İNTERNET SERVİS SAĞLAYICISI İÇİN İPTAL ANALİZİ MODELİ

MEHMET GÖK

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ

TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ARALIK 2014

ANKARA

Fen Bilimleri Enstitü onayı

_____

Prof. Dr. Osman EROĞUL

Müdür

Bu tezin Yüksek Lisans derecesinin tüm gereksinimlerini sağladığını onaylarım.

_____

Doç. Dr. Erdoğan Doğdu

Anabilim Dalı Başkanı

Mehmet GÖK tarafından hazırlanan İNTERNET SERVİS SAĞLAYICISI İÇİN İPTAL ANALİZİ MODELİ adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

_____

Yrd. Doç. Dr. Tansel ÖZYER

Tez Danışmanı

Tez Jüri Üyeleri

Başkan : Yrd. Doç. Dr. Esra Kadıoğlu Ürtiş _____

Üye : Yrd. Doç. Dr. Tansel ÖZYER _____

Üye : Yrd. Doç. Dr. Çetin Ürtiş _____

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada orijinal olmayan her türlü kaynağa eksiksiz atıf yapıldığını bildiririm.

Mehmet Gök

Üniversitesi : **TOBB Ekonomi ve Teknoloji Üniversitesi**
Enstitüsü : **Fen Bilimleri**
Anabilim Dalı : **Bilgisayar Mühendisliği**
Tez Danışmanı : **Yrd. Doç. Dr. Tansel ÖZYER**
Tez Türü ve Tarihi : **Yüksek Lisans – Aralık 2014**

Mehmet GÖK

İNTERNET SERVİS SAĞLAYICISI İÇİN İPTAL ANALİZİ MODELİ

ÖZET

İptal analizi müşterilerin davranış örüntülerinin modellenerek, gelecekte iptal eğilimi gösteren aboneler hakkında öngörülerin belirlendiği müşteri ilişkileri yönetimi sürecidir. Yeni müşterinin kazanımı, mevcut müşterinin sistemde tutulmasından çok daha fazla maliyetlidir. Bu bağlamda iptal analizi ile yapılan tahminler mevcut müşterinin iptale gitmemesi için yapılacak tutundurma faaliyetlerine yardımcı olmaktadır. Günümüzde telekomünikasyon firmaları iptal analizini çeşitli uygulamalarla sistemli bir süreç halinde iyileştirerek sürdürmektedirler. Bu çalışmada da telekomünikasyon sektöründe faaliyet gösteren bir internet servis sağlayıcısının müşteri bilgileri ve davranışları incelenerek gerçekleştirilmiştir. Yapılan literatür araştırmaları sonucunda belirlenen bir bilgi keşif süreci çerçevesinde veri madenciliği uygulamalarının yardımı ile iki fazlı çözüm modeli oluşturulmuştur. Geliştirilen iki fazlı çözüm modeli zaman serisi kümeleme ve sınıflandırma algoritmaları ile birlikte en uygun çalışacak şekilde tasarlanmıştır. Zaman serisi kümeleme uygulaması için k-ortalama ve hiyerarşik kümeleme algoritmaları, sınıflandırma için ise destek vektör makineleri ve özyinelemeli bölümleme algoritmaları karşılaştırmalı olarak performans ölçütleri değerlendirilmiştir.

Anahtar Kelimeler: Müşteri ilişkileri yönetimi, iptal analizi, bilgi keşfi, veri madenciliği, zaman serisi kümeleme, k-ortalama kümeleme, hiyerarşik kümeleme, sınıflandırma, destek vektör makineleri, özyinelemeli bölümleme

University : **TOBB Economics and Technology University**
Institute : **Institute of Natural and Applied Sciences**
Science Programme : **Computer Engineering**
Supervisor : **Assistant Associate Professor Tansel ÖZYER**
Degree Awarded and Date : **M.Sc. – December 2014**

Mehmet GÖK

CHURN PREDICTION FOR INTERNET SERVICE PROVIDER

ABSTRACT

Churn prediction is a customer relationship process that specifies predictions for customers who are inclined to churn in future through modelling customer behavior patterns. It costs more to acquire a customer than to retain a customer. In this sense, the predictions which are made with churn prediction support promotion activities executed to avoid subscription cancellation of existing customers. Nowadays, telecommunication companies maintain churn prediction with various applications as a systematic process. Also this thesis is written on the basis of customer data and behavior analysis of an internet service provider operating in telecommunication sector. Within the knowledge discovery process framework, explored as a result of realized literature survey, two phased solution model is created with the help of data mining applications. Developed two phased solution model is designed to run effectively with time series clustering and classification algorithms. Performance indicators are evaluated comparatively with respect to k-means, hierarchical clustering algorithms for time series clustering and support vector machines, recursive partitioning for classification algorithms.

Keywords: Customer relationship management, churn prediction, knowledge discovery, data mining, time series clustering, k-means clustering, hierarchical clustering, classification, support vector machines, recursive partitioning

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Yrd. Doç. Dr. Tansel ÖZYER'e yine kıymetli tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü öğretim üyelerine ve bu süreçte motivasyon anlamında bana destek olan sevgili eşim Müne Duygu GÖK'e teşekkürü bir borç bilirim.

İÇİNDEKİLER

ÖZET ..... **IV**

ABSTRACT ..... **V**

TEŞEKKÜR ..... **VI**

İÇİNDEKİLER ..... **VII**

ÇİZELGELERİN LİSTESİ ..... **IX**

ŞEKİLLERİN LİSTESİ ..... **X**

KISALTMALAR ..... **XI**

SEMBOL LİSTESİ ..... **XII**

1. GİRİŞ **1**

1.1 Problem Tanımı **2**

1.2 Araştırma Amacı **2**

1.3 Telekomünikasyon Endüstrisinde İptal Analizinin Önemi **3**

2. LİTERATÜR ARAŞTIRMASI **4**

2.1 Müşteri İlişkileri Yönetimi (CRM): Ana Kavramlar **4**

2.2 Literatürde Veri Madenciliği Uygulamaları **7**

2.3 İptal Analizi Uygulamaları **9**

2.4 Zaman Dizisi Kümeleme Uygulamaları **10**

2.4.1 Zaman Dizisi Eşleme 11

2.4.2 Zaman Dizisi Uzaklık Ölçüm Yöntemleri 12

2.4.3 Zaman Serisi Kümeleme Algoritmaları 12

2.5 Sınıflandırma Uygulamaları **15**

2.5.1 Destek Vektör Makineleri 16

2.5.2 Özyinelemeli Bölümleme 16

2.6 Verinin Anlamlandırılması **17**

2.6.1 Ortalama ve Yeniden Ölçeklendirme 18

2.6.2 Yüzdelik Alma 18

3. BİLİMSEL ARAŞTIRMA YÖNTEMİ **19**

3.1 Bilimsel Araştırma Yaklaşımı **19**

3.1.1 Nitel ve Nicel Yaklaşım 19

3.1.2 Tümevarım ve Tümdengelim Yaklaşım 21

3.2 Bilimsel Araştırma Süreci **21**

3.2.1 Problemin Belirlenmesi 22

3.2.2 Verinin Belirlenmesi 23

3.2.3 Veri Hazırlanması 25

3.2.4 Veri Madenciliği 29

3.2.5 Sonuçların Değerlendirilme Yöntemi 30
4. ANALİZ VE SONUÇLAR 34
4.1 Deney I: Basit Sınıflandırma İşlemi 34
4.2 Deney II: İki Fazlı Çözüm Modeli 37
4.2.1 K-ortalama Kümeleme Uygulaması 41
4.2.2 Hiyerarşik Kümeleme Uygulaması 42
4.3 Deney III: Hiyerarşik İF Çözüm Modeli'nin Uygulamasına Yapılan Ekler 45
4.3.1 Küme Merkezlerine Olan Uzaklıklar 46
4.3.2 Öznitelik Eliminasyonu 47
4.4 Deney IV: İF Çözüm Modelinin Gerçek Veri İle Denenmesi 48
4.5 Özniteliklerin Değerlendirilmesi 49
5. SONUÇ VE YAPILABİLECEK DİĞER ÇALIŞMALAR 53
5.1 Sonuç 53
5.2 Yapılabilecek Çalışmalar 55
KAYNAKLAR 57
ÖZGEÇMİŞ 59

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge 2.1 Güdümlü ve Güdümsüz Veri Madenciliği Uygulama Örnekleri 8

Çizelge 3.1 Nicel ve Nitel Araştırma 20

Çizelge 3.2 Ham Veri Yapısı 24

Çizelge 3.3 Ortalama ve Ölçeklemenin Öklidyen Uzaklığına Olan Etkisi 27

Çizelge 3.4 Özelliklerin Korelasyon Matrisi 28

Çizelge 3.5 Hata Matrisi (Confusion Matrix) 32

Çizelge 4.1 Her Katta Hesaplanan F-Ölçüsü 35

Çizelge 4.2 Basit Sınıflandırma Prosedürü 36

Çizelge 4.3 Alım Yönü Endeksi İçin Örnek Veri 38

Çizelge 4.4 İki Fazlı Çözüm Modeli 39

Çizelge 4.5 K-Ortalama Küme Sayısı Test Sonuçları 41

Çizelge 4.6 Hiyerarşik Zaman Serisi Kümeleme Algoritması 43

Çizelge 4.7 İptal Etmeme Kararı Verilen En İyi Beş Kural 51

Çizelge 4.8 Öznitelik Betimleyici İstatistikleri 51

Çizelge 4.9 İptal Etme Kararı Verilen En İyi Beş Kural 52

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil 2.1 Müşteri Yaşam Döngüsü Süreci 6

Şekil 2.2 Zaman Dizisi Örnekleri 10

Şekil 2.3 Zaman Dizisi Eşleme Problemi Gösterimi 11

Şekil 2.4 DVM En Geniş Ayrım Gösterimi 16

Şekil 2.5 Ani Kalp Durması Verisi 17

Şekil 3.1 Altı-Adım Bilgi Keşif Süreci 22

Şekil 3.2 Ortalama ve Ölçeklemenin Veriye Etkisi 26

Şekil 3.3 Özelliklerin Korelasyon Dağılım Grafiği 29

Şekil 3.4 İptal Analizi İçin Örnek Veri Setleri 31

Şekil 4.1 Deney I Sonuçları 37

Şekil 4.2 K-ortalama Algoritması İle Deney II Sonuçları 42

Şekil 4.3 Örnek Hiyerarşik Kümeleme Dendrogramı 44

Şekil 4.4 Hiyerarşik Kümeleme Sonuçları 45

Şekil 4.5 Küme Merkezlerine Olan Uzaklıklar 46

Şekil 4.6 İF Çözüm Modelinin Gerçek Veri Sonuçları 48

Şekil 4.7 Özyinelemeli Bölümleme Örnek Karar Ağacı 50

KISALTMALAR

Kısaltmalar	Açıklama
DVM	Destek Vektör Makineleri (Support Vektör Machines)
ÖYB	ÖzYinelemeli Bölümleme (Recursive Partitioning)
İSS	İnternet Servis Sağlayıcı
CRM	Müşteri İlişkileri Yönetimi (Customer Relationship Management)
CART	Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees)
ROC	Alıcı İşletim Karakteristiği (Reciever Operating Characteristic)
İF Çözüm Modeli	İki Fazlı Çözüm Modeli
PCA	Prensip Bileşen Analizi (Principal Component Analysis)

SEMBOL LİSTESİ

Bu çalışmada kullanılmış olan simgeler açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	Açıklama
$\boldsymbol{nitelik_{ij}}$	Müşterinin statik parametreleri
$\boldsymbol{davranış_{ijk}}$	Müşterinin zaman serisi türünde davranış parametreleri
$\boldsymbol{stDavranış_{ijk}}$	Ortalanmış ve ölçeklenmiş zaman serisi türünde davranış parametreleri
ö$\boldsymbol{lçü mK}$ü$\boldsymbol{mesi}$	Ölçüm parametre kümesi
$\boldsymbol{davranışK}$ü$\boldsymbol{mesi}$	Davranışsal zaman serisi türleri
$\boldsymbol{dk_{ij}}$	Davranışsal zaman serilerinin küme sonuçları
$\boldsymbol{performans_{ij}}$	Her kat için performans kümesinin sonuçları
$\boldsymbol{uzaklikMatrisi_{ii}}$	Objelerin birbirlerine olan uzaklıkları matrisi
$\boldsymbol{merkezMatrisi_{kj}}$	Objelerin küme merkezlerine olan uzaklıkları matrisi
$\boldsymbol{kumeSonucu_{i}}$	Objelerin küme bilgisi

1. GİRİŞ

Her iş içerisinde yeni müşteri kazanımı olmasının yanı sıra mevcut müşterilerin tutundurma faaliyetleri de önemlidir. Telekomünikasyon sektörü gibi iş pazar hacminin büyümesi günden güne azalan doygun sektörler için yeni müşterinin kazanılması yerine mevcut müşterilerin tutundurulması hem maliyet açısından hem de pazar payının korunması ve artırılması açısından araştırma konusu olarak çekici olmasına neden olmaktadır.

Her ne kadar iptal analizi üzerine gerçekleştirilen araştırmalar telekomünikasyon endüstrisinde mobil ve sabit hatlarda konuşma üzerine yaygınlaşmış olsa da internet servis sağlayıcılar için gerçekleştirilen araştırmalar son yıllarda yaygınlaşmaktadır.

Sektör bazında düşünüldüğünde pazarın doygunluğa ulaşması ve pazardaki rekabetin artması, müşteri tutundurma faaliyetlerini önemli hale getirmektedir. Şirketlerin hem prestij açısından, hem de maliyet açısından müşterilerin tutundurulması önemlidir. Sadece maliyet açısından değerlendirdiğimizde, yapılacak küçük ancak yerinde tutundurma faaliyetlerinin maliyetleri yeni kazanılacak müşterilerin pazarlama, satış ve kurulum maliyetleri açısından çok daha düşük olduğu görülecektir. Özellikle telekomünikasyon sektöründe rekabetin fazla olduğu bugünlerde kaybedilen müşterilerin maliyeti oldukça yüksek seviyelerdedir. Kaybedilen müşterilerin maliyetlerini azaltmak isteyen ve doymuş sektörde varlığını sürdürmek isteyen firmaların tutundurma faaliyetleri kapsamında iptal analizi gibi bir karar destek mekanizmasına ihtiyaçları bulunmaktadır.

Buradaki sorun telekomünikasyon sektörü gibi pazarda çevik hareket edilen sektörlerde müşteriye ait verilerin çokluğu ve karmaşasıdır. Bu karmaşanın içerisinden iptal analizi için gerekli müşteri verisinin çıkarılması ve uygulanacak çözüm metodolojisine uygun bir şekilde veri yorumlanabilmelidir. Müşterinin davranışlarının yer aldığı herhangi bir veri uygun bir şekilde yorumlandığı takdirde iptal analizi için değerli olmalıdır.

1.1 Problem Tanımı

Müşterileri iptale veya hizmet/mal almamaya götüren nedenler şüphesiz pazarda yerini korumak isteyen ve çevik hareket ederek müşteriyi kazanmak isteyen şirketler için bir araştırma ve aynı zamanda gelir kaynağı olarak değerlendirme konusu olmuştur. Diğer tüm sektörler bir yana, telekomünikasyon sektöründe aylık iptal oranı %2.2'lerdedir [1]. Yıllık olarak düşünüldüğünde yıllık hiç yeni abonenin olmaması durumunda toplam müşteri sayısının dörtte biri civarında bir kayıp söz konusudur. Bu durum şu şekilde özetlenebilir; "İptal oranının bu kadar yüksek olması sızdıran bir kaba su koymaya çalışmak gibidir". Problem tam olarak firmada gerçekleşen iptal oranıdır. Yüksek olması abone kazanmak için harcanan eforun boşa gittiğinin göstergesi olabilir. [2]

İptal oranının yüksek olması canlı bir popülasyonla karşılaştırıldığında ölüm oranının yüksek olması ile eş değerdir. Eğer doğum oranı ölüm oranından düşükse bu durum popülasyonun azalmasına sebep olacaktır. Sektör içerisinde bir firma için düşünülecek olursa firmanın bu durumda müşteri potansiyelini koruyamadığı ve gün geçtikçe müşteri sayısında düşüş olduğu görülür. Bu çalışmanın problemi doğum oranını artırmak değil ölüm oranını azaltmaktır. Böylece yeni müşteri kazanmak için harcanılan maliyet boşa gitmemelidir.

1.2 Araştırma Amacı

Araştırmanın amacı bir İSS (İnternet Servis Sağlayıcı) şirketin müşteri bilgileri ve zaman içerisinde gerçekleşen hareket verileri kullanarak kabul edilebilir bir seviyede iptal edebilecek abonelerin tahmininin yapılabileceği modeli kurgulamaktır. Bu çalışmanın çıktıları tutundurma süreçlerine temel teşkil etmelidir. Bu bağlamda bilgi keşif süreci kapsamında veri seçilmesi, ön hazırlığının yapılması, kullanılacak veri madenciliği uygulaması, performans kriterleri, ölçümleme ve iyileştirme işlemlerinin belirlenmesi amaçlanmaktadır. Yapılacak aktiviteler sonucunda, veriyi bilgiye

dönüştürebilen, veri madenciliği uygulamalarını etkin kullanabilen ve faydasının ölçümlenebileceği bir modelin oluşturulması hedeflenmektedir.

