**T.C.**
**İSTANBUL AYDIN ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE MÜŞTERİ KAYBI ANALİZİ

**YÜKSEK LİSANS TEZİ**

**Buket ÖNAL**

**(Y1513.010013)**

**Bilgisayar Mühendisliği Anabilim Dalı**

**Bilgisayar Mühendisliği Programı**

**Tez Danışmanı: Doç.Dr. Metin ZONTUL**

**Ekim, 2017**

T.C.
İSTANBUL AYDIN ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜ

## Yüksek Lisans Tez Onay Belgesi

Enstitümüz Bilgisayar Mühendisliği Ana Bilim Dalı Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı **Y1513.010016** numaralı öğrencisi **Buket ÖNAL** 'ın **"MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE MÜŞTERİ KAYBI ANALİZİ"** adlı tez çalışması Enstitümüz Yönetim Kurulunun 12.09.2017 tarih ve 2017/20 sayılı kararıyla oluşturulan jüri tarafından oybirliği ile Tezli Yüksek Lisans tezi olarak kabul edilmiştir.

| Öğretim Üyesi Adı Soyadı | İmzası |
|---|---|
| **Tez Savunma Tarihi : 13/10/2017** | |
| **1)Tez Danışmanı:** Doç. Dr. Metin ZONTUL | ..... |
| **2) Jüri Üyesi :** Prof.Dr. Ali GÜNEŞ | ..... |
| **3) Jüri Üyesi :** Yrd. Doç. Dr. Ferdi SÖNMEZ | ..... |

Not: Öğrencinin Tez savunmasında **Başarılı** olması halinde bu form **imzalanacaktır**. Aksi halde geçersizdir.

## YEMİN METNİ

Yüksek Lisans tezi olarak sunduğum “Makine Öğrenmesi Yöntemleri ile Müşteri Kaybı Analizi” adlı çalışmanın tezin projesi safhasından sonuçlanmasına kadarki bütün süreçlerde bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurulmaksızın yazıldığını ve yararlandığım eserlerin Bibliyografya ‘da gösterilenlerde oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve onurumla beyan ederim. (13/10/2017)

**Buket ÖNAL**

***Aileme,***

## ÖNSÖZ

İlk olarak tez çalışmamın hazırlanmasında her türlü yardımı esirgemeyen ayrıca değerli görüş ve yorumları, rehberliği ve desteği için değerli danışman hocam sayın Doç.Dr. Metin ZONTUL'a teşekkür ederim. Bu tez çalışma süresi boyunca sabrı, anlayışı ve desteği ile bana yardımcı olan annem, babam ve kardeşime sonsuz Teşekkür ederim.

**Ekim 2017** **Buket ÖNAL**

**İÇİNDEKİLER**

**Sayfa**

**KISALTMALAR**

| | |
|---|---|
| **CART** | :Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees) |
| **CHAID** | :Ki-kare Otomatik Etkileşim Algılama (Chi-squared Automatic Interaction Detection) |
| **COCOMO** | :Yapıcı maliyet modeli (Constructive Costing Model) |
| **DVM** | :Destek Vektör Makineleri (Support Vektör Machines) |
| **GUI** | :Grafiksel Kullanıcı Arayüzü (Graphical User Interface) |
| **PUK** | : Pearson VII kerneli |
| **RA** | :Regresyon Analizi |
| **ROC** | :Alıcı İşletim Karakteristiği (Reciever Operating Characteristic) |
| **SMO** | :Sosyal Medya Optimizasyonu |
| **VM** | :Veri Madenciliği (Data Mining) |
| **QUEST** | :Hızlı, Tarafsız, Verimli İstatistiki Ağaç (Quick, Unbiased, Efficient Statistical Tree) |
| **WEKA** | :Bilgi Analizi için Waikato Ortamı (Waikato Environment for Knowledge Analysis) |

## ÇİZELGE LİSTESİ

**Sayfa**

## ŞEKİL LİSTESİ

**Sayfa**

# MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE MÜŞTERİ KAYBI ANALİZİ

## ÖZET

Günümüzde teknoloji geçmişe nazaran hızla gelişme göstermekte, artan ihtiyaçlara cevap verecek nitelikte fayda sağlamaya çalışmaktadır. Büyük veriler elektronik ortamlarda saklanmakta, kalabalıklaşan nüfusla birlikte saklanan verilere daha erken ulaşma ihtiyacı başı çekmektedir. Veri madenciliği sınıflandırma algoritmaları kullanılarak sağlanan çalışmalarla kalabalık verilerin gruplanması yapılmakta, bu veriler etiketlenerek gerektiğinde çok kısa bir süre içerisinde karşımıza çıkabilmektedir. Günümüzde gelişme gösteren hemen hemen her sektörde olduğu gibi lojistik sektöründe de rekabet koşulları hızla artmaktadır. Gelişmiş ülkelerin çoğunun entegre olduğu ve her geçen gün hızla gelişen lojistik sektörü ülkemizde 1980' li yıllarda hizmet vermeye başlamış, 1990' lı yıllarda tam anlamıyla rayına girmiştir. Bütün dünyada hızla gelişen bu sektör için müşteri sadakati çok büyük bir öneme sahiptir. Bu nedenle müşteri kayıplarını minimize etmek için müşteri ilişkileri yönetimine daha fazla önem verilmesi gerekmektedir. Artan e-ticaret sektörü ile birlikte daha fazla müşteri potansiyelinin yükseldiği görülmektedir. Aynı zamanda bununla birlikte lojistik sektörüne bir çok yeni firma dahil olmuştur. Bu koşullar doğrultusunda mevcut müşterileri elde tutmak, başka firmaya geçme eğilimi gösteren müşterileri tespit etmek önem kazanmıştır. Müşterilerin şirketten beklentilerini anlamak ve firmanın davranışlarını daha iyi takip etmek, buna yönelik stratejiler geliştirmek sektörde tutunabilmenin temeli haline gelmiştir.

Kaybedilen müşterilerin geç tespit edilmesi, maliyetlerin artmasına sebep olmaktadır. Yeni müşteri kazanımı var olan mevcut müşterilerin elde tutulmasından daha fazla süre ve daha fazla maliyet gerektirmektedir. Buna bağlı olarak müşterilerin sergiledikleri davranışlar dikkate alınarak elde edilen veriler modellenerek, iptal eğilimi gösteren müşterilerin tespiti sağlanabilmektedir.

Veri madenciliği büyük veri setleri içerisinde gizli kalmış anlamlı bilgiyi ortaya çıkarma faaliyetidir. Veri madenciliği hem klasik istatiksel yöntemleri hem de makine öğrenmesi yöntemlerini kullanabilir. En çok kullanılan alanlardan biri müşteri kaybı analizinde müşterileri segmentasyonlara ayırarak kaybedilecek müşterileri tahmin etmektir. Bu çalışmada, Türkiye' de faaliyet gösteren bir lojistik firması ile çalışan müşterilerin geçmiş iki yıldaki gönderi bilgileri incelenerek, kaybedilmiş müşteri davranışları ortaya çıkarılmaya çalışılmıştır.

Veri madenciliği prensiplerine uygun olarak hazırlanan veriler üzerinde makine öğrenmesi algoritmalarından biri olan destek vektör makinesi (DVM) algoritması üzerinde uygulanmıştır. Firmadan alınan veriler içerisinden 2.000 adet müşteri uygulamamızda kullanılmıştır. Doğrusal olmayan bu veriler için en uygun sınıflandırma yöntemi DVM algoritması tercih edilmiştir. Bu müşterilerin geçmiş iki yılına ait veriler 3'er aylık dönemlere ayrılmıştır. Toplam 8 çeyrek üzerinde müşteri kaybı analizi yapılarak, firmadan ayrılma eğilimi gösteren müşterilerin gelecek üç aydaki kayıp analiz tahminlemesi yapılmaya çalışılmıştır.

**Anahtar Kelimeler**: *Müşteri Kaybı Analizi, Veri Madenciliği, Sınıflandırma, Destek Vektör Makinası.*

## CHURN ANALYSIS WITH MACHİNE LEARNING ALGORITHMS


## ABSTRACT

Technology develops faster than the past in our time and tries to provide benefits in order to meet the increasing needs. Larger data are stored in electronic media, and the need to reach the stored data as quicker as possible is of great importance in an overcrowded population environment. Immense data are grouped by using data mining and classification algorithms which these data are labeled so that we can find in a very short time when necessary.

The competition conditions are increasing rapidly in the logistics sector in a manner similar to nearly all developing sectors of today. The logistics industry, which is rapidly integrated by many developed countries, has started to deliver services in our country in the 1980s, and has been fully adapted in the 1990s. Customer loyalty for this fast-growing sector means tremendous importance all over the world. That is why greater emphasis should be placed on customer relationship management to minimize customer losses. With the increasing share of e-commerce sector, it is observed that more and more customer potential is in rise. At the same time, many new companies have joined the logistics sector. Under these conditions, it has become vitally important to retain existing customers and to identify customers who tend to move another companies. It has become the basis for standing in the sector to be able to understand the expectations of customers from the company, to follow up the company's attitudes better and accordingly to develop strategies in line with these findings.

Late awareness about lost customers results in cost increases. New customer acquisition requires more time and more costs than keeping existing customers. Accordingly, the data obtained by taking into consideration the behaviors exhibited by the customers can be modeled and the customers who have a tendency to escape can be determined.

Data mining is the activity of revealing meaningful information that is hidden in large data sets. Data mining can use both traditional statistical methods and machine learning methods. One of the most used areas is to estimate the customers that will be lost by segmenting the customers in customer loss analysis. In this study, we tried to reveal lost customer behaviors by analyzing the shipping information of past two years about customers of a logistics company operating in Turkey.

It is applied on the support vector machine (SVM) algorithm which is one of the machine learning algorithms on the data prepared according to the principles of data mining. 2.000 customers among the data received from the company have been used in our application. Support vector machine algorithm, being the most suitable classification method, is used for this nonlinear data. The data for the past two years of these customers are divided into quarterly periods. Customer loss analysis was conducted on a total of 8 quarters and loss analysis estimation for the customers who tend to leave the company was made for the next three months.



**Keywords:** *Churn Analyisis, Data Mining, Classification, Support Vector Machine.*

# 1 GİRİŞ

Bugünler de çeşitli iş alanlarında faaliyet gösteren birçok şirket küreselleşme ve artan rekabet koşulları karşısında mevcudiyetlerini sürdürebilmek için daha fazla mücadele etmek zorunda kalmışlardır. Bu rekabet ile müşteri satın alma davranışları ve hizmet beklentileri büyük ölçüde değişmiştir. İşletmeler de bu isteklere cevap verebilmek için farklı yöntemlere başvurmak durumunda kalmışlardır. Bu yöntemlerin doğru bir etki yaratabilmesi için işletmelerin değişen müşteri davranışlarını izleyebilmesi, işletme amaç ve hedeflerini oluşturabilmesi ve müşteri ilişkileri yönetimini doğru bir şekilde sağlayabilmesi ile mümkün olabilecektir. Müşteri merkezli bir anlayış ile müşteriye değer katmak ve müşteri bağlılığını sağlamak işletmenin bu doğrultuda yararına olacaktır.

Diğer taraftan bilişim sektöründeki gelişmeler sayesinde işletmeler, birçok veriyi saklayabilir, kolay erişebilir, bu verileri işleyebilir ve işlenmiş verileri anlamlı bilgilere dönüştürebilme imkânına sahip olabilmektedirler. İşletmeler, müşterilerin satın alma davranışlarının yanı sıra müşteriye ait çeşitli özellikleri de bilgi depolama alanlarında saklamaktadırlar. Kayıt altına alınan veriler sayesinde veri madenciliği teknikleri kullanılarak veri yığınları içerisindeki anlamlı ve gizli bilgiler tespit edilip ortaya çıkarılarak işletme için yararlı bilgiler elde edilebilmektedir. Elde edilen veri madenciliği sonuçları, müşteri odaklı birçok uygulamanın geliştirebilmesine ve müşteri kaybı için önlemler alınmasına ilişkin fikir sunmaktadır.

Veri madenciliği, günümüzde hemen hemen her sektörde elde edilen yüksek düzeydeki veri guruplarında gizli halde var olan örüntü ve yatkınlıkları ortaya çıkarma, bu verilerden yararlı örüntüler çıkarma sürecini içeren bir teknolojidir. Birçok veri madenciliği yöntemi arasından ulaşılmak istenilen bilgi ve şekline bağlı olarak en uygun tercih edilerek uygulanmaktadır.

Veri madenciliği birçok alanda uygulanabilmektedir. En yaygın kullanılanlardan biri, müşteri bölümleri oluşturulup rekabet içerisinde olunan firmaya geçme

eğilimi gösteren şahıs veya firma profillerini ortaya çıkarma uygulamasıdır. Müşteri segmentasyonu, işletmenin hitap ettiği pazar payında benzer karakteristikler taşıyan müşterileri gruplandırma işlemi olarak tanımlanır. Her müşteri grubu farklı davranış özelliklerine sahiptir. Bu gruplama işletme için benzer davranışlı müşteri gruplarına özgü öneriler ve pazarlama uygulamaları sunmasını sağlar. İşletmelerin veri madenciliği kullanımlarının asıl gayesi büyük fayda sağlayan müşterileri kaybetmeme isteğidir. Kaybedilme eğilimi gösteren müşteriler belirlendikten sonra bahsi geçen müşterinin bulunduğu segmentasyon da dikkate alınarak çeşitli pazarlama stratejileri oluşturmak mümkündür.

## 1.1 Problem Tanımı

İşletmelerin bulundukları pazar payında bulunan yerlerini koruma, aynı zamanda yeni müşteriler kazanma gerekliliği mevcuttur. Her işletme yıllık olarak geçmiş verileri göz önüne alarak değerlendirme yapar ve yeni yıldaki planlarını buna göre oluşturur. Kazanılan yeni müşterilerin yanı sıra kaybedilen müşterilerin de analizlerinin yapılması alınacak önlemlerde önemli rol oynamaktadır.

E-ticaret sektörünün gelişmesi ile birlikte kargo şirketlerinin piyasadaki mali değerleri artmıştır. Bununla birlikte pazara birçok yeni kargo taşımacılığı yapan firma da eklenmiştir. İşletmeler yapılan değerlendirmeler için geçmiş bir yılın verileri dikkate almaktadırlar. Bu nedenler yeni kazanılan müşteriler için müşteri kaybı analizi yaparken yaşadıkları en büyük zorluklardan birisi yeterli bilgi kaynağının, yani veri setinin olmayışıdır. Veri setinin az olması doğru ve sağlıklı müşteri eğilimi tespitinde yanılmalara sebep olmaktadır. Kısa süreli de olsa elimizdeki verilerle başka bir firmaya geçme eğiliminin tespiti edilmesi gerekmektedir. Aynı şekilde uzun süredir çalışılan yüksek karlı müşterilerin kaybetme eğilimin tespiti için 1 yıl gibi uzun bir sürenin beklenmesi yıl içinde kaybedilmesine bunun görülemeyip önlem alınamaması da ayrı bir problem teşkil etmektedir.

## 1.2 Çalışmanın Amacı

Müşteri kaybı analizinde yaşanan problemler dikkate alındığında ve daha önce yapılan müşteri kaybı analizi çalışmaları dikkate alındığında bu çalışmanın amacı lojistik sektöründe kısa süreli müşteri kaybı analizi yapmaktır. Bunu yaparken öncelikle elimizdeki veri seti, anlam ifade etmeyen ve önemli olmayan verilerden temizlenecektir. Daha sonra veri madenciliği sınıflandırma yöntemi kullanılarak müşteriler üçer aylık gönderi miktarları ve toplam navlun ücretleri dikkate alınarak, önceki kaybetme eğilimlerine de bağlı olarak sınıflandırılacaktır. Destek vektör makineleri algoritmasından yararlanılarak müşterinin önümüzdeki üç aylık dönemde kaybetme eğilimi hesaplanmaya çalışılacaktır.

# 2 LİTERATÜR

Tosun (2006), Yapı Kredi bankasının kredi kartı kullanmayı bırakan müşterilerin analizini yapmıştır. Karar ağacı kullanarak farklı kurallar doğrultusunda eşik değerlerini belirlemiş ve incelemelerde bulunmuştur. Bu incelemeyi yaparken sonuçları olumsuz etkileyen kuralları sonradan budama yöntemi ile elemiştir.

Akbulut (2006), bir kozmetik markasının verilerini baz alarak inceleme yapmıştır. WEKA kullanarak kümeleme ve sınıflandırma yapmış ve sınıflandırma algoritmalarından verileri için J.48 karar ağacının uygun olduğunu söylemiştir. Bu algoritmayı kullanarak müşteri davranış modelini geliştirmiştir. Elde edilen sonuçlar doğrultusunda müşteriler alışveriş yaptıkları tutarlar doğrultusunda iki küme olacak şekilde ayrılmış ve bu müşterilerin karşı firmalara geçmesindeki tercih sebepleri göz önüne alınarak önermelerde bulunulmuştur.

Özmen (2006), Telekom sektöründe kontörlü bir hat için ilgili hattın bakiye eklendikten sonra 6 aylık dönem içerisinde tekrar ekleme yapılmazsa, bu hat sahibinin sözleşmesinin sonlandırılmasından yola çıkarak müşterilerin ayrılma olasılığını hesaplama üzerine çalışmıştır. Karar ağaçları ve regresyon modellemesi kullanmıştır. Çağrı merkezine yapılan şikâyet telefonları en önemli nitelik olarak görülmüştür.

Asilkan (2008), ikinci el otomobil fiyatlarını yapay sinir ağları algoritması yardımı ileriki döneme ait fiyatları tahmin etmeye çalışmıştır. Zaman serisi analiz yöntemleri, sonuçların karşılaştırılmasına yardım etmiştir. Yapay sinir ağın tasarlanması kolaydır. Aynı zamanda probleme hızlı olarak uyarlanabilirler. Eldeki az miktarda veriye rağmen sonuçları başarılı çıktı üretebilmektedir. Bu nedenle otomobil sektöründe de kullanılabilirliği gösterilmiştir.

Bilgen (2009), Analitik Hiyerarşi yöntemlerinden yararlanmıştır. Model olarak elektronik bankacılık servislerini tercih etmiştir. Müşteri kullanım bilgisini ile

birlikte beklentilerini öngören bir yapı kurmuştur. Yapılan çalışma sonucunda müşteri kaybetmesine yönelik öngörülerde bulunulmuştur.

