**T.C.**

**İstanbul Üniversitesi**

**Sosyal Bilimler Enstitüsü**

**İşletme Anabilim Dalı**

**Sayısal Yöntemler Bilim Dalı**

**Doktora Tezi**

# Web Kullanım Madenciliği ile Satın Alma Davranışının Modellenmesi ve Bir E-Ticaret Uygulaması

**Serra ÇELİK**

**2502090313**

**Tez Danışmanı**

**Prof. Dr. M. Erdal BALABAN**

**İstanbul 2015**

**T.C.**

**İstanbul Üniversitesi**

**Sosyal Bilimler Enstitüsü**

**İşletme Anabilim Dalı**

**Sayısal Yöntemler Bilim Dalı**

**Doktora Tezi**

# Web Kullanım Madenciliği ile Satın Alma Davranışının Modellenmesi ve Bir E-Ticaret Uygulaması

**Serra ÇELİK**

**2502090313**

**Tez Danışmanı**

**Prof. Dr. M. Erdal BALABAN**

**İstanbul 2015**

Bu Doktora Tezi, İstanbul Üniversitesi Bilimsel Araştırmalar Proje Birimi (BAP) tarafından desteklenmiştir. Proje No: 44823

T.C.
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ

# DOKTORA
# TEZ ONAYI

**ÖĞRENCİNİN;**

| | | | | | |
|---|---|---|---|---|---|
| **Adı ve Soyadı** | : | SERRA ÇELİK | **Numarası** | : | 2502090313 |
| **Anabilim Dalı / Anasanat Dalı / Programı** | : | SAYISAL YÖNTEMLER | **Danışmanı** | : | PROF.DR M.ERDAL BALABAN |
| **Tez Savunma Tarihi** | : | 13.11.2015 | **Saati** | : | 14:00 |
| **Tez Başlığı** | : | **WEB KULLANIM MADENCİLİĞİ İLE SATIN ALMA DAVRANIŞININ MODELLENMESİ VE BİR E-TİCARET UYGULAMASI** | | | |

**TEZ SAVUNMA SINAVI**, İÜ Lisansüstü Eğitim-Öğretim Yönetmeliği'nin **50. Maddesi** uyarınca yapılmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin **KABULÜNE** OYBİRLİĞİ / ~~OYÇOKLUĞUYLA~~ karar verilmiştir.

| JÜRİ ÜYESİ | İMZA | KANAATİ (KABUL / RED / DÜZELTME) |
|---|---|---|
| 1-PROF.DR M.ERDAL BALABAN | | Kabul |
| 2- PROF.DR SEVİNÇ GÜLSEÇEN | | Kabul |
| 3- PROF.DR ZUHAL TANRIKULU | | KABUL |
| 4- PROF.DR HALDUN AKPINAR | | KABUL |
| 5- DOÇ.DR SEDA TOLUN | | KABUL |

| YEDEK JÜRİ ÜYESİ | İMZA | KANAATİ (KABUL / RED / DÜZELTME) |
|---|---|---|
| 1- PROF.DR RAUF NİŞEL | | |
| 2-DOÇ.DR ÇİĞDEM ARICIGİL | | |

# ÖZ

## WEB KULLANIM MADENCİLİĞİ İLE SATIN ALMA DAVRANIŞININ MODELLENMESİ VE BİR E-TİCARET UYGULAMASI

**SERRA ÇELİK**

Günümüzde veri artışı inanılmaz boyutlara ulaşmıştır. Gelişen teknolojiyle birçok farklı sektörde daha kolay veri elde edilebilmektedir. Bu noktada veri madenciliği bu veri yığınlarından anlamlı bilgiye dönüşüm sürecini hızlandırmıştır. Veri madenciliği, ilk başta veri tabanlarından bilgi çıkarımı olarak ortaya çıksa da günümüzde geliştirilen yeni yöntemler ve teknolojilerin desteği ile tahmin gücünden daha fazla yararlanılmaktadır. Tez çalışmasında veri madenciliği sınıflandırma yöntemlerinden destek vektör makineleri, web kullanım madenciliği verisi olan web günlük dosyaları (web log files) üzerine uygulanmıştır. Kullanılan veri seti bir e-ticaret sitesinin 812 güne ait web günlük dosyalarıdır. Web günlük dosyaları yapılandırılmamış veri içermektedir ve bu tip verinin analizi yapılandırılmış veriye göre daha zordur. Bu nedenle analiz öncesinde verinin temizlenmesi gerekmiş ve bu süreç çalışmada uzun bir süre almıştır. Çalışmada satın alma davranışının eğilimini belirlemek hedeflenmiştir. Destek vektör makineleriyle sınıflandırma yapılmış sonuçlar lojistik regresyonla elde edilen sonuçlarla karşılaştırılmıştır. Destek vektör makineleri ile bir e-ticaret sitesi uygulamasında daha doğru sınıflandırma yapılabildiği görülmüştür.

**Anahtar Kelimeler:** Web Kullanım Madenciliği, Destek Vektör Makineleri, E-Ticaret, Satın Alım Davranışı, İnternet

# ABSTRACT

## MODELING PURCHASE BEHAVIOR WITH WEB USAGE MINING AND AN E-COMMERCE APPLICATION

**SERRA ÇELİK**

Today, size of data has reached amazing amounts. Recent advances in technology collecting data in many different sectors is getting easier. At this point, data mining has accelerated the process of transforming data to information. In the beginning, data mining has been known as information extraction from databases, but recently it is more useful for prediction by the help of new methods and technologies developed. In this study web usage mining will be performed with classification methods of data mining using web log files. The data used is an e-commerce web site's log files of 812 days. Web log files contain unstructured data and it is very difficult to analyze it in conventional ways. Before analyzing data it has to be cleaned and this process takes long time. The aim of this study is finding the way of purchase behavior. First, analysis are made by support vector machines, then results are compared with the results obtained by logistic regression. For implementation to an e-commerce web site, it can be stated that support vector machines can classify more accurately.

**Keywords:** Web Usage Mining, Support Vector Machines, E-Commerce, Purchase Behavior, Internet

# ÖNSÖZ

Günümüz dünyasında internet yaşamın önemli bir parçası haline gelmiştir. Şirketler ürünlerini sadece fiziksel mekânlarda değil sanal ortamda da pazarlama yoluna gitmeye başlamıştır. İnternet ile fiziksel mağazaya erişilebilirlik sınırları kalkmış, müşteriler istedikleri zaman internet aracılığıyla istedikleri ürünü satın alabilir hale gelmiştir. Elektronik ticaret ile şirketler de pazarlama stratejilerini geliştirme ihtiyacı duymuşlardır. Bu çalışma, şirket yöneticilerinin e-ticaret faaliyetlerini gerçekleştirirken, site ziyaretçilerinin müşteriye dönüşme sürecini makine öğrenmesi sınıflandırma yöntemlerinden destek vektör makineleri ile araştıran bir analiz içermektedir.

Çalışmanın ilk bölümünde web kullanım madenciliğiyle ilgili temel bilgiler verilmiştir. Analizde kullanılan web verisinin özellikleri açıklanmıştır. Bu tip verilerin temizlenme aşaması analizin büyük bir kısmını oluşturduğundan veri önişlem bilgileri de bu bölümde yer almıştır.

İkinci bölümde danışmanlı öğrenme yöntemlerinden destek vektör makinelerinin teorik altyapısına yönelik tanımlar verilmiştir. Ayrıca sınıflandırma yöntemlerinin performansını ölçmeyi sağlayan değerlendirme yöntemleri ele alınmıştır.

Son bölümü oluşturan uygulama bölümünde ise web günlük dosyalarının temizlenme, önişleme ve analiz aşamaları anlatılmıştır. E-ticaret sitelerinde ziyaretçiler çok fazla sayıda sayfa dolaşmakta ancak satın alınan ürün buna kıyasla çok düşük olmaktadır. Bu da düşük cevap oranına yani veri dengesizliğine neden olmaktadır. Son yıllarda farklı alanlarda uygulamalarına rastladığımız destek vektör makinelerinin ise bu alanda fazla kullanılmadığı görülmüş olup çalışma, web günlük verisine de uygulanabilirliğini göstermek amacındadır.

Çalışmanın her aşamasında göstermiş olduğu sabır, anlayış ve yönlendirmeleri için danışman hocam Prof. Dr. M. Erdal Balaban'a teşekkürlerimi sunarım.

İzleme komitesinde yer alan/almış hocalarım Prof. Dr. Sevinç Gülseçen, Doç. Dr. Çiğdem Arıcıgil Çilan ve Doç. Dr. Seda Tolun'a her zaman yanımda olup akademik gelişimime yaptıkları katkılardan dolayı teşekkür ederim.

Değerli meslektaşlarım Öğr. Grv. Dr. Murat Gezer ve Yrd. Doç. Dr. Burak Şişman'a ve analiz için gerekli ekipmanları sağlayan İ.Ü. Bilimsel Araştırma Projeler Birimi'ne teşekkür ederim. Sevgili Yrd. Doç. Dr. Mutlu Gürsoy'a çalışma süresince vermiş olduğu moral ve destek için teşekkürü borç bilirim. Sevgili meslektaşım ve dostum Arş. Grv. Emre Akadal'a desteğini hiç esirgemediği için sonsuz teşekkürlerimi sunarım.

Her zaman yanımda olan aileme teşekkür ederim.

Annem Emine Çelik'in anısına.

Serra Çelik

**İçindekiler**

# TABLOLAR LİSTESİ

# ŞEKİLLER LİSTESİ

# KISALTMALAR LİSTESİ

**WWW:** World Wide Web

**DVM:** Destek Vektör Makineleri

**MİY:** Müşteri İlişkileri Yönetimi

**DARPA:** Defense Advanced Research Projects Agency

**CERN:** Conseil Europeen pour la Recherche Nuclaire

**HTML:** HyperText Markup Language-Hipermetin İşaretleme Dili

**URL:** Universal Resource Locator – Evrensel Kaynak Konumlayıcı

**SGML**: Standart Generalized Markup Language

**DTD:** Document Type Definiton

**ISP:** Internet Service Provider

**ADSL:** Asymmetric Digital Subscriber Line

**LR:** Logistik Regresyon

# GİRİŞ

Bilgisayar ortamındaki dosyalarda ve veri tabanlarında tutulan veri miktarları her geçen gün hızla artmaktadır. Öte yandan kullanıcılar ise, tutulan bu verilerden anlamlı bilgiler çıkarılmasını beklemektedirler. Örneğin bir pazarlama yöneticisi basit bir ürün veya satış listesiyle yetinmeyerek, müşterilerin geçmiş harcamalarıyla ilgili veriden yola çıkarak gelecek satışların tahminini öğrenmek isteyecektir. Veri madenciliğine de bu ihtiyaçları karşılamak için başvurulmaktadır. Veri madenciliği genel olarak bir veri tabanındaki gizli bilgiyi bulmak/ortaya çıkarmak olarak tanımlanmaktadır. Alternatif olarak keşfedici veri analizi (exploratory data analysis) veya veri temelli keşif (data driven discovery) olarak da isimlendirilmektedir (Dunham, 2003).

Çok kısa zamanda büyük miktarlara ulaşan veri, her geçen gün hızla artmaya devam etmektedir. Buna karşın, toplanan verinin çoğu kez sadece küçük bir kısmı kullanılmaktadır. Veri miktarı büyüdükçe ve karmaşıklık arttıkça, veriyi iyi çözümleme tekniklerine olan gereksinim de artmaktadır. Başarıyla gerçekleştirilen bilgi keşfi süreci sonucunda elde edilen faydalı bilgi, organizasyonların karar verme sürecini geliştirmek amacıyla kullanılabilecektir.

Veri madenciliği; telekomünikasyon, bankacılık, sigortacılık ve sağlık başta olmak üzere birçok sektörde etkin şekilde kullanılmaktadır. Telekomünikasyon alanında özellikle müşteri ilişkileri yönetimi kapsamındaki çalışmalarda -müşterilerin satın alım örüntülerinin belirlenmesi, müşterilerin demografik özellikleri arasındaki bağıntıların bulunması, mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması-; bankacılık alanında, farklı finansal göstergeler arasında gizli kalmış ilişkilerin bulunması, kredi kartı dolandırıcılıklarının tespiti, kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi, kredi taleplerinin değerlendirilmesinde; sigortacılık alanında, yeni poliçe talep edecek müşterilerin tahmin edilmesi, sigorta dolandırıcılıklarının tespiti, riskli müşteri örüntülerinin belirlenmesi ve satış tahmininde; sağlık alanında ise hastane kuruluş yerinin belirlenmesi, hasta odaklı

sağlık hizmeti sunumu ile kalitenin geliştirilmesi, hastaneler için finansal erken uyarı sisteminin geliştirilmesi, kronik hastalıklar için erken uyarı sinyallerinin tespiti, tıbbi teşhis, laboratuvar testleri için hata ve suistimal tespiti, başlıca veri madenciliği uygulama örnekleridir. Bunun dışında, hisse senedi fiyat tahmini, genel piyasa analizleri, alım-satım stratejilerinin optimizasyonu, endüstri (kalite kontrol analizleri, lojistik, üretim süreçlerinin optimizasyonu), bilim ve mühendislik (ampirik veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesi) problemlerinde de uygulamalar mevcuttur.

Bu çalışmada sınıflandırma yöntemlerinin tahmin güçlerinden yararlanılarak veri madenciliğinin World Wide Web verileri üzerinde bir uygulaması olan web madenciliği (web mining) çalışması yapılmıştır. Web madenciliği, bilgi çıkarımı ya da öğrenme amaçlı web tabanlı veriye veri madenciliği yöntemlerinin uygulanmasıdır. Verilere ulaşım sürecinde kullanıcıyı doğru ve hızlı olarak yönlendirmek hem web sitesinin etkili kullanımı açısından hem de elektronik ticaret sitelerinin amaçlarına ulaşmaları açısından önemlidir. Web kullanım madenciliği; web yapı madenciliği ve web içerik madenciliğinden farklı olarak, web üzerinde doğrudan erişilebilen veriyi kullanmak yerine, kullanıcıların web üzerinde dolaşırken hareketlerinden oluşturdukları veriden bilgi üretmektedir. Bu konudaki çalışmalar genel web sitesi güncelleme sistemleri, sistem iyileştirme ve kişiselleştirme, sahtekârlık ve izinsiz giriş tespiti, kullanıcının bir sonraki faaliyetini öngörme, cep belleğe alma başlıkları altında toplanabilmektedir (Gündüz ve Adalı, 2004).

Günümüzde web, insan hayatının önemli bir parçası haline gelmiştir. Web; insanların kolaylıkla iletişim kurmasına, dünyanın neresinde olursa olsun kolaylıkla düşünce ve görüşlerini paylaşmasına izin verebilmektedir. İyi tasarlanmış web sitelerinden işletmeler de fayda sağlayabileceklerdir. Neredeyse her şey çevrimiçi mağazalardan satın alınabilmektedir. İşletmelerin ürün ve hizmetlerini çevrimiçi satması işletme başarılarında önemli bir rol oynamaktadır. Çoğu firma için bir perakende web sitesi firma ve müşterileri arasında etkin bir iletişim hattıdır. Bütün ürün ve hizmetleri web sitelerinde yer almasa dahi işletmeler, müşterilerinin ihtiyaçlarını görüp rakipleri üzerine avantaj sağlayabileceklerdir. Başarılı bir web

sitesi işletmenin ihtiyaçlarını karşılayacak satış ve pazarlama aracı olmada önemlidir. Bir web sitesi, kullanıcının ihtiyaçlarını karşılıyorsa, kullanıcının site içerisinde dolaşımını kolaylaştırıyorsa, kullanıcının hedef sayfasına herhangi bir arama ya da tahmin yapmadan kısa sürede ulaşmasını sağlayabiliyorsa iyi yapılandırılmıştır. Sahibi açısından bakıldığında ise bir web sitesi; işletme kârını ve kullanıcının işletmeye güvenini arttırıyorsa iyi yapılandırılmıştır. Özetle hem sahiplerinin hem de kullanıcılarının gereksinimlerini buluşturan web siteleri başarılıdır.

Web kullanıcıları site üzerinde gezdiklerinde bir yol çizmiş olurlar. Bu yolun keşfiyle çoğu işletme için pazarlama ve satış açısından önemli bir ticari değere sahip bilgi elde edilmiş olur. Web tabanlı işletmeler, kullanıcıların davranış örüntülerinden promosyon belirleme, potansiyel riskleri tespit etme ve stratejik kararlar almada yararlanabilmektedirler. Web kullanıcılarının davranışı için modelleme ve tahmin, popüler bir konu olup farklı yaklaşımlar önerilmiştir. Genelde üzerinde durulan, bir kullanıcının izleyeceği sayfa sırasının benzer bir kullanıcının önceden görüntülemiş olduğu sayfalar aracılığıyla tahmin edilmesi problemidir. Bu problemin çözümü için web sayfaları özetlenip kategorilere ayrılabilir ve bir kullanıcının izleyeceği kategori sırası tahmini olarak belirlenebilir. Ancak yine de bu, teknik ve pratik noktaların görülmesini gerektirecek karmaşık bir makine öğrenmesi problemidir.

Bu tez çalışmasında web günlük verisiyle (web log data) web kullanım madenciliği uygulaması gerçekleştirilmiştir. Ziyaret edilmeyen web sayfalarının, ziyaret edilen sayfalardan fazla olması web günlük verisinde “seyreklik” denilen probleme neden olmaktadır. Bu problem nedeniyle, web bilgi tavsiyesi, ziyaret edilecek bir sonraki web sayfasının ve web sayfasında kalış süresinin tahmini gibi görevler için web kullanım madenciliğinin uygulanması zorlaşmaktadır. Bu şekildeki veriyi basit istatistiksel yöntemler ile analiz etmek zordur. Bu durumda destek vektör makineleri gibi makine öğrenmesi tekniklerinden yararlanılabilmektedir.

# BÖLÜM 1

# 1 WEB MADENCİLİĞİ

Bu bölümde web madenciliği kavramı üzerinde durulmuştur. Web madenciliği sınıflamalarıyla, web madenciliğinde kullanılan veri tipleri hakkında bilgi verilerek, web günlük (log) analizi ayrıntılarıyla açıklanmıştır.

## 1.1 WWW (The World Wide Web) ve İnternetin Doğuşu

Web kullanımı yaygınlaşmadan önce bilgiye ulaşmak bir arkadaşa ya da uzmana sormak, okumak için bir kitap edinmek anlamına gelmekte iken, web ile her şey ev ya da işyeri rahatlığında, sadece birkaç tık uzaklığa dönüşmüştür. Web ile sadece aranılan bilgiye ulaşılmamakta, bilgi ve deneyimler de kişilerle paylaşılabilmektedir (Liu, 2007).

İşletmeler için de önemli bir kanal haline gelen web, fiziksel mekâna gitmeye gerek kalmaksızın çevrimiçi dükkânlardan satın alıma imkân sağlamaktadır. Web'in aslında bir sanal toplum olduğunu söylemek mümkündür (Sen v.d., 2006).

World Wide Web resmi olarak "büyük belge evrenine, evrensel erişimi sağlama amacındaki hiper ortamda bilgi çıkarma girişimi" olarak tanımlanmaktadır. Basitçe web, internet adı verilen dünya çapında ağ (world-wide network) aracılığıyla depolanmış bilgiye erişim için bir bilgisayarın, diğer kullanıcıların erişimine izin veren internet tabanlı bir bilgisayar ağıdır (Liu, 2007).

Web gerçekleştirimi, standart bir istemci-sunucu modelini izlemektedir. Bu modelde bir kullanıcı, verinin saklandığı sunucu denilen bir uzak makineye bağlanmak için istemciye ihtiyaç duymaktadır. Web üzerinde gezinti, Internet Explorer, Firefox, Chrome gibi tarayıcı denilen istemci programlar ile yapılmaktadır. Web tarayıcılar bilgi için uzak sunuculara istek göndermekte, ardından HTML'de yazılmış isteğe

dönen belgeleri yorumlamakta ve bunları istemci tarafında kullanıcının bilgisayar ekranına metin ve grafik olarak yansıtmaktadır (Spiliopoulou, 1999).

Web işletimi (web operation), hipermetin (hypertext) belgelerinin yapısına bağlıdır. Hipermetin, web sayfa yazarlarına belgelerini dünyanın herhangi bir yerindeki bilgisayarlardaki ilişkili belgelerle bağlamalarına izin vermektedir. Bu belgeleri görüntülemek için basitçe bağlantılar (hyperlinks) takip edilmektedir (Liu, 2007). Hipermetin düşüncesi Ted Nelson (1965) tarafından önerilmiştir. Nelson ayrıca iyi bilinen hipermetin sistemi Xanadu'yu da oluşturmuştur. Hipermetin, hipermedya denilen diğer medya öğelerine (görüntü, ses ve video dosyaları gibi) de izin vermektedir (Liu, 2007).

İnternetin temeli soğuk savaş dönemine uzanmaktadır. Pentagon Defense Advanced Research Projects Agency (DARPA) tarafından finanse edilerek, birbirine bağlı bilgisayar sistemleri ile hızlı ve güvenilir bilgi dağıtımı üzerine çalışılmıştır. 1969 yılında Kaliforniya Üniversitesi'nde ARPANET ağının ilk düğümü açılmıştır. Ağ genişlemeye başlamış ve ikinci düğüm Stanford Araştırma Enstitüsü'ne yüklenmiştir (Velasquez ve Palade, 2008).

1990'ların başında İsviçre'de Avrupa Parçacık Fiziği Laboratuvarı (Conseil Europeen pour la Recherche Nuclaire - CERN) araştırmacılarından Tim Berners-Lee hipermetin paylaşımı düşüncesini geliştirmiştir ve böylelikle web kavramı ortaya çıkmıştır (Velasquez ve Palade, 2008). Tim Berners-Lee "World Wide Web" terimini bulmuş, ilk World Wide Web sunucusu httpd'yi ve ilk istemci programı (bir tarayıcı ve editör) olan WorldWideWeb'i yazmıştır (Liu, 2007).

Bilgisayarların diğer bilgisayarlarla birleşmesi düşüncesi daha güçlü bir yaklaşım ile bilgisayar ağlarının diğer bilgisayara bağlanmasına dönüşmüştür. Bu bağlantı, ağa ismini vermiştir: Inter-Net. Başlangıcından itibaren internet, bilgisayarlarla haberleşmeyi kolaylaştırmış, bu da kişiler arasında iletişim için elektronik posta gibi yeni etkileşim araçları ortaya çıkarmıştır. İnternet ilk yıllarında, internet hizmetlerini kullanmak için gereken karışık talimatları anlayabilecek sayıda kişinin kullanımıyla kısıtlanmıştır (Velasquez ve Palade, 2008).

1989 yılında Tim Berners-Lee CERN'de "Information Management: A Proposal" adlı bir çalışmayla hiyerarşik enformasyon organizasyonunun dezavantajlarından bahsederek hipermetin tabanlı sistemin avantajlarını vurgulamıştır. Bu çalışmada Web'in basit mimarisi olan dağıtılmış hipermetin sistemi (distributed hypertext system) taslağını ortaya atmıştır (Xu, Zhang ve Li, 2011).

1990 yılında Berners-Lee ve ekibi yaptığı değişikliklerle sunucu, tarayıcı, istemci ve sunucu arasında iletişim için kullanılan HTTP (HyperText Transfer Protocol- Hipermetin Transfer Protokolü), web belgelerine izin için kullanılan HTML (HyperText Markup Language-Hipermetin İşaretleme Dili) ve URL'yi (Universal Resource Locator – Evrensel Kaynak Konumlayıcı) ortaya çıkarmışlardır (Liu, 2007).

## 1.2 Web Belgeleri

***HTML:*** HTML (HyperText Markup Language), belge içerisine gömülü işaretin (markup) nasıl yorumlanacağını tanımlayan uluslararası bir standart olan, Standart Genelleştirilmiş İşaretleme Dilinin (Standart Generalized Markup Language – SGML) bir uygulamasıdır (Baldi, Frasconi ve Smyth, 2003).

Mevcut web kaynaklarının önemli bir oranı HTML belgesidir. HTML, bir web tarayıcısı kullanıldığında tecrübe edilen metin ve görüntülerin yapısı ve tasarımı için temel etiketleme desteği sağlayan bir biçimdir (Xu v.d., 2011).

Bir HTML belgesi aşağıdaki üç parçadan oluşmaktadır (Baldi, Frasconi ve Smyth, 2003):

*Versiyon bilgisi:* Belgede hangi DTD'nin (Document Type Definition) kullanıldığını belirleyen açıklayıcı bölümdür. W3C üç adet mümkün DTD tanımlamaktadır. Bunlar; metin (string), geçiş (transitional) ve çerçevedir (frameset).

*Header:* Metadata (belgeyle alakalı veri) içeren açıklayıcı bölümdür. Bu bölüm <head> ve </head> içerisinde ilişkilendirilmiştir.

*Body:* Gerçek belge içeriğidir. Bileşen <body> ve </body> içerisinde ilişkilenmiştir.

## 1.3 Web Madenciliği Tanımı

WWW'nun hızla gelişmesiyle internet, faydalı bilgilerin bulunabileceği bir kaynak olmuştur. Bu kaynak aynı zamanda bir veri kaynağıdır. Kullanıcılar, internette sayfalarda gezinirken, ürün/hizmet satın alırken, sosyal medyada paylaşımda bulunurken izler bırakmaktadırlar. Bu izler web sunucularında depolanmakta ve karşımıza yapılandırılmamış veri olarak çıkmaktadır.

Veri madenciliği, kabul edilebilir hesaplama limitleri içerisinde veride örüntü ya da modeller bulan bilgi keşfi sürecidir. Basitçe veri madenciliği, veriden daha önceden bilinmeyen değerli ve kullanışlı bilginin çıkarımıdır (Nisbet, Elder ve Miner, 2009).

Web üzerinde üretilen ve depolanan bu veri web madenciliği ile analiz edilebilmektedir. Web madenciliği, web verisine veri madenciliği yöntemlerinin uygulanmasıdır. Daha açık bir ifadeyle web verisinden kullanıcı örüntüleri keşfetmede veri madenciliği yöntemlerinin uygulanması olarak tanımlanabilmektedir (Srivastava v.d., 2000). Web madenciliği terimi ilk kez Etzioni (1996) tarafından kullanılmıştır.

Web kullanım madenciliği ilk kez Cooley v.d. (1997) tarafından, web sunucularından kullanıcı erişim örüntülerinin otomatik keşfi olarak tanımlanmıştır.

Web günlük analizleri web site yöneticilerinin yeterli bant genişliği ve sunucu kapasitesi sağlamada bir yol olarak sunulmasıyla başlamıştır. Bu analiz alanı, geçen sürede büyük gelişmeler yakalamıştır. E-şirketler, web günlük dosyalarını ziyaretçi profilleri ve satın alım faaliyetleri hakkında bilgi edinmek amaçlı kullanmaya başlamışlardır (Agosti ve Di Nunzio, 2007).

Geçmiş web günlüklerinden ortaya çıkartılmış örüntüler ve mevcut gezinti örüntülerinin analiziyle kullanıcı davranışını tahmin etmek mümkündür. Kullanıcıların davranışı, kişiselleştirme, sistem geliştirme ve kullanıcı ilgilerine göre sistem tasarımı, bir web sitesi modifikasyon sürecidir (Chitraa ve Davamani, 2010).

Web madenciliğinin en çok kullanılan sınıflama yöntemleri, Kosala ve Blockeel (2000) tarafından önerilmiş olan web'in hangi alanına madencilik

yapılacağına ilişkin, üç ana kategoriye ayrılan sınıflamadır. Bu üç kategori; web içerik madenciliği, web yapı madenciliği ve web kullanım madenciliğidir.

### 1.3.1 Web İçerik Madenciliği (Web Content Mining)

Web içerik madenciliği web içerikleri ve belgelerinden kullanışlı bilginin keşfi olarak tanımlanmaktadır (Kosala ve Blockeel, 2000). Örneğin, web içerik madenciliğiyle web sayfaları konularına göre otomatik olarak kümelenip sınıflandırılabilmektedir. Ürün tanımları, forum paylaşımları gibi verilerin çıkarımıyla müşteri görüşleri irdelenip tüketici hassasiyeti keşfedilebilmektedir (Jiawei, Kamber, ve Pei, 2011), (Liu, 2007).

### 1.3.2 Web Yapı Madenciliği (Web Structure Mining)

Web yapı madenciliği bağlantılardan (hyperlinks), web yapısını temsil eden kullanışlı bilginin keşfi olarak tanımlanmaktadır. (Örneğin, ortak ilgileri olan kullanıcı topluluklarının keşfi) (Xu v.d., 2011).

### 1.3.3 Web Kullanım Madenciliği (Web Usage Mining)

Web kullanım madenciliğinde hedef, bir web sitesiyle ilişkili kullanıcı profillerini ve davranışsal örüntüleri analiz etmek ve modellemektir (Liu, 2007). İnternetin gelişmesi mevcut bilginin yayılmasına öncülük etmiş ve bu bilginin kişiselleştirilmesi ihtiyaç haline gelmiştir (Batista ve Silva, 2001).

E-ticaret, web hizmetleri ve web tabanlı bilgi sistemlerinin sürekli gelişimi ve yayılmasıyla, web-tabanlı organizasyonların günlük operasyonlarında toplanan tıklama akışı ve kullanıcı verisi hacmi astronomik boyutlara ulaşmıştır. Bu büyüklükte veriyi analiz etmek, yaşam boyu müşteri değerini sağlamak, web-tabanlı uygulamaların fonksiyonelliğini optimize etmek, ziyaretçilere göre içeriğin kişiselleştirilmesini sağlamak, web alanı için en etkin mantık yapısını bulmak, etkin kampanyalar düzenlemek gibi ürün ve hizmetlere yönelik pazarlama stratejileri tasarımına yardımcı olabilmektedir (Liu, 2007).

Web kullanıcılarının tercihlerini öğrenerek elde edilen bilgi, kullanıcı davranışlarına web bilgi yapısını uyarlayarak web site etkinliğini geliştirmede kullanılabilmektedir (Batista ve Silva, 2001).

Ham web günlük dosyalarını analiz etmek için ticari yazılımlar olsa da çoğu oldukça yavaş, esnek olmayan, pahalı ve verdikleri sonuçlar açısından sınırlıdır (Zaiane, Xin ve Han, 1998).

Web verisinden kullanıcı davranış örüntüleri keşfi yapabilmek için veri madenciliği yöntemlerindeki gelişmeler erişim günlüklerinden kullanıcı profilleri çıkarımını mümkün hale getirmiştir (Batista ve Silva, 2001).

Kullanıcı, bir tarayıcı ile web sunucularına erişen bir birey olarak tanımlanmaktadır. Ancak gerçekte tekil bir kullanıcıyı tanımlamak zordur. Kullanıcı, farklı makineler ile internete girmiş ya da bir kerede birden fazla tarayıcı kullanmış olabilir. Sayfa görüntüleme, tek bir kullanıcı tıklamasıyla ilişkili bir kerede kullanıcının tarayıcısında görüntülenen dosyalardır. Bir tıklama akışı (click-stream), istek yapılan sayfa görüntülemelerinin ardışık serisidir. Burada istemci ya da vekil seviyesinde cepleme aracılığıyla erişilip görüntülenen herhangi bir sayfanın sunucu tarafında kaydedilmeyeceğine dikkat edilmelidir. Bir sunucu oturumu ya da ziyareti özel bir web sitesi için tek bir kullanıcının tıklama akışıdır. Sunucu oturumunun sonu, o web sitesinde kullanıcının tarama oturumunun sonlandığı noktadır (Xu, Zhang ve Li, 2011). Bir kullanıcı, web sitesinin kullanışlı olup olmadığına karar vermek için ortalama 3 sayfa ziyaret etmektedir (Li, Liechty ve Montgomery, 2002).

## 1.4 Veri Madenciliği Süreci

Veri madenciliği gibi web kullanım madenciliğinde de CRISP-DM (Cross-Industry Standard Process for Data Mining) süreci kullanılabilmektedir. CRISP-DM'e göre bir veri madenciliği projesi altı aşamadan oluşan bir yaşam döngüsüne sahiptir. Sıradaki aşama önceki aşamanın çıktılarıyla ilişkilidir (Markov ve Larose, 2007). Süreç Şekil 1.1'de de özetlenmiştir.