1.3 Telekomünikasyon Endüstrisinde İptal Analizinin Önemi

Günümüzde iletişim ihtiyaçları çoğu zaman ekstra bir maliyet olarak değil ihtiyaç olarak değerlendirilmektedir. Telekomünikasyon sektörü, bu sektör için tam rekabet piyasasına sahip ülkelerde çoğu zaman doygunluk eğilimindedir. Aslında başka bir deyişle pazarın büyüme hızı gün geçtikçe daha az bir ivmeye yönelmektedir. Bu değerlendirmeler ışığında yeni abone kazanmak gün geçtikçe zorlaşmaktadır. Bu durum şirketleri iptal analizi gibi çalışmalar gerçekleştirerek sistemde tutmak için faaliyetler gerçekleştirmeye yönlendirmektedir. Rekabet ortamında iptal oranının aylık %2.2'lerde [1] olduğu düşünülürse bu faaliyetlerin ne kadar önemli olduğunu tahmin etmek zor olmayacaktır. Aşağıdaki nedenler dikkate alındığında iptal analizinin telekomünikasyon sektöründe neden önemli olduğu anlaşılmaktadır;

- Aylık %2.2 oranında müşterilerin iptal etmesi demek yıllık %25 civarında iptalin olduğunu göstermektedir.
- Bir telekomünikasyon firması için her yıl yaşanan iptallerle cironun çeyreği gizli bir maliyet oluşturmaktadır.
- Tutundurma maliyeti yeni müşteri kazanma maliyetine göre 5 kat daha az maliyetlidir. [3]

Müşterilerin abonelik ömürleri ne kadar uzun olursa firma için o kadar kar olduğu düşünülmektedir. Firmalar kısa dönem müşteri ilişkileri yerine uzun dönemli sözleşme, kampanya veya tarifeleri tercih etme eğilimindedir. Sadık müşterilerin değerli olmasından ötürü sadakatin oluşturulabilmesi için iptal analizi uygulanmaktadır.

2. LİTERATÜR ARAŞTIRMASI

2.1 Müşteri İlişkileri Yönetimi (CRM): Ana Kavramlar

Müşteri ilişkileri yönetimine karşı ilginin 1990'lı yıllarda başladığı belirtilmektedir. İş dünyasında CRM sözcüğü sıkça kullanılmaktadır [4]. Fakat süreçlerin herhangi bir yerinde müşterinin dahil olması ile CRM sözcüğü yer alabilmektedir. CRM için yaygın olarak kabul görmüş bir tanım olmamakla birlikte çeşitli tanımlar yapılmaktadır.

CRM belirli müşterilerle uzun vadeli ve karlı bir ilişki kurulmasını destekleyen sistemlere olanak sağlayan iş stratejisi ve süreçler bütününü oluşturmaktadır. [4] En basit hali ile CRM bir davranış, zihniyet, işinize katmış olduğunuz değer ve bunların müşteri ile ilişkisidir. Pazar payında ve aynı zamanda müşterinin aklında organizasyonunuzun oluşmasını ve gelişmesini sağlayan metodolojilerdir. [5] CRM, müşteri kazanımını, müşteri tutundurmayı, müşteri sadakatini müşteriden elde edilen karı iyileştirebilmek için anlamlı iletişim yolları ile müşteri davranışlarını anlamak ve ona tesir etmek için uygulanan kurumsal bir yaklaşımdır. [6] Yukarıdaki tanımlamaların ortak özelliği müşteri ilişkileri yönetiminin organizasyonun faydasına müşteri ile gerçekleşecek etkileşimlerde kullanılacak strateji, süreç ve metodoloji olduğudur.

Berry ve Linoff [7] Şekil 2.1'de gösterildiği üzere müşterileri 5 ana gruba ayırmaktadır;

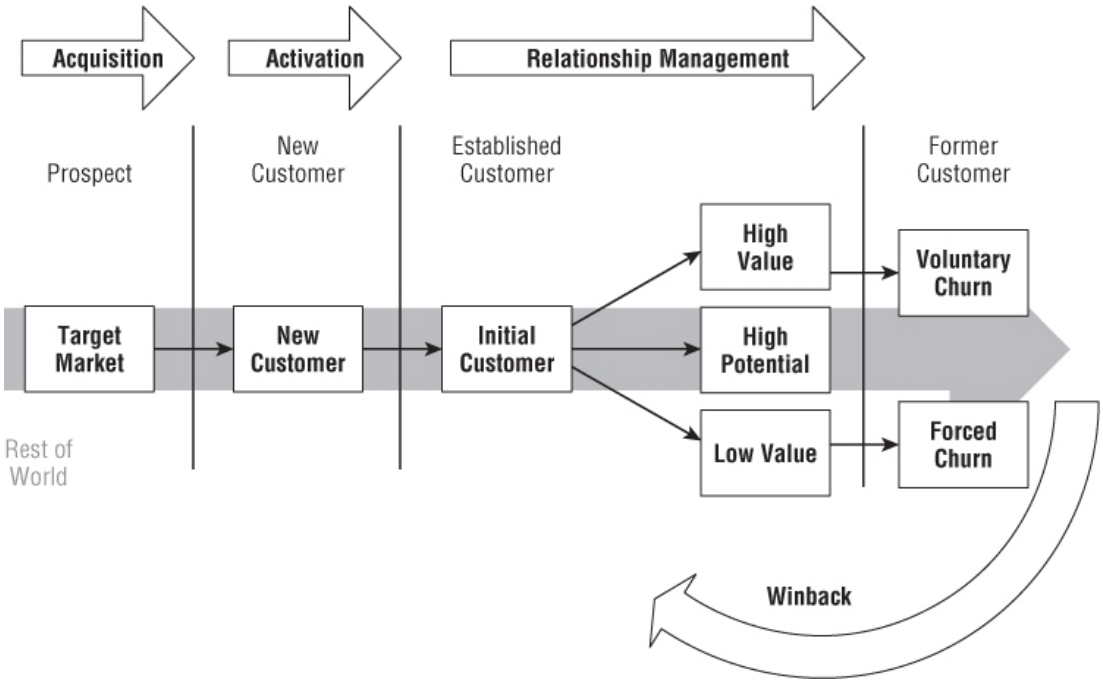
1. **Olası (Prospects):** Hedef market içerisinde olan fakat henüz kazanılmamış müşteriyi belirtmektedir.
2. **Cevap Veren (Responders):** Olası müşterilerin içerisinde ilgisi çekilebilen müşterilerdir. Genelde çevirim içi form doldurulması, satış anketlerine

katılmak veya ürün hakkında bilgi almak için herhangi bir yol ile iletişime geçen müşterilerdir.

3. **Yeni Müşteri (New Customers):** Bir sözleşme ile taahhüt altına giren veya müşteri olmak için form dolduran ilk satışın yapıldığı müşterilerdir.
4. **Kazanılan Müşteri (Established Customers):** Bu müşteri segmenti ile firmanın ilişkileri gelişmiştir. Daha fazla satışın yapılabildiği ve ilişkinin derinleştiği müşteriler olup süre anlamında da uzun olan müşterilerdir.
5. **Eski Müşteri (Former Customers):** Artık müşteri olmayan eski müşterilerdir. Bunlar ya gönüllü olarak ayrılmışlardır (başka bir rakip müşterinin ilgisini çekmiş olabilir veya müşteri ürün üzerinde daha fazla değer görmüyordur), ya zorla ayrılmaları sağlanmıştır (faturalarını ödemiyor olabilir), ya da beklenen bir ayrılma işlemi (taşınma gibi zorunlu haller doğrultusunda) gerçekleşmiştir.

Burada müşteri tanımları yapılan işe göre farklılık arz edebilir. Yapılan tanımlar telekomünikasyon sektörü için değerlendirilecek olursa;

- Olası müşteriler rakip firmalarda hizmetlerini alan ve firmanın hizmet verebileceği ancak kazanılmamış bütün müşterilerdir.
- Cevap veren müşteriler çağrı merkezini hizmet, kampanya, tarife gibi bilgileri almak için firma web sitesine girmiş, çağrı merkezini aramış durumda olan tüm müşterilerdir.
- Yeni müşteriler hizmet almak için satış kanallarının herhangi biri yoluyla müracaatta bulunmuş, dijital hizmeti açılarak (eğer kurulum gerekiyorsa kurulumu yapılarak) faturalama yapılabilen müşterilerdir.
- Kazanılan müşteriler aslında mevcut yerini koruyan ve daha fazla kar edilebilen müşterilerdir. Bu tanımdan yola çıkılarak up-sell (pahalı tarife satışı) ve cross-sell (cihaz veya destekleyici hizmet satışı) yapılarak daha fazla ürün satılabilen müşterilerdir.
- Eski müşteriler hizmetini kapatmış veya borçtan iptal durumunda kapatılmış müşterilerdir.



Şekil 2.1 Müşteri Yaşam Döngüsü Süreci (Kaynak: [7])

Müşterilerin telekomünikasyon sektöründeki durumları ele alındığında iptal analizi için araştırma konusu olabilecek müşteriler eski müşteriler, yeni müşteriler ve kazanılan müşterilerdir. Eski müşterilerin eğilimleri incelenerek aynı örüntüde olan yeni ve kazanılan müşterilerin kaybedilmeden tespit edilebilmesi ve iptal nedenlerin ortaya konulabilmesi sektör bağımsız bütün firmalar için nitelikli ve katma değeri yüksek bilgilerdir.

Müşteri yaşam döngüsü süreci içerisinde kazanılan müşterilerin özellikle yüksek gelir getirenlerinin geri kazanımı şirketler için önem taşımaktadır. Kazanılan müşteriden eski müşteriye geçişlerin nedenleri örnek olarak Şekil 2.1'de gösterilmiştir. Ancak bu ayrım zenginleştirilerek müşteri ayrımı yapılabilmelidir. İptal analizinin başka bir boyutu da iptal nedenlerinin belirlenmesidir. İptal nedenlerinin ve davranış gruplarının belirlenmesi yapılacak tutundurma faaliyetlerine yardımcı olmalıdır. Ancak bu çalışma kapsamında iptal nedenlerinin araştırılması yoktur. Bu çalışmanın ışığında yapılabilecek bir araştırma konusu olarak değerlendirilebilir.

2.2 Literatürde Veri Madenciliği Uygulamaları

Veri madenciliği sürecini bir teknik süreç olarak değerlendirerek iş problemlerini tanımlamaktan öteye iş problemlerini veri madenciliği problemlerine dönüştürmeye taşımaktadır. [7] Bu aşamada işin kendisi ile ilgilenmek yerine probleme çözüm üretecek modelin nasıl bir teknik süreçten geçirileceği değerlendirilmelidir.

Aslında tüm veri madenciliği görevleri iki ayrı kategoride değerlendirilmektedir: betimleyici, kestirimci. [8] Bu ifadeler şu şekilde açıklanabilir;

1. Betimleyici: Veri tabanındaki verilerin karakteristiklerini çıkararak özet bilgi veren uygulamalardır.
2. Kestirimci: Öngörüde bulunabilmek için mevcut veri üzerinden çıkarsama yapılmasıdır.

Bu iki grup hakkında daha detaylı bilgi verilmesi gerekirse literatürde daha çok betimleyici istatistik ve kestirimci modelleme şeklinde yer almaktadır. Ortalama, medyan, standart sapma ve sapkın gözlem tespiti gibi istatistiksel işlemler betimleyici istatistik araçlarıdır. Betimleyici istatistiğin yanı sıra kestirimci analitik işlemleri de çözüm üretmek üzere birçok araç sunmaktadır. Kestirimci analitik istatistik, modelleme, veri madenciliği ve makine öğrenme tekniklerini, güncel ve tarihsel veriyi çalışmak için kullanır. Böylece analistlere gelecek hakkında öngörü verebilme olanağı sağlar. [9]

Liao ve arkadaşları [10] 2000 ve 2011 yılları arasında yazılan makalelerden derledikleri çalışmada veri madenciliği tekniklerini uygulamaları ile birlikte değerlendirerek bilgi tipi, analiz tipi ve mimari tiplerine göre sınıflandırmışlarıdır. Buna göre veri madenciliği teknikleri dokuz ayrı grupta değerlendirilmektedir. Sinir ağları, algoritma mimarisi, dinamik kestirimci, sistem mimarisi analizi, akıllı etmen

sistemleri, modelleme, bilgi bazlı sistemler, sistem optimizasyonu ve bilgi sistemleridir.

Veri madenciliği basit bir açıklama ile; elde bulunan verilerden yola çıkarak işe yarar örüntüler tanımlayıp gelecek hakkında öngörü edinmeyi sağlar. Edinilecek öngörü belirtilen durumun gerçekleşeceği anlamına gelmez. Çıktı olarak üretilen öngörüyü üretmek için Berry ve Linoff tarafından üç farklı yöntem belirtilmektedir;

- Hipotez testi
- Güdümlü Veri Madenciliği
- Güdümsüz Veri Madenciliği

Hipotez testindeki amaç veriyi bir soruya yanıt aramak veya genel anlamı çıkarmak için kullanmaktır. Güdümlü veri madenciliğinde amaç bir veya birden fazla hedef değişkeni tahmin eden veya açıklayan modeli geliştirmektir. Güdümsüz veri madenciliğinde ise değişken bağımsız veride bulunan örüntüleri çıkarmak amaçlanmaktadır. Güdümlü ve güdümsüz veri madenciliği uygulamaları Çizelge 2.1.'de örneklenmiştir. [7]

Çizelge 2.1 Güdümlü ve Güdümsüz Veri Madenciliği Uygulama Örnekleri

Güdümlü V.M. Uygulama Örnekleri	Güdümsüz V.M. Uygulama Örnekleri
- Sınıflandırma - Kestirim - Tahminleme	- Kümeleme - Görselleştirme - İlişki Kuralı Madenciliği

Literatür içerisinde veri madenciliği uygulamaları ile ilgili birçok sınıflandırma çeşitleri bulunmaktadır. Aynı zamanda sadece veri madenciliği uygulamalarının sınıflandırılması üzerine birçok makale bulunmaktadır.

2.3 İptal Analizi Uygulamaları

Bu kısımda literatürde yer alan iptal analizi uygulamaları özetlenecektir. Bakış açısı kazandırması anlamında aynı konu üzerinde hangi yöntemlerin denendiğinin bilinmesi faydalı olacaktır.

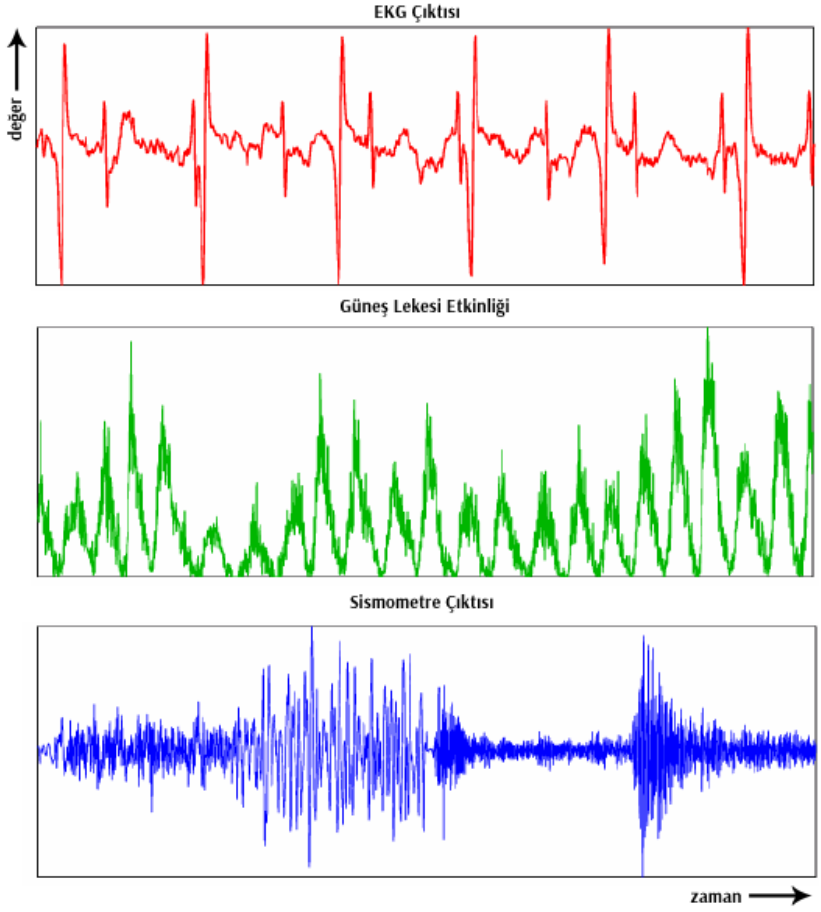
İlk uygulama [11] telekomünikasyon sektörü üzerinde yapılan bir çalışmadır. Demografik bilgilerin yetersizliğinden ötürü abonelik sözleşmesi ve görüşme detayları iptal analizinde veri olarak kullanılmıştır. Deneylerinden edindikleri sonuçlar ışığında önerilen karar ağaçları ile çalışan model daha hızlı uygulanabilirliği ve verinin eğitiminde geçen zamanın daha iyi olduğundan yapay sinir ağları ile kıyaslandığında daha verimli olduğu belirtilmiştir.

Bir diğer makale ise [12] gazete aboneliğinin verileri kullanılarak gerçekleştirilmiş bir uygulamadır. Bu uygulamada ise iptal analizi için iki parametre seçim tekniği ve destek vektör makineleri tekniği kullanılmıştır. Gerçekleştirilen deneylerin bir parametre optimizasyonuna gereksinim duyduğu görülmüş ve bir optimizasyon prosedürü geliştirilmiştir. Destek vektör makinelerinin parametreleri doğru verilmediği takdirde iyi sonuç vermediği belirtilmiştir.

Müşterilerin yatay davranış verileri, durağan verileri ile birlikte genelde değerlendirilememektedir. Öngörünün performansını artırmak için özellikle yatay davranış verileri durağan verilere dönüştürülerek işlem yapılmaktadır. [13] Belirtilen çalışmada çözüm önerisi olarak sunulan modelde klasik destek vektör makineleri yerine hiyerarşik çoklu çekirdekli destek vektör makineleri adı altında bir algoritma önerisinde bulunulmuştur. Belirtilen algoritma girdi olarak müşterinin hem durağan bilgilerini almakta hem de yatay davranış verilerini almaktadır. Üç ayrı fazda öğrenme işlemini tamamlayan algoritma öznitelik seçme işlemi de yapmaktadır. Birçok algoritmanın karşılaştırıldığı bu makalede belirtilen uygulamalar çeşitli parametre ve veri setlerinde uygulanarak sonuca ulaşılmıştır.

2.4 Zaman Dizisi Kümeleme Uygulamaları

Uygulama olarak zaman dizisi kümeleme algoritmaları günümüzde birçok problemde kullanılmaktadır. Zaman dizisi kullanımına sağlık, borsa, yerbilim uygulamaları, makine durum gözlemleme, mekan-zamansal veri uygulamaları gibi alanlar örnek olarak verilebilir [14]. Şekil 2.2.'de gösterilen sağlıkta kullanılan kalp kasının ritmi, meteorolojide kullanılan güneş lekesi etkinliği ve rasathanelerde kullanılan sismometrenin çıktısı zaman serisi kullanımlarına örnek teşkil etmektedir.



