T.C.
HALİÇ ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ SEGMENTASYONU VE MÜŞTERİ KAYIP ANALİZİ

YÜKSEK LİSANS TEZİ

Hazırlayan
Ramis BAŞKAL

Danışman
Dr. Öğr. Üyesi Ülviye HACIZADE

İstanbul – 2019

T.C.
HALİÇ ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ SEGMENTASYONU VE MÜŞTERİ KAYIP ANALİZİ

YÜKSEK LİSANS TEZİ

Hazırlayan
Ramis BAŞKAL

Danışman
Dr. Öğr. Üyesi Ülviye HACIZADE

İstanbul – 2019

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans Programı Öğrencisi Ramis BAŞKAL tarafından hazırlanan ***"Telekomünikasyon Sektöründe Müşteri Segmentasyonu ve Müşteri Kayıp Analizi"*** konulu çalışması jürimizce Yüksek Lisans olarak kabul edilmiştir.

Tez Savunma Tarihi: 27.06.2019

(Jüri Üyesinin Ünvanı, Adı, Soyadı ve Kurumu): İmzası

Jüri Üyesi : Dr.Öğr.Üyesi Ülviye HACIZADE
: Haliç Üniversitesi (Danışman)

Jüri Üyesi : Prof.Dr.Mübariz EMİNLİ
:Haliç Üniversitesi

Jüri Üyesi : Dr.Öğr.Üyesi Faruk BULUT
: İstanbul Rumeli Üniversitesi

Bu tez Enstitü Yönetim Kurulunca belirlenen yukarıdaki jüri üyeleri tarafından uygun görülmüş ve Enstitü Yönetim Kurulunun kararıyla kabul edilmiştir.

Prof.Dr. Nur TUNALI
Vekil Müdür

Ramis Başkal

ORIJINALLIK RAPORU

%12	%8	%1	%7
BENZERLIK ENDEKSI	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCI ÖDEVLERI

BIRINCIL KAYNAKLAR

1	turk.net İnternet Kaynağı	%2
2	Submitted to Istanbul Bilgi University Öğrenci Ödevi	%1
3	www.cigir.com İnternet Kaynağı	%1
4	ceaksan.com İnternet Kaynağı	%1
5	bilgisayarkavramlari.sadievrenseker.com İnternet Kaynağı	%1
6	Submitted to Bahcesehir University Öğrenci Ödevi	<%1
7	Submitted to TechKnowledge Turkey Öğrenci Ödevi	<%1
8	ogzhnyldrm.com İnternet Kaynağı	<%1
9	hasanyavuz.ozderya.net İnternet Kaynağı	<%1

görüldü

27/06/2019

TEZ ETİK BEYANNAMESİ

Yüksek Lisans Tezi olarak sunduğum "TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ SEGMENTASYONU VE MÜŞTERİ KAYIP ANALİZİ" başlıklı bu çalışmayı baştan sona kadar danışmanım Dr. Öğr. Üyesi Ülviye HACIZADE'nin sorumluluğunda tamamladığımı, verileri/örnekleri kendim topladığımı, deneyleri/analizleri ilgili laboratuvarlarda yaptığımı/yaptırdığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim.

(İmza)

Ramis BAŞKAL

ÖNSÖZ

Veri madenciliği ve veri ambarı son dönemlerde hızla gelişerek teknolojinin olduğu sektörlerde kullanımı hızla artmıştır. Telekomünikasyon söktörünün dünyada en hızlı gelişen sektörlerin başında gelmesi ve bu sektörde kullanılan sistemlerin teknolojik ağırlıklı olması nedeniyle veri madenciliği ve veri ambarı sistemlerinin kullanımın bu alanda kullanımı artarak devam etmektedir.

Sürekli gelişen teknoloji ile birlikte, telekomünikasyon sektöründede çok ciddi bir şekilde gelişmeler ve devrim yaşnmaktadır. 1 Nisan 2016 tarihinde hayatımıza giren 4.5G teknolojisiyle birlikte daha geniş ve yeni teknolojiler hayatımıza girmiştir. Bununla birlikte farklı teknolojilerin kullanımına ve geliştirilmesine başlanmıştır.

Bu tez çalışmam sırasında Türkiye'de telekomünikasyon söktöründe öncü firmalardan olan bir firmanın müşteri verileri kullanılarak müşteri davranışlarının tespiti, bu müşterilerin şikayet ve memnuniyetinin tespiti ve bu sayede memnuniyetsiz müşterileri memnun etmek, memnun müşterilerin ise memnuniyetini arttırarak firmaya bağlılığı arttırılmaya çalışılacaktır.

Yapmış olduğum bu tez çalışmasın her aşamasında yardımlarını esirgemeyen, öneri, görüş ve yönlendirmeleri ile hazırlamış olduğum bu tezimin her aşamasında değerli yardımlarını esirgemeyen danışmanım Dr. Öğr. Üyesi Ülviye HACIZADE hocamıza teşekkürü borç bilirim.

Bölüm başkanımız Sayın Prof. Dr. Mübariz EMİNLİ hocamıza ders ve ders dışındaki destek ve yardımlarından dolayı teşekkürü borç bilirim.

Yüksek lisans tezimi hazırlarken kullanmış olduğum veriler için TurkNet İletişim Hizmetleri firmasına ve hertürlü destek ve yardımlarından dolayı TurkNet İletişim Hizmetleri Bilgi Teknolojileri bölüm yöneticisi Yunus Sami ÇELEBİLER'e teşekkürü borç bilirim.

2019 Ramis BAŞKAL

İÇİNDEKİLER

Sayfa No

KISALTMALAR

BTK: Bilgi Teknolojileri Kurumu
CART: Classification and Regression Trees
CLV: Müşteri Ömür Boyu Değeri
DBMS: Database Management System
E/K: Erkek/kadın
ERP: Kurumsal Kaynak Planlama
ETL: Extract-Transform-Load
FNR: Yanlış Negatif Oranı
FPR: Yanlış Pozitif Oranı
GSM : Global System for Mobile
k-means: K Ortalamalar
k-nn: K-En Yakın Komşu
ODS: Operasyonel Veri Merkezi
OLTP: Online Analytical Processing
PSTN: Sabit Telefon
SAP: Sistem Uygulama ve Ürünler
SOM: Self-Organizing Maps
SQL: Structure Query Language
TUIK: Türkiye İstatistik Kurumu
VA: Veri Ambarı
VTBK: Veri Tabanı Bilgi Keşfi
YSA: Yapay Sinir Ağları

SEMBOLLER

b.t : Bilinmeyen tarih

C : Sınıf uzayı

Ch : Elde bulundurma maliyeti

ERR : Hata Oranı

m : Örneklerin sayısı

M : Model

n : Değişkenlerin sayısı

N : Seçenek fiyat sayısı

P : Satın Alma fiyatı

t : Dönem sayısı

TNR : Belirleyicilik

txt : Metin dosya uzantısı

v.d. : Ve diğerleri

Y : Çıktı uzayı

y.n. : Yanlış negetif

y.p. : Yanlış pozitif

ŞEKİLLER

Sayfa No

ÇİZELGELER

Sayfa No

ÖZET

TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ SEGMENTASYONU VE MÜŞTERİ KAYIP ANALİZİ

Günümüzde firmaların yeni müşteriler kazanmak için yaptığı yatırım ve çalışmaların, sahip olunan müşterileri elde tutmak için yapılan çalışmalardan daha fazla masraflı ve zararlı bir süreç olduğu bilinmektedir. Bu nedenle mevcut müşteriyi elde tutmak firmalar açısından daima karlı bir süreç olmaktadır. Bu çalışmanın amacı, telekomünikasyon sektöründe faaliyet göstermekte olan firmalara ait müşteriler temel alınarak, firmanın müşterilerinin, şirketin sunduğu hizmet ve ürünleri terk etmeden önce bu durumun farkına varmak ve müşteri kayıplarını önleyici faaliyetlerde bulunmaktır. Müşterilerin hangi koşullarda ne tür sorunlarla karşılaştığını ve bu sorunların, müşterilerin firmanın sunduğu hizmetlerden duyduğu memnuniyetsizliği ortadan kaldırarak müşterilerin rakip firmalardan hizmet alma ihtimallerini ortadan kaldırmak amaçlanmaktadır. Yapılan bu tez çalışmasında telekomünikasyon sektöründe faaliyet göstermekte olan ve sektörde ön sıralarda bulunan bir firmanın büyük boyutlu müşteri verilerine ulaşılarak, veri madenciliği teknolojilerinin yardımı ile bir müşteri segmentasyonu modeli geliştirilip gerekli sınıflandırmanın yapılması ve bu doğrultuda müşteri kayıp analizinin yapılarak müşterilere farklı kampanya ve ödeme seçenekleri sunup müşteri memnuniyeti arttırılarak müşteri kaybı ortadan kaldırılması amaçlanmıştır. Yazılımı gerçekleştirilen bu modelin telekomünikasyon sektöründe kullanılabilirliği test edilmiştir. Bu kapsamda firmanın sektördeki yapısına uygun bir sistem kurulması ve rakip firmalar ile olan rekabet gücünün arttırılarak daha iyi üst bir seviyeye çıkartılmasında katkı sağlamaya çalışılmıştır.

Anahtar Kelimeler: Müşteri Segmenti, Telekomünikasyon sektörü, Veri madenciliği,Python programı,

ABSTRACT

CUSTOMER SEGMENTATION AND CUSTOMER CHURN ANALYSIS IN TELECOMMUNICATION SECTOR

Nowadays, it is known that the investments and works made by companies to gain new customers are more costly and harmful than the efforts made to hold the customers. Therefore, keeping the existing customer is always a profitable process for the firms. The purpose of this study is to realize this situation and to prevent customer losses before leaving the company's customers and the services and products offered by the company, based on the customers of the companies operating in the telecommunication sector. It is aimed to eliminate the chances of customers to receive services from their competitors by eliminating the customer's problems and what kind of problems they face and eliminating the customer's dissatisfaction with the services offered by the company. In this thesis, a customer segmentation model has been developed with the help of data mining technologies by developing a customer segmentation model with the help of data mining technologies. aimed at eliminating customer loss by increasing customer satisfaction. This model has been tested with Python software and its usability has been tested in telecommunication sector. In this campaign, it was tried to make a system suitable for the structure of the company in the sector and to increase the competitiveness of the competitor companies to a higher level by increasing their competitiveness.

Keywords: Customer Segmentation, Telecommunications industry, Data mining, Python program,

1. GİRİŞ

Günümüzde veri tabanlarının önemini her geçen gün daha da iyi anlamakta ve veri tabanı pazarının sürekli ve hızlı bir şekilde büyümesiyle birlikte öğrenilecek yeni bilgi ve teknolojilerin ortaya çıktığını görmekteyiz. Büyük yada küçük her türlü işletme veya kuruluş bir veritabanına sahip olmakla birlikte sahip oldukları bu veritabanı giderek büyüyen bir yapıya dönüşmektedir. Daha önceden bilgileri saklamak için sayfalarca döküman kullanılmakta iken günümüzde gelişmiş veritabanı teknolojileri sayesinde bu bilgiler çok daha kolay bir şekilde saklanmakta ve istenildiğinde kolayca ulaşılmaktadır.

Veritabanları sadece tamamlanmış bitmiş işlemleri veya olayları tutmamakla birlikte yapılacak olan işlemlerin kayıtlarınıda tutmaktadır. Operasyonel işlemler olarak adlandırılılan bu veritabanları, özellikle daha büyük ölçekli firmalarda anlık erişim gerektiren işlemleri başarıyla gerçekleştirmekte ve firmalara ve iş hayatına büyük kolaylıklar sağlamaktadır.Veri tabanları sadece yapılan ve biten bir işlemi kayıt altına almaz yapılan bir işlemden farklı işlemler yapılmasına imkanlar sağlamaktadır. Özellikle geleceğe yönelik yönelik çalışmalarda müşterilerin yapacağı işlemlerin tahmin edilmesinde ve bu tahminler ile müşterilere yönelik yeni çalışmalar yapılmasında firmalara çok büyük kolaylık sağlamaktadır.Ancak operasyonel bir veritabanından anlık olarak rapor alarak müşterilere yönelik çalışmalar yapmak veritabanında performans sorunlarına neden olacaktır. Bu tür performans sorunlarına engel olmak amacıyla veriambarları oluşturulmaktadır (Usgurlu, b.t.).

Veri madenciliği sahip olduğumuz bilgi ve verilerileri kullanarak müşterilerin hareketlerini tahmin etmek, müşterileri yaptıkları davranışlara göre sınıflandırmak ve gruplara ayırarak onlara yönelik çalışmalar yapmaya yardımcı olmaktadır. Örneğin telekomünikasyon sektöründe faaliyet göstermekte olan bir firma müşteri şikayetlerinden yola çıkarak müşterilerin rahatsız olduğu ve müşteri memnuniyetsizliğine neden olan olayları tespit ederek sonrasında bu memnuniyetsizliği ortadan kaldırmak amacıyla müşterilerine farklı kampanya ve hizmetler sunabilir. Yada mevcutta sunduğu kampanya veya hizmete uygun olmayan müşterileri belirleyerek farklı hizmetler sunabilir. Bu sayede müşteri memnuniyetini

kazanmaya çalışarak müşteri kayıplarının önüne geçmeye çalışarak firmanın zarar etmesinin önüne geçilmekte hatta karını arttırabilmektedir (Usgurlu, b.t.).

Müşteri segmentasyonu, bir müşteri tabanını belirli şekillerde benzer veya birbirinin aynısı olan insan gruplarına bölme uygulamasıdır. Müşteri segmentasyonu ile yapılan gruplama işlemi sayesinde farklı müşteri gruplarına o grubun ihtiyaçları ve uygunluğu doğrultusunda farklı değer önerileri sunabiliriz. Müşteri segmentasyonu yapılırken oluşturulacak olan müşteri segmentleri genellikle kişisel özellikleri tercihler veya müşterinin karlılığını arttıran davranışlarla aynı olması gereken davranışlar gibi benzerlikler ile belirlenmektedir (Harrold, D. 2000, s. 9-10).

Bir müşteri segmentasyon modeli, pazarlama kaynaklarının etkin bir şekilde tahsis edilmesine, çapraz ve yukarı yönde satış fırsatlarının en üst düzeye çıkartılmasına izin verir. Bir grup müşteriye ihtiyaçları doğrultusunda özel bir e-posta gönderildiğinde şirketlerin bu müşterilere özel teklifler sunması ve bu teklifleri kabul ettirmesi daha kolay olmaktadır (Veri Tabanı Modelleri, Anonim, b.t.).

Birçok şirket 80/20 olarak bilinen müşterinin brüt kar marjına uygulanabilecek kuralı kullanmaktadır. Bu kuralı açıklayacak olursak müşterilerin %20'sinden elde edilen karın %80'ine dağıldığı anlamına gelmektedir. Bununla birlikte 80/20 kuralı bazı durumlarda şirketler açısından zararlı olabilir çünkü iç maliyet tek bir boyuta uymamaktadır. Şekil 1.1.'de brüt kar marjı hesaplanamsı için kullanılan 80/20 kuralı gösterilmiştir (Harrold, D. 2000, s. 9-10).

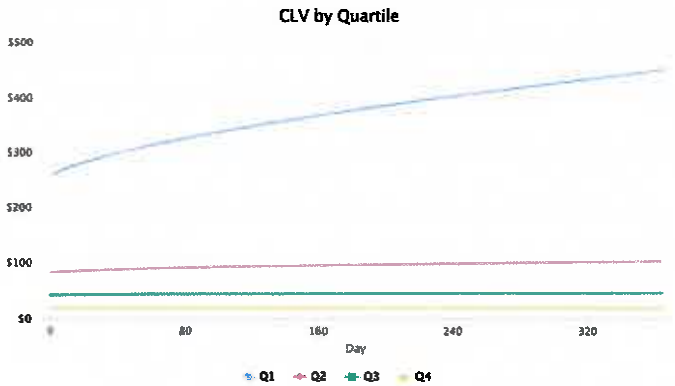
Şekil 1.1. 80/20 Kuralı

Şirketlerin belirli müşteri segmentlerinde diğer segmentlerine yaptığı maliyetlerden daha fazla maaliyette bulunması doğal bir durumdur. Müşteri karlılığı çok basit olmamakla birlikte; bilirli bir müşteri segmentine hizmet sunmak için şirket içerisinde hangi kaynakların kullanıdığını ve tüketildiğini çok iyi bir şekilde izlemek

ve belirlemek gerekmektedir. En azından, şirketler müşteri karlılığını yeni bir anlayış seviyesine yükseltmek için satış, pazarlama ve müşteri hizmetleri maaliyetlerini tahsis etmelidirler (Veri Tabanı Modelleri, Anonim, b.t.).

Müşterilerin sağladığı değeri anlamak için bir başka yaklaşım olan müşteri ömür boyu değeri CLV (Customer Lifetime Value) benimsenmiştir.CLV, sahip olduğu müşterileri varlık olarak değerlendiren bir yaklaşım benimsemiştir. CLV, bir müşteriyi kazanmanın maliyetinin gelecekte kazandıracağı geliri olan bir yatırımı temsil ettiğini kabul etmektedir. CLV'nin hesaplanması, müşteri kazanımı ve elde tutulması oranlarının, her bir segment için satın alma modellerinin ve maliyetlerinin anlaşılmasını ve ardından değerleri günümüze indirmesini gerektirir (Harrold,D. 2000, s.9-10).

Segmentasyon zor ve karmaşık olabilir. Bununla birlikte müşterilerin segmentlere ayrılması 'herkese uyan bir yaklaşım' ilkesiyle ele alındığında şirketlere çok büyük kazançlar sağlayabilmektedir. Şekil 1.2'de Müşteri ömür boyu değeri (Customer Lifetime Value) dönemsel olarak gösterilmiştir (Harrold,D. 2000, s.9-10).



Şekil 1.2. CLV(Müşteri ömür boyu değeri)

2. LİTERATÜR ARAŞTIRMASI

Telekomünikasyon sektöründeki firmalar müşterilerinin detaylı arama kayıtlarına sahitptirler. Firmalar müşterilerine sunacakları hizmet ve ürünler için fiyat ve promosyon stratejilerini belirlemekte sahip oldukları bu müşteri arama kayıtlarını kullanabilirler ve müşterilerini bölümlere ayırabilirler.

Veri madenciliği yöntemleri kullanılarak herhangi bir abonenin faturalarını ödememe durumu veya ilgili firmadan ayrılarak farklı bir firmanın hizmet ve ürünlerini kullanabilme ihtimali önceden tespit edilebilir. Ve bu sayede firmanın maddi ve manevi kayıtları önlenebilir. Bu tür analizler yapmak için standart sapma yöntemi kullanılabilir. Kullanım şekilleri ve alışkanlıklarına göre aboneler belirli gruplara ve kümelere ayrılırlar. Tutarsız özelliklere sahip olan ve tutarsız özellikler gösteren aboneler belirlenerek farklı bit gruba ayrılırlar. Veri madenciliği yöntemleri kullanılarak, uluslararası dolaşım sözleşmeleride optimize edilebilir (Tezcanlar,2007).

Veri madenciliği algoritmaları ve bu sayede elde edilen bilgiler, ticaret, astronomi, güvenlik ve telekomünikasyon, jeolojik araştırmalar gibi birçok farklı alanda başarıyla uygulanmıştır. Bu kullanımlara örneklerden bazıları aşağıda açıklanmıştır.

Ren, Zheng ve Wu çalışmalarında, telekomünikasyon sektörü müşterilerinin davranışlarından yola çıkarak onları genetik algoritmaya dayalı bir yöntem ile kümelere ayırmışlardır. Öncelikle telekomünikasyon müşterilerinin çeşitli özelliklerini (müşteri arama durumları, hizmet kullanımları vb.) ortaya çıkartmışlardır. Daha sonra ise telekomünikasyon müşterilerinin çok boyutlu özellik vektörleri arasındaki benzerlikler , iki boyutlu bir düzleme indirgenerek birbirleri arasındaki mesafeler hesaplanır. Ve son olarak mesefeler genetik algoritma yöntemleri ile kademeli olarak birbirlerine benzer olanlar ve yakın benzerlikte olanlar gruplanır (Harrold,D. 2000, s.9-10).

Müşteri kayıp analizi (Churn) ve yönetimi mobil operatörler için çok önemli bir konuma gelmiştir. Mobil operatörler sahip oldukları müşteri ve aboneleri korumak için müşterilerinin ihtiyaçlarını çok iyi bir şekilde karşılamak zorundadırlar. Müşteri kayıp analizinin zorluğuna çözüm olarak Chang'in veri madenciliği yöntemlerini kullanarak yapmış olduğu müşteri kayıp analizi yöntemi kullanılabilir. Tayvan'ın en büyük telekomünikasyon firmalarından olan bir firmanın müşterilerin faturalama bilgileri, müşteri çağrı merkezi arama kayıtları ve günlük işlem loglarını kullanarak Yapay Sinir Ağları uygulamaları ile müşteri kayıp analizi konusunda çok önemli ve verimli bir çalışma yapmıştır (Harrold,D. 2000, s.9-10).

Büyük bir yazılım sisteminde hangi dosyaların hatalı olduğunu bilmek proje yöneticileri için çok önemli ve değerli bir bilgidir. Sahip oldukları bu bilgileri yazılımın test edilmesinde ve kaynakların bu yönde nasıl verimli bir şekilde kullanılması gerektiği hususunda kullanabilirler. Turhan, Koçak ve Bener çok büyük bir telekomünikasyon sistemi üzerinde yaptıkları çalışma ile yirmibeş proje geliştirmişlerdir. Modellerinin hata eğilimlerini tahmin etmek için, modellerini halka açık olan Nasa MDP verileri üzerinde eğitmişlerdir. Yaptıkları deneylerde yöntemleri ve hata tahmin seviyelerini belirlemek için statik çağrı grafiğine dayalı sıralama yöntemini ve en yakın komşu (K-nn) modellerini örnek olarak kullanmışlardır (March ve Hevner, 2005, s.1).

Telekomünikasyon sektöründe birçok dolandırıcılık yöntemi mevcuttur. Hilas'ın yazmış olduğu makalesi büyük telekomünikasyon firmalarındaki sahte telekom aktivitelerinin tespiti konusunda yazılmış güzel bir makaledir. Bu makalede, üst üste gelen dolandırıcılıkların tespitine odaklanılmıştır.Yapılan bu çalışmada hem ağ yöneticisinin uzmanlık bilgisi hemde veri madenciliği tekniklerinin gerçek dünyadaki veriler üzerine uygulanması sonucunda elde edilen bilgileri içeren uzman bir sistemin kurulmasıyla ortaya çıkar. Hilas yaptığı çalışmada 22000 adet telefon çağrısını 5541 günde uzman sistemler yardımıyla analiz etmiştir (March ve Hevner, 2005, s.1).

Daskalaki ve arkadaşlarının yaptığı çalışmada ise, büyük bir telekomünikasyon firmasının müşteri kayıplarının nedenleri ele almak ve bu kayıpları önlemek için bir karar destek sistemi kurarak neden ve sonuçları ile birlikte müşteri kayıp projesi hakkında çalışma yapmışlardır. Yapmış oldukları bu çalışmanın temel amacı karar ağaçları ve yapay sinir ağları yöntemlerini kullanarak

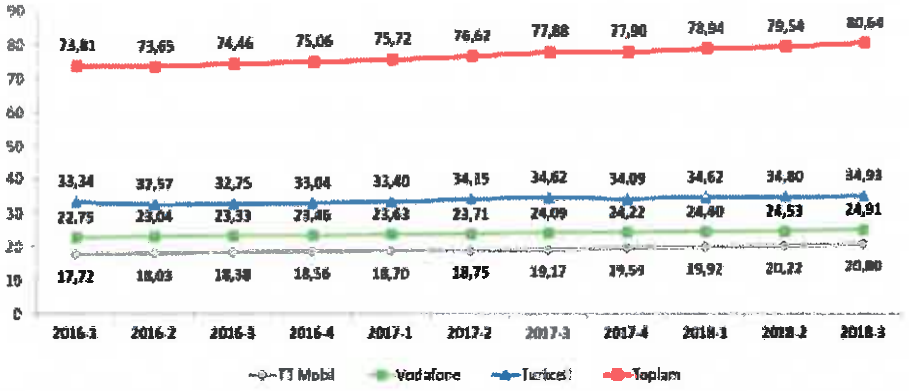
farklı firmalara churn etmek isteyen müşterileri tespit etmektir (March ve Hevner, 2018, s).

Wei ve Chiu ise abonelerden elde edilen ve yapılan sözleşmeye dayalı bilgiler ve müşterilerin yaptıkları çağrılardan elde ettikleri bilgiler üzerinde müşteri kayıp analizi tahminleme yöntemlerini deneyerek müşteri kayıplarını önleyici bir yöntem sunmuşlardır. Bu önerilen yöntem, belirli bir sürede sahip olunan potansiyel müşterilerden sözleşme sonlandırma eğiliminde olanları belirleyebilmekte etkili olmuştur. Wei ve Chiu bu çalışmada, Tayvan'da 21 milyon abonesi olan ve ülkenin en büyük müşteri potansiyeline sahip olan bir telekomünikasyon firmasının verilerini kullanmışlardır. Müşterilerin arama kayıtları ne kadar fazla ise müşteri churn analizinin yapılmasından elde edilen sonuçların doğruluk oranlarıda o kadar doğru olduğunu belirtmişlerdir (March ve Hevner, 2005, s.1).

Günümüzde Türkiye'de bulunan tüm telekomünikasyon firmalarının veri madenciliği yöntemlerini kullanarak müşteri segmentasyonu ve müşteri kayıp analizi konularında çalışmalar yaptıkları apaçık ortadadır. Örneğin Telsim firmasını satın alan Vodafone, satış, pazarlama, finansal yönetim, gelecek tahmini ve birçok farklı ihtiyaçları için veri madenciliği yöntemlerini kullanmaktadır. Vodafone firması sahip oldukları veritabanlarını kullanarak en yoğun işlem saatleri belirler ve iletişimde herhangi bir aksama veya kesinti olmaması için bu saatlerde daha yoğun bir mesai ve işgücü planlaması yapar. Turkcell iş zekası ve veri madenciliği tekniklerini kullanarak müşteri bilgilerini inceledikten sonra müşterilerine yeni tarife ve hizmetler sunmakla birlikte mevcut müşterileri için yeni kampanyalar geliştirmektedir. Turkcell ayrıca müşteri sadakatini arttırıcı programlarda geliştirmektedir (Gray ve Watson,1998, s.8-9).

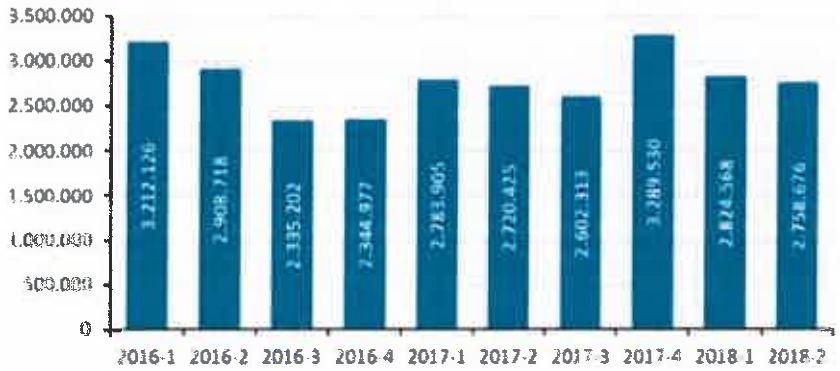
Turknet firması ise sahip olduğu büyük boyutlu müşteri verilerini çeşitli veri madenciliği yöntemleri kullanarak churn tahmin modelleri oluşturmuş ve müşteri memnuniyetini arttırıcı çalışmalar yapmıştır. Bununla birlikte Turknet'ten ayrılma eğilimi bulunan müşterilerini duyarlı kuş sürüsü optimizasyonu tabanlı bir sınıflandırma yöntemiyle sınıflandırdıktak sonra ayrılma eğilimi bulunan müşterilerine farklı kampanya ve ödeme kolaylıkları sunarak bu müşterilerin firmada kalmasını sağlamış bu sayede müşteri sadakatinide arttırmıştır (Gray ve Watson,2018, s.8-9).

Şekil 2.1.'de 2018 yılı 3. Çeyrek soru itibari ile Türkiye'de mobil işletmeci bazında abone sayıları gösterilmiştir (Gray ve Watson,2018, s.8-9).