1. Araştırmayı anlama:

a) Proje amaçlarını ve gerekliliklerini açıkça ortaya koymak
b) Amaçları ve kısıtlamaları veri madenciliği problem tanımına çevirmek
c) Bu amaçlara ulaşmak için bir öncül strateji hazırlamak

2. Veriyi anlama:

a) Veri toplama
b) Keşfedici veri analizi kullanarak ilk görüşleri keşfetmek
c) Veri kalitesini değerlendirmek
d) Gerekliyse, işlemeye uygun örüntü içeren alt setleri seçmek

3. Veriyi hazırlama:

a) İlk ham veriden bütün aşamalarda kullanılacak son veri setini hazırlamak
b) Analiz edilecek birimleri ve değişkenleri seçmek
c) İhtiyaç halinde belli değişkenleri dönüştürmek
d) Modelleme için ham veriyi temizlemek

4. Modelleme:

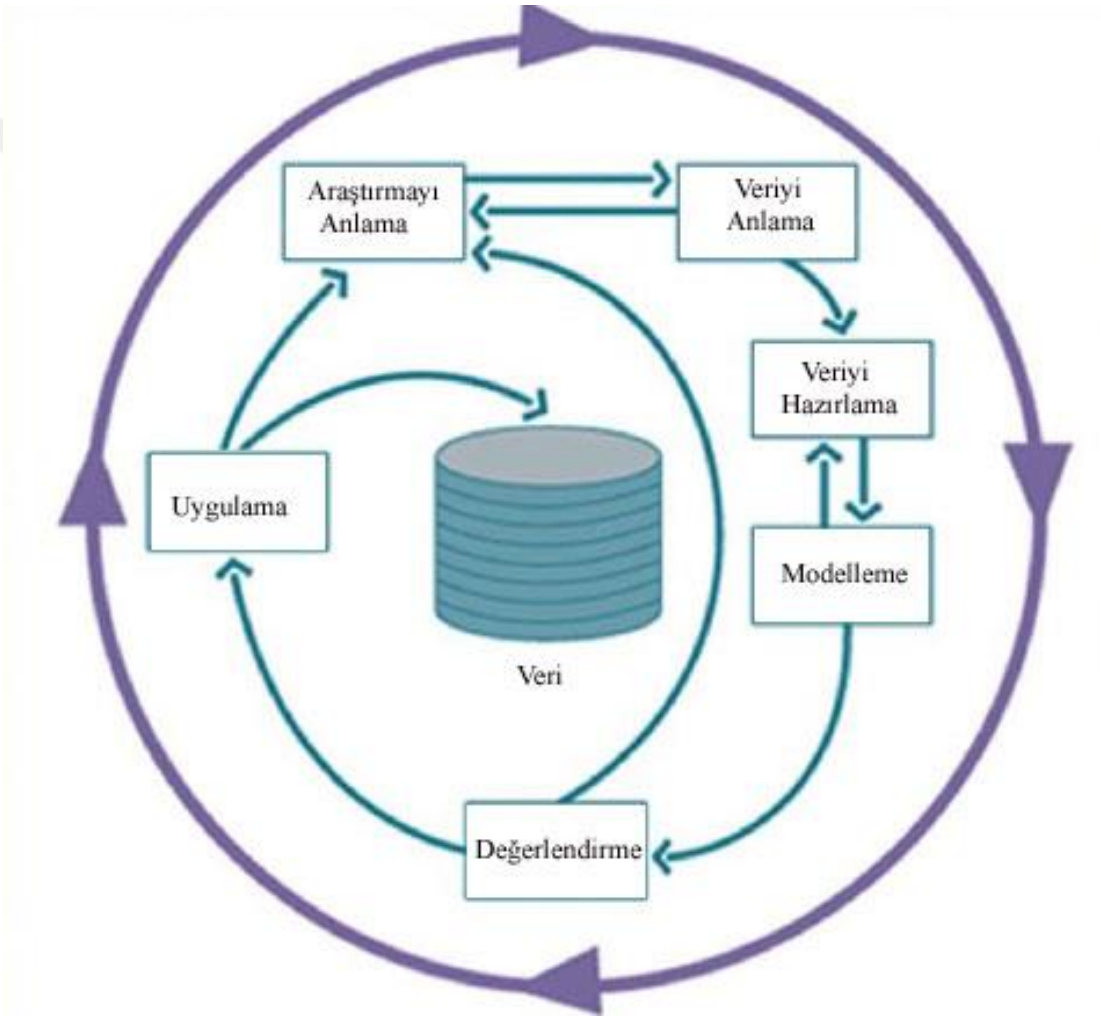
a) Uygun modelleme yöntemleri seçmek ve uygulamak
b) Sonuçları optimize etmek için model ayarlarını kalibre etmek
c) Aynı veri madenciliği problemi için farklı yöntemleri kullanmak
d) Veri hazırlama aşamasına dönerek veri biçiminde belli gereklilikleri gerçekleştirmek

5. Değerlendirme:

a) Modelleme aşamasında ortaya çıkan modelleri kullanmadan önce kalite ve etkinlikleri için değerlendirmek
b) Birinci aşamadaki amaçlara ulaşabilmek için model kurmak
c) Bazı önemli görevleri ya da araştırma problem özelliklerini ortaya çıkarmak
d) Veri madenciliğiyle elde edilen sonuçların kullanımıyla bir karara ulaşmak

6. Uygulama:

a) İşletme amaçlarına göre oluşturulmuş modelleri kullanmak
b) Basit bir uygulama örneği sağlamak ve bir rapor oluşturmak
c) Çok karışık uygulama örneği sağlamak, farklı bir birimde paralel bir veri madenciliği süreci gerçekleştirmek
d) Modele dayalı uygulama gerçekleştirmek




**Şekil 1-1 CRISP-DM süreç aşamaları (Chapman v.d., 2000)**

Web kullanım madenciliği için bir diğer çerçeve ise Srivastava v.d. (2000)'nin önerdiği süreç olup dört aşamadan oluşmaktadır. Girdi (input) aşaması, ön işleme aşaması (preprocessing stage), örüntü keşfi aşaması (pattern discovery stage), örüntü analizi aşaması (pattern analysis stage).

1. Girdi: Girdi aşamasında erişim günlükleri (access logs), referans günlükleri (referrer logs), vekil günlükleri (agent logs) olmak üzere üç tip ham web günlük dosyası kullanılmaktadır.

2. Ön İşleme: Ham web günlükleri veri madenciliğine olanak sağlayıcı bir biçimde değildir. Bundan dolayı ön işleme aşaması en önemli aşamalardan biridir. En yaygın kullanılan veri ön işleme görevleri (1) veri temizleme ve filtreleme, (2) robottan arındırma (de-spidering), (3) kullanıcı tanımlama, (4) oturum tanımlama, (5) yol tamamlamadır.

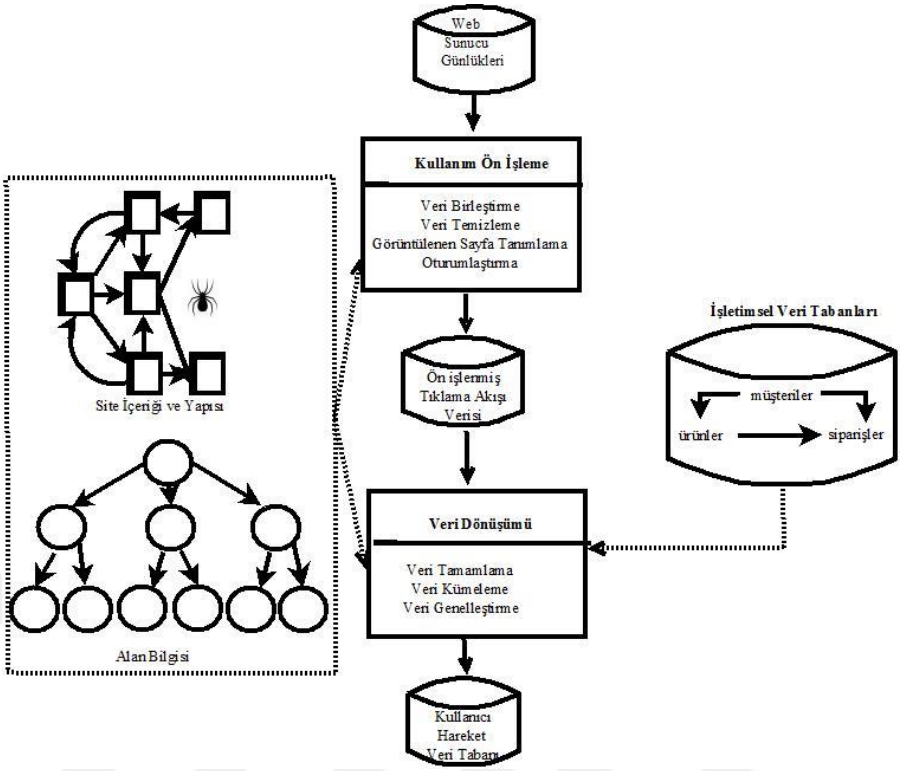
3. Örüntü Keşfi: Ön işleme tamamlandığında web verisi istatistiksel uygulamalar ve veri madenciliği yöntemleri için hazır hale gelmektedir. Bu yöntemler (1) standart istatistiksel analiz, (2) kümeleme algoritmaları, (3) birliktelik kuralları, (4) sınıflandırma algoritmaları, (5) ardışık örüntülerdir.

4. Örüntü Analizi: Örüntü keşif aşamasında meydana çıkarılan örüntülerin hepsi kullanışlı olmayabilir. Amaca uygun olarak seçilmelidir.

Veri madenciliği uygulamalarında önemli bir adım, veri madenciliği ve istatistiksel algoritmaların uygulanabileceği uygun hedef veri setinin oluşturulmasıdır. Bu, tıklama akışı verisinin kendine has özelliğinden ve farklı kaynaklardan toplanmış veriyle ilişkili olduğundan web kullanım madenciliğinde oldukça önemlidir. Bu sürece veri hazırlama denilmektedir. Veri hazırlama süreci web kullanım madenciliğinde en fazla zaman harcayan ve yoğun hesaplama gerektiren aşamadır. Süreç, veriden kullanışlı örüntülerin başarılı şekilde çıkarımı için kritik öneme sahiptir (Liu, 2007).

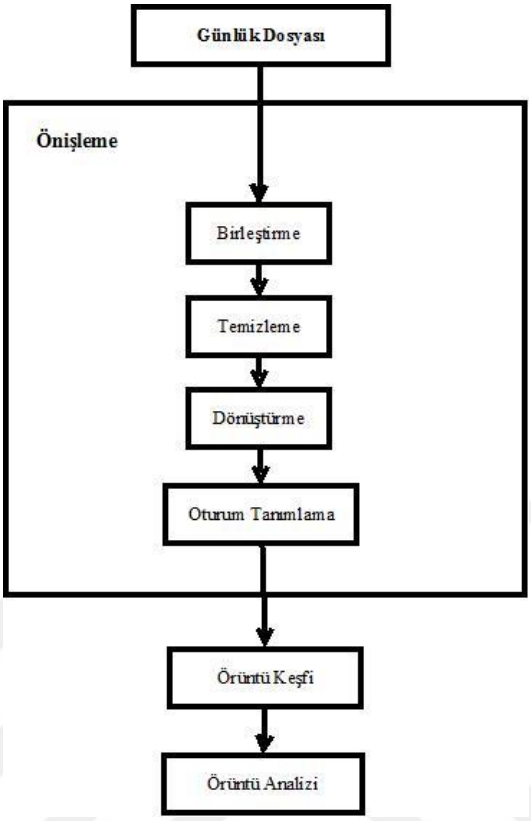
Bir e-ticaret sitesinin veri analizinde ön işleme yöntemlerinin doğru uygulanması, kullanıcı ve site metriklerini ortaya çıkarıcı öneme sahiptir (Kohavi v.d., 2004).

Web kullanım madenciliği bir ya da daha fazla web sitesi üzerindeki web kaynaklarıyla kullanıcı etkileşiminin sonucu olarak üretilen ve toplanan veriyle ilişkili örüntülerin keşfi ve analizidir. Web kullanım madenciliği süreci; veri toplama ve ön işleme, örüntü keşfi ve örüntü analizi olmak üzere üç aşamaya ayrılmaktadır (Thakare ve Gawali, 2010).




**Şekil 1-2 Web Kullanım Madenciliğinde Veri Hazırlama Adımları (Liu, 2007:451)**

Ön işleme; veri kaynaklarından toplanan, örüntü keşfi için gerekli veriyi düzenleme sürecidir. Uç değerler, hatalar ve web taramadan kaynaklı meydana gelebilen tamamlanmamış veri tespit edilmektedir. Sunucu günlüklerinde kaydedilen veri, kullanıcıların web sitesine erişimlerini yansıtmaktadır. Vekil (agent) ve IP adresleri, kullanıcıları ve oturumlarını tanımlamaktadır. Bununla birlikte bazı sayfa görüntülemeleri kullanıcının tarayıcısında ya da vekil sunucusunda önbelleklenmiş olabileceğinden, sunucu günlükleriyle toplanan verinin tümüyle güvenilir olamayacağı gözden kaçmamalıdır. Bir web sunucu günlüğünde bir vekil sunucusundan tüm istekler aynı tanımlayıcıya sahiptir. Web sunucu, çerezler (bireysel istemci tarayıcılarının site ziyaretçilerini otomatik izlemesi için web sunucu tarafından üretilmiş işaretçiler) gibi diğer kullanışlı bilgileri de saklayabilmektedir (Batista ve Silva, 2001).




**Şekil 1-3 Web Kullanım Madenciliği Süreci**

Her kullanıcı tanımlandıktan sonra her kullanıcı için tıklama akışı oturumlara bölünmelidir. Kullanıcının ne zaman web sitesini terk ettiği bilinemediğinden, kullanıcı oturumları oluşturulurken bir kesme noktası olarak genellikle zaman aşımı kullanılmaktadır (Joshi, Joshi ve Yesha, 2003).

Sonraki aşama ise örüntü keşif aşamasıdır. Bu aşamada kullanılan yöntemler ve algoritmalar istatistik, makine öğrenmesi ve veri tabanı gibi farklı alanlarda geliştirilmiştir. Web kullanım madenciliğinin bu aşamasında kullanılan üç yöntem vardır. Bunlar; birliktelik (hangi sayfalara birlikte erişildi), kümeleme (kullanıcı, işlem ve sayfa gruplarının bulunması) ve sıralı analiz (web sayfalarına hangi sırada erişildi) (Batista ve Silva, 2001).

Örüntü analizi web kullanım madenciliğinin son aşamasıdır. Bu aşamada önceki aşamada bulunan ilgisiz örüntü ve kurallar elenmektedir. Görselleştirme

teknikleri keşfedilmiş örüntüleri analiz etmede oldukça kullanışlıdır (Brusilovsky, Kobsa ve Nejdl, 2007).

## 1.5 Veri Toplama ve Ön İşleme

Bir kullanıcı internet üzerinde dolaştığında (bir istemci, tarayıcı vasıtasıyla bir web sunucusuna eriştiğinde) izler bırakmaktadır. Bu izler bilinçli olduğu gibi çoğu zaman da farkında olmadan bırakılmaktadır. Örneğin eğer bir internet kullanıcısı bir sosyal ağda resim ya da metin paylaşır ise bu kendi isteğiyle o sosyal ağ üzerinde bıraktığı izdir. Öte yandan kullanıcı bir arama motorunda sorgu yaptığında, arama motorunun o kişiyi IP bilgisi aracılığı ile tespit ederek neler sorguladığını öğrenmesi, kişinin istemeden iz bırakmasına örnektir.

Tez çalışmasına konu olan web kullanım madenciliğinde ise kullanılan izler, sunucu tarafından tutulan günlük kayıtlarıdır. Günlük kayıtları, ilgili web sitesinin sunucularında tutulan işlem kayıtlarıdır. Siteye giriş yapan her kullanıcının gerçekleştirdiği her bir istemci talebi kaydedilmektedir. Günlük kayıtlarını daha ayrıntılı açıklamak amacıyla bir internet kullanıcısının bir web sitesini bulma ve sitede dolaşma sürecini incelemek gerekmektedir. Kullanıcı, web sitesine farklı yollarla ulaşmak isteyebilir. Bir yol, eğer web site adresini eksik hatırlıyor ya da hiç hatırlamıyorsa arama motoruna başvurmasıdır. Bir diğer yol, adresi hatırlıyor ise tarayıcıya adresi yazarak siteye ulaşmasıdır. Farklı bir yol da bir başka web sitesinden verilen bağlantı aracılığıyla o web sitesine ulaşmasıdır. Sunucular ilk adım olarak kullanıcının siteye hangi yol ile geldiğini kaydetmektedir. Sonrasında ise hangi sayfaları ziyaret ettiği, sayfalarda ne kadar süre harcadığı, web sitesine hangi sayfayla giriş yapıp hangi sayfadan çıktığı gibi kullanıcı işlemleri sunucularda saklanmaktadır. Ancak bu tür veri yapılandırılmamış veri (unstructured data) olup oldukça karmaşıktır. Günlük kayıtlarından bir kesit Şekil 1.8'de verilmiştir. Birinci kullanıcının siteye ilk giriş yaptığı tarih ve sayfa ilk satırda yer almaktadır. Site üzerinde devam eden kullanıcı hareketleri ayrı satırlar olarak kaydedilmiştir. Her hareket bir satır olarak işlenmektedir. Burada önemli bir nokta her satırın yararlı bilgi içermemesidir. Günlük kayıtları bu bağlamda yararlı veriden çok gürültülü veri içermektedir.

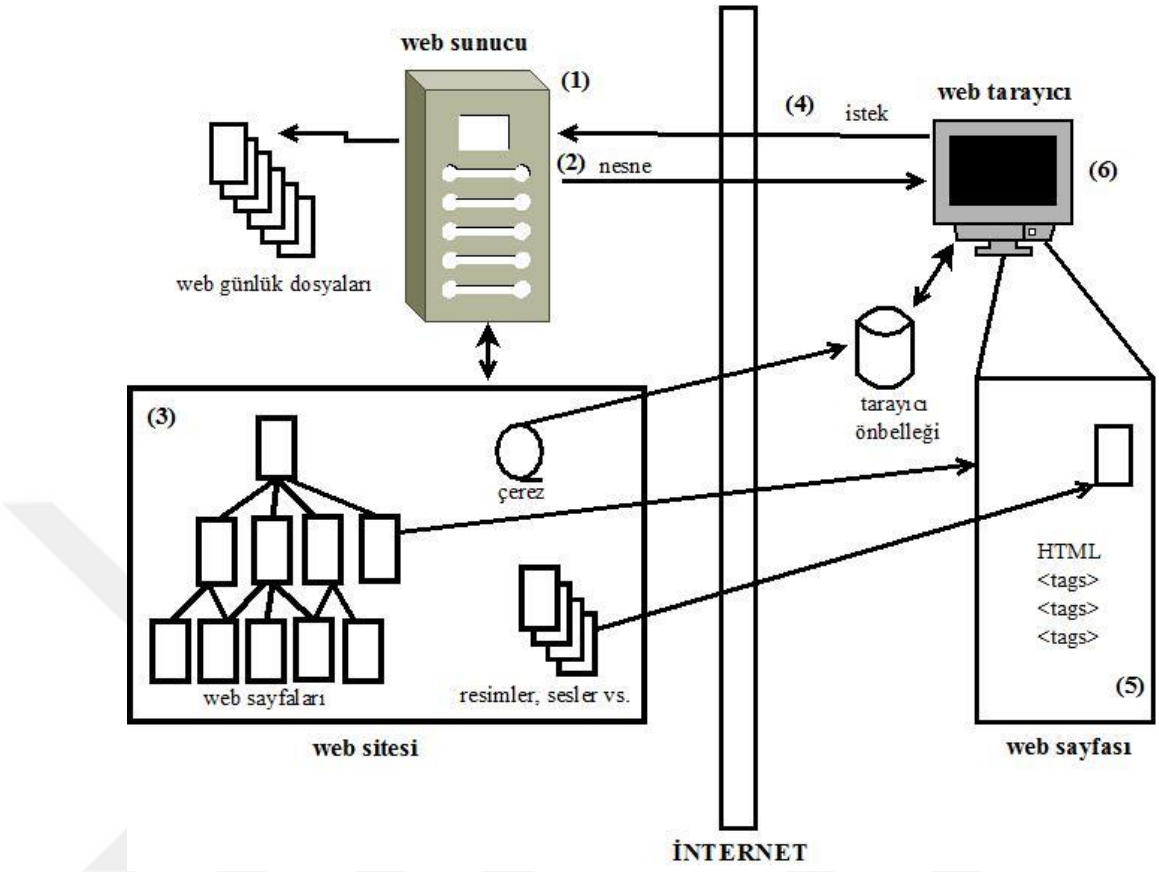
Yapılandırılmış veri, işletme kurallarına göre tanımlanan tablolar halinde organize edilmiş geleneksel veri tabanlarında bulunan veridir. Yapılandırılmış veri genellikle çalışması en kolay veri tipidir. Bu tip veri; tanımlanmış, dizilenmiş ve kolayca ulaşılıp ayrıştırılabilirdir. Yapılandırılmamış veri ise; tablolar halinde düzenlenmemiş, uygulamalar ile kullanılamamış ya da bir veri tabanı aracılığıyla yorumlanmamış veridir (Ohlhorst, 2013).

Bazı araştırmacılar ise yapısal olmayan veri teriminin aldatıcı olduğunu, çünkü her dokümanın oluşturulduğu yazılım içerisinde kendine özel yapısı ya da biçimi olduğunu savunmaktadır. Bununla birlikte dokümanın iç yapısı gerçekte yapılandırılmamıştır (Bramer, 2013).

Web günlük dosyaları ilk başta yapılandırılmamış veri tipine benzese de veri ön işleme aşamasıyla yapılandırılmış veri haline getirilmektedir. Örnek bir web günlük dosyası Şekil 1.8'de verilmiştir.

## 1.6 Web Verisi

Farklı web veri kaynakları arasında çoğu web madenciliği algoritmaları için girdi olarak kullanılan web günlükleri, web sayfaları ve web site bağlantı yapıları dikkat çekmektedir. Web kullanım madenciliğinde birincil veri kaynakları, web sunucu erişim günlükleri ve uygulama sunucu günlüklerini içeren sunucu günlük dosyalarıdır (Velasquez ve Palade, 2008).




**Şekil 1-4 Web Sunucu - Web Tarayıcı Etkileşimi**

Web işletim süreci bir web istemcinin bir web sunucusundan hizmet talep etmesine dayanmaktadır. Bu etkileşim, bir kullanıcı web tarayıcıda bir web adresi (URL) yazarak ya da diğer bir web sayfasındaki bir bağlantıya tıklayarak bir web sitesine erişmek istediğinde başlamaktadır. Ardından http kullanan web sunucuya bir istek gönderilmekte, web sunucu da istenen cevabı döndürmektedir. Şekil 1.4'te web sunucu ile tarayıcı arasındaki etkileşim gösterilmiştir. Web sunucu (1), özel bir port (genellikle 80 portu) aracılığıyla alınan isteği (4) geride çalıştıran bir programdır. Bu program, web sitesi (3) olarak bilinen yapısal bir dosya setini yönetmektedir. Bu dosyaların bir bölümü resim, ses, film ya da diğer sayfalar gibi veri dosya tiplerine linkler içeren web sayfalarıdır. Web sunucusunun ana fonksiyonu web sayfa isteklerini gerçekleştirmektir. Web sayfası (5) HTML kullanılarak yazılmıştır (www.w3.org/MarkUp/). HTML, kullanıcı ekranında tarayıcı aracılığıyla istenen nesnelerin nasıl gösterileceği ve web sunucudan diğer nesnelerin nasıl okunacağı

üzerine ilişkili yöntemler, etiket denilen sıra dizgeleri içermektedir. Bu talepler kullanıcı ekranında nesneleri gösteren tarayıcıyla yorumlanmaktadır.

Belge bir kere tarayıcı tarafından okunduğunda, özel iç etiketler yorumlanabilmektedir. Etiketleri yorumlanan tarayıcı, görüntü gibi bir nesne hakkında bir referans bulduğunda HTTP bunu tarayıcıya yollamaktadır. İşlem son etiket yorumlandığında ve sayfa kullanıcıya gösterildiğinde tamamlanmaktadır.

Sayfanın tarayıcı hızı birkaç faktöre bağlıdır. En önemlisi tarayıcı ve sunucu arasındaki mevcut bağlantının bant genişliği (bandwith) olup, ISP (internet service provider – internet servis sağlayıcı) ve bağlantı için kullanılan cihaza (modem ya da ADSL - asymmetric digital subscriber line- sistem) bağlı olarak değişmektedir.

İkinci faktör bilgisayarın hızıdır ve kısmen tarayıcıyla ilgilidir. Çoğu sayfanın içeriği karışıktır (Java Applets, Java Scripts, Dynamic HTML, vs. içerebilir). Bu durumlarda CPU, RAM ya da Hard Disk gibi bilgisayar kaynakları yeterli gelmeyecektir. Çoğu ticari site için bu büyük bir problemdir.

## 1.7 Web Günlük Dosyalarının Ön İşleme Süreci

Veri hazırlama süreci genelde en çok zaman alan web kullanım madenciliği sürecidir. Süreç, orijinal veriyi ön işleme, çok sayıda kaynaktan veriyi entegre etme ve belli veri madenciliği işlemleri için uygun biçime getirmek amacıyla entegre edilmiş verinin dönüştürülmesi şeklindedir (Thakare ve Gawali, 2010).

Web sunucu günlüğündeki bir satır, sunucuya yapılan bir isteği temsil etmektedir. Genellikle bu istek ortak günlük biçiminde (common log format) olup; uzak sunucu (host), istemci, id, isteğin yapıldığı tarih ve zaman, HTTP isteği, durumu, sayfa boyutu, referans ve kullanıcı temsilini içermektedir. Veri ön işleme aşaması birkaç adımı içermektedir. Bunlar; veri temizleme, kullanıcı tanımlama, kullanıcı oturum tanımlama, yol tamamlama ve işlem tanımlamadır (Prasad, Reddy ve Acharya, 2010).

### 1.7.1 Veri temizleme

Bir günlük girişi, web sunucusuna ulaşan bir isteğin otomatik olarak eklenmesidir. Tipik bir temizleme süreci, web örümcekleri (web spiders), dizin oluşturucular (indexers), bağlantı kontrol ediciler (link checkers) ya da belleğe önceden yüklenmiş sayfalardaki diğer akıllı temsilciler gibi web temsilciler aracılığıyla üretilen günlük girişlerini elemeye dayanmaktadır (Zaiane, Xin ve Han, 1998).

Bir günlük dosyası; uzak sunucu alanı (remote host field (IP adresi)), tarih/zaman alanı, HTTP isteği, statü kod alanı gibi alanlar içermektedir. Web günlük ön işlemede ilk adım değişken çıkarımıdır. Değişken çıkarım adımları aşağıdaki gibidir (Markov ve Larose, 2007):

1. Tarih/zaman alanından tarih değişkeninin çıkarımı
2. Tarih/zaman alanından zaman değişkeninin çıkarımı
3. HTTP istek alanından istek yönteminin çıkarımı
4. HTTP istek alanından sayfa (URI) çıkarımı
5. HTTP istek alanından protokol versiyonunun çıkarımı

Değişken çıkarımının ardından zaman damgası oluşturma adımı gelmektedir. Zaman damgası, aşağıdaki adımlarla oluşturulmaktadır ve kullanıcı ziyaretinin süresini tahmin etmek için gereklidir.

1. Web günlük giriş tarihi ile yazılımın tarihi arasında geçen gün sayısının bulunması
2. Gün sayısının 24 saatlik bir günde saniyeye karşılık gelen 86400 ile çarpılması
3. Web günlük girişinde zaman ile temsil edilen, gece yarısından itibaren geçen saniyelerin hesaplanması
4. İkinci adıma üçüncü adımın eklenmesi

Veri temizleme web kullanım madenciliği ön işleme sürecinin ilk adımıdır. Veri temizleme süreçleri aşağıdaki şekilde sıralanabilir (Prasad, Reddy ve Acharya, 2010):

### a. Web günlük dosyalarının örümceklerden arındırılması

Arama motorları tarafından üretilen erişim kayıtları tanımlanmalı ve web günlük dosyalarından kaldırılmalıdır. Yöntemler sezgisel olabileceği gibi (Berendt ve Spiliopoulou, 2000) sınıflandırma tekniklerine de (Tan ve Kumar, 2000) dayalı olabilmektedir.

Web arama motorları, world wide web üzerindeki en güncel bilgiyi kullanıcılarına sağlamaktadırlar. Bunu gerçekleştirirken web sitelerinin kapsamlı araştırılması için web site içeriklerini inceleyebilen (crawl) yazılımlar (örümcekler (spiders), crawlers, robotlar (bots)) göndermektedirler. Robotların bu davranışı insan davranışından farklıdır. Örneğin robot, sırayla web sitesindeki her mümkün bağlantıya arka arkaya istekte bulunabilmektedir. Bu davranış bir web kullanım madenciliği bakış açısı olarak ilgi çekici değildir. Aslında bu davranış web günlük dosyasında tutulmuş ise insanların web sitesini nasıl kullandığının tahmin edilmesinde analiz, doğru sonuç vermeyecektir. Web kullanım madencisi bu gibi yazılımlarla gerçekleştirilen (nonhuman) erişim davranışı tiplerini web günlük dosyalarından arındırmalıdır (Markov ve Larose, 2007).

Web günlük dosyasından robotlar ve örümcekleri en açık eleme yöntemi kullanıcı temsil (user agent) alanında örümceğin ismini tanımlamaktır. İletişim amaçları için, robotlar sıklıkla bir URL ya da bir e-posta adresi içermektedirler. Site yöneticisi, web sayfasını inceleyen yazılımların bant genişliği gibi sunucu kaynaklarını gereğinden fazla tüketmemesi için, web sitesinin bazı bölümlerinin bilgi toplamaya kapanmasını isteyebilir (Kreps, 2015). Web sayfasını inceleyen yazılımlara bir örnek Google robotudur.

### b. Sayfa uzantı keşfi ve filtrelenmesi

Görüntü dosyaları ve özel bir sayfayla ilişkili ilgisiz diğer dosyalar için istekler, web günlük dosyasında sayfa uzantı keşfiyle web günlük dosyalarından elenir (Das ve

Turkoglu, 2009). Değişken çıkarımı, zaman damgalama ve veri filtreleme için genelleştirilmiş bir algoritma Arumugam ve Suguna (2008) tarafından sunulmuştur.

### 1.7.2 Kullanıcı Tanımlama

Bu adımın amacı her bir ayrı kullanıcıyı tanımlamaktır. İdeal olan kullanıcının web sitesine erişimini kullanıcı adı ve şifresi gibi kayıt bilgileriyle gerçekleştirmesidir. Ancak çok sayıda kullanıcının çok sayıda web sitesine anonim olarak eriştiği internetin biçimden bağımsız yapısı, kayıt bilgisini mümkün kılmamaktadır (Markov ve Larose, 2007). Kullanıcı işbirliği gerektiren Java tabanlı uzak temsil (remote agent) kullanılabileceği gibi (Shahabi v.d., 1997), IP adresi, vekil kullanıcı (user agent) ve URL'i çalıştıran sezgisel yöntemler de (Cooley, Mobasher, ve Srivastava, 1999) bulunmaktadır. Genelleştirilmiş kullanıcı tanıma algoritmaları Arumugam ve Suguna (2008) ile Huiying ve Wei (2004) tarafından önerilmiştir.

Kullanıcı tanımlamada diğer bir yol çerezleri kullanmaktır. Bir çerez, sistematik bir metin yapısında olup, mevcut web sayfasına erişimlerin önceki web sayfasına erişimlerine bağlantı gerçekleştirmede kullanılmaktadır. Bununla birlikte çoğu kullanıcı, mahremiyetlerine bir tehdit olduğundan endişe etmektedirler. Bu da kullanıcı tanımlama için farklı stratejiler bulmayı gerektirmektedir (Markov ve Larose, 2007).

Uzak sunucu alanı (remote host field) ya da IP adres alanı prensipte kullanıcı tanımlama için kullanılabilmektedir. Ancak vekil sunucuların yaygın kullanımı, şirket güvenlik duvarları ve yerel bellekler, kullanıcı tanımlamada IP adresinin kullanımını güç hale getirmektedir. Örneğin birkaç kullanıcı aynı siteye bir vekil sunucu kullanarak erişiyor olsun. Bu durumda vekil sunucu her kullanıcı için aynı IP adresli web sunucu sağlayacaktır (Markov ve Larose, 2007).

Hipotetik bir web sitesi için kurgulanmış bir web günlüğünden bir parça Şekil 1.5'te verilmiştir. İlk bakışta bütün IP adreslerinin aynı olduğu, bütün girdilerin aynı kullanıcıdan geldiği görülecektir. Ancak bu görüş yanlıştır. Şu şekilde sezgisel bir varsayım göz önüne alınabilir: Temsil alanı iki web günlük girişi için farklılaşıyorsa, istekler iki farklı kullanıcıdandır. Bu varsayım aynı makine üzerinde iki farklı

tarayıcıyla aynı web sitesine erişen kullanıcıları göz ardı etmesine karşın, bu davranış çeşidi seyrektir (Xu, Zhang ve Li, 2011).

| IP Address | Time | Method | Referrer | Agent |
|---|---|---|---|---|
| 987.654.32.1 | 00:00:02 | "GET A.html HTTP/1.1" | — | Mozilla/4.0 (Windows NT 5.1, MSIE6.0) |
| 987.654.32.1 | 00:00:05 | "GET B.html HTTP/1.1" | A.html | Mozilla/4.0 (Windows NT 5.1, MSIE6.0) |
| 987.654.32.1 | 00:00:06 | "GET A.html HTTP/1.1" | — | Mozilla/5.0 (Linux 1.0, Firefox/0.9.3) |
| 987.654.32.1 | 00:00:10 | "GET E.html HTTP/1.1" | B.html | Mozilla/4.0 (Windows NT 5.1, MSIE6.0) |
| 987.654.32.1 | 00:00:17 | "GET K.html HTTP/1.1" | E.html | Mozilla/4.0 (Windows NT 5.1, MSIE6.0) |
| 987.654.32.1 | 00:00:20 | "GET C.html HTTP/1.1" | A.html | Mozilla/5.0 (Linux 1.0, Firefox/0.9.3) |
| 987.654.32.1 | 00:00:27 | "GET I.html HTTP/1.1" | — | Mozilla/4.0 (Windows NT 5.1, MSIE6.0) |
| 987.654.32.1 | 00:00:36 | "GET G.html HTTP/1.1" | C.html | Mozilla/5.0 (Linux 1.0, Firefox/0.9.3) |
| 987.654.32.1 | 00:00:49 | "GET O.html HTTP/1.1" | I.html | Mozilla/4.0 (Windows NT 5.1, MSIE6.0) |
| 987.654.32.1 | 00:00:57 | "GET M.html HTTP/1.1" | G.html | Mozilla/5.0 (Linux 1.0, Firefox/0.9.3) |
| 987.654.32.1 | 00:03:15 | "GET H.html HTTP/1.1" | — | Mozilla/5.0 (Linux 1.0, Firefox/0.9.3) |
| 987.654.32.1 | 00:03:20 | "GET N.html HTTP/1.1" | H.html | Mozilla/5.0 (Linux 1.0, Firefox/0.9.3) |
| 987.654.32.1 | 00:31:27 | "GET E.html HTTP/1.1" | K.html | Mozilla/4.0 (Windows NT 5.1, MSIE6.0) |
| 987.654.32.1 | 00:31:34 | "GET L.html HTTP/1.1" | E.html | Mozilla/4.0 (Windows NT 5.1, MSIE6.0) |

**Şekil 1-5 Hipotetik bir web sitesi için örnek web günlük dosyası**

Şekil 1.5'te bu sezgisel yaklaşım uygulandığında en az iki kullanıcının algılandığı görülmektedir. Biri Windows NT ve MS Internet Explorer, diğeri ise Linux ve Firefox kullanmaktadır. Buna göre her bir kullanıcı tarafından web sitesi üzerinde izlenen yollar aşağıdaki şekildedir (Markov ve Larose, 2007):

Kullanıcı-1: A – B – E – K – I – O – E – L

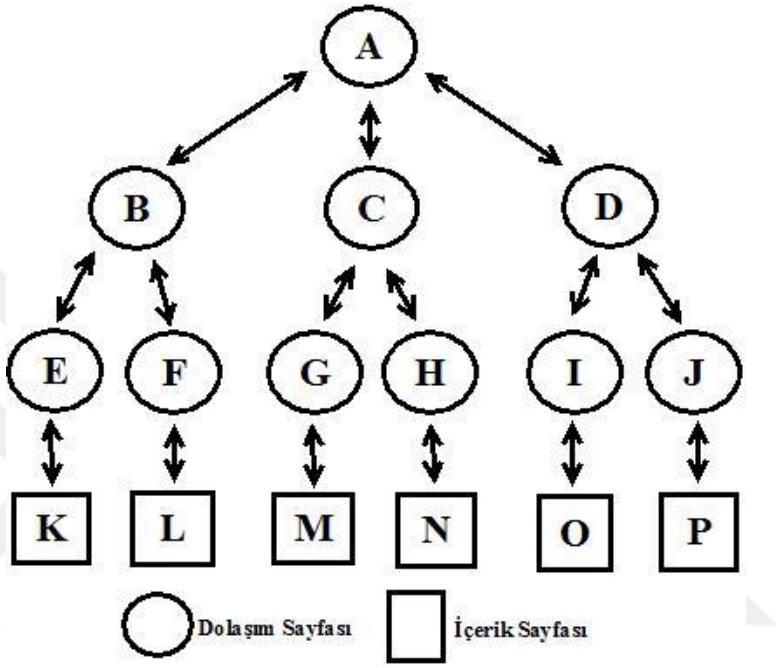
Kullanıcı-2: A – C – G – M – H – N

Referans alanından elde edilen bilgiye göre, web sitesi topolojisiyle (Şekil 1.6) Kullanıcı-1'in aslında iki kişi olduğu görülmektedir. Kullanıcı-1'in web sitesinde referans alana göre takip ettiği yolu incelemek gerekmektedir. B.html'ye erişimin A.html'den, E.html'ye erişimin B.html'den ve K.html'ye erişimin ise E.html'den referans edildiği gözükmektedir. Ancak I.html sayfası için bir referans yoktur. Web site topolojisine bakıldığında oklar yönlü bağlantıyı ifade etmektedir. K.html ve I.html arasında doğrudan bağlantı yoktur. Buradan, I.html sayfası için doğrudan o sayfaya erişen (muhtemelen aynı tarayıcı versiyon ve işletim sistemi kullanan, tarayıcıya doğrudan URL'yi giren) üçüncü bir kullanıcıdan bahsedilebilmektedir. O.html'ye de tek erişim I.html'dendir. Referans bilgisi bu üçüncü kullanıcının I.html'den O.html'ye tıkladığı sonucunu çıkarmayı desteklemektedir. Sonuç olarak, bu web günlük dosyasında üç farklı kullanıcının varlığı için kanıtımızın olduğu görülmektedir (Markov ve Larose, 2007).