Kişioğlu (2009), telekomünikasyon sektöründe iptal analizi üzerine çalışmıştır. Bu çalışmada Bayes ağları metodu ile sözleşmesini iptal etme eğilimi gösteren abonelerin davranış özellikleri modellenmiştir. Bayes ağ sistemleri kesikli değişkenler kullanmaktadır. Buna bağlı olarak sürekli değişkenleri kesikli hale getirebilmek için CHAID karar ağacı algoritmasından yararlanılmıştır. Korelasyon analizi ve eş doğrusallık testi yapılarak değişkenler arasındaki ilişki gösterilmiştir. Elde edilen sonuçlar ile birlikte uzman görüşleri alınmıştır. Bunlar birleştirilerek bir nedensel harita oluşturulmuş. Önemli değişkenler, ortalama konuşma süreleri, fatura tutarları ortalaması, ortalama diğer operatörleri arama sıklığı ve tercih ettikleri tarife türünü değiştirme veya iptal etmesi olarak belirlenmiştir. Çalışma sonucunda üç farklı senaryo elde edilmiştir. Abonelerden tarife türünü iptal edenlerin özellikleri bulunmuştur. İyileştirme için iptal oranını düşürmek için yeni tarife grupları tavsiye edilmiştir. Bu çalışma ile birlikte daha önce telekomünikasyon sektöründe yapılan iptal yönetimi çalışmalarından farklı olarak Bayes ağlarından yararlanılarak yeni bir yöntem geliştirilmiştir. İptal analiz yöntemine farklı bir bakış açısı yaklaşımı benimsenmiştir.

Koçtürk (2010), bireysel emeklilik müşterileri kayıp analizini irdelemiştir. Bunun için önce veriler temizlenmiş ve istenilen formata dönüştürülmüştür. SAS Enterprise Miner kullanarak müşteri bağlılık davranışı modellenmiştir. Entropy tabanlı 21 yapraklı karar ağacı ile analiz yapılmıştır. En kaliteli değerin sağlandığı 4 dallı ve 46 yapraklı hiyerarşinin sağladığı görülmüştür.

Arifoğlu (2011), GSM operatörünün verilerini incelemiş ve Navie Bayes, Support Vector Machine, Probabilistic Neural Network ve C-means algoritmaları karşılaştırılmıştır. Kullanılan veri seti için Adaptif Ağ Tabanlı Bulanık Çıkarım Sistemi (ANFIS) ile birlikte kullanılan kurallarda C-means algoritmasının en iyi sonucu verdiği görülmüştür.

Alizadeh (2011), fiyat geçmişi bilgilerini kullanarak fiyat tahminlemesi üzerine çalışmıştır. Narx yapay sinir ağ modelini kullanmıştır. Veri tipleri, sinüs ve

kosinüs fonksiyonları formatında sisteme verilmiştir. Oluşturulan sistem, istenilen çıktıları verdiği gözlemlenmiştir.

Karataş (2011), yazılım projesi maliyetini yapay sinir ağlarını kullanarak tahmin etme üzerine çalışmıştır. Bunun için COCOMO veri kümesini tercih etmiş ve yapay sinir ağlarının eğitiminde ve test edilmesinde kullanılmıştır. Maliyet tahmini için bir yapay sinir ağı modeli oluşturmuştur. Bu yapay sinir ağını oluştururken XOR bilinmeyeninin çözüm sisteminden faydalanmıştır.

Karahan (2011), yapay sinir ağlarını kullanarak kuru kayısı ihracat talep tahmini üzerine çalışmıştır. Malatya ili verilerini kullanmıştır. İstatistiksel talep tahmin metotlarının ileri beslemeli geri yayılım metodunu kullanmıştır. Uygulamada karşılaşılan zorluklara dikkat çekmiştir. Uygulamanın aşırı eğitim, yanlış mimari kurulması vb. gibi problemlerinden arındırılması gerektiğini vurgulamıştır. Problemler giderildiği takdirde modelinin tahmin oranının yükseldiğini ve ekonomik olduğunu belirtmiştir.

Tufan (2012), telekomünikasyon sektöründe yer alan bir firmanın verilerinden yararlanarak, ürün grupları (telefon, telefon + internet, telefon + internet + televizyon) ve tarifeler arasındaki müşteri geçişleri ile bu geçişlere neden olan başlıca faktörler tespit etmiş ve firma için önerilerde bulunmuştur. Kesikli seçim modelinden faydalanmıştır. Modelleme yapılırken değişken olarak tarife özellikleri (tarife ücreti), müşterilerin demografik özellikleri (gelir düzeyi, hane halkı sayısı, konut özelliği) ve kullanım bilgileri (konuşma süresi, veri indirme miktarı) kullanılmıştır.

Gök (2014), internet servis sağlayıcı verilerini kullanarak iptal analiz modeli ortaya koymuştur. Bu çalışması toplam dört basamaktan oluşmaktadır. İlk basamakta basit sınıflandırma işlemi ile k katlı çapraz doğrulama işlemi için her müşteriye çeşitli öznitelikler ile etiket oluşturmuştur. İkinci basamakta bu veriler davranış biçimleri gözlemlenerek k-ortalama kümeleme algoritması ve hiyerarşik kümeleme algoritması kullanılarak sınıflandırmıştır. Üçüncü basamakta verilerin tespit edilen küme merkezlerine olan uzaklıkların bulunması ve korelasyonu yüksek öznitelikler çıkarılarak modele ekler yapılmıştır. Son olarak gerçek veriler ile deneme yapılmış ve sonuca ulaşılmıştır

Yabaş (2014), Orange Telecom tarafından “Knowledge Discovery and Data Mining 2009"(KDD) yarışması için sunduğu gerçek ve kullanıma açık bir veri kümesi kullanmıştır. Bu yarışmada elde edilen sonuçlara yakın değerler elde etmeyi hedeflemiştir. Bunun için toplu sınıflandırıcı teknikleri üzerine yoğunlaşmıştır. Tek ve güçlü sınıflandırıcılar ile en son toplu sınıflandırıcıları karşılaştırıp, bu metotların performanslarını attırmak için performans gösteren sınıflandırıcılar seçerek bunları oylayıcı sınıflandırıcılar ile birleştirilmiştir. Elde edilen sonuçlar yarışmadaki sonuçlar ile yakındır.

Ercan (2015), ilişki tabanlı filtreleme yöntemi kullanılarak verileri seçmiş ve Bayesian Network, Logistic Regression, SMO ve Simple CART algoritmalarının sonuçlarını karşılaştırarak erken tahmin modeli oluşturmuştur. Elde edilen modelin doğruluğunu arttırmak için ensemble yöntemleri uygulanmış. Sonuçlar Simple CART algoritmasının oyunu terk edecek oyuncuları tahmin etmede daha başarılı olduğunu göstermiştir. Geliştirilen model oyunu bırakacak oyuncuları %68.20 doğrulukla tespit etmiştir.

Çimenli (2015), lojistik sektöründe müşteri kaybı analizi yapmıştır. Bunun için yapay sinir ağları ve karar ağaçlarını karşılaştırmıştır. Bu doğrultuda karar ağaçları 81% doğru tahmin yaparken, yapay sinir ağları 97 doğru sonuç verdiğini gözlemlemiştir.

Karaağaç (2015), bir bankadaki müşteri kaybı analizinde karar ağacı algoritması ve lojistik regresyon kullanmıştır. Analizinde kaybedilen müşterilerin yanı sıra kaybedilmemiş fakat kar olarak azalmış müşterilere de dikkat çekmiştir.

Kılıç (2015), yemekhane günlük talep tahmini yapmaya çalışmıştır. SPSS programında veriler analiz edilerek veriler arasında ilişkileri incelemiştir. Verilerin eğitimi ve test için yapay sinir ağlarından çok katmanlı ve radyal tabanlı fonksiyon tercih edilmiştir. Bu çalışma ile günlük yemek miktarı tahmininin yapılabileceği gösterilmiştir. Sonuçlar olumlu olarak yemekhaneye öngörü oluşturmuştur. Matlab programında GUI tasarlanarak günlük yemek tahmini için yapılmıştır.

Kalabalık (2016), çoklu makine öğrenmesi algoritmalarının, birleştirmeli sınıflandırma yöntemlerini mevcut tahmin etme metotlarının ölçü doğruluğunu artırmak için kullanarak birleştirilmesini incelemiştir. Bagging, boosting ve

random forest birleştirmeli sınıflandırma yöntemlerini kullanmıştır. İyi bir sınıflandırma tabanı ile kullanılan birleştirmeli sınıflandırma yöntemlerinin müşteri ayrılma analizi tespitinde etkili olduğu görülmüştür.

Sarı (2016), motor yataklarının satış talep tahmini yapay sinir ağlarını kullanarak incelemiştir. Elde edilen sonuçlar Regresyon Analizi (RA) ve zaman serileri ile yapılan tahmin sonuçlarıyla karşılaştırılmıştır. Sonuç olarak yapay sinir ağları ile gerçeğe daha yakın tahminler elde edilmiştir.

# 3 VERİ MADENCİLİĞİ

Günümüzde bilgisayar sistemlerine olan ihtiyaç hızla artmaktadır. Gün geçtikçe ucuzlayan ve gelişen bilgisayar sistemleri, hayatın her alanında kullanılmaya çalışılmaktadır. Hızla artan dünya nüfusunun arz talep dengesini sağlayabilmek için işletmeler de bilgisayar sistemlerini en üst düzeyde kullanmaktadırlar. İşletmelere, hem maliyet hem de zaman açısından çok büyük fayda sağlamaktadır. Artan nüfus ile birlikte bilgiler eski yöntemler ile değil artık bilgisayar ortamlarında saklanmaktadır. Hızla artış gösteren bu durum, sayısal bilgi miktarının artmasına sebep olmaktadır. Bu teknolojik gelişmeler ile bilgiye ulaşmak daha da kolaylaşmış ve çok daha fazla bilgiyi büyük veri tabanlarında saklama imkanı artmıştır. Teknolojinin hızla gelişmesi büyük faydalar getirdiği gibi bu gelişme bazı problemler doğmasına sebep olabilmektedir. Veri tabanlarında saklanan bilgilerin sayıları arttıkça bunları anlamlı bir hale getirmek ve verileri yönetmek sorun haline dönüşebilmektedir.

Bu noktada veri analizi devreye girmektedir. Tek başına bir anlam taşımayan, işlenmemiş verileri, işleyip anlamlı bir hale getirmek gerekmektedir. Milyonlarca veri arasından ihtiyacımız olan bilgiye erişmek için bu çok önemlidir. Faydalı sonuçlar elde edebilmek için büyük veri depoları ve gelişmiş bilgisayar sistemleri tek başına yeterli değildir. Bu doğrultuda veri madenciliği ile ihtiyacımız doğrultusunda ham bilgiyi işleyip anlamlandırarak kullanılabilir hale getirebiliriz.

Veri madenciliği verinin olduğu hemen hemen her yerde kullanılmaktadır. Kullanıcıya büyük kolaylık sağlayan bu sistem özellikle bilim ve mühendislikte, eğitimde, ulaşımda, tıbbi araştırmalarda, bankacılık ve sigortacılıkta, pazarlamada, elektronik ticarette ve telekomünikasyonda sıklıkla kullanılmaktadır. Pazarlamada mevcut müşterilerin satın alma alışkanlıkları, benzer özelliğe sahip olan müşterilerin tercihlerindeki benzerlikler, mevcut müşterilerin kaybedilmemesi adına yapılan çalışmalar ve yeni müşterilerin kazanılmaya çalışılması gibi faaliyetler bu alanda yapılan veri madenciliği

örnekleri olarak karşımıza çıkmaktadır. Bankacılık ve sigortacılıkta dolandırıcılıkların tespit edilmesi, harcama şekillerine göre müşterilerin gruplandırılması, ihtiyaçlarının değerlendirilmesi, iyi kötü müşteri analizinin yapılması, risk gurubunda olan müşterilerin belirlenmesi gibi çalışmalar bu alanda yapılan çalışmalardan bazılarıdır. Kullanıcıların talepleri doğrultusunda web sitelerinin güncellenmesi, siber saldırıların çözümlenmesi, elektronik ticaret yapan kullanıcıların ihtiyaçlarının göz önüne alınması ve çözümlenmesi elektronik ticaret alanında yapılan veri madenciliği çalışmalarından bazılarıdır. Telekomünikasyon alanında ise müşteri davranış şekillerine göre ihtiyaç duyulan yeni hizmetlerin sunulması, yasal olmayan kaçak kullanımların tespiti, hatlar üzerinde problem yaşanılan bölgelerin düzeltilmesi ve kullanıcı yaklaşım ve davranışlarının tespit edilmesi veri madenciliğinde faydalanılan alanlardan bazılarıdır. Sağlık sektöründe de sıklıkla kullanılan veri madenciliği, hastalık tanılarının konulması, gelişen dünyada gelecek için sağlık politikalarına yönelik yaklaşımlar, dna ve rna içerisindeki genlerin sıra düzenlerinin belirlenmesi, hastalık haritalarının yapılması gibi birçok konuda ciddi faydalar sağlamaktadır.

Veri madenciliği hem klasik istatiksel yöntemleri hemde makine öğrenmesi yöntemlerini kullanır. Klasik istatiksel yöntemlere Lineer Regresyon, Lojistik Regresyon ve K-Means algoritmaları örnek verilebilir. Makine öğrenmesi yöntemlerine ise Destek Vektör Makineleri, Genetik Algoritmalar ve Yapay Sinir Ağları örenek gösterilebilir.

Verilerin çok fazla miktarda olduğu durumlarda, istenilen sonuca ulaşılması için bu verilerin elle işlenmesi mümkün değildir. Aynı zamanda büyük verilerin analizinin yapılması da zorlaşmaktadır. Hedef geçmişteki verileri işleyerek gelecek için tahminlerde bulunmaktır ve bu problemleri çözmek için Makine Öğrenmesi (machine learning) yöntemleri geliştirilmiştir. Makine öğrenmesi yöntemleri geçmişteki verileri kullanarak analiz eder ve yeni veri için en uygun modeli bulmaya çalışır.

Makine öğrenmesi, bilgisayar yazılımlarıyla mevcut verilerden elde edilen deneyimlerin, gelecekteki olayları tahmin etmesine ve modelleme yapmasına imkân veren bir yapay zekâ alanıdır. Sunulan veriler ve parametrelere bağlı olarak benzetimler yaparak, programlamadıklarımızı ortaya çıkarabilen, kendi

kendini eğitebilen bir sistemdir. Makine öğrenmesinin avatajlarından bir tanesi öğrendiği bilgiler ile programın davranışı değiştirir.

Verinin incelenip, içerisinden işe yarayan bilginin çıkarılmasına da Veri Madenciliği (data mining) adı verilir (Kostek, 2014). En genel manada veri madenciliğini tanımlayacak olursak; çok sayıdaki veriler arasından değerli sayılan ve kullanılabilir olanların ayrıştırılması olarak ifade edilir. Veri madenciliğinde bilgiler toplanır, bilgi keşfinin amacına bağlı olarak bir takım istatistiksel yöntemler kullanılarak sadeleştirilir, ham veri elde edilir ve bu ham veri üzerinden gerekli çıkarımlar yapılır.

Veri madenciliği hakkında birbirini tamamlayan tanımlar mevcuttur. Jacobs, 1999 yılında, veri madenciliğini, sadece ham bilgilerin yeterli olmadığında istenilen veriye meydana getiren veri analizi safhası biçiminde belirtmiştir. Hand ise, 1998 yılında, veri tabanı teknolojisi, veri madenciliğini istatistik, örüntü tanıma, makine öğrenme ile etkileşimli yeni bir disiplin ve geniş veri tabanlarında önceden tahmin edilemeyen ilişkilerin ikincil analizi olarak tanımlamıştır. Gartner Group' a göre veri madenciliği, yığılmış veriler üzerinde bir takım istatiksel ve matematiksel teknikler kullanarak anlamlı yeni veri setlerinin ortaya çıkarılması, bu veri setlerindeki örüntülerin ve istenilen eğilimlerin keşfedilme sürecidir (Larose, 2005). Davis ise 1999 yılında, veri madenciliğinin çok büyük miktarlardaki bilgilerdeki birbiri ardındaki bağlantıları analiz eden, matematiksel algoritmalar kullandığını belirtmiştir. Davis' in tanımına göre ise, veri madenciliği varsayımları ortaya çıkarır, bu varsayımlardan elde edilen sonuçları insan yeteneğini kullanarak anlamlandırır. DuMouchel' in 1999' da ki açıklamasında, veri madenciliği için geniş veri tabanlarındaki nitelikler kullanılarak birliktelik çıkarımlarını araştırdığını belirtmiştir. Wang ve Kitler 1998' de belirttikleri gibi, veri madenciliği tanımını, tahmin yeteneği yüksek önem arz eden verilerin çok sayıda potansiyel veriden ayrımının yapılmasını gerçekleştirme becerisi şeklinde belirtmişlerdir. Bransten ise 1999 yılında, veri madenciliğini tanımlarken insan zekasının tespit edilmesi mümkün olmayan bilgilerin ortaya çıkarmayı sağladığını şeklinde açıklamıştır.

Veri madenciliğini özetleyecek olursak daha önceden bulunmamış veriler arası ilişkilerin ortaya çıkarılabilmesi amacıyla elimizdeki yüksek sayıdaki bilgiyi

inceleyen bir metottur. Veri analizinin yapılabilmesine olanak sağlayan büyük kapasiteli bilgisayarlar ve yazılımlara rahat ve az maliyetle sağlanabilmesi, bu teknolojik yapının günümüzde kolay uygulanabilmesini sağlamıştır. Veri madenciliği çalışmalarının bilgisayar üzerinde yapay zekâ çalışmaları ile birleştirilerek etkin bir şekilde uygulanması zaman tasarrufunu da beraberinde getirmiştir.

## 3.1. Veri Madenciliği Süreci

Veri madenciliği uygulaması bir süreçtir ve bu süreç birbirine bağlı adımlardan oluşur. Kaliteli sonuç veren veri madenciliği alanındaki çalışmalar için yapılması gerekli olan adımlar şu şekildedir:

### 3.1.1. Problemin tanımlanması

Veri madenciliği hiyerarşisinin en önemli ve ilk kısmı olan bu aşamada, çalışmanın amacını, var olan halinin incelenmesini, veri madenciliğinin hangi amaçla kullanıldığını ve bu çalışma için yapılan planlama aşamalarının saptanmasını içermektedir. (Koçtürk, 2010).

Problemin tanımlanması aşaması mevcut iş probleminin çözümünde gerekli nasıl bir sonuç istendiğinin, ortaya çıkacak neticenin maliyet ve fayda arasındaki yapının incelenmesinin faydalı bir şekilde çözümlenmesi yani ortaya çıkan sonucun ilgili firma için kıymetinin faydalı bir şekilde incelenmesi gerekmektedir (Akbulut, 2006).

