanır. Matriste satırlar test kümesindeki gerçek değerleri; sütunlar ise oluşturulan modelin tahminlenmesini ifade eder**.**

Çizelge 3.2. Karışıklık Matrisi

	Öngörülen Sınıf (Predicted Class)		
Gerçek Sınıf (Actual Class)		Sınıf = 1	Sınıf = 0
	Sınıf = 1	True positive (TP)	False negative (FN)
	Sınıf = 0	False positive (FP)	True negative (TN)

3.12.1. Sınıflandırma Modelini Değerlendirme

Sınıflandırma modeli sonucunda, oluşturulan modelin başarımını ölçmek için bir takım ölçüler kullanılır.

- Doğruluk
- Hata oranı
- Kesinlik
- Duyarlılık
- F – ölçütü

Doğruluk (Accurucy): Doğru sınıflandırılmış örnek sayısının, tüm örneklerin sayısına oranıdır.

$$\text{Doğruluk} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{3.11}$$

Hata Oranı (Error Rate): Yanlış sınıflandırılmış örnek sayısının, tüm örneklerin sayısına oranıdır.

$$\text{Hata Oranı} = \frac{\text{FP+FN}}{\text{TP+TN+FP+FN}} \tag{3.12}$$

Kesinlik (Precision): Doğru sınıflandırılmış pozitif örnek sayısının, pozitif sınıflandırılmış örneklerin sayısına oranıdır.

$$\text{Kesinlik} = \frac{\text{TP}}{\text{TP+FP}} \tag{3.13}$$

Duyarlılık (Recall): Doğru sınıflandırılmış pozitif örnek sayısının, pozitif örneklerin sayısına oranıdır.

$$\text{Duyarlılık} = \frac{\text{TP}}{\text{TP+FN}} \tag{3.14}$$

F-Ölçütü (F-Measure): Kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar verir. Bunun için f-ölçütü tanımlanmıştır. F-ölçütü, kesinlik ve duyarlılığın harmonik ortalamasıdır.

$$\text{F} - \text{Ölçütü} = \frac{\text{2 x Kesinlik x Duyarlılık}}{\text{Kesinlik+Duyarlılık}} \tag{3.15}$$

4. TELEKOM SEKTÖRÜNDE AYRILAN MÜŞTERİ ANALİZİ

4.1. Ayrılan Müşteri Analizi

Mevcut müşterilerin kaybını önceden tahmin etmeye dayanan analiz yöntemidir. Genellikle telekom, bankacılık veya sigortacılık sektörlerinde kullanılır. Bu analiz yöntemi ile müşteri kaybının engellenmesi amaçlanmaktadır. Telekom firmasının mevcut müşterileri arasından; ayrılan müşterileri oranına, ayrılma denilir (Akça, 2016).

Müşteri kaybı; telekom, bankacılık, sigorta gibi müşteri sirkülasyonunun çok olduğu sektörlerde büyük öneme sahiptir. Çünkü yeni müşteri elde etmek; mevcut müşteriyi elde tutmaktan daha maliyetli operasyonel adımlar gerektirir. Bir sektörün değeri, sahip olduğu aktif müşteri sayısı ile doğru orantılıdır. Bir sektörün karlılığı, büyüklüğü, maliyetleri, yatırım araçlarının kapasitesi, nakit akışı gibi parametreleri aktif müşteri sayısına bağlıdır. Aynı zamanda sahip olunan müşterilerin sadakatine de bağlıdır. Bazı şirketler, uzun süreli müşterilerinden büyük kar elde ederler. Kar oranı hesabı için Müşteri Ömrü (customer lifetime value - CLV) gibi yöntemler kullanılır. Müşterinin toplamda aldığı hizmetler ve hangi hizmetten ne kadar sürede yararlandığı gibi parametreler ele alınır. Bazı firmaların yatırımları, müşteriden elde edilen kazanca bağlı ise ayrılma analizinin önemi büyüktür.

Müşteri memnuniyetsizliği, rekabet, kalitesiz hizmet, daha uygun fiyatlı kampanyalar, regülasyon, ekonomik durumlar ayrılma oranını etkileyen nedenler arasındadır. Yeni müşteri kazanmak her zaman zor ve maliyetlidir. Bu nedenle mevcut müşterileri elde tutmak için müşteri memnuniyeti sağlanmalıdır. Bir şirket kazandığı müşteri oranında müşteri kaybediyorsa o şirketin kar oranı azalmış olacaktır.

Bugün tüm sektörlerde, ayrılma eğilimi gösteren müşteriler için ilgili departmanlar bulunmaktadır. Ayrılma eğilimi gösteren müşteriye, özel teklifler sunarak müşteriyi vazgeçirmeye çalışıyorlar. Genellikle firmalar müşteri kazanmak için şirket gelirinin büyük bir kısmını müşteriyi çekecek kampanyalara ayırıyor. Firmalar eldeki müşterilerin hizmet kalitesi sürekliliğini sağlar, ayrılma eğilimi gösteren müşterilere uygun kampanya geliştirirse ayrılma oranını en aza indirgerler.

4.1.1. Müşteri kayıp çeşitleri

Müşteri kaybı, "gönüllü kayıp (voluntary churn)" ve "gönülsüz kayıp (involuntary churn)" olmak üzere iki farklı gruba ayrılır.

4.1.1.1. Gönüllü kayıp (Voluntary Churn)

Müşterinin mevcut hizmetini kendi isteği ile bırakıp, farklı işletmeye ait hizmetten yararlanmasına, gönüllü kayıp denir. Örneğin, telekom sektöründeki bir müşterinin kendi isteği ile mevcut tarifesini iptal edip, diğer bir firmanın hizmetinden yararlanması, gönüllü müşteri kaybıdır. Bu durum, müşterinin daha iyi, daha ekonomik ve avantajlı hizmet alma inanacı doğrultusunda gerçekleşir.

4.1.1.2. Gönülsüz kayıp (Involuntary Churn)

Müşterinin genellikle kendi isteği dışında gerçekleşen, ekonomik ve çevresel şartlardan etkilenen kayıptır. Örneğin müşterinin fiber paketi var ve başka ile taşındı. Yeni yerleşim yeri fiber alt yapısını desteklemiyor. Bu nedenden dolayı müşterinin mevcut paketini iptal edip, başka firmanın hizmetini tercih etmesi, gönülsüz kayıptır. Bu durum, müşterinin yaşamış olduğu çevresel etkenlere bağlıdır. Veri madenciliğinde müşteri kayıp analizi için geliştirilen yöntemlerde bu tarz gönülsüz kayıp durumları göz ardı edilir. Çünkü, istek dışı gelişen olaylardan dolayı müşteri kaybı önlenemez.

4.2. Müşteri Kaybını Engellemek İçin Yapılması Gerekenler

Ayrılma analizinde amaç, ayrılma ihtimali yüksek olan yani ürün ve hizmet kullanımını bırakma oranı yüksek olan müşterileri belirlemektir. Ayrılma analizi de bu oranı elde etmek için kullanılan yöntemdir. Günümüzde bir çok birbirine yakın hizmetler veren, aynı ürün ya da hizmeti piyasaya süren rakip firmalar vardır. Rekabet ortamında yeni müşteri elde etmek, mevcut müşteriyi elde tutmaktan daha maliyetli bir operasyondur. Bu nedenle, mevcut müşteri firmalar için çok değerlidir. Özellikle de finans hizmetleri gibi müşteri tercihlerinin çok sık değişmediği ortamlarda müşteri kaybı önemli gelir kayıpları anlamına gelebilir (Akça, 2016).

Ayrılma analizi ile müşterinin ayrılma eğilimini belirleyerek, tahminlerde bulunabiliriz. Bunun için müşterileri segmentlerine göre ayırabilir ve müşterinin ihtiyaç duyduğu hizmete göre analizler yapabiliriz. Bu analizler sonucunda, müşteri kaybını azaltmak ve müşteri sadakati için uygun kampanyalar ve önlemler geliştirilir. Daha geniş kitleye yayılarak, firmaların karlılık oranını arttırılmış olur.

4.3.Müşteriyi Elde Tutma Yöntemleri

Şirketler için yeni müşteri kazanmak, mevcut müşterileri elde tutmaktan daha maliyetlidir. Memnuniyetsizlik yaşayan müşteriyi geri döndürmek mevcut müşterileri elde tutmaktan daha da maliyetlidir (Kotler, 2002). Şirketler, elindeki mevcut müşteriyi kaybetmemek için stratejiler geliştirmelidir.

4.3.1. Çapraz satış

Çapraz satışta amaç, müşterinin satın alma durumunu analiz ederek, müşteriye birbiri ile ilişkilli kampanya veya hizmete yönlendirmektir. Bu şekilde, müşteri ile şirket arasında bir bağ oluşur ve müşterinin ayrılma eğilimi önlenmiş olur.

4.3.2. Müşteri yaşam ömrü değeri

Müşterinin ömrü boyunca aldığı tüm hizmetler analiz edilerek ortaya çıkan net değerdir. Bu analizde önemli olan noktalar, müşterinin aldığı hizmet tutarı ve bu hizmetten ne kadar yararlandığıdır.