Kullanıcı-1: A – B – E – K – E – L

Kullanıcı-2: A – C – G – M – H – N

Kullanıcı-3: I – O




**Şekil 1-6 Hipotetik bir web sitesi için bağlantılar**

Genellikle kullanıcı tanımlamada aşağıdaki süreç takip edilmektedir:

1. Web günlük dosyalarını IP adreslerine göre sıralamak ve ardından zaman damgalamak
2. Her bir farklı IP adresi için farklı bir kullanıcıya ait olan bir temsil tanımlamak
3. İkinci adımda tanımlanan her bir kullanıcı için referans alanından toplanan yol bilgisi ve bu davranışın iki ya da daha fazla kullanıcıyla gerçekleşip gerçekleşmediğini anlamak için site topolojisini uygulamak
4. Her bir kullanıcıyı tanımlamak için mevcut çerez ve kayıt bilgisiyle 1'den 3'e adımları birleştirmek

### 1.7.3 Kullanıcı Oturum Tanımlama

Oturum tanımlama kullanıcı hareketleri sırasında toplanmış IP adreslerinin tüm sayfa görüntülenme kayıtlarının gruplanmasından oluşmaktadır.

Oturum, belli bir amaç için belli bir kullanıcının görüntülediği web sayfaları seti olarak tanımlanabilmektedir (Markov ve Larose, 2007), (Batista ve Silva, 2001). Oturum tanımlama her kullanıcının erişim günlüklerinin bireysel erişim oturumlarına ayrılması sürecidir (Prasad, Reddy ve Acharya, 2010). Zamanın uzun dönemlerine yayılan günlükler için, kullanıcılar web sitesini bir defadan fazla ziyaret edecektir. Oturum tanımlama için çeşitli stratejiler takip edilmektedir (Das ve Turkoglu, 2009). Deneysel bulgulardan keşfedilen; toplam oturum süresi için 30 dakikalık bir eşik, kullanıcıların hareketsiz kalma süresi olarak çoğu uygulamada kullanılmış ve tavsiye edilmiştir (Spiliopoulou v.d., 2003). Baglioni v.d. (2003) referans uzunluğu yaklaşımını tavsiye etmektedirler. Bu yaklaşımda bir kullanıcının bir web sayfasında harcadığı zamanın, sayfa içeriğine duyduğu ilgiyle orantılı olduğu varsayılmaktadır. Referans uzunluğu yaklaşımı, sadece bağlantıların olduğu içeriğin olmadığı gezinti sayfaları ile kullanıcının istediği bilgiyi barındıran içerik sayfaları arasındaki ayırımın yapılmasını sağlamaktadır.

Zaman gecikmesi $d_{A\to B}$, A sayfası için istek ile B sayfası için istek arasındaki zaman farkı olarak tanımlanmaktadır. Diğer bilgilerin eksikliğinde verilen bir $t$ eşik zamanı için $d_{A\to B} > t$ ise A bir içerik sayfası olarak, aksi durumda ise bir navigasyonel sayfa olarak sınıflanacaktır. Baglioni v.d. (2003) bir kullanıcı oturumunu "bir içerik URL'si ile takip edilen navigasyonel URL sırası" olarak tanımlamıştır. Ancak bu tanım aynı oturumda kullanıcının içerik sayfasından navigasyonel sayfaya, sonrasında tekrar içerik sayfasına hareket edebileceğinden kısıtlıdır.

Kullanıcı oturumu, bir siteye hangi sıklığın bir "ziyaret" olarak karşılık geleceğini kestirme açısından önemlidir. Örneğin bir kullanıcı 24 saat içinde bir web sitesini iki kez ziyaret etsin. Ziyaretler 6 saat arayla yapılsın. Eğer bu 24 saatlik dönem için kullanıcı tanımlama yöntemleri erken uygulanırsa, iki ziyaret beraber sıralanacak ve bu kullanıcıyla tanımlanacaktır. Ancak bu iki ziyaret arasında fark olması gerekir. Bu, oturum tanımlama olarak adlandırılmaktadır. Oturum tanımlama uzun bir zaman aralığında belli bir kullanıcı tarafından yapılan birleşmiş sayfa isteklerinin, bireysel oturumlara parçalanma sürecidir (Markov ve Larose, 2007).

En kolay oturum tanımlama yöntemi ise, kullanıcının son isteğinden belli bir zaman geçtikten sonra zaman aşımı uygulanmasıdır. Çoğu web analist ve ticari uygulamalar, zaman aşımı eşiğini 30 dakika olarak belirlemektedir. Bir oturum tanımlama algoritması web günlük verisine bu eşik zaman aşımı uygulandıktan sonra uygulanabilmekte ve yeni bir oturum, eşiği geçen iki istek arasında fark olduğunda başlayacak şekilde tanımlanmaktadır (Markov ve Larose, 2007), (Das ve Turkoglu, 2009), (Xu v.d., 2011).

Örnek günlük dosyasında Kullanıcı-1 için, K.html sayfası (00:00:17 zamanında) ve E.html sayfası (00:31:27) için istekler arasındaki gecikme 30 dakikadan fazladır. Oturum tanımlama algoritması E.html sayfası için gerçekleşen bu ikinci isteği yeni bir oturum olarak tanımlamaktadır. Buradan da 4 oturum tanımlanabilmektedir.

Oturum 1 (Kullanıcı-1): A – B – E – K

Oturum 2 (Kullanıcı-2): A – C – G – M – H - N

Oturum 3 (Kullanıcı-3): I - O

Oturum 4 (Kullanıcı-1): E – L

Oturum tanımlama süreci;

1. Daha önce tanımlanmış her bir farklı kullanıcı için tekil bir oturum ID'si atamak,
2. Bir $t$ zaman aşımı eşiği tanımlamak,
3. Her kullanıcı için;
   a) Her ardışık iki web günlük girişi arasındaki zaman farkını bulmak,
   b) Bu fark $t$ eşiğini aşıyor ise, bir sonraki girişe yeni bir oturum ID'si tanımlamak,
4. ID'lerine göre oturumları sıralamak

olarak sıralanmaktadır.

### 1.7.4 Yol tamamlama ve işlem tanımı

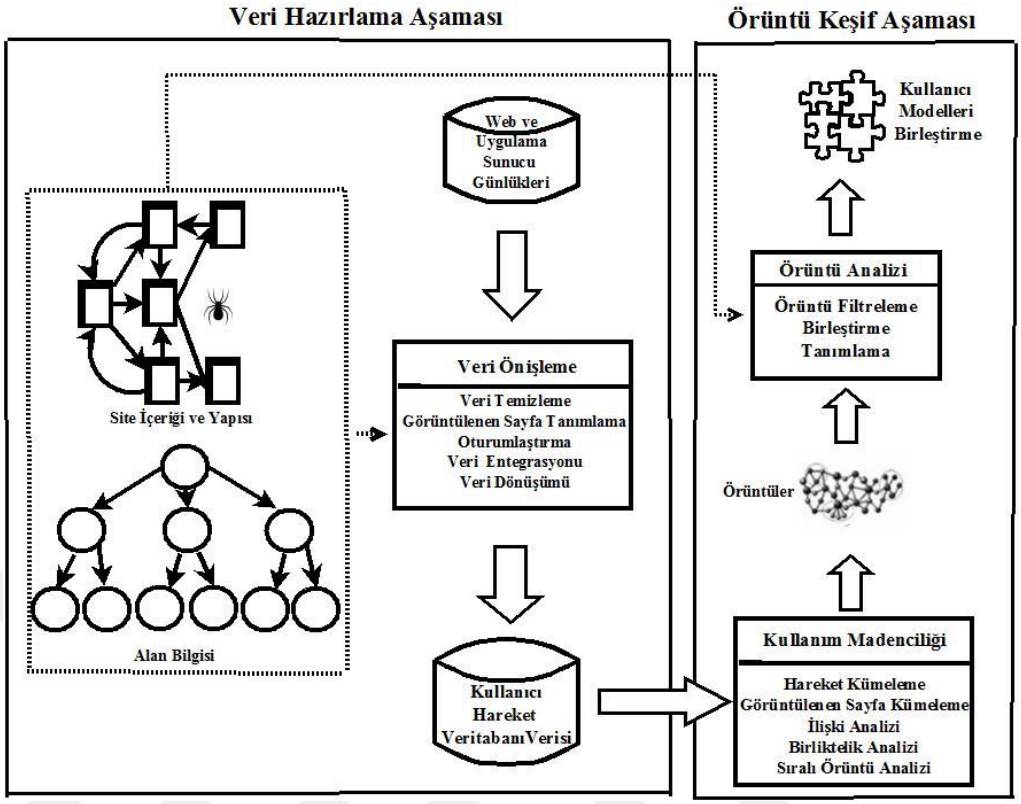
Yol tamamlama, tarayıcı ve vekil sunucu belleğine erişim günlüğünde kayıp olan önemli sayfa erişim kayıtlarını kapsamayı temsil etmektedir (Prasad, Reddy ve Acharya, 2010). Bu amaçla, referans günlüğü ve site topolojisine dayalı sezgisel yöntemler geliştirilmiştir (Cooley, Mobasher ve Srivastava, 1999).

İşlem tanımlamanın hedefi her kullanıcı için anlamlı referans kümeleri oluşturmaktır. İki tane işlem tanımlama yaklaşımı bulunmaktadır (Das ve Turkoglu, 2009). İlki maksimum ileri referans (her işlem, bir kullanıcının oturum açtığı ilk sayfadan, bir geri başvuru yapmadan önceki sayfaya kadar olan yoldaki sayfa seti olarak tanımlanmaktadır), ikincisi ise referans uzunluğu (bir kullanıcının bir sayfada harcadığı zaman miktarına dayalı olup sayfa, o kullanıcı için yardımcı ya da içerik sayfası olarak sınıflanmaktadır) yaklaşımıdır. Bir kaynak üzerine harcanan zaman çok uzun ise kaynak bir içerik olarak, diğer durumlarda ise yardımcı referans olarak tanımlanır (Cooley, Mobasher ve Srivastava, 1999).

## 1.8 Web Günlük Dosyaları

Web kullanım madenciliği uygulamaları üç ana kaynaktan toplanan veriye dayanmaktadır. Bu kaynaklar; web sunucuları, vekil sunucular ve web kullanıcılarıdır (Facca ve Lanzi, 2005).

Web sunucuları genellikle en zengin ve en yaygın veri kaynağıdır. Günlük dosyalarında ve kullandıkları veri tabanlarındaki günlük dosyalarında büyük miktarlarda bilgi biriktirebilmektedirler. Bu dosyalar genellikle uzak sunucunun ismi ve IP numarası, tarihi, istek yapılan zaman, müşteriden gelmiş istek satırı gibi temel bilgileri barındırmaktadırlar. Bu bilgi genellikle “ortak günlük biçimi” (common log format), “genişletilmiş günlük biçimi” (extended log format) ve “LogML” gibi standart biçimde sunulmaktadır. Bazen veri tabanları metin dosyalarının günlük bilgisini depolamak yerine, günlük depo sorgulamalarını geliştirmede kullanılmaktadır.




**Şekil 1-7 Web Kullanım Madenciliği Süreci (Liu, 2007:450)**

Web kullanım madenciliğinde kullanılan birincil veri kaynakları web sunucu erişim günlükleri ve uygulama sunucu günlüklerini içeren sunucu günlük dosyalarıdır. İlave veri, müşteri taraflı ya da vekil sunucudan elde edilebilir. Bir sitedeki içerik verisi nesnelerin ve kullanıcılara nakledilen ilişkilerin birikimidir. Yapılandırılmış veri site içerisinde tasarımcının içerik kurgusunu sunmaktadır. Bu kurgu sayfalar arasında bağlantı yapısıyla, bağlantıların yansıması olarak yakalanabilmektedir. Site için operasyonel veri tabanı ilave kullanıcı profil bilgisi içerebilmektedir (Thakare ve Gawali, 2010).

Web kullanım bilgisi, web sunucu günlük dosyalarından alınmaktadır. Bir kullanıcının tarayıcısından bir web sunucusuna gelen her bir istek için, bir cevap otomatik olarak üretilmektedir. Bu cevap web sunucusu üzerinde bir ASCII metin dosyasına eklenmiş olarak basit bir tek-satırlık kayıt biçimindedir. Metin dosyası virgül-sınırlı (comma-delimited), boşluk-sınırlı (space-delimited) ya da sekme-sınırlı (tab-delimited) olabilmektedir (Markov ve Larose, 2007).

Örnek bir web günlük verisi olan EPA'ya ait bir kesit Şekil 1.8'de verilmiştir. EPA web günlük verisi Internet Traffic Archive (http://ita.ee.lbl.gov/html/traces.html) adresinde mevcuttur. Bu dosyadaki her satır North Carolina Research Triangle Park'taki EPA web sunusu aracılığıyla alınmış, bir kullanıcının tarayıcısıyla istenmiş belli bir hareketi temsil etmektedir. Her satır (kayıt) aşağıdaki şekilde tanımlanmaktadır (Markov ve Larose, 2007).

```
141.243.1.172 [29:23:53:25] "GET /Software.html HTTP/1.0" 200 1497
query2.lycos.cs.cmu.edu [29:23:53:36] "GET /Consumer.html HTTP/1.0" 200 1325
tanuki.twics.com [29:23:53:53] "GET /News.html HTTP/1.0" 200 1014
wpbfl2-45.gate.net [29:23:54:15] "GET /default.htm HTTP/1.0" 200 4889
wpbfl2-45.gate.net [29:23:54:16] "GET /icons/circle logo small.gif HTTP/1.0" 200 2624
wpbfl2-45.gate.net [29:23:54:18] "GET /logos/small gopher.gif HTTP/1.0" 200 935
140.112.68.165 [29:23:54:19] "GET /logos/us-flag.gif HTTP/1.0" 200 2788
wpbfl2-45.gate.net [29:23:54:19] "GET /logos/small ftp.gif HTTP/1.0" 200 124
wpbfl2-45.gate.net [29:23:54:19] "GET /icons/book.gif HTTP/1.0" 200 156
wpbfl2-45.gate.net [29:23:54:19] "GET /logos/us-flag.gif HTTP/1.0" 200 2788
tanuki.twics.com [29:23:54:19] "GET /docs/OSWRCRA/general/hotline HTTP/1.0" 302 -
wpbfl2-45.gate.net [29:23:54:20] "GET /icons/ok2-0.gif HTTP/1.0" 200 231
tanuki.twics.com [29:23:54:25] "GET /OSWRCRA/general/hotline/ HTTP/1.0" 200 991
tanuki.twics.com [29:23:54:37] "GET /docs/OSWRCRA/general/hotline/95report HTTP/1.0" 302 -
wpbfl2-45.gate.net [29:23:54:37] "GET /docs/browner/adminbio.html HTTP/1.0" 200 4217
tanuki.twics.com [29:23:54:40] "GET /OSWRCRA/general/hotline/95report/ HTTP/1.0" 200 1250
wpbfl2-45.gate.net [29:23:55:01] "GET /docs/browner/cbpress.gif HTTP/1.0" 200 51661
dd15-032.compuserve.com [29:23:55:21] "GET /Access/chapter1/s2-4.html HTTP/1.0" 200 4602
```

**Şekil 1-8 EPA Web sitesinden örnek web günlükleri**

### 1.8.1 Tarih / Zaman Alanı

EPA, web günlük [GG:SS:DD:SaSa] biçimini kullanmaktadır. Burada GG, ayın gününü ve SS:AA:SaSa ise 24 saatlik zamanı temsil etmektedir. Örnekte, 1995 yılının Ağustos ayına ait gerçekleşen web günlüğünün giriş bölümü sunulmaktadır. Bununla birlikte yaygın tarih/zaman alanı şu biçimdedir: "GG/Ay/YYYY:SS:DD:SaSa offset". Konum (offset), yerel sunucunun Greenwich zamanından (Greenwich Mean Time (GMT)) ne kadar saat geride ya da ileride kaldığını gösteren pozitif ya da negatif sabittir. Örneğin, "09/Jun/1998:03:27:00 -0500"nın tarih/zaman alanı; isteğin 9 Haziran 1998, saat 03:27'de GMT'nin 5 saat gerisinde gerçekleştiğini göstermektedir (Markov ve Larose, 2007).

### 1.8.2 Uzak Sunucu Alanı

Bu alan istekte bulunan uzak sunucunun internet IP adresini içermektedir (141.243.1.172 gibi). Eğer uzak sunucu, alan adına sahip ise (wpbfl2-45.gate.net gibi)

DNS aracılığıyla bu bilgi elde edilebilecektir. IP adresi yerine uzak sunucu alan adını elde etmek için sunucu, internet alan adı sistemini (DNS), IP adresini bir uzak sunucu ismine çevirmede kullanarak bir istek göndermelidir. İnsanlar alan adlarıyla çalışmayı tercih ettiklerinden ve bilgisayarlar IP adresleriyle çok etkin olduklarından DNS sistemi insan ve bilgisayarlar arasında önemli bir ara yüz sağlamaktadırlar (Kreps, 2015).

### 1.8.3 HTTP İstek Alanı

HTTP İstek Alanı, istemcinin tarayıcısının web sunucusundan talep ettiği bilgiden oluşmaktadır. Tüm HTTP istek alanı tırnak içerisindedir. Aslında bu alan; 1) istek yöntemi, 2) tek biçimli kaynak tanımlayıcı (uniform resource identifier - URI), 3) başlık (header), 4) protokol (protocol) olmak üzere dört alana bölünmüştür. En yaygın istek yöntemi GET'dir. GET, URI aracılığıyla tanımlanarak alınan veri isteğidir. Örneğin, Şekil 1.8'de ilk kayıttaki istek alanı; "GET /Software.html HTTP/1.0,"dır. Software.html web sayfasını sağlayan web sunucu için istemci tarayıcısından gelen bir isteği temsil eder. GET'in yanı sıra HEAD, PUT ve POST istekleri de bulunmaktadır (Markov ve Larose, 2007).

URI, sayfa ya da belge ismini içermekte; dizin yolu istemci tarayıcısıyla istenilmektedir. URI, web kullanım madencileri tarafından sayfa ve dosyalar için ziyaretçi istekleri frekansını analiz etmek için kullanılmaktadır. Başlık sekmesi, tarayıcı isteğine ilişkin bilgiyi içermektedir. Bu bilgi web kullanım madencisi tarafından örneğin arama motorlarından ziyaretçilerin hangi anahtar kelimeleri kullanarak siteye yöneldiğini belirlemek için kullanılabilmektedir. HTTP istek alanı protokol sekmesini de içermektedir. Bu, HTTP'nin hangi sürümünün istemci tarayıcısıyla kullanıldığını belirtmektedir. Örneğin: HTTP/1.1 (Thakare ve Gawali, 2010).

### 1.8.4 Statü Kodu Alanı

Bütün tarayıcı istekleri başarılı değildir. Statü kodu alanı, web sunusundan istemcinin tarayıcısına, isteğin başarılı olup olmadığına dair üç haneli bir cevap sağlamaktadır. "2xx" şeklindeki kodlar başarılı, "4xx" biçimindeki kodlar ise hatayı

göstermektedir. Şekil 1.8'deki kayıtlar için "200" kodu görülmekte olup, bu isteğin başarılı olduğunu göstermektedir (Markov ve Larose, 2007).

Bir web sunucunun mümkün statü kod örnekleri:

*Başarılı Gönderimler (Successful transmission (200 series))*

İstemciden isteğin alınmış, anlaşılmış ve tamamlanmış olduğunu göstermektedir.

_ 200: başarılı (success)

_ 201: oluşturuldu (created)

_ 202: kabul edildi (accepted)

_ 204: içerik yok (no content)

*Tekrar Yönlendiriliyor (Redirection (300 series))*

İstemcinin isteğini tamamlamak için gereken daha uzak hareketi göstermektedir.

_ 301: kalıcı olarak taşındı (moved permanently)

_ 302: geçici olarak taşındı (moved temporarily)

_ 303: değiştirilmemiş (not modified)

_ 304: use cached document

*İstemci Hataları (Client error (400 series))*

Hatalı sözdizimi ya da kayıp bir dosyadan dolayı gerçekleştirilemeyen istemci isteğini gösterir.

_ 400: kötü istek (bad request)

_ 401: yetkisiz (unauthorized)

_ 403: yasaklanmış (forbidden)

_ 404: bulunamadı (not found)

*Sunucu hataları (Server error (500 series))*

Web sunucusunun görünürde geçerli bir isteği yerine getirmede başarısız olmasıdır.

_ 500: iç sunucu hatası (internal server error)

_ 501: gerçekleşmemiş (not implemented)

_ 502: hatalı ağ geçidi (bad gateway)

_ 503: hizmete erişilemiyor (service unavailable)

### 1.8.5 Hacim Transfer Alanı

Hacim transfer alanı (bant genişliği), istemcinin tarayıcısına web sunucusuyla gönderilen bayt türünde dosya (web sayfası, grafik dosyaları, v.b.) boyutunu göstermektedir. Sadece başarılı bir şekilde tamamlanan GET isteklerinin (Statü=200) hacim transfer alanında pozitif bir değeri vardır. Aksi halde alan bir tire (-) ya da sıfır değerini alacaktır. Bu alan network trafiğini izlemek için yardımcı olup, 24 saatlik çevrimlerde ağ ile taşınan yüktür (Markov & Larose, 2007).

## 1.9 Yaygın Günlük Biçimi

Web günlükleri, web sunucusunun konfigürasyonuna bağlı olarak değişik biçimlerde gelmektedir. Yaygın Günlük Biçimi web sunucu uygulamalarının bir çeşidiyle desteklenmekte ve aşağıdaki yedi alanı içermektedir (Markov ve Larose, 2007).

_ Uzak sunucu alanı (remote host field)

_ Tanımlama alanı (identification field)

### 1.9.1 Tanımlama Alanı

Bu alan sadece web sunucu bir kimlik doğrulama gerçekleştiriyorsa, istemci tarafından sağlanan kimlik bilgisini saklamada kullanılmaktadır. Tanımlama alanı bilgisi güvenle şifrelenmiş bir biçim yerine düz metin olduğundan çok seyrek kullanılmaktadır. Bu nedenle genellikle boş değeri temsil eden bir tire içermektedir (Markov ve Larose, 2007).

_ Yetkili alanı (authuser field)

### 1.9.2 Yetkili Alanı

Bu alan, gerekirse yetkilendirilmiş istemci kullanıcı adını saklamaktadır. Yetkili alanı, şifre korumalı klasörlere erişim kazanmayı sağlamak için bir istemcinin

ihtiyaç duyduğu yetkili kullanıcı bilgisini içerecek şekilde tasarlanmıştır. Eğer bu bilgi sağlanmıyorsa alan bir tire olarak varsayılacaktır. (Markov ve Larose, 2007)

_ Tarih/zaman alanı (date/time field)

_ HTTP alanı (HTTP request)

_ Statü kod alanı (status code field)

_ Hacim transfer alanı (transfer volume field )

## 1.10 Genişletilmiş Yaygın Günlük Biçimi

Genişletilmiş yaygın günlük biçimi, yaygın günlük biçiminin bir çeşididir. Referans alanı ve vekil kullanıcı alanı olarak iki ilave alan daha kayda eklenmektedir. İki günlük biçimi de Ulusal Süper Hesaplama Uygulamaları Merkezi (http://www.ncsa.uiuc.edu/) tarafından oluşturulmuştur (Markov ve Larose, 2007).

### 1.10.1 Başvuru Alanı

Başvuru alanı, mevcut sayfayla bağlantılı istemci tarafından daha önce ziyaret edilen site URL'lerini listelemektedir. Görüntüler için başvuru, görüntünün gösterildiği web sayfasıdır. Başvuru alanı kişilerin web sitesini nasıl bulduklarına dair izler taşıdığından pazarlama amaçlı önemli bilgi içermektedir. Bu bilgi yok ise bir tire işareti kullanılmaktadır (Markov ve Larose, 2007).

### 1.10.2 Kullanıcı Vekil Alanı

Kullanıcı vekil alanı istemcinin tarayıcısı, tarayıcı versiyonu ve işlemci sistemi hakkında bilgi sağlamaktadır. Önemli olan bu alan, robotlara ilişkin bilgiyi de içermektedir. Web geliştiriciler bant genişliğini koruma amaçlı robotların web sitesinin belli bölümlerine engeli için bu bilgiyi kullanabilmektedir. Bu alan web kullanım madencisine, site erişiminin bir insan tarafından mı yoksa bir robot tarafından mı yapıldığının belirlenmesinde yardımcı olacaktır. Böylelikle, geliştiricilerin gerçek ziyaretçilerin davranışlarıyla ilgilendiği varsayımıyla, robotların ziyareti analizden çıkarılacaktır (Markov ve Larose, 2007).

### 1.10.3 Bir Web Günlük Kaydı Örneği

Genişletilmiş bir yaygın günlük biçimi örneği aşağıda verilmiştir. Mahremiyet nedeniyle URL kısmı kısmen maskelenmiştir (Markov ve Larose, 2007).

149.1xx.120.116 -- smithj [28/OCT/2004:20:27:32-5000] ``GET /Default.htm HTTP/1.1" 200 1270 ``http:/www.dataminingconsultant.com/" ``Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.0)"

Açıklamaları:

_ Remote host: 149.1xx.120.116
_ Identification: –
_ Authuser: smithj
_ Date/time: [28/OCT/2004:20:27:32 -5000]
_ Request: "GET /Default.htm HTTP/1.1"
_ Status code: 200
_ Transfer volume: 1270
_ Referrer: "http:/www.dataminingconsultant.com/"
_ User agent: "Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.0)"

Şekil 1.9, wpbfl2-45.gate.net kullanıcısı için Şekil 1.8'deki kayıtların bir alt setini içermektedir. İlk kayıt default.htm'yi sağlayan bir istektir. default.htm genellikle bir web sitesinin ana sayfası için kullanılmaktadır. Kayıtlar incelendiğinde sonraki 5 saniyede gerçekleştirilen istekler .gif görüntü dosyalarına aittir. Bu altı görüntü dosyası sunucu üzerinde saklanmaktadır (Markov ve Larose, 2007).

wpbfl2-45.gate.net [29:23:54:15] "GET /default.htm HTTP/1.0" 200 4889
wpbfl2-45.gate.net [29:23:54:16] "GET /icons/circle logo small.gif HTTP/1.0"200 2624
wpbfl2-45.gate.net [29:23:54:18] "GET /logos/small gopher.gif HTTP/1.0" 200 935
wpbfl2-45.gate.net [29:23:54:19] "GET /logos/small ftp.gif HTTP/1.0" 200 124
wpbfl2-45.gate.net [29:23:54:19] "GET /icons/book.gif HTTP/1.0" 200 156
wpbfl2-45.gate.net [29:23:54:19] "GET /logos/us-flag.gif HTTP/1.0" 200 2788
wpbfl2-45.gate.net [29:23:54:20] "GET /icons/ok2-0.gif HTTP/1.0" 200 231

**Şekil 1-9 Tıklamanın çoklu isteğe dönüşümü**

Buradan, bir web sayfasının gerçekte web sunucusuna yapılan isteklere ait nesneler topluluğu olduğu görülmektedir. Gelişmiş web siteleri; web sunucusu, reklam sunucuları, alışveriş sepeti gibi diğer sunuculardan istekte bulunabilmektedir. Bir kullanıcı tek bir web sitesi üzerindeki bir bağlantı üzerine tek bir tıklama yaptığında, bu istek web sunucu üzerinde (istenen her bir doküman ya da nesne için günlük dosyasında bir satır) çoklu istek olarak sonuçlanmaktadır (Markov ve Larose, 2007).

Şekil 1.9'da istek topluluğu tek bir sayfa içindir. Ayrıca Şekil 1.8, web sunucunun neredeyse aynı zamanda, diğer istemcilerden istek alabileceğini göstermektedir. İstekler özel bir istemciyle ya da sırayla kaydedilmemiş özel bir sayfayla ilişkilidir. Web kullanım madencisi ilk önce veri içerisindeki sayfa görüntülemelerini ortaya çıkarmak için veri ön işleme gerçekleştirmelidir (Markov ve Larose, 2007).

## 1.11 Microsoft IIS Günlük Biçimi

Ortak ve genişletilmiş günlük dosya biçimlerinin yanında başka günlük dosya biçimleri de vardır. Bunlardan biri de Microsoft IIS günlük biçimidir ve aşağıdaki alanları içermektedir:

_ Vekil IP adresi (Client IP address)
_ Kullanıcı adı (user name)
_ Tarih (date)
_ Zaman (time)
_ Hizmet (service and instance)
_ Sunucu ismi (server name)
_ Sunucu IP bilgisi (Server IP)
_ Geçen süre (elapsed time)
_ İstemcinin gönderdiği bayt (client bytes sent)
_ Sunucunun gönderdiği bayt (server bytes sent)
_ Statü kodu (service status code)
_ Pencere statü kodu (windows status code)
_ İstek tipi (request type)
_ İşletim sistemi (operating system)
_ Parametreler (parameters)

IIS biçimindeki kayıtlar diğer biçimlerden daha fazla alan içermekte, doğal olarak da daha fazla bilgi ortaya çıkarmaktadır. Web sunucu yöneticileri bu biçimlerden hangisi amacına uyuyorsa onu seçmektedir (Markov ve Larose, 2007).

### 1.11.1 Yardımcı Bilgi

Web günlüklerinin yanı sıra, daha fazla yardımcı bilgi; kullanıcı kayıt bilgisi, kullanıcı demografik bilgisi gibi biçimlerde de bulunabilmektedir. Bu veri genellikle web günlük verisinden ayrı sunuculardaki bulunmaktadır ve ön işleme gerçekleştirilmeden önce web günlükleriyle birleştirilmelidir. Sonuç olarak, yol tamamlama olarak bilinen ön işlemenin gerçekleştirilmesi için, analistin web sitesinin topolojik yapısını, hiyerarşi ağı ve web sayfaları arasındaki ilişkileri bilmeye ihtiyacı vardır (Markov ve Larose, 2007).

## 1.12 Dinamik Tavsiye Sistemleri İçin Kullanıcı Profilleme

Günümüzde web, bilgi değişimi ve dünya çapında yayılım için muazzam bir kanaldır. Kurumlar çoğunlukla şirket hakkında belli bilgilerin bulunduğu, bir çeşit "sanal kartvizit" olarak bu kanalda yer almaktadır (Velasquez ve Palade, 2008).

E-şirketler ziyaretçi profilleri ve satın alım hareketleri hakkında bilgi edinebilmek için web günlük dosyaları kullanma yolları aramaktadırlar (Facca ve Lanzi, 2005). Çoğu şirket/kurum için bir web sitesine sahip olmak ve yüksek kalitede ürün ya da hizmete sahip olmak yeterli olmayacaktır. Bir e-ticaret sitesinin başarı ve başarısızlığının arasındaki fark, web sitesinin ilgi çekme ve ziyaretçilerinin muhafaza edilme potansiyelidir. Bu potansiyel; web site içeriği, tasarımı ve teknik yönler (diğer sitelere kıyasla sayfaların yüklenme süresi gibi) ile sağlanmaktadır (Velasquez ve Palade, 2008).

Web sitesi ziyaretleri temel bir anahtar veri iken, ziyaretçi ile müşteri arasında kesin bir ayrım vardır. Ziyaretçi; bir web sitesine giriş yapan ve kendisi hakkında (ziyaret ettiği sayfa ve site üzerinde harcadığı süre harici) bilgi bırakmadan siteden ayrılan kişidir. Müşteri ise bir kullanıcı kodu/şifre ile web sitesinin herkese açık olmayan bölümüne erişerek, şirket tarafından tanımlanmış yaş, cinsiyet, doğum tarihi gibi bazı demografik bilgiler yardımıyla şirket tarafından tanımlanmış, belirleyiciliği olan kişidir Kullanıcılar ise hem ziyaretçileri, hem müşterileri hem de site yöneticilerini içerir (Velasquez ve Palade, 2008).

Rekabet ortamında şirketler, kullanıcıların aradıkları bilgiye kolayca erişebilecekleri güncel web sitelerine ihtiyaç duymaktadır. Ancak çoğu durumda web sitesinin yapısı (sitenin içinde bir yerde olmasına rağmen) kullanıcıya arzuladığı bilgiyi bulmasında yardımcı olmamaktadır (Velasquez ve Palade, 2008). Web günlük dosyaları kullanılarak sistem performans analizi, sistem tasarımı geliştirme, web trafiğinin doğasını anlama ve kullanıcı tepkilerini anlama üzerine çalışmalar yoğunlaşmıştır. Web sunucular istek yapılan URL, IP adresi ve zaman damgasını her bir erişim için bir web günlük olarak kaydetmektedir. Günlük dosyalarının boyutunun analiz amacıyla azaltılması için başarısız isteklerin ya da görsel istek içeren kayıtların filtrelenmesi gerekmektedir. Web kaynaklarına erişimi değerlendirmede en sık kullanılan yöntem, sayfa erişimlerini ya da istekleri saymaktır. Ancak bu yeterli değildir. Web sunucu günlük dosyaları analiz için yeterli veri barındırmamaktadır. Ancak iyi tasarlanmış veri madenciliği sistemleri ile yararlı bilginin keşfi mümkündür (Zaiane, Xin ve Han, 1998).