### 3.1.2. Verilerin hazırlanması

Veri madenciliğinde en önemli aşamalarından bir diğeri, veri guruplarının işlem öncesi hazırlanması aşaması veri toplamakla başlamaktadır. Veri gurupları hazırlama aşaması, el değmemiş bilgiden başlayarak elde edilecek sondaki bilgiye kadar yapılması elzem olan tüm hazırlık düzenlemelerini içermektedir (Koçtürk, 2010). Bu düzenlemeler veri hazırlama, tablo, kayıt, veri dönüşümü ve modelleme araçları için veri temizleme gibi özellikleri içermektedir. Ortak özelliklere sahip bilgileri yan yana toplama ve kendi özelliklerini ortaya çıkarma, ardından bu bilgileri ortaya çıkarma ve gizli verileri özelliklerine göre guruplara ayırma işlemleri şeklinde devam eder (Akbulut, 2006).

#### 3.1.1.1 Veri toplama

Problemde kullanılacak verilerin kaynaklarının belirlenmesi aşamasıdır. Bu adımda işlenmemiş bilgilerde var olan farklılıklar en aza indirilmeye çalışılır, yanlış veya yapılacak incelemenin hatalı sonuçlar vermesine neden olabilecek bilgilerin temizlenmesi sağlanır. Hatalı bilgi girişinden veya sadece bir defa gerçekleşen bir durumun analizi ne kadar etkilediği dikkate alınarak veriler ana kümeden çıkarılır.

#### 3.1.1.2 Veri dönüştürme

Veriler kullanılacak olan model ve algoritma çerçevesinde istenilen formata uyacak şekilde düzenlenir. Bir örnekle açıklamak gerekirse, kredi riski çalışması için değişik iş çeşitlerinin, sağlanan gelir miktarı ve yaş seviyesi gibi parametrelerin kodlar halinde kümelenmesinin daha sağlıklı sonuçlar vereceği düşünülmektedir (Akbulut, 2006).

### 3.1.3. Modelin Kurulması

Veri madenciliğinde bilgi keşfi yapma ve tahminleme işlemlerini içeren bu yöntem modelleme olarak adlandırılır. Modelleme, mevcut problem için cevapları bilinen soruların olduğu hallerden bazı kurallar ve sonuçlar çıkarılır. Cevapları bilinmeyen soruların bulunduğu hallerde ise bu kuralların ve sonuçların mevcut bilgi kullanılarak işlenmesidir. (Tosun, 2006).

Problemin çözümünde kullanılacak olan en yararlı biçimin bulunabilmesi, imkan verdiğince yüksek miktarda örneğin entegre edilerek denenmesi ile sağlanabilir. Bundan dolayı ham bilgi hazırlama ve örnek oluşturma süreçleri birlikte ilerleyen ve yüksek fayda sağladığı düşünülen örneğe ulaşılıncaya kadar tekrar edilen bir durumdur.

Veri madenciliğinde denetimli (supervised) ve denetimsiz (unsupervised) öğrenim mevcuttur. Denetimli öğrenme kısmında ilgili kişi mevcut sınıfları daha önce tespiti yapılan bir ölçüt doğrultusunda ayırır ve bütün sınıflar için türlü modeller sunar. Burada amaç her sınıfın kendi özelliklerini ortaya çıkarmaktır. Öğrenme süreci bittikten sonra öğrenilen kurallar bütünü sağlanan yeni modeller üzerinde çalıştırılır. Model tarafından yeni örnekler ilgili sınıflara atanır. Denetimsiz öğrenmede sisteme verilen örnekler gözlenir. Örneklerin özellikleri arasındaki benzerlikler keşfedilerek sınıfların tanımlanması sağlanır.

Denetimli öğrenimde bilginin belli bir miktarı örneğin eğitilmesi için kullanılır. Diğer bir miktarı ise örneğin uygunluğunun analiz edilmesi amacıyla alınır. Modelin öğrenimi, bu kümedeki veriler üzerinden uygulandıktan sonra, test kümesindeki veriler üzerinden bu örneğin kalite miktarı tespit edilir. Bir sınıflama modelinde hatalı bir şekilde gruplandırılan durum miktarının, gerçekleşen bütün durum miktarına bölünmesiyle hata oranını temsil eder. Hatasız olarak gruplanan durum miktarının bütün gerçekleşen durum miktarına bölünmesiyle doğruluk oranı tespiti sağlanır (Akpınar, 2000).

Modelin kuruluş amacına bağlı olarak yani istenilen sonucun elde edilebilmesi için aynı teknikle farklı parametreler kullanılabilir, başka algoritmalar denenebilir ve ya farklı araçların kullanılıp denendiği değişik modeller oluşturulabilir. En uygun modelin seçilebilmesi için farklı modeller kurarak doğruluk derecelerine göre birçok deneme yapılması önem arz eder.

Kurallarla oluşturulan model her ne kadar doğruluk derecesi yüksek olsa da sonuçları gerçek dünyada kesin sonuç garantisi vermez. Çünkü yapılan testlerde bir takım varsayımlar mevcuttur. Aynı zamanda bu modelin uygulandığı veriler doğruluk payını arttırmıştır. Örneğin modelin kurulması sırasında varsayılan dolar kurunun değişkenlik göstermesi alıcının satın alma davranışında değişikliğe sebep olacaktır.

### 3.1.4. Modelin Kullanılması

İleri çözümleme yöntemlerinin kullanıldığı aşamadır. Modelleme aşaması olarak ta adlandırılan bu aşamada tekniğin belirlenmesi, test örneklemesinin oluşturulması, modelin geliştirilmesi ve tahmin yapma süreçlerini barındırır. Parametreler modellere elverişli formata dönüştürülür. Eğer parametreler seçilen yönteme uygun olmadığı görülürse ya da özel tanımlamalar gerektirirse veri hazırlama aşamasına geri dönülür.

Model tek başına bir uygulama şeklinde tasarlanabileceği gibi farklı bir uygulamanın bir kısmı olacak şekilde de çalıştırılabilir.

### 3.1.5. Modelin Değerlendirilmesi

Uygun model belirlendikten sonra elde edilen sonuçların problemin amaçlarını gerçekleştirip gerçekleştirmediğinin değerlendirildiği aşamadır. Ayrıca veri madenciliği sürecinin ele alınması söz konusudur. Bu aşamada veri madenciliği

yardımı ile ortaya çıkarılan veriler var olan benzer sorunların çözümlenmesi aşamalarında da fayda sağlamaktadır. Ulaşılan sonuçlar doğrultusunda sonraki adımların neler olacağı çıkarılır. Bu aşamanın sonunda, veri madenciliği sonuçlarının kullanılıp kullanılmayacağa konusunda karar verilir. Elde edilen bilgileri uygulamada gerekli planın hazırlanması, ilerleyen dönemlerde gözden geçirilmesi ve gerekli durumlarda bakım faaliyetlerinin gerçekleşmesini içerir. Ayrıca araştırma raporunun yazılması bu aşama kapsamındadır.

### 3.1.6. Modelin İzlenmesi

Gün geçtikçe sistemlerin özelliklerindeki ortaya çıkan değişiklikler ürettikleri verilerin de değişmesine sebep olmaktadır. Bundan dolayı kurulan modellerin devamlı izlenmesi gerekmektedir. Eğer gerekli ise yeniden düzenlenmesi de gerekecektir. Model sonuçlarının izlenebildiği, ihtimalleri düşünülen ve gerçekleşen parametrelerle olan ilişkisi sonucu ortadaki farklılığı yansıtan grafikler kullanışlı bir yöntemdir (Shearer, 2000).

Modelleme, veri madenciliğinin bulgu ve tahminleme edinme tekniğine denilmektedir. Modelleme tekniği, doğru yanıtların var olan yapılardan kurallar ve buna bağlı olarak sonuçlar sağlanarak, cevapları tahmin edilemeyen yapılar bu kurallar ve buna bağlı neticelerin bilgi ile kullanılarak işlenmesidir (Tosun, 2006).

Veri madenciliği modelleri işlevlerine göre 3 ana başlık altında toplanır:

1- Sınıflama (Classification) ve Regresyon (Regression)

2- Kümeleme (Clustering)

3- Birliktelik Kuralları (Association Rules)

## 3.2. Sınıflandırma Algoritmaları (Classification Algorithms)

Veri madenciliğinde, bir bilgi kümesi içerisinde tanımlanmış çeşitli sınıflara göre belirli özellikler baz alınarak sınıfı belli olmayan verilerin bu sınıflara dağıtılması sınıflandırma olarak tanımlanır. Sınıflandırma mevcut belirlenmiş bilgiler yardımı ile yeni sağlanan bilgi kümelerinin hata yapılmadan işaretlenmesinin yapıldığı işlemlerdir. Bu yöntemin amacı eldeki veriler kullanılarak sınıflandırılmamış verilerin yüksek oranda hangi sınıfa ait olduğu tahmin edilerek sınıflandırıcıları oluşturabilmektir.

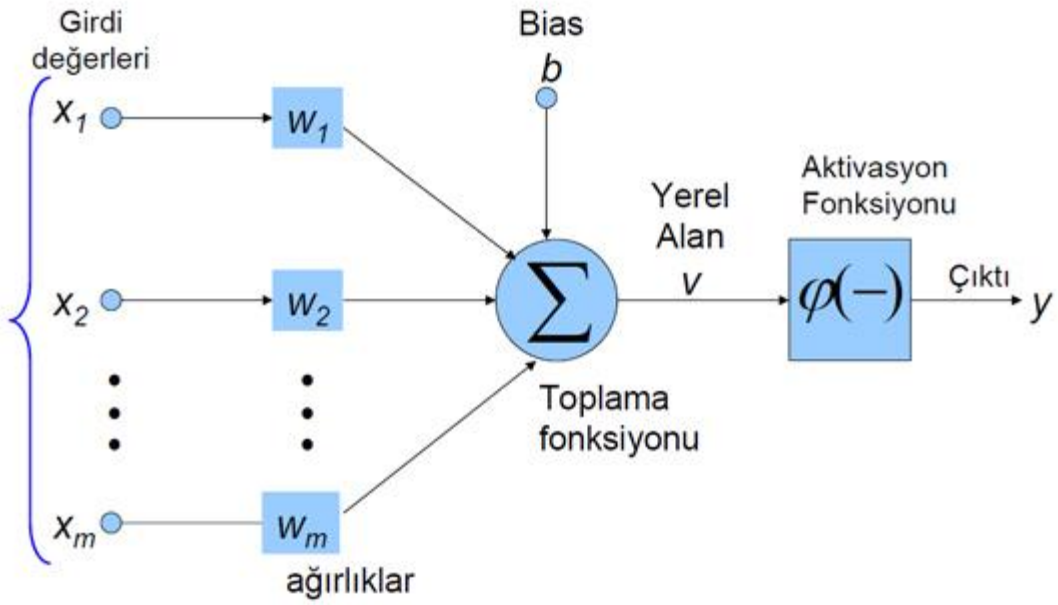
Sınıflandırma algoritmasının bankacılık sektöründe sahtecilik, telekom sektöründe müşteri sadakati analizi, tıp alanında hastalık tespitleri gibi çeşitli kullanım alanları mevcuttur. Her bir sektörde farklı sınıflandırma algoritmaları kullanabilmektedir.

### 3.2.1. Yapay Sinir Ağları (Artificial Neural Networks)

Yapay sinir ağları (YSA), insan beyninin çalışma prensiplerini benzeterek yapma şeklinden yola çıkarak birbirine nöronlarla bağlı sistemler olarak tanıtılan, insan sinir sisteminden esinlenerek yeni sistem oluşturmaya çalışan, sayısal olarak modellenen yapılar olarak tanımlanır (Zupan, 2003).

Yapay sinir ağları, birçok basit işlemci elemandan oluşur ve bu elemanlar sayısal olarak ifade edilebilen "bağlantılar" veya "ağırlıklar" ile birbirlerine bağlı olarak farklı formlarda ifade edilebilir (Perendeci, 2004). Sinir ağları, örüntü tanıma, belirleme, sınıflandırma uygulamaları ile ses, görüntü ve kontrol sistemlerini içeren çok çeşitli alanlarda karmaşık fonksiyonları çözümlemek için eğitilirler (Bilgin,2008).

Günümüzde kullanılan yazılım teknolojisi ile çözülemeyen birçok problem yapay sinir ağları ile çözüme ulaşabilmektedir. Özellikle eksik, normal olmayan, belirsiz bilgileri işleyerek çeşitli sonuçlar elde edilmesini sağlar.




**Şekil 3.1:** Yapay sinir ağı yapısı (Etika, 2009)

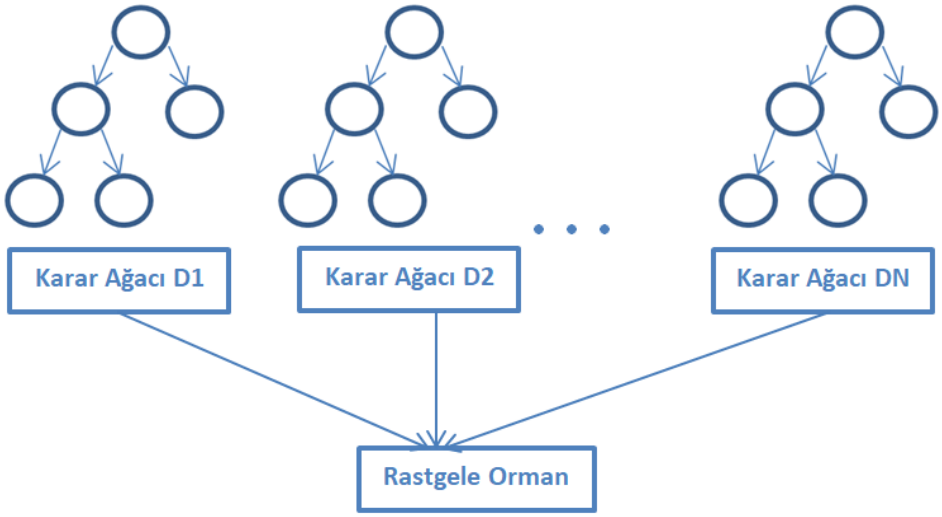
Yapay sinir hücresinin kısımları; girişler, ağırlıklar, toplama işlemi, aktivasyon fonksiyonu ve çıkış olarak toplam beş aşamadan meydana gelmektedir. Yapay sinir ağlarında öğrenmenin sağlanabilmesi işlemi için girişlerin olması gerekmekte ve bu şekilde sağlanmaktadır. Çalışma şekli olarak önceki katmandan veya dış dünyadan sağlanan veriler yeni bir giriş elemanı olarak yapay sinir hücrelerine iletilir. Ağırlık katsayıları, sağlanan yeni veri girişlerini yapay sinir hücrelerine olan etki seviyesini belirler ve öğrenme kısmının sağlanmasına etki eder. Sağlanan bu yeni veri girişinin hesaplanması, toplama işlemi olarak nitelendirilir. Bu aşamada sıklıkla kullanılan yöntem ise ağırlıklı toplamı bulma işlemidir. Bahsedilen işlemi sağlayabilmek için hesaplanan her ağırlığın bulunduğu yerdeki girişler ile çarpılmasıyla elde edilen sonucun toplamlarına eşik değeri eklenir. Gerçek sinir hücrelerinin yüzeyinde bulunan potansiyel farkın yapay sinir hücrelerinde benzer şekilde sağlamak amacı ile kullanılan kat sayılara eşik değeri denmektedir. Her bir yapay sinirin tek çıkış noktası vardır ve bu çıkış mevcut daha sonra gelen yapay sinirler için bir giriş siniri olarak kullanılabilir. Çıkış kısmında sonuçlar dışa aktarılır. Yapay sinir ağlarının giriş bilgileri sıklıkla hesaplanır çünkü yalnızca sayısal giriş verileri ile desteklenirler (Haykin, 2005).

### 3.2.2. Rastgele Orman Algoritması (Random Forest Algorithm)

Rastgele orman, sınıflandırma yöntemlerinden, toplu sınıflandırma gurubunda yer almaktadır. Birden fazla sınıflandırıcı ile bu sınıflandırıcıların tahminleri sonucunda elde edilen sonuçlarla sonuca ulaşan algoritmalar olarak tanımlanırlar.

Sıklıkla kullanılan toplu sınıflandırıcılardan olan torbalama algoritması, değiştirilmemiş eğitim veri gurubu ile birden fazla, önyüklemeli eğitim veri gurupları hazırlanır. Önyüklemeli her bir eğitim amaçlı veri gurubu için yeni bir ağaç tasarlanır. Artarda sıralanan ağaçlar bir önceki ağaçtan bağımsız bir şekilde davranır ve tahminleme amacı ile en büyük oy temel alınır. Orman yöntemi, torbalama yöntemi için rastgele özellik seçim kısmı eklenmesi ile geliştirmiştir. Rastgele orman, Breinman ve Cutler tarafından geliştirilirmiştir. En iyi dal ile her düğüm seçeneğini dallara ayırarak değil, her düğüm için rastgele seçilin verilerden en iyi olanı kullanıp her düğümü dallara ayırarak yapar. Üretilen her veri gurubu orijinal veri gurubunu kullanarak yer değişmeli

olarak sağlanır. Rastgele özellik seçimi yardımı ile ağaçlar geliştirilerek kullanılır (Archer, 2008).



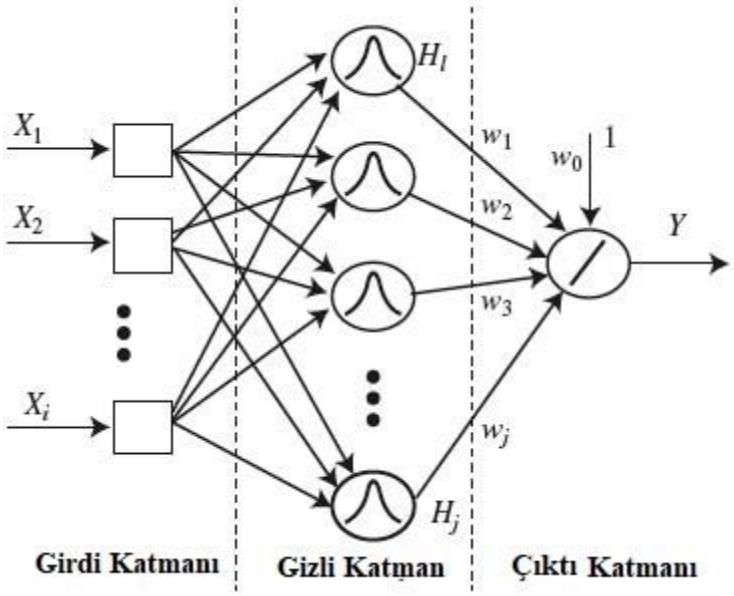

**Şekil 3.2:** Rastgele orman algoritması yapısı [1]

Şekil 3.2' de görüldüğü gibi, Rastgele Orman sınıflandırıcısının gelişimi, her veri gurubu için sağlanan karar ağaçları ve bu karar ağaçlarının birleşimleri ile oluşmaktadır.