Şekil 2.2 Zaman Dizisi Örnekleri (Kaynak: [15])

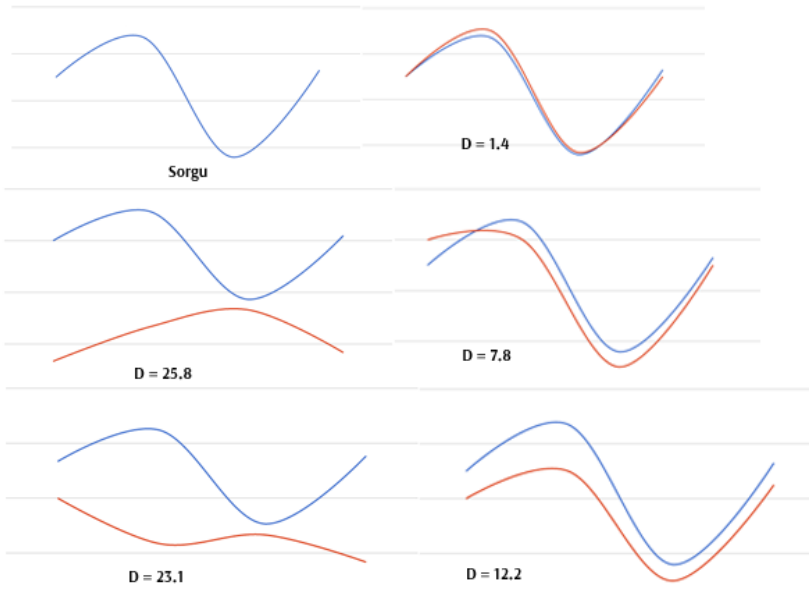
Literatür'e bakıldığında görüntülerden videolara çok çeşitli veri tiplerinin zaman dizisi şeklinde ifade edilebildiği ve çalışmalarda kullanıldığı görülebilir. Aşağıdaki farklı türdeki veri örnekleri zaman dizisi olarak nasıl dönüştürüldüğünü göstermektedir;

- Görüntü eşlemede renk histogramının zaman dizesine dönüştürülerek kümeleme uygulamasının gerçekleştirilmesi. [16]
- Kaplumbağa ve boynuzlu kertenkele kafatası fotoğraflarının dış yüzey şekillerinin zaman serisine dönüştürülmesi [17]
- Hareket algılayıcılar sayesinde hareketli noktaların 2 boyutlu ve 3 boyutlu zaman dizisi olarak yerlerinin tespiti ile animasyonların gerçekleştirilmesi

örneklerinde olduğu gibi farklı türde veri kaynakları zaman dizisi olarak ifade edilebilir.

2.4.1 Zaman Dizisi Eşleme

Zaman dizisi eşleme problemi bir zaman dizisinin diğer zaman dizilerine olan uzaklığı veya benzerliğini belirleme işlemidir. Ortaya çıkan yakınlık ve benzerlik ölçümleri zaman serileri arasındaki ilişkiyi belirler ve gruplamaya yardımcı olur. Şekil 2.3.'te gösterildiği üzere sorgu olarak belirlenen zaman dizisi üzerinden veri tabanında bulunan diğer zaman dizeleri üzerinde sorgulama yapılmıştır. Problemin sonucu olarak benzer olan zaman dizilerinin uzaklıkları daha az çıkmaktadır.



Şekil 2.3 Zaman Dizisi Eşleme Problemi Gösterimi

2.4.2 Zaman Dizisi Uzaklık Ölçüm Yöntemleri

Literatürde aynı düzlemde yer alan iki zaman serisi arasındaki uzaklığı ölçmek için özellikle dinamik zaman kayması ve en uzun alışılmış alt sözcük gibi çeşitli uzaklık ölçümlerinden bahsedilmektedir. [15] Dinamik zaman kayması zaman serisinin belirli bölgesinde gürültülü bir veri olması durumunda bunu egale ederek pürüzsüz bir şekilde asıl trende yoğunlaşarak uzaklıkları elde etmektedir. En uzun alışılmış alt sözcük algoritmasında ise zaman serisinin bazı parçalarında iki ayrı seride benzerlik gösteren kısımlar olması durumunda daha yakın olacağını varsayarak işlem yapılmaktadır. Bu iki yöntem de zaman serilerini şekilsel bazda değerlendirerek işlemlerini gerçekleştirirler. Bu ölçümler zaman serilerinde özellikle bazı karakterdeki verilerde oldukça başarılı olduğu gibi bazılarında ise uygulamanın performansını düşürecek şekilde uygulamayı yönlendirebilmektedir.

$$u_{\text{ö}} = \sqrt{\sum_{k=1}^{p} (x_{ik} + v_{ik})^2} \tag{2.1}$$

Diğer uygulamaların yanı sıra zaman serisi uzaklık ölçümlerinde literatürde çokça ismi geçen uygulamalardan Öklidyen uzaklık ve Öklidyen olmayan uzaklık uygulamaları da kullanılabilmektedir. Denklem 2.1 Öklid uzaklığını vermektedir.

2.4.3 Zaman Serisi Kümeleme Algoritmaları

Diğer tüm veri madenciliği uygulamalarında olduğu gibi benzerlik veya uzaklık üzerine sorgulama yapılabilen veri kümelerinde sınıflandırma ve kümeleme algoritmalarının pratiklerini çalışmak mümkündür. Zaman serisi sınıflandırma ve kümeleme üzerine literatürde birçok makale bulmak mümkündür.

Zaman dizisi kümeleme algoritmaları genel olarak iki ana formülasyon altında toplanabilir [18] ;

- *İlişkisel bazda online kümeleme*: Belirlenen farklı veri kümeleri üzerinden gerçek zamanlı olarak gelen verilerin sınıflandırılması için kullanılmaktadır. Daha çok borsada önceden belirlenen bazı davranışların gerçek zamanlı bilgilere dayalı olarak benzerlik gösterip göstermediği değerlendirilir.
- *Şekilsel bazda offline kümeleme*: benzer görünümdeki zaman dizisi verilerinin gruplanmasında kullanılmaktadır. Benzerlik göreceli bir kavram olduğundan burada kullanılacak benzerlik fonksiyonu önem teşkil etmektedir.

Online kümeleme bir akış üzerinde daha çok borsa gibi sürekli değişen ve anlık analiz gerektiren yerlerde kullanılmaktadır. Problem tanımı kapsamında incelenmesi gereken bölüm şekilsel bazda offline kümeleme algoritmalarıdır. Şekilsel bazda kümeleme işlemi için benzerlik veya uzaklık ölçüm yöntemi önemlidir. Liao, zaman dizilerinin arasındaki benzerlik ve uzaklık ölçümü yöntemlerini 3 ayrı grupta değerlendirmiştir. [19] Bunlar;

- Ham Veri Bazlı
- Özellik Bazlı
- Model Bazlı

Ham veri bazlı kümeleme doğrudan zaman serisini alır ve kümeleme işlemini gerçekleştirir. Özellik bazlı kümeleme ise zaman serisinden özellik seçimi yapıldıktan sonra kümeleme işlemini gerçekleştirir. Model bazlı kümelemede de kümeleme işlemi öncesinde bir modelleme işlemi gerçekleştirilir. Modelleme işlemi ile kalan parametre ve katsayılar kullanılarak kümeleme işlemi tamamlanır.

Formülasyon olarak şekilsel bazda ve kurgu olarak da ham veri bazlı kullanılabilecek offline kümeleme tekniklerinden öne çıkan hiyerarşik kümeleme ve k-ortalama kümeleme uygulamalarıdır.

2.4.3.1 K-ortalama Kümeleme Algoritması

K-ortalama kümeleme algoritması yaklaşık 35 yıl önce 1979'da Hartigan ve Wong tarafından yayınlanmıştır. [20] Ancak günümüzde birçok uygulamada değişik versiyonları kullanılmaktadır. K-ortalama kümeleme algoritmasının amacı N boyutlu M noktayı kümeler içerisinde kareler toplamı minimum olacak şekilde K kadar kümeye bölümlemektir. Algoritma noktaları K küme $S = \{S_1, S_2, S_3, \dots, S_k\}$ arasında geçişlerini sağlayarak yerel optimum değeri bulmaya çalışmaktadır. Buna göre denklem 2.2'deki amaç fonksiyonunu minimize etmeye çalışmaktadır. Her S_i kümesinin elemanlarının ortalaması μ_i'yi belirtmektedir.

$$\text{minimize} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{2.2}$$

2.4.3.2 Hiyerarşik Kümeleme Algoritması

Zaman serisi kümeleme algoritmalarından bir diğeri de hiyerarşik kümeleme algoritmasıdır. Bu algoritmanın başlangıcı ve ilk tanımı Ward tarafından 1963 yılında yapılan algoritma çok geniş olarak kullanılmaktadır. [21]

Hiyerarşik kümeleme algoritması veriyi (burada zaman serisi) kümelerden oluşan bir ağaç yapısında gruplamaya çalışmaktadır. [19] Literatürde iki tip hiyerarşik kümeleme algoritması bulunmaktadır. Bunlardan ilki yığınsal hiyerarşik kümeleme, diğeri ise bölen hiyerarşik kümelemedir. Yığınsal, bölenden çok daha popülerdir. Her

obje kendi kümelerine atanarak başlanır. Belirlenen küçük kümeler birbirleri ile birleştirilerek daha büyük kümeler elde edilmiş olur. Tüm objeler tek bir sınıf oluşturuncaya kadar veya belirli koşulları sağlayana kadar sınıflar birleştirilir. Bu tekil (bütün) bağlantı algoritması iki sınıf arasındaki benzerliği, aradaki uzaklığın en kısa olana göre sınıflandırma işlemini gerçekleştirmektedir.

Ward tarafından tanımlanan algoritma da yığınsaldır ve iki kümeyi birleştirirken varyansların kareleri toplamında olacak değişimin minimum olmasına bakılmaktadır. Varyansların kareleri toplamı hesabı her küme için gerçekleştirilir ve minimum hesabı çıkartılmaktadır. Bu da yığınsal algoritmaların karmaşıklıklarını $O(n^3)$ yapmaktadır. Büyük veriler için süreyi oldukça uzatmaktadır.

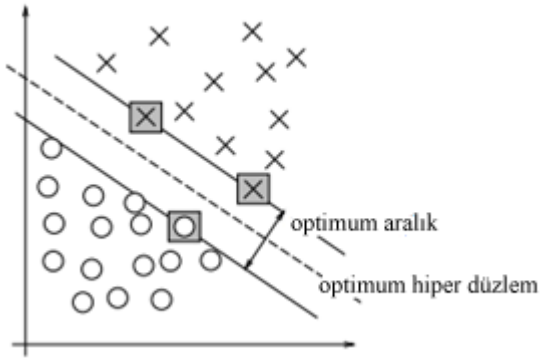
2.5 Sınıflandırma Uygulamaları

Sınıflandırma uygulamaları eğitim verisi ile öğrenme işlemini yaparak sonucu bilinmeyen objelerin öğrenilenlere göre hangi sınıfta olabileceğini tahmin eden ve verileri tamamıyla ayrı gruplara yerleştiren uygulamalardır. Literatürde karar ağacı, K-yakın komşu, Naive Bayes, destek vektör makineleri, özyinelemeli bölümleme ve sınıflandırma için genetik algoritma gibi çeşitli uygulama alanları bulunan sınıflandırma algoritmaları mevcuttur. Aslında her birinin yapmış olduğu iş verinin davranışını öğrenerek bir öngörü seti içerisinde öğrenilen veri davranışları üzerinden tahminini gerçekleştirmektir. Ancak her birinin çeşitli uygulamalarda verinin karakteristiğine ve uygulanan modele göre davranışları değişebilmektedir.

Geniş sınıflandırma literatürü içerisinden son günlerde popüler olan iki yöntem seçilerek çözüm modeli içerisinde yer verilmiştir. Bunlardan ilki destek vektör makineleri ve diğeri ise sınıflandırma için öz yinelemeli bölümleme uygulamalarıdır.

2.5.1 Destek Vektör Makineleri

Destek vektör makineleri (DVM) Cortes ve Vapnik tarafından 1995'te ikili sınıflandırma yapmak üzere geliştirilmiştir. Şekil 2.4'te görülebileceği gibi DVM ile iki ayrı sınıfın en yakın noktaları arasında destek vektörleri tanımlanmaktadır. Algoritma tanımlanan destek vektörleri arasındaki uzaklığın en fazla olduğu sonucu bulmaya çalışmaktadır. Bu destek vektörleri arasından geçtiği düşünülen optimum hiper düzlem ise sınıflandırmanın öğretisi olarak alınmaktadır. Yapılacak testlerde ayrım amacıyla belirlenen optimum hiper düzlem kullanılarak sınıflandırma yapılmış olur.



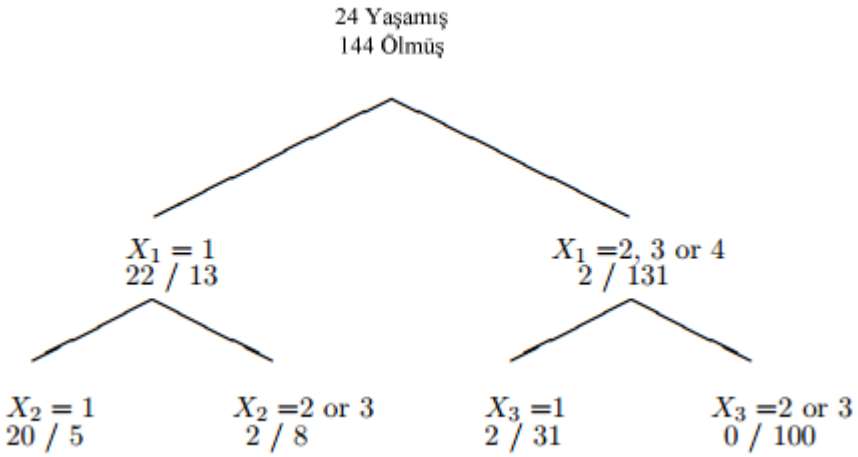
Şekil 2.4 DVM En Geniş Ayrım Gösterimi (Kaynak: [22])

Belirlenen bu algoritmanın yanı sıra üzerinde yapılan çalışmalar neticesinde lineer bir hiper düzlem bulunamayan durumlarda veriyi daha büyük bir düzlem uzayında tanımlayarak gerçekleştirilen çözümler bulunmaktadır.[23]

2.5.2 Özyinelemeli Bölümleme

Özyinelemeli bölümleme (literatürde "RPART – Recursive Partitioning" olarak geçmektedir) ismi aslında çoğu uygulamalarında sınıflandırma ve regresyon ağaçları (literatürde "CART – Classification and Regression Trees") olarak geçmektedir. tekniklerinden esinlenerek geliştirilmiştir. Literatürde CART olarak geçen ve bir

yazılımın parçası olarak marka haline gelen CART ismi bu yüzden kullanılmamıştır. [24]



Şekil 2.5 Ani Kalp Durması Verisi (Kaynak: [24])

Özyinelemeli Bölümleme (ÖYB) algoritması iki fazdan oluşmaktadır. [24] İlk fazda Şekil 2.5'te gösterildiği gibi bir ikili ağaç oluşturarak başlanmaktadır. Öncelikle hangi değişkenin veriyi ikiye en iyi ayırdığı tespit edilir. Yeni bir iyileştirme yapılamayana kadar gruplara ayrılmış veri içerisinde özyinelemeli olarak minimum büyüklüğe ulaştığında durdurulur. İkinci faz ise çalışmayı durdurma işleminin nasıl tespit edilmesi gerektiğidir. Çapraz doğrulama işlemi yapılarak bir risk değeri hesaplanır. Belirlenen risk faktörü en az olacak şekilde ağaç budanır ve en az riski taşıyan alt ağaç sınıflandırmada dikkate alınır.

2.6 Verinin Anlamlandırılması

Bu bölümde verinin daha iyi gösterimi üzerine aslında günlük hayatta dahi sıkça fark edilmeden kullanılan bazı uygulamalardan bahsedilecektir. Gerçekleştirilen tüm makine öğrenme algoritmaları veri ile çalışmaktadır. Girdi olarak algoritmanın anlayacağı dilden bir veri olmadığı takdirde işlemlerin çalışmadığı, eksik veya yanlış çalıştığı görülecektir.

2.6.1 Ortalama ve Yeniden Ölçeklendirme

Ortalama veya merkezileştirme işlemi her bir değerden tüm değişkenlerin ortalamasını çıkarıldığında elde edilen yeni özelliktir. Bu yeni işlenmiş değerin özelliği ortalamasının 0 olması ve orijinal değişken ile aynı ölçek özelliklerini taşımasıdır. Bu işlem ortalamayı bilmeyi gerektirmeksizin verinin yorumlanmasını sağlar. Bu işlem her ne kadar yeni bilgi vermese de daha kolay anlaşılmayı sağlamaktadır.

Yeniden ölçeklendirme işlemi ise ortalanmış verilerin standart sapmaya bölünmesi ile gerçekleştirilmektedir. Sonuç olarak ortalaması 0 ve standart sapması 1 olan veri kümesi elde edilmiş olunur.

2.6.2 Yüzdelik Alma

Ebeveynler arkadaşlarına genellikle bebeklerinin uzunluk, ağırlık veya baş çevresi ölçülerinin yüzde 95'lik dilimde olduğunu söylemeyi severler. [7] Çünkü burada anlamlı olan kısım normale ne kadar yakın olduğu bilgisidir. Bebeğin uzunluğunun ne kadar olduğu aslında tek başına yetersiz bir bilgidir. Burada bebeğin yaşına göre olması gereken sınır değer bilgisi ile yüzdelik alınması işlemi sayesinde aslında bebeğin ne kadar normal gelişim gösterdiği bilgisi elde edilmiş olur. Mantıklı değerlerin bu şekilde oranlanması katma değeri yüksek veriler sağlayabilmektedir.