Web sunucu günlük dosyaları; isteğin alan adını (ya da IP adresi), isteği oluşturanın kullanıcı ismini (uygulanabilir ise), isteğin tarih ve zamanını, istek yöntemini (POST ya da GET), istek dosyasının adını, istek sonucunu (başarılı, başarısız, hata, vb.), gönderilen verinin boyutunu, referans sayfasının URL'sini ve vekil istemci bilgilerini içermektedir (Zaiane, Xin ve Han, 1998).

Web, alışverişten öğrencilerin ev ödevlerini yapmalarına kadar her şeyi değiştirmektedir. Ana bir web ağı, tedarikçiler ve tüketiciler arasındaki ilişkiyi değiştirmiştir. Geleneksel olarak aktörler arasında ürün ve hizmet dağıtımında aracı zinciri bulunmaktadır. Zincirdeki her bağlantı ürün değerine gereksiz olarak bir üretim maliyeti eklemektedir. Zincir ve adımları dağıtım için gerekli olup tüketiciler ya toptan alacak ya da doğrudan dağıtım için ekstra ödeme yapacaklardır. Web tüketicilerin doğrudan ürün talebinde bulunmalarını sağlamıştır (Velasquez ve Palade, 2008).

Bu değişimler yeni web tabanlı işletmelerin doğmasına neden olmuştur. Yirminci yüzyılın sonunda çoğu web tabanlı şirket değer kazanmıştır. Çoğu şirket web teknolojileri kullanarak potansiyel müşterilere ürün alternatifleri sunabilen karışık satış sistemleri oluşturmuştur (Velasquez ve Palade, 2008).

Pazar payı yeni müşteri kazanımına ve mevcut olanların muhafaza edilmesine bağlıdır (Velasquez v.d., 2013). Bunun için, web şartlarında bir müşterinin satın alım davranışını anlamak şirketler için hayati önem taşımaktadır. Karmaşık olan bu probleme birçok farklı yoldan yaklaşılmaktadır. Elektronik ticaret uzmanları, anahtar meselenin web site içeriği olduğuna işaret etmektedir (Nielsen, 1999).

Elektronik ticaretin işletmeden işletmeye (B2B), işletmeden müşteriye (B2C), denkler arası (P2P) ve bu modellerin farklı tipleri, potansiyel kullanıcıların isteklerini karşılayacak yapı ve içeriği otomatik ayarlayacak gelişmiş kapılara (portal) ihtiyaç duymaktadır.

Başarılı web sitelerinin ana özelliği, kullanıcı için doğru içeriği doğru zamanda sağlayabilme yeteneğidir. Gelişme geçmişi dikkate alınarak, web sitesinin yapısı ve içeriğinde düzeltmeler yapılmalı, hızlı ve etkin bir şekilde aradıklarını bulmaları için kullanıcılara yardımcı olacak ipuçları hazırlanmalıdır (Velasquez ve Palade, 2008).

Web zekâsı (web intelligence), yapay zekâda araştırma konuları seti tasarlamaktadır. Özel bir web zekâsı alanı, bir web sitesinde kullanıcı tercihlerinin daha iyi anlaşılması amaçlı ve web site yapı ve içeriğini daha cazip hale getirici kullanıcı davranışı çalışmalarıdır (Velasquez v.d., 2003). Ana araştırma alanı web kullanım madenciliği olup kullanıcıların web sitesinde gezerken bıraktığı bilgiyi kullanarak web sitelerinin keşif ve analizine yardım etmektedir (Kosala ve Blockeel, 2000). Bu amaç için pek çok algoritma ve sistem önerilmiştir (Eirinaki ve Vazirgannis, 2003).

## 1.13 Web Kullanıcı İlgilerinin Kümelemeyle Modellenmesi

Web kümeleme; web sayfa ya da kullanıcı oturumu gibi benzer web birimlerini birleştirmek için web madenciliğinde sıklıkla kullanılan yöntemlerden biridir. Karşılıklı vektör uzaklıklarını ölçerek nesne grupları oluşturmaktadır (Xu v.d., 2011).

Kümeleme, web kullanıcıları ve web sayfalarını gruplamada kullanılmaktadır. Web kullanıcı gruplarıyla, kullanıcıların navigasyonel davranış örüntüleri ortaya çıkmaktadır. Web sayfalarının kümelenmesiyle de web organizasyonu için görev odaklı işlevsellik üretilmektedir (Xu v.d., 2011).

Web kullanıcılarının karakterlerini keşfetmek web site tasarımcıları için çok önemlidir. Web kullanıcılarının navigasyon yolu sunucuda mevcut ise kullanıcı ilgileri hakkında değerli bilgi taşımaktadır. Kullanıcılar hakkında benzer ilgileri bulmada amaç, kullanıcı profilinden bilgi keşfetmektir. Bir web sitesi iyi tasarlanmış ise gezinti yolları içerisindeki benzerlik ve kullanıcı ilgileri içindeki benzerlik ilişkisi güçlü olacaktır (Xu v.d., 2011).

Bir internet sunucu sitesinin (S) yapısı bağlanabilir çizge olarak adlandırılan bir yönlü çizge olarak özetlenir. Çizgenin düğüm seti sitenin bütün web sayfalarını içermektedir. Sayfalar arasındaki hipermetin bağlantılar, her biri başlangıç ve bitiş sayfasına sahip yönlü çizge kenarları olarak alınmalıdır. Bağlantıların bazıları için başlangıç ya da bitiş noktaları sitenin dışında olabilmektedir. Bağlanabilir çizge karışık olabilmektedir. Bir sitede kullanıcıların gezinti yolu sınırlıdır. İnternet tarayıcı kayıtlarından bir web kullanıcısı hakkında, hiper sayfa kullanımı sıklığı, seçtiği bağlantıların listeleri, iki bağlantı arasında geçen zamanı, bireysel web kullanıcısı olarak eriştiği sayfa sıraları gibi bilgiler edinilebilmektedir (Xu v.d., 2011).

Bir $S$ web sitesi için, belli zaman aralığında, $n$ farklı web sayfasına $P = p_1, p_2, \dots, p_n$ erişen $m$ sayıda kullanıcı $U = \{u_1, u_2, \dots, u_m\}$ olsun. Her $p_i$ sayfası ve her $u_j$ kullanıcısı, $kullanım(p_i, u_j)$ olarak ifade edilen bir kullanım değeriyle ilişkilendirilmiş olup aşağıdaki şekilde tanımlanmaktadır:

$$kullanım(p_i, u_j) = \begin{cases} 1, & u_j \; ile \; erişilen \; p_i \; varsa \\ 0, & diğer \; durumlarda \end{cases}$$

“Kullanım" vektörü sitenin günlük kayıtlarına erişilmesiyle sağlanabilmektedir. İki kullanıcı aynı sayfalara erişir ise, aynı bilgi (haber, ürün vs.) hakkında benzer ilgileri var olabilmektedir. Eriştikleri ortak sayfa sayısı bu benzerlikle ölçülebilmektedir. Bu ölçü aşağıdaki şekilde tanımlanabilmektedir:

$$Benzerlik1(u_i, u_j) = \frac{\sum_k (kullanım(p_k, u_i) * kullanım(p_k, u_j))}{\sqrt{\sum_k kullanım(p_k, u_i) * \sum_k kullanım(p_k, u_j)}}$$

$\sum_k kullanım(p_k, u_i)$, $u_i$ kullanıcının eriştiği toplam sayfa sayısı olmak üzere; $\sum_k kullanım(p_k, u_i) * \sum_k kullanım(p_k, u_j)$, $u_i$ ve $u_j$ kullanıcılarının eriştiği ortak

sayfa sayısıdır. İki kullanıcı aynı sayfaya erişirse, benzerlikleri 1 olacaktır. Bu şekilde tanımlanan benzerlik ölçüsü "kullanım tabanlı ölçü" (usage based measure) olarak adlandırılmaktadır.

Genellikle iki kullanıcı arasındaki benzerlik bütün sitelerdeki ortak sayfalara erişim sayılarıyla hesaplanmaktadır. Buradan ölçü şu şekilde hesaplanmaktadır:

$$Benzerlik2(u_i, u_j) = \frac{\sum_k \sum_s \left( erişim_s(p_k, u_i) * erişim_s(p_k, u_j) \right)}{\sqrt{\sum_k kullanım(p_k, u_i) * \sum_k kullanım(p_k, u_j)}}$$

$erişim_s(p_k, u_i)$, S sitesindeki $P_k$ sayfasına erişen $u_i$ adet kullanıcının toplam sayısıdır. Bu ölçü "sıklık tabanlı ölçü" (frequency based measure) olarak adlandırılmaktadır.

İki kullanıcı arasındaki benzerlik görüntülenen her web sayfasında kullanıcının harcadığı gerçek zaman göz önüne alınarak daha değerli bir şekilde ölçülebilmektedir. $t(p_k, u_j)$, $u_j$ kullanıcısının $p_k$ sayfasını görüntülerken ($u_j$ kullanıcısı $p_k$ sayfasına erişmemişse $t(p_k, u_j) = 0$ olarak kabul edildiği farz edilsin) harcadığı süre olsun. Bu durum için; kullanıcılar arasındaki benzerlik ölçüsü aşağıdaki gibi hesaplanmaktadır:

$$Benzerlik3(u_i, u_j) = \frac{\sum_k (t(p_k, u_i) * (p_k, u_j))}{\sqrt{\sum_k (t(p_k, u_i))^2 * \sum_k (t(p_k, u_j))^2}}$$

$\sum_k (t(p_k, u_i))^2$, web sitesinde görüntülenen sayfalarda $u_i$ kullanıcısının harcadığı sürenin kareleri toplamı; $\sum_k \left( t(p_k, u_i) * t(p_k, u_j) \right)$ ise $u_i$ ve $u_j$ kullanıcılarının ortak görüntülediği sayfalarda harcadıkları süreye ait iç çarpımdır. İki kullanıcı da aynı sayfaya erişir ise benzerlikleri bu durum için 1'den daha az olmaktadır. Bu ölçü "zaman tabanlı görüntüleme ölçüsü" (viewing-time based measure) olarak adlandırılmaktadır.

Bazı uygulamalarda bir kullanıcının sayfalara erişim sırası, her sayfanın görüntülenme süresinden daha önemli olabilmektedir. Bu durumda, iki kullanıcı gerçekte aynı sırada web sayfalarına eriştiklerinde aynı ilgilere sahip oldukları kabul edilmektedir. Kullanıcılar arasındaki benzerlik, bu durum için, gezinti yollarında web

sayfalarının erişim sıralarının kontrolüyle ölçülebilmektedir. $q_i$, $1 \leq i \leq r$, sırayla erişilen sayfa olmak üzere, $Q = q_1, q_2, \dots, q_r$ bir gezinti yolu olsun. $Q$'yu bir r-hop (r-durak) yolu olarak adlandırılmaktadır.

$Q_1$ 'i $Q$ 'nun mümkün tüm $l$ -hop ( $l$ -durak) alt yolları $(1 \leq r)$ olarak tanımlanmaktadır. Örneğin, $Q_1 = \{q_i, q_{i+1}, \dots, q_{i+l-1} | i = 1,2, \dots, r - l + l\}$ gibi. $Q_1$ 'in $Q$ 'daki tüm sayfaları içerdiği açıktır. $f(Q) = \bigcup_{l=1}^{r} Q_l$, $Q$ yolunun özellik uzayıdır. Bir çevrimsel yolun, bir kereden daha fazla alt yollarının çoğunu içerdiğine dikkat edilmelidir ($Q \subseteq f(Q)$).

$Q^i$ ve $Q^j$ sırasıyla $u_i$ ve $u_j$ kullanıcılarının eriştikleri gezinti yolları olsun. $u_i$ ve $u_j$ kullanıcıları arasındaki benzerlik, $\underset{\sim}{Q}^i$ ve $\underset{\sim}{Q}^j$ yolları arasındaki doğal açı ($\cos(\underset{\sim}{Q}^i, \underset{\sim}{Q}^j)$ gibi) kullanılarak tanımlanmaktadır.

$$Benzerlik4 = \frac{\langle \underset{\sim}{Q}^i, \underset{\sim}{Q}^j \rangle_l}{\sqrt{\langle \underset{\sim}{Q}^i, \underset{\sim}{Q}^i \rangle_l \langle \underset{\sim}{Q}^j, \underset{\sim}{Q}^j \rangle_l}}$$

$l = min\left(uzunluk\left(\underset{\sim}{Q}^i\right), uzunluk\left(\underset{\sim}{Q}^j\right)\right)$ ve $\langle \underset{\sim}{Q}^i, \underset{\sim}{Q}^j \rangle_l$, $\underset{\sim}{Q}^i$ ve $\underset{\sim}{Q}^j$ yollarının özellik uzayı üzerinde iç çarpımıdır. $\langle \underset{\sim}{Q}^i, \underset{\sim}{Q}^j \rangle_l = \sum_{k=1}^{l} \sum_{q \in O_k^i \cap O_k^j} uzunluk(q) * uzunluk(q)$.

Yukarıdaki tanımlara göre iki kullanıcı arasındaki benzerlik, sayfalara aynı sırada erişiyorlarsa "1", ortak sayfaları yok ise "0" olacaktır. Bütün $l \leq min\left(uzunluk\left(\underset{\sim}{Q}^i\right), uzunluk\left(\underset{\sim}{Q}^j\right)\right)$ ler için $\langle \underset{\sim}{Q}^i, \underset{\sim}{Q}^j \rangle_l$ 'nin aynı olduğuna dikkat edilmelidir. Bu ölçü "ziyaret sırası tabanlı ölçüdür" (visited-order based).

# BÖLÜM 2

# 2 WEB MADENCİLİĞİNDE ÖĞRENME

Web madenciliğinde çeşitli veri madenciliği araçları sınıflandırma ve kümeleme için kullanılmaktadır. Bu bölümde öncelikle danışmanlı ve danışmansız olmak üzere iki farklı öğrenme stratejisinden bahsedilmiş, ardından danışmanlı öğrenme yöntemlerinden olan ve tezin uygulama kısmında da kullanılan sınıflandırma gücü yüksek destek vektör makinelerinin teorik altyapısı incelenmiştir. İkinci bölümde son olarak, sınıflandırma görevi için performans ölçümlerinin değerlendirilmesi üzerinde durulmuştur.

## 2.1 Öğrenme Stratejileri

Öğrenme "zaman içinde yeni bilgilerin keşfedilmesi yoluyla davranışların iyileştirilmesi süreci" olarak tanımlanabilmektedir (Haykin, 1999: 318). Makine öğrenmesi ise örnek veri ya da geçmiş tecrübeyi kullanarak performans kriterini optimum hale getiren bilgisayar programlamadır. Makine öğrenmesinin temel görevi bir örneklemden çıkarım yapılması olduğundan, matematiksel modellerin kurulmasında istatistik teorisi kullanılmaktadır (Alpaydın, 2010: 3).

Örneklerden öğrenme stratejileri danışmansız ve danışmanlı öğrenme olarak iki temel başlıkta incelenmektedir. Öğrenmeyi gerçekleştirecek olan sistem ve kullanılan öğrenme algoritması bu stratejilere bağlı olarak değişmektedir (Öztemel, 2003: 24).

### 2.1.1 Danışmansız (Denetimsiz) Öğrenme

Bazı uygulamalarda veri, sınıf etiketlerine sahip değildir. Bu durumda veri içerisindeki yapının keşfedilmesi gerekmektedir. Danışmansız öğrenmede sisteme sadece girdi değerleri gösterilmektedir. Ve örneklerdeki parametreler arasındaki ilişkileri sistemin kendi kendine öğrenmesi beklenmektedir. Ancak doğru çıktıların ne olduğu hakkında bilgi verilmemektedir (Öztemel, 2003). Bu öğrenme türü etiketlenmemiş eğitim verilerinde kurallar bulmaya, küme etiketlerini ve bazen de

kümes sayılarını çıkarmaya çalışmaktadır. Kümeleme, birliktelik kuralları, sıralı örüntüler danışmansız öğrenme yöntemlerinden bazılarıdır (Liu, 2007: 117).

Kümeleme analizi en çok kullanılan danışmansız öğrenme yöntemi olup veri madenciliğinde veri segmentasyonu (data segmentation), makine öğrenmesinde ise sınıf keşfi (class discovery) olarak da adlandırabilmektedir (Izenman, 2008: 408). Kümeleme, bir veri setinin k adet grup ya da kümeye ayrılma sürecidir. Küme içerisindeki elemanlar analizde dâhil edilen özelliklere göre aldıkları değerler açısından yüksek benzerliğe, diğer kümelerdeki elemanlardan ise yüksek farklılığa sahip olmalıdır. Bir başka ifadeyle kümeler kendi içinde homojen, birbirleri arasında heterojen bir yapıdadır (Han ve Kamber, 2012: 443), (Parabhu ve Venkatesan, 2007: 35).

Birliktelik kuralları, x koşulunda y niteliğinin, bir $p(y|x)$ koşullu olasılığı biçiminde öğrenmedir. Y, daha önce satın alındığı bilinen ürün ya da ürün grubu olabilmektedir (Alpaydın, 2010: 4). Birliktelik kurallarının özel bir uygulaması olan pazar sepet analizi sıklıkla kullanılmakta, müşterilerin satın alma davranışlarını anlamak için bu analizden yararlanılmaktadır. Birliktelik kuralları ayrıca bilişim (ağ saldırı tespiti) ve biyoenformatik gibi alanlara da uygulanmaktadır.

Bir diğer danışmansız öğrenme yöntemi de sıralı örüntülerdir. Sıralı örüntü madenciliği (sequential pattern mining) sıralı bir veri tabanında ardıl örüntüleri keşfetmektedir. Sıralı veri tabanı, zamana bağlı ya da zamandan bağımsız kaydedilmiş olaylar dizisini barındırmaktadır. Müşteri işlem kayıtları (her müşteri için bir ay süresince her hafta için satın alınan ürün bilgisi) bu veri tabanına örnek olarak verilebilmektedir.

### 2.1.2 Danışmanlı (Denetimli) Öğrenme

Danışmanlı öğrenmede tüm eğitim verisi etiketlenmiştir. Bu etiketlenme işlemi genellikle dış bir mekanizma (genellikle insan) tarafından yapılmaktadır ve bu nedenle danışmanlı öğrenme olarak adlandırılmaktadır. Eğitim kümesinin girdi-çıktı çiftlerinden (eğitim örneği) oluşturulması danışmanlı öğrenme örneğidir. Danışmanlı terimi, danışmanın girdi vektörlerine $p(y|x)$ dağılımına göre bir çıktı değeri ataması

düşüncesinden gelmektedir (Vapnik, 1998). Bir başka ifadeyle her örnek için hem girdiler hem de o girdiler karşılığında oluşturulması gereken çıktılar sisteme gösterilmektedir. Sistemin görevi girdileri danışmanın belirlediği çıktılara haritalamaktır. Böylece durumların girdileri ile çıktıları arasındaki ilişkiler öğrenilmektedir (Öztemel, 2003: 24).

Birden fazla danışmanlı öğrenme görevi bulunmaktadır. Çıktı alanının sürekli değerler alması durumunda bir regresyon görevi, kesikli değerler alması durumunda ise bir sınıflandırma görevi gerçekleştirilir. Bu tez çalışmasında ele alınan problem bir ikili sınıflandırma görevi olarak kurgulanmıştır. Öğrenmede kullanılan veri kayıtları setinde, $|A|$; $A = \{A_1, A_2, \dots, A_{|A|}\}$ nitelikler seti olarak tanımlanmaktadır. Veri seti, sınıf niteliği olarak adlandırılan $C$ özel hedef niteliğine sahiptir. $C$'nin $A$'da olmadığı, özel niteliklerine göre $C$'nin, $A$'daki niteliklerden ayrı olduğu varsayılmaktadır. Sınıf niteliği $C$, $|C|$ sınıf sayısı ve $|C| \geq 2$ olmak üzere, $C = \{c_1, c_2, \dots, c_{|C|}\}$ kesikli değerler almaktadır. Sınıf değeri ayrıca sınıf etiketi (class label) olarak da adlandırılmaktadır. Öğrenme için bir veri seti basitçe bir ilişkisel tablodur. Her veri kaydı bir parça "geçmiş deneyimi" tanımlamaktadır. Makine öğrenmesinde ve veri madenciliği literatüründe bir veri kaydı; örnek (example), olay (instance), durum (case) ya da vektör (vector) olarak da isimlendirilebilmektedir. Bir veri seti basitçe örnekler ya da olaylar setidir (Liu, 2007).

Bir $D$ veri seti için öğrenmenin amacı, $A$'daki nitelik ve $C$'deki sınıf değerleriyle ilişkili bir sınıflandırma fonksiyonu üretmektir. Bu fonksiyon, gelecek verisinin sınıf değerleri/etiketlerini tahmin etmede de kullanılabilmektedir ve sınıflandırma modeli, öngörü modeli ya da sınıflandırıcı olarak adlandırılmaktadır. Fonksiyon (model), karar ağacı, kurallar seti, bayescil model ya da hiperdüzlem olabilmektedir (Liu, 2007).

Danışmanlı öğrenme gerçek dünya uygulamalarında büyük bir başarıya sahiptir. Metin ve web madenciliği dâhil olmak üzere neredeyse her alanda kullanılmaktadır. Danışmanlı öğrenme aynı zamanda "tümevarımsal öğrenme (inductive learning)" olarak da isimlendirilmektedir. Bu tip öğrenmede geçmiş tecrübelerden kazanılan

(geçmişte toplanılan veriden) bilgiyle gerçek dünya uygulamalarına yönelik yetenek geliştirilmektedir (Bramer, 2013).

Öğrenme için mevcut veri genellikle eğitim seti (training set) ve test seti (test set) olmak üzere iki ayrık alt kümeye ayrılmaktadır. Bir öğrenme algoritmasıyla eğitim verisinden bir model öğrenildikten ya da oluşturulduktan sonra, model doğruluğunu değerlendirmek için test verisi ya da çıktı değeri bilinmeyen veri kullanılmaktadır (Alpaydın, 2009).

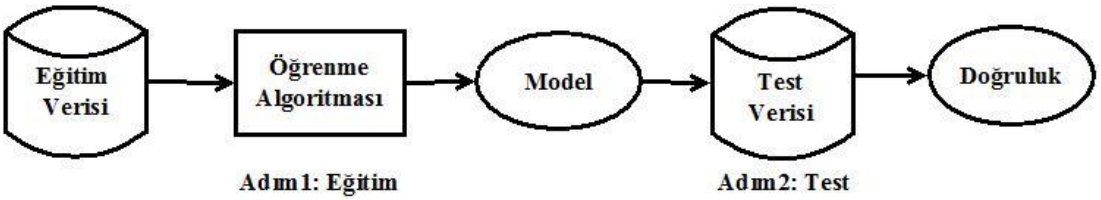
Test verisinin sınıflandırma modelini öğrenmede kullanılmadığına dikkat edilmelidir. Test verisindeki örnekler de etiketlere sahiptir. Her test örneği için, model aracılığıyla tahmin edilen sınıfın gerçek sınıfla aynı olup olmadığı karşılaştırılabildiğinden, test verisi öğrenme modelinin doğruluğunu değerlendirmede kullanılabilmektedir (Liu, 2007).

Bir test seti üzerindeki sınıflandırma modelinin doğruluğu ise;

$$Doğruluk = \frac{Doğru\ sınıflamaların\ sayısı}{Test\ durumlarının\ toplam\ sayısı}$$

olarak hesaplanabilmektedir. Doğru sınıflandırma, öğrenilmiş modelin test durumundaki orijinal sınıf ile aynı sınıfı tahmin etmesi anlamına gelmektedir. Sınıflandırma modelinin doğruluğu için başka ölçüler de kullanılabilmektedir.

Eğitim seti ile test seti arasındaki ilişkinin nasıl olması gerektiği makine öğrenmesinin ana varsayımıyla cevaplanabilmektedir. Bu varsayım, eğitim örneklerinin dağılımının test örneklerinin dağılımıyla benzer olduğunu belirtir. Uygulamalarda bu varsayım sıklıkla belli bir derecede ihlal edilmektedir. Test verisi eğitim verisinden çok farklı davranıyor ise güçlü ihlaller zayıf sınıflandırma doğruluğu olarak sonuçlanacak, öğrenmiş model test verisi üzerinde iyi performans gösteremeyecektir. Test verisi üzerinde yüksek doğruluk oranı elde edebilmek için eğitim verisi test verisini etkin şekilde temsil edebilmelidir (Liu, 2007).




**Şekil 2-1 Temel Öğrenme Süreci: Eğitim ve Test (Liu, 2007:58)**

Şekil 2.1'de gösterildiği üzere birinci adımda öğrenme algoritması, bir sınıflandırma modelini üretmede eğitim verisini kullanmaktadır. Bu adım "eğitim aşaması" ya da "eğitim adımı" olarak adlandırılmaktadır. İkinci adımda, öğrenmiş model test verisini kullanarak sınıflandırma doğruluğunu elde etmek için test edilmektedir. Bu adım "test adımı" ya da "test aşaması" olarak adlandırılmaktadır. Öğrenen modelin test verisi üzerindeki doğruluğu tatmin ediciyse, model gerçek dünya problemlerinde yeni durumların sınıflarını tahmin etmede kullanılabilir. Eğer modelin doğruluğu tatmin edici değilse geriye dönüp farklı bir öğrenme algoritması seçmek ve/veya veriyi işlemek gereklidir. Bir pratik öğrenme görevi tatmin edici bir model oluşturulmadan önce bu adımların kaç defa yineleneceğidir. Ayrıca verideki yüksek rastlantısallık derecesi ya da mevcut öğrenme algoritmalarının sınırlarından dolayı tatmin edici bir model kurulamaması da mümkündür (Liu, 2007).

Metin ve web uygulamaları gibi bazı öğrenme problemlerinin, eğitim ve test verisi üzerinde bir takım dönüşümler yapılarak, öğrenmeye hazır hale getirilmesi gerekmektedir. Söz konusu uygulamalardaki biçimlerin ve niteliklerin açık olmamasından dolayı, genellikle ham veriyi toplayıp nitelikleri tasarlamak ve ham veriden nitelik değerlerini hesaplamak gerekmektedir (Liu, 2007).

En sık kullanılan danışmanlı öğrenme yöntemleri olarak yapay sinir ağları, karar ağaçları, bayescil öğrenme, destek vektör makineleri sayılabilmektedir. Yapay sinir ağları, insan nöronlarının modeline dayalı karışık bir modellemedir. Bir sinir ağı, verilen bir girdi setiyle bir ya da daha fazla çıktı tahmin etmektedir (Bramer, 2007: 6). Kolay yorumu ve anlaşılabilirliği açısından karar vericiler için avantaj sağlayan karar ağaçlarında ise öğrenilen model, sınıflandırma kuralları veya karar ağacı olarak gösterilir. Bağımsız özellikler setini böl ve elde et (divide-and-conquer) algoritmasıyla sınıflara ayırmaktadır (Witten ve Frank, 2005: 62). Amaç, ağaç oluşturulurken

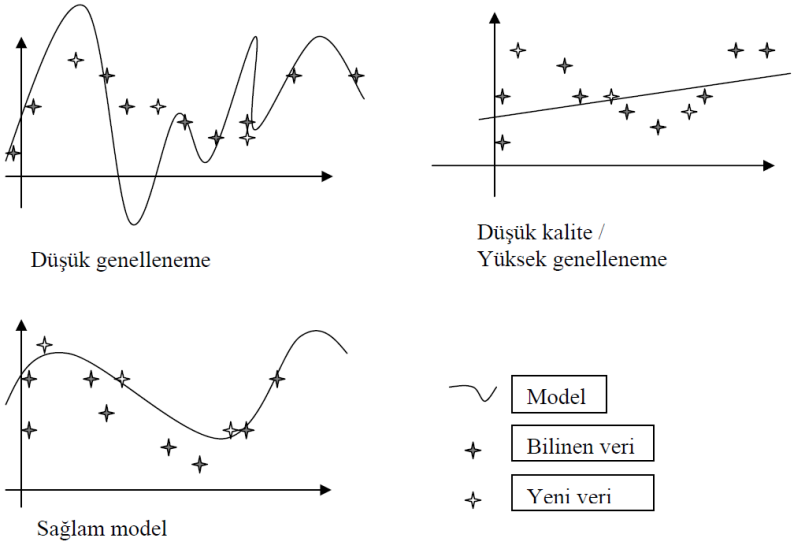
mümkün olduğunca birbirine benzeyen veriler gruplanırken ağaç derinliğinin minimum seviyede tutulmasıdır. Böylece karmaşıklık (entropi) minimuma indirilirken maksimum seviyede kazanç (elde edilen bilgi) elde edilir. Bayescil yöntemler mümkün kümeleme dağılımı hesaplamaktadır (Han ve Kamber, 2012, 459). Destek vektör makineleri bu tez çalışmasının uygulama bölümünde de kullanılan bir yöntem olduğundan ayrı bir başlıkta incelenmiştir.

## 2.2 Destek Vektör Makineleri

Temelleri ilk olarak Vladimir Vapnik ve Alexey Chervonenkis tarafından, hesaplanabilir öğrenme teorisinin önemli parçasını oluşturan ve öğrenmenin temel teorisi olarak bilinen Vapnik-Chervonenkis Teorisi kapsamında, 1960'lı yıllarda atılan destek vektör makineleri (DVM); 1992 yılında Vladimir Vapnik, Berhard Boser ve Isabelle Guyon tarafından sunulmuştur (Boser, Guyon ve Vapnik, 1992). Diğer sınıflandırma yöntemleri ile karşılaştırıldığında, eğitim süresi oldukça uzun olmasına rağmen, yüksek güvenirliği, ezbere öğrenmeye olan dayanıklılığı ve doğrusal olmayan sınıflandırmadaki başarı düzeyleri ile DVM tercih edilen bir yöntem olmuştur. DVM karar doğrusuna bağımlı olarak belirlenen destek noktaları (support vectors) arasındaki genişliği (margin) maksimize etmeyi amaçlayan danışmanlı bir öğrenme algoritmasıdır (Akpınar, 2014: 268).

Öğrenen makinenin genelleme kabiliyeti, öğrenen makineyle gerçeklenmiş fonksiyonlar setinin kapasitesine bağlıdır. Genelleme hatası tüm veri kümesi için gerçek risk ile eğitim örnek kütleleri için ampirik risk arasındaki fark olup örnek kütle sayısı arttıkça ve ağın kapasitesi azaldıkça azalmaktadır (Akaho, 1993: 493). Bu zayıf genelleme durumu aşırı uyumdan kaynaklanmaktadır. İstatistikte aşırı uyum çok fazla parametreye sahip istatistiksel modeli uygun hale getirmektedir. Garip ve yanlış bir model eldeki veri büyüklüğüyle karşılaştırıldığında yeterli karmaşıklığa sahipse mükemmel uyum gösterebilir. Buna aşırı uyum (overfitting) sorunu adı verilmektedir ve şu şekilde tanımlanmaktadır: Bir $h$ hipotezi eğitim verisine, eğer eğitim verisi üzerinde $h$'nin $h'$'den daha küçük hataya sahip olduğu bir başka $h'$ hipotezi varsa, aşırı uyum sağlar. Ancak $h'$, test verisi üzerinde $h$'den daha küçük hataya sahiptir. Bu problem tüm öğrenme algoritmaları için geneldir (Wang ve Zhang, 2003).

Aşırı uyum kavramı makine öğrenmesinde önemlidir. Eğitim süresince durumlara çıkarımsal ikileme dayanarak makinenin diğer örnekler için doğru çıktıyı öngörebilecek bir duruma ulaştığı varsayılmaktadır. Ancak makine hedef çıktıyla ilişkisi bulunmayan, eğitim verisinin nadir rassal özellikleri üzerine eğitimi düzenleyebilmektedir. Bu aşırı uyum sürecinde eğitim verisi üzerindeki performans artarken test verisi üzerindeki performans kötüleşmektedir (Şekil 2.2).




**Şekil 2-2 Modelin Genellemesi (Tolun, 2008: 61)**

Destek vektör makineleri ile öğrenen makinelerin genelleme kabiliyetini kontrol etmek için (test hatası üzerinde en küçük sınırı elde etmek için) eğitim hatalarının sayısı minimize edilerek en küçük VC boyutlu fonksiyonlar setinin kullanılması fikri doğmuştur (Vapnik, 1998: 10).

Eğitim hata sayısını minimize etmek için küçük VC boyutlu dar bir fonksiyonlar setinden ziyade geniş bir fonksiyonlar setinden seçim yapılmalıdır. Buradan da en garanti çözümü bulmak, eğitim verisinin yaklaşık doğruluğu ile hata sayısını minimize etmede kullanılan makinenin kapasitesi (VC boyutu) arasında bir uzlaşma sağlamaktır. Birbiriyle çelişen iki karşıt faktörün kontrol edilerek test hatasının minimize edilmesi düşüncesi bir tümevarım prensibi olan yapısal risk minimizasyonu (structural risk minimization) kavramını ortaya çıkarmıştır. Tümevarımsal çıkarımdaki uzlaşmanın varlığı düşüncesi, Occam'ın 14.yy'da önerdiği

Occam'ın usturası olarak bilinen *"nesnelerin sayısı zorunlu değilse çoğaltılmamalı"* prensibine dayanmaktadır. *"En basit açıklama en iyisidir"*, Occam'ın usturasının en yaygın kullanılan açıklamasıdır. Yapısal risk minimizasyonu ise: *"en küçük kapasiteli (VC boyutlu) makineyle açıklama en iyisidir"* iddiasındadır (Vapnik, 1998: 11).

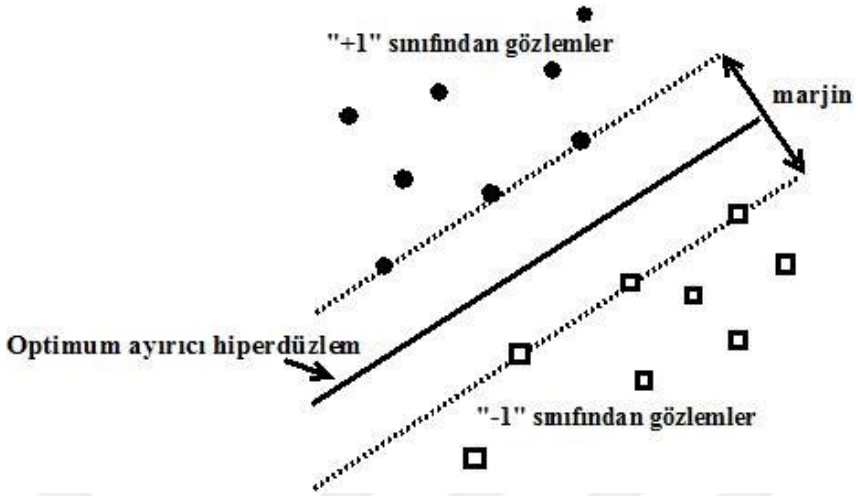
İkili sınıflandırma için genelleme sınırı;

$$R(\alpha_n) \leq R_{amp}(\alpha_n) + \Omega(\frac{h}{l}, \frac{ln\eta}{l})$$

şeklinde yazılabilmektedir. Eşitsizliğin sağ tarafındaki ilk terim $R_{amp}(\alpha_n)$ ampirik risk, ikinci terim ise gerçek risk VC boyutu $h$ (öğrenen makinenin karmaşıklığı) ve eğitim örneği sayısı $l$ ile ilişkisi olan güven aralığı olmak üzere gerçek riskin sınırı iki bölümden oluşmaktadır.