### 3.2.3. Radyal Tabanlı Fonksiyon Ağları (Radial Basis Function Networks)

Bir optimizasyon uygulaması olan, katmanlı Yapay Sinir Ağları' nın gelişiminde geriye yayılımlı öğrenme algoritması kullanılmaktadır. Moody ve Darken tarafından 1989 yılında geliştirilmiştir. Danışmalı öğrenme şeklinde çalışan ileri beslemeli Yapay Sinir Ağları modelidir. RTFA' yı çok boyutlu uzayda eğri uydurma yaklaşımı olarak görmekteyiz (Mahanty, 2004).

İleri beslemeli bir ağ yapısında bulunan Radyal Tabanlı Fonksiyon Ağları, üç katmandan oluşmaktadır. Bunlar; giriş katmanı, gizli katman ve sonuç katmanıdır. Diğer ağ yapılarından farklı olarak bu ağlarda, giriş katmanından gizli katmana geçerken doğrusal olmayan bir kümeleme analizi ve radyal tabanlı aktivasyon fonksiyonları kullanılmaktadır. Ağırlıkların hesaplanması için bunlarla ilgili öğrenme algoritmalarının tercih edildiği ve bunlarla ilgili ağırlıkların belirlendiği kısım ise gizli katman ile sonuç katmanı arasındaki kısımdır (Bolat, Küçük, & Yıldırım, 2004).




**Şekil 3.3:** Radyal tabanlı fonksiyon ağ yapısı (Şen, 2004)

Şekil 3.3' te radyal tabanlı fonksiyon ağ yapısı ve katmanları görülmektedir.

İnterpole etme problemi için eğri uydurma teorisi kullanılmaktadır. Bu ağlarda kullanılan bazı parametreler vardır. Bunlar; çıkış katman ağırlıkları, merkez vektörleri ve radyal fonksiyon genişliği olarak sıralanabilir. Doğrusal optimizasyon veya eğim düşme yöntemleri ile ağırlıklar kolay bir şekilde hesaplanabilir çünkü çıkış katmanının doğrusal bir yapıdadır. Radyal tabanlı fonksiyon ağlarının performansını arttırmak amacı ile merkez vektörlerinin ve bu ağların genişliğinin istenilen seviyeye getirilmesi sebebiyle birçok yöntem geliştirilmiştir. Bu ağlardaki merkezler, giriş verileri arasından rastgele veya değişken olmayan şeklinde belirlenebilmektedir. Merkez yöntemlerini belirlemek için iki yöntemimiz vardır. Birinci yöntem, dik en küçük kareler yöntemi ile eğitim düşme yöntemi eğiticili öğrenme algoritması ile uyarlanarak belirlenir. İkinci yöntem ise, giriş verilerinden gruplama yaparak kendiliğinden düzenlemeli yöntem ile belirlenir.

### 3.2.4. Karar Ağaçları Algoritmaları (Decision Tree Algorithms)

Veri madenciliğinde, sınıflandırma ve tahmin çalışmaları için karar ağaçları sıklıkla kullanılmaktadır. Sınıflandırma için yapay sinir ağları ve diğer yöntemlerin de kullanılabilmesine karşın, kolay anlaşılabilir olması kullanıcılar için büyük kolaylıklar sağlamaktadır (Chien & Chen, 2008).

Karar ağaçlarının sıklıkla tercih edilmesinin sebeplerinden bazıları; veri tabanları ile bağlantısının kolay yapılabilmesi, güvenilirliğinin yüksek olması, maliyetinin yüksek olmaması, anlaşılması ve yorumlanmasının kolay olması şeklinde sıralanabilir. Bu tekniğin kullanılması ile veri sınıflandırması iki kısımdan oluşur ve bunlar öğrenme ve sınıflamadır. Model oluşturmak için bir eğitim verisi, sınıflama algoritması yardımı ile analiz edilir. Elde edilen bu model, karar ağacı olarak ortaya çıkar. Sınıflama aşamasında ise, test işlemi için gönderilen veri karar ağacının doğruluk seviyesini göstermek için kullanılır. Yapılan test işleminden sonra, doğruluk seviyesi istenilen düzeyde ise, yapılan yeni veri girişleri bu kurallar üzerinden yapılır. Karar ağacı oluşturulması sırasında uygulanacak aşamaların sırası belirlenmelidir. Bu aşamaların sırasının belirlenmesi için yaygın olarak Entropi ölçümü kullanılmaktadır. Entropi ölçüm sonucunun yüksek çıkması ile bu sonuca bağlı olarak işlenen veriler, o derece tutarsızdır ve bu veriler karar ağacının tepesinde kullanılır. Entropi ölçüm sonucunun düşük çıkması sonucunda ise, işlenen veriler karar ağacının kökünde kullanılırlar. Girilen bir $A_k$ alanı için Entropi ölçümünü yapan formüller aşağıdaki gibidir.

$$E(C|A_k) = \sum_{j=1}^{M_k} p(a_k, j) x \left[ -\sum_{i=1}^{N} p(c_i|a_k, j) \log_2 p(c_i|a_k, j) \right] \quad (1.1)$$

Formül 1.1 eşitliğindeki ifadeler aşağıdaki gibidir:

$E\ (C \setminus Ak)$ = $Ak$ alanının sınıflama özelliğinin Entropi ölçüsü.

$p\ (ak, j)$ = $ak$ alanının $j$ değerinde olma olasılığı.

$p\ (ci \setminus ak, j)$ = $ak$ alanı $j$. değerindeyken sınıf değerinin $ci$ olma olasılığı.

$Mk$ = $ak$ alanının içerdiği değerlerin sayısı.

$N$ = farklı sınıfların sayısı.

$K$ = alanların sayısı.

Bir S kümesinin elemanları, gurupsal anlamda $C1, C2, C3, \ldots, Ci$ guruplarına ayrıştırılırlarsa, $S$ kümesindeki bir elemanın gurubunun hangisi olduğunu belirlemek amacıyla gereken bilgi 1.2.de gösterilen formül yardımı ile hesaplanmaktadır:

$$I(S) = -(p_1 \log_2(p_1) + p_2 \log_2(p_2) + \ldots + + p_i \log_2(p_i)) \quad (1.2)$$

Formül 1.2' deki pi, Ci sınıfına ayrılma ihtimalidir. Entropi denklemi aşağıdaki şekilde gösterilebilir:

$$E(A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} x I(S_i) \quad (1.3)$$

Formül 1.3' te A alanı ile işlenecek dallanma işleminde, bilgi kazancı Formül 1.4' e göre hesaplanmaktadır:

$$Kazanç\ (A) = I\ (S) - E(A) \quad (1.4)$$

Yani Kazanç (A), A alanının değerini biliyor olmanın sağladığı entropideki azalma durumudur. Karar ağaçlarında kullanılan bazı algoritmalardan yararlanılmaktadır. Bunlardan bir kaçı; ID 3, C 4.5, C 5.0, CART, CHAID ve QUEST bu algoritmalardan bazılarıdır.

C 4.5 ve C 5.0 Algoritmaları: Karar ağacı için sıklıkla kullanılan Quinlan' ın ID3 algoritmasının gelişmiş versiyonu olarak C 4.5' i gösterebiliriz (Quinlan,1993).

C 4.5 algoritmasının geşilmiş versiyonu ise C 5.0 algoritmasıdır ve bu algoritma daha büyük veri gurupları için kullanılmaktadır. Boosting algoritmaları da kullanıldığı için C 5.0 algoritmaları Boosting ağaçları olarak da adlandırılır. C 4.5 ve C 5.0 algoritmaları aynı sonuçları vermektedir. C 5.0 algoritması C 4.5 algoritmasına göre daha hızlı çalışmaktadır ve biçimsel açıdan daha verimli karar ağaçları sunmaktadır.

CART Algoritması : Breiman ve arkadaşları tarafından 1984 yılında Automatic Interaction Detection isimli karar ağacı algoritmasının devamı olarak geliştirilmiştir. Sınıflandırma ve regresyon problemleri için kullanılan bu algoritma, hem sayısal hem de nominal veri türlerini desteklemektedir.

CHAID Algoritması: Sıklıkla kullanılan karar ağacı algoritmalarından olan CHAID algoritması, optimal bölünmelerin tespit edilebilmesi için ki-kare analizini kullanmaktadır. Bu algoritma bölümlendirme amacı ile kullanılmakta olup, güçlü bir analiz tekniğidir.

QUEST Algoritması: Bu algoritma diğerlerinden farklı olarak, ikili karar ağacı yapısı kullanmaktadır. Loh ve Shih tarafından 1977 yılında önerilmiştir. Doğrudan durma ve budama işlemlerinin yapılabiliyor olması, ikili ağaç kullanılmasının başlıca sebebidir. Diğer algoritmalarda olduğu gibi bölünme işlemlerini aynı zamanda yapmaz, her birini ile tek tek inceler.

### 3.2.5. Genetik Algoritmalar (Genetic Algorithms)

"En faydalı olan hayatta kalır" ilkesine dayanan ve stokastik bir arama yöntemi olan bu algoritma, John Holland tarafından bulunmuştur. Bu algoritma biyolojik sistemlerin ilerleme aşamalarını modellemektedir (Holland, 1975). Holland' ın çalışma arkadaşları ve öğrencileri tarafından geliştirilerek Holland tarafından 1975 yılında kitap haline getirildi ve yayınlandı. Genetik algoritmalar sınıflandırma işlemlerinde de kullanılan arama algoritmalarıdır. Evrimleşme mekanizmasına dayanan bu algoritmalar, rastgele üretilmiş kromozom, durum ve çözüm hipotezi ile başlar. Belli bir boya sahip dizilerden oluşmuş bir topluluğa sahiptir. Bu topluluğa ait tüm kromozomlar, çözüm kümesi için bir düğümü belirtir. Bu bireyler üreme yoluyla hayatlarını devam ettirmeye adaydır. İşleyiş yönünden bu algoritma, bireyler arasından sonuç vermeye elverişli olmayanları elemek, sonuç vermeye daha elverişli bireyleri seçmek ve bu verimli bireylerden üreme yoluyla yeni bireyler ortaya çıkarmaya dayanır. Aşamalı olarak çalışan bu algoritmanın asıl amacı, bireyler arasından sonuç vermeye elverişli olmayanları elemeye dayanmaktadır. Bir sonraki aşama ise verimli olduğu düşünülen bireylerden faydalanılarak, yeni çözümlere ulaşmayı sağlamaktır. Bahsedilen bu sistem, bilgi alışverişi ile yürütülmektedir. Rastgele olarak yapılan bilgi alışverişi, yapılan arama işlemi için daha verimli yerlerde sürdürülmesine olanak sağlar (Telcioğlu, 2002).

Holland, bir bilgisayara yardımı ile bilgisayar sistemlerine anlayamadığı çözüm metotlarının öğretilebileceği fikrine sahipti. Eldeki bir problem için en iyi sonucu bulabilmesinin garanti edilmediği fakat sezgisel bir algoritma olması sebebi ile çözümü zor olan veya çok zaman alan problemler için en iyiye yakın sonuçlar verebildiği görüşmüştür.

### 3.2.6. Naive Bayes Algoritması (Navie Bayes Algorithm)

Bu algoritma, istatistiksel bir metot olan Bayes teoremini temel alarak işlemektedir.

Bu teoremde rastgele seçilen bir değişken için, olasılık dağılımı yapılırken marjinal ve koşullu olasılıkların ilişkilerini izah eder. Bayes teoremi, bağımsızlık önermesi ile sadeleştirilerek Naive Bayes algoritması oluşturulmuştur. Bağımsızlık önermesi, elde bulunan örüntülerin özelliklerini kullanıp önceden öğrendiği haliyle bu örüntüleri sınıflandırarak kullanılacak bütün kriterlerin istatistiksel olarak saltık halde ve aynı seviyede önemli görülmesi ihtiyacını ortaya koyar.

$$P(C_i|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|C_i)P(C_i)}{\mathrm{P}(\mathbf{X})} \tag{2.1}$$

Bayes kuralı Formül 2.1' de gösterilmektedir. Bu formülde bulunan P($X$ | $C_i$) i sınıfından bir elemanın X olabilme ihtimalini, P($C_i$) i sınıfı için ilk ihtimalini, P($X$) herhangi bir elemanın X olabilme ihtimalini ve P($C_i$ | $X$ ) ise $X$ olan bir elemanın i sınıfından olma ihtimalini yani son ihtimali belirtmektedir.

Sınıflandırıcı için en yüksek son ihtimali gösteren max(P($C_i$ | $X$ )) istenilmektedir. Sınıfa dahil edilecek veri, en yüksek ihtimali veren test verisi olacaktır. P($X$) ihtimali bütün sınıflar için değişmediğinden, Formül 2.2' deki ihtimal için en yüksek değer istenir.

$$P(C_i|\boldsymbol{X}) = P(\boldsymbol{X}|C_i)P(C_i) \tag{2.2}$$

Bu yöntemde herhangi bir bilinmeyen $X$ sınıflandırılırken tüm $C_i$ sınıfı için ayrı ayrı P($X$ | $C_i$)P($C_i$) hesaplama işlemi yapılır. Bu hesaplama ile $X$ elemanı için $S_i$ sınıfı, en değerli olarak işaretlenir. Karşılaştırma yapmak amacıyla basite indirgenen bu durum, ayırt edici bütün durumların birbirinden saltık olması durumunda P($X$ | $S_i$) ise Formül 2.3' teki gibi belirtilebilir.

$$P(x|C_i) \approx \prod_{k=1}^{L} P(x_k|C_i) \tag{2.3}$$

$X$ durumu Formül 2.4' teki büyüklük belirten durum da olduğu gibi $C_i$ sınıfına dahil olma ihtimali diğer ihtimalden yüksek ise bu durumda Ci sınıfına dahil olur.

$$P(C_i)\prod_{k=1}^{L} P(x_k|C_i) > P(C_j)\prod_{k=1}^{L} P(x_k|C_j) \quad (2.4)$$

### 3.2.7. Regresyon Analizi (Regression Analysis)

Bir veri gurubu içerisindeki değişkenlerin birbirleri ile olan ilişkiyi tespit etmek amacıyla yapılan analize regresyon analizi denmektedir. Bu analizdeki değişkenlerin birbirleri ile olan ilişkilerini tespit edebilmek için bağımlı değişkenler ve bağımsız değişkenler olmak üzere iki ayrı gurup vardır.

Regresyon analizi $y = f(x) + Є$ formülü yardımı ile yapılmaktadır. Formülde bulunan $y$ eşitliğin sol tarafında yer almakla birlikte bağımlı değişkenleri ifade etmektedir. Formülde bulunan $x$ değişkeni ise, eşitliğin sağ tarafında yer almakla birlikte bağımsız değişkenleri ifade etmekte ve ağırlıklandırılırlar. Formülde bulunan bu ağırlık dereceleri bağımlı ve bağımsız değişkenler arasındaki ilişkiyi belirlemede rol oynamaktadır. Formülde bulunan bir diğer değişken olan Є ise, analiz yapılırken ki hata payını belirtmektedir.

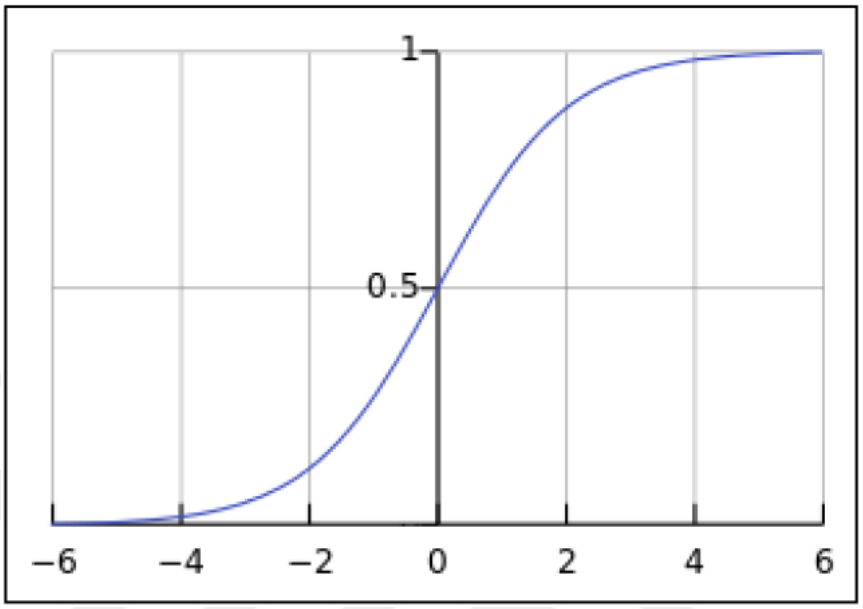
Bağımlı değişkenler de kendi aralarında doğrusal ve lojistik regresyon olmak üzere iki guruba ayrılmaktadır. Doğrusal regresyon incelenirken bu veri ile ilgili bağımlı değişken miktarı ölçülmektedir. Lojistik regresyon incelenirken ise, bağımlı değişkenlerin aralarından bir değişkenin seçilme ihtimali tespit edilmektedir. Lojistik regresyon konusu çalışmamızı yakından ilgilendirdiği için bu konu daha detaylı olarak aşağıdaki kısımda anlatılacaktır.

#### *3.2.7.1. Lojistik Regresyon Analizi (Logistic Regression Analysis)*

Nitel bağımlı değişkenlerinin bulunduğu analizlerde kullanılan lojistik regresyon, sınıflandırma problemleri açısından uygun bir yöntem olarak kullanılmaktadır. İkili ve çoklu sınıflandırma problemleri için kullanılabilen bu analiz türünün çalışmamızda ikili sınıflandırma ile alakalı aşamaları anlatılmaktadır.