Şekil 2.1. Mobil İşletmeci Bazında Toplam Abone Sayıları, Milyon

Türkiye'de 2008 yılı sonunda kurulan numara taşıma hizmeti ile birlikte telefon sağlayıcıların sayısıda giderek artmıştır. Rakiplerinden toplam 463000 müşteri alan Avea, Turkcell'den 335000 ve Vodafone'den 128000 müşteriyi kendi müşteri ağına katmıştır. Bu dönemde Avea rakip operatörlerden abone alan tek operatördür. Şekil 2.2.'de 2018 yılı 2. Çeyrek soru itibari ile BTK'dan alınan Türkiye'deki mobil numara taşıma adetleleri gösterilmiştir (Theodoratos ve Sellis, 1999, s 279-301). Şekil 2.2.'den anlaşılacağı üzere numara taşıma işlemleri operatörler için çok önemli bir müşteri kaynağı olarak görülmektedir.



Şekil 2.2. Toplam Mobil Numara Taşıma Sayıları

3. TELEKOMÜNİKASYON SEKTÖRÜ

Hem ekonominin hemde toplumun tamamı için vazgeçilmez bir öneme sahip olan bilgi ve iletişim sektörü özellikle son yıllarda ve dönemlerde çok büyük gelişmeler göstermiş ve insanlık için vazgeçilmez bir yer edinmiştir. Dünya üzerindeki bütün sektörlerin teknolojik sistem ve altyapılarını geliştirmelerini ve müşteri deneyimlerini arttırarak sorunsuz ve kusursuz bir hale getirebilmeleri, diğer firmalar ile olan rekabet gücünü arttırmak için çok önemli bir ön koşul ortaya çıkmıştır. Telekomünikasyon alanı ve sektörü için teknoloji ise var olabilmenin en önemli ve tek koşulu haline gelmiştir. Hizmet kalitesi ve teknolojik altyapı gibi faktörler telekomünikasyon sektörünün tüm zamanlardaki en önemli önceliği olan ana başlığı oluşturmaktadır (Han ve Kamber, 2000).

Örnek verecek olursak düne kadar sadece bir problem olan ağ bağlantısı günümüzde ise insan yaşamı üzerinde tehlike yaratabilecek bir öneme sahiptir. Henüz süreçler tamamlanmamış olabilir ancak çok kısa ve yakın bir gelecekte her bir bireyin ve her bir nesnenin birbirine bağlantılı hale geleceğini açıkça görebiliyoruz.

Bağlantı yapısının kaliteli olması yaşamın tehlikesiz ve kesinti olmaksızın devam etmesini sağlayacak önemli bir faktördür. Günümüzde hergeçen gün artan ağ trafiği ve birbirine bağlantılı cihaz sahısı tüm sektörlerin içerisinde telekomünikasyon sektörünün en önde olmasını sağlayan en önemli etkenlerdir. Telekomünikasyon sektörünün ilk adımları sahip oldukları ağlarını 5G teknolojisine yükselterek gerekli bant genişliği ve gerekli hızı elde etmek ve tüm bu yatırımlar ile birlikte sahip oldukları müşteri ve abone sayılarını koruyarak en kaliteli ve en iyi müşteri deneyimi konusunda taviz vermemektir. Yapılan yeni yatırımlar müşteri deneyimini ve buna bağlı olarak müşteri deneyimi ise firmaların Pazar paylarını dğrudan etkilemektedir (Han ve Kamber, 2000).

Tüm dünyanın merkezinde ve odağında olan büyük verinin ise telekomünikasyon firmaları açısından çok farklı bir önemli ve anlamı bulunmaktadır. Ayrıca tüm bu büyük verilerin büyük ve önemli bir kısmının telekomünikasyon sektörleri taradından sağlandığı bilinmelidir. Firmaların sahip oldukları aboenelerin

kişisel verilerini bir kenara koyacak olsak dahi telekomünikasyon firmalarının ve mobil ağların günlük işlemleri, performans hesaplamaları ve yapılan hertürlü işlemlerin ağ operasyonları tarafından üretilen çok geniş ve zengin bir veri kaynağına sahip olduklarının en önemli göstergesidir.

Sahip olunan tüm bu verilerin büyük bir özen ve titizlikle analiz edildikten sonra sisteme dahil edilmesi telekomünikasyon sektörünün hayatının devam etmesi ve bu sektörün yarınlara güçlü ve emin adımlarla taşınmasını sağlayacaktır. Yapay zeka, dijital ödeme sistemleri, nesnelerin interneti, bulut teknolojileri ve mobil teknolojilerin temelinde büyük verilerin analiz edilmesi bulunmaktadır (Yarımağan, 2000, s.291).

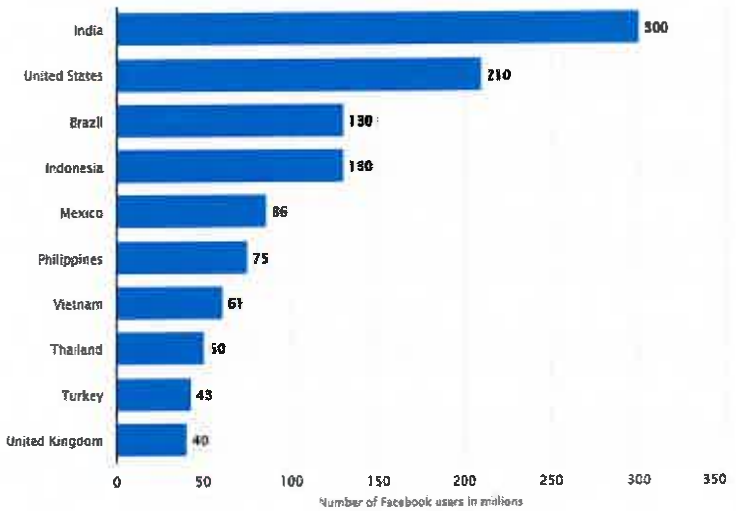
Bu özelliklerin yanında ise pazarlama açısından hız, kolaylık, müşteri memnuniyeti gibi terimlerin yanı sıra ileri düzeyde bir deneyimin ötesinde güvenlikte gerekmektedir. Çünkü sürekli olarak birbiri ile iletişim halinde olan ve birbirine bağlı olan yeni dünya sisteminde güvenlik insan hayatınında içerisinde olduğu riskleri önlemek anlamına gelmektedir. İşte bu nedenle kriz çözme becerisinin öneminden ziyade çıkabilecek olan krizleri önceden tahmin edebilmek, görebilmek ve çıkan krizleri çözüme kavuşturma yeteneğinin gelişimi dahada önemli bir yer tutmaktadır.

Saymış olduğumuz tüm bu konuların merkezinde ise kurumsal kültür yer almaktadır. Telekomünikasyon sektöründeki bireylerin daima kurumsal kültürlerini öncelikli olarak masaya yatırmaları gerekmektedir. Nesnelerin internetini, daha güvenli ağları, yapay zekayı, büyük veriyi ve mükemmel bir müşteri deneyimini inşaa edecek olan bu kültürü ise liderler ve çalışanlar oluşturacaktır. Tüm yapılan robotik çalışma ve teknolojiler insanlık ve iş gücüne karşı bir tehdit olarak görülsede henüz günyüzüne çıkmamış birçok yeni meslek dalının ortaya çıkmasında çok büyük bir rol oynamaktadır. Bu nedenle telekomünikasyon firmaları geleceklerini bu yönde planlamalıdırlar (Yarımağan, 2000, s.291).

Telekominikasyon söktörü sadece ülke genelinde olmamak ile birlikte tüm dünyada sürekli olarak gelişim ve değişim göstermektedir. Finlandiya, İsveç ve İngiltere gibi bazı Avrupa ülkelerinde söktörde liberalleşme şeklinde pazarlar açılmıştır. Diğer ülkelerin etkileriyle birlikte onbeş Avrupa ülkeside liberalleşmiştir. Diğer gelişmekte olan Peru, Malezya ve Şili gibi bazı ülkeler devletlerin sahip oldukları telefon tekelini sonlandırmışlardır. Tüm bu yaşanan gelişmeler ve yeni

oluşumlar sayesinde ülkelerde ve dünyada yeni yeni pazar fırsatları ortaya çıkmıştır (Yarımağan, 2000, s.291).

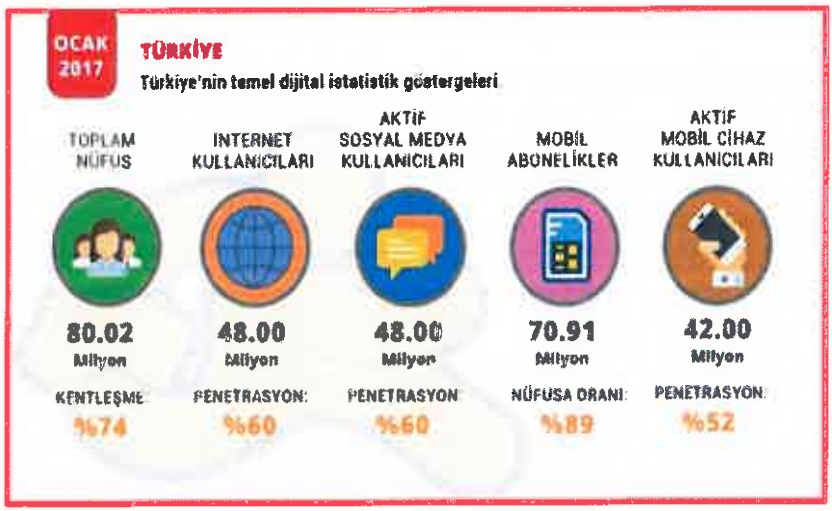
Telekomünikasyon söktöründe yaşanan yeni gelişmeler sadece dünyada değil Türkiye'de de yaygınlaşmış ve internet ve sosyal medya kullanımını önemli derecede arttırmıştır. Günümüzde Türkiye'deki nüfusun yaklaşık yüzde 48'i internete rahat bir şekilde erişebilmekte ve aktif olarak kullanmaktadır. Yapılan araştırmalardan yola çıkılarak internet kullanımının en büyük nedeni sosyal medya olduğu görülmektedir. Çok büyük bir kullanıcı kitlesine sahip olan facebook kullanıcı ülkeler açısından sınıflandırılacak olursa Türkiye bu sıralamada dokuzuncu ülke olarak en fazla kullanıcıya sahip ülke konumunda yer almaktadır. Şekil 3.1.'de 2018 yılı sonu itibariyle facebook kullanıcılarının ülkelere göre gruplandırılmış adetleri gösterilmektedir (Yarımağan, 2000, s.291). Şekil 3.1'de görüldüğü gibi Türkiye'deki facebook kullanıcı sayıları diğer ülkelere göre düşük seviyede görülmektedir.



Şekil 3.1. 2018-Dünyada ve Türkiye'de Facebook kullanım oranları

İnternet kullanıcı sayısı dünyada olduğu gibi Türkiye'de de çok hızlı bir şekilde artmaktadır.2015 yılı sonu itibariyle 11.5 milyon olan internet kullancı sayısı 2017 yılı itibariyle 48 milyon internet kullanıcı ve penetrasyon oranı olarak %60 seviyesinde gerçekleşmiştir. Mobil hizmet kullanan abone sayısı yaklaşık 71 milyon seviyesinde ve nüfus ile oranı %89 sevisesindedir. Aktif mobil cihaz kullanım sayısı ise 42 milyon ve penetrasyon oranı ise %52 seviyesinde gerçekleşmiştir. Tüm bu değerler göz önünde bulundurularak genel bir değerlendirme yapıldığında ise Türkiye'nin telekomünikasyon sektöründe sahip olduğu potansiyeli ve bu

potansiyelin her geçen gün hızla artarak devam ettiğini açıkça görebilmekteyiz. Bununla birlikte Türkiye'nin geleceğin ekonomisinde söz sahibi olmak adına telekominikasyon söktöründe yapacağı yatırımların ne kadar önemli olduğu görünmektedir. Şekil 3.2.'de Ocak 2017 itibariyle Türkiye'deki internet ve sosyal medya gibi telekom söktörüne ait kullanıcı sayılarının nüfusa oranları gösterilmiştir. Şekil 3.2.'de görüleceği üzere Türkiye'de dijital ortam kullanım oranları nüfusa oranlar yüksek seviyelerde görülmektedir (Yarımağan, 2000, s.291).



Şekil 3.2. Türkiye'nin temel dijital istatistik göstergesi

2016 yılı son çeyrek itibariye Türk Telekom ve mobil şebeke işletmecileri ve diğer işletmecilerin net satış gelirleri 10,4 milyar TL seviyesinde gerçekleşmiştir. Yaşanan tüm bu gelişme ve büyümeler telekomünikasyon sektöründeki yapılması gereken yatırımların ne kadar gerekli ve önemli olduğunu göstermektedir. Telekominikasyon sektöründe yer alan öncü ilk 5 firmanın yapmış olduğu yatırım değerleri 2016 yılının ilk üç ay sonundaki değeri 823 milyon TL olarak gerçekleşmiştir (Gray ve Watson, 1998, s.14).

3.1. TURKNET'in Telekominikasyon Sektöründeki Yeri

1996 yılında kurulan TurkNet, kurumsal telekom pazarında bireylere ve işletmelere,kendilerine özel ihtiyaçlarını karşılamak üzere tasarlanan, telefon, toplu internet, IP MPLS VPN (özel sanal ağ), veri merkezi ve barındırma hizmetleri sunan, Türkiye'nin en güçlü bağımsız yeni nesil telekom operatörüdür. TurkNet'in, yaygın ve yüksek kapasiteli ulusal veri omurgası Türkiye'nin her bölgesindeki kişi ve kurumlara yüksek kalite ve yüksek güvenilirlikli hizmetleri verebilmek için

tasarlanmıştır. Türkiye'nin en büyük 100 bilişim şirketi arasında yer alan TurkNet yerel, ulusal ve uluslararası bağlantıları için büyük ölçekli fiber optik kablo tesis yatırımları gerçekleştirmiştir ve bu yatırımları sürdürmektedir (Gray ve Watson, 1998, s.14). Çizelge 3.1'de TurkNet'in yıllara göre kilometre taşları gösterilmiştir.

Çizelge 3.1. TurkNet in yıllara göre kilometre taşları

1996	TurkNet, NetOne Telekom adıyla, kurumlara Internet erişimi sağlamak amacıyla kuruldu.
2002	TurkNet (NetOne Telekom) kurumsal pazara IP, MPLS, VPN hizmeti sağlayan ilk operatör oldu.
2007	Fiber kablo yatırımları için altyapı lisansını alan TurkNet (NetOne Telekom) tüm Türkiye'de geçerli altyapı lisansını alan ilk operatörlerden oldu.
2007	TurkNet Veri Akışı Erişimi (VAE) üzerinden ADSL ve G.SHDSL hizmetlerini sunan ilk operatör oldu.
2007	NetOne Telekom, tasfiye halindeki Sabancı Telekom'dan Turk.net markasını satın alarak TurkNet İletişim Hizmetleri A.Ş. unvanını aldı.
2009	TurkNet, Türkiye'nin ilk veri merkezini kuran ve işleten SATKO Telekom Oteli'ni satın alarak, Telekom Odaklı Veri Merkezi'ni de bünyesine kattı.
2010	Şubat ayında yeni nesil operatörler UMTH yerine, şehir içi yönünün de dahil olduğu Sabit Telefon Hizmeti (STH) vermeye başladı.
2010	TurkNet Yerel Ağın Paylaşıma Açılması (YAPA) kapsamında Türk Telekom santrallerine kendi altyapısını kurmaya başladı.
2011	Yeni nesil operatörler numara taşıması ve numara tahsisi yapabilmeye başladı.
2014	TurkNet Türkiye genelinde yeni nesil şebeke üzerinden, tamamen kendi altyapısıyla erişim ve telefon hizmetleri sunuyor ve İstanbul'un merkezinde operatör bağımsız bir veri merkezi işletiyor.
2014	Bireysel abone sayısı 3 kat artarak 140.000'i geçti.
2015	Online Satış Kanalı aktif hale geldi.
2016	Taahhütsüz Özgür İletişim tarifesi bireysel pazara sunuldu.
2016	Kendi cihazlarımızla santrallerde hizmet verdiğimiz YAPA (Yerel Ağın Paylaşımı) abone sayısı 30.000'e yaklaştı.

TurkNet'in telekomünikasyon sektöründeki yeri ise şu şekildedir. 2015 yılı sonu itibariyle11.493.057 sabit telefon abonesi bulunan Tuknet' in Türkiye'de penetrasyon oranı bir önceki çeyreğe göre 0,5 puan azalarak yaklaşık %14,6 seviyesine düşmüştür. Sabit telefon hizmetleri (STH) abone sayısı açısından ilk üç sırada yer alan işletmeler arasında %9,9 pazar payı ile TURKNET 3. sırada yer almaktadır. Net satışlarına göre işletmecilerin pazar paylarını incelediğimizde ise %3,5 ile Turknet 7. sıradadır. Ayrıca Turknet' in internet servis sağlayıcısı olarak abone sayısı bakımından pazar payı %2 olup net satışları bakımından pazar payı %1,4' tür (Gray ve Watson, 1998, s.14).

Telekomünikasyon sektörünün bu kadar hızlı bir şekilde büyüyüp gelişmesi insanların kullandıkları akıllı telefonlar, bilgisayarlar, tabletler veya bilgisayarlar gibi

dijital cihazlardan sürekli olarak çok büyük boyutta data üretmesine olanak sağlamaktadır. Telekomünikasyon sektöründe üretilen veriler çok büyük ve karmaşık boyutta olduğundan veri madenciliği teknikleri kullanarak bu verilerin işlenerek anlamlı hale getirilmesinde çok büyük zorluklar meydana çıkmaktadır. Bununla birlikte bu devasa boyuttaki veriler zor ve karmaşık olmalarının yanı sıra içlerinde çok önemli ve değerli bilgilerde barındırmaktadırlar.

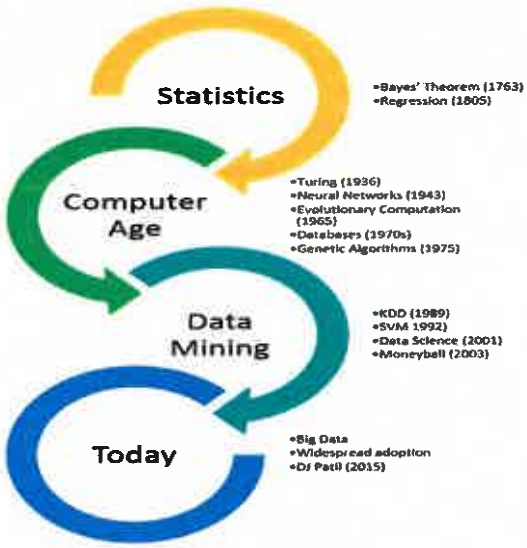
Normal bir yapıda kullanılan geleneksek veri madenciliği yöntemleri karmaşık ve büyük boyuttaki verileri işleyerek bunlardan anlamlı veriler çıkarmakta yetersiz kalabilir. Bunun nedeni ise geleneksel veri madenciliği teknikleri büyük veriler ile gelen heterojik hacim, hız, doğruluk ve gizlilik ile başa çıkamamaktadırlar. Bu nedenle yapılan çalışmalar son zamanlarda bu konu üzerine yoğunlaşmışlardır. Büyük boyutta ve karmaşık olan verileri eş zamanlı olarak büyük veri madenciliği yöntemleri ile işleme gereksinimi ortaya çıkmıştır. Tüm bu bilgiler doğrultusunda telekomünikasyon sektöründe faaliyet göstermekte olan şirketler müşterilerinden topladıkları büyük ve karmaşık boyuttaki verileri veri madenciliği yöntemleri ile işleyerek anlamlı hale getirmişler ve birçok problemin çözümüne fayda sağladığını keşfetmişlerdir. Bu hususta veri madenciliği yöntemlerini etkin bir şekilde kullanmaktadırlar (Gray ve Watson, 1998, s.14).

4. VERİ MADENCİLİĞİ

Veri madenciliğinin tanımı ile ilgili olarak birçok tanım olmasına rağmen genel anlamda birbirinden çok farklılıkları bulunmamaktadır. En genel anlamıyla veri madenciliği büyük ve karmaşık miktardaki verilerin içerisinden geleceğe yönelik tahminlerin yapılmasına yardımcı olacak anlamlı ve yararlı bilgilerin bilgisayar programları aracılığıyla ortaya çıkartılması ve analiz edilmesi anlamınada gelmektedir (Meyer ve Cannon, 1998, s.20-21).

4.1. Veri Madenciliğinin Tarihi Gelişimi

Veri madenciliği terimi yeni teknoloji haberleriyle sık sık anılıyor olması nedeniyle veri madenciliğinin tarihinin çok kısa bir süre önce başladığı düşünülebilir. Ancak veri madenciliği 1700'lü yıllarda Bayes Teoremi ve 1800'lü yıllardaki Regresyon analizi gibi çoğunlukla veri içindeki paternleri tanımlayan erken veri madenciliği metodlarıyla başlayan bir terimdir (Meyer ve Cannon, 1998, s.20-21).



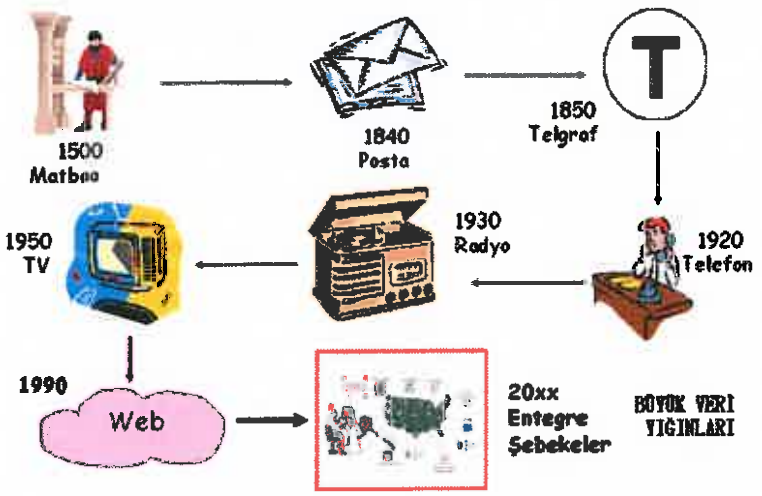
Şekil 4.1. Veriden bilgi edinmenin tarihçesi

Şekil 4.1.'de veriden bilgi edinme sürecinin tarihçesi ve veri madenciliğinin tarihi gelişimi gösterilmiştir (Meyer ve Cannon, 1998, s.20-21)..

1960'larda veriler elektronik ortamda toplanmaya ve geçmiş veriler bilgisayarlar ile analiz edilmeye başlanmıştır. 1980'lerde ise bağlantılı (relational) veritabanları ve SQL (Select Query Language) yapısal sorgulama dili ile verilerin dinamik ve anlık olarak analiz edilmesine olanak sağlanmıştır (Gray ve Watson, 1998, s.37).

Sinir ağları, kümeleme, genetik algoritmalar(1950), karar ağaçları (1960) ve destek vektör makineleri (1990) gibi bilgisayar bilimlerinde yaşanan değişimler de veri madenciliğinin gelişiminde etki etmiştir.

Veri madenciliği üç farklı disiplinden beslenmektedir; istatistik (veriler arasındaki sayısal ilişkilerin ortaya çıkarılması), yapay zekâ (yazılım veya makineler tarafından insan benzeri istihbarat üretme) ve makine öğrenmesi (verilerden öğrenerek tahminler çıkarabilen algoritmalar). Şekil 4.2'de gelişen bilişim teknolojisi ve veri oluşum süreci tarihleri ile gösterilmiştir (Gray ve Watson, 1998, s.37).



Şekil 4.2. Gelişen bilişim teknolojisi ve veri oluşumu

2000 li yıllarda veri ambarlarının kullanımıyla birlikte veri madenciliği giderek yaygınlamaya başlamıştır.Ayrıca veri madenciliği çok büyük ve karmaşık miktardaki verileri inceleyerek bunlar arasındaki bağlantıları ortaya çıkartmaya ve veri tabanı içindeki kayıtlı bilgilerden gizli kalmış olanların ortaya çıkartılmasına fayda sağlayan bir veri analiz yöntemidir.

4.2. Veri Madenciliğinin Diğer Disiplinlerle İlişkisi

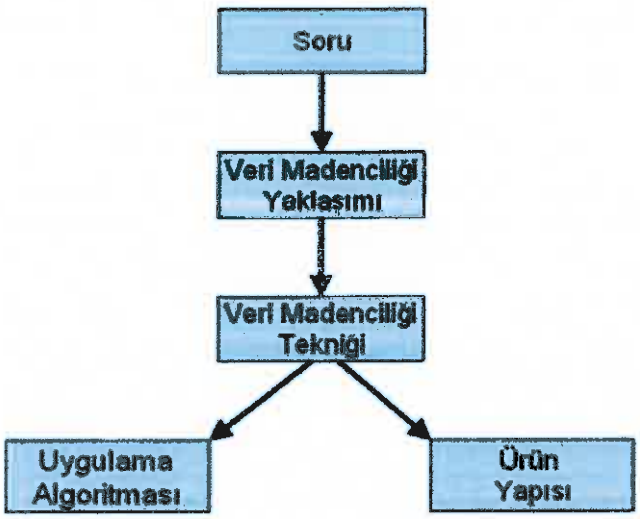
Günümüzde belirli bir konu hakkında karar verme sürecine ihtiyaç duyulan birçok alanda veri madenciliği uygulamaları yaygın olarak kullanılmaktadır. Örneğin pazarlama, biyoloji, bankacılık, sigortacılık, borsa, perakendecilik, telekomünikasyon, genetik, sağlık, bilim ve mühendislik, kriminoloji, sağlık, endüstri, istihbarat vb. birçok alanda başarılı bir şekilde veri madenciliği yöntemlerinin uygulandığı görülmektedir. Son 20 yılda Amerika Birleşik Devletleri'nde gizli dinlemelerden vergi kaçakçılığının ortaya çıkartılmasına kadar birçok alanda veri madenciliği algoritmalarının kullanıldığı bilinmektedir.Bununla birlikte kaynaklar yapılan çalışmalar incelendiğinde veri madenciliği yöntemlerinin en çok tıp, biyoloji ve genetik alanlarında kullanıldığı görülmektedir (Gray ve Watson, 1998, s.37).

Veri madenciliği ile ilgili işlemler yapılırken birçok alan kullanılmaktadır.Bu alanlar içerisinde Şekil 4.3.'de gösterildiği gibi, veri tabanı sistemleri, veri görselliği, yapay sinir ağları, istatistik, yapay öğrenme, vb. gibi disiplinler bulunmaktadır (Gray ve Watson, 1998, s.37).

Şekil 4.3. Veri Madenciliği Uygulama Alanları

Varlığıyla bulunduğu yerde bir anlam ifade eden verilerin bir disiplin altında toplanarak belirli bir teknik veya teknikler kullanılarak işlenmesiyle artıklarından arındırılması sonucunda anlamlı bilgileri gün yüzüne çıkaran ve geleceğe yön verebilmemizi sağlayan ve bu sonuçlara matematiksel ve istatistiksel yöntemlerle mutlaklık kazandıran bir süreçtir.

Veri madenciliği araçları kullanılarak, işletmelerin daha etkin kararlar almasına yönelik karar destek sistemlerinde gerekli olan eğilimlerin ve davranış kalıplarının ortaya çıkarılması mümkün olmaktadır. Geçmişteki klasik karar destek sistemlerinin kullanıldığı araçlardan farklı olarak, veri madenciliğinde çok daha kapsamlı ve otomatize edilmiş analizler yapmaya yönelik, birçok farklı özellik bulunmaktadır.



Şekil 4.4. Veri madenciliğinin yapısı

Şekil 4.4'te belirtilen veri madenciliği yapısının soru aşamasında karşı karşıya olunan sorular net bir şekilde ortaya konmalı ve bu sorulara cevap bulmak için istek ve kararlar oluşturulmalıdır (Gray ve Watson, 1998, s.37). Örneğin hangi müşteriler firmaya sadıktır?

Veri madenciliği yaklaşımı aşamasında ise tahmine veya tanımlamaya yönelik yöntemler belirlenmelidir. Gruplama, regresyon ve sınıflandırma vb.

Veri madenciliği tekniği aşamasında ise veriler arasındaki ilişkileri ortaya çıkarabilecek matematiksel ve sezgisel yöntemlerin uygulandığı aşamadır.