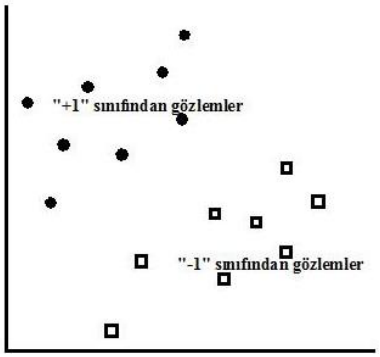
### 2.2.1 Marjlar ve Uzaklıklar (Doğrusal Ayrılabilir Durum)

Destek vektör makinelerinin ardında yatan düşünce, verileri optimal olarak iki sınıfa ayıran bir sınıflandırıcı oluşturulması problemidir (Clarke, Fokoué ve Zhang, 2009). İyi bir ayırıcı düzlem düşüncesi, veriden çok uzakta geniş bir marj (maksimal marj) sınıflandırıcı kavramında şekillenmektedir. Marj, iki veri bulutunu ayıran boş şerit genişliğini ifade etmektedir. Şekil 2.3'te iki sınıfın mükemmel bir şekilde ayrıldığı bir örnek verilmiştir. Bu veri seti için bile gözlemleri ayırabilen sonsuz sayıda hiperdüzlem bulunmaktadır. Ancak istenen, gelecekteki gözlemleri de doğru sınıflandıracak bir sınırın seçilmesidir. Şekil 2.4'teki iki sınıfı ayıran sınırın nasıl seçileceğine dair makul tercihler Şekil 2.5'te gösterilmiştir. Şekil 2.5d'deki sınır en geniş marjı göstermekte olup, tercih edilmektedir (Clarke, Fokoué ve Zhang, 2009).

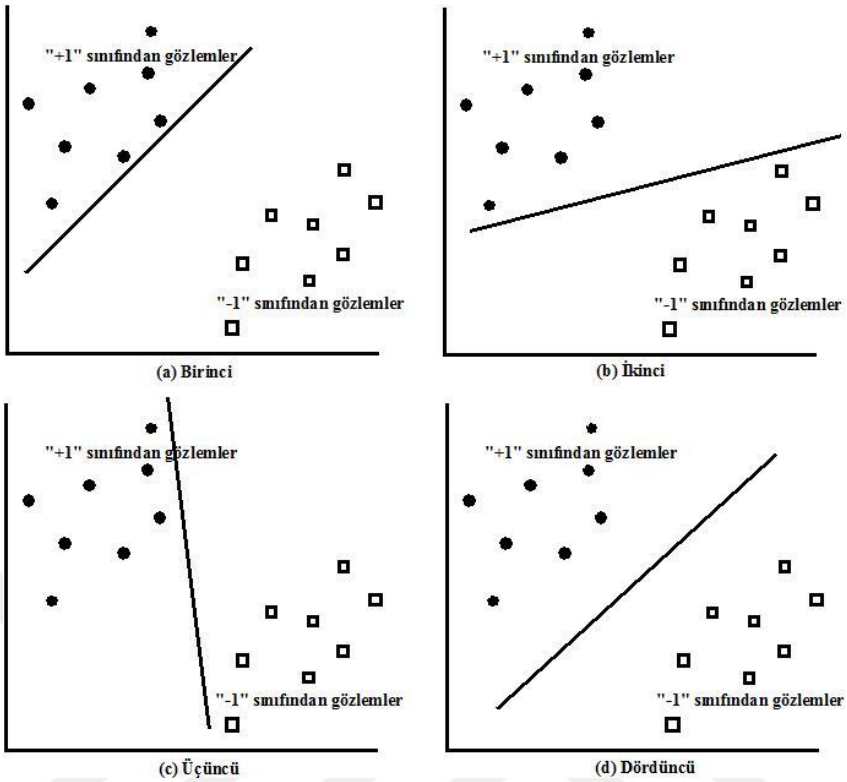



**Şekil 2-3 İdeal durum. İki veri bulutundaki noktalar arasındaki minimum (dik) uzaklık marjdır.**

Marj, hiperdüzlem ve gözlemler arasındaki minimum (dik) uzaklık olarak tanımlanmaktadır. Amaç, en küçük (minimal) uzaklık kriterini sağlayan düzlemler arasındaki marjın maksimizasyonudur (minimum uzaklıkların maksimumunun bulunması problemi).




**Şekil 2-4 İki sınıf için x-değerlerini gösteren (+1 ve -1 olarak etiketlenen) içi dolu daireler ve boş kareler. İki sınıfın doğrusal olarak ayrılabildiği durum.**




**Şekil 2-5 (a)'da sınır sadece bir sınıfın etrafında bir yol çizmektedir. (b) ve (c)'de sınır, veri bulutunun temsilinde olmayan noktalar tarafından etkilenmektedir. (d)'de ise sınır, veri bulutları arasında marjın ortasında yer almaktadır.**

Şekil 2.5'teki kesikli çizgiler en yakın iki hiperdüzlemi göstermektedir. Veri noktaları en küçük (dik) uzaklığı gerçeklemektedir. Marj, dıştaki iki hiperdüzlem arasındaki uzaklıktır. Düz çizgi optimum ayırıcı hiperdüzlemi (geniş marj sınıflandırıcı) göstermektedir (Clarke, Fokoué ve Zhang, 2009).

Merkezi marjın konumunu biçimlemek için dört işlem bulunmaktadır:

1) a noktası ve ayırıcı hiperdüzlem arasındaki uzaklığı hesaplamak

2) verilen gözlemler setinin minimum uzaklığını sağlamak

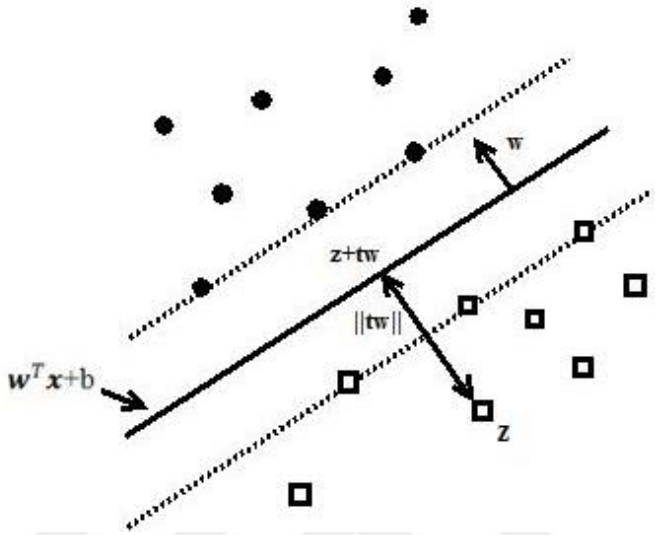
3) iki paralel hiperdüzlem arasındaki uzaklığı hesaplamak

4) ayırıcı hiperdüzlemden eşit uzaklıktaki iki hiperdüzlemi sağlamak

Bu adımlar geometrik olarak sezgisel görünse de bazı tanımlamalar gerektirmektedir.

$w = \left(w_1, w_2, \ldots, w_p\right)^T \in \mathbb{R}^p$ bir vektör katsayısı ve $b \in \mathbb{R}$ bir sabit olsun.

$h$: $\mathbb{R}^p \to \mathbb{R}$ olmak üzere $h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$+b doğrusal fonksiyonu yazılır.




**Şekil 2-6 z noktası w yönünde çizgiyle gösterilen hiperdüzlemden t birim uzaklıktadır.**

Verilen bir $c \in \mathbb{R}$ sabiti için,

$$H_c(\boldsymbol{w}, b) = \{\boldsymbol{x}: h(\boldsymbol{x}) = c\}$$

bir $(p-1)$ boyutlu hiperdüzlemdir. Eğer $c = 0$ ise, $H_{c=0}(\boldsymbol{w}, b)$, $H(\boldsymbol{w}, b)$'dir.

Bir vektör hiperdüzleme paralel ise hiperdüzlemin yönlü vektörü, hiperdüzlemin tüm mümkün yönlü vektörlerine dik ise hiperdüzleme normal vektördür. Açıkça $\boldsymbol{w}$ vektörü herhangi $c$ için $H_c(\boldsymbol{w}, b)$'dir. Aslında $\forall \boldsymbol{x}_i, \boldsymbol{x}_j \epsilon H_c(\boldsymbol{w}, b)$ için $\boldsymbol{w}^T\boldsymbol{x}_i + b = c = \boldsymbol{w}'\boldsymbol{x}_j + b$ ve buradan $\boldsymbol{w}^T\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right) = 0$ dır.

DVM sınıflandırıcı formülasyonu, $\mathbb{R}^p$ 'de bir nokta ile bir hiperdüzlem arasındaki dikey uzaklığın ifadesini gerektirmektedir.

**Teorem:** $H_c(\boldsymbol{w}, b)$ hiperdüzlemi ve $z \epsilon \mathbb{R}^p$ olmak üzere bir nokta arasındaki dikey uzunluk $d\left(z, H_c(\boldsymbol{w}, b)\right)$ olmak üzere;

$$d\big(z, H_c(\boldsymbol{w}, b)\big) = \frac{|\boldsymbol{w}^T z + b - c|}{\|\boldsymbol{w}\|} \qquad (1)$$

dir.

**Teorem:** $H_c(\boldsymbol{w}, b)$ ve $H_{c'}(\boldsymbol{w}, b)$ iki paralel hiperdüzlemi arasındaki dikey uzaklık,

$$d\left(H_c(\boldsymbol{w}, b), H_{c'}(\boldsymbol{w}, b)\right) = \frac{|c - c'|}{\|\boldsymbol{w}\|} \qquad (2)$$

dir.

### 2.2.2 İkili Sınıflandırma ve Risk

İkili sınıflandırmanın genel problemi keşfedici değişken değerlerini temsil eden $\mathcal{X} \subseteq \mathbb{R}^p$ girdi alanı ve etiket olarak davranan $\mathcal{Y} = \{-1, +1\}$ sınıf alanı ile başlamaktadır. Y değerlerinin, Y= 1, 2 değil +1 ya da -1 olduğuna dikkat edilmelidir. $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ bir veri seti olsun. Doğrusal ayrılabilir durumdaki ikili sınıflandırmada hedef, $p$ boyutlu vektör $w = (w_1, w_2, \dots, w_p)^T$'yi tahmin eden veriyi kullanmaktır.

$b$ sabitiyle $H(\boldsymbol{w}, b) = \{\boldsymbol{x}: \boldsymbol{w}^T \boldsymbol{x} + b = 0\}$ hiperdüzlemi $\boldsymbol{x}_i$ gözlemlerini -1 ve +1 sınıf etiketlerine ayırmaktadır. Hedef; bir $f: \mathcal{X} \to \{-1, +1\}$ fonksiyonunu bulmaktır. Öyle ki;

$\boldsymbol{x}_i \epsilon\, \mathcal{D}$ için $sınıf(\boldsymbol{x}_i) = f(\boldsymbol{x}_i) = sign(\boldsymbol{w}^t \boldsymbol{x}_i + b) = \begin{cases} +1, \boldsymbol{w}^T \boldsymbol{x}_i + b > 0, \\ -1, \boldsymbol{w}^T \boldsymbol{x}_i + b \leq 0, \end{cases}$ dir.

Bir DVM sınıflandırıcı beş adımlı bir düşünceye dayanır:

1) Seçilen karar sınırından sınıf fonksiyonları tanımlamak (x doğrusal fonksiyonları)

2) Bir aday karar sınırı ile her sınıftaki noktalar arasındaki minimal uzaklık olan marjı tanımlamak

3) Birinci adımda tanımlanan sınıftan karar sınırını seçmek

4) Eğitim setinde seçilmiş karar sınırının performansını değerlendirmek

5) Yeni veri noktaları üzerinde beklenen sınıflandırma performansını değerlendirmek (genelleştirme hatası)

$\mathbb{R}^p$ deki hiperdüzlemlerden oluşturulan $\mathcal{F}$ sınıflarını göz önüne alarak,

$$\mathcal{F} = \{f: \mathbb{R}^p \to \{-1, +1\}, s.t. \forall \boldsymbol{x} \in \mathcal{X}, f(\boldsymbol{x}) = sign(h(x)), h, \mathbb{R}^p de\ hiperdüzlemdir\}$$

Bir değerlendirme hatası olarak, sıklıkla 0-1 kaybı kullanılmaktadır. $\boldsymbol{x}$'in gerçek sınıfı $y$ ve sınıflandırıcı $f(\boldsymbol{x})$, $y$'den farklı ise yanlış sınıflandırma hatası oluşur ve kayıp fonksiyonu ile gösterilir.

$$\ell(y, f(\boldsymbol{x})) = \boldsymbol{I}(f(\boldsymbol{x}) \neq \boldsymbol{y}) \qquad (3)$$

Sınıflandırıcı doğru ise kayıp 0, yanlış ise 1'dir. Makine öğrenmesinde sınıflandırıcı $f$ sıklıkla hipotez, $f$'lerin dahil olduğu $\mathcal{F}$ sınıfı da hipotez sınıfı olarak isimlendirilmektedir. $f^*$ genelleme hatasını göstermekte olup aynı zamanda öngörü hatası (prediction error) ya da yapısal risk olarak da adlandırılmaktadır.

$X \epsilon \mathcal{X} \subseteq \mathbb{R}^p$ ve $Y \epsilon \mathcal{Y} = \{-1, +1\}$ ve $\mathcal{X}x\mathcal{Y}$ üzerinde genellikle bilinmeyen bir $\mathbb{P}(X, Y)$ olasılığı olmak üzere, $\mathcal{P}$ iki değişkenli bir popülasyon olsun.

Bu $f$, $\mathcal{X}$ üzerinde tanımlanan bir sınıflandırıcı (hipotez) ise, $f$'in genelleştirme hatası;

$$R(f) = \mathbb{E}[\ell(Y, f(\boldsymbol{X}))] = \int_{\mathcal{X}x\mathcal{Y}} \ell(y, f(\boldsymbol{x}))d\mathbb{P}(\boldsymbol{x}, y)\text{'dır.}$$

Bu risk, $f$'in $\mathcal{X}$'den tesadüfi olarak seçilen bir üyesini yanlış sınıflandırma olasılığıdır. En iyi $f$'i bulmak;

$$f^* = \arg\inf_f R(f)$$

ile gerçekleşecek olsa da, $f^*$'ı bulmak, $R(f)$ için kapalı formları (closed-form expressions) türetme kabiliyeti ve $\{-1, +1\}^{card(\mathcal{X})}$ fonksiyonlar uzayını arayan bir yol varsaymaktadır. Mümkün hipotezlerin sınıfını kısıtlamak için bazı öncül bilgileri kullanmak yardımcı olacaktır. Örneğin;

$$f^{\circ} = \arg \inf_{f \epsilon \mathcal{F}} R(f)\text{'yi}$$

arayan $\mathcal{F}$ sınıfı üzerinde minimizasyon yapılmıştır.

0-1 kaybı altında risk, $R(f)$, tesadüfi örnekte popülasyonun küçük bir kesiti var olduğundan gerçekte mevcut değildir. Bununla birlikte $R(f)$ yanlış sınıflandırmanın olasılığı da olduğundan, $n$ büyüklüğündeki örnekten hesaplanan deneysel (ampirik) risk ile ya da eğitim hatası (training error) $\hat{R}_n(f)$, ile de tahmin edilebilmektedir. Bir $\mathcal{H}$ hipotez uzayı sınıflandırıcıları için, $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_n, y_n)\}$ bir veri seti ve $f \epsilon \mathcal{H}$ olmak üzere,

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{I}(f(\boldsymbol{x}_i) \neq y_i) \qquad (4)$$

dir.

$\hat{R}_n(f)$, $f$ ile yanlış sınıflandırılmış noktaların deneysel kısmıdır. (4)'ü minimize etmek, deneysel risk fonksiyonunu minimize eden $\hat{f}$ sınıflandırıcısını bulmak anlamına gelmektedir.

$$\hat{f} = \arg\min_{f \epsilon \mathcal{F}} \hat{R}_n(f) \qquad (5)$$

(4) ve (5) sadece 0-1 kaybı için tanımlanmış olsalar dahi herhangi bir kayıp fonksiyonu için genelleştirilebilmektedir.

Görünürde (5) daha makul gözükmektedir. Bununla birlikte $\hat{f}$, $\hat{R}_n(\hat{f})$'i mümkün olduğu kadar küçük yapacak şekilde seçilmektedir. $\hat{R}_n(\hat{f})$, $R(\hat{f}_{gerçek})$ gerçek hata oranını (true error rate) azımsama eğilimindedir. Azımsama (underestimate) derecesini ölçmek, $\hat{f}$ ile ilişkili gerçek risk $R(\hat{f})$, $R(f^0) = \min_{f \in \mathcal{F}} R(f)$'e mümkün olduğu kadar yakın, $\mathcal{F}$'de kazanılan en düşük risk, ve olabildiği kadar küçük olan $\hat{R}_n(\hat{f})$'ı sağlama arasında bir dengelemeye (trade-off) ihtiyaç duymaktadır.

Bir başka ifadeyle yazılacak olursa deneysel risk, $\hat{R}_n(f)$, gerçek riske $R(f)$ yakınsamalıdır. Buradan $f \epsilon \mathcal{F}$ üzerinde düzgün olarak $\delta$ ve $\varepsilon$ önceden atanmalıdır.

$$\mathbb{P}\left(\left|\hat{R}_n(f) - R(f)\right| < \varepsilon\right) \geq 1 - \delta, \quad \forall f \in \mathcal{F} \qquad (6)$$

Temelde (6) genelleştirilmiş varyans-taraflı dengelemedir (generalized variance-bias trade-off). Diğer bir ifadeyle, aranılan sıfır deneysel risk değil, iki düşük tahmin hatasının kombinasyonu olan küçük bir yaklaşık hataya karşılık gelen en küçük deneysel risktir.

### 2.2.3 Doğrusal Olmayan Sınıflandırma

Doğrusal olmayan sınıflandırma destek vektör makineleriyle yumuşak marj (aylak değişken) ve kernel hilesi (kernel trick) yöntemi olmak üzere iki şekilde ele alınabilmektedir. Maksimal marj sınıflandırıcıyla ilgili temel sorun her zaman mükemmel şekilde tutarlı yani eğitim hatası olmayan bir hipotez üretmesidir. Marj kavramına bağımlılık, gürültünün her zaman var olabileceği gerçek verilerde kırılgan bir tahminleyene neden olabilmektedir. Yumuşak marjlar hatayı minimize ederek doğrusal olarak ayrılamayan sınıfları çözmeye yardımcı olmaktadırlar. Kernel hilesi yöntemi ise doğrusal olmayan DVM'ler için kullanılmaktadır ve girdi vektörlerini özellik uzayına haritalamaktadır.

***<u>Yumuşak Marj DVM:</u>***

Gürültünün var olduğu veya sınıfların üst üste bindiği durumlarla başa çıkabilmek için marjların yerleştirilmesinde; marj içerisindeki tüm veriler, ister ayırıcı doğrunun yanlış tarafında ister doğru tarafında olsun ihmal edilerek, hatayla işbirliği yapmak mümkündür. Doğrusal ayrılabilir olmayan durumda optimal hiperdüzlem geometrik marjı maksimize eden ve hata fonksiyonunu minimize eden bir hiperdüzlem olarak tanımlanmaktadır. Kısıtların kontrollü olarak ihlal edilmesine izin veren marj aylak değişkenlerinin $\xi_i$ eklenmesi yoluyla $\mathrm{y_i}\left((\mathrm{x_i} * \psi) + \mathrm{b}\right) \geq 1$ , $\mathrm{i} = 1,2, \ldots . \ell$ kısıtları $\mathrm{y_i}\left((\mathrm{x_i} * \psi) + \mathrm{b}\right) \geq 1 - \xi_i$, $\xi_i \geq 0$ olarak değişmektedir. Bu durumda hem marj maksimizasyonunu hem de hata minimizasyonunu birlikte ifade eden optimizasyon problemi (7) (Vapnik, 1998: 411) şu şekilde ifade edilmektedir:

$$Min\ \phi(\psi, \xi) = \frac{1}{2}(\psi * \psi) + C\left(\sum_{i=1}^{\ell} \xi_i\right) \qquad (7)$$

Öyle ki,

$$y_i\big((x_i * \psi) + b\big) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, i = 1,2, \ldots . \ell$$

Eğer $\xi_i > 1$ ise $x_i$ veri noktası hiperdüzlem tarafından yanlış sınıflandırılmıştır. Ceza faktörü olarak bilinen C, eğitim hatasını minimize etmek ile marjı maksimize etmek arasındaki eşiği kontrol etmektedir ve kullanıcı tarafından belirlenmektedir.

Lagranj tekniği ile çözülen optimizasyon problemi (7) dual formda da ifade edilebilmektedir. Dual form problemin iç çarpımlar cinsinden ifade edilmesine fırsat tanıdığı için orijinal formülasyonu tercih edilmektedir.

$$Maks\ W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (x_i * x_j)$$

Öyle ki,

$$0 \leq \alpha_i \leq C, i = 1, \ldots, \ell,$$

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0$$

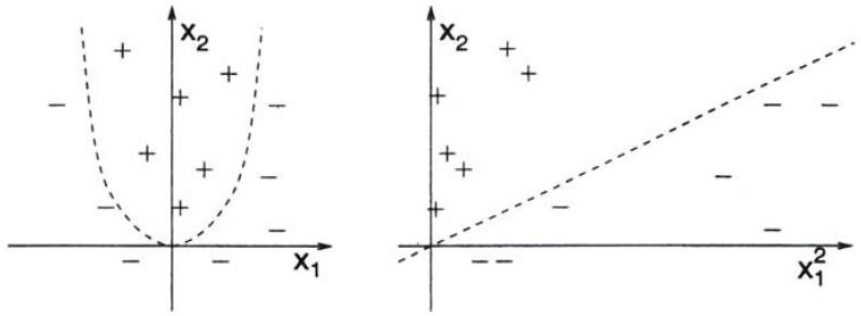
$\alpha_i > 0$ olan tüm eğitim örnekleri daha önce olduğu gibi destek vektörleri olarak adlandırılmaktadır. Destek vektörleri artık ya da marj üzerinde [$0 < \alpha_i < C$'ya sahip $(y_i f(x_i) = 1)$] (sınırlandırılmamış destek vektörleri) ya da marj alanının içerisinde uzanmaktadırlar [$\alpha_i = C$'ya sahip $y_i f(x_i) < 1$)] (sınırlandırılmış destek vektörleri).

***Kernel Hilesi:***

DVM'ler doğrusal olmayan öğrenenlere kolaylıkla dönüştürülebilmektedir. Bu süreçte orijinal verilerden sınıflandırma özelliklerini çıkarmak için doğrusal olmayan bir haritalama kullanılmaktadır (Pöyhönen, 2004). Haritalama fonksiyonunda ana

fikir, verinin daha üst boyutlarda bir uzaya dönüştürülmesidir (Akpınar, 2014: 277). Bir kernel (çekirdek) fonksiyonu genellikle veriyi özellik uzayına $\phi(x)$ fonksiyonu ile haritalayarak onun boyutsallığını arttırmaktadır. DVM daha sonra özellik uzayında maksimal marj doğrusal sınıflandırma kuralını öğrenmektedir. Sınıflandırma kuralı özellik uzayında doğrusal olsa da, orijinal girdi uzayında doğrusal değildir.

Şekil 2.7'de sol taraf $(x_1, x_2)$'de doğrusal ayrılabilir olmayan eğitim kümeleri gösterilmektedir. Sağ tarafta ise aynı problem, $(x_1^2, x_2)$ düzlemi üzerine yansıtılarak doğrusal olmayan bir dönüşüm sonrasını göstermektedir. Bu yeni uzayda eğitim örnekleri doğrusal ayrılabilirdir.




**Şekil 2-7 Doğrusal olmayan haritalama (Joachims, 2002).**

Doğrusal olmayan haritalama $\phi$; destek vektörleri ile özellik uzayındaki örüntü vektörü arasındaki iç çarpımları hesaplamak için kernel fonksiyonlarını kullanmaktadır:

$$K(x_i, x_j) \equiv (\emptyset(x_i)^T \emptyset(x_j))$$

Literatürde en sık kullanılan kernel tipleri ve parametreleri Tablo 2.1'de gösterilmiştir.

**Tablo 2-1 Kernel tipleri ve parametreleri**

| Kernel Tipi | Eşitlik | Parametre |
|---|---|---|
| Doğrusal | $K(x_i, x_j) = x_i^T x_j$ | - |
| Polinominal | $K(x_i, x_j) = (\gamma(x_i^T x_j + c))^p$ | $\gamma, c, p$ |
| Gaussian (RBF) | $K(x_i, x_j) = \exp\left(-\left\|x_i - x_j\right\|^2/\sigma^2\right)$ | $\sigma$ |
| Sigmoid (MLP) | $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + c)$ | $\gamma$, c |

## 2.3 Sınıflandırıcı Değerlendirme

Bir sınıflandırıcı oluşturulduğunda doğruluğunun da hesaplanması gerekmektedir. Bir sınıflandırıcının yaklaşık doğruluğunu bilmemek, gerçek dünya problemlerinde kullanılamamasına neden olduğundan, etkin değerlendirme için önemlidir (Liu, 2007).

Bir sınıflandırıcıyı değerlendirmek için birden fazla ölçü bulunmaktadır. Temel ölçü, test setinde doğru sınıflanan örnek birim sayısının, test setindeki birimlerin toplam sayısına bölünmesiyle elde edilen sınıflandırma doğruluğudur (doğruluk oranı). Hata oranı ise, 1'den doğruluk oranının çıkarılmasıyla elde edilir. Birden fazla sınıflandırıcı mevcut ise en yüksek doğruluğu veren tercih edilmelidir. Aynı eğitim ve test setleri için hangi sınıflandırıcının daha iyi olduğu istatistiksel anlamlılık testleri ile de kontrol edilebilir (Liu, 2007).

En sık kullanılan yöntemler, doğru sınıflandırma oranı (correct classification rate), hata oranı (error rate), sınıflandırma doğruluğu (classification accuracy) ve ROC eğrileridir (duyarlılık-belirlilik eğrileri) (He, Kong ve Shen, 2005), (Ren, Shen ve Ma, 2004).

DVM tabanlı sınıflandırma için performans, destek vektör sayısı ve diğer performans kriterleri ile öğrenmiş modelin büyüklüğüyle değerlendirilmektedir. Performans kriterleri bir kontenjans tablosu ile hesaplanmaktadır. Doğruluk oranı sadece bir sınıfın tahmini değil tüm sınıflandırmayı temsil eden bir ölçüdür. Bununla birlikte kesinlik (precision), duyarlılık (recall) ve belirlilik (specificity) sadece bir sınıfın tahmin performansını ölçmektedir. Kappa istatistiği sınıflandırma kalitesini

ölçmek için kullanılmaktadır. Doğruluk gibi kappa da tek değerli bir kontenjans tablosu ile gösterilmektedir ve 1'e yaklaştıkça sınıflandırıcının performansı artmaktadır.

### 2.3.1 Holdout Set

Mevcut D verisi, $D_{test}D = D_{eğitim} \cup D_{test}$ ve $D_{eğitim} \cap D_{test} = \emptyset$ olmak üzere, $D_{eğitim}$ eğitim seti ve $D_{test}$ test verisi olarak iki ayrık alt sete bölünmektedir. Test seti holdout seti olarak da adlandırılmaktadır. Holdout (dışarıda tutma), daha çok D veri seti büyük olduğunda kullanılmaktadır. Orijinal veri seti D'deki örneklerin sınıflar ile etiketlenmiş olmasına dikkat edilmelidir (Liu, 2007).

Eğitim seti bir sınıflandırıcı öğrenmek için, test seti ise sınıflayıcı sonucunu değerlendirmek için kullanılmaktadır. Sınıflandırıcı değerlendirmek için test verisi yerine eğitim verisi kullanılması durumunda eğitim setinde çok yüksek doğruluk yakalanırken, modelin test verisi için genellemesi düşük olmakta, bir başka ifadeyle test verisinde düşük doğruluk oranı söz konusu olmaktadır. Bu durum sınıflandırıcının eğitim setine aşırı uyum (overfit) göstermesi olarak ifade edilmektedir. Görülmeyen test seti kullanmak tarafsız sınıflandırma doğruluğu tahmini vermektedir. Eğitim ve test için veri setinin ne kadarlık yüzdelerinin kullanılacağı veri setinin büyüklüğüne bağlıdır (Liu, 2007).

D'yi eğitim ve test seti olarak bölmek için temel iki yaklaşım kullanılmaktadır:

1. $D$'den tesadüfi olarak eğitim için örnek çekmek ve geri kalanını test için ayırmak,
2. Veri, zaman ile toplanmışsa verinin eski tarihli bölümünü öğrenme/eğitim için, sonraki bölümlerini ise test için ayırmak.

### 2.3.2 Çoklu Tesadüfi Örnekleme

Mevcut veri seti küçük olduğunda test seti temsil edilmek için oldukça küçük olacaktır. Bu sorun ile baş edebilmek için n defa tesadüfi örneklem seçilebilmektedir. Her seferinde farklı bir eğitim seti ve farklı bir test seti ile n adet doğruluk oranı

üretilmektedir. Veri üzerinde nihai tahmin doğruluğu ise, n adet doğruluk oranının ortalaması alınarak hesaplanır (Liu, 2007).

### 2.3.3 Çapraz Değerleme

Veri seti küçük olduğunda n kat çapraz değerleme (*n*-fold cross-validation) yöntemi sıklıkla kullanılmaktadır. Bu yöntemde, mevcut veri n sayıda eşit boyutlu ayrık alt setlere ayrılmaktadır. Her alt set test seti olarak kullanılmakta ve kalan *n*-1 alt set sınıflandırıcıyı öğrenmek için eğitim seti olarak birleşmektedir. Bu süreç *n* defa çalışır. Bu da *n* sayıda doğruluk oranı verir. Veri setinden son tahmin edilmiş öğrenme doğruluğu *n* sayıdaki doğruluğun ortalamasıdır. Uygulamalarda, sıklıkla n=5 veya n=10 olarak kullanılmaktadır (Jiawei, Kamber ve Pei, 2011).

Çapraz değerlemenin özel bir durumu ise birini dışarda bırak (leave-one-out) yöntemidir. Bu yöntemde, çapraz değerlemenin her katı sadece tek bir test örneğine sahiptir ve verinin geri kalanı eğitimde kullanılır. Buradan orijinal veri m örneğe sahip ise, bu *m*-kat çapraz değerlemedir. Bu yöntem genelde mevcut veri seti çok küçük ise uygulanmaktadır. Büyük veri seti için *m* sayıda sınıflandırıcının kurulması etkin değildir (Liu, 2007).

### 2.3.4 Kesinlik, Duyarlılık, F-skor

Özellikle metin ve web uygulamalarında sadece tek bir sınıf ile ilgilenilmektedir. Ayrıca ağ ihlali (network intrusion) ve finansal dolandırıcılık tespiti gibi yüksek veri dengesizliği içeren sınıflandırmada da sadece azınlık sınıfıyla ilgilenilmektedir. Kullanıcı sınıfı pozitif sınıf, geri kalan ise negatif sınıf olarak adlandırılmaktadır. Bu gibi durumlar için uygun bir ölçü olmayıp, doğruluk oranı oldukça yüksek olacak ancak tek bir ihlali tanımlayamayacaktır. Örneğin, bir ağ ihlalinin tespiti için kullanılan veri setinde birimlerin %99'u normaldir. Bir sınıflandırıcı basitçe bir şey yapmaksızın, her test birimini "ihlal değil" olarak sınıflandırarak %99'luk bir doğruluk oranı elde edebilmektedir ancak bu da yararsızdır (Liu, 2007), (Alpaydın, 2011). Kesinlik (precision) ve duyarlılık (recall) pozitif sınıf üzerine ne kadar hassas ve ne kadar eksiksiz sınıflandırma yapıldığının ölçüleridir ve bu tarz uygulamalar için daha uygundur. Kesinlik ve duyarlılık ölçüleri kontenjans

tablosu kullanılarak gösterilebilmektedir. Bir kontenjans tablosu, bir sınıflandırıcıyla verilen gerçek ve tahmin edilmiş sonuçları içermektedir (Liu, 2007), (Balaban ve Kartal, 2015), (Jiawei, Kamber ve Pei, 2011).

**Tablo 2-2 Kontenjans Tablosu**

| | | Gerçek | | Toplam |
|---|---|---|---|---|
| | | Pozitif | Negatif | |
| Tahmin | Pozitif | Doğru Pozitif (dp) | Yanlış Pozitif (yp) | tPoz |
| | Negatif | Yanlış Negatif (yn) | Doğru Negatif (dn) | tNeg |
| | Toplam | Poz | neg | m |

dp (tp): pozitif örneklerin doğru sınıflandırma sayısı

yn (fn): pozitif örneklerin yanlış sınıflandırma sayısı

yp (fp): negatif örneklerin yanlış sınıflandırma sayısı

dn (tn): negatif örneklerin doğru sınıflandırma sayısı

tPoz: pozitif tahmin örneklerinin toplam sayısı

tNeg: negatif tahmin örneklerinin toplam sayısı

Kontenjans tablosuna (Balaban ve Kartal, 2015) göre pozitif sınıfların kesinlik (p) ve duyarlılığı (r) şu şekilde tanımlanmaktadır.

$$pozitif\ öngörü\ değeri\ (PPV) = p = \frac{dp}{tPoz} = \frac{dp}{dp + yp}$$

$$duyarlılık\ (TPR) = r = \frac{dp}{poz} = \frac{dp}{dp + yn}$$

Kesinlik (p), doğru olarak sınıflandırılan pozitif örneklerin pozitif olarak sınıflandırılan toplam örneklere bölünmesidir. Duyarlılık (r) ise doğru olarak sınıflandırılan pozitif örneklerin test setindeki gerçek pozitif örneklerin toplam sayısına bölünmesidir. Anlamları açık olmasına karşın bu iki ölçüye göre sınıflandırıcıları kıyaslamak zordur. Bir test seti için kesinlik çok yüksek ancak duyarlılık çok düşük ya da tam tersi çıkabilir (Liu, 2007). Teoride kesinlik ve duyarlılık ilişkili değildir. Ancak yüksek kesinlik genelde düşük duyarlılıkta ve yüksek duyarlılık düşük kesinlikte kazanılmaktadır. Bir uygulamada hangi ölçünün önemli olduğu uygulamanın doğasına bağlıdır. Eğer farklı sınıflandırıcıları kıyaslamak için tek bir ölçüye ihtiyaç duyuluyorsa sıklıkla F-skor kullanılmaktadır (Jiawei, Kamber ve Pei, 2011).

$$F = \frac{2pr}{p + r}$$

F-skor kesinlik ve duyarlılığın harmonik ortalamasıdır. F-skorun yüksek olması için p ve r değerleri de yüksek olmalıdır.

$$F = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

# BÖLÜM 3

# 3 WEB GÜNLÜK DOSYALARININ WEB KULLANIM MADENCİLİĞİ İLE ANALİZİ ÜZERİNE UYGULAMA

Çalışmanın son bölümünde, web günlük verileri üzerine Destek Vektör Makineleri (DVM) ve uygulanmıştır. Öncelikle verinin ön analizi hakkında bilgi verilmiş ardından verinin temizleme ve dönüşüm süreci anlatılmıştır. Son olarak da sınıflandırma yöntemi DVM ve karşılaştırma yöntemi olarak kullanılan Logistik regresyon ile farklı modeller denenerek satın alım sınıflandırma sonuçları değerlendirilmiştir.