4.3.3. Tepki modelleme

Müşteri ile doğru pazarlama iletişimi kurulmalıdır. Gerekli olan bilgi ve mesajlar doğru iletişim araçları ile müşteriye iletilmelidir. İletişim ne kadar etkili olduğu ölçülmelidir. Bu nedenle, müşterilerin bu iletişime verdiği tepki analiz edilmelidir**.**

5. UYGULAMA

5.1. Problem Tanımı

Telekom sektörüne ait verilerle ayrılma eğilimi gösteren müşteriler belirlenir. Bu müşterilerin neden mevcut hizmeti bırakıp, rakibe yöneldiği sorularına yanıt aranır. Elde edilen sonuçlara göre müşteriyi elde tutmak için uygun stratejiler belirlenir ve geliştirilir.

5.2.Veri Madenciliği Süreci

Öncelikle müşterilere ait veriler veritabına kaydedilir, tüm veriler toplanır. Veri kümesinde boş ve eksik değer içeren alan olmadığı için veri temizleme işlemi yapılmadı. Veri kümemiz, eksik boş bilgi içermemektedir. Veriler tek değer içerdiği için veri dönüştürme adımı da yapılmamıştır.

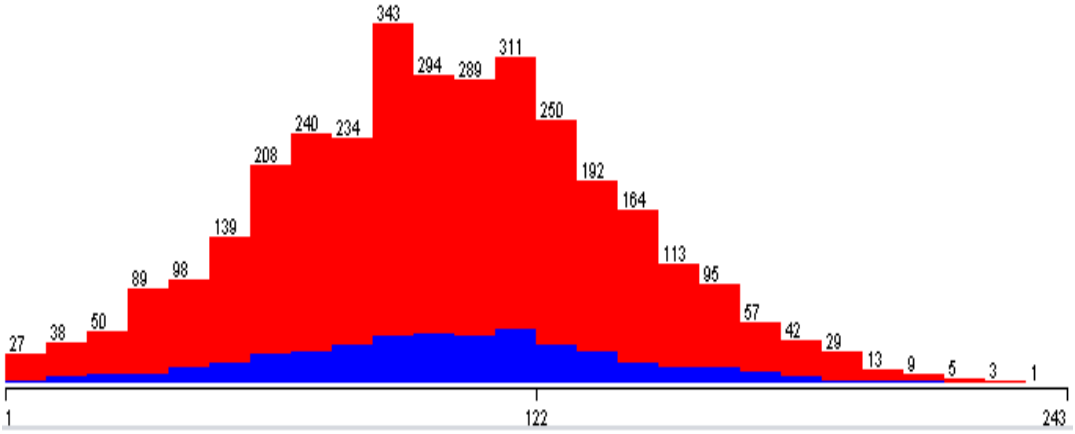
Veri kümemiz bütünüyle anlamlı veri içermektedir. Bu şekliyle, veri kümesine çeşitli sınıflandırma algoritmaları uygulanıp, problem çözümüne adım adım devam edilir. Ayrılma eğilimi gösteren müşterilere uygun hizmet planlaması ve müşteriyi elde tutma stratejileri geliştirilir.

5.3.Veri Kümesinin Tanıtımı

Churn(???) veri kümesi bigml.com sitesinden alınmıştır. Veri kümesine ulaşmak için "https://bigml.com/user/francisco/gallery/dataset/530c28ec67dc090932002584" linkinden yararlanılır. Veri kümemiz 18 aylık bilgilerin toplamıdır. 3333 adet müşterilere ait veri içerir. Veri kümesinde 21 adet özellik vardır. Bu özelliklere ait bilgileri tek tek açıklayalım. Ayrıca, müşterilerin hangi aralıkta ayrılma eğilimi gösterdiği grafiklerde belirtilmiştir. Ayrılanlar mavi, ayrılmayanlar kırmızı renk ile gösterilmiştir.

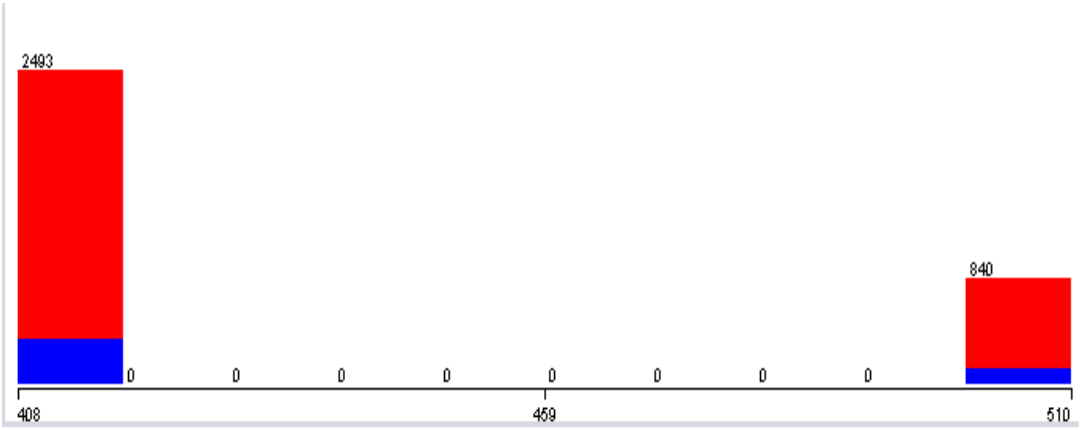
State: 51 adet Kolombiya' daki şehir isimlerini içerir. Bu alanın veri madenciliği analizine etkisi yoktur.

Account Length*:* Hesabın ne zamandır aktif olduğu bilgisini içerir. Account length alanına göre ayrılma dağılımı Şekil 5.1'deki gibidir.



Şekil 5.1. Account Length Alanına Göre Ayrılma Dağılımı

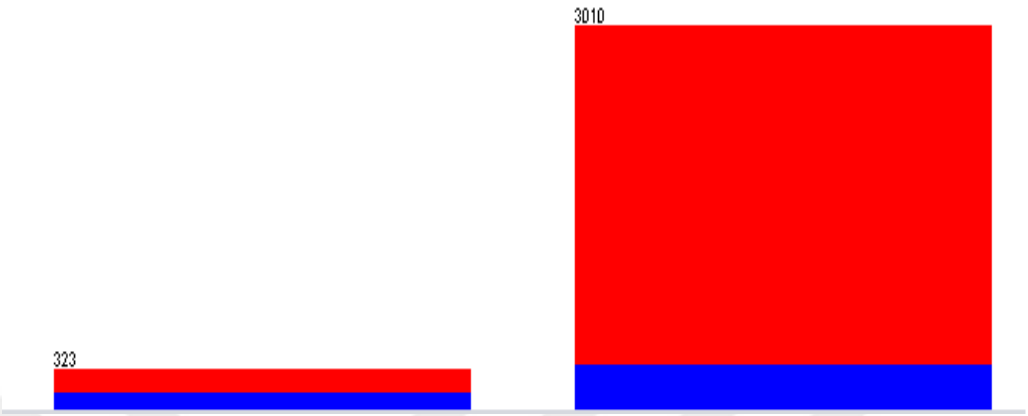
Area Code: Şehirlere ait alan kodunu ifade eder. Alan kodlarına göre ayrılma dağılımı Şekil 5.2'deki gibidir.



Şekil 5.2. Area Code Alanına Göre Ayrılma Dağılımı

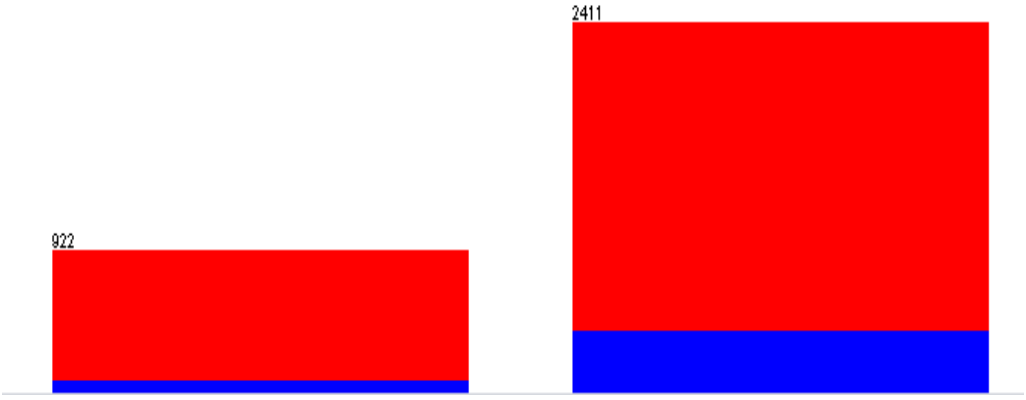
Phone Number: Müşterilerin telefon numarası bilgisini içerir. Bu alanın veri madenciliği analizine etkisi yoktur.

International Plan: Müşterinin uluslararası bir planı olup olmadığının bilgisini içerir. International Plan tercihine göre ayrılma dağılımı Şekil 5.3'teki gibidir.