4.3. Veri Madenciliği Süreçleri

Data mining Türkçeye veri madenciliği olarak çevrilmiş olsa da temelinde yatan anlam Veri tabanlarında bilginin keşfi (The Knowledge Discovery in Databases KDD) olarak temsil edilmesi daha doğru bir kavrayışa olanak vermektedir. Bilginin keşfi yolculuğunda önde gelen metodoloji CRISP-DM (Cross

Industry Standard Process for Data Mining)'de aşağıdaki belirtilen maddelerden oluşmaktadır.

- Problemi/İşi/Sorunsalı Kavrama (Business Understanding)
- Veri/Veri Setlerini Anlama (Data Understanding)
- Datayı Hazırlama/Ön İşleme (Data Preparation)
- Modelleme (Modeling)
- Değerleme (Evaluation)
- Yayınlama/Canlıya Alma (Deployment)

Özet olarak yapılan işlemler üç başlık altında toplanabilir; ön işleme, veri madenciliği ve sonuçların doğrulanması.

4.3.1. İşi Anlama Aşaması

Veri madenciliği algoritmaları uygulanmadan önce hedef veri seti yekpare olmalıdır. Veri madenciliği, yalnızca verilerin içinde bulunan desenleri ya da özellikleri (attribute) ortaya çıkarabileceğinden, hedef data setleri, bu desenleri içerecek kadar geniş ve kabul edilebilir bir süre içerisinde sonuç verebilecek kadar büyüklüğe sahip olmalıdır. Data temizliği, kirlilik ve kayıp verinin nasıl işleneceği (missing value), manipülasyon, data mart ve veri tabanı uygulamaları bu aşamanın içerisindedir. Veri madenciliği yapılan çalışmalarda başarılı bir sonuç elde etmenin en önemli şartı kurulacak olan uygulamanın hangi kuruluş amacı için yapılacağına açık ve net bir şekilde karar vermektir. Bu doğrultuda çalışmanın amacı sorunlar üzerine odaklanmış ve açık bir şekilde net olarak ifade edilmiş olmalı ve çalışma sonrasında elde edilecek sonuçların başarı düzeyinin nasıl ölçüleceği tanımlanmalıdır. Yapmış olduğumuz veri madenciliği çalışmasında kullandığımız veriler eğer ki sorunlarla uyuşmuyorsa yapılan çalışma sorunu çözmeyeceği gibi ortaya farklı problemlerin çıkmasına neden olacaktır.Bununla birlikte yanlış kararlar sonucunda katlanılacak olan maliyetlere ve doğru tercihler ile elde edilecek olan kazanımlara yönelik öngörülerede bu safhada yer verilmelidir.

Bir veri madenciliği çalışması yapmak için uygulamamız gereken en öncelikli durum iş probleminin veya durumunun tanımlanmasıdır. Bu maddeye örnek olarak durmadan müşteri kaybı veya iptali yaşayan bir şirketin bu kayıpların nedenlerini öğrenmek ve müşteri kaybı yaşamadan önce bunları belirleyerek engellemek istemesidir. Yapılacak olan bu tür çalışmalara churn ismi verilmektedir. Müşteri segmentasyonu veya churn gibi çalışmalar için öncelikle yapılması gereken iş

çalışmanın amacının iyi bir şekilde anlanmasıdır. Yapılacak olan çalışma için iş anlayış safhası veri madenciliği çalışmalarının birinci aşamasıdır ve bu aşamanın doğru olması bundan sonraki aşamalarında sağlıklı ve doğru bir şekilde ilerlemesini sağlayacaktır. Bu safhada; hedefler, amaç ve ön stratejiler belirlenir. Şekil 4.5.'te iş anlayış safhasının adımları gösterilmiştir (Meyer ve Cannon, 1998, s.35).

Şekil 4.5. İş anlayış safhası

İş anlama safhasının adımları şu şekilde sıralanmaktadır;

- İş hedeflerini belirlemek
- Mevcut durumu değerlendirmek
- Veri madenciliğinin amacını belirlemek
- Proje planı oluşturmak

4.3.1.1. İş Hedeflerini Belirleme

Çalışmayı yapacak olan analistin ilk amacı öncelikle müşterinin isteklerini analiz ederek tam anlamıyla ne istediğini anlamaktır. Çalışmayı yapacak olana analist projenin sonuçlarını etkileyecek önemli kriterleri belirler.

Temel iş faktörleri belirlendikten sonra bu faktörlerin müşterilerle ilişkilendirilecek diğer sorularda türeyebilecektir. Buna örnek olarak internet aboneliğini iptal ettiren bir internet müşterisi ilk olarak kullandığı internetin düşük internet hızıyla müşteriye verilmesinden ne kadar rahatsız?

4.3.1.2. Mevcut Durum Değerlendirme

Bu safhada çalışmada kullanılacak datanun belirlenerek gerekli olan tüm kurallar, kaynaklar ve diğer etkenler detaylı bir şekilde incelenir. Yapılacak olan çalışmada kullanılacak sistemin yazılım ve donanım bilgileri ve veriler dikkatle incelenir. Çalışmanın sonucunun başarısız olmasına sebep verebilecek etkenler listelenir ve terminoloji belirlenir. Yapılacak olan çalışmaya yapılacak maliyetler belirlenir ve proje sonunda sağlanacak olan kar ve faydalar ile karşılaştırılır.

4.3.1.3. Veri Madenciliğinin Amacını Belirleme

Yapılacak olan çalışmanın amacı net olarak belirlenir. Örneğin sahip olunan mevcut müşterilerin memnuniyetini arttırarak firmaya bağlılığını arttırmak veya daha fazla müşteri hedefine ulaşarak müşteri sayısını arttırmak gibi.

4.3.1.4. Proje Planı Oluşturma

Yapılacak olan veri madenciliği projeleri için amaçların ve hedeflerin belirlenmesi amacıyla bir çalışma planı ortaya çıkartılır.

4.3.2. Veriyi Anlama Aşaması

Veriyi anlama aşamasında, yapılacak olan çalışma planı doğrultusunda projenin hedefine uygun bir biçimde kullanılacak olan datanın belirlenmesi toplanması ve veri kalitesine yönelik analizlerin yapılması gerekmektedir. Bu aşamada yapılacak olan işlemler Şekil 4.6.'da gösterilmiştir (Meyer ve Cannon, 1998, s.35).

Şekil 4.6. Veriyi anlama safhası

4.3.2.1. İlk Verilerin Toplanması

Verilerin toplanması aşamasında yapılacak olan çalışmaya uygun bir biçimde verileri toplanır. Projede kullanılacak olan araçlar varsa belirlenir. Eğer çok fazla kaynaktan toplanan bir veri var ise tek bir araç kullanarak çalışma yapmak, yapılacak olan çalışmanın kolaylığı açısından daha uygun olacaktır. Kaynakları ile birlikte veriler listelenir. Bu safhada bazı problemler çıkabilir ve bu problemlere uygun çözümler bulunarak sorunlar çözülür.

4.3.2.2. Verinin Açıklanması

Çalışmada kullanılacak olan verilerin özellikleri incelenerek sonuçlarla birlikte raporlanır. Veri türleri ve formatları, veri adetleri ve verilerin bulunduğu tablolar detaylı bir şekilde incelenerek bağlantılı olan ve ihtiyaç duyulabilecek her kriter detaylı bir şekilde belirlenir.

4.3.2.3. Verinin Keşfedilmesi

Verinin keşfedilmesi safhasında belirli rapor ve sorgulama yöntemleri kullanılarak veri madenciliği ile ilgili sorulara erişilir. Bu sorular içerisinde anahtar kelimelerin dağılımı, küçük agresyonların belirlenmesi gibi basit statik analizler olabilir. Yapılan bu analiz çalışması doğrudan veri madenciliği amacına varabilir yada farklı bir veri hazırlanmasındaki adımlara yol gösterici olabilir. Yapılan bu ilk analizlere göre bir ön hipotez hazırlanır.

Veri tabanında bilgi keşfi yolcuğu en bilinen altı modelleme seçeneği ile ön planda yer almaktadır.Bilgi keşfi yolculuğunda en yaygın olarak kullanılmakta olan modelleme seçenekleri Çizelge 4.1'de maddeler halinde sırasıyla gösterilmiştir (Meyer ve Cannon, 1998, s.35).

Çizelge 4.1. Bilgi Keşfi Modelleme Seçenekleri

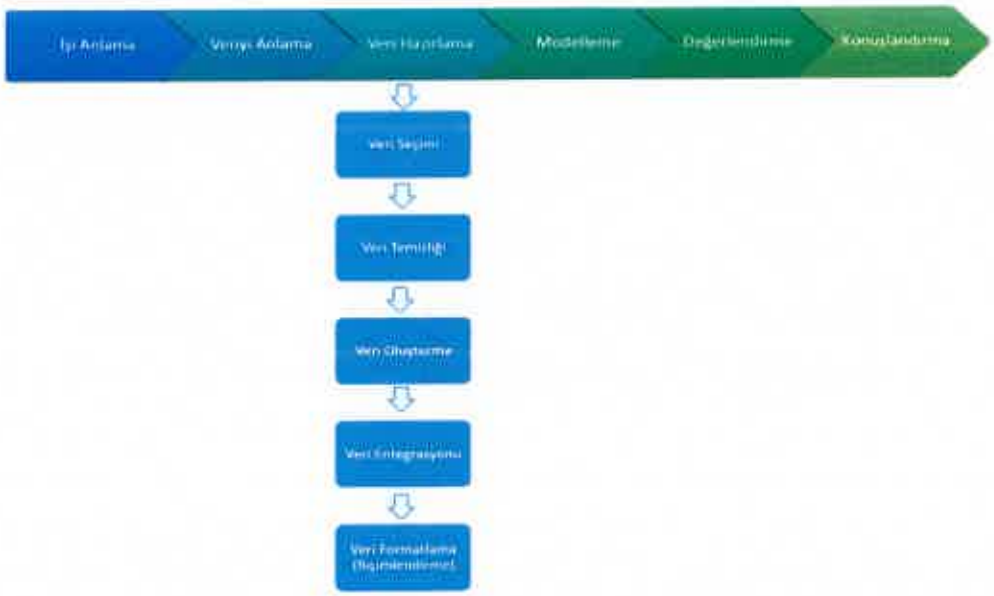
1	Anomali Tespiti (Outlier/Değişim/Sapmalar) – Anomaly Detection
2	Birliktelik Kuralı (Bağımlılık Modellemesi) – Association Rule
3	Kümeleme – Clustring
4	Sınıflandırma – Classification
5	Regresyon – Regression
6	Özetleme – Summarization

4.3.2.4. Veri Kalitesinin Teyit Edilmesi

Verinin kalitesinin teyit edilmesi aşamasında veri kaliyesiyle ilgili sorular analiz edilerek bu sorulara cevap aranır. Örneğin hazırlanılan datanın doğruluk oranı nedir, datanın boş ve dolu kısımları nelerdir bu boş olan kısımlar sonucumuzu ne oranla etkileyebilir, verilerin hatalı olma olasılığı nedir eğer hata var ise bu hataların diğer verilerle benzer özellikleri varmıdır şeklinde analizler yapılarak listelenir.

4.3.3. Veri Hazırlama Aşaması

Model oluşturma sırasında belirebilecek olan her türlü sorun veri hazırlma aşamsına dönülmesine ve sorunların çözümü için dataların baştan yeniden düzenlenmesine sebepiyet verecektir. Model kurulumu esnasında ortaya çıkan sorunların tespit edilerek yeniden datanın düzenlenmesi amacıyla veri hazırlama aşamasına geri dönülmesi gerek modelin kurulması gerek veri keşif aşamasının süresi gerek harcanan efor bakımından negatif yönde çalışmayı etkileyebilmektedir. Veri hazırlama aşamasının adımları Şekil 4.7.'de gösterilmiştir (Meyer ve Cannon, 1998, s.35).



Şekil 4.7. Veri hazırlama safhası

Veri madenciliği çalışmalarında kullanılan verilen herzaman tek bir kaynak üzerinde bulunmayabilir. Çok farklı kaynaklardan çok farklı veriler toplanılarak veri madenciliği çalışmaları yapılabilir. Bu nedenle kullanılan verilerde birbirleriyle uyumsuzluklar meydana gelebilir.Bu uyumsuzluklara örnek vermek istersen

hazırlanan verilerin farklı zaman dilimlerinde hazırlanması, verilerde yapılan güncelleme problemleri ve hataları, kullanılan data formatlarının farklı türde olması, yazılan kodlarda verilerin farklı şekilde tutulması(cinsiyet ile ilgili müşteri verisinin bir databasede E/K olarak tutulurken farklı bir databasede 0/1 şekilden belirtilmesi), ölçü birimlerinin ve varsayımların farklı şekilde belirtilmesidir. Bununla birlikte tüm bu dataların nerede nasıl ve ne tür şartlarda elde edildiğide çalışmanın sağlıklı olabilmesi ve verinin doğruluğu açısından çok büyük önem taşımaktadır. Eğer ki yapacak olduğumuz veri madenciliği çalışmasında güvenilir olmayan bir kaynaktan veri toplayarak bir çalışma yapacak olursak yapacak olduğumuz bütün veri madenciliği çalışmasınında güvenilirliği etkilenmiş olacaktır. Veri hazırlama aşamasındaki adımları şu şekilde inceleyebiliriz.

4.3.3.1. Veri Seçimi

Veri madenciliği çalışmasında kullanılacak olan modellerin kurulması için datasetler belirnir. Veri seçimi aşamasında yapılacak olan projenin analiz edilmesi ve modelde kullanılacak olan datasetler hazrlanır. Analiz çalışmasında kullanılacak olan veriye karar verilir ve hedeflerimiz doğrultusunda kullanılacak olan değerler ve data tipleri incelenir.

4.3.3.2. Veri Temizleme

Veri temizleme aşamasında veri madenciliği çalışmamız için seçmiş olduğumuz veriler için yapılması gereken temizleme kriterleri belirlenir ve bu temizleme sonrasında kaybolacak data tahmini yapılır. Karşılaşabileceğimiz sorunlar için yapmış olduğumuz temizleme çalışmaları tanımlanır.

4.3.3.3. Veri Oluşturma

Veri oluşturma aşamasında veriyi seçtiğimiz çalışmalarda kullanılan formüller değişkenler belirlenir. Topladığımız verileri kullanarak oluşturduğumuz yeni veri değerleri ve setleri veya sonradan oluşturulan veriler belirlenir.

4.3.3.4. Veri Entegrasyonu

Veri ambarlarında bulunan veriler mutlaka bütünleştirilmiş olmalıdır. Farklı veritabanlarından gelen bilgilerde, aynı değeri ifade etmek için farklı semboller/kısaltmalar kullanılabilir. Bu türden farklılıklar yok edilmeli ve veriler

alınmadan önce mutlaka dönüştürme ve standartlaştırma işlemi yapılmalıdır (Meyer ve Cannon, 1998, s.35).

4.3.3.5. Veri Formatlama (Biçimlendirme)

Yapılan veri madenciliği çalışmalasında oluşturulmak istenilen model için bir modelleme aracı kullanmak gerekiyorsa veriye yeni bir boyut verilir. Yapmış olduğumuz çalışmadaki modelde veri tabanının boyutunun çok büyük olması halinde verilerin random olarak seçilmesini bozmayacak şekilde örnekler belirlenmesi daha doğru olacaktır. Bununla birlikte seçmiş olduğumuz örneklerin kullanmış olduğumuz datanın tamamı ile uygun özelliklere sahip olup olmadığıda kontrol edilmesi gerekmektedir. Günümüzde kullanılan paket programlar ve işletim sistemleri her ne kadar gelişmiş ve ileri düzeyde olurlarsa olsunlar büyük veri tabnlarındaki büyük veriler üzerinden bir modelleme çalışması yapılmak istenildiğinde zamanlama kriterlerinin sınırlı olması nedeniyle sorunlar oluşabilmektedir. Bu nedenle tüm veri tabanı üzerinden sınırlı sayıda modelin oluşturulması yerine rastgele oluşturulmuş örnek bir veri tabanından çok sayıda modelin kurularak denenmesi ve bu mdeller içerisinden en güçlü ve en sağlıklı modelin seçilerek kullanılması daha sağlıklı olacaktır. Yani özet olarak kurulan modellerin performansları en uygun karar yöntemi ile test edilmeli ve sonuçlar analiz edilmelidir (Gray ve Watson, 1998, s.67).

4.3.4. Modelleme Aşaması

Veri madeninciliği çalışmalarının modelleme aşamasında birçok farklı modelleme tekniklerinden, elimizdeki veriye en uygun yöntem olan modelleme yöntemi seçilmelidir. Belirlenen sorunların çözümü için en doğru modelin belirlenmesi ancak mümkün olduğunca çok fazla modelleme tekniğinin kurularak denenmesi ile mümkün olacaktır. Bu nedenle veri hazırlama ve model oluşturma süreçleri, en doğru ve en sağlıklı model elde edilene kadar sürekli güncellenen aşamalardır (Gray ve Watson, 1998, s.67).

Geliştirilen örneklerden öğrenme olarak adlandırılan denetimli öğrenmede kullanılacak olan sınıflar ve kriterler bir denetçinin kontrölü ile önceden belirli sınıflara ayrılarak çeşitli sınıflar oluşturulur. Kurulan sistemin gayesi sisteme verilen örneklerden yola çıkarak her bir sınıfa ait özelliklerin belirlenmesi ve bu özelliklerin kurallar şeklinde ifade edilmesi şeklinde olmalıdır.

Modelin öğrenme aşaması tamamlandığında, oluşturduğumuz kural cümleleri sisteme verilen yeni örnekler üzerinde de denenir. Bu sayede kurulan model sayesinde yeni örneklerin hangi sınıfa ait olduğu belirlenmiş olur (Gray ve Watson, 1998, s.67).

Modellerin oluşturulmasındaki denetimsiz öğrenme aşamasında, kümeleme analizine benzer bir yapı olana örneklerin gözlenmesi ve birbiri ile benzer özellik gösterenlerin sınıflandırılması şeklinde bir yapı kurulmaya çalışılır.

Denetimli öğrenmede ise öncelikle veriler kurulan algoritmaya uygun bir biçimde hazırlanır. Sonrasında ise dataların belirli bir kısmı oluşturulan modelin öğrenmesi için kalan kısmı ise modelin uygunluğunun denenmesi için ayrılır. Öğrenme için ayrılan veri kümesi kullanılarak modelin öğrenmesi gerçekleştikten sonra test için ayırdığımız veri kümesi ile oluşturulan modelin doğruluk seviyesi belirlenir. Modelleme aşamasının adımları Şekil 4.8.'de gösterilmiştir (Gray ve Watson, 1998, s.67).

Şekil 4.8. Modelleme safhası

4.3.4.1. Modelleme Tekniğinin Seçilmesi

Bu safhada modelleme yapmk için teknikler belirlenir. Modelleme tekniğinin seçilmesi işlemi iş anlayış safhasında da yapılmış olabilir. Yapılacak olan veri madenciliği çalışmasında kullanılacak modelin dökümantasyon işlemlerinin yapılması gerekmektedir. Her bir modelleme tekniği kendine özgü değerler ve özelliklere sahip olabilir. Örneğin verilerin boş olan kısımları elenerek bunlar verilerden çıkartılabilir.

4.3.4.2.Test Tasarımının Oluşturulması

Yapacak olduğumuz çalışmada kullanılacak modeli oluşturmadan önce, verilerimizdeki hatalı ve eksik kısımları belirlemek adına testler yapmalıyız. Bu testler sayesinde en doğru ve en uygun modelleme yapısı kurulmaya çalışılır (Gray ve Watson, 1998, s.67).

4.3.4.3. Model İnşaa Edilmesi

Bu aşamada veri madenciliği projesi için hazırlamış olduğumuz veriler üzerinden model inşaa edilerek çalıştırılır.

4.3.4.4. Modelin Değerlendirilmesi

Yapmış olduğumuz modelleme çalışmasının sonucunda elde edilen çıktılar incelenerek yaomış olduğumuz veri madenciliği çalışmasının iş hedefleri ile uyumlu olup olmadığı değerlendirilir.

4.3.5. Değerlendirme Aşaması

Kurmuş olduğumuz veri madenciliği modelinin sonuçlarının doğruluklarının kontrol edilmesinde birkaç farklı yöntem mevcuttur. Bunlardan en basit ve en kullanışlı olanı basit geçerlilik testi olarak adlandırılan kontrol sistemidir. Basit geçerlilik testinde modelde kullanılan dataların %5 ve %33 'lük kısımlarında kalan verilerin belirli bir bölümü test etmek amacıyla ayrılır. Daha sonra geriye kalan veriseti üzerinden modelin öğrenmesi için çalışmalar yapılır ve bu kısım üzerinden gerekli test işlemleri gerçekleştirilir (Gray ve Watson, 1998, s.67).

Ancak kullanacak olduğumuz veriler sınırlı bir sayıda ise bu tür durumlarda çapraz geçerlilik testleri ile değerlendirme yapılır. Çapraz geçerlilik testleri ile yapılacak olan değerlendirmede datalar rastgele iki eşit kümeye ayrılır. Öncesinde bir parça üzerinden modelin öğrenmesine yönelik çalışmalar yapılırken kalan diğer parçada ise test adımları uygulanarak değerlendirme yapılır. Daha sonrasında ise diğer ikinci küme üzerinde modelin öğrenme çalışmaları yapılıp kalan kısım üzerinde test işlemleri yapılır. Değerlendirme aşamasının adımları Şekil 4.9.'da gösterilmiştir (Türkmen, b.t.).

Şekil 4.9. Değerlendirme safhası

4.3.5.1. Sonuçların Değerlendirilmesi

Bu safhada veri madenciliği çalışmamızda kullanmak üçin kurulan modelin sonuçları değerlendirilir ve proje sonucunda elde etmek istediğimiz sonuçlara ne kadar yakın değerler elde edildiğinin analizi yapılır. Sahip olduğumuz gerçek veriler ve sonuçlarla karşılaştırma yapılarak doğruluk oranları tespit edilir.

4.3.5.2. Sürecin Gözden Geçirilmesi

Sürecin gösden geçirilmesi safhasında yapmış olduğumuz çalışmaları değerlendirerek yeniden gözden geçirilir. Modelin doğruluğu ve daha sonra yapılacak olan analiz çalışmalarında da kullanılabilirliğinin uygunluğu tespit edilir (Singh, 1998, s.159).

4.3.5.3. Sonraki Adımların Belirlenmesi

Daha önceki süreçler değerlendirilerek çalışmanın geleceği ile ilgili tespitler ve kararlar ortaya konulur. Oluşturulan bu modelleme farklı veri madenciliği çalışmaları içinde kullanılacak mı, modelin kullanım süresine kararlar verilir.

4.3.6. Konuşlandırma Aşaması

Veri madenciliği çalışmamız için oluşturmuş olduğumuz modelleme çalışması doğrudan kendisi bir uygulama olarak kullanılabileceği gibi farklı bir programın içerisinde de kullanılabilir. Kurulan modeller örneğin bir bankada risk analizi, dolandırıcılıkların tespiti, müşterilere verilecek olan krediler için bir değerlendirme mekanizması gibi alanlarda direk olarak kullanılabilir. Bununla

birlikte promosyonların planlanması için simülasyonlarada dahil edilebilir. Konuşlandırma aşamasının adımları Şekil 4.10'da gösterilmiştir (Singh, 1998, s.159).

Şekil 4.10. Konuşlandırma safhası

4.3.6.1. Planın Konuşlandırılması

Bu aşamada proje ve modelin yayınlanması için bir yöntem ve strateji belirlenir. Yapmış olduğumuz çalışmanın adımlarını barındıran bir yöntem hazırlanır.

4.3.6.2. Planın Gözetlenmesi, Bakımı ve Sürdürülmesi

Yapmış olduğumuz çalışma ve kurulan modelin tamamlanması çalışmalarımızın tamamen bittiği anlamına gelmemektedir. Uzun soluklu bir proje yapmak istiyorsak planlamalarımızın gözetlenmesi ve sonrasında devam etmesi amacıyla bakımlarının da yapılması gerekmektedir. Bu aşamada güncel olarak proje ile ilgili konuları takip etmek ve olası sorunlar ve problemler karşısında bakımlarının yapılması gerekmektedir. Bununla birlikte eklenmesi gereken yeniliklerin de takibinin yapılarak çalışmamıza dahil edilmesi gerekebilir (Singh, 1998, s.159).

4.3.6.3. Final Raporunun Oluşturulması

Yapılan çalışma sonrasında çalışmanın amaçlarını ve sonuçlarını içeren bir rapor çıkartılarak müşterilere görsel ve yazılı olarak sunumlar yapılır. Bu sayede müşteriler proje ile yapılması hedeflenen amaçlar doğrultusunda sonuçlara ulaşılıp ulaşılmadığı ve ne kadar faydalı bir çalışma yapıldığı konusunda bilgilendirilmiş olur.

4.3.6.4. Projenin Gözden Geçirilimesi

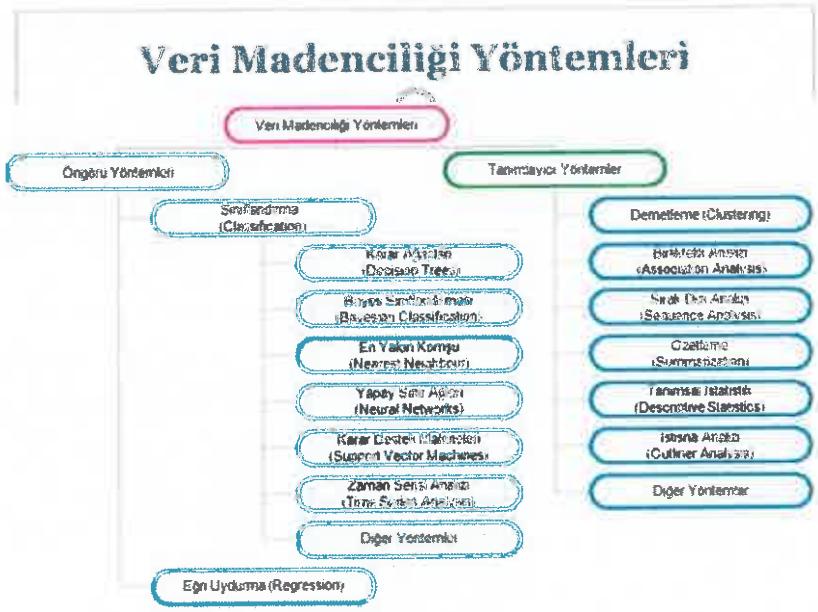
Yapılan bu çalışmalar sonrasından çalışmayı yapan ekip tecrübe kazanmış olacaktır. Daha sonrasında bu projeye benzer projelerinde yapılması ihtiyacı duyulduğunda gerekli bilgi ve tecrübeye sahip olmuş olacaklardır.

İlerleyen sürelerde kullanılan sistemlerdeki değişiklikler ortaya çıkan datalarda değişimlere neden olacağından dolayı kurmuş olduğumuz modelleme çalışmasının sürekşi olarak gözlemlenmesi ve güncellenmesi gereken durumlarda müdahalede bulunarak gerekli değişikliklerin yapılması gerekecektir. Gözlemlenen değişiklikler ile tahminler doğrultusunda elde edilen verilerin arasındaki farkları gösteren farklılık grafikleri kurmuş olduğumuz modelin sonuçlarınun takip edilmesinde bizlere kolaylık sağlayacaktır (Singh, 1998, s.159).

5. VERİ MADENCİLİĞİNDE KULLANILAN YÖNTEMLER

Veri madenciliği çalışmaları için geliştirilmiş olan ve kullanılan bir çok algoritma ve yöntem vardır. Bu yöntemler genellikle istatistiksel tabanlı yöntemlerdir. İşletmeler son dönemlerde kaybettiği müşteri verilerinden yola çıkarak bu müşteriler için anlam ifade eden ortak özellikleri belirlemek bu özelliklerden de yola çıkarak daha sonra kaybedebilecekleri müşterileri belirlemeye çalışırlar. Bu yöntemlere örnek olarak veri madenciliği çalışmalarına aşağıdaki modelleme örnekleri verilebilir.Veri madenciliği yöntemlerine işlevleri açısından bakacak olursak, veri madenciliği yöntemleri iki sınıf altında toplanmaktadır. Bunlar: öngörülü (Predictive Methods)yöntemler ve tanımlayıcı (Desciptive Methods) yöntemlerdir. Şekil 5.1.'de veri madenciliği yöntemleri gösterilmiştir (Meyer ve Cannon, 1998, s. 24-25).