3. BİLİMSEL ARAŞTIRMA YÖNTEMİ

Bu bölümde bilimsel araştırma sınıfları içerisinde yapılacak olan araştırmanın nerede yer aldığını ve çözüm aşamasında nasıl yol izleneceği konusunda genel bilgiler verilecektir. Verinin nasıl anlamlandırılacağı, hangi süreçlerin gerçekleştirileceği ve hangi araçları kullanarak araştırma stratejisinin uygulanacağı konusunda değerlendirilmelere yer verilecektir. Araştırma metodolojisi verinin nasıl toplanacağı ve analiz edileceğini belirler. Bu doğrultuda araştırma sürecinin adımların daha detaylı belirtilmesi amaçlanmaktadır.

3.1 Bilimsel Araştırma Yaklaşımı

Sonuca ulaşmak için yapılan çalışmaların bilimsel araştırma yaklaşımı ana fikrinin değerlendirmesi ve belirlenmesi bu bölümde yapılmıştır.

3.1.1 Nitel ve Nicel Yaklaşım

Nicel yaklaşım, ölçülebilen değerler üzerinden bahsedilen teorinin değerlendirilmesi, sayılarla ifade edilebilmesi ve istatistiksel tekniklerin kullanımı ile gerçekleştirilmektedir. Nicel yaklaşımın amacı teorinin doğruluğunu sayılarla ispat etmektir. Nitel yaklaşım için kabul gören genel bir açıklama yapmak mümkün olmamakla birlikte gözlem, görüşme ve doküman analizi gibi nitel veriler ışığında açıklamaları, deneyimleri, nedensellikleri, düşünce ve görüşleri değerlendirmektir.

Nitel ve nicel yaklaşımın özellik bakımından amaç, örneklem, veri toplama, veri analizi ve çıktıları arasında önemli farklar bulunmaktadır. Çizelge 3.1'de bu karşılaştırmalar bulunabilir.

Çizelge 3.1 Nicel ve Nitel Araştırma (Kaynak: [25])

Nicel Araştırma	**Nitel Araştırma**
Varsayım	
- Gerçeklik nesneldir - Asıl olan yöntemdir - Değişkenler kesin sınırlarıyla saptanabilir ve bunlar arasındaki ilişkiler ölçülebilir - Araştırmacı olay ve olgularla dışarıdan bakar, nesnel bir tavır geliştirir	- Gerçeklik oluşturulur - Asıl olan çalışılan durumdur - Değişkenler karmaşık ve iç içe geçmiştir ve bunlar arasındaki ilişkileri ölçmek zordur - Araştırmacı olay ve olguları yakından izler, katılımcı bir tavır geliştirir
Amaç	
- Genelleme - Tahmin - Nedensellik ilişkisini açıklama	- Derinlemesine betimleme - Yorumlama - Aktörlerin bakış açılarını anlama
Yaklaşım	
- Kuram ve denence ile başlar - Deney, manipülasyon ve kontrol - Standardize edilmiş veri toplama araçları kullanma - Parçaların analizi - Uzlaşma ve norm arayışı - Verilerin sayısal göstergelere indirgenmesi	- Kuram ve denence ile son bulur - Kendi bütünlüğü içinde doğal - Araştırmacının kendisinin veri toplama aracı olması - Örüntülerin ortaya çıkarılması - Çokluluk ve farklılık arayışı - Verinin, derinliği ve zenginliği içinde betimlenmesi
Araştırmacı Rolü	
- Olay ve olguların dışında, yansız ve nesnel	- Olay ve olgulara dahil, öznel bakış açısı olan ve empatik

Bu değerlendirmeler ışığında abonelerin iptal analizi gibi veri madenciliği konuları kendine has istatistiksel veri, değerlendirme, analiz, yöntem ve çıktılar içerdiğinden veri madenciliği ile ilgili yöntemlerin geliştirilmesi nicel bilimsel araştırma konuları arasında değerlendirilmelidir.

3.1.2 Tümevarım ve Tümdengelim Yaklaşım

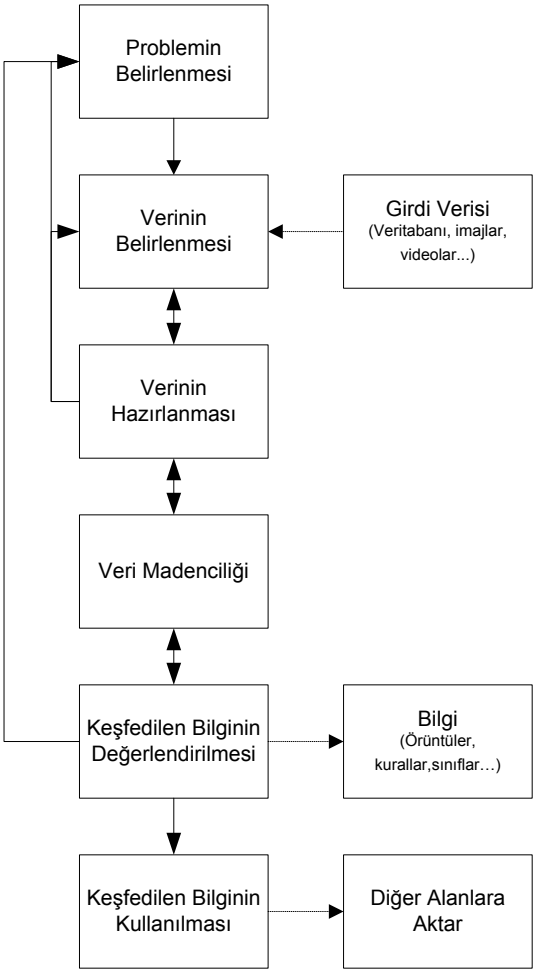
Tümevarım (aşağıdan yukarıya) araştırma, özel bir önermeden genel bir önermeye gidişi sağlayan düşünce biçimidir. Tümdengelim (yukarıdan aşağıya) araştırma ise genel bir önermeden yola çıkarak hipotezin doğruluğunu ortaya koymak amaçlanmaktadır.

Bu araştırma tümevarım ve tümdengelim yaklaşımlarını abonelerin davranışları ile veri madenciliği tekniklerini kullanarak bir model geliştirmeyi ve modelin sonuçlarını değerlendirmeyi amaçlayarak kullanmaktadır. Veriler tamamen abone davranışlarından elde edilerek sayısal veya kategorik biçimde ifade edilmesi yöntemi ile eğitim ve test verileri oluşturulacaktır.

3.2 Bilimsel Araştırma Süreci

Veri madenciliği, bilgi keşfi ve veri tabanında bilgi keşfi terimlerinde bazen karışıklık olduğundan öncelikle bu terimlerin tanımının yapılması gerekmektedir. Her ne kadar bazı araştırmacılar veri madenciliğini bilgi keşfi ile eş anlamda kullansa da; veri madenciliği, bilgi keşfi sürecinin sadece bir aşamasıdır. [26]

Bu tanımlardan yola çıkıldığında bu tez çalışması bir bilgi keşif çalışmasıdır ve Şekil 3.1'de belirtilen altı adım bilgi keşif süreci [27] kullanılarak yapılacağı ve iteratif süreci takip edeceği konusunu bu bölümde belirtmek gerekmektedir.



Şekil 3.1 Altı-Adım Bilgi Keşif Süreci (Kaynak: [27])

3.2.1 Problemin Belirlenmesi

Bilgi keşfi sürecinin ilk aşaması problem anlayışının ortaya konulmasıdır. Burada alan bilgisi olan kişilerle çalışılmalı ve çalışmaya konu olacak ve çözüm için model üretilecek problemin belirlenmesi gerekmektedir. Veri madenciliğinin hedefleri ve iş amaçları belirlenmelidir.

Daha önce de değinildiği üzere telekomünikasyon sektöründe iptal eden müşterilerin geri kazanılması için yapılacak çalışmalara girdi olarak iptal olasılığı yüksek abonelerin tespit edilmesi ile müşteri segmentasyonunun yapılması amaçlanmaktadır. Burada problem belirli bir davranış sergileyen müşterilerin davranış örüntülerinin

belirlenebilmesidir. Bu davranış örüntülerinin tespit edilmesi ile iptal sürecine gidebilecek abonelerin ayırt edilebilmesi gerekecektir.

3.2.2 Verinin Belirlenmesi

Bu aşamada örneklem verinin toplanması, katma değeri olabilecek verilerin tespit edilmesi, verinin formatı, büyüklüğü ve alabileceği değerlerin belirlenmesi gerekmektedir. Ayrıca, verinin bütünlüğü, artıklığı, eksikliği, güvenirliliği gibi verinin kalitesini etkileyen etmenler kontrol edilmelidir.

Her satır bir aboneliği işaret etmekle birlikte örneklem veri için 6.000 civarında abonelik seçilmiştir. Gerçek veri testleri için bölgesel olarak pilot bir bölge seçilerek yaklaşık 70.000 abone verisinin toplanılması sağlanmıştır. Gerçek veri olarak bahsedilen veri canlı örnekler olup iptal örneklerini çoğaltmak ve kısıtlamak adına çalışmanın yapıldığı günden 1 yıl öncesine kadar hizmetini iptal ettirmiş aboneler seçilmiştir. Örneklem veri içinse iptal oranını koruyacak şekilde gerçek veri içerisinden küçük bir örneklem seçilmiştir.

Bu çalışmada kullanılacak ham veri işletme tarafından kullanılan PostgreSQL veritabanından PL/pgSQL kullanılarak elde edilmiştir. Bu kadar büyük bir verinin canlı bir sistem üzerinden tek sorgu ile çekilmesi çok mümkün görünmediğinden SQL içerisinde Limit ve Offset kullanılarak JAVA ile sayfalama tekniği ile verilerin loglanması sağlanmıştır. Veri analizlerinin birkaç format değişikliği dışında tamamı R üzerinde gerçekleştirilmiştir. Hazırlanan işlenmemiş veri tipleri Çizelge 3.2'de detaylı bir şekilde gösterilmiştir.

Çizelge 3.2 Ham Veri Yapısı

Veri Adı	Veri Türü	Alınan Değerler	Açıklama
Hizmet Tekil Numarası	Integer	Tam Sayı	Müşteriye ait hizmete verilen tekil numara
Cinsiyet	Kategorik	Erkek, Bayan	Hizmet sahibinin cinsiyeti
Yaş	Integer	(18,)	Hizmet sahibinin yaşı
Kota Grubu	Kategorik	Kotalı, Adil Kullanımlı, Limitsiz	3 ayrı gruptan oluşan internet hizmetinin kota tipi
Kota Limiti	Integer	Tam Sayı	Kotalı ve adil kullanımlı aboneler için anlamlı olan download kotası (Birimi: GB)
Hizmet Hızı	Integer	Tam Sayı	İnternet hizmetinin hızı (Birimi: GB)
Taahhüt	Boolean	True: Taahhütlü False: Taahhütsüz	Hizmetin iptal anında veya analiz yapılan günde taahhüdünün olup olmadığı bilgisi
Toplam Gecikme	Integer	Tam Sayı	Faturalarında yapmış olduğu gecikme miktarı
Abonelik Yaşı	Integer	Tam Sayı	Abonenin ne kadar süredir abone olduğu (Birimi: Gün)
Alım Yönü Endeksi	Zaman Serisi (Nümerik)	Rasyonel Sayı	Abonenin son aylarda yapmış olduğu harcamalar. (6 aylık dönem)
Download Miktarı	Zaman Serisi (Nümerik)	Rasyonel Sayı	Abonenin son günlerde yapmış olduğu download miktarı (10'ar günlük dönemler halinde son 2 ay)
Arıza Sayısı	Zaman Serisi (Integer)	Tam Sayı	Abonenin son günlerde açmış olduğu arıza sayıları (10'ar günlük dönemler halinde son ay)
İptal Tercihi	Boolean	True: İptal Abone False: Devam Eden Abone	Abonenin iptal tercihini belirtmektedir.

3.2.3 Veri Hazırlanması

Ham veri belirli işlemlerden geçmeden veri madenciliği uygulamalarına hazır hale getirilemez. Bu aşamada hangi verinin kullanılacağına karar verilir. Veri anlamlı hale getirilerek veri madenciliği uygulamalarında kullanılmak üzere işlemlerden geçirilir. Sadece ham veri kullanılarak yapılacak işlemlerle veri daha anlamlı hale getirilebilir. Ortalama, yeniden ölçeklendirme, yüzdelik dilim belirleme ve kategorik verilerin anlamlı bir şekilde nümerik verilere dönüştürülmesi gibi birçok işlem uygulanabilir.

Öte yandan korelasyon testi, önemlilik testi, örneklem alma ve veri bütünlüğü, gürültü temizliği ve eksik verilerin uygun bir yöntemle eklenmesi gibi veri temizliği gerektiren işlemler de bu aşamada gerçekleştirilmelidir. Temizlenen veri daha sonra yeni verilerin üretilmesi, özelliklerin seçilmesi, yeni özelliklerin türetilmesi gibi bir çok işlemden geçirilebilir.

3.2.3.1 Verinin Anlamlandırılması

Literatür araştırmasında bahsedildiği üzere bu aşamada veriyi daha anlamlı kılmak için bazı işlemler gerçekleştirilmiş ve sonuca faydası olduğu görülmüştür. Bunlardan ilki yüzdelik almak olarak özetleyebileceğimiz kullanım bilgisinin elde edilmesi ile gerçekleştirilmektedir.

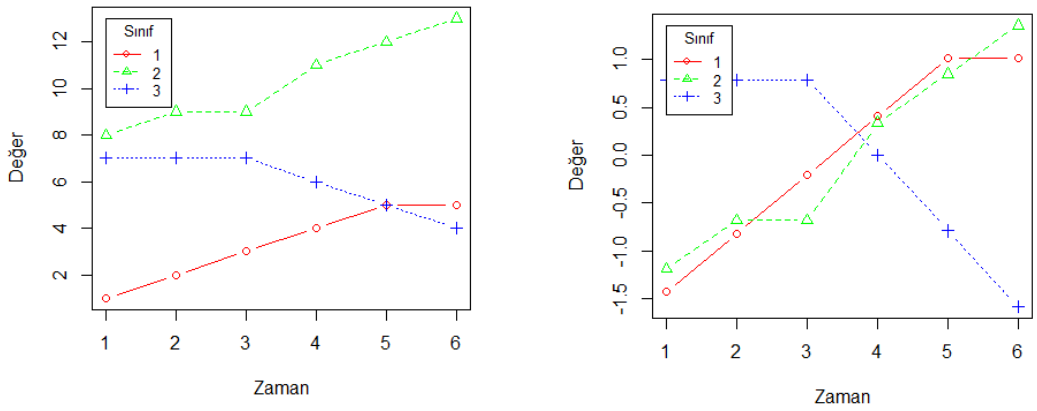
$$kullanım = \frac{download\ miktarı}{hız * 0.1029} \tag{3.1}$$

Burada belirtilen çarpan abonenin 10 günde en fazla ne kadar download yapabileceğini göstermektedir. Böylece ham veride yer alan iki veriden abonenin kullanım miktarını belirleyen bir özellik elde edilmiştir. Belirlenen kullanım özelliği

10 günlük periyotlar halinde abonenin hat kapasitesinin % kaçını kullandığını belirtmektedir ve bu özellik de bir zaman serisi olarak kullanılacaktır. Verinin anlamlandırılması için verilen yeni yüzdelik bilgi daha anlamlı ve kullanışlı olacağı düşünülmektedir.

Verinin daha fazla anlamlandırılması için kullanılan ikinci yöntem ortalama ve yeniden ölçeklendirme işlemidir. Burada gerçekleştirilecek işlem satırlar üzerinden verinin standart hale getirilerek istatistiksel olarak eşit koşullarda analize girmelerini sağlamak amaçlanmaktadır. Bu işlem özellikle zaman serisi olarak ifade edilen veriler üzerinde gerçekleştirilmektedir. Bu işlem sayesinde zaman serisi grafikleri eşit koşullarda değerlendirilebilir ve her müşterinin davranışı birbirine yakınlık olarak aynı ölçekte değerlendirilebilir.

Bu aşamada ortalama ve ölçeklendirme işleminin zaman serisi verileri üzerinde nasıl katkıda bulunduğu ve ne gibi etkileri olduğu küçük bir örnek ile açıklanmaktadır. Şekil 3.2'de 3 ayrı gözlemin ortalama ve ölçekleme öncesi ve sonrası şekilsel olarak nasıl göründüğü belirtilmektedir.



Ham Veri Ortalama ve Ölçekleme Sonrası

Şekil 3.2 Ortalama ve Ölçeklemenin Veriye Etkisi

Şekil 3.2'de 1. ve 2. gözlemin artan bir yönelim gösterdiği ve şekilsel olarak birbirine çok yakın olduğu görülmektedir. Tabii ki ham veri kendi başına önemli bilgiler içermektedir, ancak veri işlendikten sonra literatürde bahsedilen DTW metodu ile zaman serileri üzerinde aynı skala üzerinden değerlendirme yapıldığında şekilsel olarak önemli bilgiler edinilmektedir. Çizelge 3.3'de görüleceği üzere 1. ve 2. gözlem ham veri ile 57 birim uzaklıkta iken işleme sonrası uzaklık 2.12'ye düşmektedir.