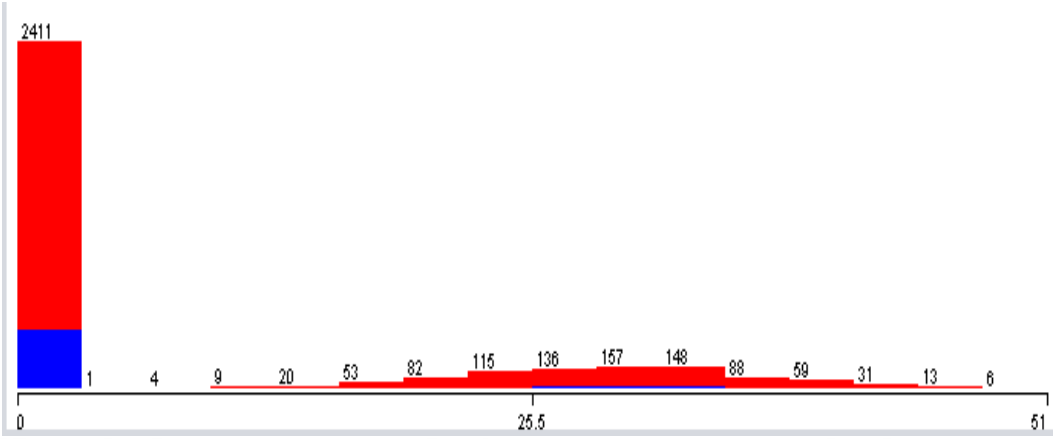
Şekil 5.3. International Plan Alanına Göre Ayrılma Dağılımı

VoiceMail Plan: Müşterinin sesli mail planı olup olmadığının bilgisini içerir. VoiceMail Plan tercihine göre ayrılma dağılımı Şekil 5.4'teki gibidir.



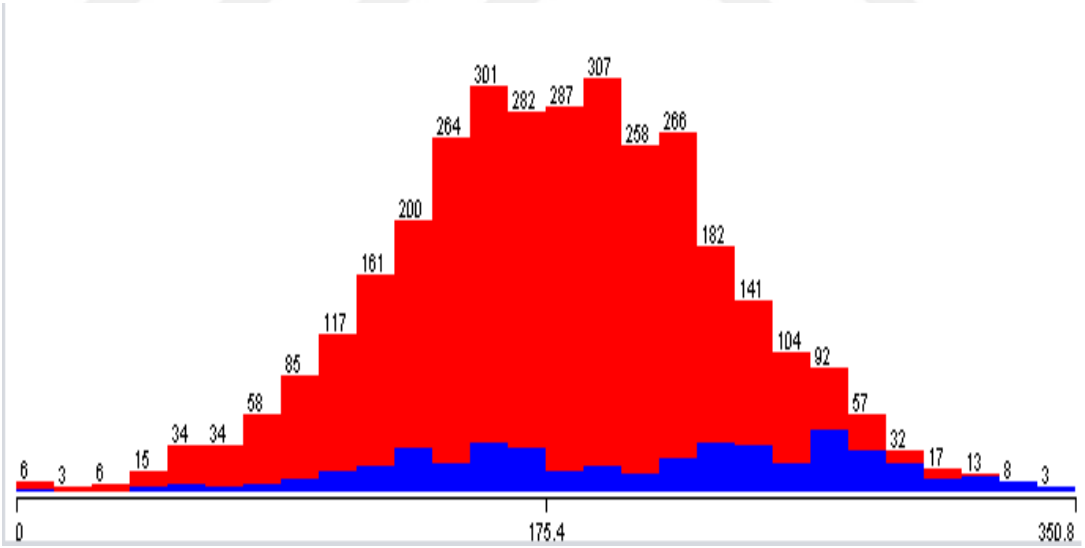
Şekil 5.4. VoiceMail Plan Alanına Göre Ayrılma Dağılımı

Number Of Voice Mail Messages: Müşterinin sesli mail mesaj sayısını ifade eder. Atılan VoiceMail Messages sayısına göre ayrılma dağılımı Şekil 5.5'teki gibidir.



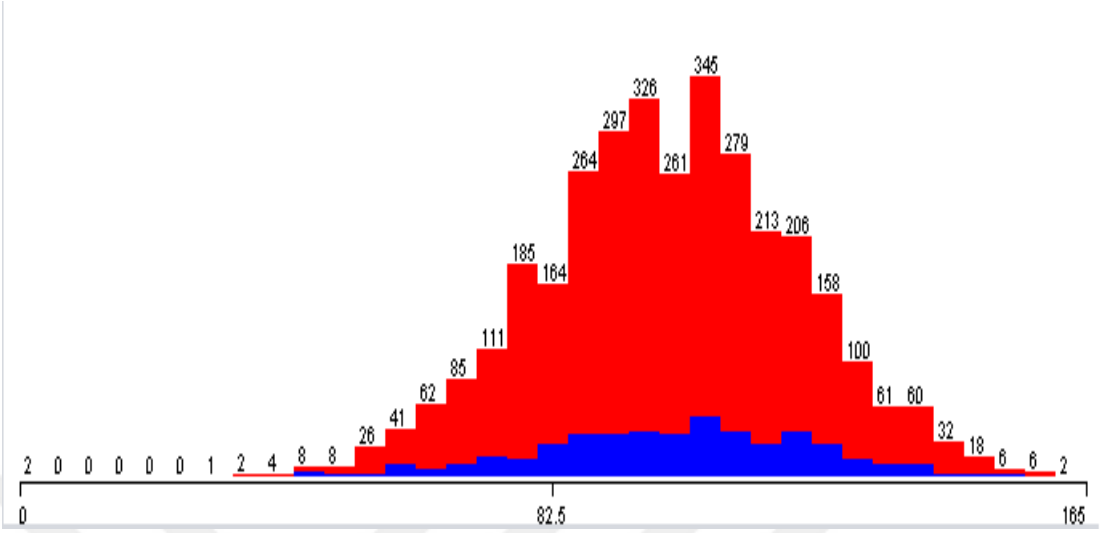
Şekil 5.5. Number Of Voice Mail Messages Alanına Göre Ayrılma Dağılımı

Total Day Minutes: Günlük konuşulan dakika bilgisini içerir. Günlük konuşulan dakika sayısına göre ayrılma dağılımı Şekil 5.6'daki gibidir.



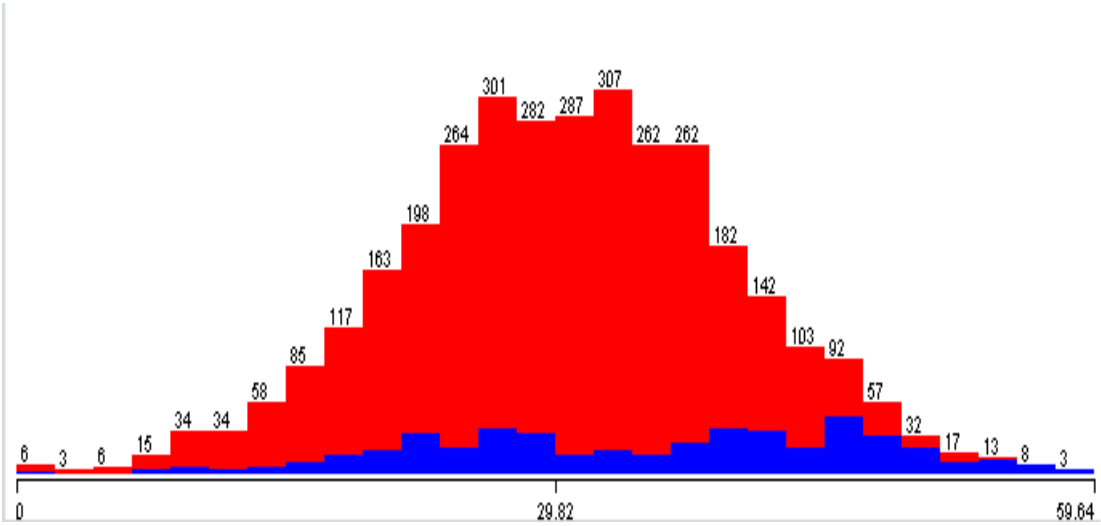
Şekil 5.6. Total Day Minutes Alanına Göre Ayrılma Dağılımı

Total Day Calls: Günlük yapılan arama sayısı bilgisini içerir. Günlük yapılan arama sayısına göre ayrılma dağılımı Şekil 5.7'deki gibidir.



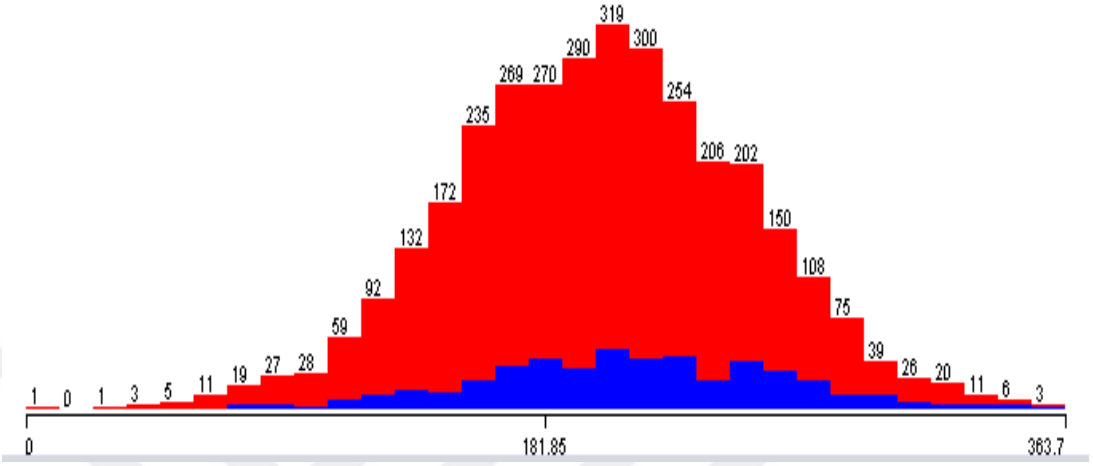
Şekil 5.7. Total Day Calls Alanına Göre Ayrılma Dağılımı

Total Day Charge: Günlük toplam konuşma tutar bilgisini içerir. Günlük tutar alanına göre ayrılma dağılımı Şekil 5.8'deki gibidir.