1. Öngörülü – Tahminleyici Modeller (Predictive Methods)
2. Tanımlayıcı Yöntemler (Desciptive Methods)

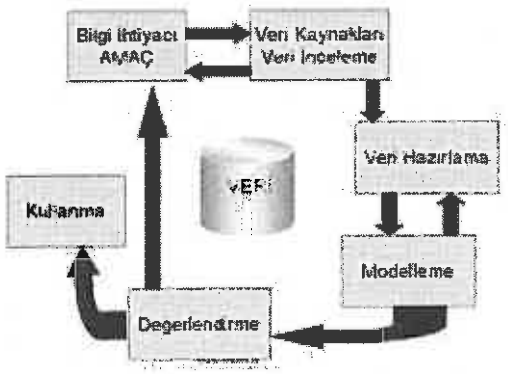


Şekil 5.1. Veri Madenciliği Yöntemleri

5.1. Öngörülü – Tahminleyici Modeller

Tahminleme modellemeleri ile sonuçları bilinen datalardan yola çıkılarak bir model kurulur ve bu model üzerinden sonuçları tahmin edilmeyen yeni data kümelerinin oluşturulmasında kullanılacak tahmin değerlerinin oluşturulması hedeflenmektedir. Örnek verecek olursak bir telekomünikasyon firmasındaki firmanın sahip olduğu müşteriler için faturalarını zamanında ve sürekli olarak ödeyen müşteriler bağımsız değişken olarak nitelendirilirken yeni kazanılan müşterilerin faturalarını zamanında ödeyip ödemeyeceği gibi özellikler bağımlı değişkenler olarak ifade edilir. Bu datalar üzerinden oluşturulacak bir model ile yeni kazanılan bir müşterinin özelliklerine göre faturalarının ödenip ödenmeyeceği tahminlemesi işlemleri yapılabilir.

Öngörülü yöntemlerde veri tabanları sayesinde elde edilen veriler geleceğe yönelik tahminleme işlemlerinde kullanılır. Bu yöntem kullanıcılara ait tüm alan bilgilerine sahip olmasalar bile bu bilgilerin kayır edilmelerine olanak sağlamaktadırlar. Öngörülü yöntemde veriler boş olsa dahi daha önceki verilere bakarak bir tahminleme yöntemiyle boş olan kısımlar doldurulur. Şekil 5.2.'de Öngörülü – Tahminleyici Modellin yapısı gösterilmiştir (Meyer ve Cannon, 1998, s. 24-25).



Şekil 5.2. Öngörülü – Tahminleyici Modeller

Sahip olduğumuz datalar üzerinden bir sonuç elde etmek amacıyla tahminde kullanmak için ortak bir özellik üzerinden formüller ve kriterler oluşturulur. Oluşturulan bu formülleri sahip olduğumuz veriler üzerinde çalıştırarak sonuçlar elde edilir. Tahminleme yapacağımız ödellikler bizim için bağımlı değişkeni ifade ederken bu tahminleme çalışmasının oluşmasında kullanılacak diğer özellikler ise bağımsız değişkenler olarak adlandırılır.Tüm bu tahminleme yöntemleri için formül

oluşturabilmek adına bazı yöntemler kullanılmaktadır. Bu yöntemlerin en önemlileri ise Sınıflandırma (Classification) yöntemleri ile Regresyon-Eğri Uydurma analizi yapmak ve zaman serileridir.

5.1.1. Sınıflandırma

Veri madenciliği çalışmalarında Sınıflandırma (Classification) yöntemi çok fazla kullanılan bir yöntem olmakla birlikte gizli örüntülerin veri tabanlarından elde edilmesinde kullanılmaktadır. Daha basit anlamda ise sınıflandırma yöntemi, veri tabanlarından elde ettğimiz veri setleri üzerindeki tanımlanmış çeşitli sınıflar arasında verilerin paylaştırılması işlemidir. Sınıflandırma teknikleri olarak tanımlanan bu teknik kurulan modele sunmuş olduğumuz eğitim kümeleri sayesinde yapacakları dağılımın şeklini öğrenirler. Bu öğrenmeden sonra sınıflarının belirli olmadığı bir data geldiği zaman daha önce öğrenmiş oldukları sınıflandırma tekniği ile uygun şekilde sınıflandırma işlemini yaparlar.

Bir diğer açıklama olarak ise sınıflandırma teknikleri var olan veritabanı üzerindeki verilerin bir kısmını eğitim amaçlı olarak ayırarak bu veriler üzerinden sınıflandırma kurallarının oluşturulmasını sağlarlar. Daha sonra ise bu kurallar sayesinde sınıfları belirli olmayan bir veri geldiğinde nasıl karar verileceğini ve hareket edilmesi gerektiğine karar verirler. Örneğin Çizelge 5.1'deki gibi, veri tabanı üzerinden elde edilmiş bir veri kümesi olsun.

Çizelge 5.1. Eğitim verileri

Müşteri	Yaş	Boy	Kilo	Cinsiyet
1	20	175	70	Erkek
2	21	179	80	Erkek
3	19	162	50	Kız
4	22	169	55	Kız
5	20	183	90	Erkek
6	19	181	75	Erkek
7	21	171	57	Kız

Bu örnek veriler üzerinden problem şu şekilde tanımlanacak olursa; Çizelge 5.1'deki verilerini kullanarak kilo, boy ve yaş kriterlerine bakarak müşterilerin cinsiyetini bulacak bir sınıflandırma tekniği kurmak istenilebilir. Yani Çizelge 5.1'deki sahip olduğumuz veriler bizim için bir eğitim kümesi olacaktır. Ve bundan sonra gelecek olan yeni müşterileri için yaş, boy ve kilo değerlerine bakılarak

kurmuş olduğumuz sınıflandırma kuralları yöntemi ile müşterilerin cinsiyetleri tahmin edilecektir.

Bir çok sayıda sınıflandırma algoritması mevcuttur. Bu algoritmalardan basit bir yöntemi seçilecek olursa, sınıflandırma algoritması verilen sütundaki değerlerin ortalamasını alacak ve sonrasında bu ortalama değer, öğrenilen değer olacak. Ardından ise yeni gelen test verileri için ortalama değerlerini bulacak ve ortalama uzaklıkları hesaplayarak en yakın olduğu etiketten kabul edilecek. Çizelge 5.1'deki veri kümelerini cinsiyet kriterine göre 2 kümeye ayırarak her bir grup için ortalama değerlerini hesaplayacak (Levene ve Loizou, 2003, s.235-240).

Çizelge 5.2. Erkek müşteriler için öğrenme işlemi sonucu

Müşteri	Yaş	Boy	Kilo	Cinsiyet
1	20	175	70	Erkek
2	21	179	80	Erkek
5	20	183	90	Erkek
6	19	181	75	Erkek
Ortalama	20	179,5	78,75	

Aynı sınıflandırma yöntemi kız müşterileri üzerinden de uygulanabilir. Çizelge 5.3'te bu sınıflandırma yönteminin sonuçları gösterilmiştir.

Çizelge 5.3. Kız müşteriler için öğrenme işlemi sonucu

Müşteri	Yaş	Boy	Kilo	Cinsiyet
3	19	162	50	Kız
4	22	169	55	Kız
7	21	171	57	Kız
Ortalama	20,66667	167,3333	54	

Nihai olarak kurulan algoritma erkeklerde (20, 179.5, 78.5) değerlerini öğrenirken kız müşteriler için öğrenme değerlerimiz (20.66, 167.33, 54) olarak belirlenmiş olmaktadır. Ve bu aşamadan sonra gelecek olan yeni test verimizde yaş değeri 21 boy değeri 165 ve son olarak kilo değeri 60 olduğunu ele alalım. Algoritmamızın daha önceki test verileinden öğrenmiş olduğu değerlere göre yeni gelen müşteri verilerini değerlendirerek müşterin,n cinsiyetini thmin etmeye çalışalım. Bu hesaplamada da bir çok hesaplama algoritması bulunmasına karşın temel seviyede hesaplama ypmak adına öklit mesafesi (euclidean distance) yöntemini kullanarak her bir değere olan mesafeyi hesaplmaya çalışılır (Levene ve Loizou, 2003, s.235-240).

Kurulan algoritmanın erkek müşteriler için tanımlanan öğrenme değerleri ile mesafesinin hesaplanması ile alttaki sonuç elde edilir. Denklem (5.1)'de erkek (5.2)'de kız nitelikler için öklit uzaklık hesaplanmıştır (Gasson, ve diğ.2005).

$$\text{ÖklitUzaklık(Erkek)}=\sqrt{((20-21)^2+(179.5-165)^2+(78.75-60)^2)}=23.72 \quad (5.1)$$

Benzer işlemleri cinsiyeti kız olarak tanımlanan müşteri verilerinden öğrendilen değerler için yapılacak olursa alttaki sonuca ulaşılır.

$$\text{Öklit Uzaklık(Kız)}=\sqrt{((20.66-21)^2+(167.33-165)^2+(54-60)^2)}=6.44 \quad (5.2)$$

Bu hesaplamalar sonucunda algoritma erkeklere olan mesafenin 23.72 sonucunu ve kızlara olan mesafenin ise 6.44 olduğunu göstermektedir. Bu sonuçlardan yola çıkarak yeni gelen müşterinin cinsiyetinin kız olduğu ortaya çıkmaktadır. Sınıflandırma algoritmalarının yapısı yapılan bu örnektede olduğu gibi en basit yöntemiyle şu 2 basit yapı ile oluşturulur (Levene ve Loizou, 2003, s.235-240).

- Eğitim verisi üzerinden öğrenme
- Öğrenilen değerlerle test verisi üzerinde sınıflandırma

Sınıflandırma (Classification) yönteminde kullanılan teknikler işe şu şekildedir;

- Karar Ağaçları (Desicion Trees)
- Bayes Sınıflandırması (Bayesian Classification)
- K-En Yakın Komşu (Nearest Neighbour)
- Yapay Sinir Ağları (Neural Networks)
- Karar Destek Makineleri (Support Vector Machines)
- Zaman Serisi Analizi (Time Series Analysis)

5.1.1.1. Karar Ağaçları

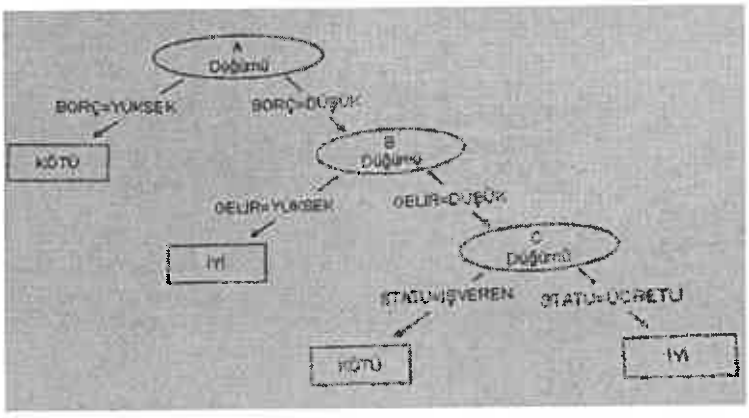
Karar ağaçları ile sınıflandırma yönteminin en çok tercih edilme yöntemlerinin başlıcaları ucuz olması, veritabaları ile uyumlu olarak çalışabilmesi ve yorumlanabilirliğinin kolay olmasıdır. Kara ağaçlarının yapısı düğüm ve dallardan oluşmakta ve bu yapı anlaşılabilirliğini kolaylaştırmaktadır. Karar agacının yapısındaki her bir dalın bir olasılığı vardır. Karar agacının bu yapıda olmasının istediğimiz köke veya son dallardan köke ulaşabilecek bir hesaplama yapılmasına imkan sağlmaktadır (Levene ve Loizou, 2003, s.235-240).

Örnek:

Çizelge 5.4. Eğitim verileri

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	YÜKSEK	YÜKSEK	İŞVEREN	KÖTÜ
2	YÜKSEK	YÜKSEK	ÜCRETLİ	KÖTÜ
3	YÜKSEK	DÜŞÜK	ÜCRETLİ	KÖTÜ
4	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
5	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
6	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ
7	DÜŞÜK	YÜKSEK	ÜCRETLİ	İYİ
8	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
9	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
10	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ

Bir bankadan kredi kullanan müşterilerine ait risk durumlarını karar ağacı tekniğiyle bulmak istediğini düşünelim. Bu teknik sayesinde belirlediğimiz özellikler sayesinde yeni gelecek olan kredi taleplerinin karar ağacı bilgilerine dayanılarak kredi verilip verilmeyeceğine karar verilecektir. Çizelge 5.4'te gösterilen verileri karar agacı modeli ile eğitim datası olarak kullanılacak verilerdir. Bu verileri kullanarak karar ağaçlarını oluşturmak amacıyla veri madenciliğinde birçok yöntem ve algoritma mevcuttur. Örnek verilecek olursa C4.5 algoritması yöntemiyle oluşturulacabilecek yöntem Şekil 5.3.'te gösterilmiştir (Bozkır v.d.,2018).



Şekil 5.3. Eğitim verilerine uygun karar ağacı

Şekil 5.3.'te kurulan karar ağacı yöntemiyle karar kuralları oluşturulmuştur. Bu kurallar yöntemiyle kredi talebinde bulunan müşterilere kredi verilip

verilmeyeceğine karar verilebilir. Bu karar ağacını kurallarından yararlanılarak aşağıdaki kurallar oluşturmuştur.

KURAL 1:

Eğer BORÇ=YÜKSEK **ise** RİSK=KÖTÜ;

KURAL 2:

Eğer BORÇ=DÜŞÜK **ise ve**

Eğer GELİR=YÜKSEK **ise** RİSK=İYİ;

KURAL 3:

Eğer BORÇ=DÜŞÜK **ise ve**

Eğer GELİR=DÜŞÜK **ise ve**

Eğer STATÜ=İŞVEREN **ise** RİSK=KÖTÜ;

KURAL 4:

Eğer BORÇ=DÜŞÜK **ise ve**

Eğer GELİR=DÜŞÜK **ise ve**

Eğer STATÜ=ÜCRETLİ **ise** RİSK=İYİ;

Bu eğitim verilerinden elde edilen kurallar kullanılarak yeni gelen müşteri kredi taleplerine yönelik risk durumları değerlendirip kredi verilip verilemyeceğine karar verilebilir.

5.1.1.2. Bayes Sınıflandırması

Bayes sınıflandırma olarak adlandırılan bu sınıflandırma yöntemi ismini Matematikçi Thomas Bayes'den almış bir sınıflandırma algoritmasıdır. Bayes sınıflandırması olasılık ilkelerinden yola çıkılarak tanuımlanmış hesaplama kuralları ile sisteme sunmuş olduğumuz verilerin kategori ve sınıflarını belirlemeyi amaçlayan bir yöntemdir (Data Mining, Anonim, b.t.).

Bayes sınıflandırma yönteminde sisteme öncelikle belirli bir miktarda öğretilmiş veri sunulur (Örneğin 500 adet). Ancak sisteme sunulan bu verilerin bir sınıfı, kategorisi bulunmak zorundadır. Sisteme sunulan bu öğretilmiş veriler ile yapılan olasılık işlemleri kullanılarak, sisteme sunulan veriler daha önce test edilerek elde edilen olasılık değerleri doğrultusunda işletilir ve verilen test verilerin hngi sınıfa dahil olması gerektiğine karar verilmeye çalışılır. Sisteme sunulan öğretilmiş veri sayısı ne kadar çok ise test verisinin doğru kategorisinin tespit edilme oranıda o kadar yüksek olacaktır.

Bayes sınıflandırma yönteminin kullanıldığı birçok alan mevcuttur. Ancak burada hangi verilerin sınıflandırıldığından ziyade nasıl sınıflandırıldığı önem taşımaktadır. Örneğin sisteme sunduğumuz ve öğretim işlemi yapmak istediğimiz verileri formatları string veya integer formatta olabilirler ve burada asıl önemli olan verilerin türlerinden ve nasıl veriler olduğundan ziyade bu dataların birbirleriyle nasıl ve hangi oranda ilişki kurulduğu önem kazanmaktadır (Data Mining, Anonim, b.t.).

Örneğin Çizelge 5.5'te verilen değerler bayes yöntemi ile sınıflandırma yapmak için eğitim verilerimiz olsun. Bu verilerde sadece yazılım ve muhasebe departmanlarında çalışan çalışanlara ait yaş, maaş ve iş tecrübesine ait veriler mevcutur. Bu verilerden yola çıkarak sınıflandırma yapmak istenildiğinde ve yeni bir çalışan verisi modele sunulduğunda hangi departmanda çalıştığının tahmin edilme işlemi yapılacaktır.

Çizelge 5.5. Eğitim verileri

DEPARTMAN	MAAŞ	YAŞ	İŞ TECRÜBESİ
Yazılım	3000	26	4
Muhasebe	1500	22	2
Yazılım	5000	30	9
Muhasebe	2000	30	7
Muhasebe	500	18	3
Yazılım	2000	20	2
Yazılım	7000	29	5
Muhasebe	6000	45	15

Çizelge 5.5'te verilen değerler doğrultusunda; Maaş: 3000, Yaş : 30, Tecrübe: 5 Yıl kriterlerine sahip bir kişinin hangi departmanda çalıştığı bulunacaktır.

Bu bilgiler doğrultusunda öncelikle eğitim modelinin kurulması gerekmektedir. Eğitim modeli üzerinden ise test işlemleri yapılacaktır. Varyans ve ortalama değerlerini departman bilgileri üzerinden hesaplayacak olursak Çizelge 5.6'daki değerler elde edilmiştir (Data Mining, Anonim, b.t.).

Çizelge 5.6. Eğitim verileri

DEPARTMAN	MAAŞ	YAŞ	İŞ TECRÜBESİ
Muhasebe	2500	28.75	6,75
Yazılım	4250	26.25	5
Muhasebe	5833333,33	142.25	34,91666667
Yazılım	4916666,67	20.25	8,666666667

Çizelge 5.6 da verilen değerler için ilk satır verileri ortalama değerleri ifade ederken ikinci satır verileri ise varyans değerleri ifade eder. Bu değerler kullanılarak beklenen değerlerin hesaplanması işlemi yapılacaktır. Modele sunulan yeni test verisinin Yazılım derpartmanı mı yoksa Muhasebe departmanına mı ait olduğunu gösteren beklenen durum değerlerini hesaplayacağız. Bu hesaplama da ise Bayes sınoflandırma yöntemini kullanıyor olacağız. Bayes sınıflandırma yöntemi özet olarak tüm koşullar üzerinden olasılıkların çarpımı ile hesaplanır. Bayes sınıflandırmada kullanılan formüller (5.3), (5.4) ve (5.5)' te gösterilmiştir (Gasson, ve diğ.2005).

$$sınıflandırma(S_1, S_2 \dots, S_n) = azami_c p(K = k) \prod_{i=1}^{n} p\,(S_i = s_i | K = k) \qquad (5.3)$$

Yukarıdaki formülde görüldüğü gibi sınıflar arasında bir seçim yapılmak istenildiğinde bu sınıflara ait olasılık değerleri ve k koşulları için çarpımlarından farkları olmamaktadır. Her sınıf için bir koşul olasılğı değeri mevcuttur ve bu olasılıklar yardımıyla test verisinin hangi kategoriye ait olduğu bulunmaya çalışılır.

$$beklenti(Yazılım) = \frac{P(Yazılım)\; p(maaş|yazılım)\; p(Yaş|yazılım)\; p(tecrübe|yazılım)}{normalleştirme} \qquad (5.4)$$

Test verisindeki kişinin yazılım departmanı kategorisine ait olduğunu bulmak için öncelikle yazılım departmanında olan kişilere ait oranlar hesaplanır. Bu departmandaki kişiler için maaş yaş ve tecrübe kriterlerine göre olasılıklar hesaplanarak normalleştirilmiş değerlerine bölünür. Aynı hesaplama yöntemi muhasebe departmanındaki kişiler içinde altta belirtilen formülde olduğu gibi hesaplanır.

$$beklenti(Muhasebe) = \frac{P(Muhasebe)\; p(maaş|Muhasebe)\; p(Yaş|Muhasebe)\; p(tecrübe|Muhasebe)}{normalleştirme} \qquad (5.5)$$

Ancak bu hesaplamanın farkı muhasebe departmanı kategorisine ait olan kişiler için olasılığının koşullu olarak alınması gerekir. Model de bulunan tüm ihtimallerin normalleştirilmesi ise normalleştirme değeri ile yapılır. Aşağıdaki örnek üzerinden basit şekilde olasılık hesabı yapılacak olursa:

P (Yazılım) =0.5 (8 kişiden 4'ü yazılım kısmında)

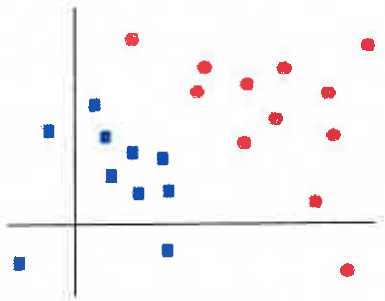
Benzer şekilde,

P (Muhasebe) = 0.5 olarak bulunur.

5.1.1.3. K-En Yakın Komşu

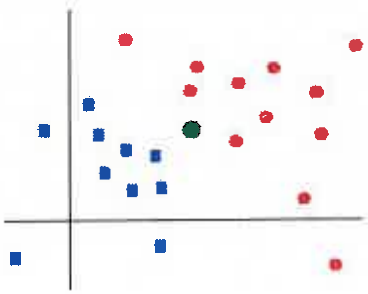
Regresyon ve sınıflandırma yöntemlerinden bir yöntem olan En yakın komşu algoritması çok yaygın olarak kullanılan bir sınıflandırma yöntemidir. Sınıflandırma yöntemiyle elde edilen daha önceki benzerliklerden yola çıkılarak daha sonra verilen test verilerinin bu özelliklerden hangisine en yakın olduğuna bakılması yöntemidir.

Örnek verilecek olursa k=5 olarak kabul gördüğü durumda yeni bir değerin sınıflandırma işlemi yapılarak hangi gruba daha yakın olduğu bulunmaya çalışılır. Sınıflandırma işlemi yapılan örneklerden 3 adet örnek alınarak yeni verinin bunlardan hangisine yakın olduğu hesaplanmaya çalışılır. Bu hesaplamada en çok kullanılan teknik ise öklit yöntemiyle mesafe hesabının yapılmasıdır (Veri Madenciliği, Anonim, b.t.).



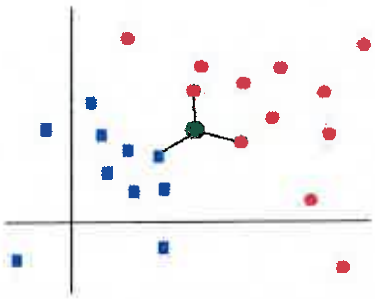
Şekil 5.4. K-En Yakın Komşu Eğitim verileri

Şekil 5.4'te gösterilen koordinat düzleminde örnekler 2 boyutlu düzlem üzerinde gösterilmiştir. Bu örnek üzerinden yola çıkılarak en yakın komşu algoritması yöntemiyle sınıflandırma yapıldığında Şekil 5.5'te olduğu gibi yeni bir üyenin diğer üyelere olan mesafesi hesaplanır (Veri Madenciliği, Anonim, b.t.).



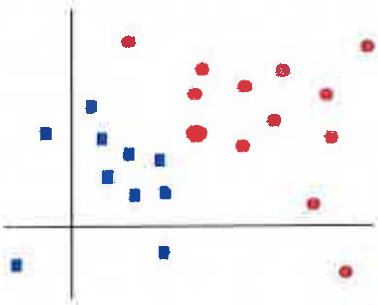
Şekil 5.5. K-En yakın komşu algoritması yeni değer

Şekil 5.5.'te yeşil renk ile gösterilen yeni üyenin diğer üyelere olan yakınlık mesafelerinin hesapanması yapıldığında Şekil 5.6.'da olduğu gibi bir görüntü elde edilir (Veri Madenciliği, Anonim, b.t.).



Şekil 5.6. K-En yakın komşu yeni elemanın diğer elemanlara olan mesafesinin gösterimi

Bu mesafelerin hesaplanmasına bakılacak olursa yeni değerin 2 adet kırmızı değere daha yakın mesafede olduğu görünmektedir. Bu sonuç doğrultusunda yeni değerin yuvarlak ve kırmızı değerler aynı sınıfa dahil edildiği görünmktedir (Veri Madenciliği, Anonim, b.t.).



Şekil 5.7. Yeni değerin k-nn ile sınıflandırılmış gösterimi

Tez çalışmasının uygulama kısmında kullanılanacak olan algortimada bir en yakın komşu (K-NN) yöntemi algoritmasıdır (Veri Madenciliği, Anonim, b.t.).

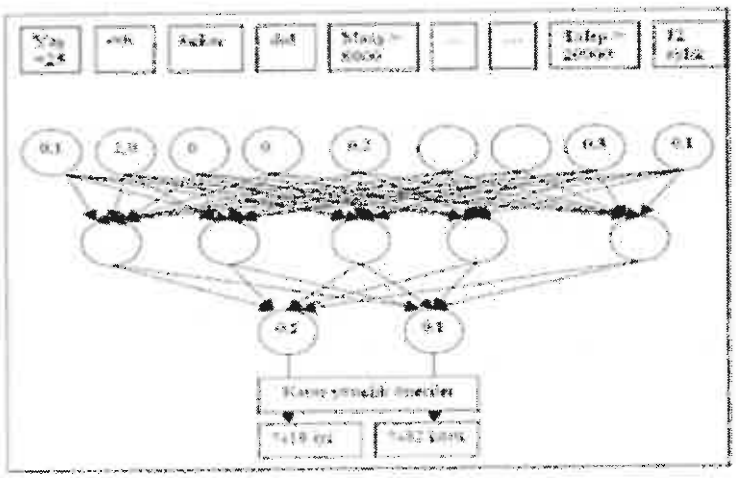
5.1.1.4. Yapay Sinir Ağları (Neural Networks)

Gelişen bilgisayar teknolojileri sayesinde günümüzde insanlar ihtiyaç duydukları işlemlerinin büyük bir kısmını teknoloji sayesinde yapmaktadırlar. 1980'li yıllarda ortaya çıkan makinelerin insanlar gibi düşünerek hareket etmelerinin sağlanması fikri ile birlikte yapay sinir ağları terimi ortaya çıkmıştır. 1990'lı yıllardan itibaren ise yapay sinir ağları oluşumu giderek önem kazanmış ve teknolojinin gelişmesinde çok büyük yer edinmiştir.

Yapay sinir ağları (YSA) çalışma presnsibi olarak insan beyninin sinir sistemine olan etkisini kendine çalışma yöntemi olarak hedefleyen bir modeldir. Bir nevi insan beyninin aynı özelliklerde çalışan bir kopyası şeklindedir. Nasılsa insan beyni yeni bilgileri öğrenme yoluyla elde ediyor ve edinmiş olduğu bu bilgiler sayesinde düşünme karar verme işlemlerini yardımcı faktörler olmadan yapabiliyorsa yapay sinir ağlarıda aynı yapıda sistemler geliştirmek amacıyla yapılmıştır. Yapay sinir ağları hesaplamalarında adım adım çalışan bir yönteme ihtiyaç yoktur. Yapay sinir ağları çalışma mantığı olarak kendi iç kurallarını kendi üreten ve üretmiş olduğu bu kuralları kullanarak sonuçlar elde ederek örnekler üzerinden karşılaştırma yaparak düzenlenir. Yapay sinir ağları modelleri deneme yanılma yöntemi ile işin en doğru şekilde yapılması şeklinde çalışır. Bilgilerin saklanması işlemi ise eğitim özelliklerinden elde edilen sonuçlar üzerinden sağlanır.

Yapay sinir ağları yapı olarak bir amaca yönelik model oluşturulur ve tıpkı insanlarda olduğu gibi yeni örnekler sayesinde öğrenme yeteneğini geliştirir. Tıpkı tüm canlılarda olduğu gibi sinir sistemlerinin adapta olabilme mantığı ile çalışmayı hedefleyen bir yapıdadırlar.

Şekil 5.8.'de yapay sinir ağları uygulaması ile bir bankanın sahip olduğu müşterileri için bir risk analizi ve tahminlemesi yapmayı hedefleyen bir örnek verilmiştir. Bankalar müşterilerinin sadece gelir düzeyi, yaş bilgileri medeni durumları gibi bilgilere değil bunların dışında boçlarıyla ilgili daha detaylı bilgilerede sahiptirler. Bankalar sahip oldukları tüm bu verileri kullanarak müşteri risk seviyesini belirlemeye yönelik bir modelleme çalışması yapabilirler (Veri Madenciliği, Anonim, b.t.).