Bu yöntemin amacı, bağımsız değişkenleri kullanıp sıradan bir bağımlı değişkeni sağlayabilmenin ihtimalini bulabilmektedir. Bir örnek ile açıklamak gerekirse, $K$ sınıf etiket değerini ifade etmesi durumunda, ikili sınıflandırma problemi için $K$ değişkeninin değeri 1 ile -1 arasında bir değer alabilmektedir. Buradaki amaç, $P(K=1|X=a)$ ihtimalini $a$ değişkeni ile belirtilen bağımsız değişkenler kullanılarak tespit edebilmektir (Shalizi).

Buradaki $K$ ihtimali, en çok olabilirlik yöntemi kullanılarak elde edilebilir. $P(a)$ fonksiyonunu doğrusal regresyon analizi ile elde etmeye çalıştığımızda sonucumuz 0 ile 1 arasında çıkmayacağı için bu sorunu aşabilmek amacıyla sonucumuzun 0 ile 1 arasında çıkması için bir yönteme ihtiyaç duyulmaktadır.




**Şekil 3.4:** Lojistik regresyon fonksiyonu [2]

Şekil 3.4' te de görüldüğü gibi, $a$ değişkeni hangi değeri alırsa alsın istenilen sonuç 0 ile 1 arasında olacaktır. Bu şekilde etiketi belirlenemeyen bir verinin bu şekle göre bakıldığında değeri belirlenebilecektir. Lojistik regresyon metoduna ait algoritmanın çok karmaşık olmaması, basit ve uygulanabilir olması, büyük akan verilerde kullanılabilme olasılığını arttırmaktadır.

### 3.2.8. K En Yakın Komşu Algoritması (K Nearest Neighborhood Algorithm)

Bu algoritmada bahsedilen K değeri, komşu olan verilerin sayısını ifade etmektedir. Demetlenmenin bir türü olan bu algoritmalar, girilen n adet ilk örnek örüntüye ve bu örüntülerin doğru bir şekilde sınıflandırılmasına göre sınıflandırılmamış bir örüntüyü en yakınında bulunan komşu guruba bağlamaktadır. K değerinin artış göstermesi ile bu sınıflandırmanın doğruluk derecesi artmaktadır. Bu yöntemin doğru sonuç verilmesi, girilen verinin bu algoritmaya uygunluğu ve kalitesi ile doğrudan ilgili olup ayrıca önceki veri guruplarına da ihtiyaç duymaktadır (Han & Kamber, 2001).

Bu algoritmanın adımları;

1- Test kümesinde bulunan her bir verinin öğrenme kümesinde bulunan verilere olan yakınlığı tespit edilir.

2- Her bir verinin öğrenme kümesinde bulunan verilere olan mesafeleri sıralanıp ilk “n” tanesi alınarak ortalamaları hesap edilir.

3- Ortalama değerleri, belirlenen seviyeden büyük olanlar iyi, küçük olanlar ise kötü olarak guruplandırılır.

Bu algoritmanın performansı; komşu sayısına, eşik değerine, benzerlik ölçümü işlemine ve öğrenme gurubunda bulunan verilerin iyi anlamda tanımlanan davranışlarının yeterli miktarda olmasına bağlıdır.

# 4 DESTEK VEKTÖR MAKİNELERİ

Destek vektör makinesi (DVM), orijinal adı support vector machine olarak anılır. Lineer ve lineer olmayan giriş verilerini analiz ederek aralarındaki örüntü problemini çözebilen, yaygın olarak kullanılan bir algoritmadır. Temelinde istatiksel öğrenme teorisi ve yapısal risk minimizasyonu vardır. Değişkenler arasındaki ilişki tespit ederek bunları sınıflandırmada kullanılan bir eğiticili öğrenme yöntemidir. Destek vektör makineleri sınıflandırma problemlerinde en çok kullanılan yöntemlerden birisidir. Doğrusal ve doğrusal olmayan verileri modelleyebilmesi, yüksek seviyede doğru sonuç vermesi, birbirinden bağımsız çok sayıda değişken ile çalışabilmesi, iyi bir sınıflandırma yapabilmesi tercih edilme sebeplerindendir. DVM yöntemi VM' de kullanılan diğer algoritmalar ile kıyaslandığında daha az karmaşık oluşu ve gelişme aşamasında yapılan işlem miktarının düşük olması sebebi ile diğer yöntemlerden farklıdır (Osowski, Siwekand, & Markiewicz, 2004). Bundan dolayı büyük verilerin sınıflandırmasında diğer yöntemlere göre daha uygundur.

DVM önceleri iki guruptan oluşan doğrusal bilgilerde kullanılmıştır. Sonraki aşamalarda birden fazla guruptan oluşan doğrusal olmayan bilgilerin sınıflandırılmasında kullanılmıştır. DVM'nin temelindeki en iyi yöntem ve amacı sınıfları birbirinden ayıran destek vektörlerini maksimum uzaklıkta bulunan optimum hiper düzlemi bulmaktır. Farklı guruplarda bulunan mesafe olarak birbirine en az mesafeli iki bilginin aralarındaki uzaklığın en yükseğe çekilmesi ile hiper düzlemin verilerin en iyi şekilde ayırmasını sağlayarak bu işlemi gerçekleştirilir.

DVM, bilgi gurubunun doğrusal bir şekilde ayrılıp ayrılamama haline bakılarak esas şekilde doğrusal olan ve doğrusal olmayan DVM kekinde toplan iki bölümde incelenebilir.

## 4.1 Doğrusal ayrılabilen veriler için Destek Vektör Makineleri

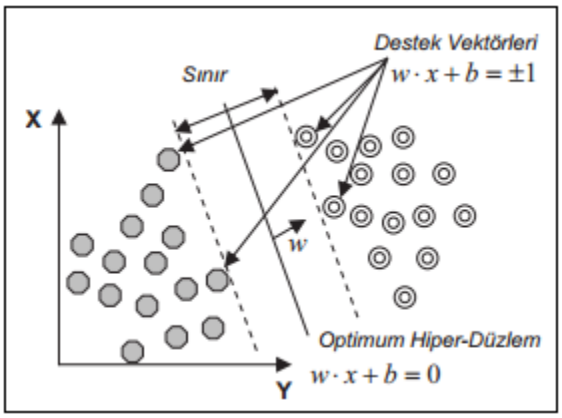
Destek vektör makineleri yardımıyla gruplandırma yapılırken doğrusal olarak ayrılabilen veriler için, geliştirilme aşamasında olan bilgi yardımı ile elde sağlanan karar fonksiyonu kullanılarak birbirlerinden ayrılan iki sınıf oluşturabilme amaçlanır. DVM girdi olarak seçilen veri kümesi içerisindeki fonksiyonlara göre bu fonksiyondan sağlanan çıktıları iki sınıfa ait olma şartı ile sınıflandırmaktadır (Meyrueıs, Soubarı, Guessoum, & Namane,2014). Elde edilen iki sınıfa ait birbirine aralarındaki mesafenin en az olduğu iki node' un mesafesi en yükseğe çekilerek hiper düzlem elde edilir. Bu sayede görünmeyen verilerde bu ayırımda en iyi şekilde yayılması sağlanır. Hiper düzlemin uzaklığı, verilerin en geniş sınırlarda ve en iyi şekilde sınıflandırıldığı anlamına gelir. Elde edilen hiper düzlem birden fazla olabilir. Aralarındaki mesafenin en az olduğu düğümlerin birbirlerine olan mesafesi en yükseğe çekilerek en yüksek seviyedeki mesafeyi veren hiper düzlem tercih edilir ve buna optimum hiper düzlem denir. Oluşan iki sınıfın sınırlarını belirleyen noktalardan geçen vektörler destek vektörleri ilgili gurubun çizgilerini tayin eder ve ayırma işlemi hiper düzlemine paralel bir düzlem hattında bulunur (Burges, 1998).

Doğrusal olarak ayrılabilen veriler için iki sınıflı oluşturulacak algoritmanın eğitilmesinde i = 1,2,…,n olarak kabul edildiğinde her bir $\{x_i, y_i\}$ için optimum hiper düzleme ait eşitsizlikler Formül 3.1' de gösterildiği gibidir. $x \in R^N$ ise N boyutlu bir uzayı ,$y \in \{+1, -1\}$ ikili sınıf etiketlerini, w ağrılık vektörünü ve $\beta$ eğilim değerini gösterir.

$$wx_i + \beta \geq +1, \quad y_i = +1 \; için$$

$$wx_i + \beta \leq -1, \quad y_i = -1 \; için \tag{3.1}$$

Optimum hiper düzleme paralel olan ve iki sınıf arasındaki sınırlarını oluşturan iki hiper düzlemin belirlenmesi optimum hiper düzlemin oluşabilmesi için gereklidir. Destek vektörleri $wx_i + \beta = \pm 1$ şeklinde ifade edilir (Vapnik,1995).




**Şekil 4.1:** Doğrusal olarak ayrılabilen veriler için optimum hiper düzlemin tayin edilmesi (Vapnik, 1995)

$$\frac{\|w\|}{2} \tag{3.2}$$

Formül 3.2' de gösterilen formül ile destek vektörleri arasındaki geometrik uzaklık hesaplanır. Optimum hiper-düzlemin sınırının maksimum olması gerekir. Bunun için *w* ifadesinin minimum hale getirilmesi gerekir.

$$\frac{2}{\|w\|} \tag{3.3}$$

Destek vektörleri arasındaki mesafeyi maksimum yapmak, Formül 3.3' te gösterilen ifadenin en aza indirgenmesi ile mümkün olacaktır. Buna göre en uygun hiper düzlemin seçilmesi için sınırlı optimizasyon probleminin çözümü gereklidir ve Formül 3.4' te gösterilen kısıtlara sahip Formül 3.5 'de gösterilen optimizasyon problemi ile bulunabilir (Vapnik, 1995).

Optimizasyon Problemi;

$$min\frac{[\|w\|]}{2} \tag{3.4}$$

Kısıtlar;

$$y(wx_i + \beta) - 1 \geq 0 \; ve \; y \in \{+1, -1\} \tag{3.5}$$

Elde edilen optimizasyon problemi Lagrange denklemleri kullanılarak çözülebilir.

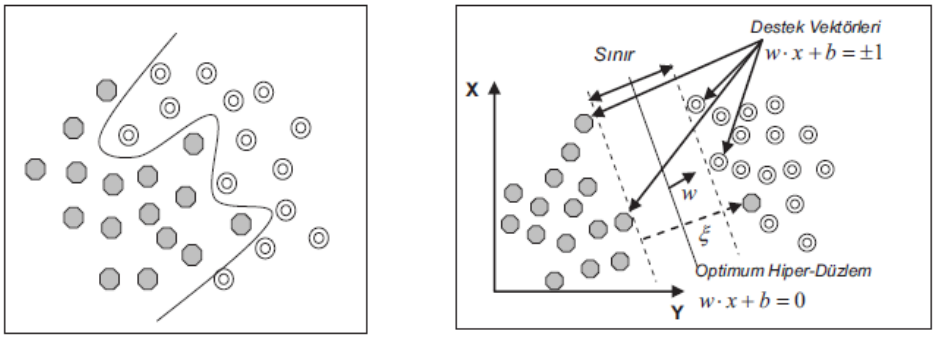
$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{k} \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^{k} \alpha_i \tag{3.6}$$

Buradan yola çıkarak Formül 3.6' da görülen eşitlik elde edilir. Bu eşitliğin sonucunda doğrusal olarak ayrılabilen iki sınıflı bir problem için karar fonksiyonu Formül 3.7' de gösterildiği gibi ifade edilebilir (Osuna, Freund, & Girosi, 1997).

$$f(x) = sign\left( \sum_{i=1}^{k} \lambda_i y_i (x \cdot x_i) + b \right) \tag{3.7}$$

## 4.2 Doğrusal olarak ayrılamayan veriler için Destek Vektör Makineleri

Veri madenciliğinde veriler her zaman doğrusal olarak ayrılamayabilirler. Bu durumda ortaya doğrusal olarak ayrılamayan verilerin sınıflandırılması ile ilgili problem ortaya çıkar. Doğrusal olarak ayrılamayan veri kümesine ait eğitim verilerinin bir kısmının optimum hiper düzlemin diğer tarafında kalmasından kaynaklanan problemi çözmek için pozitif yapay değişken adı verilen ( $i\ \xi$ ) 'nin tanımlanması gerekmektedir. Doğrusal olarak ayrılabilen DVM yönteminde olduğu gibi vektörler arasındaki mesafeyi yani sınırın maksimum hale getirilmesi ve yanlış sınıflandırma hatalarının minimum hale getirilmesi arasındaki denge pozitif değerler alan ve C ile gösterilen bir denge parametresi ($0 < C < \infty$) tanımlanması ile gerçekleşir (Cortes & Vapnik,, 1995).




**Şekil 4.2:** (a) Doğrusal olarak ayrılamayan veri seti, (b) Doğrusal ayrılamayan veri setleri için hiper-düzlemin belirlenmesi (Kavzaoğlu & Çölkesen, 2010)

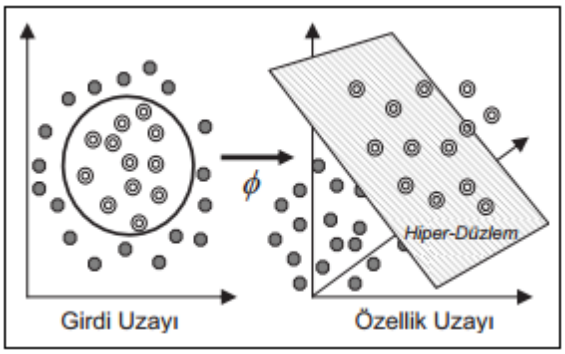
Şekil 4.2’ de gösterilen doğrusal olarak ayrılamayan veri setlerinde hiper düzlemin belirlenmesi için optimizasyon problemi Formül 3.8’ de gösterilmiştir.

$$min\left[\frac{\|w\|^2}{2}+C\cdot\sum_{i=1}^{r}\xi_i\right] \tag{3.8}$$

Buna bağlı sınırlamalar ise Formül 3.9’ da gösterilmiştir.

$$y_i(w\cdot\varphi(x_i)+b)-1\geq 1-\xi_i$$
$$\xi_i\geq 0 \quad \text{ve} \quad i=1,\ldots,N \tag{3.9}$$

Girdi uzayında ayırımı doğrusal olarak yapılamayan veriler, optimizasyon probleminin çözümü için Şekil 4.3’ de gösterilen özellik uzayı şeklinde nitelendirilen çok boyutlu uzaya dönüştürülerek boyutu yüksek bir uzayda gösterimi gerçekleştirilir. Bu şekilde yapılan analizlerde sınıflarının birbirinden ayrılabildiği, optimum hiper düzlem belirlenebilmekte ve verilerin ayrımı doğrusal olarak yapılabilmektedir.




**Şekil 4.3:** Kernel fonksiyonu ile verinin daha yüksek bir boyuta dönüştürülmesi (Kavzaoğlu & Çölkesen, 2010)

$$K(x_i,x_j)=\varphi(x)\cdot\varphi(x_j) \tag{3.10}$$

Destek vektör makineleri Formül 3.10’da gösterilen kernel fonksiyonu olarak bilinen, matematiksel formülü kullanırlar. Bu formül yardımı ile doğrusal olmayan dönüşümler yapabilirler. Böylelikle veriler yüksek boyutta doğrusal olarak sınıflara ayrılabilirler.

$$f(x) = sign\left( \sum_i \alpha_i y_i \varphi(x) \cdot \varphi(x_i) + b \right) \tag{3.11}$$

Doğrusal olarak ayrılamayan iki sınıflı verilerde kernel fonksiyonu kullanılarak oluşturulan karar kuralı Formül 3.11' de gösterilmiştir. (Osuna vd. , 1997).

En sık kullanılan kernel fonksiyonları polinom, normalleştirilmiş polinom, radyal tabanlı fonksiyon ve Pearson VII (PUK) fonksiyonudur. Kernel fonksiyonu, DVM ile sınıflandırma işlemleri için kullanılırsa optimum parametrelerin belirlenmesi gerekmektedir. Bu fonksiyona ait formüller aşağıda sırası ile verilmiştir ve formüllerde görülen ($d$) polinom derecesi, ($\gamma$) kernel boyutu ve ($\sigma$ , $\omega$ ) pearson genişliği parametreleridir.

$$K(x,y) = ((x \cdot y) + 1)^d \tag{3.12}$$

$$K(x,y) = \frac{((x \cdot y) + 1)^d}{\sqrt{((x \cdot x) + 1)^d ((y \cdot y) + 1)^d}} \tag{3.13}$$

$$K(x,y) = e^{-\gamma \left\| (x - x_i) \right\|^2} \tag{3.14}$$

$$\frac{1}{\left[ 1 + \left( \frac{2 \cdot \sqrt{\left\| x - y \right\|^2 \sqrt{2^{(1/\omega)} - 1}}}{\sigma} \right)^2 \right]^{\omega}} \tag{3.15}$$

Matematiksel olarak ifade edilen eşitsizliklerde polinom ve radyal tabanlı kernel formüllerinin daha anlaşılabilir olduğu görülmektedir. Polinom kernel fonksiyonunda d ile ifade edilen polinom derecesindeki artış algoritmanın karmaşık bir şekil almasına sebep olmaktadır. Bunun sonucunda sınıflandırma doğruluğu düşebilmekte ve işlem süresi uzamaktadır. Radyal tabanlı kernel fonksiyonunda ise kernel boyutu ( $\gamma$ ) aldığı değerlerde sınıflandırma performansına etkinin az olduğu gözlemlenmiştir. (Hsu, Chang, & Lin,2010).

Normalleştirilmiş polinom fonksiyonu polinom kernel fonksiyonunun genelleştirilmiş halidir. Veri setinin normalleştirilmesi yerine polinom kernel için kullanılan formülün normalleştirilmesi ile oluşturulur. PUK kerneli

diğerlerinden farklı olarak iki parametre alır ve bu da diğer kernel fonksiyonlarına göre daha karmaşık bir yapıda olmasını açıklar. Sınıflandırma doğruluğuna etki eden bu parametreler için uygun çiftin bulunması önemlidir.

Optimum hiper düzlem belirlenirken destek vektör makineleri için kullanılan C parametresinin alacağı değerler sınıflandırma doğruluğuna doğrudan etki eder. Alacağı değerler normalden çok yüksek seçilirse ya da çok düşük seçilirse sınıflandırma doğru yapılamaz. Eğer $C = \infty$ ise, bu parametre ile sadece doğrusal ayrılabilen veri setleri kullanılabilir.