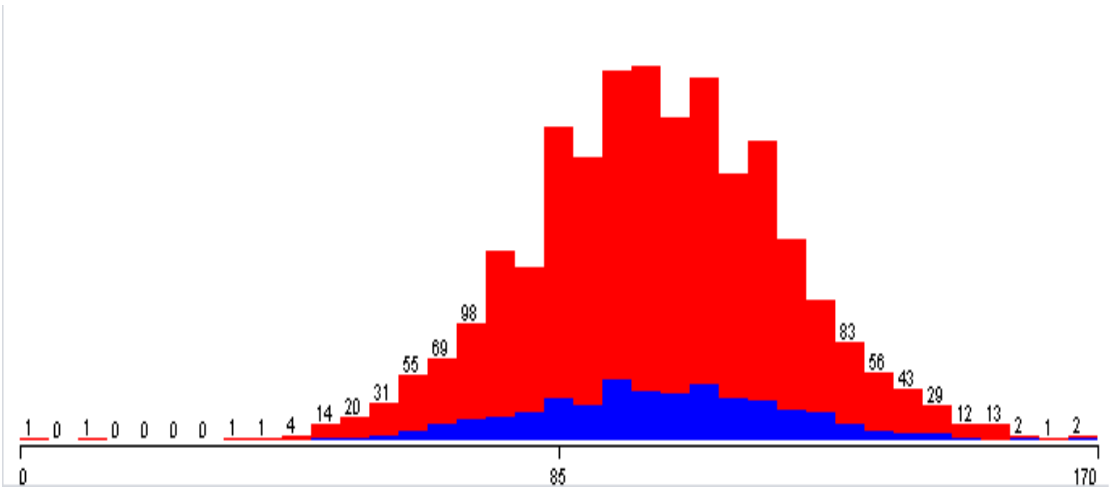
Şekil 5.8. Total Day Charge Alanına Göre Ayrılma Dağılımı

Total Evening Minutes*:* Akşam konuşulan dakika bilgisini içerir. Akşam konuşulan dakika sayısına göre ayrılma dağılımı Şekil 5.9'daki gibidir.



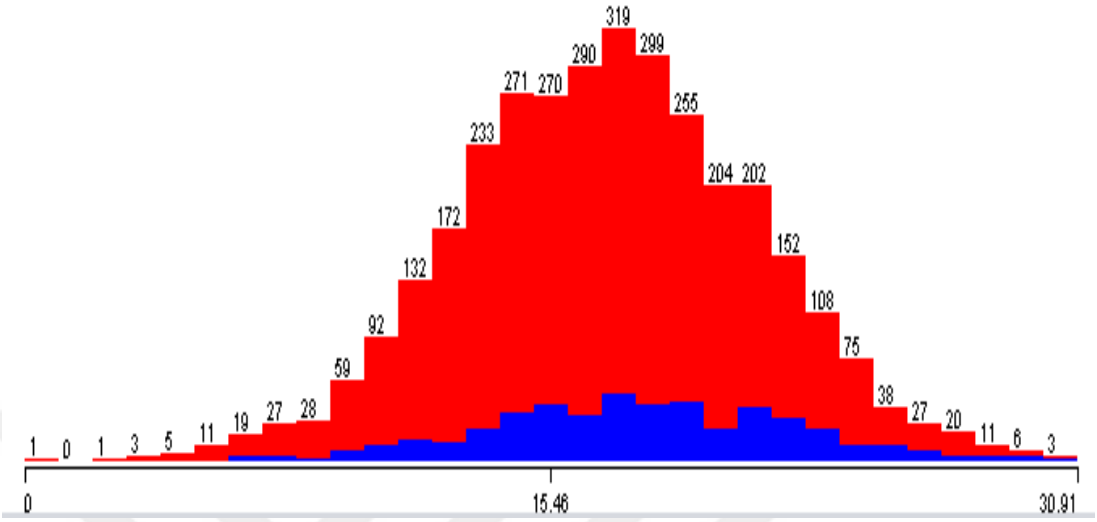
Şekil 5.9. Total Evening Minutes Alanına Göre Ayrılma Dağılımı

Total Evening Calls: Akşam yapılan arama sayısı bilgisini içerir. Akşam yapılan arama sayısına göre ayrılma dağılımı Şekil 5.10'daki gibidir.



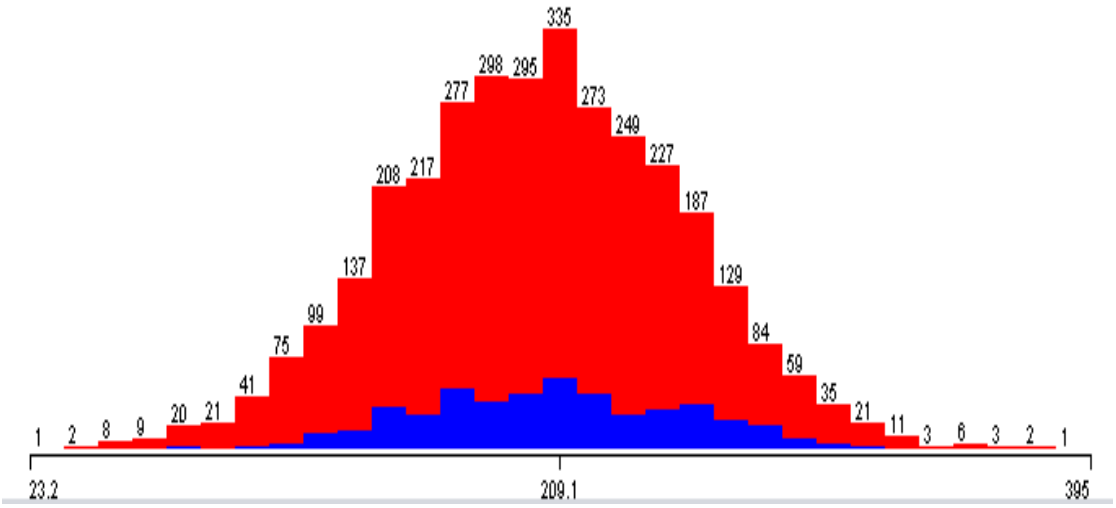
Şekil 5.10. Total Evening Calls Alanına Göre Ayrılma Dağılımı

Total Evening Charge: Akşam yapılan toplam konuşma tutar bilgisini içerir. Evening charge alanına göre ayrılma dağılımı Şekil 5.11'deki gibidir.



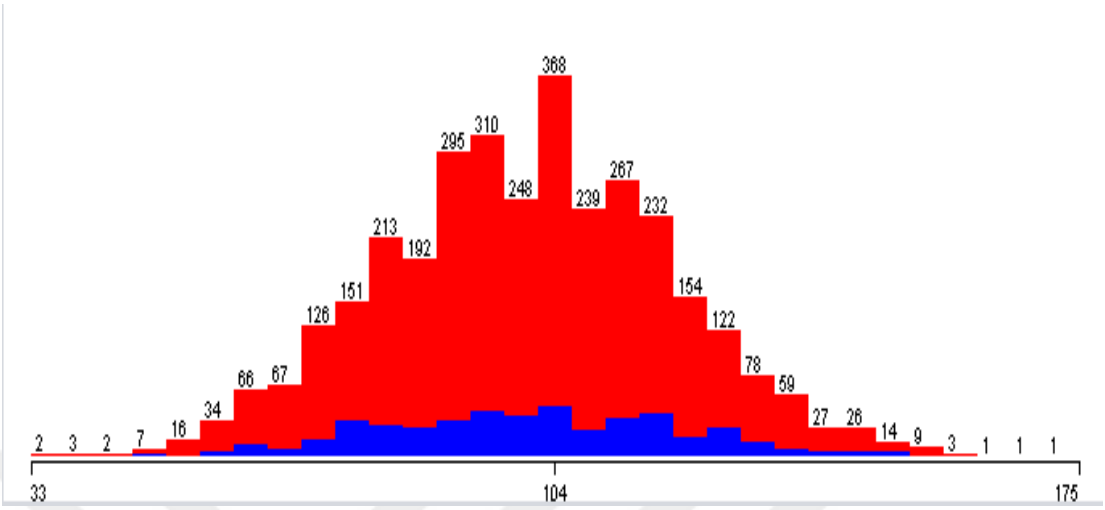
Şekil 5.11. Total Evening Charge Alanına Göre Ayrılma Dağılımı

Total Night Minutes: Gece konuşulan dakika bilgisini içerir. Gece konuşulan dakika sayısına göre ayrılma dağılımı Şekil 5.12'deki gibidir.