Şekil 5.8. Yapay sinir ağları uygulaması

5.1.1.5. Karar Destek Makineleri (Support Vector Machines)

Vladimir Vapnik ve Alexey Chervonenkis ikilisi ile birlikte 1963 yılında ortaya çıkan ve Destek Vektör Makineleri (DVM) olarak adlandırılan istatistiksel hesaplamalar ile öğrenimini yapan bir gözetimli öğrenme algoritmasıdır. Bu yöntem temelde 1960 lı yıllarda ortaya çıkmış olsada asıl geliştirme islemleri 1963'lü yıllarda Berhard Boser, Vladir Vapnik ve Isabella Guyon'un çalışmaları ile gün yüzene çıkmış ve önem kazanmıştır (Veri Madenciliği, Anonim, b.t.).

Bu tekniğin temelinde ise iki farklı sınıfa ait olan dataları en uygun yöntem ile birbirlerinden ayırma çalışmalarında kullanılmaktadır. Bu ayrım için ise karar sınırlamalarına ihtiyaç durulmaktadır (Veri Madenciliği, Anonim, b.t.).

Günümüzde ise plaka tanımlama sistemleri, yüz tanım sistemleri, parmak izi okuma sistemleri ve seslerin analiz edilmesi gibi bir çok önemli çalışma ve uygulamalarda kısa adıyla DVM olarak ifade edilen Destek Vektör Makineleri ile yapılan sınıflandırma algoritmaları kullanılmaktadır. Bu yöntemin avantajlarından kısaca bahsedecek olursak;

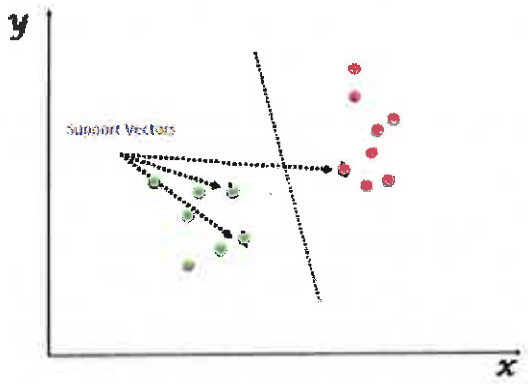
- Yüksek boyutlu uzaylarda etkilidirler (Veri Madenciliği, Anonim, b.t.).
- Boyut sayısının, örnemlem sayısından fazla olduğu durumlarda etkilidirler.
- Karar fonksiyonunda bir takım eğitim noktaları kullanılır ("support vectors"). Dolayısıyla bellek verimli bir şekilde kullanılmış olur.
- Çok yönlü: Karar fonksiyonu için çok farklı çekirdek fonksiyonları ("kernel functions") kullanılabilmektedir (Veri Madenciliği, Anonim, b.t.).

Destek vektör makineleri ikiye ayrılmıştır.

Doğrusal Destek Vektör Makineleri

Bu yöntem ile sınıflandırmalara örnek verecek olursak iki farklı kümeye ait örneklerin doğrusal bir şekilde ayrılmış olduğunu düşünecek olursak bu iki örnek kümesinin bir eğitim verisi sayesinde elde edilmiş olan bir karar mekanizması ile birbirlerinden ayrıştırmaları hedeflenmektedir.

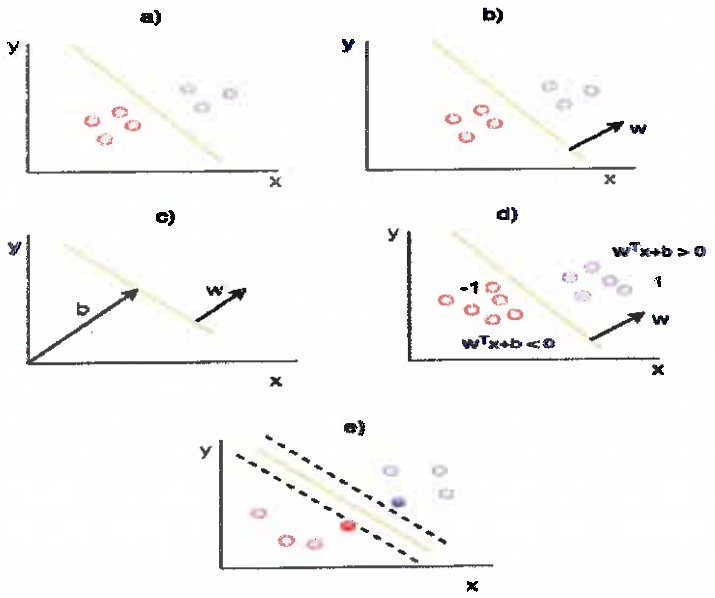
Şekil 5.9'da gösterildiği gibi veri kümesini ikiye bölen çizgi ise karar doğrusu olarak adlandırılmaktadır. Bu yöntemde sonsuz adet karar doğrusu çizebilme seçeneğimiz olsa bile asıl önemli olan en uygun, en doğru karar doğrusunu yani optimali bulabilmektir (Veri Madenciliği, Anonim, b.t.).



Şekil 5.9. Destek Vektörleri

Bu yöntemde yapılan sınıflandırma işlemlerinde sınıf etiketleri tanımlanırken genel olarak (+1, -1) şeklinde tanımlanmaktadır.

Sınır çizgisinin 2 kümeninde sınır çizgililerine en yakın uzaklıkta belirlenmesi yeni gelecek olan veriye karşı karar doğrusunun dayanıklı olmasını sağlayacaktır. Buradaki sınır çizgilerine en yakın olarak belirlenen noklarar ise destek noktaları olarak isimlendirilmektedir. Şekil 5.10.'da ise bu yönteme örnek bir yapı verilmiştir (Veri Madenciliği, Anonim, b.t.).



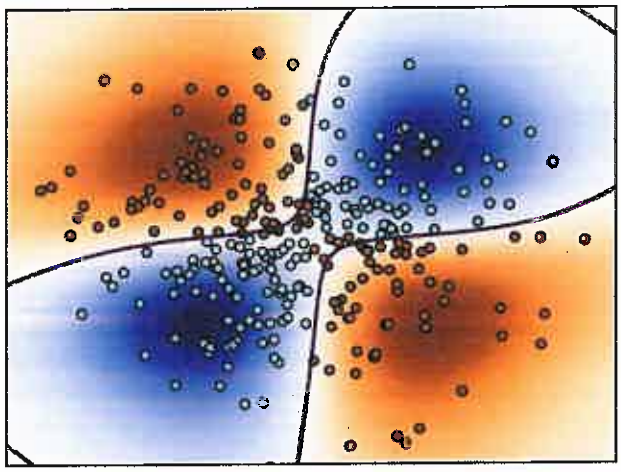
Şekil 5.10. Doğrusal Destek Vektör Makineleri

Doğrusal Olmayan Destek Vektör Makineleri

Doğrusal olmayan destek vektör makinelerinde ise veri kümelerinde doğrusal bir ayrım çizgisi çizilememektedir. Bu yüzden doğrusal olmayan destek vektör makinelerinde kernel trick olarak isimlendirilen bir çekirdek numarası kullanılır.

Çekirdek olarak adlandırılan bu yöntemin makine öğrenimlerinde kullanılması sayesinde modelin doğru bir şekilde çalışarak doğrusal olmayan dataların sınıflandırma yetenekleri arttırılmıştır. Çekirdek yönteminin en çok kullanıldığı yöntemlere örnek verecek olursak;

- Polynomial Kernel
- Gaussian RBF (Radial Basis Function) Kernel



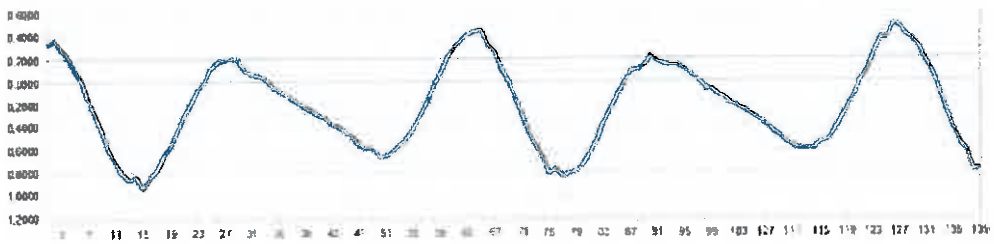
Şekil 5.11. Doğrusal olmayan Destek Vektör Makineleri

Şekil 5.11'de ise doğrusal olmayan destek vektör makinelerine ait bir örnek verilmiştir (Veri Madenciliği, Anonim, b.t.).

5.1.1.6. Zaman Serisi Analizi

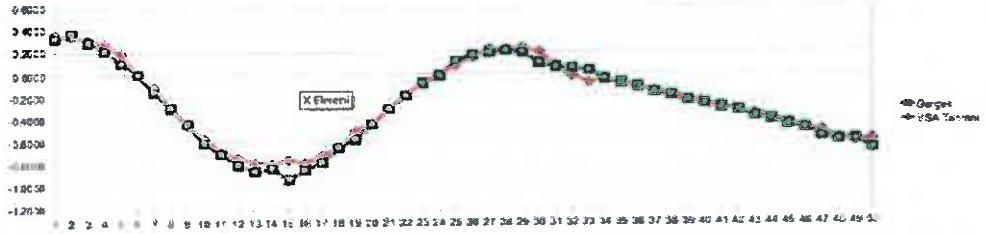
Zaman serisi yöntemiyle analizlerde zamana bağlı olan değişkenler tahmin edilmeye çalışılmaktadır. Örnek verecek olursak bir yaz döneminde ki dondurma satışlarının oranı bir önceki yaz döneminde satılan dondurma adetleri ile belirlenmeye çalışılır (Veri Madenciliği, Anonim, b.t.).

Zaman serileri ile yapılan analizlerde biir değer kendisinden önce gelen değerler sayesinde tahmin edildiğinden önceki değerle ile ilişkilidirler.

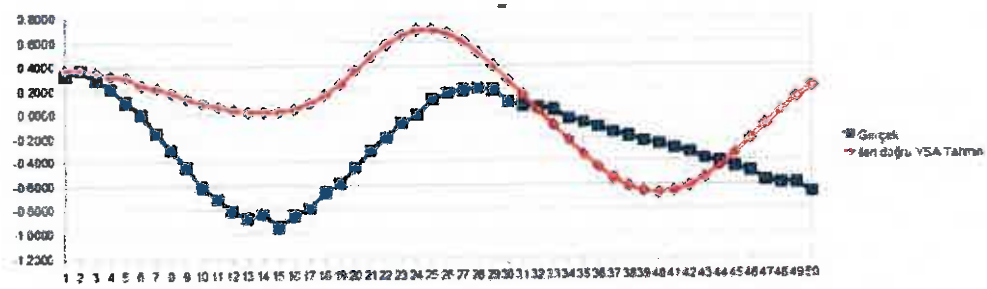


Şekil 5.12. Zaman Serisi

Şekil 5.12'deki örnek zaman serisi üzerinden çıkışı 1 ve girişi 9 olan bir model kuracak olursak Şekil 5.13'teki tahmin edilmiş grafiği elde ederiz (Meyer ve Cannon, 1998, s. 24-25).



Şekil 5.13. İleri yönde hazırlanmış zaman serisi



Şekil 5.14. İleri yönde hazırlanmış zaman serisi

5.2. Tanımlayıcı Yöntemler

Tanımlayıcı yöntemler tahmini yöntemlerdeki gibi veri desenleri üzerinden yeni veriler bulmaktan ziyade daha çok verilerdeki desenleri bulmaya yöneliktir. Karar vermeye yardımcı olacak ve mevcut datalardaki örüntülerin tanımlanması sağlanmaktadır. Tanımlayıcı yöntemler olarak adlandırılan bu yöntem bir kümeleme ve birliktelik tanımıdır.

Müşterilerin sahip oldukları ortak özellikler sayesinde kümeleme işlemlerinin yapılması yada marketlerdeki hangi ürünlerin birlikte satıldığını bulmak amacıyla yapılan modelleme çalışmaları tanımlayıcı yöntemler ile yapılan örneklerdir.

Tanımlayıcı yöntemlerde kullanılan teknikler iki gruba ayrılmaktadır. Bu teknikler şu şekildedir;

- Kümeleme Analizi
- İlişki Analizi (Birliktelik Kuralları ve Ardışık Örüntüler)

6. KULLANILAN TEKNOLOJİLER

Günümüzde artan teknolojiler sayesinde artan bilgilerin ilkel koşullarda saklanması yerine veritabanlarında saklanması yöntemlerine başvurulmaktadır. Bu aşamada çalışmada kullanılan teknolojiler anlatılmıştır. Bu teknolojiler şu şekildedir;

6.1. MS SQL (Microsoft Structured Query Language)

Veritabanları çok fazla sayıda verilerin saklandığı ve depolandığı yapılardır. Bir diğer tanımla birbirleriyle ilişki halinde olan verilerin çeşitli amaçlar doğrultusunda depolandığı ve ihtiyaç duyulduğunda kolayca erişme imkanı sunan yapılar veri tabanı olarak tanımlanmaktadır.

Veri tabanları milyonlarca müşteriye sahip olan bir bankanın müşteri verilerini kolayca saklayabildiği, yada bir hastanenin hastalarına ait verilerini depoladığı alanlardır.

SQL (Structured Query Language) Türkçe karşılığı yapılandırılmış sorgulama dili olarak tanımlanmaktadır. Sql için bilinmesi gereken en önemli madde bir programlama dili değil bir veri tabanı olduğu ve gerektiğinde sorgulama metodları ile istenilen verileri kolayca erişim imkanı sunan bir sorgulama dili olduğudur (Veri Madenciliği, Anonim, b.t.).

SQL dili ile sorgulama imklanı sunan yazılımların bazıları şu şekildedir.

- Ms Sql (Microsoft Structured Query Language)
- Oracle
- MySql
- Acces

Yazırlamış olduğumuz tez çalışmamızda sunucu tabanlı Microsoft firması tarafından geliştirilen bir yazılım olan Ms Sql kullanılmıştır.

6.2. SAP IDT

SAP (System Analysis and Program) firması tarafından geliştirilen ve bilgi tasarım aracı olarak bilinen Information Desing Tool (IDT) birçok farklı veri kaynağından farklı verileri ilişkisel bağlantı yöntemleri kullanarak tek bir yapı altında toplamak için kullanılan bir tasarım aracıdır.

Univere olarak adlandırılan yapılar kullanıcıların birçok veriyi analiz etmelerine imlan sağlayan mantıksal boyut ve nesneler koleksiyonu olarak adlandırılırlar. Nesneler ve boyutlar farklı hiyerarşi ve hesaplamaları, özel hesaplamaları ve özellikleri temsil etmektedirler.

Universe, Sql Server veya Oracle gibi ilişkisel veritabanlarından beslenerek birçok verilerin tekbir çatı altında toplanmasına imkan sağlamaktadır. Farklı veri kaynaklarında bulunan verilere bağlantılar kurularak tek bir merkezden erişim imkanı sağlanır (Veri Madenciliği, Anonim, b.t.).

Information Desing Tool kullanılarak Universe tasarlanır ve sonrasında farklı raporlama araçlarından universe erişimi sağlanılarak raporlamalara imkan sağlanır.

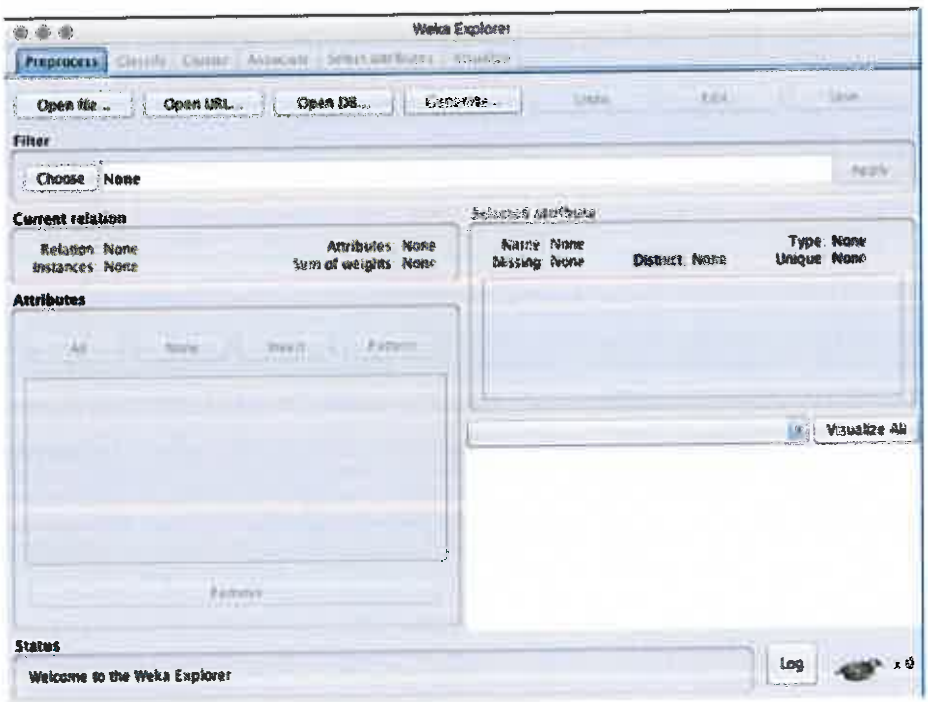
Yapmış olduğumuz tez çalışmasında müşteri verileri Sql veritabanından tutulmakta olup bu verilere erişip rapolama işlemlerini yapmak ve gerekli verilerin elde edilmesi amacıyla information desing tool aracı kullanılarak universe oluşturulmuş ve veriler bu yapı üzerinden elde edilmiştir. Şekil 6.1'de SAP Information Desing Tool aracı gösterilmiştir.

Şekil 6.1. SAP İnformation Desing Tool

6.3. Weka Explorer

Temelde makine öğrenmesi algoritmalarının ve veri ön işleme (data pre-processing) gibi gereksinimlerin bir arada sunulduğu, Waikato Üniversitesi tarafından açık kaynak olarak dağıtılan ve Java ile geliştirilen bir veri madenciliği programıdır (Veri Madenciliği, Anonim, b.t.).

Makine öğrenmesi işlemlerinde, dahil olduğum proje ve akademik araştırmalarda sıklıkla adı geçen uygulamalardan bir kaçından bahsetmek istiyorum. Bu uygulamalardan ilki Weka. Temelde makine öğrenmesi algoritmalarının ve veri ön işleme (data pre-processing) gibi gereksinimlerin bir arada sunulduğu Waikato Üniversitesi tarafından açık kaynak olarak dağıtılan ve Java ile geliştirilen bir veri madenciliği programıdır. Weka yazılımı dosya uzantısı olarak "ARFF" (Attribute Relationship File Format) formatını kullanır. Şekil 6.2'de Weka Explorer programının ekran görüntüsü gösterilmiştir.



Şekil 6.2. Weka Explorer Ekran Görüntüsü

6.4. Python

Python, ilk sürümü Guido van Rossum tarafından 1991'de ortaya konulmuş genel amaçlı bir programlama dilidir. Yorumlanan ve dinamik bir dil olan Python, esas olarak nesne tabanlı programlama yaklaşımlarını ve belli bir oranda da

fonksiyonel programlamayı desteklemektedir. Python günümüzde görece kolaylığı ile birlikte sahip olduğu geniş ölçekli standart kütüphanesi sayesinde çok fazla tercih edilen ve popüler hale gelen, çok büyük kurumlarında arasında olduğu geniş bir kullanıcı kitlesine sahip olan bir yapı haline gelmiştir.

Python, mühendislikten-finansa kadar birçok alanda kullanılmaktadır. 2000'li yıllardan itibaren bilimsel veya mühendislikle ilgili hesaplamalı çalışmalarda da çokça kullanılmaya başlamıştır. Bunda hem donanımsal hem de yazılımsal gelişmelerin etkisi olmuştur (Veri Madenciliği, Anonim, b.t.).

7. GEREÇ ve YÖNTEM

Veri madenciliği çalışmalarında kullanılan yöntemler farklılık gösterebilmektedir. Yapılan çalışmadaki yöntemler şu şekildedir;

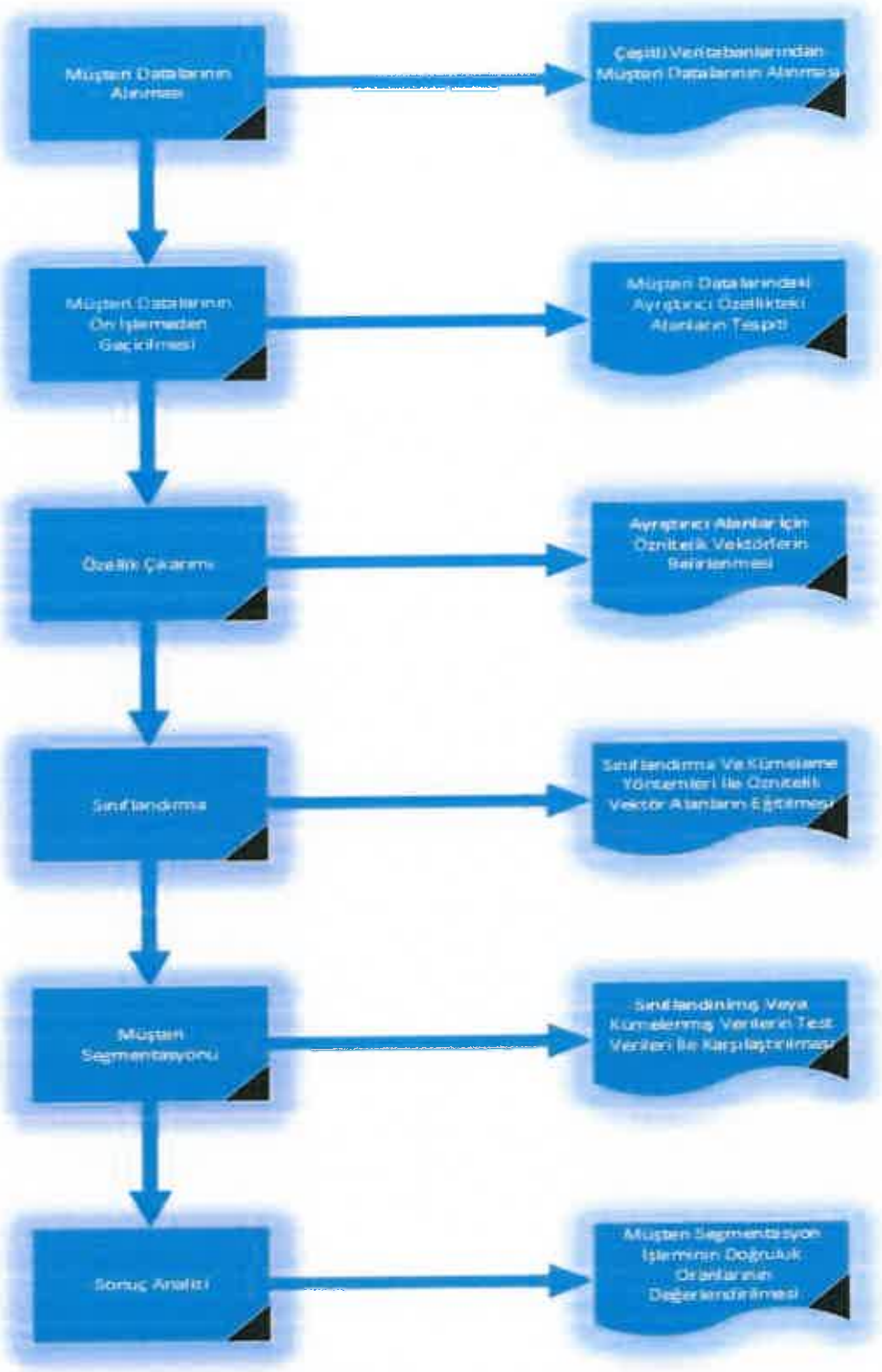
7.1. K-En Yakın Komşu Sınıflandırma Algoritması Kullanarak Müşteri Segmentasyonu

Bu bölümde veri madenciliği yöntemlerinden faydalanarak telekomünikasyon sektörü müşterilerinin segmentasyonu ve sonrasında ise müşteri kayıplarına yönelik analiz uygulamalarının yapılmasına yer verilmiştir. Müşteri segmentasyonuna yönelik çalışma yapılırken veri madenciliği süreçlerinden yararlanılmıştır(Nanbiyev ve diğ.2005).

Müşteri ilişkileri yönetimi (CRM) sistemleri günümüzde firmaların müşterilerine daha kolay ulaşarak onlara daha iyi hizmetler verebilmeyi sağlayan oluşumlardır. Müşteri ilişkileri yönetimi sistemlerinde veri madenciliği çalışmaları çok fazla kullanılan bir yöntem haline gelmiştir (Nanbiyev ve diğ.2005).

Bu tez çalışmasında müşteri segmentasyonu yapmak amacıyla müşterilerin kullanım alışkanlıkları, davranışlarının yanı sıra demografik ve coğrafik özelliklerde göz önünde bulunularak incelemeler yapılmış ve daha verimli sonuçlar elde etmek amaçlanmıştır.

Telekomünikasyon şirketi müşterilerinin verileri kullanılarak bu müşterilerin segmentasyon işlemlerinin yapılması amacıyla pyhthon programı,ve kümeleme algoritması olarak ise k-nn olarak adlandırılan k-en yakın komşu algoritması kullanılmıştır. Bu algoritmanın kullanılma amacı ise uygulanması ve eğitim işlemlerinin kolaylığıyla birlikte güvenilir sonuçlar vermesidir. Müşteri segmentasyon süreçlerini gösteren akış şeması Şekil 7.1'de gösterilmiştir (Nanbiyev ve diğ.2005).



Şekil 7.1. Müşteri Segmentasyonu Sürecinin Akış Seması

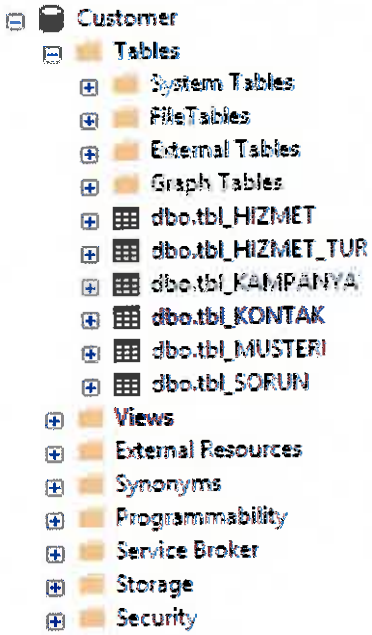
7.2. Uygulama İçin Veritabanı ve Veri Ambarı Tasarımı

Yapacak olduğumuz müşteri segmentasyonu çalışmasında telekomünikasyon sektöründe Türkiye'de öncü firmalardan olan bir şirketin müşteri dataları kullanılacaktır. Çeşitli veritabanları üzerlerinden elde edilecek olan müşteri datalarıyla örnek bir veritabanı ve veri ambarı yapısı oluşturularak çalışmalar bu veriler üzerinden yapılacaktır.

Şirketler sahip oldukları personel bilgilerini, müşteri bilgilerini, satış ve ürün bilgilerini tek bir veritabanı üzerinde tutabilecekleri gibi birden fazla veritabanı üzerinde de tutabilirler. Şirketler yapılarının büyüklüklerine göre sahip oldukları bilgileri farklı operasyonel yapılar üzerinde tutabilirler. Verilerin farklı ortamlarda tutulması gibi sebeplerden dolayı verilerin raporlanmasında sorunlar oluşabilir. Bu sorunların ortadan kaldırılması amacıyla veri ambarı yapısının oluşturulması yöntemine başvurulmuştur.

Veri ambarları bir veya birden fazla kaynaktan toplanan verileri tek bir yapıda toplamak amacıyla kullanılır. Farklı kaynaklardaki veriler çeşitli aktarım toolları yardımıyla veri ambarında toplanırlar. Örneğin ETL ve Golden Gate gibi toollar.

Bu çalışmada kullanılmak amacıyla oluşturulan veri tabanının görüntüsü Şekil 7.2.'de gösterilmiştir.



Şekil 7.2. Çalışmada kullanılan veritabanı

Müşteri segmentasyonu yapmak amacıyla kullanılabilecek olan müşteri nitelikleri Çizelge 7.1.'de gösterilmiştir.