Sonuç olarak parametre değerlerinin uygun seçimi DVM sınıflandırma performansını daha yüksek seviyeye getireceği söylenebilir. Çapraz doğrulama yaklaşımı ile sınıflandırma modelinin performansı değerlendirilir. Bu teoride veri seti iki kısımda incelenir. Birinci veri seti eğitim amaçlı kullanılır ve modelin oluşması sağlanır. İkinci kısım ise oluşturulan modelin test verilerine uygulanması ile performansın ölçülmesini sağlanır. Performans ölçümünde çapraz geçerlilik yöntemi kullanılarak en iyi kernel fonksiyonun ve uygun parametreler ile temel model oluşturulur.

# 5 UYGULAMA

Bu bölümde tez çalışması için kullanılacak veriler belirlenmiş ve bu veriler üzerinde destek vektör makinesi algoritması kullanılarak müşteri kaybı analizi tahmininin nasıl yapıldığı gösterilmektedir.

Verilerimiz, dünya çapında ve aynı zamanda ülkemizde de lojistik sektöründe faaliyet gösteren bir firmadan alınmıştır. Bu firma normal şartlar altında her ay ne kadar müşteri kaybı olduğunu tespit edebilmekte fakat bu müşterilerin kaybedileceğine dair öngörüleri yeterli miktarda bulunmamaktadır. Yapılan analizlerde ilgili müşterinin önceki yıl ve mevcut yıldaki gönderi adetleri ve navlun ücretleri karşılaştırılıp bu durum değerlendirmeye alınmaktadır. Konusunda uzman kişilerle görüşülüp daha önce yapılan müşteri kaybı analizi prosedürü incelenmiştir. Bu inceleme sonucunda 2015 ve 2016 yıllarında bu firma ile çalışan müşterilerin bilgileri alınmıştır. Çalışmamızda kullanılmak üzere, toplamda 5.000 adet müşteriye ait işlenmemiş verilerden faydalanılmıştır.

Mevcut firma müşteri kaybı analizini yıllık olarak yapmaktadır. Aldığımız veriler ay bazlı olup, her müşterinin o ay içerisindeki toplam gönderi adeti ve toplam navlun ücreti bilgisini içermektedir. Yapacağımız çalışmada müşteri kaybı analizini yıllık olarak değil, çeyrekler halinde inceleyerek önümüzdeki ilk çeyrek içerisinde, firmayı terk etme eğilimi gösteren müşterileri tespit etmeyi amaçlamaktadır.

=EĞER(**Önceki_Dönem_Hacim**=0;"CONVERSION";EĞER(**Dönem_Hacim**=0;"LOST";
EĞER(**Hacim_Farkı**=0;"NOCHANGE";EĞER(**Hacim_Farkı**<0;"DECLINER";"GAINER"))))

(5.1)

Verilerimizi öncelikle 5 farklı gruba ayrılmıştır. Gruplara ayırmak için Formül 5.1' de gösterilen matematiksel bir formül kullanılmıştır. Bu matematiksel formülde kullanılan parametreler aşağıda açıklanmıştır:

**Önceki Dönem Paket Adedi:** Mevcut müşterinin analizinin yapıldığı ilgili önceki dönem aylarına ait toplam gönderi adedini gösterir.

**Önceki Dönem Navlun Toplamı**: Mevcut müşterinin analizinin yapıldığı ilgili önceki dönem aylarına ait toplam navlun bedelini gösterir.

**Önceki Dönem Çalışılan Gün Sayısı**: Mevcut müşterinin analizinin yapıldığı ilgili önceki dönem aylarına ait toplam çalışılan gün sayısını gösterir.

**Dönem Paket Adedi:** Mevcut müşterinin analizinin yapıldığı ilgili mevcut dönem aylarına ait toplam gönderi adedini gösterir

**Dönem Navlun Toplamı:** Mevcut müşterinin analizinin yapıldığı ilgili mevcut dönem aylarına ait toplam navlun bedelini gösterir

**Dönem Çalışılan Gün Sayısı**: Mevcut müşterinin analizinin yapıldığı ilgili mevcut dönem aylarına ait toplam çalışılan gün sayısını gösterir.

**Önceki Dönem Hacim**: Müşterinin, önceki dönemdeki ilgili aylara ait ortalama paket hacmini verir. Önceki dönem paket adedi parametresinin önceki dönem çalışılan gün sayısına bölünmesi ile elde edilir.

**Dönem Hacim:** Müşterinin, mevcut dönemdeki ilgili aylara ait ortalama paket hacmini verir. Dönem paket adedi parametresinin dönem çalışılan gün sayısına bölünmesi ile elde edilir.

**Önceki Dönem Gelir**: Müşterinin, önceki dönemde ilgili aylara ait ortalama net gelir bilgisini verir. Önceki dönem navlun toplamı parametresinin önceki dönem çalışılan gün sayısına bölünmesi ile elde edilir.

**Dönem Gelir**: Müşterinin, mevcut dönemde ilgili aylara ait ortalama net gelir bilgisini verir. Önceki dönem navlun toplamı parametresinin önceki dönem çalışılan gün sayısına bölünmesi ile elde edilir.

**Hacim Farkı:** Önceki dönem ve mevcut dönem arasındaki hacim farkını temsil eder.

**Grup**: Müşterilerin gruplandırılmasında kullanılan matematiksel formül uygulanması sonucunda elde edilen grup tanımıdır.

İlgili parametrelerin aldığı değerler kullanılarak formüle uygulandığında elde edilen gruplar ve hesaplanma detayları aşağıda verilmiştir:

**No Change:** Eğer hacim farkı parametresi 0 değerini almış ise grup, no change olarak belirlenir. Yani önceki dönem ile mevcut dönem içerisindeki ilgili aylara

ait paket hacmi değişmemiş ise mevut müşteri kayıp müşteri listesine dahil edilmez.

**Conversion:** Eğer müşterinin önceki döneme ait paket hacmi 0 ise ve mevcut döneme ait paket hacminde artış gözlenmiş ise bu bir kazanılmış müşteridir ve conversion grubuna dahil edilir.

**Gainer:** Eğer mevcut müşterinin ilgili döneme ait hacim farkı değeri 0'dan büyük ise bu müşteri gönderi hacmini yükseltmiştir ve gainer grubuna dahil edilir.

**Decliner:** Eğer mevcut müşterinin ilgili döneme ait hacim farkı değeri 0'dan küçük ise bu müşteri gönderi hacmi küçülmüştür ve decliner grubuna dahil edilir. Bu gruptaki müşteriler kaybedilmeye aday müşterilerdir.

**Lost:** Eğer müşterinin önceki döneme ait paket hacmi mevcut ise ve mevcut dönem ile ilgili aya ait paket hacmi 0 değerini almış ise bu bir kaybedilmiş müşteridir ve lost grubuna dahil edilir.

Verilerimiz destek vektör makinesi algoritmasında kullanılacağı için normalleştirme işlemine tabii tutulur. Bu durumda veriler sayısal değerler alır.

**Çizelge 5.1:** Verilerin sınıf adları ve değerleri

| Sınıf Adı | Değer |
|---|---|
| No Change | 0 |
| Conversion | 1 |
| Gainer | 2 |
| Decliner | 3 |
| Lost | 4 |

Çizelge 5.1' de verilerimiz normalleştirilme yapılarak aldığı yeni değerler gösterilmektedir. 2015 yılı 4 çeyrek ve 2016 yılı 4 çeyrek verileri toplamda 8 çeyrek olacak şekilde veriler hazırlanmıştır. Oluşturulan veri setinde ilk 7 çeyrek girdi değerini, 2016 son çeyreği çıktı değerini temsil etmektedir.

Müşteri kaybı analizi yapılırken kazanılan ve kaybedilen müşterilerin davranışları önemlidir. Bir müşterinin kaybedilme eğilimi bu iki müşteri tipi karşılaştırıldığında ortaya çıkmaktadır. Bu nedenle çıktı değerleri gainer yani kazanılmış müşteri ve lost kaybedilen müşterilerin 7 çeyrek boyunca sergilediği davranış tezimize konu olmuştur.

Spyder (Python 3.6) uygulaması veri setimizi eğitecek ve test edilecek destek vektör makinesi algoritmasını geliştireceğimiz platform olarak seçilmiştir. Programımız python dilinde geliştirilmiştir.

**Çizelge 5.2:** Örnek veri seti

| Girdi 1 | Girdi 2 | Girdi 3 | Girdi 4 | Girdi 5 | Girdi 6 | Girdi 7 | Çıktı |
|---|---|---|---|---|---|---|---|
| 1 | 3 | 2 | 3 | 3 | 2 | 2 | 2 - **0** |
| 1 | 4 | 1 | 2 | 3 | 3 | 2 | 2 - **0** |
| 1 | 2 | 3 | 3 | 4 | 1 | 3 | 2 - **0** |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 - **0** |
| 1 | 2 | 2 | 2 | 3 | 3 | 2 | 2 - **0** |
| 3 | 3 | 2 | 3 | 2 | 4 | 1 | 4 - **1** |
| 1 | 1 | 1 | 1 | 1 | 3 | 2 | 4 - **1** |
| 1 | 2 | 2 | 3 | 3 | 2 | 3 | 4 - **1** |
| 1 | 3 | 4 | 1 | 1 | 1 | 1 | 4 - **1** |
| 1 | 2 | 2 | 3 | 4 | 1 | 3 | 4 - **1** |

Çizelge 5.2' de oluşturulan veri setinden örnek bir bölüm gösterilmiştir. Örnek veri setimizi incelediğimizde 2 değeri gainer müşterileri ve 4 değeri lost müşterileri temsil etmektedir. Uygulayacağımız algoritma dikkate alındığında bu iki değeri 0 ve 1 değerlerini dönüştürdüğümüzde elde edeceğimiz sonuçlar daha sağlıklı bir uygulama yapılmasını sağlayacaktır.

```
import pandas as pd
import numpy as np
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.feature_selection import SelectFromModel
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
import sklearn.preprocessing as pr
from sklearn.multiclass import OneVsRestClassifier
from sklearn.metrics import roc_curve, auc,  classification_report
import matplotlib.pyplot as plt
```

**Şekil 5.1:** Python yardımcı paketler

Şekil 5.1' de uygulamamızda bizde kolaylık sağlayacak paketler gösterilmiştir. Uygulamamızı geliştirebilmek için bu kütüphaneleri projemize referans olarak eklememiz yeterli olacaktır. Bunun için ilk önce paketler import komutu ile uygulamamıza dahil edilir.

```
churn=np.array(pd.read_csv('C:\\Users\\Buket\\.spyder-py3\\Projects\\churn8.csv',
                           sep=';'))
```

**Şekil 5.2:** Veri setinin diziye alınması

Programımız için öncelikle gerekli olan verileri projeye bir dizi halinde almamız gerekmektedir. Bunun için Şekil 5.2' de görüldüğü üzere pandas ve numpy kütüphanelerinden yararlanmaktayız.

Numpy, numerical python ifadesinin yani sayısal python söz dizisinin kısaltılmış halidir. Veri merkezli çalışmalar için en çok tercih edilen kütüphanedir. Numpy ile hızlı çalışan çok boyutlu diziler elde edilebilir. Diziler ve dizi ile çalışan matematiksel işlemler için uygundur. Dizi tabanlı veri setlerini oluşturma, okunma ve yazma fonksiyonlarını barındırır. Özellikle sayısal verilerde dizileri sıralama gibi işlemlerde en etkili ve en konforlu python çözümünü sunar.

Pandas üst düzey veri yapıları ve işleme araçlarını barındırır. Python'da hızlı ve kolay veri analizi yapmak için geliştirilmiştir. Pandas numpy'nin üzerine inşa edilmiştir. Bu nedenle numpy merkezli uygulamalarda kullanımı kolaylaştırmaktadır. Pandas, açık kaynaklı bir kütüphanedir ve BSD lisansına sahiptir. Python programlama dili için yüksek performanslı, kullanımı kolay veri yapıları ve veri analizi araçları sunmaktadır.

Pandas kütüphanesi uygulamamızda pd kısaltması ile tanımlanmıştır. Verilerimizi okumak için read_csv() fonksiyonu kullanılır. Bu fonksiyonun ilk parametresi verinin bulunduğu csv dosyasıdır. Bu dosya bulunduğu konum ile birlikte yazılır. İkinci parametrede ise oluşturulacak dizilerin hangi ayırıcı karakter ile belirleneceği bilgisi verilir.

Numpy kütüphanesi içerisinde bulunan array() fonksiyonu ile dizi tipinde veriler tanımlanabilir. Numpy'ın bir nesnesi olarak array() fonksiyonu çağrılır ve csv dosyasından okunan veriler ile dizi oluşturulur. Dizimiz churn değişkenine atanır.

```
y = churn[:,7]
X = churn[:,0:7]
```

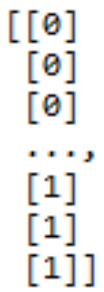
**Şekil 5.3:** Girdi ve çıktı değerlerinin sayısı

Destek vektör makinesi algoritmasında kullanılmak üzere Şekil 5.3' te gösterilen y değişkenine churn değişkeninde bulunan dizimizin çıktı değerleri atanır. Girdi değerleri ise x değişkenine atanır. Verilerimiz toplamda 2015 yılı 4 çeyrek ve 2016 yılı ilk 3 çeyrek durumları 7 girdi, 2016 son çeyrek bilgisi ise 1 çıktı oluşturacak şekilde tasarlanmıştır.

```
y = pr.label_binarize(y, classes=[0,1])
```

**Şekil 5.4:** Çıktı değerlerinin sabit sınıflara ayrılması

Uygulamamızda referans olarak yer alan sklearn.preprocessing, python projelerinde makine öğrenmesi için en çok tercih edilen bir kütüphanedir. Kütüphane pr olarak tanımlanmıştır ve label_binarize() fonksiyonunu ile çıktı değerlerimiz için sabit sınıflar oluşturulmasını sağlar. Fonksiyonun aldığı ilk parametre hangi veriler için sabit sınıf tanımlanacağıdır. İkinci parametre ise bu sınıfın hangi değerlere sahip olacağıdır. Şekil 5.4'te işlem gösterilmiştir.




**Şekil 5.5**: Çıktı değerlerinin aldığı sınıf değerleri

Bu sayede çıktı verilerimiz Şekil 5.5' te görüldüğü üzere algoritma öğrenmesinde 0 ve 1 değerleri ile temsil edilecektir.

```
n_classes =y.shape[1]
```

**Şekil 5.6:** Çıktı değerlerinin toplam sınıf sayısı

Şekil 5.6' da gösterilen shape[1] fonksiyonu y dizinin kaç kolondan oluştuğunu verir. Bu bizim çıktı değerimizin kaç sınıfa ayrıldığını gösteren bilgiyi verir ve uygulamamızda n_classes değişkenine atanmıştır.

```
clf = ExtraTreesClassifier()
clf = clf.fit(X, y)
sum=0
j=0
for i in clf.feature_importances_:
    print(str(j)+':'+str(i))
    sum+=i
    j+=1
```

**Şekil 5.7:** Değişken önem analizi

Şekil 5.7' de gösterilen ExtraTreesClassifier() fonksiyonu ile girdi verilerimiz için değişken önem analizi yapılır. Bu fonksiyon bir takım karar ağacı algoritmalarını kullanarak her değişkene bir değer atar ve gerekli nesneler clf değişkenine atanmıştır. Değişken önem analizi veriler içerisinde modele konu olmayacak verilerin çıkarılması için gereklidir. Değişken önem analizinde kullanılan karar ağacı algoritmasının test verileri clf.fit() fonksiyonu ile x ve y parametrelerini alarak tamamlanmaktadır.

```
0:0.115567994733
1:0.122133447689
2:0.0935844192411
3:0.10682893515
4:0.116744562308
5:0.261789846375
6:0.183350794505
```

**Şekil 5.8:** Değişken önem analizi sonucu

Uygulanan karar ağacı algoritmaları sonucunda veri setimizde bulunan girdi değerlerimizin önem dereceleri Şekil 5.8' de gösterilmiştir.

```
model = SelectFromModel(clf,threshold=0.11,prefit=True)
X_new = model.transform(X)
```

**Şekil 5.9:** Modelin belirlenmesi

Makine öğrenmesi için gerekli veri setimiz belirlendikten ve önem analizi yapıldıktan sonra söz konusu asıl modelin belirlenmesi Şekil 5.9' da gösterilen SelectFromModel() fonksiyonu kullanılarak gerçekleşir. Bu fonksiyonun aldığı ilk parametre clf değişkenidir. Fonksiyonda belirtilen threshold parametresi ile modelimizde özellik seçiminde önem derecesi 0.11 değerinden düşük verilerin dikkate alınmayacağını belirtir. Verilerimiz uygulamamızda doğrudan kullanılabilir ve herhangi bir dönüşüm işlemi gerektirmeyen sayısal veriler olduğu için prefit parametresini true olarak atanır. Oluşan modelde transform() fonksiyonuna girdi değerlerini barındıran x parametresini verdiğimizde, önem derecesi 0.11 değerinden yüksek olan veriler alınarak kullanılacak esas değerler ile oluşan veri dizisi x_new parametresine atanır.

```
X_train, X_test, y_train, y_test =
train_test_split(X_new, y, test_size = 0.33, random_state=2, stratify=y)
```

**Şekil 5.10:** Öğrenme ve test verilerinin belirlenmesi

Öğrenme ve test verilerinin oluşturulması için Şekil 5.10 'da gösterilen train_test_split() fonksiyonu kullanılır. Aldığı ilk parametre girdi değerleri, ikinci parametre ise çıktı değerleridir. Test olarak kullanılacak veri sayısı test_size parametresi ile belirtilir ve toplam verinin yüzde kaçının teste ayrılacağına karar verilir. 2 değerini alan random_state parametresi, test ve öğrenme verileri için her defasında aynı veri setlerinin kullanılmasını sağlar.

Fonksiyonun aldığı son parametre y değişkeninin katmanlı bir şekilde bölünmesini sağlar. Algoritma bunu sınıf etiketle için kullanır.

```
X_train=pr.scale(X_train)
X_test=pr.scale(X_test)
random_state = np.random.RandomState(0)
```

**Şekil 5.11:** Normalizasyon işlemleri

Algoritmamızda kullanılacak girdi verileri için Şekil 5.11' de gösterilen sklearn.processing (pr) kütüphanesinde bulunan scale() fonksiyonu ile normalizasyon işlemi gerçekleştirilir. İndeksleme yapmak için ratstgele sayılara ihtiyaç duyulur. Bu sebeple rastgele sayı üretmek için numpy kütüphanesinin random.RandomState() fonksiyonu kullanılır. Bu fonksiyon rastgele sayı üreticisini temsil eden random_state parametresine atanır.

```
svm=OneVsRestClassifier(SVC(C=0.1,gamma=0.3,kernel='rbf',
                            probability=True,random_state=random_state))
```

**Şekil 5.12:** Destek vektör makinesi algoritmasının uygulanması

Şekil 5.12' de görülen OneVsRestClassifier() fonksiyonu çok sınıflı, çoklu etiket stratejisini uygular. Bu stratejide her sınıf için sadece bir sınıflandırıcı yani classifier vardır. Her sınıf sadece tek bir sınıflandırıcı tarafından temsil edilir ve ilgili sınıflandırıcı ile incelenerek sınıfı hakkında bilgi sahibi olmak mümkündür. Bu çoklu sınıf, sınıflandırma için en yaygın olarak kullanılan bir stratejidir.