E-ticaret sitelerinde ziyaretçiler her zaman satın alım kararı ile web sitesini ziyaret etmemekte; ürün kıyaslama, fiyat öğrenme gibi amaçlarla da sitelere erişmektedir. Bu durum satın alım sayısının ziyaret sayısından oldukça küçük olmasına, başka bir ifadeyle müşteri tepki oranının düşük kalmasına yol açmaktadır. Müşteri tepki oranının düşük olduğu dolandırıcılık tespiti, iptalden vazgeçirme, promosyon çalışmaları v.b. alanlar Müşteri tepki modellerini ortaya çıkartmıştır. E-ticaret alanında ise müşteri tepki oranı diğer alanlara göre çok daha düşüktür. Bu doktora tezinde verisi kullanılan e-ticaret web sitesi için bu oran yüzde birden daha küçüktür. Literatürde, müşteri tepki modelleri çerçevesinde çok sayıda uygulama olsa da e-ticaret alanında web günlük verileriyle yapılmış çalışmaya rastlanmamıştır. Makine öğrenmesi yöntemleri ise müşteri tepki modellerinde oldukça seyrek kullanılmakta, ancak bu doktora tezi kapsamında bir çalışmaya rastlanılmamıştır. Bu çalışmada, makine öğrenmesi yöntemlerinden DVM'nin web kullanım madenciliği kapsamında müşteri tepki modeli olarak kullanılabilirliği test edilmiştir. Tepki modellerinde sıkça kullanılan çok değişkenli güçlü bir istatistik yöntem olan logistik regresyon ile de sonuçlar karşılaştırılmıştır. Çalışmada Srivastava v.d.'nin (2000) önerdiği web madenciliği analiz süreci (Bölüm 1) takip edilmiştir.

## 3.1 Verinin Ön Analizi

Web kullanım madenciliğinde web sunucu günlükleri kullanılmaktadır. Web sunucu günlükleri ASCII biçiminde metin dosyası olarak sunucularda saklanmakta, içeriğinde siteye giriş yapan ziyaretçilerin IP adreslerini de barındırmaktadır. IP gizliliğini esas alan şirketler web sunucu günlüklerini paylaşmamayı tercih etmektedir. Uygulamada kullanılan e-ticaret sunucu günlük dosyaları, İstanbul'da faaliyet gösteren bir e-ticaret sitesine aittir. E-ticaret sitesinin fiziksel satış mekânı bulunmamakta, sadece tek bir ürün (farklı seçeneklerinin olduğu) satışı gerçekleştirilmektedir. E-ticaret sitesinin isteği üzerine çalışmada ismi saklı tutulmuş, analiz ekranlarında IP bilgileri gizlenmiştir.

Uygulamada söz konusu e-ticaret sitesinin 13 Mayıs 2011 ile 1 Ağustos 2013 arası dönemi içeren 812 güne ait sunucu günlük dosyaları kullanılmıştır. Verinin özet istatistiklerin çıkarılması için "Nihuo Web Log Analyzer" programı kullanılmıştır. Program ücretli olmakla birlikte 30 günlük deneme süresi içerisinde özet istatistikler çıkarılmıştır.

Tablo 3.1.'de verilen genel istatistiklere bakıldığında toplam istek sayısının 23.538.873 olduğu görülmektedir. Bu sayı web günlük dosyalarındaki toplam satır sayısıdır. Her satır, ziyaretçiler tarafından sunucuya yapılmış bir isteğe karşılık gelmektedir. Bu istekler metin, resim ya da video olabilmektedir. Ayrıca ziyaretçilerden gelen isteklerden farklı olarak yazılımlar aracılığıyla yapılan istekler de bulunmaktadır. Örümcek (robot, bot, spider, crawler vb. istekler uygulamada ortak isimle örümcek olarak adlandırılmıştır ancak aralarında farklılıklar bulunmaktadır) adını verdiğimiz bu ziyaretler, başta Google olmak üzere diğer arama motorlarının ya da web üzerinden veri toplamak isteyen kişilerin oluşturduğu yazılımlar aracılığıyla yapılmış insan kaynaklı olmayan isteklerdir.

Ceplenmiş istekler, istemci üzerinde kaydedilmiş isteklerdir. Bir tarayıcı daha önce istenmiş dosyanın kopyasını yedeklemiş ise sunucuya güncel olanı değil yedeği gönderir. Kaybedilen istekler ise hatayla sonuçlanan isteklerdir.

Sayfa görüntüleme, web sitesinin tek bir web sayfasına erişimini ifade etmektedir. İstek sayısı web sunucuda bulunan bir dosyaya (resim, metin, v.b.) erişime karşılık gelirken, sayfa görüntüleme sayısı belli bir zamanda erişilen sayfa sayısıdır. Örneğin bir web sayfası 5 resim içeriyorsa o sayfaya olan ziyaret 6 istek olarak günlük dosyasında kayda alınacaktır. Bir istek web sayfasına, 5 istek ise resimlere olacaktır. Bir ziyaretçi 10 sayfa dolaşmış ve her sayfa 10 resim içeriyorsa web sunucu 110 istek kaydedecektir. Sayfa görüntüleme sayısı ise 10 olacaktır. Ziyaret sayıları, sayfa görüntülemenin IP adreslerine göre gruplanmış hali (tekil kullanıcıların ziyaretleri) olarak ifade edilebilir.

Tablo 3-1'de verilen genel istatistiklere bakıldığında toplam istek sayısının 23.538.873 olduğu görülmektedir. Bu isteklerin 605.239'u örümcek denilen makine üretimidir. Örümcekleri, insanların değil çoğunlukla arama motorlarının web sayfalarına indeksleme amaçlı tesadüfi yaptıkları ziyaretler olarak tanımlamak mümkündür. Sayfa görüntülemelerine bakıldığında 2 milyonun üzerinde sayfa görüntülendiği görülmektedir. Günde ortalama 2.724 sayfa görüntülenmiştir. Ziyaret başına ise ortalama 4,87 sayfa görüntülenmiştir.

**Tablo 3-1 İstek Sayısı, Ziyaret, Sayfa İstatistikleri**

| **İstek Sayıları** | |
|---|---|
| Toplam İstek Sayısı | 23.535.873 |
| Normal İstek Sayısı | 22.930.634 |
| Örümcek İstek Sayısı | 605.239 |
| Gün Başına Düşen Ortalama İstek Sayısı | 29.020 |
| Ziyaret Başına Düşen Ortalama İstek Sayısı | 51,92 |
| Ceplenmiş İstekler | 1.261.170 |
| Kaybedilen İstekler | 310.900 |
| **Sayfa Görüntülemeleri** | |
| Toplam Sayfa Görüntülemesi | 2.209.288 |
| Gün Başına Düşen Ortalama Sayfa Görüntülemesi | 2.724 |
| Ziyaret Başına Düşen Ortalama Sayfa Görüntülemesi | 4,87 |
| Tekil Sayfa Görüntülemesi | 1.428.643 |
| **Ziyaretler** | |
| Toplam Ziyaret Sayısı | 453.304 |
| Normal Ziyaretler | 253.505 |
| Örümcek Ziyaretler | 199.799 |

| | |
|---|---|
| Gün Başına Düşen Ortalama Ziyaret | 558 |
| Toplam Ziyaretçi Kalış Süresi (dakika) | 46.799:57:02 |
| Ortalama Ziyaretçi Kalış Süresi (dakika) | 06:11 |

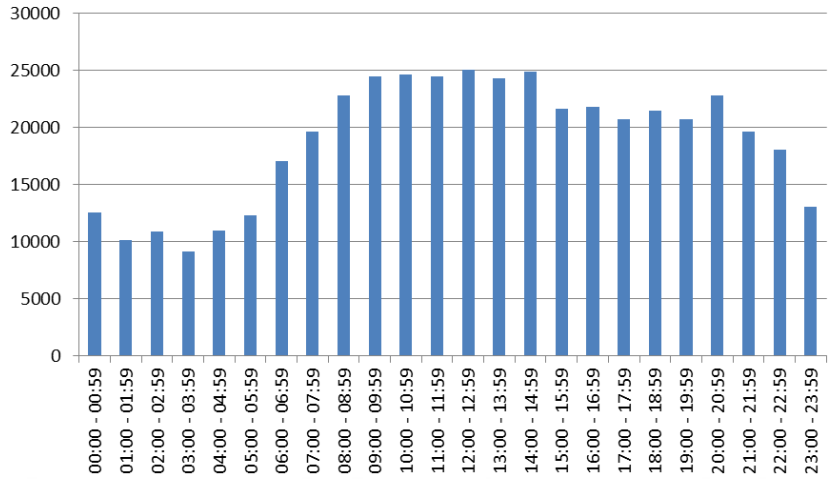
Ziyaret istatistiklerine bakıldığında toplamda 453.304 ziyaret gerçekleşmiş olmasına karşın, bunlardan sadece 253.505 adetinin bireylere ait olduğu gözükmektedir. Bir ziyaretçinin ortalama 6 dakika 11 saniye söz konusu e-ticaret sitesinde zaman geçirdiği görülmektedir.

Genel istatistiklerin devamı olarak Tablo 3-2'de yer alan aktivite bilgilerine bakıldığında hafta içi beş güne ait ortalama ziyaret sayısı (577) ile hafta sonu iki güne ait ortalama ziyaret sayılarının (509) birbirine yakın olduğu görülmektedir. Sitenin ziyaretinde en fazla tercih edilen gün Perşembe iken, en az tercih edilen gün ise Pazar olmuştur.

**Tablo 3-2 Aktivite Özeti**

| | |
|---|---|
| Hafta içi Gün Başına Düşen Ortalama Ziyaret Sayısı | 577 |
| Hafta içi Gün Başına Düşen Ortalama İstek Sayısı | 30.551 |
| Hafta sonu Gün Başına Düşen Ortalama Ziyaret Sayısı | 509 |
| Hafta sonu Gün Başına Düşen Ortalama İstek Sayısı | 25.069 |
| Haftanın En Aktif Ziyaret Günü | Perşembe |
| Haftanın En Az Aktif Ziyaret Günü | Pazar |
| En Aktif Tarih | Perşembe, 02 Şubat, 2012 |
| En Aktif Tarihteki İstek Sayısı | 421.034 |
| En Aktif Tarihteki Ziyaret Sayısı | 2.349 |
| En Az Aktif Tarih | Pazar, 22 Mayıs, 2011 |
| En Az Aktif Tarihteki İstek Sayısı | 523 |
| En Az Aktif Tarihteki Ziyaret Sayısı | 11 |
| Günün En Aktif Saati | 09:00 - 09:59 |
| Günün En Az Aktif Saati | 03:00 - 03:59 |

Saat aralıklarına göre ziyaret sayıları incelendiğinde (Şekil 3.2) en yoğun saat aralıklarının 12:00-12:59 ve 14:00-14:59 dilimleri olduğu görülmektedir. Saat aralıklarına göre istek, sayfa, ziyaret sayıları daha ayrıntılı olarak Ek 3'te tablo olarak verilmiştir.



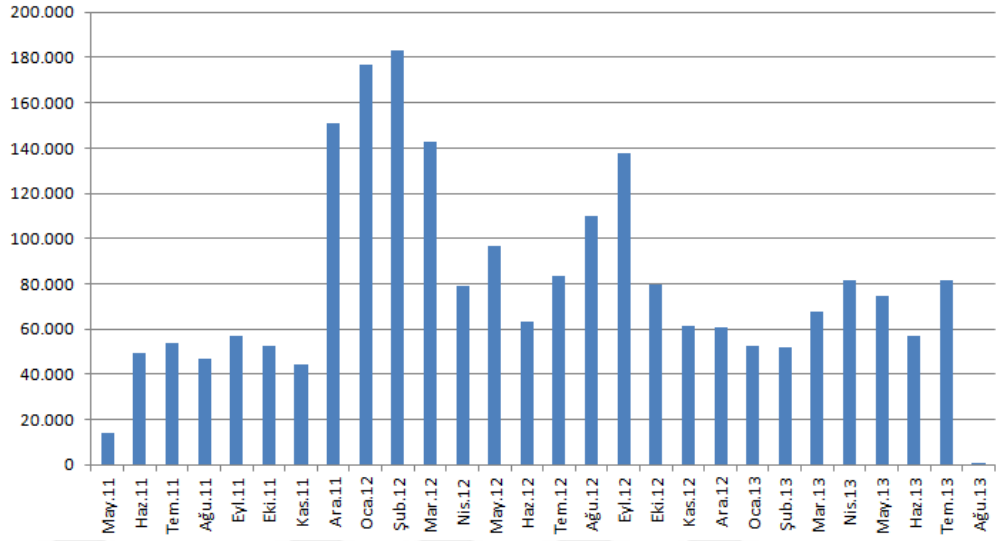

**Şekil 3-1 Ziyaretlerin saat aralıklarına göre dağılımı**

Tablo 3.3 incelendiğinde günlere göre istek, sayfa ve ziyaret sayıları görülmektedir. Günlere göre sayfa başına ziyaret sürelerinin birbirine çok yakın olduğu görülmektedir.

**Tablo 3-3 Günlere göre istek, sayfa, ziyaretçi dağılımı**

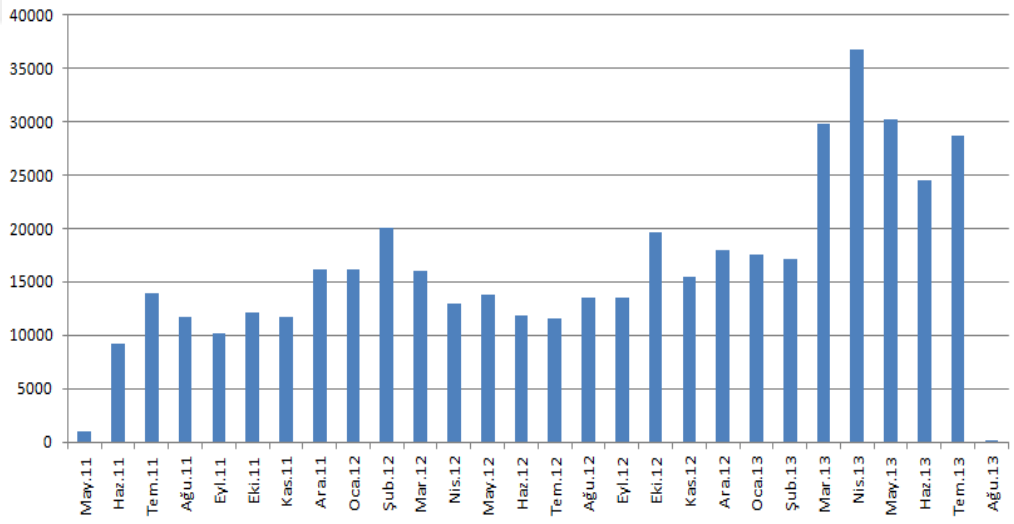
| Gün | İstek Sayısı | Sayfalar | Ziyaretler | Ziyaret Süresi/Sayfa |
|---|---|---|---|---|
| Pazar | 2.834.210 | 269.629 | 58.114 | 06:20 |
| Pazartesi | 3.500.517 | 359.157 | 67.760 | 06:04 |
| Salı | 3.671.452 | 356.751 | 67.274 | 06:04 |
| Çarşamba | 3.425.646 | 327.949 | 66.259 | 06:17 |
| Perşembe | 3.734.544 | 320.724 | 68.155 | 06:06 |
| Cuma | 3.387.524 | 307.441 | 65.655 | 06:19 |
| Cumartesi | 2.981.980 | 267.637 | 60.087 | 06:10 |
| **Ortalama** | **3.362.267** | **315.612** | **64.757** | **06:11** |
| **Toplam** | **23.535.873** | **2.209.288** | **453.304** | |

Şekil 3.2'de aylara göre görüntülenen sayfa sayıları özetlenmiştir. En fazla sayfa Ocak ve Şubat 2012 tarihlerinde görüntülenmiştir.




**Şekil 3-2 Aylara göre sayfa sayılarının dağılımı**

Şekil 3.3'te aylara göre ziyaret sayılarının dağılımı gösterilmekte, en çok ziyaretin Nisan ve Mayıs 2013 tarihlerinde olduğu gözükmektedir. Aylara göre istek, sayfa ve ziyaret sayıları EK3'te ayrıntılı olarak verilmiştir.




**Şekil 3-3 Aylara göre ziyaret sayılarının dağılımı**

Tablo 3-4'e bakıldığında ise toplam ziyaretlerin %40,85'inin Türkiye'den olduğu görülmektedir. Ayrıca 8 dakikanın üzerinde bir süre ile Amerika kaynaklı ziyaretçilerin sitede daha fazla zaman geçirdikleri görülmektedir. Çin, Japonya ve Rusya kaynaklı ziyaretler sırasıyla toplam ziyaretlerin %8,07, %2,74 ve %2,00'sini

oluşturmaktadır. Diğer ülkelerden gelen ziyaretler ise %1'in altındadır. Tablo 3-4'ün diğer ülkelerin de olduğu genişletilmiş versiyonu EK3'te yer almaktadır.

**Tablo 3-4 Ülkelere göre istek, ziyaret, Sayfa sayıları**

| | Ülke | Ziyaretler | Toplam Ziyaret Yüzdesi | İstek Sayısı | Toplam İstek Yüzdesi | Ziyaret Başına Düşen Sayfa Görüntüleme | Ziyaret Başına Düşen Kalma Süresi |
|---|---|---|---|---|---|---|---|
| 1 | A.B.D. | 189.846 | 41,88 | 835.622 | 3,55 | 1,47 | 08:35 |
| 2 | Türkiye | 185.160 | 40,85 | 21.607.502 | 91,81 | 9,51 | 04:53 |
| 3 | Çin | 36.590 | 8,07 | 105.515 | 0,45 | 1,39 | 00:35 |
| 4 | Japonya | 12.408 | 2,74 | 38.460 | 0,16 | 1,07 | 01:19 |
| 5 | Rusya | 9.049 | 2,00 | 79.812 | 0,34 | 2,36 | 20:29 |

Tablo 3-5'te yer alan şehirlere göre dağılımlara bakıldığında Ankara ve İstanbul en çok ziyareti gerçekleştiren ilk iki şehirdir. Üçüncü sırada ise Amerika Redmond yer almaktadır. Türkiye'nin diğer şehirleri ise 23.sırada İzmir, 34.sırada Antalya, 35.sırada Bursa, 38.sırada Gaziantep, 42.sırada İzmit ve 45.sırada Adana olarak sıralanmıştır. Tablo 3-5'in diğer ülkelerin de olduğu genişletilmiş versiyonu EK3'te yer almaktadır.

**Tablo 3-5 Şehirlere göre ziyaret, istek, sayfa sayıları**

| | Şehir | Ziyaretler | İstek Sayısı | Ziyaret Başına Sayfa Görüntüleme | Ziyaret başına Kalış Süresi |
|---|---|---|---|---|---|
| 1 | Ankara | 104.174 | 12.152.946 | 8,52 | 03:23 |
| 2 | Istanbul | 60.498 | 6.940.293 | 11,44 | 07:52 |
| 3 | Washington | 33.367 | 87.409 | 1,30 | 03:48 |
| 4 | California | 27.989 | 140.570 | 1,97 | 17:45 |
| 5 | Pekin | 24.820 | 28.922 | 0,95 | 00:29 |

Son yıllarda akıllı telefonların ve tabletlerin de yaygın kullanılmasıyla, bu cihazların siteye yaptıkları girişlere ulaşılabilmektedir. Tablo 3-6'da en fazla ziyaretçinin IOS tabanlı işletim sistemi kullanan cihazlardan geldiği görülmektedir. Toplam 22.801 ziyaretçinin 15.537'si iPhone kullanıcılarının oluşturduğu ziyaretler

olup, ikinci sırada yine IOS işletim sistemli tablet olan iPad kullanıcılarının oluşturduğu ziyaretler gelmektedir. Android işletim sistemi kullanan çoğu Samsung modeli ayrı yazıldığı halde IOS'un gerisinde kalmış gözükmektedir. Tablonun genişletilmiş versiyonu EK3'te bulunmaktadır.

**Tablo 3-6 Akıllı telefon ve tabletlerin ziyaret bilgileri**

| | Mobil Aletler | Ziyaretler | Toplam Ziyaretler Yüzdesi | İstek Sayısı | Toplam İstek Yüzdesi |
|---|---|---|---|---|---|
| 1 | iPhone | 15.537 | 3,43% | 900.456 | 3,83% |
| 2 | iPad | 5.707 | 1,26% | 762.406 | 3,24% |
| 3 | Samsung SGH-E250 | 500 | 0,11% | 914 | 0,00% |
| 4 | BlackBerry 9800 | 355 | 0,08% | 14.819 | 0,06% |
| 5 | BlackBerry 9700 | 275 | 0,06% | 17.679 | 0,08% |

## 3.2 Veri Temizleme ve Kullanıcı Tanımlama

Veri temizleme aşamasında bütün web günlük dosyaları phpmyadmin aracılığıyla veri tabanı oluşturularak bir araya toplanmıştır. SQL (Structured Query Language – Yapılandırılmış Sorgulama Dili) ile veri temizleme ve düzenleme işlemleri gerçekleştirilmiştir. Veri tabanından elde edilen düzenlenmiş veri .csv uzantılı olarak aktarılarak analizde kullanılabilecek hale getirilmiştir.

Nihuo Web Log Analyzer programından veri temizleme aşamasında da yararlanılmış, analizde istenmeyen verilerin elenmesi ve yazılımların yapmış olduğu sahte ziyaretler (robot) belirlenmiştir. Aşağıda programla tespit edilen bu tip isteklere örnekler verilmiştir.

/robots.txt
/robots.txt?_=1366364465299
/robots.txt?_=1366638165213
/robots.txt?_=1371732786772

Çalışmada yararlanılan verinin ait olduğu e-ticaret sitesi, günlük dosyalarını Windows IIS biçiminde saklamaktadır. Bu günlük dosyaların ilk haline ait bir kesit (4 isteğe ait) aşağıda gösterilmiştir. IP numaraları kişisel güvenlik nedeniyle saklanmıştır (IP numaraları xx.xx.xxx.xxx ile gösterilmiştir).

2011-05-28 00:17:28 W3SVC1 xx.xx.xxx.xxx GET /DXR.axd r=2_15-okUJ2 80 - xx.xx.xxx.xxx Mozilla/4.0+(compatible;+MSIE+8.0;+Windows+NT+5.1;+Trident/4.0;+GTB6.5;+.NET+CLR+1.1.4322;+.NET+CLR+2.0.50727;+InfoPath.2;+OfficeLiveConnector.1.3;+OfficeLivePatch.1.3;+.NET+CLR+3.0.04506.30;+.NET+CLR+3.0.4506.2152;+.NET+CLR+3.5.30729) 200 0 0

2011-05-28 00:17:28 W3SVC1 xx.xx.xxx.xxx GET /DXR.axd r=2_22-okUJ2 80 - xx.xx.xxx.xxx Mozilla/4.0+(compatible;+MSIE+8.0;+Windows+NT+5.1;+Trident/4.0;+GTB6.5;+.NET+CLR+1.1.4322;+.NET+CLR+2.0.50727;+InfoPath.2;+OfficeLiveConnector.1.3;+OfficeLivePatch.1.3;+.NET+CLR+3.0.04506.30;+.NET+CLR+3.0.4506.2152;+.NET+CLR+3.5.30729) 200 0 0

2011-05-28 00:17:28 W3SVC1 xx.xx.xxx.xxx GET /DXR.axd r=2_29-okUJ2 80 - xx.xx.xxx.xxx Mozilla/4.0+(compatible;+MSIE+8.0;+Windows+NT+5.1;+Trident/4.0;+GTB6.5;+.NET+CLR+1.1.4322;+.NET+CLR+2.0.50727;+InfoPath.2;+OfficeLiveConnector.1.3;+OfficeLivePatch.1.3;+.NET+CLR+3.0.04506.30;+.NET+CLR+3.0.4506.2152;+.NET+CLR+3.5.30729) 200 0 0

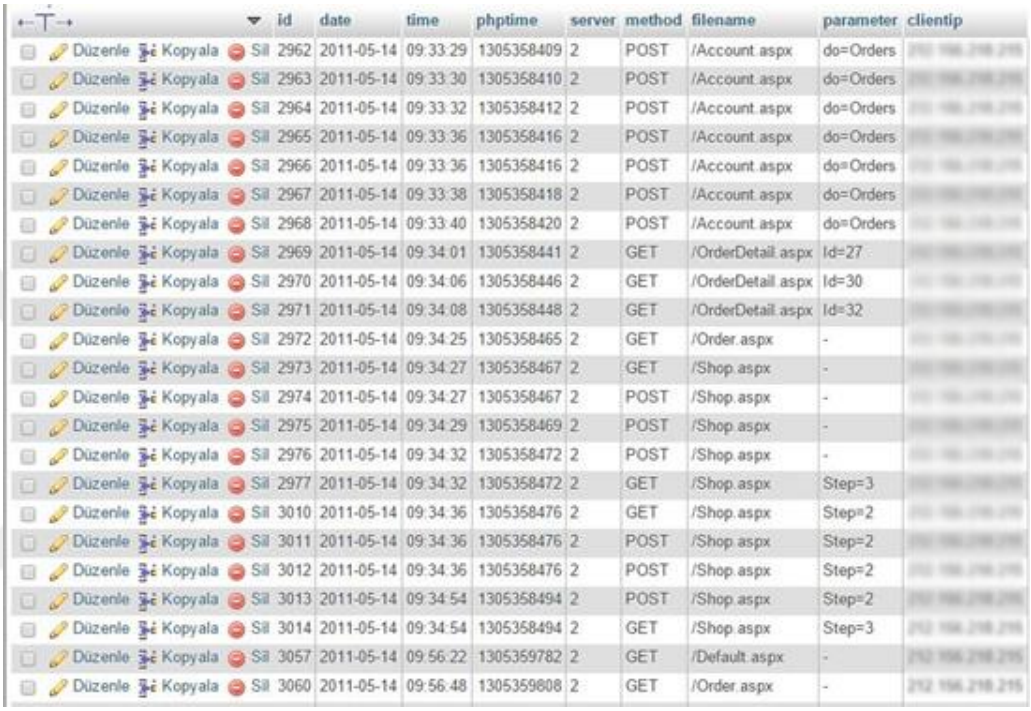
2011-05-28 00:17:28 W3SVC1 xx.xx.xxx.xxx GET /DXR.axd r=1_54-akUJ2 80 - xx.xx.xxx.xxx Mozilla/4.0+(compatible;+MSIE+8.0;+Windows+NT+5.1;+Trident/4.0;+GTB6.5;+.NET+CLR+1.1.4322;+.NET+CLR+2.0.50727;+InfoPath.2;+OfficeLiveConnector.1.3;+OfficeLivePatch.1.3;+.NET+CLR+3.0.04506.30;+.NET+CLR+3.0.4506.2152;+.NET+CLR+3.5.30729) 200 0 0

Öncelikle bu metin dosyaları “mysql” yardımıyla sütunlarına ayrılarak veri tabanına aktarılmıştır. İlk etapta 9 sütuna ayrılmıştır. Bu sütunlar Windows IIS biçimli günlük dosyasının alanlarını temsil etmektedir. Her sütun bir niteliğe karşılık gelmektedir. Şekil 3.4’te görüldüğü üzere bu özellikler id, tarih (date), zaman (time), phptime, sunucu bilgisi (server), yöntem (method), dosya adı (filename), parametre ve istemci IP (clientip)’sidir.

Ardından analiz için gerekli olmayan resim, video, animasyon gibi dosyalar ayıklanmıştır. Bu işlem için kullanılan SQL komutları EK2’de verilmiştir. Filtrelenen uzantılar .swf, .jpeg, .gif gibi medya dosyalarıdır. Bu dosyaların tespiti için de yine Nihou web analyzer programından yararlanılmıştır.

Eleme işlemleri neticesinde 23.535.873 satır 1.872.097 satıra indirilmiştir. Gereksiz verinin temizlenmesi ardından elde edilen veriye ait veri tabanından bir kesit Şekil 3.4’te görülmektedir.

Veri temizleme işlemi sonlandıktan sonra web kullanım madenciliğinde önemli olan kullanıcı oturum tanımlama sürecine geçilmiştir. Kullanıcı tanımlama veri tabanına aktarım ile IP adreslerine göre bir sütunda tanımlanarak gerçekleştirilmiştir. Şekil 3.4'te dikkat edileceği üzere bu sadece bir kullanıcının 23 dakikada yapmış olduğu işlemlere ait bir kesittir.

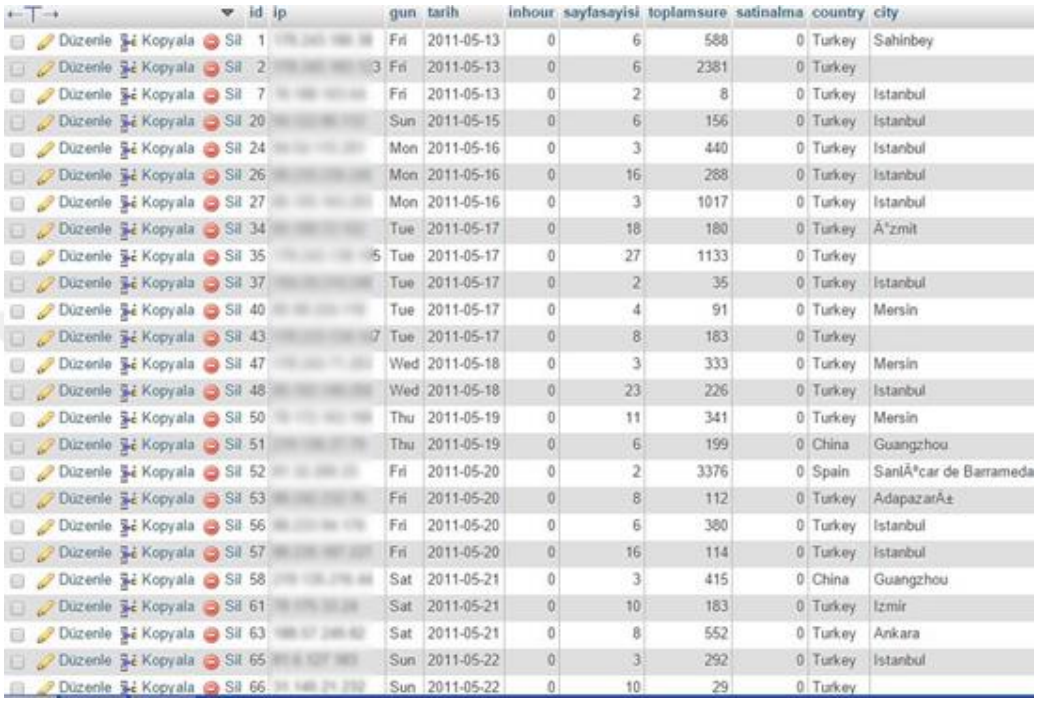


| | id | date | time | phptime | server | method | filename | parameter | clientip |
|---|---|---|---|---|---|---|---|---|---|
| Düzenle Kopyala Sil | 2962 | 2011-05-14 | 09:33:29 | 1305358409 | 2 | POST | /Account.aspx | do=Orders | [redacted] |
| Düzenle Kopyala Sil | 2963 | 2011-05-14 | 09:33:30 | 1305358410 | 2 | POST | /Account.aspx | do=Orders | [redacted] |
| Düzenle Kopyala Sil | 2964 | 2011-05-14 | 09:33:32 | 1305358412 | 2 | POST | /Account.aspx | do=Orders | [redacted] |
| Düzenle Kopyala Sil | 2965 | 2011-05-14 | 09:33:36 | 1305358416 | 2 | POST | /Account.aspx | do=Orders | [redacted] |
| Düzenle Kopyala Sil | 2966 | 2011-05-14 | 09:33:36 | 1305358416 | 2 | POST | /Account.aspx | do=Orders | [redacted] |
| Düzenle Kopyala Sil | 2967 | 2011-05-14 | 09:33:38 | 1305358418 | 2 | POST | /Account.aspx | do=Orders | [redacted] |
| Düzenle Kopyala Sil | 2968 | 2011-05-14 | 09:33:40 | 1305358420 | 2 | POST | /Account.aspx | do=Orders | [redacted] |
| Düzenle Kopyala Sil | 2969 | 2011-05-14 | 09:34:01 | 1305358441 | 2 | GET | /OrderDetail.aspx | Id=27 | [redacted] |
| Düzenle Kopyala Sil | 2970 | 2011-05-14 | 09:34:06 | 1305358446 | 2 | GET | /OrderDetail.aspx | Id=30 | [redacted] |
| Düzenle Kopyala Sil | 2971 | 2011-05-14 | 09:34:08 | 1305358448 | 2 | GET | /OrderDetail.aspx | Id=32 | [redacted] |
| Düzenle Kopyala Sil | 2972 | 2011-05-14 | 09:34:25 | 1305358465 | 2 | GET | /Order.aspx | - | [redacted] |
| Düzenle Kopyala Sil | 2973 | 2011-05-14 | 09:34:27 | 1305358467 | 2 | GET | /Shop.aspx | - | [redacted] |
| Düzenle Kopyala Sil | 2974 | 2011-05-14 | 09:34:27 | 1305358467 | 2 | POST | /Shop.aspx | - | [redacted] |
| Düzenle Kopyala Sil | 2975 | 2011-05-14 | 09:34:29 | 1305358469 | 2 | POST | /Shop.aspx | - | [redacted] |
| Düzenle Kopyala Sil | 2976 | 2011-05-14 | 09:34:32 | 1305358472 | 2 | POST | /Shop.aspx | - | [redacted] |
| Düzenle Kopyala Sil | 2977 | 2011-05-14 | 09:34:32 | 1305358472 | 2 | GET | /Shop.aspx | Step=3 | [redacted] |
| Düzenle Kopyala Sil | 3010 | 2011-05-14 | 09:34:36 | 1305358476 | 2 | GET | /Shop.aspx | Step=2 | [redacted] |
| Düzenle Kopyala Sil | 3011 | 2011-05-14 | 09:34:36 | 1305358476 | 2 | POST | /Shop.aspx | Step=2 | [redacted] |
| Düzenle Kopyala Sil | 3012 | 2011-05-14 | 09:34:36 | 1305358476 | 2 | POST | /Shop.aspx | Step=2 | [redacted] |
| Düzenle Kopyala Sil | 3013 | 2011-05-14 | 09:34:54 | 1305358494 | 2 | POST | /Shop.aspx | Step=2 | [redacted] |
| Düzenle Kopyala Sil | 3014 | 2011-05-14 | 09:34:54 | 1305358494 | 2 | GET | /Shop.aspx | Step=3 | [redacted] |
| Düzenle Kopyala Sil | 3057 | 2011-05-14 | 09:56:22 | 1305359782 | 2 | GET | /Default.aspx | - | [redacted] |
| Düzenle Kopyala Sil | 3060 | 2011-05-14 | 09:56:48 | 1305359808 | 2 | GET | /Order.aspx | - | [redacted] |