Çizelge 3.3 Ortalama ve Ölçeklemenin Öklidyen Uzaklığına Olan Etkisi

Ham Veri Uzaklıkları			Ortalama ve Ölçekleme Sonrası Uzaklık	
	1	2	1	2
2	57		2.12	
3	26	43	11.55	12.54

3.2.3.2 Verinin Davranış Analizi

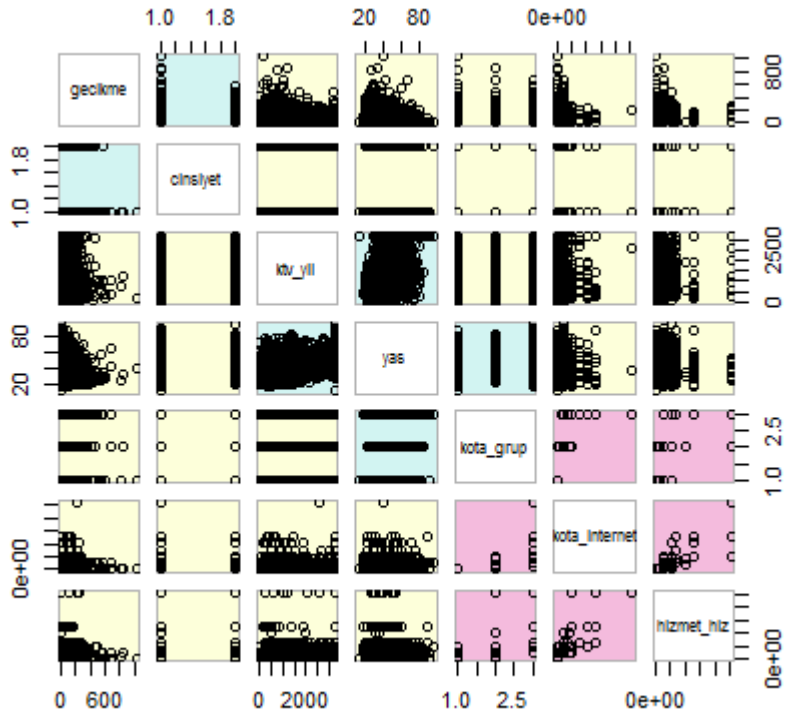
Verinin daha anlamlı kılınmasının yanı sıra verinin nasıl bir dağılım gösterdiği de analiz aşaması öncesinde incelenmesi gereken bir konudur. Bu bağlamda veri üzerinde korelasyon testleri gerçekleştirilmelidir.

Gerçekleştirilen korelasyon testlerinde Çizelge 3.4'de belirtilen şekilde bir korelasyon matrisi bulunmaktadır. Veriyi yorumlamak adına korelasyon matrisi oldukça iyi fikir verir. Öncelikle yüksek korelasyon bulunan alanlar olarak aralarında 0,5'in üzerinde negatif veya pozitif korelasyonu olan veriler incelenebilir. Şekil 3.3'de daha net anlaşılabileceği üzere kota grubu, internet kotası ve hizmet hızı arasında yüksek korelasyon bulunmaktadır. Abonelik tarifeleri belirli bir örüntüde olduğundan bu durum beklenen bir durumdur ancak gereksiz bir bilgi olduğu anlamına gelmemelidir.

Çizelge 3.4 Özelliklerin Korelasyon Matrisi

	Cinsiyet	**Yaş**	**Kota Grubu**	**İnternet Kotası**	**Hizmet Hızı**	**Gecikme Toplamı**	**Abonelik Yaşı**
Cinsiyet	1,000	-0,039	-0,085	-0,092	-0,079	0,032	-0,022
Yaş	-0,039	1,000	0,010	-0,038	-0,047	-0,192	0,491
Kota Grubu	-0,085	0,010	1,000	0,761	0,596	-0,137	-0,020
İnternet Kotası	-0,092	-0,038	0,761	1,000	0,744	-0,086	-0,062
Hizmet Hızı	-0,079	-0,047	0,596	0,744	1,000	-0,058	-0,074
Gecikme Toplamı	0,032	-0,192	-0,137	-0,086	-0,058	1,000	-0,182
Abonelik Yaşı	-0,022	0,491	-0,020	-0,062	-0,074	-0,182	1,000

Yapılan deneyler içerisinde başarımı artırıp artırmadığına bakılması amacıyla korelasyonu yüksek olan veriler ayıklanarak da işlem yapılabilir. Burada gerçekleştirilen çalışma örnek teşkil etmesi amacıyla müşteriden elde edilen ham verinin analiz edilmesi ile ortaya çıkmıştır. Daha detaylı çalışma veri madenciliği uygulamalarını çalıştırmadan önce elde edilen işlenmiş veri üzerinde yapılmalıdır. Son olarak uygulamanın bir de korelasyonu yüksek verilerin çıkarılması ile denenmesi sonuçları nasıl etkilediğini görmek için gerekli bir yöntemdir.



Şekil 3.3 Özelliklerin Korelasyon Dağılım Grafiği

3.2.4 Veri Madenciliği

Öncelikle yapılan çalışmada kullanılacak uygulamaların literatürde nerelere değdiği ve sınıflandırmalardan hangilerine dahil olduğu belirtilebilir. Bu çalışma kapsamında ham veri üzerinde hem betimleyici istatistik araçları hem de kestirimci analitik araçları kullanılmıştır. Ayrıca güdümlü ve güdümsüz veri madenciliği modelleri kullanılmıştır. Güdümsüz uygulamalar kullanılarak müşterilerin zaman içerisindeki davranışları analiz edilerek müşteri kümeleme uygulamaları gerçekleştirilmiştir. Güdümsüz uygulamaların sonuçları güdümlü uygulamalar içerisinde kullanılarak iptal analizi gerçekleştirilmiştir.

Uygulanan teknik iki ayrı fazda açıklanabilir. İlk fazda zaman serisi kümeleme uygulamaları yer almaktadır. Bu fazda müşterinin en son hareketleri incelenir ve örüntüler çıkarılır. Bu işlem bir yandan ham zaman serisi verisi için yapılırken, diğer

yandan da daha önce de bahsedilen ortalama ve ölçekleme işlemi sonrasında elde edilen zaman serisi verisi için yapılır. Sonuç olarak iki ayrı kümeleme işlemi farklı anlamlar ve gruplar içermektedir. Veri kümeleme işlemi ile zenginleştirilmiş olur. Çözüm modelinde iki farklı kümeleme algoritması karşılaştırılmalı olarak kullanılmaktadır. Bunlar;

1. K-ortalama Kümeleme Algoritması
2. Hiyerarşik Kümeleme Algoritması

İkinci fazda ise zenginleştirilmiş verinin sınıflandırılması işlemi gelmektedir. Sınıflandırma işleminin yaptığı iş basit olarak anlatmak gerekirse iki sınıflı bir kümeye oluşturulan verilerin örüntülerine göre atama yapma işlemidir. Bu işlem için yine çalışmanın literatür kısmında detaylı bir şekilde anlatılan iki ayrı sınıflandırma uygulaması denenmektedir. Bunlar;

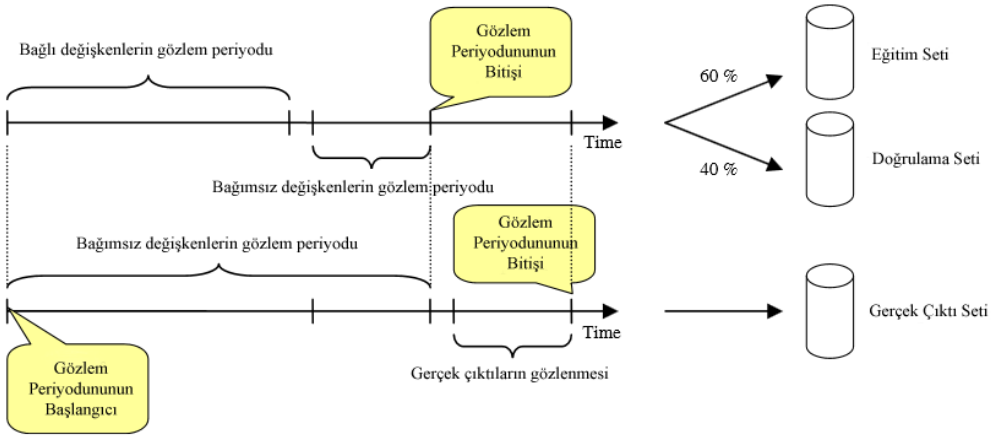
1. Destek Vektör Makineleri (DVM - Support Vector Machines)
2. ÖzYinelemeli Bölümleme (ÖYB - Recursive Partitioning)

İki fazdan oluşan model üzerinde farklı yöntemlerle çalıştırmaya da imkan sağlamaktadır. Zaman serisi kümeleme işleminden elde edilen verinin farklı şekilde kullanımları farklı yöntemleri değerlendirme açısından çalışmaya yön verecektir.

3.2.5 Sonuçların Değerlendirilme Yöntemi

Uygulanacak iki fazlı işlem sonrasında elde edilecek çıktıların hassasiyeti üzerinde ölçüm yapılabilir. Çıktılara geçmeden önce analiz aşamasında verinin nasıl ayrıştırılarak yöntemin belirlenmesi gerekmektedir. Burez ve Van den Poel finansal veya tecimsel iptallerin analizi için yapmış olduğu çalışmada [28] Şekil 3.4'de belirtilen şekilde bir değerlendirme yöntemi uygulamışlardır. Burada belirtilen elde bulunan verinin öncelikle bir kısmının eğitim seti olarak ayrılıp kalan diğer kısım ile

tahmin işleminin gerçekleştirilmesi işlemidir. Modelin performansını bu aşamada ölçümlemek yerine daha sonra gerçek çıktıların üzerinde gözlemlenmesi ile modelin ne kadar başarılı olduğu tespit edilecektir.



Şekil 3.4 İptal Analizi İçin Örnek Veri Setleri (Kaynak: [28])

Yapılan çalışmalarda, Şekil 3.4'te gösterilen örneğin aksine belirli bir orana bölmek yerine literatürde k-katlı çapraz doğrulama (k-fold cross validation) olarak geçen algoritma kullanılacaktır. Bu algoritmada örneklem kümesi k kadar rassal parçalara bölünerek bir tanesi doğrulama seti diğer kalanlar ise eğitim seti olarak değerlendirilir. Sonrasında her doğrulama seti için k kadar sonucun ortalaması alınarak modelin hassasiyet gibi birçok performans göstergeleri ölçümlenmiş olacaktır.

Sonuçların değerlendirilmesi aşamasında sınıflandırma işleminin çıktısı olarak hata matrisi (Confusion Matrix) alınmaktadır. Bu matris üzerinden performans ölçütü olarak aşağıda belirtilen precision, recall ve F-ölçüsü hesaplamaları gerçekleştirilmiştir. Bu değerler literatürde kabul görmüş performans ölçümlerinden bazılarıdır.

Hata matrisi üzerinden yapılacak bu hesaplamaların k-katlı çapraz doğrulama işlemi ile her katta hesaplanan değerlerin ortalamaları alınarak sonuca ulaşılması gerekmektedir.

Çizelge 3.5 Hata Matrisi (Confusion Matrix)

		Gerçekleşen Durum	
	Toplam Nüfus	**Pozitif Durum**	**Negatif Durum**
Test Çıktısı	**Pozitif Test Çıktısı**	Doğru Pozitif (DP)	Yanlış Pozitif (YP - Tip I Hata)
	Negatif Test Çıktısı	Yanlış Negatif (YN - Tip II Hata)	Doğru Negatif (DN)

$$Doğru\ Pozitif\ Oranı = \frac{DP}{DP + YN} \tag{3.1}$$

$$Doğru\ Negatif\ Oranı = \frac{DN}{DN + YP} \tag{3.2}$$

$$Duyarlık = \frac{DP}{DP + YP} \tag{3.3}$$

$$Geri\ Çağırım = \frac{DP}{DP + YN} \tag{3.4}$$

$$Hatasızlık = \frac{DP + DN}{DP + DN + YP + YN} \tag{3.5}$$

$$F - \text{Ölçüsü} = \frac{2}{\frac{1}{duyarlık} + \frac{1}{geri\ çağırım}} \tag{3.6}$$

F-ölçüsü, geri çağırım ve duyarlık gibi ölçümler diğer (ROC eğrisinin altında kalan alan gibi) ölçümlerin hesaplanması için kullanılmaktadır. [29] Burada yapılan çalışmanın doğası gereği doğru pozitif oranı önem arz etmektedir. Çalışma kapsamında bu oran doğru tahmin edilen iptal abonelerin iptal abonelerin toplamına olan oranını vermektedir.

Ancak tabii ki doğrulama setleri üzerinden yapılacak herhangi bir ölçüm bize gerçek performansı vermeyecektir. Bunun için belirli bir süre sistem takip edilmeli ve gerçek sonuçların yapılan tahminlerle karşılaştırılması yapılmalıdır.

4. ANALİZ VE SONUÇLAR

Bu bölümde uygulanan modelin detaylı akışı ve her adımda karşılaştırmalı olarak alınan sonuçlar değerlendirilecektir. Uygulanacak model 3. Bölümde bilimsel araştırma sürecinde belirtilen altı adım bilgi keşif sürecine göre tasarlanmış olup belirtilen her adım ayrı ayrı modele katkı sağlamaktadır.

4.1 Deney I: Basit Sınıflandırma İşlemi

Bu aşamada verinin sadece müşterinin demografik ve tanımlayıcı bilgilerini içeren özellikleri kapsam dahilinde tutularak gerçek veri ve örneklem veri üzerinde bazı testler gerçekleştirip sonuçlar paylaşılacaktır.

Çizelge 4.2'de görüldüğü üzere Algoritma 1 içerisinde bazı değişkenler ve fonksiyonlar tanımlanmıştır. Belirtilen $musteriNitelik_{ij}$ değişkeni her bir müşteri için ilgili özelliğinin değerini belirtmektedir. İndis olarak belirtilen i müşterileri belirtmektedir ve 1'den müşteri sayısı kadar değer alır. j indisi ise özniteliği belirtmektedir ve 1'den öznitelik sayısı kadar değer alır. Burada belirtilen öznitelikler içerisinde müşterinin davranışlarını içeren zaman serisi verileri bulunmamaktadır.

Diğer yandan $katlaraAyir$ fonksiyonu ise sonuçların değerlendirme yönteminde bahsedilen k-katlı çapraz doğrulama işlemi için rassal katları oluşturarak veriyi setlere ayırma işlemi için bir başlangıç oluşturmaktadır. Bu işlemde müşterilere 1'den k'ya kadar rasgele etiketleme işlemi gerçekleştirilir. Literatürde k-katlı çapraz doğrulama uygulamaları genel olarak 10 kat oluşturularak yapıldığından k 10 kabul edilebilir. Her kat rassal olarak dağıldığından eşit ölçümler çıkmamakla birlikte Çizelge 4.1'de görüldüğü üzere %4-5 arasında bir standart sapma görülmektedir.

Çalışma kapsamında yapılacak sınıflandırma işlemi $eğitimSeti$ üzerinden eğitilip, eğitim bilgileri kullanılarak $testSeti$ üzerinden de öngörü işlemi tamamlanmaktadır. Eğitilme ve öngörü tahmini işlemleri literatürde bahsedilen destek vektör makineleri ve özyinelemeli bölümleme sınıflandırma algoritmaları ile gerçekleştirilmektedir. Her kat için işlem tekrarlanmakta ve $preformans_{ij}$ değişkeni içerisinde her $i.$ performans ölçüsü (bunlar 3.2.5. bölümde Sonuçların Değerlendirilme Yöntemi konusunda yer alan hatasızlık oranı, doğru pozitif oranı, doğru negatif oranı, duyarlık, geri çağırma ve F-ölçüsüdür) için $j.$ kat hesaplamaları dikkate alınarak kriterleri belirlenmektedir.

Çizelge 4.1 Her Katta Hesaplanan F-Ölçüsü

Kat	DVM	ÖYB
1	0,400	0,550
2	0,436	0,475
3	0,367	0,432
4	0,283	0,430
5	0,340	0,543
6	0,434	0,417
7	0,318	0,478
8	0,358	0,460
9	0,438	0,486
10	0,354	0,467
St. Sapma	**0,053**	**0,045**
Ortalama	**0,373**	**0,474**

Son olarak $ortalamaPerformans_i$ içerisinde her kat için belirlenen performans parametrelerinin ortalamaları alınarak mevcut veri üzerinde bir öngörü edinilmiş olmaktadır. Çizelge 4.1'de F-ölçüsü için yer alan ortalama alanı bu ölçüm için örnek teşkil etmektedir.

Çizelge 4.2 Basit Sınıflandırma Prosedürü

Algoritma 1. Basit Sınıflandırma

$nitelik_{ij} \leftarrow Her\ i.\text{müşteri için}\ j.\text{özelliği eşitle}$

$katlar \leftarrow katlaraAyır(nitelik_{ij}, k)$

$\text{ölçümKümesi} \leftarrow \{DPO, DNO, Duyarlık, GeriÇağırım, Hatasızlık, F - \text{Ölçüsü}\}$

$\boldsymbol{for}\ i \leftarrow 1 \ldots k\ \boldsymbol{do}$

$eğitimSeti \leftarrow \text{öz}NitelikKümesiniBölümle(nitelik_{ij}, katlar[i])$

$testSeti \leftarrow \text{öz}NitelikKümesiniBölümle(nitelik_{ij}, katlar[-i])$

$sınıflandırıcı.model \leftarrow ogren(metod, eğitimSeti)$

$sınıflandırıcı.tahmin \leftarrow tahminEt(sınıflandırıcı.model, testSeti)$

$performans_{ij} \leftarrow performansHesapla(sınıflandırıcı.tahmin)\ \forall i \in \text{ölçümKümesi}\ \wedge\ \forall j \in 1 \ldots k$

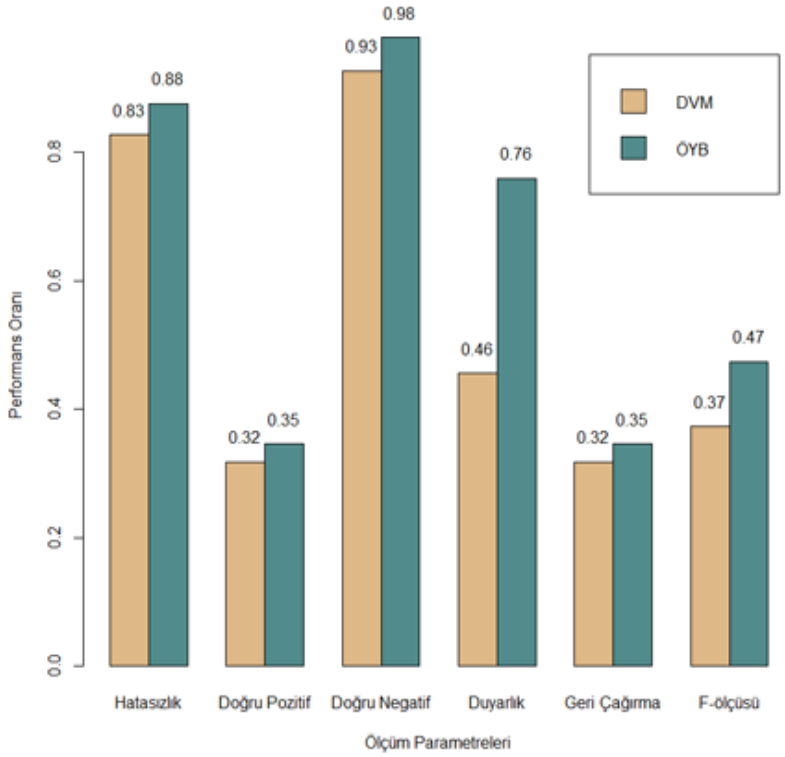
$\boldsymbol{end\ for}$

$ortalamaPerformans_i \leftarrow {\sum_{j=1}^{k} performans_{ij}}/{k}\ \ \forall i \in \text{ölçümKümesi}$

Örneklem veri ile Algoritma 1 çalıştırılmış ve Şekil 4.1'de gösterilen sonuçlar elde edilmiştir. Özyinelemeli bölümleme algoritması her ölçüm için destek vektör makineleri algoritmasından daha iyi sınıflandırma işlemi yapıyor diyebiliriz. Burada önemli nokta doğru pozitif oranının %32 ile %35 arasında olmasıdır. Bu oran daha önce bahsedildiği üzere doğru tahmin edilen iptal edecek abonelerin, iptal eden abonelerin toplamına olan oranını vermektedir. Doğru pozitif oranı her ne kadar düşük görünse de "iptal edecek" olarak sınıflandırılan bir abonenin 0,35 ihtimalle iptal etme olasılığı dahi firma için oldukça önemli bir bilgidir.