Destek vektör makinesi algoritması sklearn kütüphanesi ile bize hazır sunulmaktadır. İstenilen parametreler verildiğinde bize uygun sonucu vermektedir. DVM uygulanabilmesi için gerekli parametreler mevcuttur. İlk parametre C ile ifade edilen hata oranını temsil eder. Kernel parametresi algoritmada kullanılacak çekirdek türünü belirtir. Çekirdek türü linear, poly, rbf, sigmoid, precomputed değerlerini alabilir. Gamma parametresi rbf, poly and sigmoid çekirdek türleri için çekirdek katsayını belirtir. Probability parametresi algoritma içerisinde olasılık tahminin kullanılıp kullanılmayacağını belirtir. Verilerimiz doğrusal olarak ayrılamayan veriler olduğu için olasılık formülünün hesaba katılması doğruluğumuzu arttıracaktır. Fonksiyonda kullanılan

random_state parametresi olasılık tahmini için verileri karıştırırken kullanacak olan sahte rasgele sayı üretecini temsil eder.

```
model=svm.fit(X_train,y_train)
```

**Şekil 5.13:** Algoritma girdi ve çıktı değerlerinin tanımlanması

Sınıflandırıcılar Şekil 5.13' te görüldüğü gibi öğrenme için ayrılan girdi ve çıktı verilerine uygulanır.

```
y_pred = model.predict(X_test)
print(svm.score(X_test,y_test))
```

**Şekil 5.14:** Algoritma sonucunun değerlendirilmesi

Şekil 5.14' te görüldüğü üzere predict() fonksiyonu ile modeldeki test verilerinin çıktıları değerlendirilir. Yapılan test sonucunda elde edilen tahminleme değerleri Şekil 5.15' te gösterildiği üzere ortalama 0.73 değerinde doğru varsayımda bulunmaktadır.

```
Ortalama tahmin değeri:  0.736363636364
```

**Şekil 5.15:** Ortalama tahmin değeri

```
y_score = svm.fit(X_train, y_train).decision_function(X_test)
```

**Şekil 5.16:** Öğrenme verilerine öğrenilen kuralın uygulanması

Test verileri ile algoritma öğrenme gerçekleştirildikten sonra Şekil 5.16' da görülüğü üzere asıl veriler modelimize set edilir. Verilerin hiper düzleme uzaklığını belirtmek için ise decision_function() fonksiyonu kullanılır. Aldığı X_test parametresine daha önce uzaklık değerini tanımlamıştık.

```
print(classification_report(y_test,y_pred))
```

**Şekil 5.17:** Sınıflandırma sonucunun yazılması

**Çizelge 5.3:** Sınıflandırma sonucu değerleri

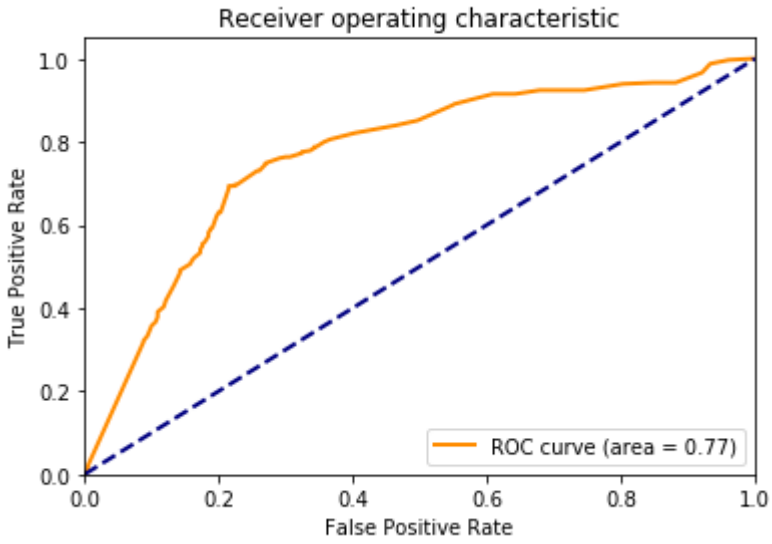
| Sınıf | Precision | Recall | F1 - Score | Örnek Sayısı |
|---|---|---|---|---|
| 0 | 0.72 | 0.78 | 0.75 | 330 |
| 1 | 0.76 | 0.69 | 0.73 | 330 |
| Ortalama/ Toplam | 0.74 | 0.74 | 0.74 | 660 |

Modelimiz öğrenmesini tamamladıktan sonra giriş değerlerini verdiğimizde Şekil 5.17' da gördülüğü üzere classification_report() fonksiyonu ile çıktı değerlerinin doğruluk oranlarını görebiliriz. Bu fonksiyon için ilk parametre öğrenme için kullanılan y_test çıktı parametresidir. İkinci parametre ise olasılık tahmini sonucunda elde edilen hedeflerdir. Buna bağlı olarak elde edilen sonuçlar Şekil Çizelge 5.3' te gösterilmiştir. Gainer ve lost olarak ayrılan verilerde 0 değerinin doğru olarak sınıflandırılması 0.72 oranında tahmin edilebiliyor iken, 1 değerinin doğru olarak sınıflandırılması 0.76 oranında tahmin edilebilmektedir. Bu durumda algortimamız bu veriler üzerinde sınıflandırma yaparken 0.74 oranında veriyi doğru sınıfına atayabilmektedir. Elde edilen sonuçlar doğrultusunda bir verinin 0 sınıfına ait olma durumu 0.75 oranında doğru tahmin edilebilmektedir. 1 sınıfına ait olma durumu ise 0.73 oranında doğru tahmin edilebilmektedir. Bu durumda bir verinin hangi sınıfa ait olduğu 0.74 oranında doğruluk payı ile yerleştirilmektedir.

```
fpr["micro"], tpr["micro"], _ = roc_curve(y_test.ravel(), y_score.ravel()
roc_auc["micro"] = auc(fpr["micro"], tpr["micro"])
plt.figure()
lw = 2
plt.plot(fpr["micro"], tpr["micro"], color='darkorange',
         lw=lw, label='ROC curve (area = %0.2f)' % roc_auc["micro"])
plt.plot([0, 1], [0, 1], color='navy', lw=lw, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.show()
```

**Şekil 5.18:** ROC grafiğinin oluşturulması

DVM sınıflandırıcı çıktı kalitesi değerlendirilmesi ve grafik olarak gösterilebilmesi için Şekil 5.18 'de gösterilen kodlardan yararlanılır. Receiver Operating Characteristic (ROC) olarak bilinen grafik ile gösterilir.




**Şekil 5.19:** Sonuçların ROC grafiğinde gösterilmesi

Verilerimiz üzerinde uygulanan DVM algoritması sonucunda oluşan ROC grafiği Şekil 5.19' de gösterilmiştir. ROC grafiğinde Y ekseni gerçek pozitif oranı, X ekseni yanlış pozitif oranını temsil eder. Eğri sol üst köşeye ne kadar yakınsa değerlendirme o kadar iyidir. Bu durumda 1.0 değerinde ROC eğrisi ideal olarak kabul edilir. Eğrinin altında kalan alan ne kadar büyük ise o kadar iyi kaliteli sonuç elde edildiği şeklinde yorumlanır. ROC eğrilerinin eğimi de önemlidir. Çünkü yanlış pozitif oranı minimuma indirirken gerçek pozitif oranı en üst düzeye çıkarmak idealdir.

ROC eğrileri tipik olarak bir sınıflandırıcının çıktısını incelemek için ikili sınıflandırmada kullanılır. ROC eğrisini ve ROC alanını çoklu sınıf veya çoklu etiket sınıflandırmasına genişletmek için çıktıyı iki katına çıkarmak gerekir. Her etiket başına bir ROC eğrisi çizilebilir. Ancak etiket gösterge matrisinin her elemanını, ikili tahmin olarak düşünerek bir ROC eğrisi çizebilir.

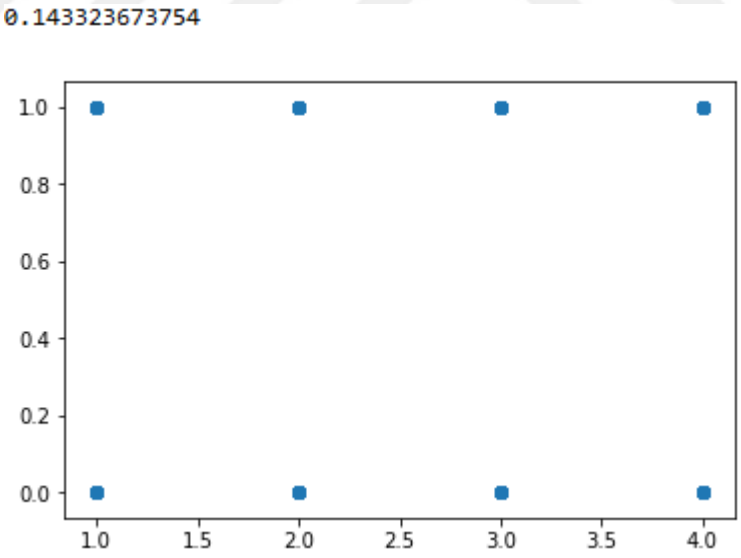
Şekil 5.19' da görüldüğü gibi çıktı kalitemiz 0.77 eğri oranına sahiptir. İdeal eğri oranına yakın bir değerde sonuç elde ettiğimiz görülmektedir. Bu da uygulamamızda yüksek oranda başarı elde ettiğimizi göstermektedir.

# 6 DENEYSEL ÇALIŞMALAR

Bu bölümde, uygulamamızda kullanılan veri setimiz için farklı girdi ve çıktı değerleri denenmiştir. Ayrıca verilerimiz üzerinde lojistik regresyon, rastgele orman ve destek vektör makinesi algoritmaları üzerinde test edilmiştir. Elde edilen sonuçlar doğrultusunda sınıflandırma ve tahminle işlemlerinde DVM algoritmasının daha iyi çalıştığı belirlenmiştir. Bu doğrultuda DVM algoritması için kullanılan parametre değerleri test edilerek en uygun sonuca ulaşılmaya çalışılmıştır.

## 6.1 Algoritmanın Seçilmesi

Verilerimiz lojistik regresyon, rastgele orman ve destek vektör makinesi algoritmaları üzerinde denenmiştir.




**Şekil 6.1:** Lojistik regresyon analizi sonucu veri dağılımı

Verilerimiz lojistik regresyon algoritması kullanılarak test edilmiş ve bunun sonucunda verilerimiz sınıflandırıldığında elde edilen tahmin değeri 0.14 olduğu görülmüştür. İkili sınıflandırma sonucunda verilerin dağılımı Şekil 6.1' de gösterilmiştir. Grafiği göre veriler belli bölgelerde toplanmış ve dağılımı istenilen ayrışmayı vermemiştir.

**Çizelge 6.1:** Lojistik regresyon analiz sonucu

| Sınıf | Precision | Recall | F1 - Score | Örnek Sayısı |
|---|---|---|---|---|
| 0 | 0.68 | 0.67 | 0.68 | 330 |
| 1 | 0.68 | 0.68 | 0.68 | 330 |
| Ortalama/ Toplam | 0.68 | 0.68 | 0.68 | 660 |

Çizelge 6.1' de görüldüğü gibi verilerimiz diğer bir algoritma olan rastgele orman kullanılarak test edilmiştir. Algoritma 0 ve 1 değerleri için verileri 0.68 oranında doğru olarak sınıflandırıldığı görülmüştür. Buna bağlı olarak ortalama tahmin değeri 0.68 oranında sonuç elde edilmiştir.

**Çizelge 6.2:** Destek vektör makinesi algoritması analiz sonucu

| Sınıf | Precision | Recall | F1 - Score | Örnek Sayısı |
|---|---|---|---|---|
| 0 | 0.72 | 0.78 | 0.75 | 330 |
| 1 | 0.76 | 0.69 | 0.72 | 330 |
| Ortalama/ Toplam | 0.74 | 0.74 | 0.74 | 660 |

Çizelge 6.2' de verilerimiz DVM algoritması kullanılarak sınıflandırılmış ve bunun sonucunda ortalama tahmin değeri 0.74 olarak elde edilmiştir. Algoritma 0 sınıfı için 0.72, 1 sınıfı için 0.76 oranında doğru yerleştirme yaptığı görülmektedir. Ortalama 0.74 oranında doğru tahmin edildiği görülmüştür.

**Çizelge 6.3:** Algoritmaların karşılaştırılması

| Algoritma | Tahmin Değeri |
|---|---|
| Destek Vektör Makinesi | 0.74 |
| Rastgele Orman | 0.68 |
| Lojistik Regresyon | 0.14 |

Çizelge 6.3' de görüldüğü üzere elde edilen sonuçlar doğrultusunda lojistik regresyon 0.14 oranında tahmin değerine sahip iken, rastgele orman algoritması 0.67 oranında doğru tahmin edebilmektedir. Buna karşın en yüksek tahmin oranı olan 0.74 değerine, DVM algoritması uygulandığında görülmüştür. Destek vektör makinesi algoritması en kaliteli sonucu verdiğinden bu algoritma kullanılarak çalışmamız geliştirilmiştir.

## 6.2 Veriler üzerinde değişimler

Verilerimiz öncelikle üçer aylık dönemlere ayrılarak hesaplama yapılmıştır. İlk aşamada tüm sınıflar dikkate alınarak hesaplama yapılmaya çalışılmıştır.

**Çizelge 6.4:** Tüm sınıflar verildiğindeki analiz sonucu

| Sınıf | Precision | Recall | F1 - Score | Örnek Sayısı |
|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 0 |
| 1 | 0.82 | 0.88 | 0.85 | 151 |
| Ortalama/ Toplam | 0.82 | 0.88 | 0.85 | 151 |

Çizelge 6.4' te görüldüğü üzere müşteri kaybı tahminlemesi yapılırken var olan tüm verileri algoritma öğrenmesine verildiğinde hepsini 1 sınıfına ait olarak yerleştirmektedir. Bundan dolayı verilerimizin net bir şekilde ayrılabilmesi için özniteliklerinin birbirinden keskin olarak ayrılması gerekmektedir. Sınıfları incelediğimizde gainer ve lost verilerinin bizim için yeterli olacağı görülmüştür.

**Çizelge 6.5:** Gainer ve Lost sınıfları verildiğinde analiz sonucu

| Sınıf | Precision | Recall | F1 - Score | Örnek Sayısı |
|---|---|---|---|---|
| 0 | 0.72 | 0.78 | 0.75 | 330 |
| 1 | 0.76 | 0.69 | 0.73 | 330 |
| Ortalama/ Toplam | 0.74 | 0.74 | 0.74 | 660 |

Uygulamamızın amacı söz konusu şirket ile çalışmayı bırakacak firmaları tahmin etmektir. Bu durumda şirket ile çalışmayı bırakmış firmaların davranışlarının izlenmesi en olası çözümdür. Buna bağlı olarak en zıt yönündeki kazanılmış müşterilerin davranışları incelendiğinde karşılaştırma yapılarak uygun sonuçlar elde edilmiştir. Sonuç olarak Çizelge 6.5' de görüldüğü üzere eğer sadece kazanılmış ve kaybedilmiş müşteri davranışları dikkate alındığında veriler doğru bir şekilde ayrılabilmiştir.

## 6.3 C parametresi değişiminin etkileri

Destek vektör makineleri için kullanılan C parametresinin alacağı değerler sınıflandırma doğruluğuna doğrudan etki eder. Alacağı değerler normalden çok yüksek seçilirse ya da çok düşük seçilirse sınıflandırma doğru yapılamaz. C parametresi için 0.1,0.3, 0.5, 0.7, 1, 2.0, 3.0 ve 4.0 değerleri verilmiş ve bu değerlere göre ortaya çıkan sonuçlar Çizelge 6.2' de gösterilmiştir. C parametresine farklı değerler verilirken diğer parametreler sabit tutulmuştur.

**Çizelge 6.6:** C parametresi değişimi ile elde edilen tahminleme sonuçları

| C | Gamma | Kernel | Tahmin |
|---|---|---|---|
| 0,1 | 0,3 | rbf | 0,74 |
| 0,3 | 0,3 | rbf | 0,74 |
| 0,5 | 0,3 | rbf | 0,74 |
| 0,7 | 0,3 | rbf | 0.73 |
| 1,0 | 0,3 | rbf | 0.73 |

**Çizelge 6.6:** (devam) C parametresi değişimi ile elde edilen tahminleme sonuçları

| | | | |
|---|---|---|---|
| 2,0 | 0,3 | rbf | 0.72 |
| 3,0 | 0,3 | rbf | 0.72 |
| 4,0 | 0,3 | rbf | 0.72 |

Çizelge 6.6' da görüldüğü gibi, en iyi sonuç C parametresi 0.1, 0.3 ve 0.5 değerlerine sahip iken alınmıştır. Ayrıca 0.3 ve 0.5 değerlerinde iken tahmin değerinin zaman zaman 0.73 değerine düştüğü görülmüştür. Bu nedenle 0.1 değeri algoritmamızda tercih edilmiştir.

## 6.4 Kernel parametresi değişiminin etkileri

Kernel parametresi çekirdek türünü ifade etmektedir. Çekirdek türü linear, poly, rbf, sigmoid, precomputed değerlerini alabilir. Çekirdek türü değişimine bağlı olarak tahmin sonuçları Çizelge 5.3'te gösterilmiştir.

**Çizelge 6.7:** Kernel parametresi değişimi ile elde edilen tahminleme sonuçları

| Kernel | C | Gamma | Tahmin |
|---|---|---|---|
| Linear | 0,1 | 0,3 | 0,55 |
| Sigmoid | 0,1 | 0,3 | 0,63 |
| Poly | 0,1 | 0,3 | 0,69 |
| Rbf | 0,1 | 0,3 | 0.74 |

Çizelge 6.7'de görüldüğü üzere linear çekirdek türü denendiğinde en düşük sonuç 0.55 değerine ulaşılmıştır. Algoritmamız sigmoid çekirdek türünde 0.63 tahmin değerine sahip iken poly çekirdek türünde 0,69 tahmin değerini verdiği görülmüştür. Rbf çekirdek türünde ise maksimum 0.74 tahmin değerine ulaşılmış ve bu çekirdek türünün algoritmamızda kullanılmasına karar verilmiştir.