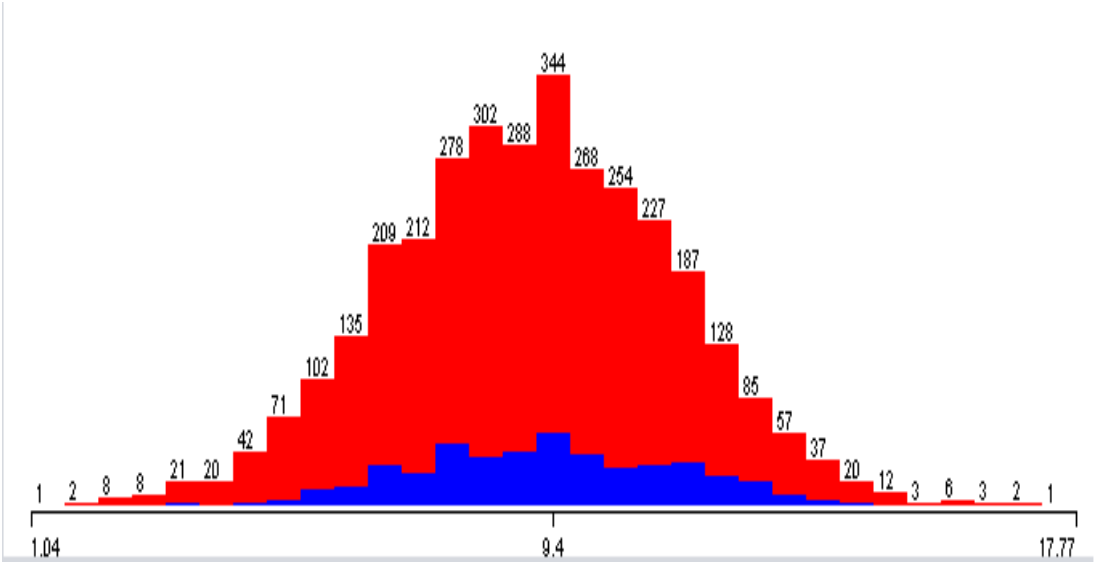
Şekil 5.12. Total Night Minutes Alanına Göre Ayrılma Dağılımı

Total Night Calls: Gece yapılan arama sayısı bilgisini içerir. Gece yapılan arama sayısına göre ayrılma dağılımı Şekil 5.13'teki gibidir.



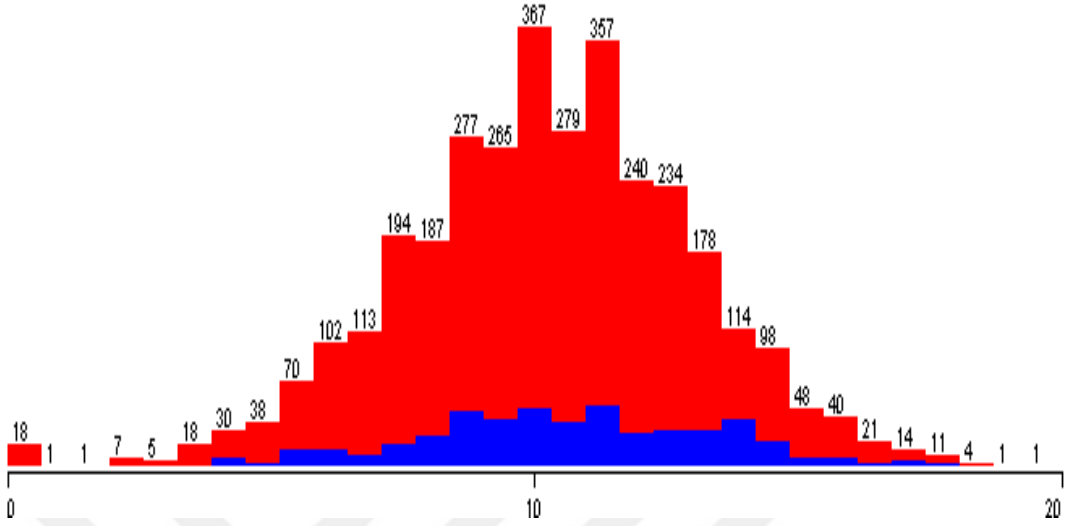
Şekil 5.13. Total Night Calls Alanına Göre Ayrılma Dağılımı

Total Night Charge: Gece yapılan toplam konuşma tutar bilgisini içerir. Night charge alanına göre ayrılma dağılımı Şekil 5.14'teki gibidir.



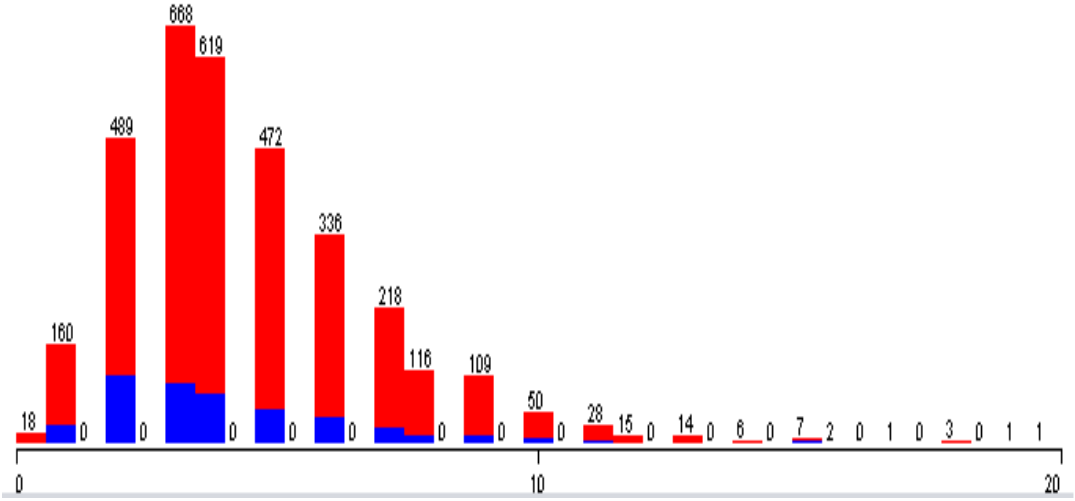
Şekil 5.14. Total Night Charge Alanına Göre Ayrılma Dağılımı

Total International Minutes: Yapılan toplam uluslararası dakika bilgisini içerir. Toplam international minutes alanına göre ayrılma dağılımı Şekil 5.15'teki gibidir.



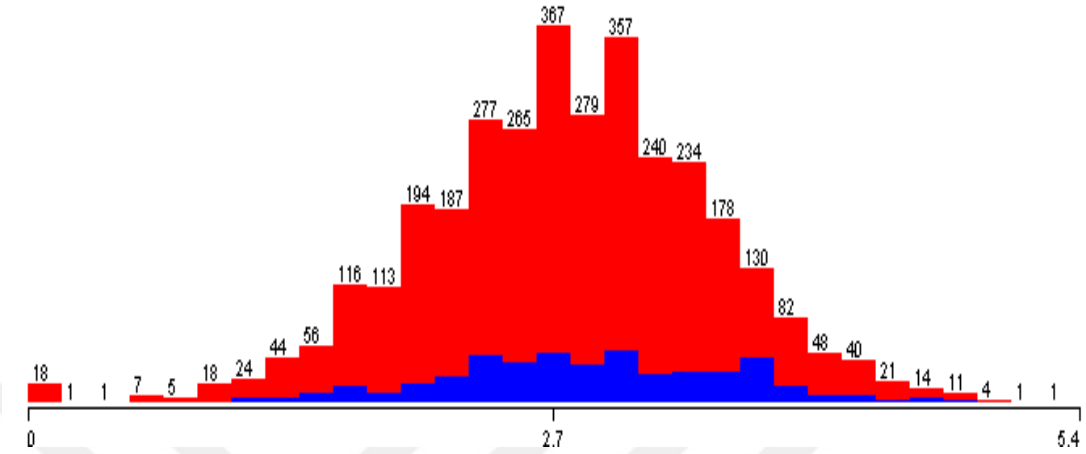
Şekil 5.15. Total International Minutes Alanına Göre Ayrılma Dağılımı

Total International Calls: Yapılan toplam uluslararası arama sayısı bilgisini içerir. Toplam international calls alanına göre ayrılma dağılımı Şekil 5.16'daki gibidir.



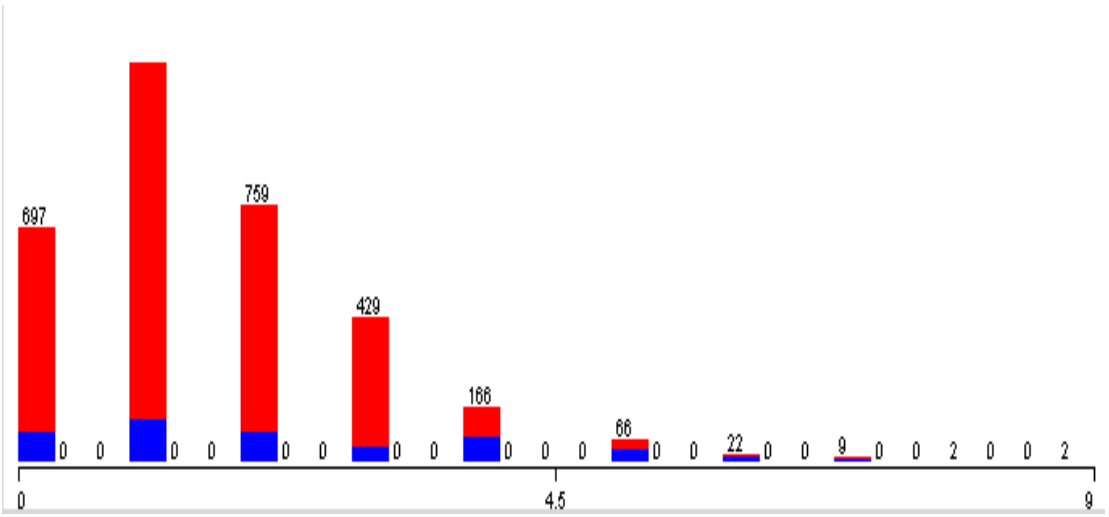
Şekil 5.16. Total International Calls Alanına Göre Ayrılma Dağılımı

Total International Charge: uluslararası yapılan aramaların toplam tutar bilgisini içerir. Toplam international charge alanına göre ayrılma dağılımı Şekil 5.17'deki gibidir.