Basit sınıflandırma deneyiminde ölçümü yapılan tüm değerler için diğer deneylerde iyileştirilebilmesi için çalışmalar gerçekleştirilecektir. Doğru pozitif oranının yanı sıra F-ölçüsü de gerçekleştirilen uygulama hakkında ortalama bir algı

oluşturmaktadır. Her ne kadar konu gereği iptal eden abonelerle ilgilensek de iptal etmeyen veya iptal etmeyecek abonelerin de doğru tahmin edilmesi algoritmanın doğru bilgi sağlaması açısından önemlidir.



Şekil 4.1 Deney I Sonuçları

Şekil 4.1'de eleştirilmesi gereken bir diğer konu ise hatasızlık oranı, doğru negatif oranı gibi bazı değerlerin yüksek çıkmasıdır. Bunun nedeni örneklem veri içerisinde gerçeği yansıtması açısından iptal eden abonelerin, toplam nüfusa olan oranının %16 olması ve bu abonelerin tahmin oranının nüfusa yakın olmasına neden olmaktadır.

4.2 Deney II: İki Fazlı Çözüm Modeli

İki fazlı çözüm modelinde basit sınıflandırma modeline ek olarak geliştirmeler gerçekleştirilerek bir çözüm metodolojisi uygulanacaktır. Belirtilen iki fazın ikincisi sınıflandırma işlemidir. Sınıflandırmadan önce yapılacak ek işlemler birinci faz olarak adlandırılmaktadır.

Birinci fazda yapılan işlemler ilk algoritmadan farklılık gösterdiğinden değişkenlerin açıklanmasına $davranış_{ijk}$ ile başlanabilir. Bu değişken içerisinde müşterilerin zaman serisi halinde davranışları yer almaktadır. Örnek verilecek olursa müşterinin fatura tutarları üzerinden belirlenen alım yönü endeksinin j. davranış kümesi için i. müşterinin k. gözlemi bu değişken içerisinde değer bulmaktadır. Davranış kümeleri verinin belirlenmesi aşamasında belirtilen zaman serisi türünde olan verilerdir.

Çizelge 4.3 Alım Yönü Endeksi İçin Örnek Veri

Müşteri	Gözlem 1	Gözlem 2	Gözlem 3	Gözlem 4	Gözlem 5	Gözlem 6
1	41,500	41,500	41,500	41,500	41,500	41,500
2	64,700	44,600	54,400	50,400	44,500	44,800
3	142,200	41,000	40,900	41,000	36,500	200,000
4	9,900	191,600	61,400	57,800	51,500	51,500
5	11,400	214,200	47,400	55,400	46,500	31,500
6	71,400	61,400	61,300	52,300	61,500	61,900

Örnek verilmesi gerekirse Çizelge 4.3'te görüldüğü üzere bazı aboneler standart yönde alım eğilimi gösterirken 3., 4. ve 5. müşteride görüldüğü üzere bazı gözlemlerde yüklü meblağda alım yaptıkları görülmektedir. Bunlar gerçek bir iptalin sinyalleri olabileceği gibi aslında standart bir kampanya alımından veya cihaz alımından kaynaklanıyor da olabilir. Konu gereği bu analizleri yapmak üzere zaman serisi sınıflandırma algoritmaları kullanılmalıdır.

Bu kısımda ek olarak $davranış_{ijk}$, $stDavranış_{ijk}$ve $tpDavranış_{ijk}$ eklenerek birinci faz oluşturulmaya başlanmıştır. Çizelge 4.3'de örneklendiği gibi her müşterinin belirli bir davranış kümesi içerisinde yapılan gözlemleri $davranış_{ijk}$ değişkeninde tutulmaktadır. Bu değişkenin literatürde bahsedilen ortalama ve ölçekleme işlemi yapıldıktan sonraki hali ise $stDavranış_{ijk}$ ve bu iki değişkenin davranış kümeleri üzerinden birleşmiş hali de $tpDavranış_{ijk}$ olarak geçmektedir.

Çizelge 4.4 İki Fazlı Çözüm Modeli

Algoritma 2. Çözüm Metodolojisi

$nitelik_{ij} \leftarrow Her\ i.$müşteri için $j.$özelliği eşitle

$davranış_{ijk} \leftarrow Her\ i.$müşteri için $j.\ davranış\ k$ümesinin $k.$gözlemini eşitle

$stDavranış_{ijk} \leftarrow satırdaOrtala$Ölçekle$(davranış_{ijk})$

$tpDavranış_{ijk} \leftarrow kBağla(stDavranış_{ijk}, davranış_{ijk})$

$katlar \leftarrow katlaraAyır(nitelik_{ij}, k)$

ölçümKümesi

$\leftarrow \{DPO, DNO, Duyarlık, Geri$ Çağırım$, Hatasızlık, F -$ Ölçüsü$\}$

$davranışK$ümesi

$\leftarrow \{Alım$ Yönü $Endeksi, Download\ Miktarı, Arıza\ Sayısı \dots\}$

for $her\ j \in davranışK$ümesi ***do***

$dk_{ij} \leftarrow kumele(metod,\ tpDavranış_{ijk})\ \forall i \in 1 \dots i\ \wedge j \leftarrow j$

$dk_{ij} \leftarrow ikiliMatriseDonustur(dk_{ij})$

$nitelik_{ij} \leftarrow kBağla(nitelik_{ij}, dk_{ij})$

end for

for $i \leftarrow 1 \dots k$ ***do***

$eğitimSeti \leftarrow$ öz$NitelikK$ümesini$B$ölümle$(nitelik_{ij}, katlar[i])\ \forall j \leftarrow i$

$testSeti \leftarrow$ öz$NitelikK$ümesini$B$ölümle$(nitelik_{ij}, katlar[-i])\ \forall j \leftarrow i$

$sınıflandırıcı.model \leftarrow ogren(metod, eğitimSeti)$

$sınıflandırıcı.tahmin \leftarrow tahminEt(sınıflandırıcı.model, testSeti)$

$performans_{ij} \leftarrow performansHesapla(sınıflandırıcı.tahmin)\ \forall i$

$\in$ ölçümKümesi $\wedge\ \forall j \in 1 \dots k$

end for

$ortalamaPerformans_i \leftarrow {\sum_{j=1}^{k} performans_{ij}}/{k}\ \ \forall i \in$ ölçümKümesi

İki fazlı çözüm modelinin birinci fazı zaman serisi türündeki davranış kümeleri veri setlerinin ayrı ayrı kümelenerek anlamlı bir şekilde öznitelik olarak değerlendirilmesi ve ortaya çıkan yeni özniteliklerin $nitelik_{ij}$ bulunan özniteliklerle aynı satırda birleştirilmesidir.

İlk aşama her davranış kümesi için çalışmaktadır. Bu aşamanın değer katan kısmı $kumele$ fonksiyonudur. Her davranış kümesi için her müşterinin sistem içerisinde gözlemlenen davranışlarını gruplayarak basit anlamda müşteriler birbiriyle kesişmeyecek şekilde ayrık kümelere atanmaktadır. Benzer şekilde davranan müşterilerin süreç içerisinde tahmin edilmeye çalışılan davranışları da benzer olmalıdır varsayımıyla bu işlemler gerçekleştirilmektedir.

Kümeleme işlemi sonucu olarak tam sayı değer olacak şekilde dk_{ij} değişkeni içerisinde her bir müşterinin küme numarası bulunmaktadır. Yapılan deneylerde görüldüğü ve verinin anlamlandırılması sürecinde bahsedildiği üzere birbirine üstünlüğü bulunmayan kategorik verilerin sayısal olarak ifadesinde birbirine üstünlüğü olmaması durumunda 1,2,3,… gibi tamsayıların verilmesi yapılacak işlemlerde sorun olmaktadır. Böyle şekilde ifade etmek yerine $ikiliMatriseDonustur$ işlevi yardımı ile kümeler, tekil küme sayısı kadar öznitelik olacak şekilde ikili matrise dönüştürülmüştür.

Bu sayede her müşteri için ayrı ayrı belirlenen davranışsal kümeleme sonuçları eğitim ve test verisi olarak ayrılabilecektir. Başka bir deyişle $kBağla$ metodu ile dk_{ij} değişkeninde yer alan kümeleme sonuçları $nitelik_{ij}$ değişkenlerine yeni öznitelik olarak eklenecektir.

İki fazında çalışması sağlanarak sonuçlar değerlendirilmiştir. İlk aşamada yukarıda bahsedilen işlemler kü$mele$ işlevinin içerisinde $metod$ olarak k-ortalama kümeleme algoritması ve hiyerarşik kümeleme algoritması kullanılmıştır.

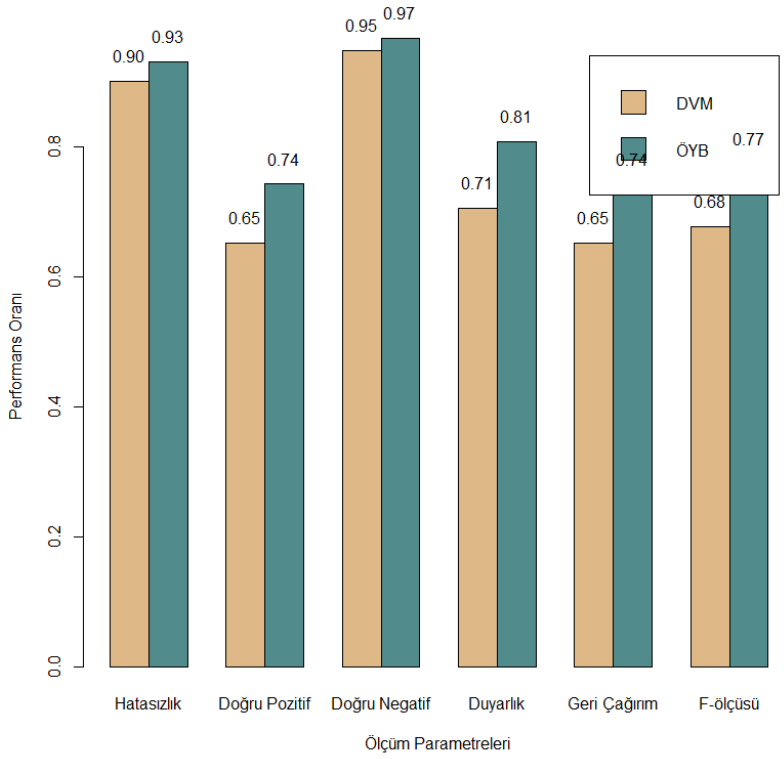
4.2.1 K-ortalama Kümeleme Uygulaması

İlk olarak k-ortalama kümeleme algoritması ile iki fazlı çözüm modelinin değerlendirilmesi gerçekleştirilmiştir. K-ortalama kümeleme algoritmasının çeşitli parametreleri ile birlikte testler yapılmış ve en iyi sonuç verecek şekilde uygulamada kullanılmıştır.

Çizelge 4.5 K-Ortalama Küme Sayısı Test Sonuçları

	Küme Sayısı								
Performans Ölçüsü	**2**	**3**	**5**	**8**	**10**	**20**	**30**	**40**	**50**
Hatasızlık Oranı	0,89	0,90	0,92	0,92	0,92	0,92	0,93	0,92	0,92
Doğru Pozitif Oranı	0,64	0,64	0,73	0,69	0,72	0,73	0,74	0,73	0,72
Doğru Negatif Oranı	0,94	0,95	0,96	0,97	0,96	0,96	0,96	0,96	0,96
Duyarlık	0,67	0,70	0,78	0,81	0,80	0,80	0,80	0,79	0,79
Geri Çağırım	0,64	0,64	0,73	0,69	0,72	0,73	0,74	0,73	0,72
F-Ölçüsü	0,65	0,67	0,75	0,75	0,75	0,76	0,77	0,76	0,75

Uygulama içerisinde örneklem veri üzerinde yapılan testler Çizelge 4.5'te belirtilmiştir. Burada verilen sonuçlar sadece özyinelemeli bölümleme sınıflandırmasına aittir ancak destek vektör makinelerinde de benzer sonuçlar alınmaktadır. Yapılan parametre testlerinde görüldüğü üzere performansı hızlı bir şekilde artırıp indirebilen etmen olarak küme sayısı belirtilebilir.



Şekil 4.2 K-ortalama Algoritması İle Deney II Sonuçları

K-ortalama uygulaması ile parametre testlerinden elde edilen en iyi sonuçlara göre iki fazlı çözüm modeli üzerinde yapılan deneylerin ikinci faz sonuçları Şekil 4.2'de belirtilmiştir. Sonuçlardan da anlaşılacağı üzere I. deney ile karşılaştırıldığında diğer performans ölçütlerinde olduğu gibi çalışmanın konusu gereği özellikle incelenmesi gereken F-ölçüsü ve doğru pozitif oranı önemli oranda değer kazanmıştır.

4.2.2 Hiyerarşik Kümeleme Uygulaması

Bu kısımda iki fazlı çözüm modeli üzerinde kümeleme işlemi için hiyerarşik kümeleme uygulaması test edilecektir. Öncelikle hiyerarşik kümeleme algoritmasının müşteri verisi gibi çok satırlı veri analizlerinde uygulamasının ek varsayımlara ihtiyaç duyduğunu söylemek gerekiyor. Çizelge 4.6'da hiyerarşik kümeleme algoritmasının kullanım şekli gösterilmiştir. Burada yapılan işlem özet olarak kümeleme işlemini tüm veri üzerinde değil daha küçük bir örneklem üzerinde

yapılmasıdır. Hiyerarşik kümeleme algoritmasının karmaşıklığı $O(n^3)$ olduğundan müşteri sayısının artmasıyla işlem süresi buna bağlı olarak artmaktadır.

Çizelge 4.6 Hiyerarşik Zaman Serisi Kümeleme Algoritması

Algoritma 3. Zaman Serisi Hiyerarşik Kümeleme

$orneklem_{ij} \leftarrow$ öz$NitelikK$ümesini$B$ölümle$(veriMatrisi_{ij}, n)$

$uzaklikMatrisi_{ii} \leftarrow uzaklık(orneklem_{ij}, uzaklıkÖlçümMetodu)$

$hAgac \leftarrow hiyerarşikKümele(uzaklikMatrisi_{ii}, metod)$

$dk_{ij} \leftarrow agaciKümelereAyir(hAgac, k)$

$merkezMatrisi_{kj} \leftarrow kumeMerkezleriniHesapla(dk_{ij}, k)$

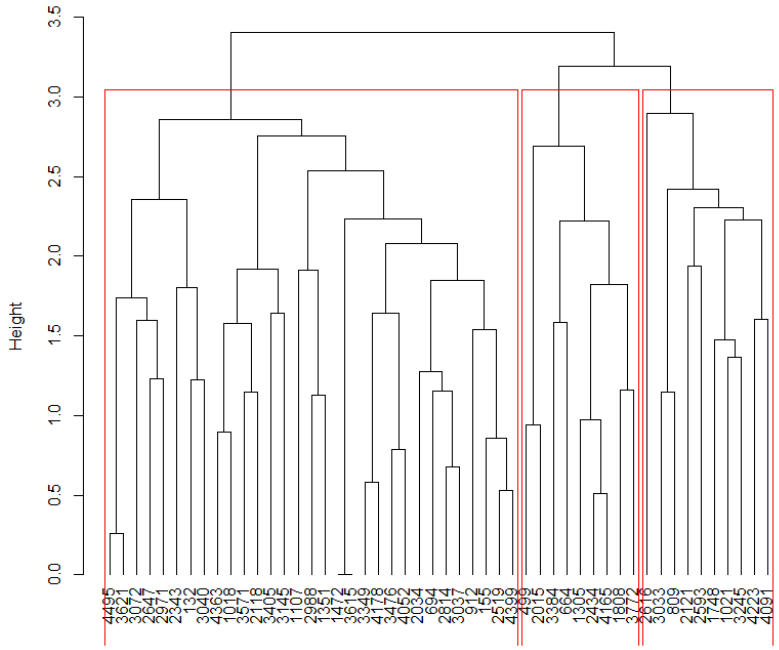
for $i \leftarrow 1 \ldots i$ **do**

$kumeSonucu_i \leftarrow enYakınKume(merkezMatrisi_{kj}, orneklem_{ij})$

end for

Karmaşıklığın $O(n^3)$ olmasının nedeni $N-1$ iterasyonda $N \times N$ uzaklık matrisine bakılmasıdır. Çizelge 4.6'da belirtildiği üzere $uzaklikMatrisi_{ii}$ her bir müşteri davranışının birbirine olan uzaklığını bulmaktadır. Burada örneklem alınmadan işlem yapılması durumunda hesapları gerçekleştirilen 6000 civarında müşteri verisi için 6000^2 büyüklüğünde bir uzaklık matrisi ile işlem yapılıyor olacaktı. Uzaklık ölçümü işlemi literatürde bahsedilen öklidyen uzaklığı ile yapılmaktadır.