**Şekil 3-4 Tek bir kullanıcıya ait günlük kaydı**

Oturum tanımlama, kullanıcı hareketleri sırasında toplanmış IP adreslerinin tüm sayfa görüntülenme kayıtlarını gruplamasından oluşmaktadır. Oturum tanımlama işleminin yapılmasının nedenini şu şekilde açıklamak mümkündür. Bir kullanıcı 24 saat içinde bir web sitesini iki kez ziyaret etmiş olsun. Ziyaretler 6 saat aralık ile yapılmış ise bu 24 saatlik periyot için kullanıcı tanımlama yöntemleri erken uygulanırsa, iki ziyaret beraber sıralanacak ve bu kullanıcıyla tanımlanacaktır. Ancak bu iki ziyaret arasında fark olması gerekir. Çalışmada ziyaretçi hareketsiz kalma süresi (ikinci oturum başlatma süresi) 60 dakika olarak belirlenmiştir. Oturumlar tanımlanırken kullanıcının siteye giriş yaptığı gün, kaç sayfa gezdiği, sitede ne kadar süre harcadığı, satın alma yapıp yapmadığı bilgileri hesaplanmıştır. Bu işlem için yine SQL'den ve ilave olarak PHP dilinden faydalanılmıştır. Oluşturulan kodlar EK1'de

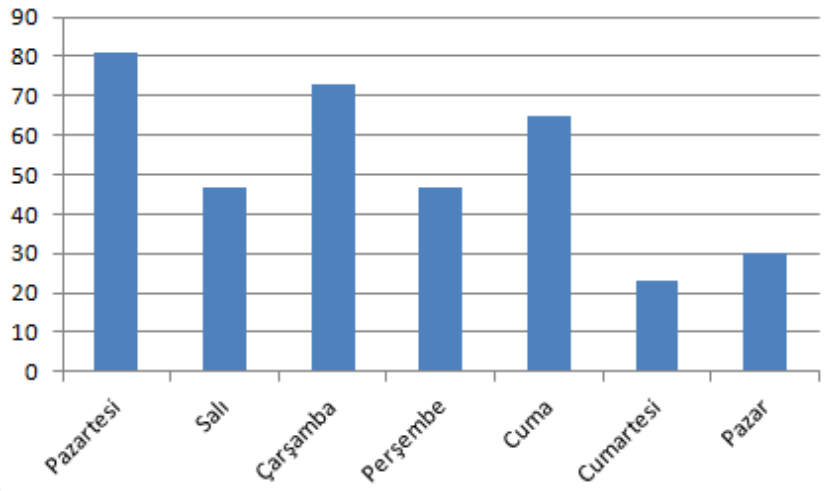
yer almaktadır. Kullanıcı ve oturum tanımlama işlemleriyle veri dönüşüm gerçekleştirilmiş ve veri tabanının son hali Şekil 3.5'te verilmiştir. Oturum tanımlama ile 1.872.097 olan satır sayısı 91.095 oturuma dönüşmüştür.



| | id | ip | gun | tarih | inhour | sayfasayisi | toplamsure | satinalma | country | city |
|---|---|---|---|---|---|---|---|---|---|---|
| Düzenle Kopyala Sil | 1 | | Fri | 2011-05-13 | 0 | 6 | 588 | 0 | Turkey | Sahinbey |
| Düzenle Kopyala Sil | 2 | 3 | Fri | 2011-05-13 | 0 | 6 | 2381 | 0 | Turkey | |
| Düzenle Kopyala Sil | 7 | | Fri | 2011-05-13 | 0 | 2 | 8 | 0 | Turkey | Istanbul |
| Düzenle Kopyala Sil | 20 | | Sun | 2011-05-15 | 0 | 6 | 156 | 0 | Turkey | Istanbul |
| Düzenle Kopyala Sil | 24 | | Mon | 2011-05-16 | 0 | 3 | 440 | 0 | Turkey | Istanbul |
| Düzenle Kopyala Sil | 26 | | Mon | 2011-05-16 | 0 | 16 | 288 | 0 | Turkey | Istanbul |
| Düzenle Kopyala Sil | 27 | | Mon | 2011-05-16 | 0 | 3 | 1017 | 0 | Turkey | Istanbul |
| Düzenle Kopyala Sil | 34 | | Tue | 2011-05-17 | 0 | 18 | 180 | 0 | Turkey | Ä°zmit |
| Düzenle Kopyala Sil | 35 | 5 | Tue | 2011-05-17 | 0 | 27 | 1133 | 0 | Turkey | |
| Düzenle Kopyala Sil | 37 | | Tue | 2011-05-17 | 0 | 2 | 35 | 0 | Turkey | Istanbul |
| Düzenle Kopyala Sil | 40 | | Tue | 2011-05-17 | 0 | 4 | 91 | 0 | Turkey | Mersin |
| Düzenle Kopyala Sil | 43 | 7 | Tue | 2011-05-17 | 0 | 8 | 183 | 0 | Turkey | |
| Düzenle Kopyala Sil | 47 | | Wed | 2011-05-18 | 0 | 3 | 333 | 0 | Turkey | Mersin |
| Düzenle Kopyala Sil | 48 | | Wed | 2011-05-18 | 0 | 23 | 226 | 0 | Turkey | Istanbul |
| Düzenle Kopyala Sil | 50 | | Thu | 2011-05-19 | 0 | 11 | 341 | 0 | Turkey | Mersin |
| Düzenle Kopyala Sil | 51 | | Thu | 2011-05-19 | 0 | 6 | 199 | 0 | China | Guangzhou |
| Düzenle Kopyala Sil | 52 | | Fri | 2011-05-20 | 0 | 2 | 3376 | 0 | Spain | SanlÃ°car de Barrameda |
| Düzenle Kopyala Sil | 53 | | Fri | 2011-05-20 | 0 | 8 | 112 | 0 | Turkey | AdapazarÄ± |
| Düzenle Kopyala Sil | 56 | | Fri | 2011-05-20 | 0 | 6 | 380 | 0 | Turkey | Istanbul |
| Düzenle Kopyala Sil | 57 | | Fri | 2011-05-20 | 0 | 16 | 114 | 0 | Turkey | Istanbul |
| Düzenle Kopyala Sil | 58 | | Sat | 2011-05-21 | 0 | 3 | 415 | 0 | China | Guangzhou |
| Düzenle Kopyala Sil | 61 | | Sat | 2011-05-21 | 0 | 10 | 183 | 0 | Turkey | Izmir |
| Düzenle Kopyala Sil | 63 | | Sat | 2011-05-21 | 0 | 8 | 552 | 0 | Turkey | Ankara |
| Düzenle Kopyala Sil | 65 | | Sun | 2011-05-22 | 0 | 3 | 292 | 0 | Turkey | Istanbul |
| Düzenle Kopyala Sil | 66 | | Sun | 2011-05-22 | 0 | 10 | 29 | 0 | Turkey | |

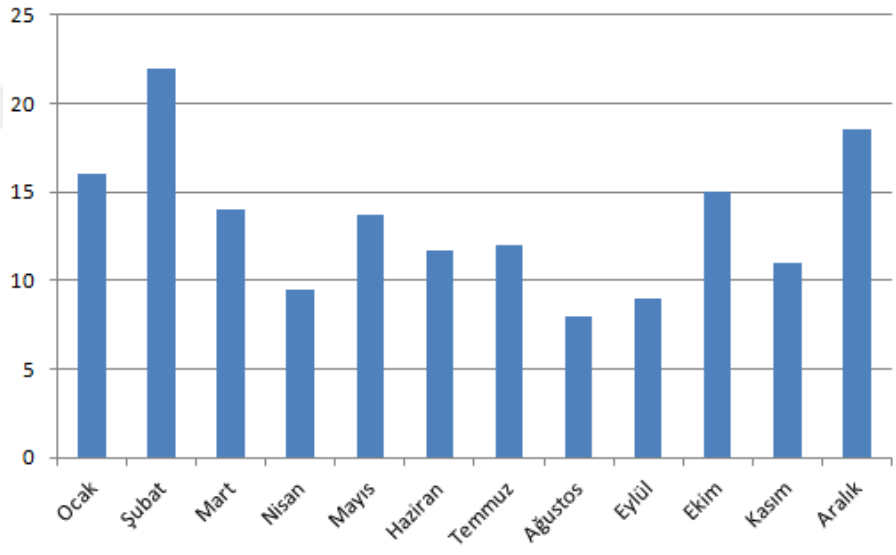
**Şekil 3-5 Veri temizleme ve dönüştürme işlemi sonrası örnek kayıt**

Satışa dönüşen ziyaretler de oturum tanımlama ile belirlenmiştir. 91.095 oturumdan 366'sı satış ile sonuçlanmıştır. Şekil 3.6'da günlere göre satış rakamları verilmiştir. En fazla satışın yapıldığı gün Pazartesi olarak görülmektedir. Bu aynı zamanda söz konusu sitenin en çok ziyaret edildiği gün ile de aynı sonucu vermektedir. Ziyaretin fazla olduğu gün satışta da artış olmuştur.



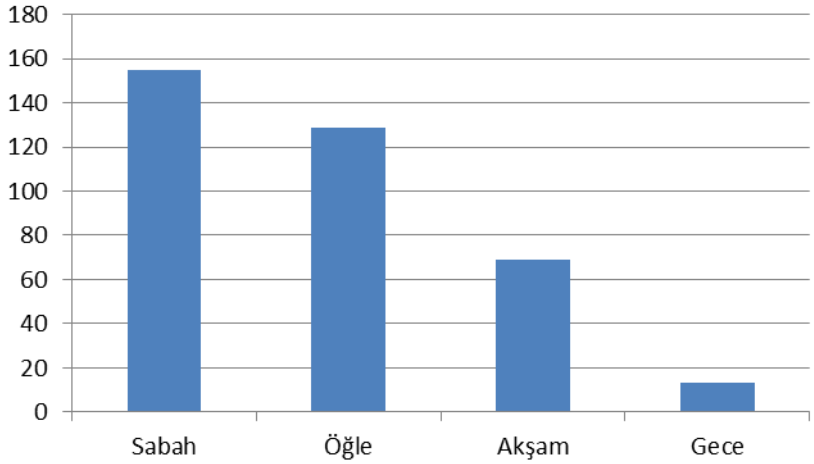

**Şekil 3-6 Günlere göre satılan ürün sayısı**

Şekil 3.7'de aylara göre ortalama satış sayıları verilmiştir. En fazla satışın yapıldığı ay Şubat olarak görülmektedir.



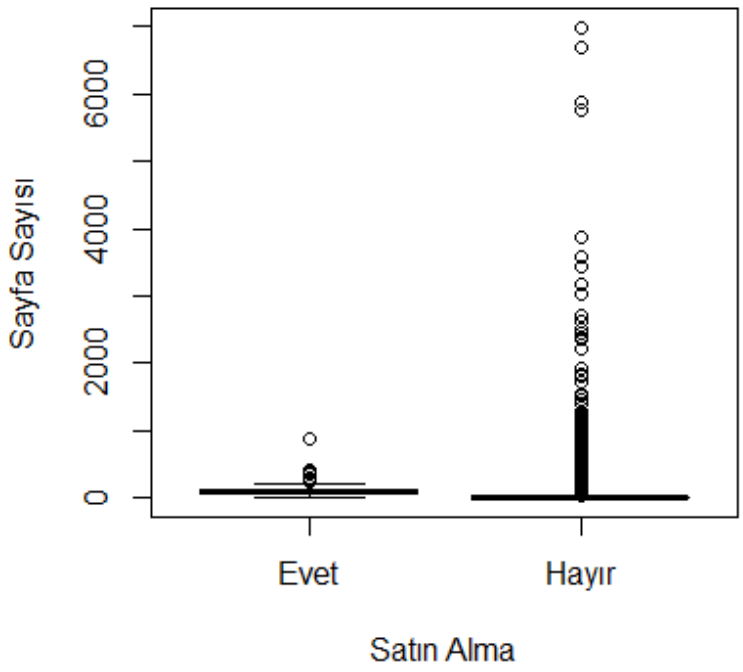

**Şekil 3-7 Aylara göre satılan ortalama ürün sayısı**

Siteye giriş yapılan saatler 4 kategori olarak 4 saat dilimine dönüştürülmüştür. Bu dört kategori Sabah (06:00-11:59 arası), Öğle (12:00-17:59 arası), Akşam (18:00-23:59 arası) ve Gece (00:00-05:59 arası) olarak tanımlanmıştır. Satılan ürün sayıları Şekil 3.8'de verilmiştir. "Sabah" saat diliminde satışların yoğunlaştığı görülmektedir.

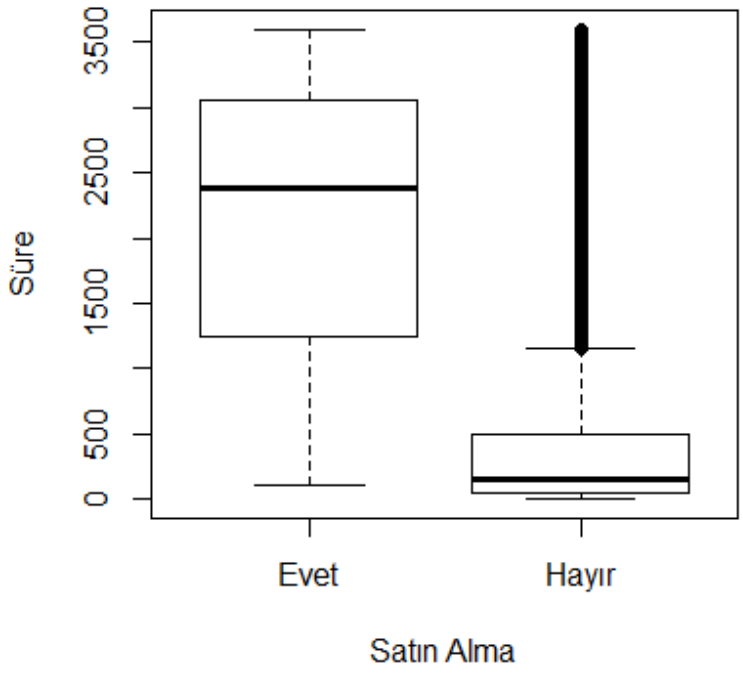



**Şekil 3-8 Saat dilimlerine göre satılan ürün sayısı**

Satın almaya bağlı olarak ziyaret edilen sayfa sayıları ve kalış sürelerine göre kutu grafikleri incelendiğinde satın alımda daha az sayfa gezildiği, buna karşın daha fazla süre harcandığı görülmektedir (Şekil 3.9 ve Şekil 3.10)




**Şekil 3-9 Satın almaya bağlı sayfa sayısı kutu grafiği**




**Şekil 3-10 Satın almaya bağlı harcanan süre kutu grafiği**

## 3.3 Veri Analizi

İnternetin artık yaşamın vazgeçilmez bir unsuru olması, e-ticaretin hızlı gelişimi ve başarısı, pazarlama yöneticilerinin karar almalarını kolaylaştırıcı modelleri kullanmalarını gerektirmektedir. Ekonomistler için rekabet, fiyat dağılımları, bilgisayar bilimleri için web site işletimlerinin optimizasyonu, internet kullanımı ve işlemlerden elde edilen verinin analizinde kullanılacak yeni algoritmaların geliştirilmesi, sosyologlar için de sosyal ağlar üzerine çalışmalar önemli hale gelmiştir.

Son dönemde internet ve e-ticaret üzerine yapılan modelleme çalışmaları; geleneksel perakendecilik, fiyatlama, promosyonlar, müşteri hizmetleri, satın alma davranışı ve müşteri sadakati üzerine yoğunlaşmıştır. Elektronik ticaret yapan şirketler web sitelerine potansiyel müşterileri (ziyaretçileri) çekmek, satın alma davranışını anlamak ve müşterilerinin sürekliliğini sağlamayı amaçlamaktadırlar (Bucklin, 2008).

Şirket kârlılığı için satışlar ve müşteri potansiyelinin tanımlanması kritik öneme sahiptir. Bu alan doğrudan, etkileşimli, hedef kitle ve veri tabanı pazarlama olarak bilinmekte ve pazarlama alanında araştırmacı ve uygulamacılar için büyük önem arz etmektedir (McCarty ve Hastak, 2007). Buradaki amaç, satın alım gibi müşterinin gelecek davranışının tahmini için müşterinin işlem ve davranışsal verisinin analiz edilmesidir (Hughes, 2005).

Şirket veri tabanlarının etkin kullanımı ve veri analiz yöntemlerindeki gelişmelerle müşteri veri analizi de önemli bir hale gelmiştir. Pazarlama çalışmalarında analiz yöntemi olarak karar ağaçları, yapay sinir ağları, logistik regresyon analizi gibi makine öğrenmesi yöntemlerini içeren müşteri tepki modelleri kullanılmaktadır (Bose ve Chen, 2009).

Müşteri tepki modellerinde "sınıf dengesizliği" gerçek uygulamalarda da ortaya çıkan önemli bir sorundur. Ancak literatürde bu problemin yeterince ilgi görmediği görülmektedir. Örneğin, satın alım olması durumu için "1", satın alım olmaması durumu için "0" değerinin atandığı iki kategorili sınıflandırma modelinin veri setinde "1"lerin yani satın alımların "0"lara oranla oldukça az olduğu gözlenmektedir. Sigorta dolandırıcılık tespiti, petrol sızıntısı tahmini, erken dönem doğum tahmini, sistem başarısızlık tahmini, afet tahmini gibi örnekler bu tür verilere örnek verilebilmektedir (Ling ve Li, 1998), (Weiss, 2004). Bu veri setlerinde veri analizi yöntemleri çoğunluk durumlarına göre taraflı davranma eğilimi gösterebilmekte ve azınlık durumları da çoğunluğa göre sınıflandırılabilmektedir. Bu problem müşteri tepki tahmini (McCarty ve Hastak, 2007) ve müşteri kayıp tahmini (Burez ve Van den Poel, 2009) gibi pazarlama verilerinde sıklıkla görülmektedir. McCarty ve Hastak (2007), %3'ten, Ling ve Li (1998) ise %1,5'tan daha az tepki oranına sahip veri setleriyle çalışmışlardır. Dengesizlikle uygun bir şekilde başa çıkılamazsa model, gerçek tepki oranının zayıf bir tahminini sunmaktadır. Daha açık bir ifadeyle, modelin tahmin doğruluğu artmakta ancak ilgi sınıfı (satın alma, sahtekârlık, ikna, vb.) hakkında yanlış bilgi vermektedir. Bu doğruluk oranının da, yanıltıcı olduğu sonucu ortaya çıkmaktadır.

Bu tez çalışmasında sınıflandırma yöntemi olarak kullanılan DVM de dengesiz veri durumunda karar sınırını azınlık sınıfına kaydırır. Çoğu veri noktası çoğunluk sınıfında sınıflanır. Bu da modelin zayıf olmasına neden olabilmektedir.

Sınıf dengesizliğiyle başa çıkabilmek için kullanılan başlıca yöntemler sınıflandırıcı değişimi (classifier change) ve yeniden örnekleme (resampling) yöntemleridir (Weiss, 2004). Sınıflandırıcıyı değişim modelinde taraflı tahminden kaçınmak için amaç fonksiyonu üzerinde farklı ağırlığa sahip tahminleyiciler oluşturulur (He ve Garcia, 2009). Yeniden örnekleme yönteminde ise, veriyi dengelemek için veri arttırılır (oversampling) ya da azaltılır (undersampling) (He ve Garcia, 2009), (Japkowicz, 2001).

Veri setini azaltan örnekleme (örnek küçültme) yönteminde veriyi dengelemek için çoğunluk sınıf verisi, tesadüfi olarak azınlık sınıfına eşit ya da katları olarak alınmaktadır. Yöntem, veri setinde sınıf dengesini kolaylıkla sağlayabilmektedir (Kim, Chae ve Olson, 2013).

Veri setini arttıran örnekleme (örnek büyütme) yönteminde ise azınlık sınıfından kopyalar üretilerek veri dengelenmek istenmekte, ancak bu da aşırı uyuma (overfitting) neden olabilmektedir (Chawla v.d., 2002), (Drummond ve Holte, 2003).

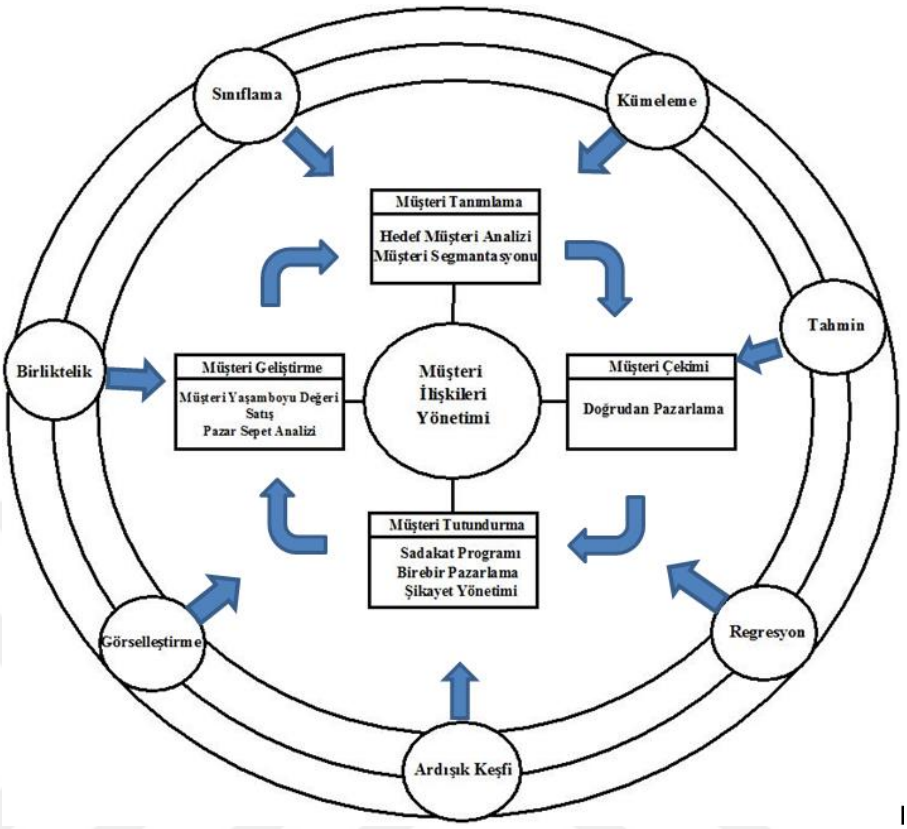
Japkowicz (2000), örnek büyütme ve örnek küçültme yöntemlerini değişik veri setleri için değerlendirmiş ve iki yöntemin de etkili olduğu sonucuna varmıştır. Örnek küçültme yöntemi, örnek büyütme yöntemine göre uygulanması daha kolay ve büyük veri setleri için daha uygun kabul edilmektedir (Drummond ve Holte, 2003), (Khoshgoftaar, Van Hulse ve Napolitano, 2011). Chawla v.d. (2002), Synthetic Minority Oversampling Technique (SMOTE) adıyla bir yöntem geliştirmişlerdir. Bu yöntem, hayalet dönüşümü ile yeni özellikler oluşturmaktadır. Her pozitif örnek için onun en yakın pozitif komşusu tanımlanmakta ve yeni pozitif örnekler oluşturularak, komşuları arasında tesadüfi olarak yer almaktadır. Burada problem, öğrenmiş sınırın pozitif özelliklere oldukça yakın olmasıdır (Provost ve Fawcett, 2001).

Ngai, Xiu ve Chau (2009) çalışmalarında yapay sinir ağları, karar ağaçları ve regresyon analizinin müşteri ilişkileri yönetiminde geniş olarak kullanıldığını

göstermiştir. Bu yöntemler müşteri segmentasyonu oluşturma ya da tepki modelinde de kullanılmaktadır (Bose ve Chen, 2009). Öte yandan MİY ve müşteri tepki modeli için DVM kullanımı çok seyrektir (Viaene v.d., 2001).

DVM ile karşılaştırma yöntemi olarak kullanılan logistik regresyon analizi ise tepki modellerine dört nedenden dolayı iyi uyarlanabilmektedir. 1) Logit modelleme, özellikle bireysel tüketici seviyesinde (Neslin v.d., 2006) pazarlama alanında sıkça kullanılmaktadır (Bucklin ve Gupta, 1992). 2) Logit yorumunun diğer yöntemlere göre (yapay sinir ağları gibi) kolaylığı önemli bir avantajdır. 3) İkna tahmini (Neslin v.d., 2006) ve kredi skorlama (Baesens v.d., 2003) gibi çalışmalarda iyi sonuçlar vermektedir. 4) Veri tabanı pazarlamada da diğer yöntemlere nazaran iyi sonuçlar vermektedir (Levin ve Zahavi, 1998).

Şekil 3.1'de MİY'de veri madenciliği yöntemleri üzerine sınıflandırma çerçevesinde bir gösterim gerçekleştirilmiştir. Bu çerçeve aynı zamanda Swift (2001), Kracklauer, Mills ve Dirk (2004), Parvatiyar ve Sheth (2001)'nin araştırmalarına dayanmakta, bu araştırmalarda MİY boyutları müşteri tanımlama, müşteri çekme, müşteri tutundurma ve müşteri potansiyelini genişletme olarak tanımlamaktadır. İlave olarak Ahmed (2004), Giraud-Carrieer ve Povel (2003), Mitra, Pal ve Mitra (2002) ve Shaw v.d. (2001) veri madenciliği modelleri olarak birliktelik, sınıflandırma, kümeleme, tahmin, regresyon, ardışıklık keşfi ve görselleştirmeyi tanımlamıştır. Sınıflandırma, müşteri tanımlamada önemli olup hedef müşteri analizi ve müşteri segmentasyonunda sınıflandırma yöntemlerinden yararlanılmaktadır.




**Şekil 3-11 Müşteri İlişkileri Yönetimi'nde Veri Madenciliği Yöntemleri için Sınıflandırma Çerçevesi (Ngai v.d. 2009:2592)**

Destek vektör makineleri ve logistik regresyon analizi için RapidMiner Studio 6 kullanılmıştır. RapidMiner, Java diliyle yazılmış, GNU altında açık kaynak kodlu bir veri madenciliği yazılım paketi olup, Excel, Access, Oracle, MySQL, SPSS gibi farklı veri kaynaklarına erişim sağlayan veri yükleme, dönüştürme, modelleme ve görselleştirmede güçlü bir araçtır. Veri temizlemeden sonraki adımlarda sınıflandırma yöntemleri RapidMiner aracılığıyla gerçekleştirilmiştir. Kurulan modellerin ekran çıktıları EK4'te verilmiştir.

Analizler İstanbul Üniversitesi Bilimsel Araştırmalar Proje Birimi (BAP) projesi ile tedarik edilen 16 GB RAM'e sahip Intel Xeon CPU 3.50 GHz işlemcili workstation ile yapılmıştır.

Analizde sınıflandırma yöntemi olarak destek vektör makineleri kullanılmıştır. Veri temizleme ve oturum tanımlama işlemleri sonucu veriden elde edilen nitelikler Tablo 3.7'de verilmiştir. Bunlar; kullanıcının sitede gezdiği **sayfa sayısı**, ziyaretçinin

sitede harcadığı **toplam süre**, oturumuna bağlı olarak sitede kaldığı **saat dilimi** [Sabah (06:00-11:59 arası), Öğle (12:00-17:59 arası), Akşam (18:00-23:59 arası), Gece (00:00-05:59 arası)], ziyaretçinin hangi **ay** siteye giriş yaptığı (Ocak, Şubat, Mart, Nisan, Mayıs, Haziran, Temmuz, Ağustos, Eylül, Ekim, Kasım, Aralık), siteye giriş yapılan **gün** (Pazartesi, Salı, Çarşamba, Perşembe, Cuma, Cumartesi, Pazar), hangi **ülke**den istek yaptığı (Türkiye ve diğer ülkeler olarak iki kategoride tanımlanmıştır) ve son olarak da ziyaretçinin ürün **satın alıp almadığına** (satın alım için Evet, satın almama için Hayır) ait niteliklerdir. Satın alma niteliği kategorik olarak tanımlanmasına karşın aynı zamanda niceldir. Web sitesi özellikli ürün sattığı için her satın alma bir ürüne denk gelmektedir.

**Tablo 3-7 Analizde kullanılan nitelikler**

| **Nitelikler** | **Özellikleri** |
|---|---|
| Sayfa Sayısı (sayfasayisi) | Nicel |
| Süre (sure) | Nicel |
| Saat Dilimi (saatdilimi) | 4 kategori:<br>Sabah (06:00-11:59)<br>Öğle (12:00-17:59)<br>Akşam (18:00-23:59)<br>Gece (00:00-05:59) |
| Ay (ay) | 12 kategori |
| Gün (gun) | 7 kategori |
| Ülke (ulke) | 2 kategori:<br>Türkiye<br>Diğer |
| Satın Alma Durumu (satinalim) | Kategorik<br>Satın alma için EVET<br>Satın almama için HAYIR |

Klasik destek vektör makineleri sınıflandırma modeli şu şekildedir:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

$x_i$ eğitim vektörleri, $\emptyset$ kernel fonksiyonu tarafaından daha yüksek boyutlu bir uzaya haritalanmaktadır. Kernel fonksiyonu, $K(x_i, x_j) \equiv (\emptyset(x_i)^T \emptyset(x_j))$ olup, analizde sigmoid fonksiyonu kullanılmıştır:

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + c)$$

DVM sınıflandırıcı ise,

$$sign(\sum_{i}^{l} \alpha_i y_i K(x, x_i) + b)$$

olarak tanımlanmaktadır.

91.095 oturuma karşılık 366 satış gerçekleşmiştir. Tepki oranı %1'in altında oldukça düşük bir yüzdeye sahiptir. Bu durun daha önce belirtildiği üzere veri dengesizliği problemini ortaya çıkarmaktadır. Veri dengesizliğini gidermek amacıyla yeniden örnekleme yönteminden yararlanılmıştır. Yeniden örnekleme yöntemlerinden örnek küçültme kullanılarak çoğunluk sınıfından (satın almama durumu), azınlık sınıfının (satın alma durumu) 1, 2, 3 ve 4 katına denk gelecek şekilde tesadüfi olarak alt örnek veri setleri oluşturulmuştur. Bu şekilde oluşturulan dört modelde sınıflandırıcı performansları karşılaştırılmıştır. Bu dört modele ilave olarak bu kez azınlık sınıfından örnekler azaltılarak, azınlık sınıfın 1, 2, 3 ve 4 katına eşit olacak şekilde çoğunluk sınıfından örnekler çekilerek de analizler tekrarlanmıştır. Toplamda 20 model ile analizler gerçekleştirilmiştir.

Destek vektör makineleri ve logistik regresyon için model oluşturulurken ilk önce veri seti satın alma (pozitif sınıf, 366 oturum) ve satın almanın gerçekleşmediği

(negatif sınıf, 90.729 oturum) durumlara göre ikiye ayrılmıştır. Negatif sınıftan pozitif sınıfın 1, 2, 3 ve 4 katına denk gelecek şekilde tesadüfi olarak örnek seçilmiştir. Negatif sınıftan örnek seçilirken frekans aralığı, tarihlere göre sıralandırılmış, yıllara göre 3 gruba ayrılmıştır. Üç yıllık veriye sahip olmamızın yanı sıra üç tam yıl olmadığından (verinin ait olduğu dönem Mayıs 2011 ile Ağustos 2013 arasıdır) yılların sahip oldukları haftalara göre bir gruplamaya gidilmiştir. Aynı şekilde satın alımın gerçekleştiği pozitif sınıf da üç gruba ayrılarak örnekleme oranları hesaplanmıştır. Tablo 3.8'de negatif sınıfa ve Tablo 3.9'da pozitif sınıfa ait oluşturulan gruplar ve her model için çekilen tesadüfi örnek sayısı verilmiştir.

**Tablo 3-8 Negatif Sınıfa Ait Örnek Oranları**

| Model | Negatif Örnek Sayısı | Gruplar (Oranlar) 1 (1-30.074) | 2 (30.075-72.325) | 3 (72.326-90.729) |
|---|---|---|---|---|
| Model 1a | 366 | 0,0035 | 0,0039 | 0,0052 |
| Model 2a | 732 | 0,0070 | 0,0078 | 0,0100 |
| Model 3a | 1098 | 0,0140 | 0,0120 | 0,0150 |
| Model 4a | 1464 | 0,0140 | 0,0160 | 0,0200 |
| Model 4b | 1200 | 0,0120 | 0,0130 | 0,0170 |
| Model 3b | 900 | 0,0086 | 0,0096 | 0,0130 |
| Model 4c | 800 | 0,0076 | 0,0086 | 0,0110 |
| Model 2b-3c | 600 | 0,0058 | 0,0064 | 0,0080 |
| Model 2c-4d | 400 | 0,0038 | 0,0043 | 0,0050 |
| Model 1b-3d | 300 | 0,0029 | 0,0032 | 0,0040 |
| Model 1c-2d-4e | 200 | 0,0019 | 0,0021 | 0,0030 |
| Model 3e | 150 | 0,0014 | 0,0016 | 0,0020 |
| Model 1d-2e | 100 | 0,0010 | 0,0010 | 0,0010 |
| Model 1e | 50 | 0,0005 | 0,0005 | 0,0007 |

Tablo 3.8'de 1. Grup örnekler 2011 yılına ait 30.074 oturum arasından, 2. Grup örnekler 2012 yılına ait 42.251 oturum arasından ve 3. Grup örnekler 2013 yılına ait 18.404 oturum arasından ilgili oranlara göre çekilmiştir. Aynı şekilde pozitif örnekler de Tablo 3.9'da gösterildiği üzere satın alımların gerçekleştirildiği oturumlar göz önüne alınarak oluşturulmuştur. 1. Grup örnekler 2011 yılına ait 103 oturumu, 2. Grup 2012 yılına ait 188 oturumu ve son olarak 3. Grup da 2013 yılındaki 75 oturumu tanımlamaktadır.