Çizelge 7.1. Müşteri segmentasyonunda kullanılan müşteri nitelikleri

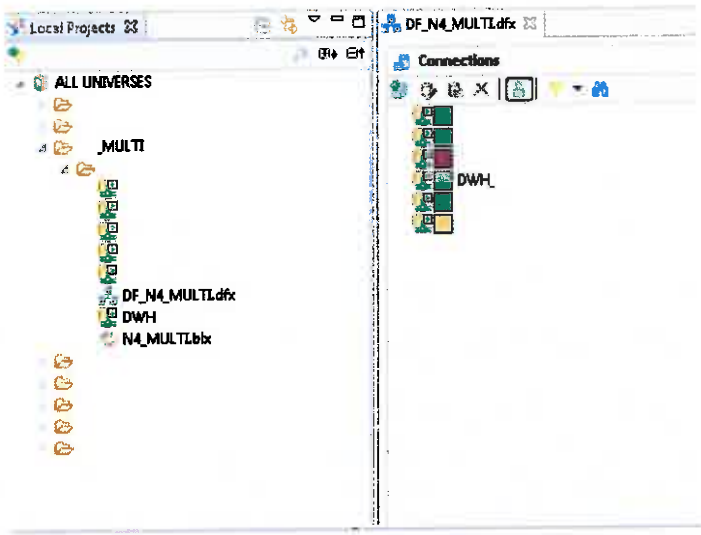
NO	NITELIK AÇIKLAMA
1	Kullanıcının aktif olduğu ay sayısı
2	Günlük arama dakikaları toplamı
3	Uluslararası arama sayıları toplamı
4	Günlük arama sayıları toplamı
5	Uluslararası arama ücretleri toplamı
6	Günlük arama ücretleri toplamı
7	Müşteri hizmetlerini arama sayısı
8	Akşam arama dakikaları toplamı
9	Akşam arama sayıları toplamı
10	Demografik bilgiler: Yaş, lokasyon, çocukların sayısı ve yaşları
11	Mobil cihaz: Yeni mi yenilenmiş mi?
12	Hane halkının yaşı
13	Müşterinin ömrü boyunca kullandığı aylık ortalama dakika
14	Sesli aramalarda paket dışında yapılan kullanımlardan elde edilen gelir
15	Paket dışında yapılan toplam kullanımlardan elde edilen gelir
16	Paket dışında kullanılan dakikalar
17	Müşterinin şimdiye kadar aylık yaptığı görüşme sayısı ortalaması
18	Sesli aramalarda paket dışında yapılan kullanımlardan elde edilen gelirin ortalaması
19	Paket dışında yapılan kullanımlardan elde edilen gelirinin ortalaması
20	Hesap harcama limiti
21	Sesli aramalarda paket dışında kullanılan dakikaların ortalaması
22	Gelir aralığı (fatura tutarı)
23	Müşterinin kaç aydır hizmet almakta olduğu
24	Toplam aylık yinelenen ücret ortalaması
25	Verilen telefon Sayısı
26	Kullanılan dakikalar
27	Yapılan sesli aramaların sayısı
28	Tamamlanan sesli aramaların sayısı
29	Yapılan aramaların sayısı
30	Tamamlanan aramaların sayısı
31	Yoğun olan zamanlarda yapılmış hem gelen hem de iden sesli aramalar sayısı
32	Aylık gelir aralığı (fatura tutarı)
33	Müşteri hizmetleri aramalarında kullanılan ortalama yuvarlatılmış dakika
34	Müşteri hizmetleri aramalarında kullanılan yuvarlatılmış dakika
35	Alınan sesli aramaların sayısı
36	Müşterinin ömrü boyunca fatura düzenlenmiş toplam dakikalar
38	Önceki 3 aylık dönemde kullanılan aylık ortalama dakikalar

Çizelge 7.1. Müşteri segmentasyonunda kullanılan müşteri nitelikleri (devam)

NO	NITELIK AÇIKLAMA
39	Toplam aylık yinelenen ücret aralığı
40	Tamamlanan sesli aramalar kullanılan yuvarlatılmamış dakika
41	Yoğun olmayan zamanlarda yapılan arama sayıları
42	Gece kullanma için isteklilik faktörü (gece arama ücretleri, gece arama dakikası)
43	Gündüz kullanma için isteklilik faktörü (gündüz arama ücretleri, gündüz arama dakikası)
44	Uluslararası kullanma için isteklilik faktörü (uluslararası arama ücretleri, uluslararası arama dakika)
45	Akşam kullanma isteklilik faktörü (akşam arama ücretleri, akşam arama dakika)
46	Kullanım süresi faktörü (aylık, günlük çağrı sürelerini vermektedir)
47	Müşteri hizmetleri faktörü (müşteri hizmetleri dakikaları ve mesaj adetler)
48	Ödenmemiş bakiyeler ise müşteri kaybı olasılığı ile doğru ilişkilidir.
49	Ödenmemiş aylık senet sayısı müşteri kaybı olasılığı ile doğru ilişkilidir.
50	Müşterinin statüsü
51	Geçiş maliyetleri, üyelik kartları
52	Yoğun kullanıcılara prim avantajlar sunmak
53	Bütün üyelere açık asgari avantajlar sunmak
54	Hiçbir avantaj sunmamak
55	Müşteri ile ilgili değişkenler (Müşteri notu)
56	Kullanım düzeyi ve mülkiyeti dayalı olarak en yüksek seviye
57	Kullanım düzeyi ve mülkiyeti dayalı olarak orta seviye
58	Kullanım düzeyi ve mülkiyeti dayalı olarak en düşük seviye
59	Cinsiyet
60	Ödeme metodu
61	Telefon Özellikleri
62	Servis kullanımı
63	Fatura tutarları
64	Ödenmemiş tutar
65	Ödenmemiş aylık fatura sayısı
66	Kontratın Tipi
67	Kontratın başlama tarihi
68	Kontratın statüsü
69	Kontratın bitiş tarihi
70	Kontrat süresi (başlangıçtan beri devam ettiği ay)
71	Müşterinin sahip olduğu aktif telefon sayısı
72	Müşterinin yaşı
73	Müşterinin segmenti

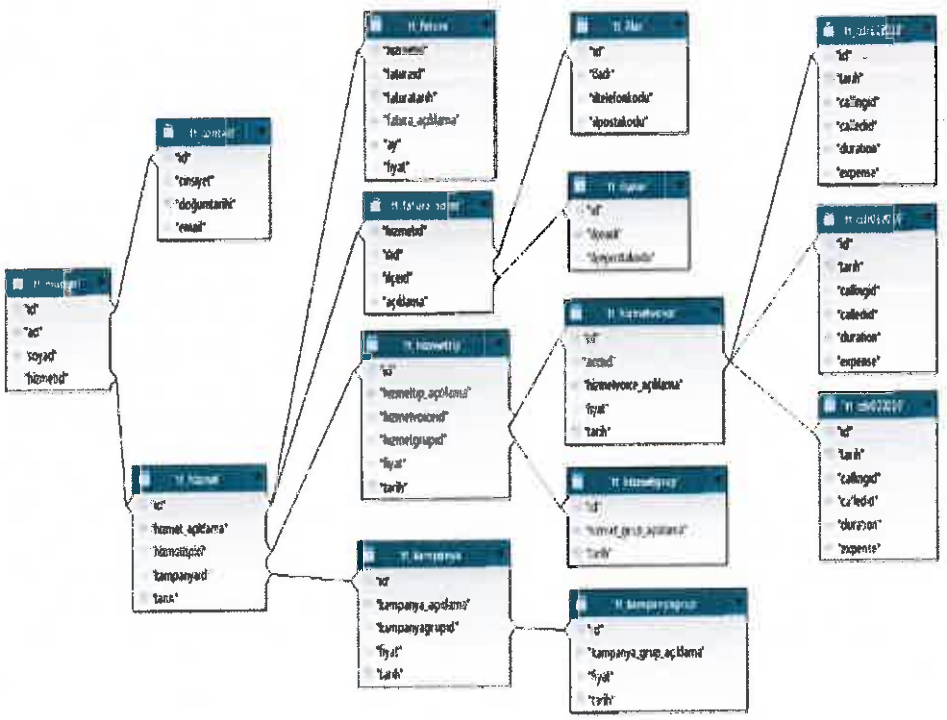
Veri tabanındaki verilerin modele uygun hale getirilmesi amacıyla ETL yapısı kurulmuştur. ETL işlemlerinin yapılması amacıyla universe tasarımında SAP toollarından olan SAP Information Desing Tool kullanılmıştır. Şekil 7.3.'te

Information Design Tool üzerinden veri tabanına erişmek amacıyla oluşturulan bağlantı yapısı gösterilmiştir.



Şekil 7.3. Information Desing Tool Bağlantı Yapısı

Bağlantı tanımlandıktan sonra veri tabanındaki tablolar Universe aktarılır ve birbirleriyle olan bağlantıları yapılır. Şekil 7.4.'te tabloların ilişkilerinin kurulmuş şekilleri gösterilmiştir.



Şekil 7.4. Universe Tablo Bağlantıları

7.3 Uygulama İçin Veri Madenciliği Süreci

Önceki bölümlerinde veri madenciliğinin süreçlerinden bahsedilmiştir. Veri madenciliği çalışmasının her bir adımı bir analiz ve araştırma gerektirmektedir. Veri madenciliği süreci kısa tanımı ile ihtiyaç duyulan verileri ve bu verilerin ne gibi avantajlar sağlayacağı bu avantajların sonucunda ortaya çıkabilecek faydalı çalışma süreçlerinin tamamı veri madenciliği süreci olarak tanımlanmaktadır. Veri madenciliği süreçleri şu şekilde incelenir (Gasson, ve diğ.2005).

7.3.1. Problem Tanımlama

Veri madenciliği çalışmalarının ilk aşaması daha öncede bahsedildiği gibi problem tanımlama safhasıdır. Problem tanımla safhasında yapılmak istenilen işler doğrultusunda ihyitaçlarımız belirlenir. Bu aşamadan yola çıkarak müşteri segmentasyonu işlemi yapılır.

Örneğin çalışmamıza konu olan telekomünikasyon firması müşterilerine çok sayıda hizmet ve kampanya sunmaktadır. Her müşteriye tanımlanmış bir Müşteri No bilgisi vardır. Bu numaralar için istenilen müşteri bilgilerine kolay bir şekilde ılaşılabilir. Ancak bir müşterinin sahip olduğu birçok bilgi farklı tablolar üzerinde tutulmaktadır. Bu nedenle müşterilerin istenilen bilgilerine ulaşabilmek amacıyla çok fazla sorgular yazmak gerekebilir. Bu sebepten ötürü tek bir çalışma ile müşterilerin sınıflandırma gereksinimi ortaya çıkar. Bu sınıflandırma çalışmasına da müşteri segmentasyonu denir (Gasson, ve diğ.2005).

Bu çalışmamızda kullanacak olduğumuz müşteri hizmet türleri müşterilere sunulan ses yani PSTN (Sabit Telefon) hizmetleridir.

7.3.2. Veriyi Anlama

Problem tanımlama safhasında veritabanı yapısı ile ilgili temel bilgiler verilmiştir. Veriyi hazırlama aşamasında ise çok daha fazla ayrıntı ve detaya girmeden tablolar arasındaki bağlantı ilişkileri ve tablolardaki veriler incelenecektir. Veritabanındaki tablolar birbirleriyle bazı kolonlar üzerinden ilişki kurmaktadırlar.

Veri tabanımızdaki bazı tablolardan örnekler alarak bu çalışmamızda kullanılacaktır. Alınan bu örnek veriler Türkiye geneli PSTN (Telefon abonelikleri) ve XDSL (İnternet) aboneliklerini kapsamaktadır (Gasson, ve diğ.2005).

Anazliz yapmak amacıyşa seçmiş olduğumuz örnek veri kümesinin seçiminde:

- Türkiye'nin önde gelen telekomünikasyon firmalarından birisinin sahip olduğu 120.000 müşteri kümesinden yaklaşık 5.000 adet örnek hizmet alınmıştır.
- Çalışma zamanında hizmet alma işlemi devam eden 4.500 müşteri bulunmuştur.
- Abonelerin fatura kalemlerinden yola çıkarak tahakkuk durumları dikkate alınmıştır.

Hazırlanan datalar tabloları anlamak adına detaylı bir şekilde incelenmiştir. Bu tabloların birbirleri ile olan bağlantı ve ilişkilerini, veri olmayan boş kolonları ve tablolardaki veri türlerini anlamak amacıyla analizler yapılmıştır. Bu analiz çlışmaları sırasında ise klasik SQL sorgulama yöntemleri ile birlikte SAP Universe Desing Tool programları kullanılmıştır.

Yapılacak olan segmentasyon analizinde Çizelge 7.2'de belirtilen nitelikler ayırt edici olarak yol gösterecektir.

Çizelge 7.2. Segmentasyon analizinde kullanılan alanlar.

NO	ÖZELLİK	AÇIKLAMA
1	Ortalama Adsl Kullanım	İnternet Kullanımları
2	Ortalama Telefon Kullanım	Telefon Kullanımları Dakika
3	Ortalama Fatura	Faturalar Tutarları
4	Ortalama Çağrı	Yapılan Görüşme Adet
5	Ortalama Arama Süresi	Yapılan Görüşme Süresi
6	Ortalama Çağrı Süresi	Ortalama Görüşme Süresi
7	Şikâyet Sayısı	Arıza Kaydı Adedi
8	Aktif Kullanım Ay	Abonelik Süresi
9	Kontratlı Olduğu Süre	Kontrat Süresi
10	Hizmet Tipi	Kullanılan Hizmet Türü
11	Kampanya	Kullanılan Kampanya Türü
12	Meslek	Mesleği
13	Gelir Düzeyi	Gelir Düzeyi
14	Ödenmemiş Aylık Fatura Adet	Toplam Ödenmemiş Fatura Adedi
15	Gecikmiş Fatura Adet	Toplam Gecikmiş Fatura Adedi
16	Hizmet Durumu	Hizmet Aktif\Pasif Durumu

7.3.3. Veri Hazırlama

Hazırlamış olduğumuz veri setine hakimiyet sağlandıktan sonraki adım ise verinin hazırlanması aşamasıdır. Veri analizi yapmadan önceki bu adım çalışmanın sağlıklı ve güvenilir olması açısından oldukça önem arz etmektedir. Bu nedenden dolayı yapılacak olan veri madenciliği çalışmalarında en fazla zaman ayrılması gereken aşamadır. Hazırlanan veri setleri bu adımda incelenerek eksik veri olan kısımlar tespit edilir, aykırı veriler içeren veriler tespit edilerek bunlar için çözümler üretilme işlemi yapılır, verilerin bütünleştirilmesi yada dönüştürülmesi gibi gereksinimler duyulması halinde bu yönde çalışmalar yapılır. Veri hazırlama aşamasında aşağıdaki çalışmalar yapılır.

7.3.3.1. Eksik Verilerin Analizi

Toplamış olduğumuz veri setinde çeşitli nedenlerden dolayı ölçülemeyen verilerin olduğu durumlar olabilir. Bu şekilde ölçülemeyen veriler eksik veri(Missing Data) olarak adlandırılmaktadır. Veri setimizde ki eksik verilerin tespiti ve tamamlanması amacıyla birçok yöntem kullanılabilir. Bu yöntemler şu şekilde incelenebilir.

- Veri setimizdeki eksik veriler çalışmamız açısından önem arzetmediği durumda eksik veriler "N/A" olarak değerlendirilir.
- Aynı örnek için çok sayıda eksik veri olması durumunda bu verileri silinerek veri setinden çıkartılır.
- Eksik verilerin olduğu alanlar numerik olması durumunda bu alanların ortalama değerleri hesaplanarak eksik verilere bu ortalama değerler atanır.
- Eksik verilerin olduğu alanlar kategorik değerler içermesi halinde en fazla tekrar eden değer eksik veri alanına atanır.
- Eksik veriler çeşitli veri madenciliği veya makine öğrenim algoritmaları ile tamamlanabilir.
- Regresyon analizi gibi yöntemler sayesinde eksik veriler tamamlanarak eksik veriler değer kazandırılır.

7.3.3.2. Aykırı Verilerin Analizi

Uç noktalar olarakta adlandırılan aykırı verilerin temizlenmesi gerekmektedir. Aykırı veriler veri madenciliği çalışmasının sağlıklı bir şekilde ilerlemesine engel

olarak yanıltıcı sonuçlar elde edilmesine sebep olabilir. Örneğin müşterilerin yaş bilgilerinin 150 olması aykırı veri olarak tanımlanabilir. Aykırı verilerin oluşma sebepleri ise şu şekildedir;

- Kodlama hatalarından veya verilerin kayıt işlemleri yapılırken hatalı olarak kayıt edilmesi.
- Olağandışı durumlar.
- Veriler doğru şekilde girilmiş olsalar dahi değişken kaynaklı durumlar nedeniyle aykırı veriler oluşabilir.

Veri setimizde oluşn aykırı verilerin düzenlenmesi amacıyla ise şu yöntemlerden yararlanılabilir;

- İnsanların denetimleri ile aykırı veriler düzeltilebilir.
- Eğri uydurma yöntemiyle aykırı veriler düzeltilebilir.
- Kümeleme yöntemi ile aykırı veriler düzeltilebilir.
- Kutulama (Binning) yöntemiyle aykırı veriler düzeltilebilir.

7.3.3.3. Normalizasyon

Veri madenciliği çalışmamız için hazırlamış olduğumuz veri setindeki numerik değerler için değişim aralıklarının çok fazla olması durumunda bu değerler için normalizasyon işleminin yapılması gerekmektedir. Nomalizasyon işlemleri için farklı yöntemler kullanılabilir. Bunların en bilindikleri ise min-max, z-score ve ondalık ölçekleme gibi yöntemlerdir. Normalizasyon yönteminde kullanılan formüller (7.1), (7.2) ve (7.3)'te gösterilmiştir (Prokoski ve Reidel 1999).

$$Min - Max\ Normalizasyon\ (V') = \frac{v - min_A}{max_A - min_A} \tag{7.1}$$

$$\text{Z} - \text{Score Normalizasyon}(V') = \frac{v - mean_A}{stand\_dev_A} \tag{7.2}$$

$$\text{Ondalık Normalizasyon}\ (v') = \frac{v}{10^j}\ \ j: Max\ (|v'|) < 1 \tag{7.3}$$

Yapılan tez çalışmasında verilerin daha anlanmlı ve birbirleriyle uyumlu olması için min-max normalizasyon yöntemi kullanılmıştır. Müşteri bilgilerine ait verilerin bir kısmı Çizelge 7.3 'te gösterilmiştir.

Çizelge 7.3. Normalizasyon Öncesi Müşteri Nitelikleri

ADSL KULLANIM GB	TELEFON KULLANIM DK	SIKAYET SAYISI	AKTIF AY	KAMPANYA ID	SEHIR KODU	YAS
0	12	4	20	4	31	35
11	10	4	20	4	42	46
27	4	3	46	6	54	57
29	4	7	32	1	53	57
31	1	4	20	3	45	49
29	3	6	42	3	53	56
0	34	4	9	5	31	35
36	1	6	7	4	33	36
0	38	5	10	5	41	45
38	0	9	10	4	36	40

Çizelge 7.3'te verilen müşteri verileri min-max normalizasyon yöntemi ile 0-1 aralığında değerlere normalize edilerek Çizelge 7.4'te gösterilmiştir.

Çizelge 7.4. Min-Max Normalizasyon İşlemi Sonrası Müşteri Nitelikleri

ADSL KULLANIM GB	TELEFON KULLANIM DK	SIKAYET SAYISI	AKTIF AY	KAMPANYA ID	SEHIR KODU	YAS
0,0000	0,1319	0,0667	0,3148	0,4286	0,1053	0,1228
0,1571	0,1099	0,0667	0,3148	0,4286	0,2982	0,3158
0,3857	0,0440	0,0333	0,7963	0,7143	0,5088	0,5088
0,4143	0,0440	0,1667	0,5370	0,0000	0,4912	0,5088
0,4429	0,0110	0,0667	0,3148	0,2857	0,3509	0,3684
0,4143	0,0330	0,1333	0,7222	0,2857	0,4912	0,4912
0,0000	0,3736	0,0667	0,1111	0,5714	0,1053	0,1228
0,5143	0,0110	0,1333	0,0741	0,4286	0,1404	0,1404
0,0000	0,4176	0,1000	0,1296	0,5714	0,2807	0,2982
0,5429	0,0000	0,2333	0,1296	0,4286	0,1930	0,2105

7.3.3.4. Veri Bütünleştirme

Veri madenciliği çalışmalarımızda kullanacağımız veriler birçok farklı veritbanı veya veri kaynağından elde edilebilir. Farklı veri kaynaklarından elde edilen verilerin veri madenciliği çalışmasında kullanılabilmesi için veri bütünleştirme işlemlerinin yapılması gerekmektedir. Yapmış olduğumuz çalışmada müşterilerin arama bilgileri ile müşterilerin karakteristik özellikleri farklı veritbanları üzerinde tutulmaktadır. Bu nedenden dolayı verilerin tek bir kaynakta toplanması

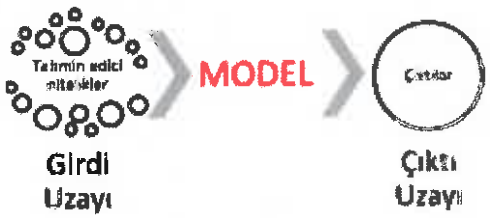
amacıyla SAP Universe Desing Tool uygulaması kullanılarak veriler tek bir kaynakta toplanmışlardır.

7.3.3.5. Veri Dönüştürme

Veri dönüştürme işleminin yapılması veri madenciliği çalışmamızın sonuçlarının daha sağlıklı olmasını sağlayacaktır. Bu aşamada veri tipleri arasındaki dönüşüm işlemleri yapılarak veriler uyumlu hale getirilir. Örneğin müşterilerin arama sürelerindeki saat veya dakika bazlı aralıklar ile sınıflandırılması ve gruplanması veya yaş değerlerinin belirlenen bir ölçek aralığında aralıklandırması veri dönüştüme aşamasının örneğidir.

7.4. Modelleme

Hazırlanan veri setinde tam yetki sağlandıktan sonra ve ön işleme adımları gerçekleştikten sonra model kurma işlemi aşamasına geçilir. Model kurma aşamasında problem ve öğrenme stratejilerine uygun bir algoritma sayesinde giriş örneklerinin istemiş olduğumuz çıktılar vermesi amacıyla dönüştürme işlemleri yapılır. Yapılacak olan müşteri segmentasyon çalışmasında veri madenciliği tekniklerinden k-en yakın komşu algoritmasını bu bölümde anlatarak çalışmada sınıflandırma yöntemi için kullılacaktır. Veri madenciliği çalışmasında modelin oluşturulması problemin çözümü anlamına gelmemektedir. Kurulan modellerin değerlendirilerek çözüme en uygun modeli seçmek gerekmektedir. Birçok model değerlendirme yöntemi sayesinde en uygun model bulunmaya çalışılır ve sonrasında kullanılır. İlgili değişkenler tahmin edici değişkenlerden oluşan bir girdi uzayı sayesinde elde edilir. Modelin doğru bir tahminleme yapması ve doğru sonuçlar üretmesi veri madenciliği teknikleri sayesinde olur. Şekil 7.5'de tahmin edici nitelikler yardımıyla elde edilen çıktıları gösteren şekil bulunmaktadır (Nixon ve diğ. 1999).



Şekil 7.5. Makine Öğrenme Modeli

M={M1, M2,.....Mn} şeklindeki küme bütün modelleri temsil eden bir küme olarak düşünülürse, modeller içerisinde en uygun performansta olanı seçmek gerekir. Kurulan modellerden hangisi üzerinde test verilerilerinin en iyi sonucu verdiğini belirlemek gerekmektedir.

7.4.1. K-En Yakın Komşu Algoritmasına Ait Uygulama Örneği

Bu algoritma yönteminde veriler örüntü uzayında tutulmaktadır. Bu örüntü uzayına bakılarak k-en yakın komşu algoritmasındaki bilinmeyen verilerin hangi sınıflara ait oldukları tespit edilmeye çalışılır. Bu algoritmada Öklid, Manhattan gibi yöntemler kullanılarak uzaklık hesaplamaları yapılarak komşular arası uzaklıklar belirlenir. K-En yakın komşu algoritmasının adımları şu şekildedir.

- Belirlenen bir nokta için k adet en yakın komşu adedi belirlenir.
- Belirlenen noktanın diğer tüm noktalara olan uzaklıkları belirlenir.
- Hesaplanan uzaklıklar sonrasında kayıtlar arası uzaklıklar belirlenir ve en küçük uzaklıktaki k alınır.
- Yapılan seçimlerden en fazla tekrar eden kategorideki kayıt belirlenir.
- Belirlenen kategori tahmin edilmek istenilen verinin kategorisi olarak belirlenir.

K-en yakın komşu algoritmasının çalışma mantığını basit bir şekilde inceleyecek olursak;

K-en yakın komşu algoritması hem regresyon hemde sınıflandırma problemleri için kullanulan bir yöntemdir. Bu algoritmada bir eğitim süreci bulunmamaktadır.

Algoritmaya yeni bir veri sunulduğunda bu verinin diğer tüm noktalara olan mesafeleri hesaplanır. *K* değerine bağlı olarak veri setindeki en yakın komşular belirlenir.

Eğer $k=1$ olduğunda tüm noktalardan minimum mesafede olan veri ile aynı kategoriye dahil edilir.

$k>1$ olduğu durumlarda ise k nın minimum uzaklıkta olan değeri kategori olarak belirlenir.

Sınıflandırma işlemi için veri setindeki k değerlerinin çoğunluğuna göre yeni kategori sınıflandırma kategorisi olarak kabul edilir.

Regresyon analizi için ise listedeki tüm değerlerin ortalaması alınır.

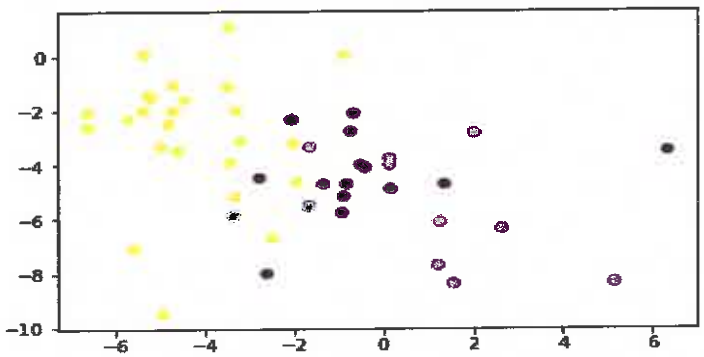
K-en yakın komşu algoritmasının veriler üzerinden çalışma şekli şu şekildedir. Şekil 7.6'da Python ortamında veri setinin oluşturulma kodu gösterilmiştir.

```
In [2]: (X,y) = make_blobs(n_samples=50,n_features=2,centers=2,cluster_std=1.95,random_state=50)
```

Şekil 7.6. Eğitim için veri seti belirlenmesi

Şekil 7.7'da veri setinin iki farklı sınıfa ayrılmış hali gösterilmiştir.

```
In [3]: plt.scatter(X[:,0],X[:,1],marker='o',c=y)
        plt.show()
```

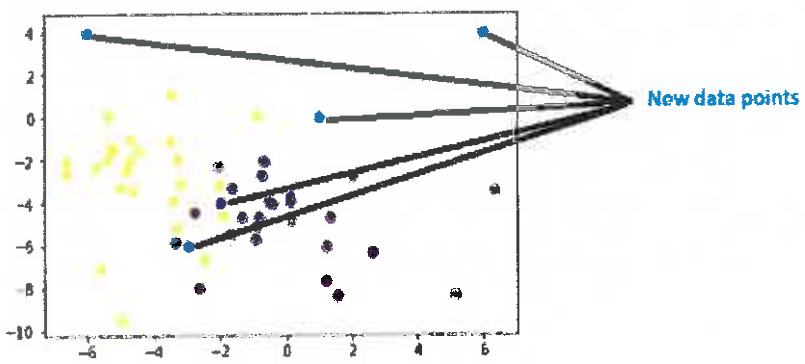


Şekil 7.7. Veri Setinin Sınıflandırılması

Test verileri üzerinden sınıflandırma işlemi tamamlandıktak sonra gerekli yeni verilerin hangi kümeye dahil edilmesi gerektiğine karar vermek gerekmektedir. Şekil 7.8'de veri setine yeni verilerin eklenme işlemi gösterilmiştir.

```
In [4]: prediction_points=[[-2,-4],[-3,-6],[1,0],[6,4],[-6,4]]
        prediction_points=np.array(prediction_points)

        plt.scatter(X[:,0],X[:,1],marker='o',c=y)
        plt.scatter(prediction_points[:,0],prediction_points[:,1],marker='o')
        plt.show()
```



Şekil 7.8. Yeni verilerin veri setine dahil edilmesi

Yeni veri noktalarının veri setine dahile edilmesinden sonra her bir veri noktasının diğer tüm verilere olan uzaklıkları hesaplanır.