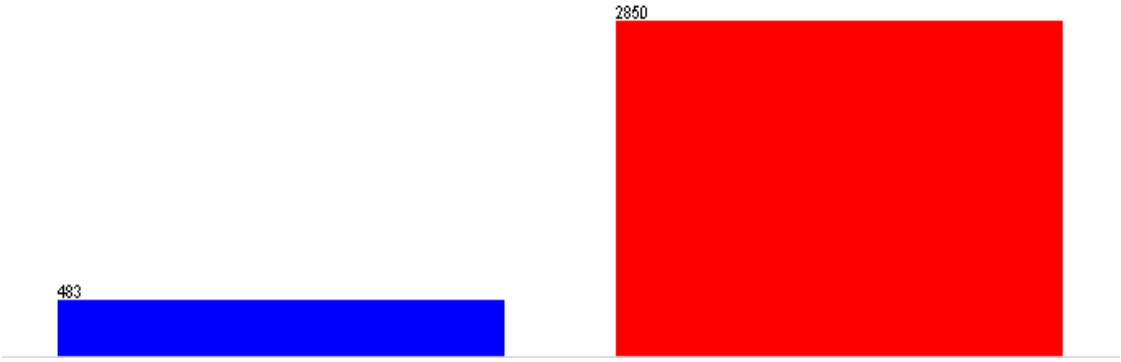
Şekil 5.17. Total International Charge Alanına Göre Ayrılma Dağılımı

Number Of Calls To Customer Service: Müşteri hizmetlerine yapılan arama sayısını ifade eder. Arama sayısına göre ayrılma dağılımı Şekil 5.18'deki gibidir.



Şekil 5.18. Number Of Calls To Customer Service Alanına Göre Ayrılma Dağılımı

Churn : Müşterinin ayrılma eğilimi gösterip göstermediği bilgisini içerir. 3333 adet veriden 483 tanesi ayrılma eğilimi göstermiş, 2850 adet müşteri mevcut hizmetini almaya devam etmiştir (Şekil 5.19).



Şekil 5.19. Ayrılma Dağılımı

5.4. Modelleme

Gereksiz verilerin temizlenip, birbiri ile ilişkili hale getirildikten sonraki aşamadır. Bu aşamada, veri kümesi üzerine farklı modeller denenerek hata payı en az olan model seçildi.

Kurulacak olan modeli belirlemek için WEKA programı kullanıldı. WEKA programı, Yeni Zellanda'daki Waikato Üniversitesi tarafından geliştirilmiş veri analiz programıdır. WEKA programı, birçok algoritmaları içinde barındırır. İşlevsel grafik yapısına sahiptir, Java ile yazılmıştır ve açık kaynak kodludur. Ayrıca bu veri analiz programı; sıra ilişkilendirme kuralları, regresyon, kümeleme, veri ön işleme, sınıflandırma ve görselleştirme araçları içerir. Weka' da ARFF (Attribute-Relation File Format) denilen değişken tanımlanmasına izin veren ASCII metin dosyaları kullanılır. ARFF dosyasının ilk kısmında, değişkenler, bu değişkenlerin birbiri arasındaki ilişkiler, herbir değişkenin türü ve sonrasında tanımlanan değişkenlerin alacağı değer bulunur. İlgili verilerin başında @DATA bulunur.

Değişkenler, ilişkileri ve dataların .arrf dosyasında tanımlama şekli aşağıdaki gibidir:

@relation churn
@attribute State string
@attribute AccountLength numeric
@attribute AreaCode numeric
@attribute Phone string
@attribute IntlPlan {yes,no}
@attribute VMailPlan {yes,no}
@attribute VMailMessage numeric
@attribute DayMins real
@attribute DayCalls numeric
@attribute DayCharge real
@attribute EveMins real
@attribute EveCalls numeric
@attribute EveCharge real
@attribute NightMins real
@attribute NightCalls numeric
@attribute NightCharge real
@attribute IntlMins real
@attribute IntlCalls numeric
@attribute IntlCharge real
@attribute CustServCalls numeric
@attribute Churn? {True,False}

@data
KS,128,415,3824657,no,yes,25,265.100000,110,45.070000,197.400000,99,16.78000
0,244.700000,91,11.010000,10.000000,3,2.700000,1,False
OH,107,415,3717191,no,yes,26,161.600000,123,27.470000,195.500000,103,16.620
000,254.400000,103,11.450000,13.700000,3,3.700000,1,False

WEKA üzerinde yapılan çalışmada, veri kümesi analizinde k-fold cross validation yöntemi kullanılmış ve k değeri 10 olarak seçilmiştir.

İlk olarak, veri madenciliği analizine etkisi olmayan PhoneNumber ve Account değişkenleri silindi. Kalan 19 değişkenlere çeşitli sınıflandırma algoritmaları

uygulandı. Ham dataya uygulanan sınıflandırma algoritma sonuçları Deney A kategorisinde Çizelge 5.2'de gösterilmiştir.

İkinci adım olarak da, doğruluk oranına en iyi etki edecek niteliklerin belirlenmesi için veri kümesi, nitelik seçimi algortimaları olan InfoGain, GainRatio ve Correlation algoritmalarına sokuldu. Nitelik seçimi algoritma sonuçları Çizelge 5.1'deki gibidir. Bu şekilde, sonuca en çok etki edecek nitelikler belirlenerek veri kümesine tekrar aynı sınıflandırma algoritmaları uygulanmıştır. Uygulama sonucunda, 19 değişkenden 11 inin doğruluk oranına daha çok etki ettiği görülmüştür.

Çizelge 5.1. Nitelik Seçim Sonuçları

Attributes	**InfoGain**	**GainRatio**	**Correlation**
IntlPlan	4	2	1
VMailPlan	5	9	5
VMailMessage	6	8	8
DayMins	2	4	3
DayCharge	1	3	4
EveMins	10	10	6
EveCharge	11	11	7
IntlMins	9	6	10
IntlCalls	7	7	11
IntlCharge	8	5	9
CustServCalls	3	1	2

Nitelik seçiminden sonraki adımda yapılan sınıflandırma algoritma sonuçları; ham data için Deney A, InfoGain için Deney B, GainRatio için Deney C, Correlation için Deney D sütununda belirtilmiştir. Bu sonuçlar Çizelge 5.2'deki gibidir.

Çizelge 5.2. Sınıflandırma Algoritma Sonuçları

	Naive Bayes	**Logistic Regresyon**	**SMO**	**J.48**	**K – star**
Deney A (Ham Data)	88.2988 %	86.1386 %	85.5086 %	94.4194 %	82.9283 %
Deney B (Info Gain)	87.4287 %	85.5686 %	85.5086 %	94.8095 %	86.7687 %
Deney C (Gain Ratio)	86.8587 %	86.8587 %	86.8587 %	91.5092 %	87.5188 %
Deney D (Correlation)	87.6388 %	85.7486 %	85.5086 %	92.1392 %	89.529 %

Veri kümesine sırasıyla uyguladığımız Naïve Bayes, Lojistik Regresyon, SMO, J.48, K-star sınıflandırma algortimalarının TP, FP, Kesinlik, Duyarlılık değerleri de Çizelge 5.3'te verilmiştir. Bu çizelgeden de anlaşılacağı üzere doğruluk oranı yüksek olan, ham veri kümesine uygulanan J.48 algoritmasıdır (İşler, Narin, 2012).

Çizelge 5.3. Doğruluk Sonuçları

Algoritmalar	**Deney Veri Kümeleri**	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Naive Bayes	**Deney A**	0,883	0,117	0,816	0,665	0,799	0,834
	Deney B	0,874	0,126	0,800	0,656	0,779	0,825
	Deney C	0,869	0,131	0,796	0,651	0,781	0,845
	Deney D	0,876	0,124	0,806	0,658	0,789	0,847
Logistic Regresyon	**Deney A**	0,861	0,139	0,777	0,643	0,759	0,814
	Deney B	0,856	0,144	0,765	0,638	0,747	0,808
	Deney C	0,869	0,131	0,796	0,651	0,781	0,845
	Deney D	0,857	0,143	0,768	0,639	0,749	0,812
SMO	**Deney A**	0,855	0,145	0,676	0,637	0,713	0,500
	Deney B	0,855	0,145	0,676	0,637	0,713	0,500
	Deney C	0,869	0,131	0,796	0,651	0,781	0,845
	Deney D	0,855	0,145	0,676	0,637	0,713	0,500
J.48	**Deney A**	0,944	0,056	0,887	0,726	0,867	0,833
	Deney B	**0,948**	**0,052**	**0,891**	**0,730**	**0,801**	**0,929**
	Deney C	0,915	0,085	0,856	0,697	0,837	0,832
	Deney D	0,921	0,079	0,863	0,703	0,839	0,792
K – star	**Deney A**	0,829	0,171	0,742	0,611	0,734	0,700
	Deney B	0,868	0,132	0,789	0,650	0,760	0,836
	Deney C	0,875	0,125	0,801	0,657	0,777	0,816
	Deney D	0,895	0,105	0,831	0,677	0,806	0,844

En yüksek doğruluk oranı içeren yöntem, diğer algortima sonuçları ile karşılaştırmalar sonucunda belirlendi. Veri madenciliğinde kullanılacak olan model belirlenmiş oldu.