## 6.5 Gamma parametresi değişiminin etkileri

Gamma parametresi rbf, poly and sigmoid çekirdek türleri için çekirdek katsayını belirtir. Gamma parametresi için 0.3, 0.5, 0.7, 1, 2.0, 3.0, 4.0 ve 5.0 değerleri verilmiş ve bu değerlere göre ortaya çıkan sonuçlar Çizelge 6.4'te

gösterilmiştir. Gamma parametresine farklı değerler verilirken diğer parametreler sabit tutulmuştur.

**Çizelge 6.8:** Gamma parametresi değişimi ile elde edilen tahminleme sonuçları

| Gamma | C | Kernel | Tahmin |
|---|---|---|---|
| 0,3 | 0,1 | rbf | 0,74 |
| 0,5 | 0,1 | rbf | 0,74 |
| 0,7 | 0,1 | rbf | 0,73 |
| 1,0 | 0,1 | rbf | 0.73 |
| 2,0 | 0,1 | rbf | 0.72 |
| 3,0 | 0,1 | rbf | 0.72 |
| 4,0 | 0,1 | rbf | 0.67 |
| 5,0 | 0,1 | rbf | 0.66 |

Çizelge 6.8'de görüldüğü üzere çekirdek katsayısını ne kadar arttırırsak tahmin değeri o kadar düşmektedir. Parametre 0.3 ve 0.5 değerlerine sahip iken uygun sonuca ulaşılabilmektedir. Fakat çekirdek katsayısı ne kadar artar ise algoritma öğrenme süresi o kadar artacağından 0.3 değeri algoritmamız için doğru değer olarak seçilmiştir.

# 7 SONUÇ

Veri madenciliğinin en yaygın kullanıldığı alanlardan biri de müşteri kaybı analizi yöntemidir. Müşteri kaybı analizi hemen hemen her sektörde karşımıza çıkmaktadır. Farklı öznitelikler ile karşımıza çıkan veriler, veri madenciliği sayesinde ayrıştırılabilmektedir. Kullanılan çeşitli algoritmalar ile anlamlı hale dönüştürülen veri yığınları kullanılarak başka firmaya geçme eğilimi gösteren müşterileri tespit edilebilmektedir.

Yapılan bu çalışmada Dünya' da ve Türkiye' de faaliyet gösteren bir lojistik firmasından alınan veriler, veri madenciliği teknikleri kullanılarak sınıflandırılmış ve destek vektör makinesi algoritmasından yararlanılmıştır. Bu lojistik firmasını önümüzdeki 3 ay içerisinde terk etme ihtimali yüksek olan müşterilerinin tespit edilmesi sağlanmıştır.

Firmadan alınan veriler, gönderi bilgilerine bağlı olarak sınıflandırılmıştır. Seçilen her müşterinin 2015 ve 2016 yıllarına ait gönderi kaydının olmasına dikkat edilmiştir. Firmanın yıl bazlı kayıp analizi yaparken kullandığı mevcut sınıflandırma yöntemi dikkate alınmıştır. Buna bağlı olarak, müşteri kaybı analizini kısa sürede tespit edebilmek amacıyla yeni veriler oluşturulurken 3'er aylık dönemler halinde değerlendirilerek sınıflandırma yapılmıştır.

Verilerimiz toplamda 5 sınıftan oluşmaktadır ve her bir sınıfa sayısal bir değer atanmıştır. Kazanılan ve kaybedilen müşteri davranışı için çıktı değerleri 2 farklı sayısal değer ile ifade edilmektedir. Fakat giriş değerlerinde tüm sınıflar ve bu sınıflara ait değerler dikkate alınmıştır. Sınıf tahmini için birçok yöntem denenmiş olmakla beraber en iyi sonuç makine öğrenmesi yöntemlerinden destek vektör makinesi algoritması ile elde edilmiştir.

Lojistik sektöründe müşteri kaybı analizi üzerinde yapılan çalışmalar oldukça azdır. Çalışmamızda yapılan testler sonucunda %74 oranında başarı sağlanmıştır. Bu da çalışmamızın başarılı bir şekilde sonuçlandığını göstermektedir. Firmadan ayrılma eğilimi gösteren müşterilerin önümüzdeki 3 ay içerisinde yüksek oranda tahmin edilmesi yapılacak iyileştirme

çalışmalarında firma için katkı sağlayacaktır. Kısa bir süre içerisinde kayıpların tespit edilebilmesi firmanın mevcut müşterilerini sistemde tutmasını sağlayacaktır.

Çalışmada kullanılan veriler 2000 adet müşteriyi temsil etmektedir. Uygulama daha fazla müşteri ve daha fazla nitelik ile daha da geliştirilebilir. Ayrıca yıllık müşteri kaybı analizi çeyrek dönemde yüksek oranda tespit edilebilir iken, ileride yapılacak çalışmalarda aylık bazlı tespitler de yapılabilir. Çalışmanın, bu konuda araştırma yapmak isteyen bilim insanlarına yardımcı olması beklenmektedir.

## KAYNAKLAR


**Akbulut S.** (2006), *Veri Madenciliği Teknikleri İle Bir Kozmetik Markanın Ayrılan Müşteri Analizi Ve Müşteri Segmentasyonu*, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara.

**Akpınar H.** (2000), "Veritabanlarında bilgi keşfi ve veri madenciliği", *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 29: 1-22.

**Alizadeh M.** (2011), *Yapay Sinir Ağları İle Fiyat Tahmin Analizi*, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

**Archer K.J**. (2008). "Emprical characterization of random forest variable ımportance measure", 52(4),2249-2260

**Arifoğlu E.** (2011)**,** *Churn Management By Using Fuzzy C-Means*, Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

**Asilkan Ö.** (2008)**,** *Veri Madenciliği Kullanılarak İkinci El Otomobil Pazarında Fiyat Tahmini*, Akdeniz Üniversitesi Sosyal Bilimler Enstitüsü, Doktora Tezi, Antalya.

**Bilgen O.** (2009)**,** *Churn Management Of Electronıc Bankıng Customers*, Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

**Bilgin, S**. (2008), *Kalp Hızı Değişkenliğinin Dalgacık Dönüşümü Ve Yapay Sinir Ağları Kullanarak Analizi*, Doktora Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya.

**Bolat, B., Küçük, Ü., Yıldırım, T.** (2004), Aktif Öğrenen PNN İle Konuşma/Müzik Sınıflandırma, Akıllı Sistemlerde Yenilikler Ve Uygulamalar Sempozyumu, 187-189.

**Bransten L.** (1999)**,** "Technology – power tools – looking for patterns: data mining enables companies to better manage the ream of statistics they collect; the goal: spot the unexpected", *Wall Street Journal*, 27 (12): 16- 20.

**Burges, C. J. C.** (1998), "A tutorial on support vector machines for pattern recognition, data mining and knowledge discovery", *Kluwer Academic Publishers*, 2 (2), 121-167.

**Chien, C. F., Chen, L. F.** (2008), "Data mining to ımprove personnel selection and enhance human capital: a case study in high-technology ındustry," *Expert Systems with Applications*, vol. 34, p. 280-290.

**Cortes, C., Vapnik, V.** (1995), Support-Vector Network, Machine Learning, 20(3): 273–297.

**Çimenli S**. (2015), *Churn Analysis And Prediction With Decısion Tree And Artificial Neural Network*, Kadir Has Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

**Davis B.** (1999)**,** "Data mining transformed", *Information Week*, 751: 86.

**DuMouchel W.** (1999)**,** "Bayesian data mining in large frequency tables, with an application to the FDA spontaneous", *American Statistician*, 53 (3): 177.

**Ercan P.** (2015)**,** *Detection of Churners in Internet Games Using Crm Approach: A Case Study on Pishti Plus*, Orta Doğu Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara.

**Etika E.** (2009), *Yapay Sinir Ağı Temelli Model Esaslı Kontrol Algoritmasının Bir Polimer Reaktörüne Uygulanması*, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara.

**Gök M**. (2014), *İnternet Servis Sağlayıcısı İçin İptal Analizi Modeli*, TOBB Ekonomi ve Teknoloji Üniversitesi Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, Ankara.

**Han, J. and Kamber , M.** (2001), "Data mining concepts and techniques", *Academic Press*, New York.

**Hand D.J.** (1998)**,** "Data mining: statistics and more ?", *The American Statistician*, 52:112-118.

**Haykin, S.S.** (2005). "Neural networks: a comprehensive foundation", *USA: Pearson Prentice Hall*.

**Holland, J. H.** (1975)**,** *Adaptation in Natural and Artificial Systems*,University of Michigan Press, Ann Arbor.

**Hsu, C.W., Chang, C.C., Lin, C.J.** (2010), A Practical Guide to Support Vector Classification, Erişim: 10 Mayıs 2017, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf,

**Jacobs P.** (1999)**,** "Data mining: what general managers need to know", *Harvard Management Update*, 4 (10): 8.

**Kalabalık G.** (2016)**,** *A Comparison Of The Performance Of Ensemble Classification Methods In Telecom Customer Churn Analysis*, Yaşar Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İzmir.

**Karaağaç Ş.S**. (2015), *Churn Analysis And Churn Prediction In A Private Bank*, Marmara Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

**Karahan M.** (2011)**,** *İstatistiksel Tahmin Yöntemleri: Yapay Sinir Agları Metodu İle Ürün Talep Tahmini Uygulaması*, Selçuk Üniversitesi Sosyal Bilimler Enstitüsü, Doktora Tezi, Konya.

**Karataş E.K.** (2011)**,** *Yapay Sinir Ağları İle Yazılım Projesi Maliyet Tahmini*, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

**Kavzaoğlu, T., Çölkesen, İ.** (2010), "Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi", *Harita Dergisi, Sayı 144,*73-82.

**Kılıç G.** (2015)**,** *Yapay Sinir Ağları İle Yemekhane Günlük Talep Tahmini*, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Denizli.

**Kişioğlu P.** (2009)**,** *Telekomünikasyon Sektöründe İptal Analizi*, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

**Kitler R., Wang W.** (1998)**,** "The emerging role of data mining", *Solid State Technology*, 42 (11): 45.

**Koçtürk Y.** (2010)**,** Veri Madenciliğinde Bağlılık, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

**Kostek M.** (2014)**,** *Makine Öğrenme Yöntemlerinin Araştırılması ve Uygulanması*, Bilecik Şeyh Edebali Üniversitesi Mühendislik Fakültesi, Bilecik.

**Larose, D. T.** (2005), "Discovering knowledge in data: an ıntroduction to data mining", *John and Wiley Sons Incorporated*, USA.

**Mahanty, R.N. And Gupta, P.B.D.** (2004), "Application of RBF neural network to fault classification and location in transmission lines.", *IEEE Proceedings of Gener. Transm.Distrib.*, 151, 201-212.

**Namane, A., Guessoum, A., Soubarı, E. H. ve Meyrueıs, P.** (2014), "CSM neural network for degraded printed character optical recognition", *Journal of Visual Communication and Image Representation*, 25, 1171-1186.

**Osowski, S., Siwekand, K., and Markiewicz, T**. (2004), "MLP and SVM Networks", *Proceedings of the 6th Nordic Signal Processing Symposium*. (pp.37-40)

**Osuna, E.E., Freund, R., Girosi, F.** (1997),*Support Vector Machines: Training and Applications. (Teknik Rapor)*,Massachusetts Institute of Technology and Artificial Intelligence Laboratory, Massachusetts.

**Özmen M.** (2006)**,** *Churn Modeling In Telecommunications Sector*, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

**Perendeci, A**. (2004), *Şeker Fabrikası Anaerobik Atık Su Arıtma Tesisinin Yatışkın Olmayan Durumda Modellenmesi Çalışmaları*, Doktora Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

**Quinlan, J. R.** (1993), "C4.5: Programs for Machine Learning", *Morgan Kaufmann Publishers*, USA.

**Sarı M.** (2016)**,** *Yapay Sinir Ağları ve Bir Otomotiv Firmasında Satış Talep Tahmini Uygulaması*, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Sakarya.

**Shalizi, C**. *Logistic Regression Ders Notu*, Department of Statistics, Carnegie Mellon University.

**Shearer C.** (2000)**,** "The Crisp-DM model: The new blueprint for data mining", *Journal of Data Warehousing*, 5 (4): 13-23.

**Şen, Z.** (2004),"Yapay sinir ağları ilkeleri", *İstanbul: Su Vakfı Yayınları*.

**Telcioglu M.B.** (2002), *Montaj Hattı Dengeleme Problemlerinin Genetik Algoritma Teknigi Kullanılarak Çözülmesi ve Bilgisayar Programı Uygulaması Pegasus Yazılımı,* Yüksek Lisans Tezi, Erciyes Üniversitesi.

**Tosun T.** (2006)**,** *Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi*, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

**Tufan E**. (2012), *Telekomünikasyon Sektöründe Müşterilerin Ürün Grupları Ve Tarifeler Arası Geçiş Analizi*, TOBB Ekonomi ve Teknoloji Üniversitesi Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, Ankara.

**Vapnik, V.N.** (1995),*The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

**Yabaş U.** (2014)**,** Customer Churn Prediction For Telecommunications Industry, İzmir Ekonomi Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İzmir.

**Zupan, J**. (2013), "Basics of artificial neural networks", *Data Handling in Science and Technology*, 23, 199-229.

[1]**Breiman L.,Cutler A.,** Random forest, Erişim: 5 Kasım 2013, http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[2]Regression Vikipedia. Erişim: 28 Kasım, 2015, https://en.wikipedia.org/wiki/Logistic_regression

**EKLER**

**Ek A:** Makine Öğrenmesi Yöntemleri ile Müşteri Kaybı Analizi Tezimde Kullanılan Veri Kaynağı Kısmı Cd'de teslim edilecektir

**Ek B:** Makine Öğrenmesi Yöntemleri ile Müşteri Kaybı Analizi Tezimde Algoritmanın Uygulandığı Kod Kısmı Cd'de teslim edilecektir

## ÖZGEÇMİŞ

**Buket ÖNAL**
0535 351 02 16
bukettonal@gmail.com

**KARİYER HEDEFİM**

Bilgisayar programlama alanında kazandığım uzmanlık ve deneyimi kalitelilik ve verimlilikle yazılım projelerinde uygulayabilmek, süreci yönetebilmek ve bilgimi daha da geliştirebilmek.

**TEKNİK**

**Diller**

- C#, HTML, CSS ,JavaScript ,Jquery, , Entity Framework, Linq

**Teknolojiler**

- ASP.NET, Web Forms , WCF, Xamarin , Telerik, Devexpress, MVC , Windows Forms

**Yazılım**

- Veritabanı : Microsoft Sql Server
- Kod Kontrol Sistemleri : TFS
- Geliştirme Araçları : Microsoft Visual Studio

**DENEYİM**

**Uygulama Geliştirme Uzmanı**
01.11.2017
*LC WAIKIKI*

- E Ticaret, ürün, stok ve müşteri uygulamalarının oluşturulması ve geliştirilmesi.

**Yazılım Uzmanı**
17.10.2016 – 31.10.2017
*UPS*

- Şirket içi, Yurtdışı ve Müşteri istekleri doğrultusunda MVC,WCF,Web Forms teknolojilerini kullanarak programların geliştirilmesi.

**Yazılım Geliştirici**
10.06.2014 – 15.10.2016
TREEM A.Ş

- Aizen, Web Tabanlı MES Yazılımı, Tübitak Destekli Yazılım Projesi, Üretim sistemlerinde veri toplama ve analiz için gerekli tanımların yapılması, alınan verilerin analizi ve grafiğe dökülmesi üzerine C# dili ile MVC,Telerik teknolojisi ile geliştirilmiş web tabanlı yazılım projesinin geliştirilmesi.
- CCI, Online Monitör, Coca Cola İçecek A.Ş., Web Tabanlı Online Hat İzleme ve Raporlama Projesi, Coca-Cola üretim tesislerinde toplanan verileri anlık olarak izleme ve raporlama projesidir. C# dili kullanılarak MVC ,Devexpress teknojisi ile birlikte geliştirilmiştir.

- Hes Kablo, Web Tabanlı Bakım Yönetimi Projesi, Hes Kablo üretim tesislerinde bakım yönetimi için C# dili ve MVC kullanılarak geliştirilmiş yazılımdır.
- CCI, (Xamarin)Android Tabanlı Fabrika-Hat Yönetim Projesi, Coca-Cola üretim tesislerinde tabletler aracılığıyla hat yönetimini sağlamak ve günlük,aylık kontrollerin yapılması, ISG ve Görev atama işlemlerinin yapılabildiği C# dili kullanılarak Android tabanlı Xamarin platformunda geliştirilmiştir. Ayrıca projede gerekli tanımların yapılabileceği MVC ile geliştirilmiş bir web projesi de yer almaktadır.
- CCH, Online Monitör, Coca Cola Hungary., Web Tabanlı Online Hat İzleme ve Raporlama Projesi, Coca-Cola Hungary üretim tesislerinde toplanan verilerin anlık olarak izlenebilmesi için geliştirilmiştir. Coca-Cola İçecek için geliştirilen projeden farklıdır. Shift tanımlama, SAP entegrasyonu gibi işlemler bu sistem üzerinden gerçekleşmektedir. C# dili kullanılarak MVC, Devexpress ile geliştirilmiştir.
- Mert Yazılım,Sensorium, Isı Takip Sistemi, Sıcaklık ve nem değerlerinin anlık olarak izlendiği bir sistemdir.İstenen sıcaklık ve nem değerlerinin dışına çıkılması durumunda ilgili kişileri anında, email ve sms ile uyarmaktadır. . C# dili kullanılarak MVC ile geliştirilmiştir.

## EĞİTİM

İstanbul Aydın Üniversitesi 2015-2017
*İstanbul, Türkiye*

- Bilgisayar Mühendisliği Tezli Yüksek Lisans Eğitimi

Yıldız Teknik Üniversitesi 2010 – 2014

- Bilgisayar ve Öğretim Teknolojileri Eğitimi Öğretmenliği Lisans Eğitimi

İstanbul,Türkiye

## REFERANSLARIM

Ümit Değirmenci (AR-GE Müdürü) (0533 638 11 05)
***Treem A.Ş- Trex DCAS***
Hayrullah Çetinkaya (Yazılım Departmanı Müdürü) (0533 516 01 40)
***UPS***