Bir noktanın diğer noktalara olan mesafesinin hesaplanmasında ise şu yöntemler kullanılır. Noktalar arası mesafelerin hesaplanmasında kullanılan formüller (7.4), (7.5) ve (7.6)'da gösterilmiştir (Yang ve Huang, 1994).

Uzaklık Fonksiyonları

$$d(Euclidean) = \sqrt{\sum_{i-1}^{n}(x_i - y_i)^2} \tag{7.4}$$

$$d(Manhattan) = \sum_{i-1}^{n}|x_i - y_i| \tag{7.5}$$

$$d(Minkowski) = \left(\sum_{i-1}^{n}(|x_i - y_i|)^q\right)^{1/q} \tag{7.6}$$

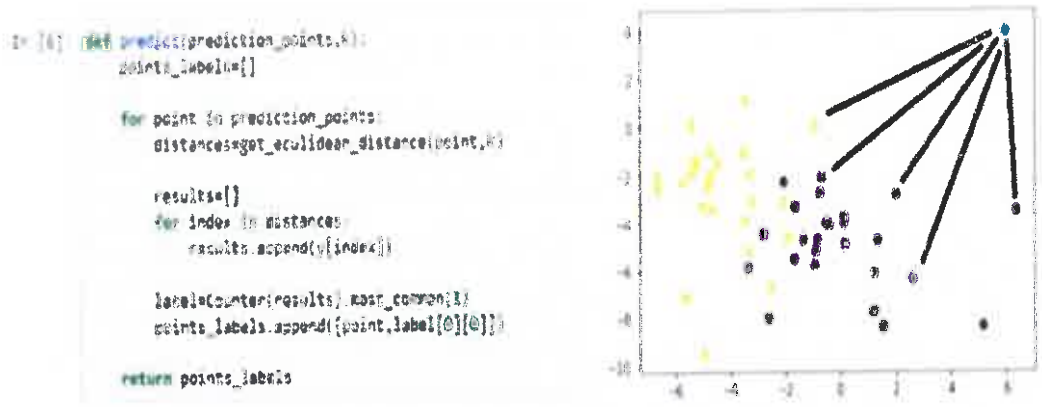
Not: d iki nokta arasındaki uzaklıktır. Yapılmış olan bu çalışmada öklit mesafe hesaplama yöntemi kullanılmıştır.

Örnek olarak öklit mesafe hesaplama yöntemini kullandığımızı düşünürsek ve k değerinin 1 olarak kabul edersek Şekil 7.9'de yeni verilerin diğer tüm noktalara olan uzaklıkları öklit mesafe yöntemi ile hesaplanması gösterilmiştir.

```
In [5]: def get_eculidean_distance(point,k):
            euc_distance = np.sqrt(np.sum((X - point)**2 , axis=1))
            return np.argsort(euc_distance)[0:k]
```

Şekil 7.9. K-nn Öklit Mesafe Hesaplama Formülü

K değerini 1'den büyük kabul ederek 5 olarak belirlersek bu durumda 5 minimum mesafedeki noktayı bazalırız. Sonrasında ise bu 5 nokta içerisinden en fazla olan kategoriye dahil ederiz. Yeni veri noktasının ait olduğu sınıf Şekil 7.10'da gösterilmiştir.



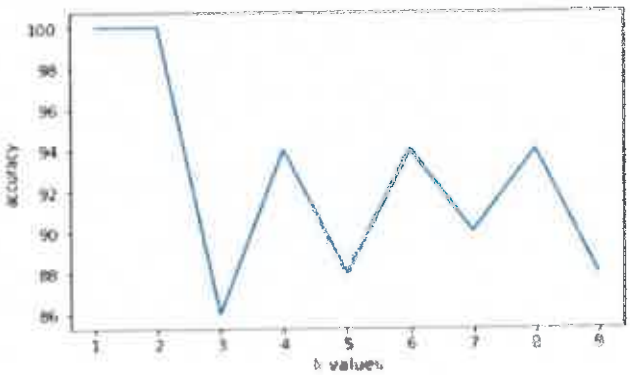
Şekil 7.10. Yeni veri Noktasının sınıflandırılması

Şekil 7.10'da görüldüğü üzere yeni veri noktamız mor renklı kategorilerin çoğunlukta olması nedeniyle bu gruba dahil edilmiştir.

Yeni veri noktasının doğru kategoriye dahil edildiğini anlamak için k değerinin bir çok farklı değer ile hesaplanması gerekmektedir. Farklı k değerleri uygulanarak yapılan sınıflandırma Şekil 7.11'de gösterilmiştir.

```
In [16]: #for different K values
         acc=[]
         for k in range(1,10):
             results=predict(X,k)
             predictions=[]
             for result in results:
                 predictions.append(result[1])
             acc.append([get_accuracy(predictions),k])
         plotx=[]
         ploty=[]
         for a in acc:
             plotx.append(a[1])
             ploty.append(a[0])

         plt.plot(plotx,ploty)
         plt.xlabel("k values")
         plt.ylabel("accuracy")
         plt.show()
```



Şekil 7.11. Farklı k değerleri ile mesafe hesaplama

Şekil 7.11 'de görüldüğü gibi mesafe değerlerinin hesaplandığı sonuçlar k değerinin 1 ve 2 olduğu durumlarda en iyi sonucu vermektedir.

7.5. Karışıklık Matrisi (Confusion Matrix)

Uygulanan k değerlerinin değerlendirilmesi amacıyla hata matrisi (confusion matrix) kullanılmıştır. Makine öğrenmesinde kullanılan sınıflandırma modellerinin performansını değerlendirmek için hedef niteliğe ait tahminlerin ve gerçek değerlerin karşılaştırıldığı hata matrisi sıklıkla kullanılmaktadır. Her ne olursa olsun sınıflandırma tahminleri şu dört değerlendirmeden birine sahip olacaktır (Davidson, 2002):

- Doğruya doğru demek (True Positive – TP) DOĞRU
- Doğruya yanlış demek (True Negative – TN) YANLIŞ
- Yanlışa doğru demek (False Positive – FP) YANLIŞ
- Yanlışa yanlış demek (False Negative – FN) DOĞRU

En yüksek doğruluk oranına sahip, iki veya daha fazla olduğu durumda makine öğrenmesi, sınıflandırmanın performansını ölçebilmektedir. Tahmin edilen ve gerçek değerlerle 4 farklı kombinasyonlu bir tablodur. Şekil 7.12'de Karışıklık Matrisinin (Confusion Matrix) yapısı gösterilmiştir (Davidson, 2002).

		TAHMİN		
		YOK	VAR	TOPLAM
GERÇEK	YOK	TN 100	FP 20	120
	VAR	FN 10	TP 200	210
	TOPLAM	110	220	

Şekil 7.12. Karışıklık Matrisi (Confusion Matrix) Yapısı

Yapılan çalışmada müşterilerin memnuniyetsizlikleri sonucunda firmadan ayrılma eğilimlerinin hesaplanması işlemleri ve uygun sınıfa dahil edilip edilmediklerinin kontrolü amacıyla confusion matrix kullanılmıştır.

True Positive (Doğru Pozitif) : Olumlu tahmin ettiniz ve bu doğru. Müşterinin firmadan ayrıldığı doğru tahmin ettiğimiz ve müşterinin gerçekten ayrıldığı durum (Davidson, 2002).

True Negative (Doğru Negatif): Olumsuz tahmin edilen ve doğru. Müşterinin firmadan ayrılmadığını tahmin ettiğimiz ve gerçekten ayrılmadığı durum.

False Positive (Yanlış Olumlu) : Olumlu tahmin edilen ve yanlış.

False Negative (Yanlış Olumsuz) : Olumsuz tahmin edilen ve yanlış.

Karışıklık matrisinden hesaplanan bazı oranlar vardır. TP, TN, FP ve FN'nin birbirleriyle ilişkisini gösteren bu terminolojiler (7.7), (7.8) ve (7.9)'da verilmiştir (Davidson, 2002).

$$Toplam = TP + FP + FN \quad (7.7)$$

$$GerçekPozitifler = TP + FN \quad (7.8)$$

$$GerçekNegatifler = TN + FP \quad (7.9)$$

Doğruluk Oranı (Accuracy Rate): Genel olarak, sınıflayıcının ne sıklıkta doğru tahmin ettiğinin bir ölçüsüdür. (7.10)'da doğruluk oranı formülü gösterilmiştir. (Davidson, 2002).

$$DoğrulukOranı = (TP + TN)/TOPLAM \quad (7.10)$$

Yanlış Sınıflandırma Oranı (Misclassification Rate): Genel olarak, sınıflayıcının ne sıklıkta yanlış tahmin ettiğinin bir ölçüsüdür. Hata Oranı (Error Rate) olarak da bilinir. (7.11)'de formülü gösterilmiştir (Davidson, 2002).

$$YanlışSınıflandırmaOranı = (FP + FN)/TOPLAM \quad (7.11)$$

Gerçek Pozitif Değerlerin Oranı (True Positive Rate): Sınıflayıcının ne kadar gerçek pozitif değeri doğru tahmin ettiğinin bir ölçüsüdür. (7.12)'de formülü gösterilmiştir (Davidson, 2002).

$$GerçekPozitifDeğerlerinOranı = TP/GERÇEKPOZİTİFLER \quad (7.12)$$

Gerçek Negatif Değerlerin Oranı (True Negative Rate):Sınıflayıcının ne kadar gerçek negatif değeri doğru tahmin ettiğinin bir ölçüsüdür. (7.11)'de hesaplama formülü gösterilmiştir (Davidson, 2002).

$$GerçekNegatifDeğerlerinOranı = TN/GERÇEKNEGATİFLER \quad (7.13)$$

Yanlış Pozitif Değerlerin Oranı (False Positive Rate): Gerçek değeri 0 olmasına karşın 1 olarak tahmin edilenlerin oranıdır. Yan ürün (Fall-out) olarak da bilinir. (7.14)'te hesaplama formülü gösterilmiştir (Davidson, 2002).

$$YanlışPozitifDeğerlerinOranı = FP/GERÇEK\ NEGATİFLER \quad (7.14)$$

Yanlış Negatif Değerlerin Oranı (False Negative Rate): Gerçek değeri 1 olmasına karşın 0 olarak tahmin edilenlerin oranıdır. Kayıp oranı (Miss Rate) olarak da bilinir. (7.15)'te hesaplama formülü gösterilmiştir (Davidson, 2002).

$$YanlışNegatifDeğerlerinOranı = FN/GERÇEK\ POZİTİFLLER \quad (7.15)$$

Hassasiyet (Precision): Tüm sınıflardan, doğru olarak ne kadar tahmin edildiğinin bir ölçüsüdür. Mümkün olduğu kadar yüksek olmalıdır. (7.16)'da hesaplama formülü gösterilmiştir (Davidson, 2002).

$$Hassasiyet = TP/TP + FP \quad (7.16)$$

Hata Oranı (Null Error Rate): Çoğunluk sınıfına ait değer (1 veya 0) sürekli tahmin edilseydi ne oranda yanlış tahminleme yapıldığının bir ölçüsüdür. Bu, sınıflandırıcıların karşılaştırılması için yararlı bir temel metrik olabilir. Bazen en iyi sınıflandırmayı yapan modelin hata oranı, boş hata oranından daha yüksek olabilir; buna Doğruluk Paradoksu (Accuracy Paradox) denir.

Cohen's Kappa: Sınıflandırıcının aslında ne kadar iyi performans gösterdiğinin bir ölçüsüdür. Cohen's Kappa sadece iki sınıflandırıcı arasında karşılaştırma yapmaya yarar (Davidson, 2002).

F Puanı (F Score): Bu, gerçek pozitif değerlerin oranının (recall) ve hassasiyetin (precision) harmonik ortalamasıdır. (7.17)'de hesaplama formülü gösterilmiştir (Davidson, 2002).

$$F\ Puanı = \frac{2 * Hassasiyet * Gerçek\ Pozitif\ Değerlerin\ Oranı}{Hassasiyet + Gerçek\ Pozitif\ Değerlerin\ Oranı} \quad (7.17)$$

ROC Eğrisi (ROC Curve): Bu, sınıflandırıcının tüm olası değerler üzerinde performansını özetlemek için kullanılan bir grafiktir. Belirli bir sınıfa gözlem atanması eşiğini değiştirdiğinizde Gerçek Pozitif Değerlerin Oranına (Hassasiyet) (x ekseni) karşı Yanlış Pozitif Değerlerin Oranını (Özgüllük) (y ekseni) çizerek oluşturulur. ROC Eğrisi, Hassasiyet / Özgüllük (Sensitivity / Specificity) raporu

oluşturmaya yarar. Şekil 7.13'te ayrılan müşterilere ait karışıklık matrisi kontrolü gösterilmiştir (Davidson, 2002).

		Gerçekleşen	
		Devam Eden Müşteri	Ayrılan Müşteri
Tahmin Edilen	Devam Eden Müşteri	tp	fn
	Ayrılan Eden Müşteri	fp	tn

Şekil 7.13. Ayrılan Müşteri Karışıklık Matrisi Kontrolü

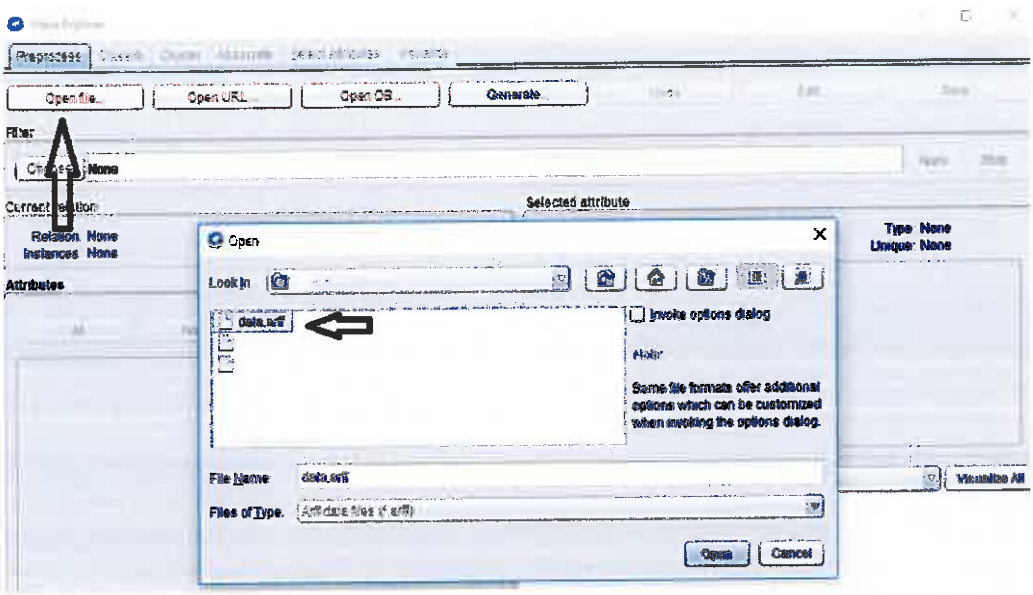
7.6. Uygulama

Yapılan çalışma sonucunda Weka programı ile müşteri segmentasyonu yapmak için aşağıdaki veriler wekaya yüklenmiştir.

Çizelge 7.5. Müşteri Segmentasyonu için kullanılan nitelikler

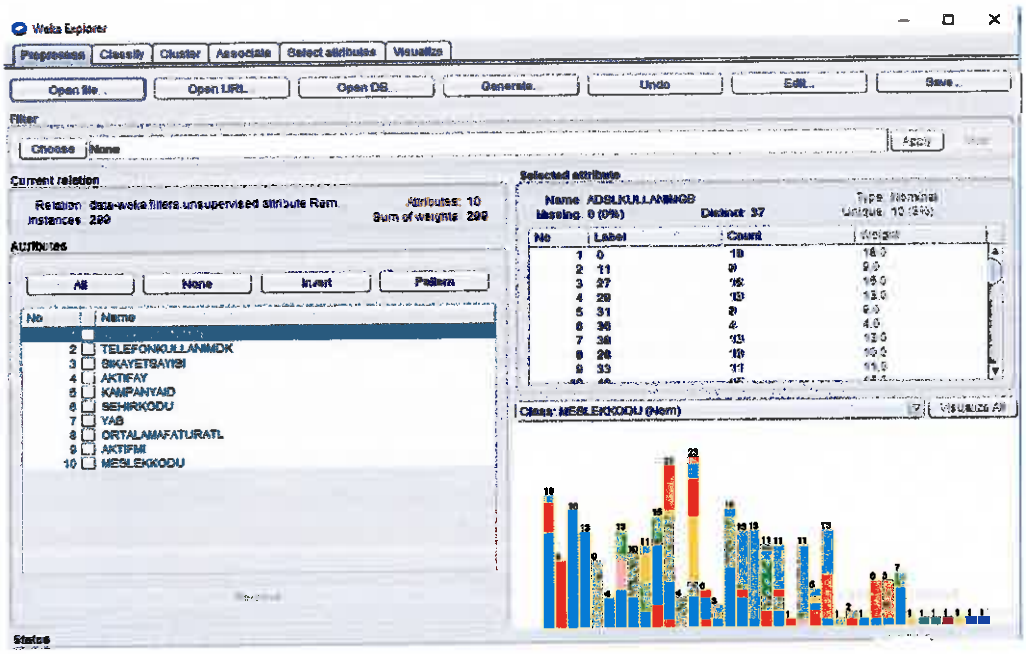
NO	ÖZELLİK	AÇIKLAMA
1	Ortalama Adsl Kullanım	İnternet Kullanımları
2	Ortalama Telefon Kullanım	Telefon Kullanımları Dakika
3	Ortalama Fatura	Faturalar Tutarları
4	Ortalama Arama Süresi	Yapılan Görüşme Süresi
5	Şikayet Sayısı	Arıza Kaydı Adedi
6	Aktif Kullanım Ay	Abonelik Süresi
7	Hizmet Tipi	Kullanılan Hizmet Türü
8	Kampanya	Kullanılan Kampanya Türü
9	Meslek	Mesleği
10	Gelir Düzeyi	Gelir Düzeyi
11	Ödenmemiş Aylık Fatura Adet	Toplam Ödenmemiş Fatura Adedi
12	Gecikmiş Fatura Adet	Toplam Gecikmiş Fatura Adedi
13	Hizmet Durumu	Hizmet Aktif\Pasif Durumu

Çizelge 7.5'te belirtilen müşteri niteliklerinin Weka programında eğitilmesi amacıyla eklenmesi gerekmektedir. Bu niteliklerin Weka'ya eklenme işlemi Şekil 7.14'te gösterilen menü ile yapılmaktadır.



Şekil 7.14. Weka'ya Veri Ekleme Menüsü

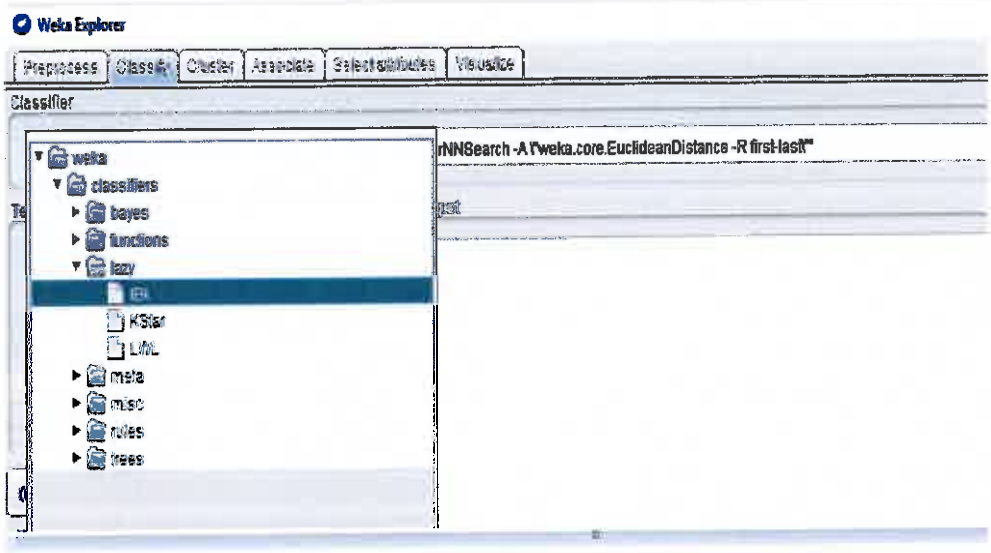
Şekil 7.14'te görüldüğü gibi Weka Explorer programı üzerinden Preprocess mesnüsü altında bulunan Open File butonu yardımıyla müşteri niteleklerine ait arff dosyası Weka'ya eklenmektedir.



Şekil 7.15. Niteliklerin Wekaya Yüklenmesi

Şekil 7.15'te nitelik verilerin weka programına eklenmiş durumları gösterilmiştir.

Weka'ya eklenen niteliklerin sınıflandırma işlemlerinin yapılması amacıyla uygulanacak olan method seçilmelidir. Bu işlem için Weka Explorer programı üzerindenki Classify menüsü altından uygulanacak sınıflandırma yöntemi seçilir. Şekil 7.16'da K-en yakın komşu algoritmasının seçilmesi işlemleri gösterilmiştir.

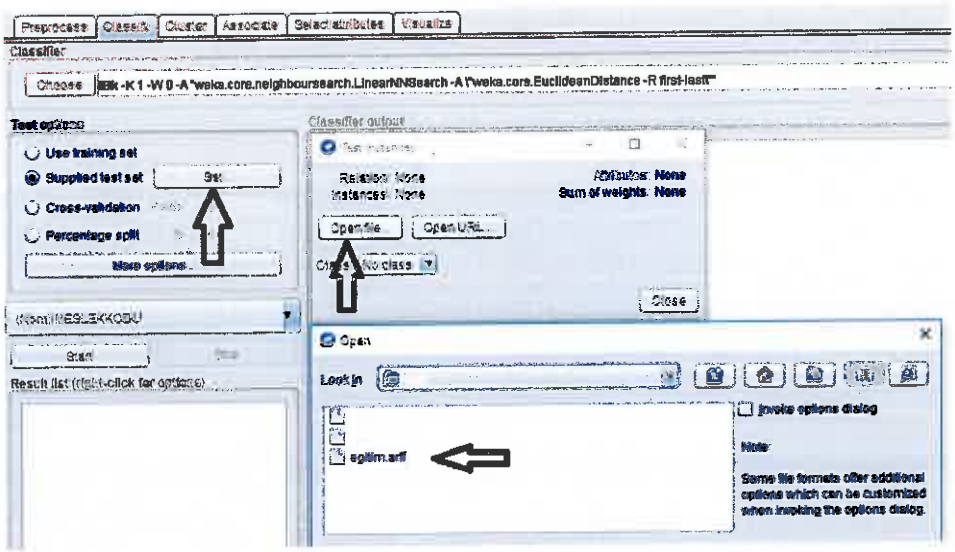


Şekil 7.16. Sınıflandırma Yönteminin Seçilmesi

Belirlenen niteliklerin Weka'ya yüklenmesinden sonra uygun sınıflandırma yöntemi seçilerek müşterilerin sınıflandırılması yapılır. Sınıflandırılma işlemlerinin yapılması amacıyla eğitim verilerininde Weka'ya yüklenmesi gerekmektedir. Çizelge 7.6'da eğitimde kullanılan test verileri ve Şekil 7.17'de bu verilerinin Weka'ya yüklenmesi aşaması gösterilmiştir.

Çizelge 7.6. Müşteri Segmentasyonunda Kullanılan Test Verileri

ADSL KULLANIM GB	TELEFON KULLANIM DK	SIKAYET SAYISI	AKTIF AY	KAMPANYA ID	YAS	ORTALAMA FATURA TL	AKTIFMI
7	12	4	20	4	35	33	0
11	10	4	20	4	46	53	0
27	4	3	46	6	57	73	1
29	4	7	32	1	57	76	0
31	1	4	20	3	49	81	1
29	3	6	42	3	56	84	0
17	34	4	9	5	35	91	0
36	1	6	7	4	36	105	1
34	38	5	10	5	45	100	0
38	0	9	10	4	40	104	1



Şekil 7.17. Sınıflandırma İşlemi İçin Eğitim Verilerinin Wekaya Yüklenmesi

Eğitim verileri Weka'ya yüklendikten sonra sınıflandırma işlemi için gerekli çalışma yapılmaya başlanır.

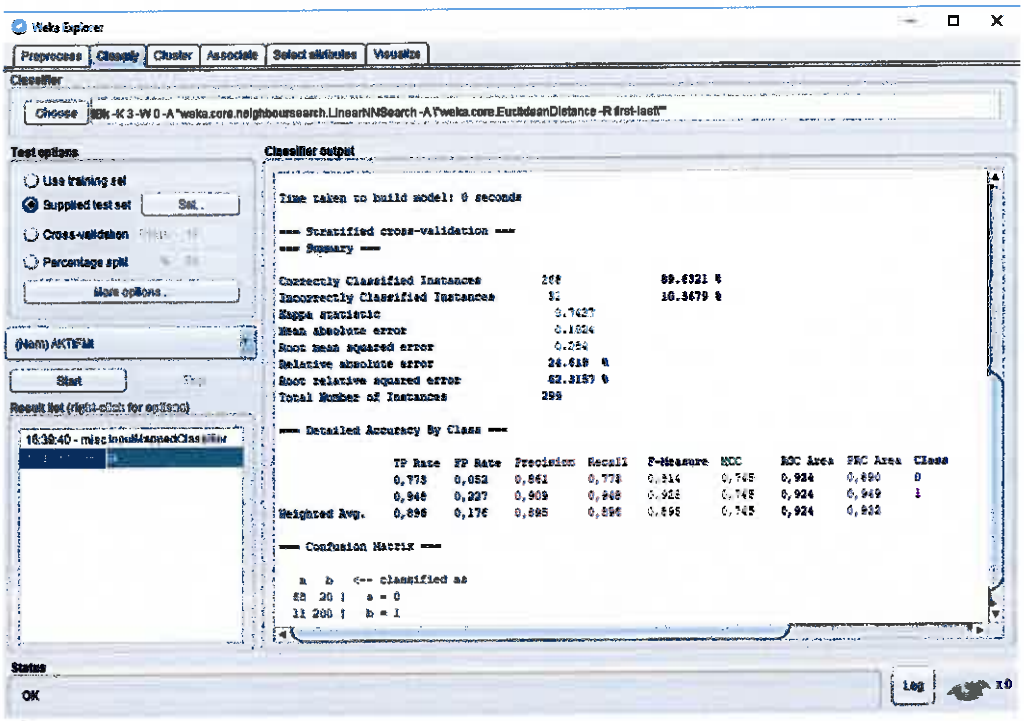
8. BULGULAR

Yapılan müşteri segmentasyonu çalışmasında müşterilerin uygun segmentlere göre sınıflandırılması amacıyla k-en yakın komşu algoritması kullanılmıştır. K-en yakın komşu algoritmasında niteliklerin birbirleriyle olan mesafelerinin doğru şekilde hesaplanması ve uygun sınıfa dahil edilmesi yapılan çalışmanın doğru sonuçlar vermesi açısından büyük önem taşımaktadır. Çizelge 8.1'de farklı *k* değerlerinin uygulanması sonucu en yüksek doğruluk oranına sahip sınıflandırma değerleri gösterilmiştir.