Sonuçlar doğrultusunda J.48 algoritmasına göre model oluşşturuldu. Bu modele göre doğruluk matrisi Çizelge 5.4. te belirtilmiştir.

Çizelge 5.4. Doğruluk Matrisi

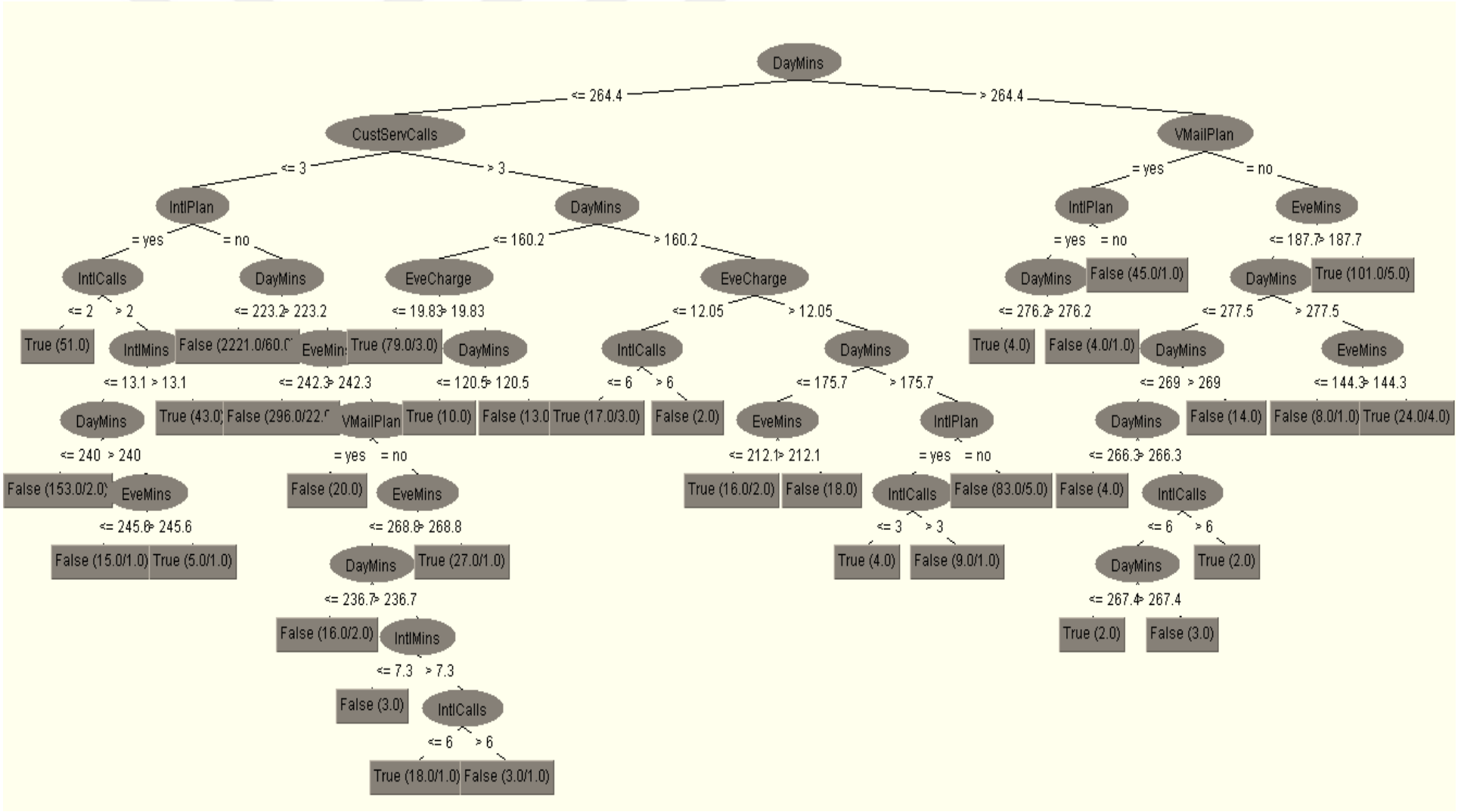
		Öngörülen Sınıf	
		Sınıf=1	Sınıf=0
Doğru Sınıf	Sınıf=1	353	130
	Sınıf=0	43	2807

Çizelge 5.4. teki değerlere göre; 353 (TP), 43 (FP), 130 (FN), 2807 (TN) doğruluk sonuçlarını hesaplarsak;

Doğruluk = (2807 + 353) / 3333 = **0,948**
Hata Oranı = (130 + 43) / 3333 = **0,052**
Kesinlik = 353 / (353 + 43) = **0,891**
Duyarlılık = 353 / (353 + 130) = **0,730**
F-Ölçütü = (2* 0,891 * 0,730) / (0,891 + 0,730) = 1,30 / 1,621 = **0,801**
değerlerini elde ederiz.

5.4.1. Ayrılan müşteri analizi

18 aylık süreç sonunda toplanan 3333 adet müşterisi bilgisi içinde ayrılma eğilimi gösteren müşteri sayısı 483’ tür. Bu sonuca ulaşmada, en yüksek doğruluk oranına sahip algoritma karar ağacı algoritmlarından J.48 olmuştur. Şimdi Şekil 5.20’de gösterilen J.48 karar ağacını yorumlayalım:



Şekil 5.20. J.48 Karar Ağacı

Günlük konuşma dakikası 264.4 üzerinde olan, sesli mail planı olan, aynı zamanda uluslararası konuşma planı olan müşteriler ayrılma eğilimi göstermiş.

Günlük konuşma dakikası 264.4 üzerinde olan, sesli mail planı olmayan, akşam arama dakikası 187.7 nin üzerinde olan müşteriler ayrılma eğilimi göstermiş.

Müşteri hizmetlerini 3 ten az arayan, uluslarası arama planı olan, uluslararası arama sayısı 2 nin altında olanlar ve uluslararası konuşma süresi 13.1 dakikadan fazla olan müşteriler ayrılma eğilimi göstermiş.

Müşteri hizmetlerini 3 ten az arayan, uluslarası arama planı olan, uluslararası arama sayısı 2 nin altında olanlar ve konuşma süresi 13.1 dakikanın altında olan, günlük konuşma süresi 240 dakikadan fazla olan ve aynı zamanda akşamları konuşma konuşma süresi 245.6 dakikadan fazla olan müşteriler ayrılma eğilimi göstermiş.

Veri kümesini state alanına göre incelediğimizde özellikle MD, MI, NJ ve TX bölgelerinde yaşayan müşterilerin diğer bölgelere oranla daha çok ayrılma eğilimi gösterdikleri görülmüştür.

Çizelge 5.5. State' lere Göre Ayrılan Müşteri Analizi

State	**intl**	**vmail**	**vmail.msgs**	**day.mins**	**day.charge**	**eve.mins**	**eve.charge**	**intl.mins**	**intl.calls**	**intl.charge**	**svc.calls**	**churn**
MD	yes	yes	41	173,1	29,43	203,9	17,33	14,6	15	3,94	0	Yes
TX	no	no	0	178,9	30,41	169,1	14,37	13,8	3	3,73	4	Yes
TX	no	no	0	210,6	35,8	249,2	21,18	12,4	1	3,35	2	Yes
NJ	no	no	0	237,9	40,44	247,6	21,05	13,9	4	3,75	1	Yes
TX	no	no	0	326,5	55,51	176,3	14,99	10,7	6	2,89	2	Yes
MD	no	no	0	250,2	42,53	267,1	22,7	13	2	3,51	1	Yes
MD	yes	no	0	312	53,04	129,4	11	10,5	2	2,84	0	Yes
MI	yes	no	0	216,9	36,87	207,4	17,63	17,5	5	4,73	1	Yes
NJ	yes	yes	37	220,2	37,43	185,3	15,75	4,1	2	1,11	0	Yes

Sonuçlara göre uluslararası planı olan ve uluslararası konuşma yapan müşteriler daha çok ayrılma eğilimi göstermişlerdir. Firma, mevcut altyapısını güçlendirerek müşterilerine daha kaliteli bir hizmet sağlarsa ayrılma oranı düşecektir. Firma, aynı zamanda günlük ve akşam yapılan konuşmalarda tarife ücreti daha az olan kampanyalar üretirse ve de günlük ek dakika gibi mevcut pakete eklenirse yine ayrılma oranında azalma görülür. Veri kümesinde ayrılma eğilimi gösteren bölgeleri sıraladığımızda; MD, MI, NJ ve TX bölgelerinde en çok ayrılma eğilimi yaşanmıştır. Bu bölgelere ek baz istasyonu kurulabilir, altyapı sorunları giderilebilir ya da altyapıya destek verecek çalışmalarda bulunulursa ayrılma oranı azalabilir.