Sonuç olarak alınan örneklem ile hiyerarşik kümeleme işlemi gerçekleştirilir ve Şekil 4.3'te dendrogram olarak gösterildiği üzere $hAgac$ değişkenine hiyerarşik kümeleme işlemi sonrasında bir ağaç objesi atanmaktadır. Bu işlem sonrasında dk_{ij} değişkeninde kümelere ayırma işlemi gerçekleştirilir. Bu işlem de görsel olarak Şekil 4.3'te kırmızı çizgiler üzerinden görülebilir. Görselde belirtilen kutucukların her biri bir kümeye karşılık gelmektedir.

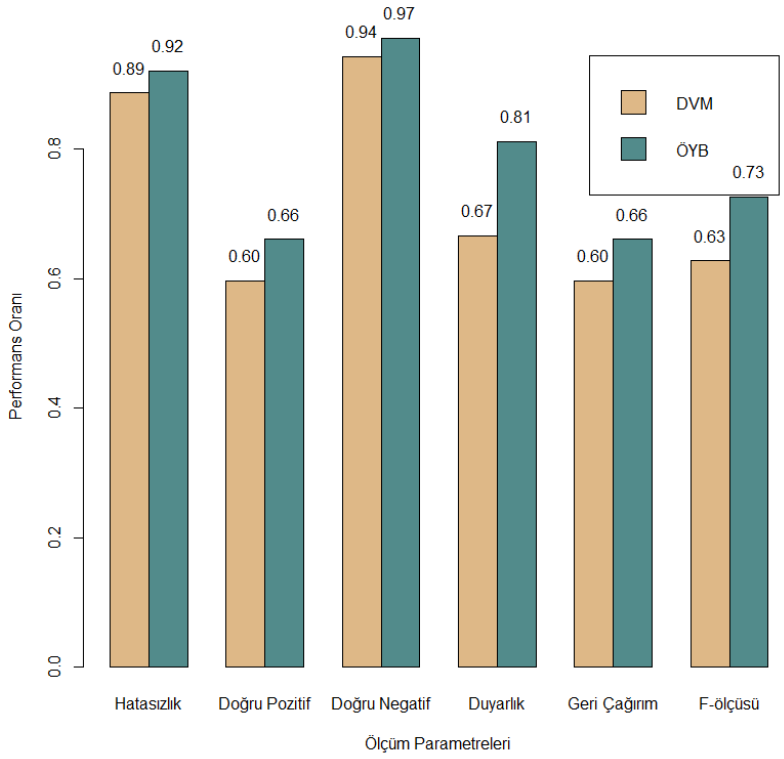


Şekil 4.3 Örnek Hiyerarşik Kümeleme Dendrogramı

Örneklem üzerinde yapılan kümeleme işlemi sonrasında veri k kadar kümeye ayrılmış olacaktır. Bu işlemde de k-ortalama kümeleme işleminde uygulanan yöntem uygulanarak çeşitli parametrelerin denenmesi ile k parametresi belirlenmiştir.

Oluşan kümelerin örneklem içerisinde dağılımlarına bakılarak merkez nokta tespiti $merkezMatrisi_{kj}$ parametresi üzerinde $kumeMerkezleriniHesapla$ metodu ile bulunur. Merkez tespiti sırasında her kümede bulunan elemanların ağırlıklı ortalamalarına bakılarak işlem yapılır. Sonuç olarak her müşterinin bir kümesi olması gerektiğinden örneklemin tamamı üzerinde hangi kümeye daha yakın olduğunun tespitinin yapılması için $enYakınKume$ metodu kullanılır. Uzaklık her müşteri için belirlenen her bir küme merkezine olan $uzaklık$ metodu ile ölçülen birimdir.

Böylece birinci faz tamamlanmış olur ve eklemeleri ile birlikte Deney I'de bahsedilen sınıflandırma işlemine geçilmiş olacaktır. Kümeleme işlemi sonucunda $kumeSonucu_i$ değişkeninde bulunan kümeler ikili matris şeklinde öznitelik olarak incelenecek veride yerini alır.



Şekil 4.4 Hiyerarşik Kümeleme Sonuçları

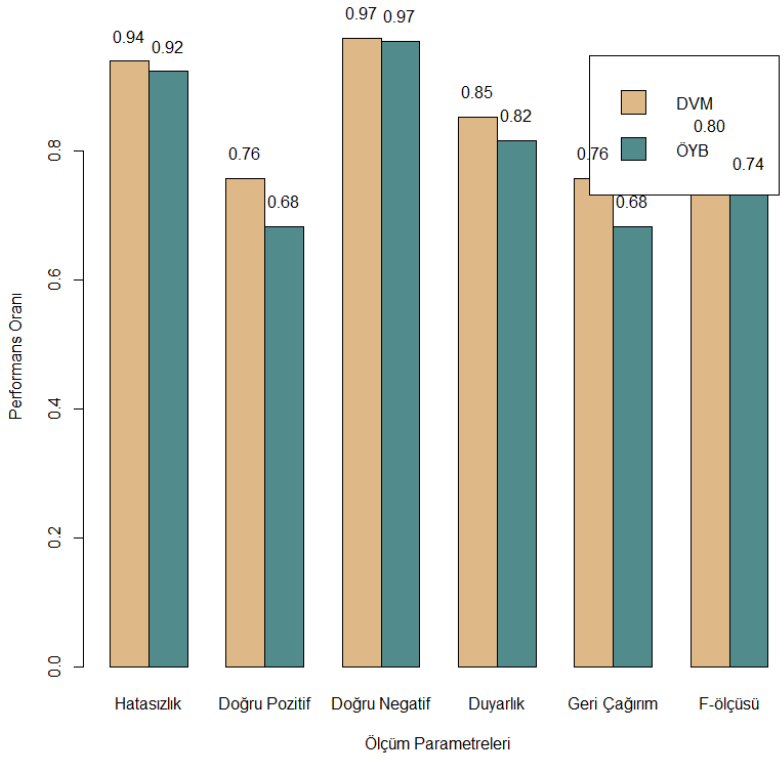
Sonuç olarak hiyerarşik kümeleme işlemi iki fazlı çözüm modelinde uygulandığında Şekil 4.4'te belirtilen şekilde bir performans göstermektedir. İki fazlı çözüm modeli içerisinde gerçekleştirilen iki ayrı kümeleme algoritması denemesi performanslar karşılaştırıldığında k-ortalama kümeleme algoritmasının özyinelemeli bölümleme sınıflama algoritması ile çalıştığında F-ölçüsü için %2 daha iyi sonuç verdiği görülmektedir.

4.3 Deney III: Hiyerarşik İF Çözüm Modeli'nin Uygulamasına Yapılan Ekler

Bu bölümde iki fazlı çözüm modeline uygulanan hiyerarşik kümeleme algoritması üzerinde bazı değişiklikler yaparak daha iyi sonuç veren bir model bulunmaya çalışılacaktır. Bunlardan birincisi belirlenen küme merkezlerine olan uzaklıkların bulunması ile ilgilidir. Diğeri ise korelasyonu yüksek özniteliklerin çıkarılarak denenmesidir.

4.3.1 Küme Merkezlerine Olan Uzaklıklar

Hiyerarşik kümeleme algoritmasının sonuçları ayrık kümelere müşterilerin atanması ile gerçekleştiriliyordu. Bu deneyde ayrık kümeler yerine küme merkezleri belirlenerek her müşterinin bu kümelere olan uzaklıkları değerlendirilecektir. Sonuç olarak ikili matris yerine küme merkezlerine olan uzaklıklar değerlendirmeye alınarak aslında müşterinin tekil bir kümede olmadığı ve yakınlık derecesine göre hangi kümede ne kadar olduğu değerlendirme kriteri olacaktır.



Şekil 4.5 Küme Merkezlerine Olan Uzaklıklar

Gerçekleştirilen yöntem Çizelge 4.6'da belirtilen hiyerarşik kümeleme prosedürü üzerinden değerlendirilecek olursa her müşteri için bulunan $enYakınKume$ fonksiyonu yerine örneklem içerisinde yer alan müşteri davranışının $merkezMatrisi_{kj}$ içerisinde yer alan küme merkezlerine olan uzaklıklarını bulan $kümeUzaklıklarınıBul$ metodu ile gerçekleştirilmektedir. Ayrıca ikili matris bu

deneyde istenmeyen bir işlev olduğundan Çizelge 4.4'te iki fazlı çözüm modelinde uygulanan $ikiliMatriseDonustur$ metodu da kullanılmamıştır.

Özniteliklerin ikili matris şeklinde değerlendirilmesi bazı algoritmaların doğası gereği işlemleri zor kılarken bazı algoritmalar içinse daha kolay kılmaktadır. Karar ağacı modellerinde özniteliklerin kategorik olması işlemleri kolaylaştırırken destek vektör makineleri gibi uygulamalarda ayrım yapmak zorlaşmaktadır. Bu hipoteze en güzel örnek olarak bu deneyin çıktıları verilebilir. Çalışma kapsamında gerçekleştirilen daha önceki deneylerin hepsinde görüldüğü üzere destek vektör makineleri, özyinelemeli bölümleme algoritmasına göre düşük sonuç vermektedir. Yapılan bu deneyde ikili matrislerin kaldırılıp bir müşterinin bir davranış kümesi içerisinde keskin çizgiler yerine küme uzaklıkları değerlendirildiğinde destek vektör makineleri Şekil 4.5'te görülebileceği gibi en iyi sonucu vermektedir. Değerlendirme kriteri olarak alınan F-ölçüsü %3 fark ile en iyi sonucu vermektedir.

4.3.2 Öznitelik Eliminasyonu

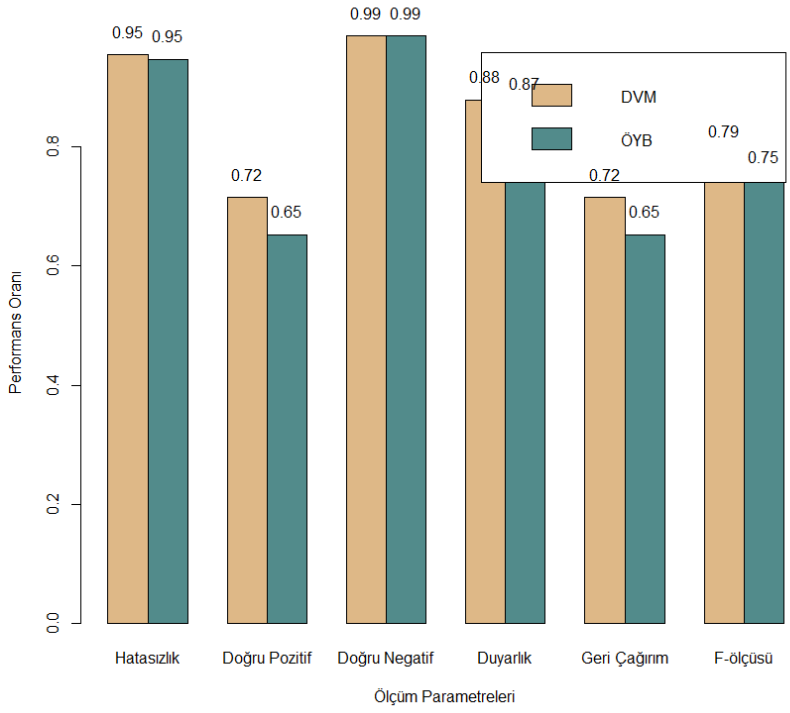
Örneklem veri ile gerçekleştirilecek son deney ise en iyi sonuç alınan küme merkezlerine olan uzaklıkların iki fazlı çözüm modeli ile uygulanması işleminde veri içerisinde sonuca negatif yönde etkisi olabilecek özniteliklerin olduğu varsayımıyla belirlenecek özniteliklerin çıkarılıp sonuca katkı sağlayan öznitelikler ile işlemin gerçekleştirilmesini sağlamaktır. Ancak bu deneyde negatif yönde öğrenmeye etkisi olabilecek veriler olduğu varsayımı önem arz etmektedir. Yanlış özniteliklerin seçilip veri kümesinden çıkarılması bu deney için yapılabilecek bir hata olup bunun değerlendirilmesi yapılmalıdır.

Bu kapsamda küme merkezlerine olan uzaklıkların hesaplanmasında yapılanlara ek olarak birinci fazda korelasyonu yüksek olan özniteliklerin çıkarılmasını sağlayan $öznitelikEliminasyonu(dk_{ij}, p)$ işlevi ile her davranış kümesi için korelasyonu p'nin üzerinde olan öznitelikleri çıkarılacaktır. Burada yüksek korelasyon olarak belirtilen p değeri için %70, %80 ve %90 değerleri denenmiştir. Çıkan sonuçlar

değerlendirildiğinde bu deneyin daha verimli bir deney olduğu söylenememektedir. Bahsedilen negatif yönde öğrenmeye etkisi bulunan özniteliklerin bulunması varsayımı, yapılan bu deneyde doğrulanamamıştır. Bunun sebeplerinden birisi hiyerarşik kümeleme algoritmasında seçilen örneklem ve küme sayısının doğru seçilmesi olabilir. K-ortalama kümeleme uygulaması konusunda belirtilen kümeleme deneylerinde alınan çıktılara göre fazla küme seçilmesi durumunda da uygulamanın başarımının düştüğü görülmüştür.

4.4 Deney IV: İF Çözüm Modelinin Gerçek Veri İle Denenmesi

Bu aşama deney gözlemlerinin son aşaması olup gerçek verinin küme merkezlerine olan uzaklıklar geliştirmesi ile birlikte iki fazlı çözüm modelinde denenmesi ile elde edilen sonuçlar Şekil 4.6'te gösterilmektedir.



Şekil 4.6 İF Çözüm Modelinin Gerçek Veri Sonuçları

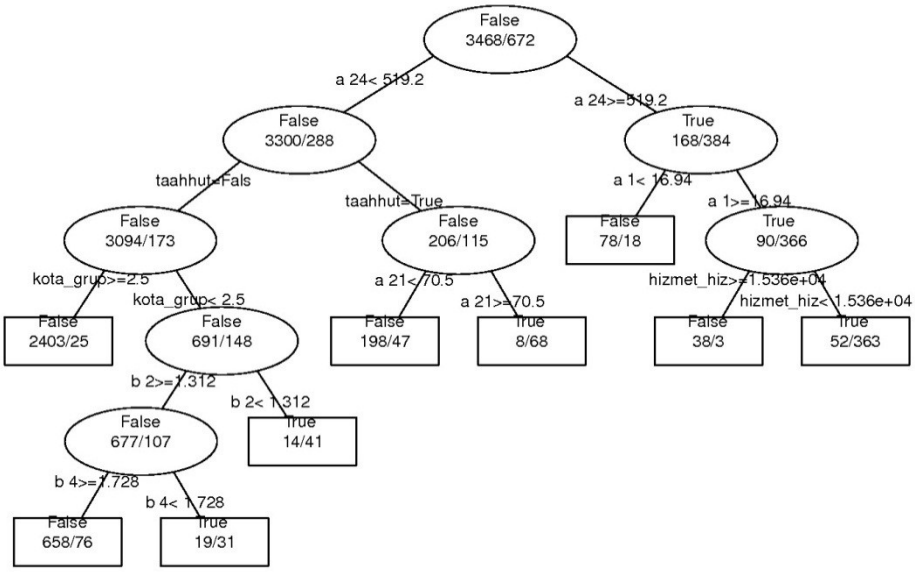
Burada gerçekleştirilen işlem bir pilot bölge alınarak toplamda 70000 civarında abonenin analizidir. Bu açıdan performans problemlerinin değerlendirilmesi

gerekmektedir. Birinci faz değerlendirilecek olursa hiyerarşik sınıflandırmaya küme merkezlerinin bulunması için örneklem veri üzerinden değerlendirme yapıldığından müşteri sayısının artması ile sadece müşterinin sınıf ataması işlemi uzun sürmektedir. İkinci faz değerlendirilecek olursa DVM ve ÖYB algoritmaları karşılaştırıldığında ÖYB algoritması DVM'ye göre oldukça hızlı çalışmaktadır. 70000 abone için ÖYB algoritması dakikalar içerisinde işlemleri bitirirken, DVM algoritması yarım günü aşan bir sürede işlemi tamamlayabilmiştir. Her ne kadar %79 gibi bir F-ölçüsü çıktı olarak alınmışsa da abone sayısı dikkate alındığında milyonlara varan bir uygulamada edinilecek bilgilerden yararlanacak olanlar için süre makul olmayabilir.

Bir diğer konu ise sistemin canlı bir şekilde değerlendirilmesi hususudur. Çizelge 3.5'te verilen hata matrisinde pozitif test çıktısı olup negatif durumda olan objeler için gözlem yapılması gerekmektedir. Performans göstergeleri Yanlış Pozitif - Tip 1 hata üzerinde zaman içerisinde doğru pozitife kaymalar yaşanabileceğinden aslında tahmin etme alanı mevcut veri ile çalışıldığından Tip 1 hata grubuna giren objeler olacaktır.

4.5 Özniteliklerin Değerlendirilmesi

Gerçekleştirilen sınıflandırma işleminde hangi parametrenin ne kadar katkıda bulunduğunu göstermek üzere özyinelemeli bölümleme algoritmasında gerçekleştirilen on katın her biri için kullanılan karar ağaçları incelenmiştir. Kullanılan her ağacın kökünden yaprağına kadar belirlenen her kural için F-ölçüsü hesaplanmıştır. Şekil 4.7'de örnek olarak gösterilen karar ağacına göre parametreler arasındaki ilişkiler çıkarılmıştır. Burada gösterilen her bir kare kutucuk (yaprak) iptal eden veya etmeyen olarak iki ayrı karar grubuna ayıracak şekilde değerlendirilmiştir. Her bir yaprak tekil olarak ele alınmış iptal eden yaprakta bulunan değerler dışında kalan diğer kararlar iptal etmeyen olarak doğru pozitif, doğru negatif, yanlış pozitif ve yanlış negatif değerleri hesaplanmıştır. Ayrıca bu bölümde yer alan analizlerin tamamı gizlilik gerekçesi ile örnek teşkil etmesi açısından örneklem veri üzerinde gerçekleştirilmiştir.