Çizelge 8.1. Farklı *k* değerleri ve doğruluk oranları

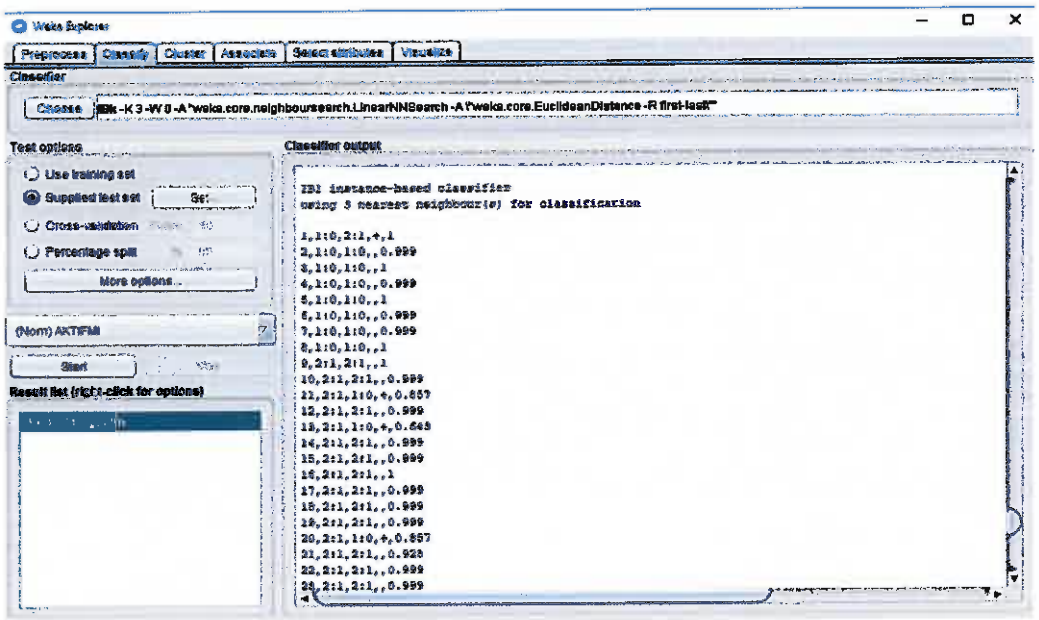
	Doğruluk Oranı		
k	*Percentage Split %66*	Cross-Validation-5-Folds	Cross-Validation-10-Folds
1	89.21	87.62	87.95
3	88.23	88.29	**89.63**
5	89.21	84.61	85.28
7	84.31	83.27	83.27
9	83.33	82.27	83.61
11	84.31	82.27	81.93
13	85.29	81.93	83.61
15	85.29	82.27	84.28
17	79.41	81.60	83.27
19	73.52	78.59	83.27
21	71.56	77.92	81.27

Çizelge 8.1'de görüldüğü gibi uygun sınıflandırma işleminin yapılması amacıyla en yüksek doğruluk oranının *k* değerinin 3 ve Cross-Validation-10-Fold yöntemiyle sağlandığı görülmektedir. Bu nedenden dolayı sınıflandırma işleminde *k* değeri 3 olarak seçilmiştir. Şekil 8.1'de K-en yakın komşu sınıflandırma yönteminin uygulanması sonucu gösterilmiştir.



Şekil 8.1. *k*=3 Değeri İçin Sınıflandırma Sonuçları

Şekil 8.1'de görüldüğü gibi sınıflandırma işleminin yapılması amacıyla K-en yakın komşu yöntemi kullanılmış ve *k* değeri 3 olarak seçilmiştir. Yapılan bu işlem sonrasında modele sunulan verilerin %89.63 oranında doğru şekilde sınıflandığı sonucu elde edilmiştir. Uygulanan sınıflandırma işleminin sonuç çıktısı Şekil 8.2'de gösterilmiştir.



Şekil 8.2. k=3 değeri için sınıflandırma işlemi sonuçları

Yapılan sınıflandırma işlemi sonrası Weka'ya eklenen veri setinin Aktif/pasif sınıfını gösteren oranlar gösterilmiştir. Bu değerlerden yola çıkılarak müşterilerinin 1 (Aktif)'e daha yakın değerler aldığı ve firmadan ayrılmayacakları görülmektedir.

Yapılan sınıflandırma işleminin doğruluğunun test edilmesi amacıyla Karışıklık Matrisi (Confusion Matrix) ile doğruluk kontrolü yapılmıştır.

Çizelge 8.2. *k*=3 Aktif/Pasif Karışıklık Matrisi Hesaplama Değerleri

Toplam Örnek	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
299	0.773	0.052	0.861	0.773	0.814	0.745	0.924	0.890	0
	0.948	0.227	0.909	0.948	0.928	0.745	0.924	0.949	1

Çizelge 8.2'deki Karışıklık Matrisi hesaplama değerleri *k*=3 şeçildiğinde elde edilen değerleri göstermektedir. Çizelge 8.2'de görüldüğü gibi 299 adetlik bir veri seti üzerinden sınıflandırma işlemi yapılmış ve bu işleme ait sonuçlar elde edilmiştir.

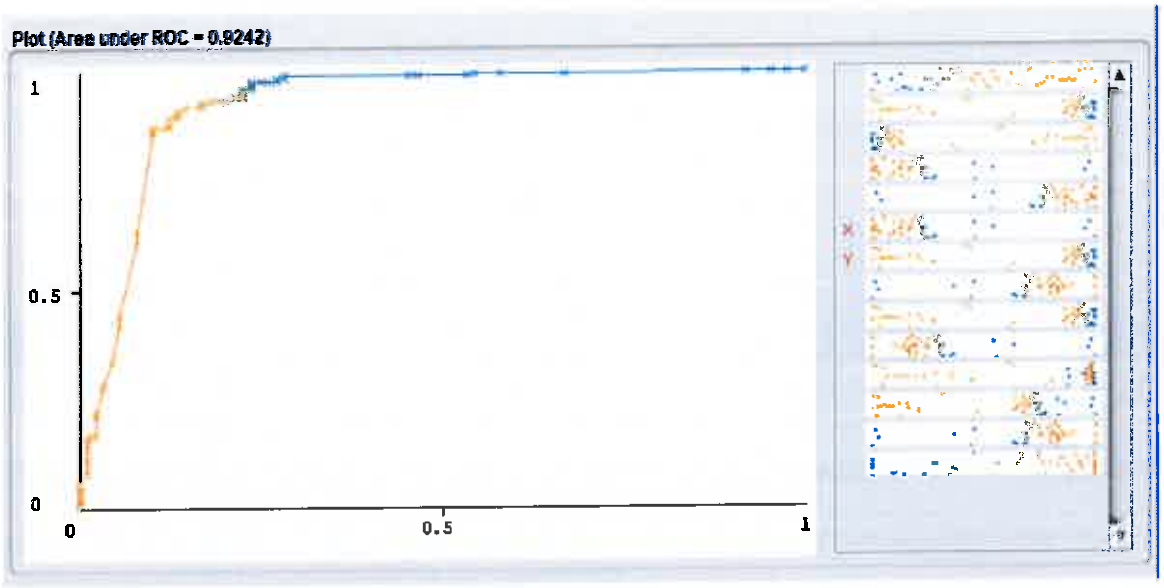
Bu değerlerden yola çıkılarak Karışıklık Matrisi hesaplanmış ve Çizelge 8.3'te gösterilen sonuçlar elde edilmiştir

Çizelge 8.3. *k*=3 için Aktif/Pasif Karışıklık Matris Sonucu

Karışıklık Matrisi		
a	b	
68	20	a=0
11	200	b=1

Çizelge 8.3'te *k*=3 şeçildiğinde elde edilen karışıklık matrisi sonucunu içeren bilgiler gösterilmiştir. Çizelge 8.3'te görülen Karışıklık Matrisine göre sınıflandırma sonucunda a=0 ve b=1 değerlerini içeren iki adet sınıf elde edilmiştir. Bu sınıflardan a=0 modele sunulan müşterilerin firmadan ayrılacağını gösterirken b=1 sınıfı ise müşterilerin ayrılmadan firmanın müşterisi olmaya devam edeceğini göstermektedir. Bu veriler doğrultusunda uygulanan modelin ayrılma eğiliminde olan 68 adet müşteriyi doğru sınıflandırdığını ancak 11 adet müşteriyi doğru sınıflandıramadığını göstermektedir. Firmada kalma eğiliminde olan 20 adet müşteri ise yanlış sınıflandırma yapılırken 200 adet müşteri doğru şekilde sınıflandırılmıştır.

Çizelge 8.3'te elde edilen değerlerin Aktif/Pasif sınıfına ait 1 değerlerini içeren ROC eğrisi Şekil 8.3'te gösterilmiştir.



Şekil 8.3. k=3 değeri için ROC eğrisi

Şekil 8.3'te gösterilen ROC eğrisinde görüldüğü gibi roc cizgisinin dik bir şekilde 1'e yakın bir şekilde olması yapılan sınıflandırma işleminin başarısının yüksek olduğu görülmektedir.

9. TARTIŞMA

Yapılan bu çalışma kapsamında veri madenciliği, ver madenciliği süreçleri ve bu süreçlerin telekomünikasyon sektöründen elde edilen müşteri verilerine uygulanmasın hakkında bilgi verilmiştir.

Literatür taraması sonrasında elde edilen veriler ile çalışma sonunda elde edilen sonuçlar farklılıklar göstermektedirler. Örneğin Turkcell firması müşterilerinin internet kullanımlarını gece ve gündüz saatlerine göre değerlendirerek gece ve gündüz tarifeleri sunarken, TurkNet firması müşterilerin toplam internet kullanımlarına göre değerlendirmiş ve kullanım limitleri erken biten müşteriler için Adil Kullanım Noktası'nı (AKN) kaldırmıştır ("BTK 2015 Pazar Verileri", (Erişim: 19.02.2019), www.btk.gov.tr). Bu farklılıkların sebebi yapılan her çalışmanın kendine özgü veri seti üzerinden kendine özgü parametreler kullanılarak yapılmış olmasından kaynaklanmaktadır. Ancak yapılan çalışmaların telekomünikasyon sektörü üzerine yapılmış çalışma olması nedeniyle ve genel olarak sınıflandırma algoritma yöntemleri kullanılması nedeniyle birbirleriyle benzerlik kurulabilir. Yapmış olduğumuz çalışmada k-en yakın komşu algoritması yöntemi ile sınıflandırma yapılmış ve bu algoritmaya uygulanan k değerleri farklı değerler için hesaplanarak sınıflandırmalar hesaplanmıştır. K-en yakın komşu algoritması ile oluşturulan model genel anlamda benzer sonuçlar vermektedir. k değerleri arttıkça modelin doğru şekilde kümeleme yapması önce giderek azalmış sonra azalmıştır.

10. SONUÇ ve ÖNERİLER

Firmaların sahip oldukları müşterilere kaliteli ve uygun bir hizmeti sunmaları başta telekomünikasyon sektörü olmakla birlikte tüm diğer sektörler içinde en önemli konuların başında gelmektedir. Bu nedenden dolayı firmalar sahip oldukları müşterilerin kullanım alışkanlıkları ve tercihlerine uygun hizmet ve kampanyalar sunarak müşteri memnuniyetini arttırıcı yada şikayetlerini giderici çalışmalar yapmaları çok önemli bir konu haline gelmiştir. Bu noktada öneri ve dilek sistemleri kullanıcılara her türlü nesnelerin sunulmasını sağlamıştır. Turkcell iş zekası ve veri madenciliği tekniklerini kullanarak müşteri bilgilerini inceledikten sonra müşterilerine yeni tarife ve hizmetler sunmakla birlikte mevcut müşterileri için yeni kampanyalar geliştirmektedir (Yapay Sinir Ağları, Anonim, b.t.). TurkNet firması ise yapılan müşteri segmentasyonu çalışması sonrasında Turkcell firmasında olduğu gibi müşteri memnuniyetini arttırıcı çalışmalar yapabileceği ve ek olarak aylık adil kullanım kotalarındaki sınırlamalarında kaldırılarak daha iyi bir müşteri memnuniyeti elde edebilecektir.

Yapılan bu tez çalışmasında telekomünikasyon sektörü müşterileri kullanım alışkanlıkları ve diğer özellik kriterleri detaylı şekilde incelenmiştir. Daha sonra ise müşteri özelliklerine göre segmente edilmesi amacıyla detaylı bir şekilde literatür araştırması yapılmış ve bu araştırma sonrasında müşteri segmentasyonuna yönelik çalışmalar özetlenerek çeşitli kriterlerde değerlendirilme işlemi yapılmıştır. Yapılan çalışmada müşterilerin yaş kriterleri, meslek bilgileri, kullanım oranları gibi kriterler parametre olarak kullanılmıştır. Belirtilen bu parametreler ile birlikte yapılan tez çalışması, müşterilerin kullanım alışkanlıkları ve özelliklerine göre çeşitli gruplara segmente edilerek müşterilere daha kaliteli ve hizmet sunmaya yönelik katkılar sağlanması amaçlanmıştır.

Yapılan tez çalışmasında telekomünikasyon sektörünün öncü kuruluşlarından olan bir firmanın müşteri verileri K-en yakın komşu algoritması yöntemi ile segmentasyon işlemine tutulmuştur. Bu bağlamda müşteri verilerinin her biri için K-

en yakın komşu algoritmasına ait skorları hesaplanmıştır. Her bir müşterinin kullanım alışkanlıkları bilgileri skorları belirlenerek K- en yakın komşu algoritması yöntemi kullanılarak kümelere ayrılma işlemi yapılmıştır. Bu çalışmas sonrasında ise müşteriler firmadaki değerleri yönünden 3 kümeye ve firmadan ayrılma eğilimlerine göre ise 2 kümeye ayrılmaları şeklinde karar verilmiştir. Mevcut müşteri verilerinin yüzde 30'luk kısmı test verisi olarak kullanılmıştır. Bu kapsam doğrultusunda müşteri verilerinden rastgele seçilen bir müşterinin dahil olabileceği müşteri kümesi belirlenmiştir. Bu tespit sonrasında ise müşterinin dahil olduğu kategöri ile ilgili müşteri memnuniyetini arttırmak amacıyla şikayet oranlarının azaltılması ve uygun kampamya ve hizmetlerin sunulmasına yönelik öneri listeleri hazırlanmıştır.

Çalışmanın bir sonraki aşamasında ise yapılan çalışma daha da genelleştirilerek toplam 300 adet müşteri üzerinde uygulanmuştır. Yapılan çalışmalar sonucunda ise firmanın sunmuş olduğu ürün ve hizmetler ile kıyaslanmış ve analiz edilmiştir. Yapılan bu çalışmaların performasını değerlendirmek amacıyla ise confusion matrix kullanılmıştır. Aynı zamanda bir diğer parametre olarak müşterilerin kullanım miktarları değerlendirme kriterlerine eklenmiştir.

Yapılan bu çalıma doğrultusunda gelecekte yapılacak olan çalışmalarda daha büyük boyuttaki veri setlerine uygulanarak müşteri segmentasyonuna yönelik daha sağlıklı bir şekilde uygulanması işlemi gerçekleştirilebilir.

Bunların yanında çalışma ve kurulan modelin geliştirilmesi adına kümeleme aşamasında müşterilere ait farklı bilgiler parametre olarak kullanılabilir. Örneğin müşterilerin ekonomik ve sosyo kültürel statüsü, müşterilerin medeni durumları, müşterilerin internette ziyaret ettikleri web siteleri veya müşterilerin kullanım zaman dilimleri gibi farklı bilgiler parametre olarak kullanılabilir. Bu hususta dikkat edilmesi gereken konu oluşturulacak müşteri kümelerinin adetlerinin yönetilebilir sayıda ve anlamlı bir seviyede tutulmasıdır.

Yapılan çalışmanın farklı bir yönden geliştirilebileceği bir diğer hususta müşteri verilerinin farklı algoritmalar ve farklı yöntemler ile denenerek daha iyi seviyede sonuçlar elde edilmesi sağlanılabilir. Ancak uygulama veri setindeki veri miktarı arttıkça her bir veri için artan sayıda parametre olması hesaplama kombinasyon sayılarının artmasına neden olacağında hesaplama yükünün ve zaman değerlerinin artmasına neden olacktır. Bu bağlamda ise günümüzde yeni teknolojiler olan bulut teknolojisi ve farklı hesaplama tekniklerinin yaygınlaşması ve kullanılması yapılacak olan çalışmaların kolaylaşmasını sağlayacaktır.

11. KAYNAKLAR

Aaker, D.(1971). Multivariate Analysis in Marketing: Theory and Application, Wadswort Publishing, California.

A. Backiel, Y. Verbinnen, B. Baesens, and G. Claeskens, "Combining local and social network classifiers to improve churn prediction," Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, Paris, France, 2019, pp. 651-658.

Acungil, M. (b.t.) Veri madenciliği. 20.06.2019, http://mustafaacungil.blogspot.com/

Agrawal, R., Imielinski, T., Swami, A. (1993). Mining association rules between sets of items in large databases. ACM SIGMOD Conference on Management of Data, Washington.

Akbulut, S. (2006). Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu, Yayınlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi

Akpınar, H. (Nisan 2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İ.Ü.İşletme Fakültesi Dergisi, Sayı:1, İstanbul

A. Amin, C. Khan, I. Ali, and A, Anwar, "Customer churn prediction in telecommunication industry: with and without counter-example," Mexican International Conference On Artificial Intelligence, Tuxtla Gutiérrez, Mexico, 2019, pp. 206-218.

Aytekin, G. (2002). Perakendecilik Sektöründe Veri Ambarı Uygulamaları Üzerine Bir Araştırma. Cilt 5, Sayı 17 .

Bergeron, B. (2002). Bioinformatics Computing, Prentice Hall PTR, U.S.A.

Berkhin, P. (2002). Survey of Clustering Data Mining Techniques, California, U.S.A.

Berry, M., Linoff, G. (2000). The Art and Science of Customer Relationship Management, Wiley Computer Publishing, U.S.A.

Bozkır, A.S., Gök, B., Sezer, E. (2008). Üniversite Öğrencilerinin İnterneti Eğitimsel Amaçlar İçin Kullanmalarını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti, BUMAT 2008: Bilimde Modern Yöntemler Sempozyumu.

BTK.(2015),www.btk.gov.tr/File/?path=ROOT%2F1%2Fdocuments %2FSayfalar%2FPazar_ Verileri%2F2015-Q3_v1.pdf , 19.02.2019

Chen, M.S., Han, J., Yu, P. S. (1996). Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering,8(6)

C. Müssel, L. Lausser, M. Maucher, and H.A. Kestler, "Multi-objective parameter selection for classifiers," Journal of Statistical Software, vol. 46, no. 5, 2012.

Duran, B.S. and P.L. Odell (1974). Cluster Analysis (Lecture Notes in Economics and Mathematical Systems, Econometrics; Managing Editors: M. Beckmann and H.P. Kunzi). Springer-Verlag: New York

Duran, M. (2002). Veri Tabanı Pazarlama. 27.06.2019. www.danismend.com

D. Yağan. (2019, 18 Haziran). Hanehalkı bilişim teknolojileri kullanım araştırması [Online]. Erişim: http://www.tuik.gov.tr/PreHaberBultenleri. do?id=21779.

Everitt, B. (1974). Cluster Analysis. Heinemann Educational Books Ltd, London. ETL Tools-METAspectrumSM Evaluation. (b.t.). metinler.19 Aralık 2018, http://www.sas.com/offices/europe/czech/technologies/enterpriseintelligenceplatf orm/MetagroupETLmarket.pdf

Fayyad, U. (March 1998) Mining Databases: Towards Algorithms for Knowledge Discovery, IEEE Bulletin of the Technical Committee on Data Engineering, Vol.21 No1

Flexer, A. (2001). On the Use of Self-Organizing Maps for Clustering and Visualization. Intelligent Data Analysis 5.

Harrold,D. (2000), What.s Your Data Telling You?,Control Engineering http://www.erpcrm.com/crm_anasf/crm_mimarisi.htm. 2005

Inmon, W.H. (2002). Building the Data Warehouse. Wiley, U.S.A. İnternet Terimleri Sözlüğü.(b.t.). metinler. 19.03.2019, http://www.ttnet.com.tr/web/208-997-1- 1/tr/ttnet/internet_terimleri_sozlugu/_internet_terimleri_sozlugu

Johnson A. R., Wichern W. D. (1998). Applied Multivariate Statistical Analysis. Prentice Hall, U.S.A.

Kiang M.Y., Kumar, A. (2001). An Evaluation of Self Organizing Map Networks as a Robust Alternative to Factor Analysis in Data Mining Applications. Information Systems Research, Vol.12. No.2.

Karakaş, M. (b.t.) Veri Ambarları Genel Yapısı, 07.03.2019, http://www.bilgiyonetimi.com

Gray, P., Watson, H. (1998). Decision Support in the Data Warehouse, Prentice Hall PTR, New Jersey.

Internet: Vikipedia https://tr.wikipedia.org/wiki/Cascading_Style_Sheets

W. Hadley, C. Winston. (2019, 30 Mart). KMggplot2 [Online]. Erişim: https://cran.r-project.org/ packages/RcmdrPlugin.KMggplot2.

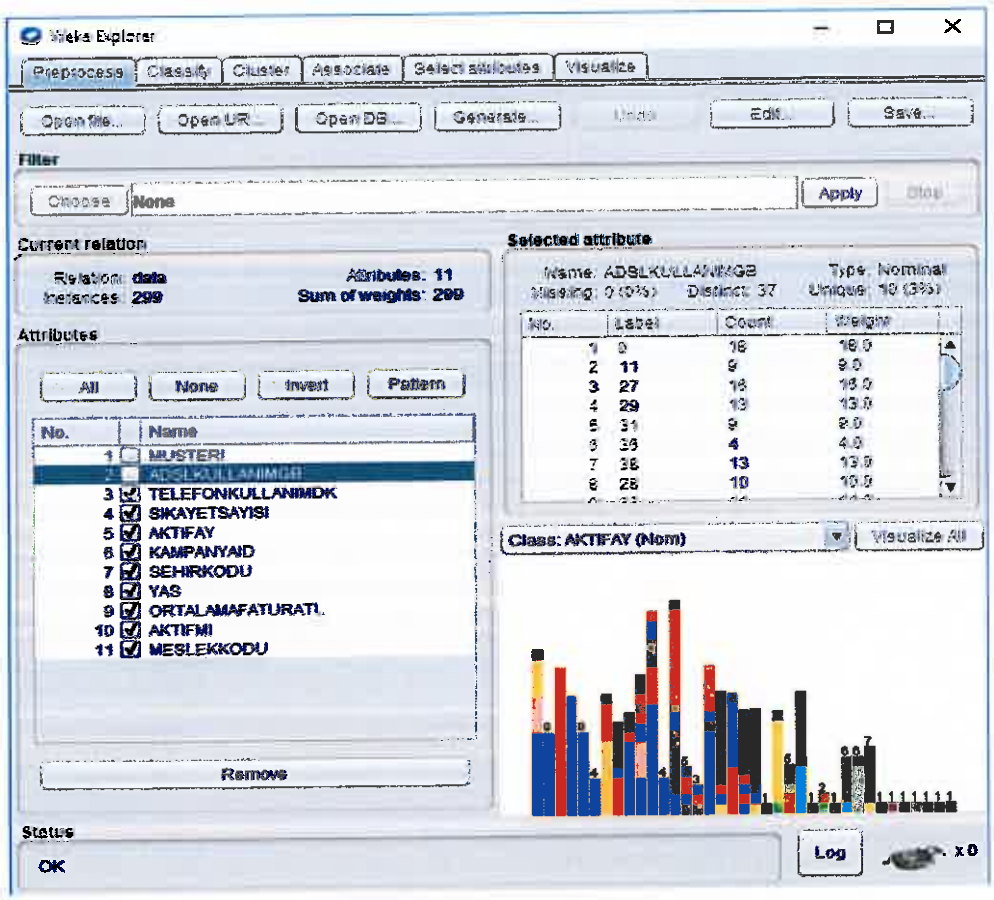
T. Therneau, B. Atkinson, B. Ripley. (2019, 24 Mayıs). Package rpart [Online]. Erişim: https://cran.r-project.org/web/packages/rpart/rpart.

Z. Martinasek, J. Hajny, and L. Malina, "Optimization of power analysis using neural network," presented at International Conference on Smart Card Research and Advanced Applications, Berlin, Germany, 2019.

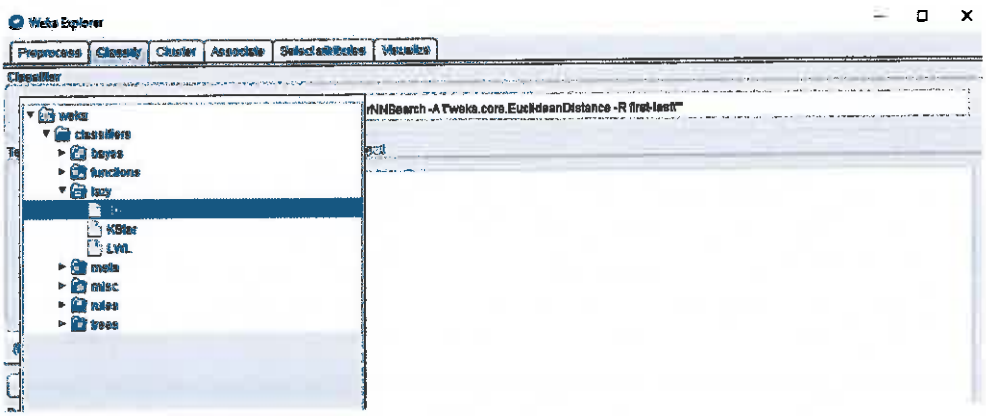
Weka- Hakkında- Weka Nedir? https://ceaksan.com/tr/weka-nedir/ 27.06.2019

12. EKLER

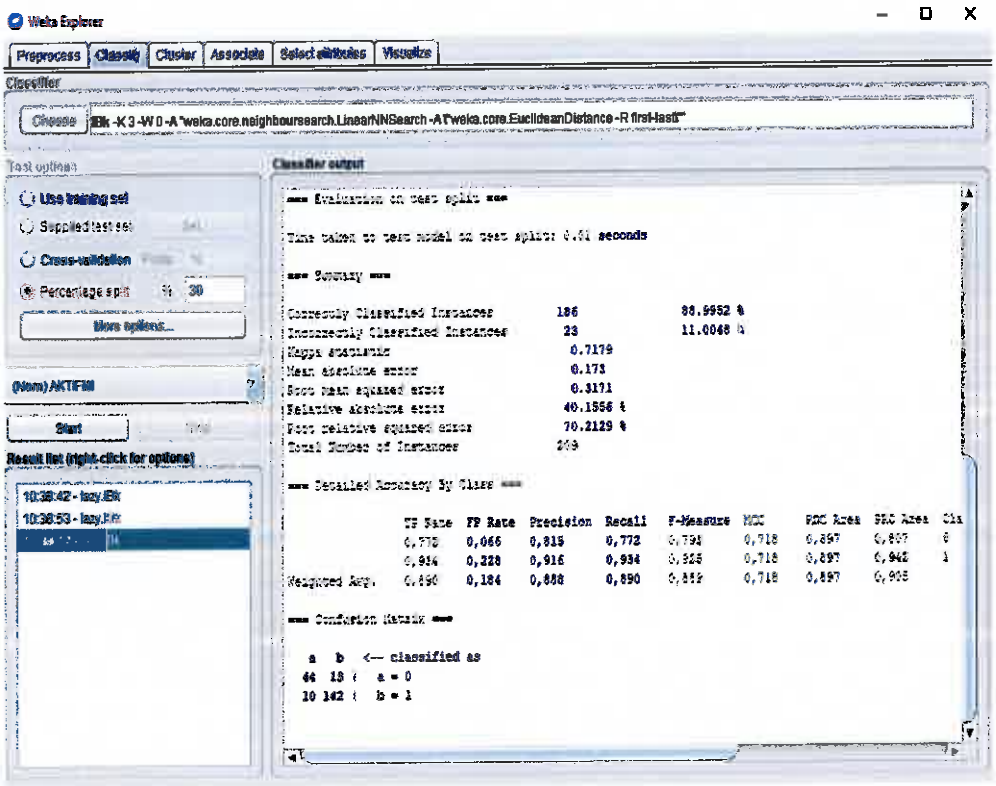
EK 1: Müşteri Verilerinin Wekaya Aktarılması



EK 2: K-en Yakım Komşu Algoritmasının Wekada Seçilmesi



EK 3: Farklı k Değerlerinin Modele Uygulanması ve Başarı Oranları



13. ÖZGEÇMİŞ,

1988 yılında Kastamonu'nun Küre ilçesinde doğdu. İlkokulunu köyünde bulunan Kayadibi Köyü Sence Mahallesi İlkokulu'nda tamamladı. Ortaokulunu Seydiler Yatılı İlköğretim Bölge Okulunda, lise öğrenimini ise Kastamonu Anadolu Ticaret Meslek Lisesinde tamamladı. Lisans eğitimini Mersin Üniversitesi Bilgisayar Teknolojisi ve Bilişim Sistemleri bölümünde tamamladı. Halen T.C. Haliç Üniversitesi Lisansüntü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim dalı Tezli Yüksek Lisans öğrencisi olarak devam etmektedir. 2011 yılında atılmış olduğu iş hayatında sırasıyla Hobim Bilgi İşlem Hizmetleri A.Ş. firmasında Raporlama Uzmanı, TurkNet İletişim Hizmetleri A.Ş firmasında iki yıl Raporlama Uzmanı üç yıl İş Zekası ve Raporlama Uzmanı olarak çalıştı. Bugün ise Marport Liman İşletmeleri Sanayi ve Ticaret A.Ş. firmasında İş Zekası Uzmanı olarak iş hayatına devam etmektedir.