6. SONUÇ VE ÖNERİLER

Veri madenciliği, gizli, önemli, önceden bilinmeyen verileri analiz edip anlamlı sonuçlar elde etme tekniğidir. Diğer bir deyişle, veri yığınlarından geçerliliği olan ve uygulanabilecek verilerin elde edilmesindeki dinamik süreçtir. Bu teknikte sayısal veriler dışında, sayısal olmayan verilerle de analizler yapılıp anlamlı bilgi ortaya çıkarılabilmektedir. Bu işlemde yapay sinir ağları, istatiksel yöntemler, bellek tabanlı yöntemler ve karar ağaçları gibi yöntemler kullanılmakta ve bu yöntemler hızla gelişmektedir. Veri madenciliği algoritmalarını destekleyen pekçok program geliştirilmiştir, WEKA da bunlardan biridir.

Çalışmada, veri madenciliği yöntemlerini kullanarak, telekom sektörüne ait veriler analiz edildi. Hangi tip müşterilerin ayrılma eğilimi gösterdiği tahminlendi ve müşterilerin tercih ettikleri hangi hizmetten kaynaklı ayrılma eğilimi yaşandığı tespit edilmiştir. Bu analizler sonucunda, telekom firması uygun kampanyalar geliştirerek müşteri ayrılma eğilimini azaltabilir.

Benzer yapılan çalışmadan farklı olarak bu çalışmada, doğruluk oranına etki edecek nitelik seçim yöntemleri denendi ve InformationGain yönteminin sınıflandırmaya en çok etki eden nitelik seçimi algoritması olduğu belirlendi. Ham data ve nitelik seçimleri yöntemlerinden ayrı deney kümeleri oluşturularak, sınıflandırma algoritmaları karşılaştırıldı. Sonuç olarak, doğruluk oranı en yüksek olan J.48 algoritması kullanıldı. Bu algoritma ile model oluşturuldu ve ayrılma eğilimi gösteren müşterilerin tahminlemesi yapılmıştır.

Teknolojinin gelişmesi, veri madenciliğinin de gelişimine katkı sağlamış olup; büyük boyutlu verilerin analizinde veri madenciliğinin önemi ortaya çıkmıştır. Yaptığım çalışmayla veri madenciliği yaklaşımı, telekom fiması müşterileri üzerinde uygulanmış ve sonuçlar belirtilmiştir. Hangi tip müşterilerin ayrılma eğiliminde olduğu tahmin edilmiştir. Müşteri kazanımı, müşteri sadakati, müşteriyi elde tutma firmanın kazancı büyük önem teşkil etmektedir. Bu çalışmam, diğer sektörlerde ayrılma analizine ışık tutacak niteliktedir.

KAYNAKLAR

Akça, M., Churn Analizi ile Müşteri Kaybının Engellenmesi. Erişim Tarihi: 28.12.2016. http://mustafaakca.com/churn-analizi/

Akyol, K., Şen, B., Çalık, E., Biyokimya ve Hemogram Laboratuvar Test Sonuçlarının Lojistik Regresyon Yöntemiyle Analizi. Erişim Tarihi: 12.12.2016. http://ab.org.tr/ab12/bildiri/23.pdf

Amanmadov, N., Veri madenciliği sınıflandırma ve kümeleme teknikleri yardımıyla Wisconsin veriseti üzerinde Göğüs Kanseri Teşhisi. Erişim Tarihi: 29.09.2016. http://docplayer.biz.tr/15128771-Veri-madenciligi-siniflandirma-ve-kumeleme-teknikleri-yardimiyla-wisconsin-veriseti-uzerinde-gogus-kanseri-teshisi-hazirlayan-nury-amanmadov.html

Aydoğan, E., Gencer, C., Akbulut, S., (2009). Veri Madenciliği Teknikleri İle Bir Kozmetik Markanın Ayrılan Müşteri Analizi Ve Müşteri Bölümlenmesi. Mühendislik ve Fen Bilimleri Dergisi, (26), 43-57

Baykal, A., Veri Madenciliği Uygulama Alanları. Erişim Tarihi: 05.01.2017. http://zgefdergi.com/Makaleler/853277941_07_09_Baykal.pdf

Cihan, P., Kalıpsız, O., Öğrenci Proje Anketlerini Sınıflandırmada En Başarılı Algoritmanın Belirlenmesi. Erişim Tarihi: 05.01.2017. http://tbv.dergipark.gov.tr/download/article-file/207241

Çelik, U., Akçetin, E., Karınca Kolonisi Optimizasyonu Sınıflandırma Algoritması Yöntemi İle Telefon Bankacılığında Doğrudan Pazarlama Kampanyası Üzerine Bir Sınıflandırma Analizi. Erişim Tarihi: 10.10.2016. http://www.journalagent.com/iuyd/pdfs/IUYD-20592 RESEARCH_ARTICLE-AKCETIN.pdf

Çıngı, H., Veri Madenciliğine Giriş. Erişim Tarihi : 05.05.2017. yunus.hacettepe.edu.tr/~hcingi/ist376a/6Bolum.doc

Çölkesen, İ., Kavzoğlu T., Örnek Tabanlı K-star Algoritması İle Uzaktan Algılanmış Görüntülerin Sınıflandırılması. Erişim Tarihi: 10.10.2016. http://www.gtu.edu.tr/Files/UserFiles/80/jeodezi/yayinlar/pdf/KavzoglucolkesenTUFUAB2011.pdf

Dolgun, M., Özdemir, T., Oğuz, D., (2009). Veri madenciliğinde Yapısal Olmayan Verinin Analizi: Metin ve Web Madenciliği. İstatikçiler Dergisi, (2), 48-58.

Gülpınar, V. (2015). Yapay Sinir Ağları Ve Sosyal Ağ Analizi Yardımı İle Türk Telekomünikasyon Piyasasında Müşteri Kaybı Analizi. Marmara Üniversitesi Iktisadi ve Idari Bilimler Dergisi, 34 (1), 331-350.

Gündüz, S., Veri Madenciliği Temel Sınıflandırma Yöntemleri. Erişim Tarihi: 12.11.2016. http://web.itu.edu.tr/~sgunduz/courses/verimaden/slides/d3.pdf

İşler, Y, Narin, A., Weka Yazılımında k-ortalama Algoritması Kullanılarak Konjestif Kalp Yetmezliği Hastalarının Teşhisi. Erişim Tarihi: 29.11.2016. http://edergi.sdu.edu.tr/index.php/tbd/article/viewFile/3431/2918

Kotler, P., Pazarlama Yönetimi. Erişim Tarihi: 16.09.2016. http://www.perspectiva.md/ro/files/biblioteca/KotlerMarketing%20Management%20Millenium%20Edition.pdf

Kalikov, A., (2006). Veri Madenciliği ve Bir E-Ticaret Uygulaması, Yayımlanmamış Yüksek Lisans Tezi, Ankara: Gazi Üniversitesi, Fen Bilimleri Enstitüsü.

Odabaş, Ö., Kasapbaşı, M., (2016). Churn and Customer Segmentation Analyses with Data Mining Techniques for a Bookstore Company. International Journal of Innovative Technology and Exploring Engineering(IJITEE), 9, 1-6.

Rokach, L., Maimon, O., (2008). Data Mining with Decision Trees Theory and Applications. Machine Perception and Artificial Intelligence, 81, 100.

Swift, R., (2000). Accelerating Customer Relationship:Using CRM and Relationship Technologies", Prentice Hall PTR, 1, 1-512.

Şeker, S., E., Müşteri Kayıp Analizi. Erişim Tarihi: 16.09.2016. http://ybsansiklopedi.com/wp-content/uploads/2016/06/musteri_kayip.pdf

Şeker, Ş. E., Bilgi Kazanımı. 2012. Erişim Tarihi: 23.11.2016. http://bilgisayarkavramlari.sadievrenseker.com/2012/11/13/information-gain-bilgi-kazanimi/

Şimşek, G., Umman, T., (2010). Telekomünikasyon sektöründe müşteri ayrılma analizi. İstanbul Üniversitesi İşletme Fakültesi Dergisi, 39, 35-49.

Timor, M., Şimşek, T., (2008). Veri Madenciliğinde Sepet Analizi İle Tüketici Davranışı Modellemesi, İstanbul Üniversitesi, İşletme Fakültesi, Sayısal Yöntemler Anabilim Dalı.

Uslu, E.,Veri Madenciliği Nedir - Neden Veri Madenciği. Erişim Tarihi: 25.12.2016. http://emrahuslu.com/post/2011/09/25/Veri-Madenciligi-Nedir-Neden-Veri Madencigi.aspx

Vis, J., Zwet, R., (2009). Churn in Telecom dataset. Databases and Datamining,

ÖZGEÇMİŞ

Adı Soyadı : Özlem ODABAŞ

Doğum Yeri ve Yılı : ÜSKÜDAR, 15/06/1990

Medeni Hali : Bekar

Yabancı Dili : İngilizce

E-posta : ozlem.odabas90@gmail.com

Eğitim Durumu

Lise : Üsküdar Lisesi (YDA), 2008

Lisans : İstanbul Ticaret Üniversitesi, Mühendislik ve Tasarım Fakültesi, Bilgisayar Mühendisliği Bölümü, 2012

Mesleki Deneyim

Universal Bilgi Teknolojileri 2012-2014

Vodafone Teknoloji Hizmetleri A.Ş. 2014-…(devam ediyor)

Yayınları

Odabaş, Ö., 2016. Veri Madenciliği Teknikleri ile Kitapevi Firması İçin Ayrılan Müşteri Analizi ve Müşteri Bölümlenmesi. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2278, 1-6.