Şekil 4.7 Özyinelemeli Bölümleme Örnek Karar Ağacı

Yapılan hesaplamalara göre her bir katta belirlenen tekil kurallar üzerinden hesaplanan F-ölçüsü değerleri hem iptal eden grubu hem de iptal etmeyen grubu için en iyi sonucu verecek 5 kural sözcüğü aynı öznitelikler tamamıyla tekrarlanmayacak şekilde Çizelge 4.7 ve Çizelge 4.8'de gösterilmiştir. Kural sözcüğünü okurken a, b, c gibi harfler zaman serisi kümelerini belirtmektedir. Alım yönü endeksinin kümeleme sonrası oluşan öznitelikleri "a" harfinin yanına küme numarası gelecek şekilde oluşturulmuştur. Aynı şekilde alım yönü endeksinin ortalanmış ve ölçeklenmiş hallerinin zaman serisi kümeleme işleminden sonra oluşan öznitelikleri "b" harfi ve küme numarasının birleşimi ile gösterilmektedir. Koşul aralarında mantıksal ve belirteci "-VE-" ile gösterilmektedir.

Çizelge 4.7'de görülebileceği gibi "a 7" özniteliği tüm denemelerde karar ağacının en başında yer almaktadır. Bu durum belirtilen özniteliğin iptal kararında iyi bir ayrıştırma gerçekleştirdiğinin göstergesidir. Bu öznitelikte belirtilen koşulda işlenmemiş ham alım yönü endeksinin betimleyici istatistiklerine bakılacak olursa;

- Yedinci kümede ortalamalar göz önünde bulundurulduğunda takip eden zamanlarda arada farklılıkların çok olmadığı görülmüştür. Sabit alım hareketliliği gösteren müşterilerin iptal etmemesi yorumu elde edilebilir.

- Aynı zamanda standart sapmanın iptal etmeyecek olarak tahmin edilen müşterilerde daha düşük olması bu aboneler için alım yönü endeksi davranışını önemli kılmaktadır. Değişken alım yönü endeksi davranışı istenmeyen bir durum olduğu gözlemlenmiştir.

Çizelge 4.7 İptal Etmeme Kararı Verilen En İyi Beş Kural

Kural Sözcüğü	**F-Ölçüsü**
a 7< 272.7 -VE- a 16< 304.5 -VE- taahhut=False -VE- kota_internet>=3072	0,55
a 7< 272.6 -VE- a 16< 299.7 -VE- ktv_yil>=130 -VE- taahhut=False -VE- kota_internet>=3072	0,55
a 7< 272.7 -VE- a 16< 306.7 -VE- taahhut=False -VE- ktv_yil>=130 -VE- kota_grup>=2.5	0,54
a 7< 272.7 -VE- taahhut=False -VE- kota_grup>=2.5	0,53
a 7< 272.7 -VE- a 16< 306.7 -VE- taahhut=False -VE- ktv_yil>=130 -VE- kota_grup< 2.5 -VE- b 2>=1.542	0,30

Diğer alım yönü endeksi özniteliklerinde de aynı şekilde sabit ve alım yönü endeksi ortalama seviyede bulunan müşteriler için davranış biçimi ve sonucu benzerlik göstermektedir.

Çizelge 4.8 Öznitelik Betimleyici İstatistikleri

	a 7 < 272.7					
	Gözlem 1	Gözlem 2	Gözlem 3	Gözlem 4	Gözlem 5	Gözlem 6
Ortalama	54.0	46.36	49.55	49.66	49.29	49.03
St. Sapma	22.41	14.71	12.94	13.18	13.33	13.16
	a 7 >= 272.7					
	Gözlem 1	Gözlem 2	Gözlem 3	Gözlem 4	Gözlem 5	Gözlem 6
Ortalama	22.18	45.64	43.50	42.96	43.08	44.30
St. Sapma	29.42	60.14	40.92	31.47	29.68	32.56

Çalışmanın konusu gereği iptal eden abonelerin belirlenmesi önem teşkil ettiğinden iptal edilenleri belirleyen kararı etkileyen öznitelikleri incelemek gerekmektedir. Çizelge 4.9'da verilen kurallar F-ölçüsü kriteri ile değerlendirildiğinde iptal etme eğiliminde olan abonelerin çoğunu kapsamaktadır. Her kat içerisinde oluşturulan ağacın ilk düğümü "a 7" ile başladığından yukarıda iptal etmeme eğiliminde kararı verilen abonelerde yapılan yorumların tamamının tersi iptal etme eğiliminde olan aboneler için de geçerlidir.

Çizelge 4.9 İptal Etme Kararı Verilen En İyi Beş Kural

Kural Sözcüğü	**F-Ölçüsü**
a 7>=272.7 -VE- a 2>=22.37 -VE- hizmet_hiz< 15360 -VE- b 21>=2.673	0,64
a 7>=272.7 -VE- a 2>=22.43 -VE- hizmet_hiz< 15360 -VE- b 28< 3.39	0,64
a 7>=272.7 -VE- b 6>=3.561 -VE- hizmet_hiz< 23040 -VE- a 9>=35.43	0,63
a 7>=272.7 -VE- kota_grup< 1.5	0,49
a 7>=272.7 -VE- kota_grup>=1.5 -VE- taahhut=False -VE- b 21>=2.976 -VE- a 3>=42.15 -VE- hizmet_hiz< 23040	0,21

Belirtilen kurallardan da anlaşılacağı üzere 15 Mbit ve altındaki hıza sahip olan abonelerin iptal etme eğilimi daha fazla görünmektedir. Tabii ki bu bilgiyi tek başına değerlendirmek yerine kurallarda belirtildiği şekilde alım yönü endeksi ile birlikte değerlendirmek sonucu etkileyecektir. Bu durum son iki kural ile birlikte açıklanabilir. Limitsiz tarifesi olan (kota_grup<1.5) aboneler tek başına iptal eğilimi gösterirken kotalı ve adil kullanımı olan aboneler taahhütlü olmaması durumunda iptal eğilimi gösterebilmektedir.

5. SONUÇ VE YAPILABİLECEK DİĞER ÇALIŞMALAR

Bundan önceki bölümlerde iptal analizi ile ilgili literatürde yer almış uygulamaların içerisinden detaylı bir analiz ve gözlem ile bilgi keşif süreci dahilinde yapılan deneyler sayesinde yeni bir metot üretilmiş ve çalışma performansı bir internet sağlayıcısı işletmenin verileri incelenerek değerlendirilmiştir.

Bu bölümde ise sonuç bağlamında gerçekleştirilen uygulamanın önemini ve çalışmada yaşanan bazı zorluklar paylaşılacaktır. Devamında ise yapılan çalışmanın konusu için devam niteliği taşıyan gerçekleştirilmesi gerektiği düşünülen gelecekte yapılabilecek çalışmalar için örnekler verilecektir.

5.1 Sonuç

Yapılan işin gereği telekomünikasyon sektöründeki bir şirket için iptal analizi, pazar payında şirketin yerini koruyabilmek adına, bir seçenekten çok bir zorunluluk haline gelmiştir. Abone kazanmanın güçlüklerinin yanı sıra mevcut abonelerin korunması için bu çalışma kapsamında önerilen çözüm uygulamaları gerçek veri üzerinde yapılan deneylerden de görüleceği gibi katkı sağlayabilecek niteliktedir.

İptal analizinin veri madenciliği uygulamalarının arasında pratik olarak somut çıktıların alınabildiği bir süreç olduğu anlaşılmıştır. Şirketlerin günümüzde üzerinde bir çok konuda uğraştığı müşteri ilişkileri yönetimi süreçlerinin önemlilerinden biri olmaya aday bir bilgi keşi süreci gerçekleştirilmiştir.

Yapılan çalışmalarda kümeleme ve sınıflandırma uygulamaları araştırılmış ve edinilen bilgiler ışığında firma verilerinden kullanılabilecek veriler imkanlar dahilinde edinilmiştir. Verinin toplanması süreci her araştırmacının yaşayabileceği dikkat edilmesi gereken unsurlardan bazıları bu çalışma kapsamında da yaşanmıştır. Bunlardan bazıları;

- Şirket politikası gereği stratejik veya gizli herhangi bir bilgi verilmemesine özen gösterilmiştir.
- Kişisel bilgi gizliliği önemsenerek kişisel veri hiçbir şekilde analizlerde kullanılmamıştır.

Verinin toplanması sürecinde ise özellikle gerçek veri üzerinde Ağustos 2014 tarihinden bir yıl öncesi dikkate alındığından bir yıllık süre zarfında sistem üzerinde yapılan değişiklikler dikkate alınmış verinin doğruluğu teyit edilerek çalışmalar yapılmıştır. Ayrıca demografik verilerden bazıları tutarsızlığa yol açabileceğinden çalışma kapsamından çıkarılmıştır.

Edinilen veriler ile yetinilmeyip anlamlandırmak için çalışmalar gerçekleştirilmiştir. Daha anlamlı verinin her zaman daha iyi sonuç vereceği bilinci bu çalışma kapsamında edinilmiştir. Böylece ham veriden elde edilemeyecek yeni özellikler çalışma kapsamına dahil edilmiştir.

Gerçekleştirilen uygulamaların sürekli iyileştirilmesinin gerektiği ve bunun ancak yaşatılan bir bilgi keşif süreci ile mümkün olduğu görülmüştür. Her sürecin girdisi, yapılan işlemleri ve çıktısı bulunmaktadır. Bilgi keşif sürecinden beklenen çıktının bilgi olduğu düşünülürse, bilginin iyileştirilmesi için yaşayan sistemde girdilerin ve yapılan işlemlerin güncel tutulması gerekmektedir.

Altı-adım bilgi keşif sürecinde veri madenciliğinin aslında sadece sürecin bir aşaması olduğu görülmüştür. Aslında bilgi keşif sürecinin veri madenciliği uygulamaları aşamasının önünde ve arkasında diğer aşamalar ile döngünün yaşatılması ile anlamlı olduğu görülmüştür.

Deneylerin analiz ve sonuçları değerlendirilecek olursa her zaman basit işlemlerle başlanıp komplike yapılara gidilmesinin doğru olduğu görülmüştür. Yapılan tüm deneyler R üzerinde gerçekleştirilmiştir. R'ye destek verenler açık kaynak kod

dünyasında başarılı işler gerçekleştirmiştir. Literatürde ismi geçen makaleler R ile pratik bir şekilde uygulanabilmektedir. Tabii ki bazen veriden veya üretilen modelden kaynaklanan problemler olmaktadır. Ancak bu problemleri aşabilmek için kütüphaneler genişletilebilmektedir.

Geliştirilen iki fazlı model sayesinde zaman serisi türünde müşteri davranış verileri ile birlikte müşterinin standart verileri aynı düzlemde değerlendirilebilir hale gelmiştir. Zaman serisi kümeleme işlemi ile müşteri davranış biçimleri kümelere ayrılmış ve tüm kümeler dikkate alınarak hiyerarşik kümeleme uygulaması üzerinde belirlenen merkezlere olan uzaklıkların hesaplanması yöntemi en iyi sonucu vermiştir.

5.2 Yapılabilecek Çalışmalar

Bu kısımda tez kapsamında değerlendirilmemiş ancak ileride denenmesi mantıklı görünen ve değer katabileceği düşünülebilen uygulamalardan bahsedilecektir.

Öncelikle bu kadar büyük verinin işlenmesi ve sonuç alınması geleneksel yöntemlerle pek mümkün değildir. Performans artırımı ve anlık cevap alabilme adına zaman alan tüm yavaş kısımlar paralel işleme ile daha fazla hızlandırılabilmelidir. Bu bağlamda literatürde birçok araştırma mevcut olup başka bir çalışmanın iptal analizi ve iki fazlı çözüm modeli üzerinden konusu olabilir.

İki fazlı çözüm modelinde algoritmaların gerektirdiği küme sayısı gibi parametrelerin optimize edilmesi gerekmektedir. Her ne kadar en iyi sonucu veren parametreler kullanıldıysa da her zaman serisi özniteliği için farklı küme sayıları kullanılması daha iyi sonuç verebilirdi. Bu şekilde parametrelerde gerçekleştirilecek optimizasyon işlemleri sonuçta büyük oranda katkı sağlayabilir.

Bir başka deney ise örneklem azaltma tekniğidir. Özellikle iki fazlı çözüm modelinin birinci fazında hiyerarşik kümeleme içerisinde gerçekleştirilerek örneklemi azaltılan veri seti üzerinden kümeleme işlem işlemi yapmak hem sonucu etkileyebilir hem de performansı etkileyebilecek bir işlemdir.

Son olarak önerilebilecek çalışma Şekil 3.1'de gösterilen altı-adım bilgi keşif süreci dahilinde sistemi canlı olarak uygulamaya alıp rutin raporlarlar çerçevesinde veri ambarı oluşturulması ve iptal sonucunu doğuran kök nedenlerinin araştırılmasıdır. Ve süreç gereği uygulanan işlemlerin ve deneylerin sıklıkla güncellenerek sonucun kullanılabileceği pazarlama, satış ve müşteri ilişkileri yönetimi gibi alanlarla paylaşılmasını sağlamaktır.

KAYNAKLAR

[1] Chih-Ping, W. ve I-Tang, C., Turning telecommunications call details to churn prediction: a data mining approach, Expert Systems with Applications, 23, 103-112, 2002.

[2] Kotler, P. ve Keller, L., Marketing Management, *Prentice Hall*, New Jersey, 2006.

[3] Kamalraj, N. ve Malathi, A., A Survey on Churn Prediction Techniques in Communication Sector, International Journal of Computer Applications, 2013.

[4] Ling, R. ve Yen, D., Customer Relationship Management: An Analysis Framework and Implementation Strategies, Journal Of Computer Information Systems, 2001.

[5] Roberts-Phelps, G., Customer Relationship Management: How to Turn a Good Business Into a Great One!, *Hawksmere*, Londra, 2001.

[6] Swift, R.S., Accelerating Customer Relationships: Using CRM and Relationship Technologies, *Prentice Hall Professional,* New Jersey, 2001.

[7] Linoff, G.S. ve Berry, M.J.A., Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, *Wiley Publishing*, Indianapolis, 2011

[8] Han, J., Kamber, M., Pei, J., Data Mining, Southeast Asia Edition: Concepts and Techniques, *Morgan Kaufmann*, San Fransisco.

[9] "Big Data Analytics: Descriptive Vs. Predictive Vs. Prescriptive" erişim adresi: http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279, erişim tarihi: 17 Aralık 2014.

[10] Liao, S.H., Chu, P.H., Hsiao, P.Y., Data mining techniques and applications – A decade review from 2000 to 2011, Expert Systems with Applications 39 (12), 11303-11311, 2012.

[11] Wei, C. P., ve Chiu, I. T., Turning telecommunications call details to churn prediction: A data mining approach, Expert Systems with Applications, 23 (2), 103–112, 2002.

[12] Coussement, K., ve Van den Poel, D., Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques, Expert Systems with Applications, 34 (1), 313–327, 2008.

[13] Chen, Z., Fan, Z. , Sun, M., A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioural data, European Journal of Operational Research, 223, 461–472, 2012.

[14] Dimitrios, K., Goce T., Dimitrios G., ve Charu C.A., Time Series Data Clustering, *CRC Press*, Florida, 2013.

[15] Michalis, V., A Practical Time-Series Tutorial with Matlab, 9th The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases PKDD/2005, Porto, Portekiz, Ekim 2005.

[16] Jessica, L., Michail, V., Eamonn, K. ve Dimitrios G., Multi-resolution time series clustering and application to images, *Springer*, Londra, 2007.

[17] Rakthanmanon, T., ve E., Keogh, Fast shapelets: A scalable algorithm for discovering time series shapelets, Proceedings of the 13th SIAM International Conference on Data Mining, Texas, USA, 2013.

[18] Dimitrios, K., Goce, T., Dimitrios, G. ve Charu, C.A., Time Series Data Clustering, *CRC Press*, Florida, 2013.

[19] Liao, T.W., Clustering of time series data - a survey, Pattern Recognition, 38(11), 1857- 1874, 2005.
[20] Hartigan, J.A., Wong, M.A., Algorithm AS 136: A K-Means Clustering Algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics), 28 (1), 100-108, 1979.
[21] Murtagh, F., Legendre, P., Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?, Journal of Classification, 31 (3), 274-295, 2014.
[22] Cortes, C. ve Vapnik, V., Support Vector Networks, Machine Learning, 20, 1-25, 1995.
[23] "Support Vector Machines - The Interface to libsvm in package e1071" erişim adresi: http://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf, erişim tarihi: 17 Aralık 2014
[24] "An Introduction to Recursive Partitioning - Using the RPART Routines" erişim adresi: http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf, erişim tarihi: 17 Aralık 2014
[25] Yıldırım, A., ve Şimşek, H., Sosyal Bilimlerde Nitel Araştırma Yöntemleri, *Seçkin Yayıncılık*, Ankara, 2013.
[26] Li, T. ve Ryan, D., An extended process model of knowledge discovery in databases, Journal of Enterprise Information Management 20 (2), 169-177, 2007
[27] Cios., K. ve Kurgan., L., Trends in data mining and knowledge discovery, *Springer*, Londra, 2005
[28] Burez, J., ve Van Den Poel, D., Seperating financial from commercial customer churn: A modelingstep towards resolving the conflict between the sales and credit department, Expert Systems with Applications, 35, 497-514, 2008
[29] Olson, David L., ve Delen, Dursun, Advanced Data Mining Techniques, *Springer*, Berlin, 2008.

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı	: GÖK, Mehmet
Uyruğu	: T.C.
Doğum tarihi ve yeri	: 10.08.1985 Mersin
Medeni hali	: Evli
Telefon	: 0 (312) 292 40 00
Faks	: 0 (312) 292 40 91
e-mail	: mgok@etu.edu.tr

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek Lisans	TOBB ETÜ/Bilgisayar Müh.	2015
Lisans	Çankaya Üniversitesi/Endüstri Müh.	2008

İş Deneyimi

Yıl	Yer	Görev
2009- Devam Ediyor	Türksat A.Ş.	Uzman Analist

Yabancı Dil

İngilizce