**Tablo 3-9 Pozitif Sınıfa Ait Örnek Oranları**

| Model | Pozitif Örnek Sayısı | Gruplar | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| | | (1-103) | (104-291) | (292-366) |
| Model 1a-2a-3a-4a | 366 | 1 | 1 | 1 |
| Model 1b-2b-3b-4b | 300 | 0,83 | 0,72 | 1 |
| Model 1c-2c-3c-4c | 200 | 0,56 | 0,48 | 0,7 |
| Model 1d-2d-3d-4d | 100 | 0,28 | 0,24 | 0,35 |
| Model 1e-2e-3e-4e | 50 | 0,14 | 0,12 | 0,17 |

Oluşturulan her gruptan tablolardaki oranlarda tesadüfi örnekler çekilerek analiz aşamasına geçilmiştir (EK4'de RapidMiner süreci ekran çıktılarıyla birlikte anlatılmıştır).

Veri setlerinden nitelik seçimi ve normalizasyon işlemleri gerçekleştirildikten sonra verinin yüzde 80'i eğitim, yüzde 20'si test seti olarak ayrılmıştır.

Doğrusal, polinominal, radyal tabanlı ve sigmoid kernel tipleri denenmiş, en iyi sonucu sigmoid kernel vermiştir. Parametre değerleri C (hata teriminin ceza parametresi (C>0)) ve gamma ($\gamma$) için optimum değerler RapidMiner'da bulunan evrimsel parametre optimizasyon tekniği kullanılarak belirlenmiştir. Aşırı uyumdan kaçınmak, tesadüfiliği artırmak amacıyla çapraz değerleme yöntemi kullanılmıştır. Eğitim seti 10 eşit alt sete bölünmüş ve setlerden 9 adedi eğitim, 1 adeti test seti olarak kullanılmıştır. Böylelikle 10 farklı eğitim ve test seti için parametre optimizasyon tekniği uygulanmıştır. Elde edilen parametreler modelin parametreleri olarak DVM eğitiminde kullanılmıştır. Nihai olarak da modeller test verisine uygulanarak sınıflandırıcı performansları elde edilmiştir. Çoğunluk ve azınlık sınıfın farklı oranlara göre oluşturulan modeller için hem DVM hem de logistik regresyon analiziyle elde edilen sınıflandırma sonuçları karşılaştırılmıştır.

### 3.3.1 Model 1

Birinci modelde bütün nitelikler dahil olmak üzere pozitif sınıf ile aynı sayıda negatif sınıftan örnekler seçilerek analizler gerçekleştirilmiştir. Model 1a'da 366 pozitif örneğe (satın alım durumu) karşılık negatif sınıftan tesadüfi olarak 366 örnek

çekilmiştir. Model 1'de tepki oranı %50'dir. Model 1b, Model 1c, Model 1d ve Model 1e'de ise pozitif sınıfa ait örnekler kademe kademe azaltılarak yine aynı sayıda negatif sınıftan örnekler çekilerek analizler gerçekleştirilmiştir. Bütün modellerde örneklerin %80'i eğitim veri seti, %20'si test seti olarak ayrılmıştır.

**Tablo 3-10 Model 1 için örnek sayıları**

| Model | Pozitif Sınıftan Çekilen Örnek Sayısı | Negatif Sınıftan Çekilen Örnek Sayısı |
|---|---|---|
| Model1a | 366 | 366 |
| Model1b | 300 | 300 |
| Model1c | 200 | 200 |
| Model1d | 100 | 100 |
| Model1e | 50 | 50 |

**Model 1a:**

Bu modelde pozitif sınıftan 366, negatif sınıftan da 366 örnek çekilerek DVM ve Logistik Regresyon yöntemleri uygulanmıştır. Modelin test veri seti üzerinde aşağıdaki performans değerleri elde edilmiştir. Bunlar; **doğruluk** (doğru tahminlerin tüm sınıflara olan oranı), **kappa (**gerçek sınıfla tahminlerin uzlaşma ölçüsü olup 0,40'dan büyük olması beklenmektedir), **duyarlılık** (gerçek pozitif oranı, doğru sınıflandırılan pozitif örneklerin toplam pozitif örnek sayısına oranı), **belirlilik** (doğru sınıflandırılan negatif örneklerin toplam negatif örneklere oranı), **kesinlik** (pozitif öngörü değeri, doğru sınıflandırılan pozitif örneklerin toplam pozitif tahmin edilen örneklere oranı) ve **F** (kesinlik ve duyarlılık ölçülerinin harmonik ortalaması)'dir.

Satın alımın tahmini önceliğe sahip olduğundan özellikle duyarlılık ve kesinlik ölçülerinin değerlerine göre karşılaştırma yapılmıştır.

DVM Sonuçları:

**Tablo 3-11 Model 1a - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 62 | 8 | 0,8493 | 0,8889 | 0,8857 | 0,869 | 0,763 | 0,867118 |
| Tahmin Negatif | 11 | 64 | | | | | | |

Tepki oranı %50 olan ilk model için doğruluk oranı %86,9 olarak hesaplanmıştır. Kappa değeri de %76,3 ile yüksektir. Gerçek pozitif sınıflandırma oranı %85 civarıdır. Vektör ağırlıklarına bakıldığında (EK5'de ayrıntılı olarak verilmiştir) süre (665,648), sayfa sayısı (663,848) sınıflandırmayı negatif yönde etkilerken, çarşamba (584,708) ve Türkiye (386,241) sınıflandırmayı pozitif yönde etkileyen, sınıflandırma tahmininde öne çıkan nitelikler olmuştur.

Logistik Regresyon (LR) sonuçları:

**Tablo 3-12 Model 1a - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 59 | 7 | 0,8082 | 0,9054 | 0,8939 | 0,8571 | 0,712 | 0,848893 |
| Tahmin Negatif | 14 | 67 | | | | | | |

Logistik regresyon ile elde edilen sonuçlara bakıldığında doğruluk oranı %85,71 ve F-skoru, %84,89 ile oldukça iyi bir değer olarak hesaplanmıştır. DVM modeline göre negatif sınıfları daha doğru tahmin etmiştir (belirlilik %90,54).

Model çıktısına bakıldığında 0.05 anlamlılık düzeyinde modelde kalan nitelikler E%'te verilmiştir.

**Model 1b:**

Bu modelde pozitif örnek sayısı azaltılarak ve aynı sayıda negatif örnek seçilerek analizler gerçekleştirilmiştir. 300 pozitif, 300 negatif örnek çekilmiştir.

DVM sonuçları:

**Tablo 3-13 Model 1b - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 56 | 9 | 0,875 | 0,8475 | 0,8615 | 0,8618 | 0,672 | 0,868198 |
| Tahmin Negatif | 8 | 50 | | | | | | |

Negatif yönde en büyük etkiyi süre (453,928), sayfa sayısı (298,438), ülke = Diğer (263,841) nitelikleri gösterirken, pozitif yönde pazartesi (269,549), sabah (273,280), ağustos (265,523) nitelikleri etkili olmuştur. Duyarlılık ve kesinlik değerlerinin harmonik ortalaması olan F-skoru ise %86,82 gibi yüksek bir değerdir.

LR sonuçları:

**Tablo 3-14 Model 1b - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 46 | 11 | 0,7541 | 0,8136 | 0,807 | 0,7833 | 0,664 | 0,779654 |
| Tahmin Negatif | 15 | 48 | | | | | | |

Logistik regresyonda pozitif sınıflar DVM'ye göre daha düşük oranda tahmin edilmiştir. F-skoru %77,96 ile DVM'nin gerisindedir.

**Model 1c:**

Bu modelde pozitif örnek azaltılarak ve yine aynı sayıda negatif örnek seçilerek analizler gerçekleştirilmiştir. 200 pozitif örneğe karşılık 200 negatif örnek seçilmiştir.

DVM sonuçları:

**Tablo 3-15 Model 1c - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 38 | 7 | 0,8636 | 0,825 | 0,844 | 0,8452 | 0,683 | 0,853688 |
| Tahmin Negatif | 6 | 33 | | | | | | |

Pozitif sınıf tahmin gücü, F-skoru %85,37 ve Kappa değeri 0,683 olarak hesaplanmıştır. Vektör ağırlıklarına bakıldığında; süre (343,750), sayfa sayısı (259,288), gece yarısı (114,034) ve ülke = Diğer (108,594) negatif yönde en etkili nitelikler olmuştur. Pozitif yönde ise ocak (188,583), aralık (168,532), ülke = Türkiye (108,594) en etkili nitelikler olmuştur.

LR sonuçları:

**Tablo 3-16 Model 1c - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 32 | 10 | 0,75 | 0,7442 | 0,7619 | 0,747 | 0,701 | 0,755903 |
| Tahmin Negatif | 11 | 30 | | | | | | |

Logistik regresyonda pozitif sınıflandırma gücü, DVM'nin gerisinde kalmıştır (%75.59). DVM'yle benzer olarak sayfa sayısı niteliğinin negatif etkili olduğu gözükmektedir.

**Model 1d:**

Bu modelde pozitif örnek azaltılarak aynı sayıda negatif örnek seçilerek analizler gerçekleştirilmiştir. Pozitif sınıftan 100, negatif sınıftan 100 örnek seçilmiştir.

DVM sonuçları:

**Tablo 3-17 Model 1d - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 19 | 4 | 0,8636 | 0,8 | 0,8261 | 0,8333 | 0,701 | 0,844434 |
| Tahmin Negatif | 3 | 16 | | | | | | |

DVM ile pozitif sınıf tahmini (F) %84,44 gibi yüksek bir değerle gerçekleşmiştir. Vektör ağırlıklarına bakıldığında haziran (84,188), Türkiye (53,626), akşam (59,880), salı (46,580) pozitif yönde sınıflandırmayı en çok etkileyen nitelikler, sayfa sayısı (81,967), cuma (74,417), ülke = Diğer (63,626) ise negatif etkileyen nitelikler olmuştur.

LR sonuçları:

**Tablo 3-18 Model 1d - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 21 | 4 | 0,9545 | 0,7895 | 0,84 | 0,878 | 0,699 | 0,893597 |
| Tahmin Negatif | 1 | 15 | | | | | | |

LR ile pozitif sınıflandırma gücü DVM'ye göre daha fazladır. F değeri % 89,36 ile DVM'nin ilerisindedir.

**Model 1e:**

Bu modelde pozitif sınıftan 50, negatif sınıftan 50 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-19 Model 1e - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 9 | 2 | 0,8182 | 0,8 | 0,8182 | 0,8095 | 0,578 | 0,8182 |
| Tahmin Negatif | 2 | 8 | | | | | | |

DVM sonuçlarına göre pozitif sınıflandırma gücü %81,82'dir. Vektör ağırlıkları incelendiğinde cumartesi (42,066), temmuz (28,452), mart (20,558) pozitif yönde, pazartesi (29,398), kasım (19,661), öğle (15,606) negatif yönde en fazla etkileyen niteliklerdir.

LR sonuçları:

**Tablo 3-20 Model 1e - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 9 | 0 | 0,8182 | 1 | 1 | 0,9048 | 0,773 | 0,900011 |
| Tahmin Negatif | 2 | 10 | | | | | | |

LR, DVM'ye göre pozitif sınıfları (F=%90) ve negatif sınıfları daha doğru tahmin (%100) etmiştir. Modelde kalan 0,05 anlamlılık seviyesindeki nitelikler ile kurulan model EK5'te verilmiştir.

### 3.3.2 Model 2

Bu modelde pozitif sınıfların iki katı kadar negatif sınıf örneği seçilerek analiz gerçekleştirilmiştir. Model 2 için tepki oranı %33,33'tür.

**Tablo 3-21 Model 2 için örnek sayıları**

| Model | Pozitif Sınıftan Çekilen Örnek Sayısı | Negatif Sınıftan Çekilen Örnek Sayısı |
|---|---|---|
| Model2a | 366 | 732 |
| Model2b | 300 | 600 |
| Model2c | 200 | 400 |
| Model2d | 100 | 200 |
| Model2e | 50 | 100 |

**Model 2a:**

Bu modelde pozitif sınıftan 366, negatif sınıftan 732 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-22 Model 2a - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 53 | 11 | 0,726 | 0,9241 | 0,8281 | 0,8578 | 0,712 | 0,773696 |
| Tahmin Negatif | 20 | 134 | | | | | | |

Pozitif sınıf tahmin gücü %77,37, negatif sınıfın tahmin gücü ise %92,41 olarak gerçekleşmiştir. Vektör ağırlıkları incelendiğinde, sınıflandırmayı; süre (1834,817), ülke=Türkiye (1173,633), çarşamba (1075,390), ocak (675,053), sayfa

sayısı (649,785) pozitif yönde, ülke = diğer (1173,633), temmuz (530,094), aralık (339,388), salı (288,617) negatif yönde etkileyen niteliklerdir.

LR sonuçları:

**Tablo 3-23 Model 2a - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 54 | 12 | 0,7397 | 0,8947 | 0,8182 | 0,8342 | 0,722 | 0,776972 |
| Tahmin Negatif | 19 | 102 | | | | | | |

LR, pozitif sınıflandırmada DVM ile benzer sonuçlar ortaya koymaktadır. Pozitif sınıfların tahmin gücü %77,69'dur.

**Model 2b:**

Bu modelde pozitif sınıftan 300, negatif sınıftan 600 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-24 Model 2b - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 42 | 9 | 0,7 | 0,925 | 0,8235 | 0,85 | 0,631 | 0,756744 |
| Tahmin Negatif | 18 | 111 | | | | | | |

Pozitif sınıf tahmini %75,67 olup, negatif sınıf tahmini daha yüksektir (%92,5). Vektör ağırlıkları; ülke = Türkiye (1167,113), haziran (744,558), çarşamba (591,198), sayfa sayısı (549,331) ile pozitif yönde; ülke = Diğer (1167,113), kasım (751,372), gece yarısı (418,033) nitelikleri ise negatif yönde etkili olmuştur.

LR sonuçları:

**Tablo 3-25 Model 2b - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 45 | 13 | 0,7377 | 0,8908 | 0,7759 | 0,8389 | 0,652 | 0,756318 |
| Tahmin Negatif | 16 | 106 | | | | | | |

Pozitif sınıf tahmini %75,63 ile DVM ile benzerlik göstermektedir, ancak negatif sınıf tahmininde LR, DVM'nin gerisine düşmüştür (%89,08).

**Model 2c:**

Bu modelde pozitif sınıftan 200, negatif sınıftan 400 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-26 Model 2c - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 33 | 10 | 0,75 | 0,875 | 0,7674 | 0,8306 | 0,67 | 0,7586 |
| Tahmin Negatif | 11 | 70 | | | | | | |

Pozitif sınıflandırma oranı %75,86, doğruluk oranı ise %83,06'dır. Vektör ağırlıkları incelendiğinde süre (589,360), pazartesi (430,482), ülke = Türkiye (413,817), sayfa sayısı (396,860) sınıflandırmayı pozitif etkileyen niteliklerdir. nisan (529,979), ülke = Diğer (413,817), temmuz (275,952), sınıflandırmayı negatif yönde etkileyen niteliklerdir.

LR sonuçları:

**Tablo 3-27 Model 2c - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 32 | 7 | 0,7442 | 0,9114 | 0,8205 | 0,8525 | 0,61 | 0,78049 |
| Tahmin Negatif | 11 | 72 | | | | | | |

DVM'ye göre benzer sınıflandırma oranları vermiştir.

**Model 2d:**

DVM sonuçları:

Bu modelde pozitif sınıftan 100, negatif sınıftan 200 örnek tesadüfi olarak çekilmiştir.

**Tablo 3-28 Model 2d - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 14 | 3 | 0,6364 | 0,925 | 0,8235 | 0,8226 | 0,668 | 0,717961 |
| Tahmin Negatif | 8 | 37 | | | | | | |

Vektör ağırlıklarına bakıldığında aralık (182,894), ülke = Türkiye (113,551), pazartesi (87,505), cumartesi (84,538), süre (91,647) sınıflandırmayı pozitif yönde, cuma (130,819), ülke = Diğer (113,551), temmuz (102,343) negatif yönde etkileyen niteliklerdir.

LR sonuçları:

**Tablo 3-29 Model 2d - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 17 | 2 | 0,7727 | 0,9487 | 0,8947 | 0,8852 | 0,756 | 0,829237 |
| Tahmin Negatif | 5 | 37 | | | | | | |

Pozitif sınıflandırma oranı DVM'ye göre daha yüksek olup F değeri %82,92 olarak hesaplanmıştır.

**Model 2e:**

Bu modelde pozitif sınıftan 50, negatif sınıftan 100 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-30 Model 2e - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 9 | 3 | 0,8182 | 0,85 | 0,75 | 0,8387 | 0,576 | 0,782617 |
| Tahmin Negatif | 2 | 17 | | | | | | |

Vektör ağırlıkları incelendiğinde süre (52,658), perşembe (41,447), akşam (48,395) ve mart (22,827) pozitif yönde, öğle (50,006), nisan (40,756) ve cumartesi (21,678) nitelikleri ise negatif yönde etkilidir.

LR sonuçları:

**Tablo 3-31 Model 2e - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 8 | 0 | 0,7273 | 1 | 1 | 0,9 | 0,763 | 0,842124 |
| Tahmin Negatif | 3 | 19 | | | | | | |

Pozitif sınıf tahmin gücü DVM'den daha fazla hesaplanmıştır. F-skoru %84,21'dir.

### 3.3.3 Model 3

Üçüncü modelde pozitif sınıfın üç katı büyüklüğünde negatif sınıftan örnek seçilerek analiz gerçekleştirilmiştir. Tepki oranı %25'tir.

**Tablo 3-32 Model 3 için örnek sayıları**

| Model | Pozitif Sınıftan Çekilen Örnek Sayısı | Negatif Sınıftan Çekilen Örnek Sayısı |
|---|---|---|
| Model3a | 366 | 1098 |
| Model3b | 300 | 900 |
| Model3c | 200 | 600 |
| Model3d | 100 | 300 |
| Model3e | 50 | 150 |

**Model 3a:**

Bu modelde pozitif sınıftan 366, negatif sınıftan 1098 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-33 Model 3a - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 53 | 10 | 0,726 | 0,9539 | 0,8413 | 0,8966 | 0,667 | 0,779409 |
| Tahmin Negatif | 20 | 207 | | | | | | |

Vektör ağırlıkları incelendiğinde süre (5383,642), sayfa sayısı (3233,743), ülke = Türkiye (3139,305), haziran (1343,452) pozitif yönde, ülke = Diğer (3139,305), akşam (1276,787), cuma (1052,279), cumartesi (888,281) nitelikleri ise negatif yönde etkilidir.

LR sonuçları:

**Tablo 3-34 Model 3a - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 54 | 12 | 0,7397 | 0,9554 | 0,8182 | 0,9094 | 0,648 | 0,776972 |
| Tahmin Negatif | 19 | 257 | | | | | | |

Performans değerleri DVM ile benzer sonuçlar üretmiştir.

**Model 3b:**

Bu modelde pozitif sınıftan 300, negatif sınıftan 900 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-35 Model 3b - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 44 | 19 | 0,7333 | 0,8927 | 0,6984 | 0,8523 | 0,663 | 0,715425 |
| Tahmin Negatif | 16 | 158 | | | | | | |

Vektör ağırlıkları incelendiğinde süre (3488,440), ülke = Türkiye (2044,862), çarşamba (1711,642), sayfa sayısı (1706,352) pozitif yönde, ülke = Diğer (2044,862), kasım (1179,747), pazar (1087,220), cumartesi (1003,048) nitelikleri ise negatif yönde etkilidir.

LR sonuçları:

**Tablo 3-36 Model 3b - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 36 | 12 | 0,5902 | 0,933 | 0,75 | 0,8458 | 0,65 | 0,660573 |
| Tahmin Negatif | 25 | 167 | | | | | | |

LR, pozitif tahminde DVM'ye göre daha düşük bir F değerine sahiptir (%66).

**Model 3c:**

Bu modelde pozitif sınıftan 200, negatif sınıftan 600 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-37 Model 3c - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 29 | 10 | 0,6591 | 0,9174 | 0,7436 | 0,8485 | 0,646 | 0,698805 |
| Tahmin Negatif | 15 | 111 | | | | | | |

Doğruluk oranı 0,8485 olan modelin vektör ağırlıklarına bakıldığında; süre (1996,191), sayfa sayısı (1031,187), pazartesi (973,736), ülke = Türkiye (758,910) nitelikleri pozitif yönde, ülke = Diğer (758,910), salı (625,649), pazar (741,099), nisan (498,976) nitelikleri ise negatif yönde etkilidir.

LR sonuçları:

**Tablo 3-38 Model 3c - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 26 | 9 | 0,6047 | 0,9244 | 0,7429 | 0,8395 | 0,645 | 0,666714 |
| Tahmin Negatif | 17 | 110 | | | | | | |

Performans değerleri DVM ile benzerlik göstermektedir.

**Model 3d:**

Bu modelde pozitif sınıftan 100, negatif sınıftan 300 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-39 Model 3d - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 15 | 2 | 0,6818 | 0,9661 | 0,8824 | 0,8889 | 0,656 | 0,769237 |
| Tahmin Negatif | 7 | 57 | | | | | | |

Vektör ağırlıkları incelendiğinde, süre (664,885), sayfa sayısı (457,279), öğle (144,993), ülke = Türkiye (148,771) niteliklerinin pozitif yönde; pazar (168,387), ülke = Diğer (148,771), salı (92,311), şubat (83,427) niteliklerinin ise negatif yönde etkili olduğu görülmektedir.

LR sonuçları:

**Tablo 3-40 Model 3d - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 14 | 1 | 0,6364 | 0,9831 | 0,9333 | 0,8889 | 0,599 | 0,756772 |
| Tahmin Negatif | 8 | 58 | | | | | | |

**Model 3e:**

Bu modelde pozitif sınıftan 50, negatif sınıftan 150 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-41 Model 3e - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 8 | 1 | 0,7273 | 0,9667 | 0,8889 | 0,9024 | 0,672 | 0,800021 |
| Tahmin Negatif | 3 | 29 | | | | | | |

Doğruluk oranının %90,24 olduğu modelde, süre (149,225), sayfa sayısı (68,960), salı (50,783), mart (38,598) nitelikleri pozitif yönde, gece yarısı (60,820), nisan (70,494), perşembe (34,548) ve temmuz (29,862) negatif yönde etkilidir.

LR sonuçları:

**Tablo 3-42 Model 3e - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 8 | 2 | 0,7273 | 0,9333 | 0,8 | 0,878 | 0,674 | 0,76192 |
| Tahmin Negatif | 3 | 28 | | | | | | |

### 3.3.4 Model 4

Son modelde azınlık sınıfın dört katı olarak negatif örnek seçilmiştir. Tepki oranı %20'dir.

**Tablo 3-43 Model 4 için örnek sayıları**

| Model | Pozitif Sınıftan Çekilen Örnek Sayısı | Negatif Sınıftan Çekilen Örnek Sayısı |
|---|---|---|
| Model4a | 366 | 1464 |
| Model4b | 300 | 1200 |
| Model4c | 200 | 800 |
| Model4d | 100 | 400 |
| Model4e | 50 | 200 |

**Model 4a:**

Bu modelde pozitif sınıfın tamamı ve negatif sınıftan 1464 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-44 Model 4a - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 55 | 14 | 0,7534 | 0,9522 | 0,7971 | 0,9126 | 0,634 | 0,774634 |
| Tahmin Negatif | 18 | 279 | | | | | | |

Vektör ağırlıkları incelendiğinde süre (12095,604), sayfa sayısı (7681,074), ülke = Türkiye (4091,640), pazartesi (1913,492) pozitif yönde, ülke = Diğr (4091,640), mayıs (2027,181), akşam (1727,347), cumartesi (1777,597) nitelikleri ise negatif yönde etkilidir

LR sonuçları:

**Tablo 3-45 Model 4a - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 42 | 15 | 0,5753 | 0,9628 | 0,7368 | 0,9034 | 0,608 | 0,646111 |
| Tahmin Negatif | 31 | 388 | | | | | | |

**Model 4b:**

Bu modelde pozitif sınıftan 300, negatif sınıftan 1200 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-46 Model 4b - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 36 | 14 | 0,6 | 0,9414 | 0,72 | 0,8729 | 0,643 | 0,654545 |
| Tahmin Negatif | 24 | 225 | | | | | | |

Negatif örnek sayısı arttıkça pozitif sınıfı tahmin oranı düşmektedir. Bu model için F-skor değeri %65,45'te kalmıştır. Süre (8539,169), sayfa sayısı (5360,566), ülke = Türkiye (2916,074) ve sabah (2033,573) pozitif sınıfı en çok etkileyen niteliklerdir. Ülke = Diğer (2916,074), akşam (1847,227), gece yarısı (1110,427) ve salı (1096,614) nitelikleri ise negatif yönde etkilidir.

LR sonuçları:

**Tablo 3-47 Model 4b - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 39 | 10 | 0,6393 | 0,9582 | 0,7959 | 0,8933 | 0,599 | 0,709056 |
| Tahmin Negatif | 22 | 229 | | | | | | |

DVM ile benzer olarak LR'de de pozitif sınıflandırma başarısı düşmüştür.

**Model 4c:**

Bu modelde pozitif sınıftan 200, negatif sınıftan 800 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-48 Model 4c - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 28 | 6 | 0,6364 | 0,962 | 0,8235 | 0,8911 | 0,608 | 0,717961 |
| Tahmin Negatif | 16 | 152 | | | | | | |

Süre (4161,522), sayfa sayısı (2353,143), ülke = Türkiye (1538,579), pazartesi (908,792) pozitif yönde, ülke = Diğer (1538,579), pazar (1309,382), gece yarısı (1004,380) ve nisan (692,310) negatif yönde etkileyen niteliklerdir.

LR sonuçları:

**Tablo 3-49 Model 4c - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 30 | 5 | 0,75 | 0,9686 | 0,8571 | 0,9246 | 0,564 | 0,799981 |
| Tahmin Negatif | 10 | 154 | | | | | | |

**Model 4d:**

Bu modelde pozitif sınıftan 100, negatif sınıftan 400 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-50 Model 4d - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 11 | 5 | 0,5 | 0,9375 | 0,6875 | 0,8431 | 0,682 | 0,578947 |
| Tahmin Negatif | 11 | 75 | | | | | | |

Süre (893,077), sayfa sayısı (482,491), ülke = Türkiye (271,570), mayıs (225,898) nitelikleri pozitif yönde en çok etkili olan, temmuz (283,817), ağustos (275,315), ülke = Diğer (271,570), gece yarısı (228,951) nitelikleri de negatif yönde etkilidir.

LR sonuçları:

**Tablo 3-51 Model 4d - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 16 | 8 | 0,7273 | 0,9 | 0,6667 | 0,8627 | 0,664 | 0,695683 |
| Tahmin Negatif | 6 | 72 | | | | | | |

LR ile DVM'den daha yüksek pozitif sınıf tahmini elde edilmiştir. F-skor değeri %69,57'dir.

**Model 4e:**

Bu modelde pozitif sınıftan 50, negatif sınıftan 200 örnek tesadüfi olarak çekilmiştir.

DVM sonuçları:

**Tablo 3-52 Model 4e - DVM için performans değerleri**

| Sınıf | Gerçek Pozitif | Gerçek Negatif | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 7 | 1 | 0,6364 | 0,975 | 0,875 | 0,902 | 0,611 | 0,736866 |
| Tahmin Negatif | 4 | 39 | | | | | | |

Süre (266,616), sayfa sayısı (128,201), aralık (94,983) ve çarşamba (93,584) sınıflandırmada pozitif yönde en çok etkili niteliklerdir. Kasım (58,730), eylül (54,782), pazar (54,164) ve nisan (50,562) negatif yöndeki niteliklerdir.

LR sonuçları:

**Tablo 3-53 Model 4e - LR için performans değerleri**

| Sınıf | Gerçek Pozitif (Satış Var) | Gerçek Negatif (Satış Yok) | Duyarlılık | Belirlilik | Kesinlik | Doğruluk | Kappa | F |
|---|---|---|---|---|---|---|---|---|
| Tahmin Pozitif | 8 | 1 | 0,7273 | 0,975 | 0,8889 | 0,9216 | 0,694 | 0,800021 |
| Tahmin Negatif | 3 | 39 | | | | | | |

DVM'den farklı olarak pozitif sınıf tahmin değeri daha fazladır. F-skor değeri %80 ve doğruluk oranı %92,16'dir.

Destek vektör makineleri ve logistik regresyon ile cevap oranı 1, 2, 3 ve 4 kat olacak şekilde modeller oluşturulmuştur. Tablo 3.54'te destek vektör makineleri (DVM) ve logistik regresyon analizi (LR)'nin doğruluk, kesinlik, duyarlılık ve belirlilik, kappa, F-skor değerleri verilmiştir.

Pozitif sınıfların doğru sınıflandırma yüzdesi, tepki oranı yüzde 50 olan modellerde en yüksek seviyeye ulaşmıştır. Negatif örnek sayısı arttıkça pozitif sınıf tahminleri düşmektedir. Örnek azaltmayla yapılan analizler (Model 1a, 2a, 3a, 4a) sonrasında pozitif sınıftan da örnek azaltılarak tekrarlanmıştır. Böylelikle tepki oranlarına göre değişimler gözlenmiştir. Tablo 3.54'te özet olarak verildiği üzere hem logistik regresyon hem de destek vektör makineleri için en iyi pozitif sınıf tahmini negatif örnek ile pozitif örneğin eşdeğer olduğu modellerde sağlanmış, negatif örnek sayısı artıkça (tepki oranı azaldıkça) pozitif sınıf tahmin gücü azalmıştır.

Düzenlenen verilerden, DVM ve LR ile kurulan modellerin hem eğitimi hem testi için rassal olarak örnek veri setleri çekilmiştir ve bu işlem her iki yöntem için de 5'er kez yinelenerek farklı rassal veri setleri üzerinde denemeler yapılmıştır. Analizler yapıldıktan sonra ise 6 adet farklı performans ölçüm kriterine göre model değerlendirmeleri elde edilmiş ve rassal veri setleri için bulunan her bir performans ölçümünün ortalama değeri hesaplanmıştır.

Sonuçlara bakıldığında (Tablo 3.54) farklı performans ölçümleri için DVM ve LR modellerinin üstünlüklerinin de farklılık gösterdiği görülmektedir. Örneğin %50 tepki oranı için sonuçlara bakıldığında doğruluk, F skoru ve duyarlılık ölçümlerine göre DVM; kappa, belirlilik ve kesinlik ölçümlerine göre ise LR üstünlük göstermektedir. % 33 tepki oranında tüm performans ölçümlerinde LR üstünlük sağlarken, %25 oranında belirlilik haricinde diğer tüm performans ölçümlerinden DVM üstünlük sağlamıştır. Son olarak %20 tepki oranında ise kappa ve belirlilik ölçümlerinde DVM; doğruluk, F skoru, duyarlılık ve kesinlik ölçümlerinde ise LR üstünlük sağlamıştır.

Dört farklı tepki oranına göre (%50, %33, %25, %20) her bir performans ölçümü için elde edilen ortalama değerlerin de ortalaması alındığında ise, hem DVM hem de LR için en iyi sonucun yüzde 50 tepki oranı ile oluşturulan model olduğu ve DVM ile LR performans ölçümlerinin (0,818) birbirine eşit olduğu görülmektedir. Bir başka ifadeyle, söz konusu ikili sınıflandırma görevi için her iki sınıfa ait örnek sayılarının eşit olduğu durumda hem DVM hem de LR ile kurulan modellerde en iyi sınıflandırma sonuçlarına ulaşılmıştır.

Yukarıda özetlenen sonuçlar ışığında, %50 tepki oranı ile kurulan modelin analiz için tercih edilmesi gerektiği açıktır. Buna göre DVM ile %50 tepki oranı için kurulan modelde tüm rassal veri setleri için vektör ağırlıkları incelendiğinde (EK5'te ayrıntılı olarak verilmiştir) "Türkiye", "sabah", "haziran" ve "ocak" niteliklerinin pozitif sınıfların (satın alım olması) belirlenmesinde; "Diğer", "kasım" ve "sayfa sayısı" niteliklerinin ise negatif sınıfların (satın alım olmaması) belirlenmesinde etkili olduğu görülmektedir.

LR için de en iyi model olan % 50 tepki oranı için rassal veri setlerinden en iyi sonucu veren 1e modelidir ve pozitif sınıfları F=%90 ve negatif sınıfları %100 kesinlikle doğru tahmin etmiştir. Modelde kalan 0,05 anlamlılık seviyesindeki nitelikler ile açıklayıcılık değeri 0,677 olan LR modeli; "= 0.71 + 0,29 * Türkiye - 0,33 * Sabah - 0,4 * Temmuz + 0,24 * Perşembe + 2,98 * Sayfa Sayısı" şeklinde ifade edilebilmektedir.

**Tablo 3-54 DVM ve LR Modellerinin Karşılaştırılmalı Performansı**

| Tepki Oranı | Model | Doğruluk | | Kappa | | F | | Duyarlılık | | Belirlilik | | Kesinlik | | Ort. Performans | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DVM | LR | DVM | LR | DVM | LR | DVM | LR | DVM | LR | DVM | LR | DVM | LR |
| 50% | 1a | 0,869 | 0,857 | 0,763 | 0,712 | 0,867 | 0,849 | 0,849 | 0,808 | 0,889 | 0,905 | 0,886 | 0,894 | | |
| | 1b | 0,862 | 0,783 | 0,672 | 0,664 | 0,868 | 0,780 | 0,875 | 0,754 | 0,848 | 0,814 | 0,862 | 0,807 | | |
| | 1c | 0,845 | 0,747 | 0,683 | 0,701 | 0,854 | 0,756 | 0,864 | 0,750 | 0,825 | 0,744 | 0,844 | 0,762 | | |
| | 1d | 0,833 | 0,878 | 0,701 | 0,699 | 0,844 | 0,894 | 0,864 | 0,955 | 0,800 | 0,790 | 0,826 | 0,840 | | |
| | 1e | 0,810 | 0,905 | 0,578 | 0,773 | 0,818 | 0,900 | 0,818 | 0,818 | 0,800 | 1,000 | 0,818 | 1,000 | | |
| **Ortalama** | | **0,844** | **0,834** | **0,679** | **0,710** | **0,850** | **0,836** | **0,854** | **0,817** | **0,832** | **0,851** | **0,847** | **0,861** | **0,818** | **0,818** |
| 33% | 2a | 0,858 | 0,834 | 0,712 | 0,722 | 0,774 | 0,777 | 0,726 | 0,740 | 0,924 | 0,895 | 0,828 | 0,818 | | |
| | 2b | 0,850 | 0,839 | 0,631 | 0,652 | 0,757 | 0,756 | 0,700 | 0,738 | 0,925 | 0,891 | 0,824 | 0,776 | | |
| | 2c | 0,831 | 0,853 | 0,670 | 0,610 | 0,759 | 0,780 | 0,750 | 0,744 | 0,875 | 0,911 | 0,767 | 0,821 | | |
| | 2d | 0,823 | 0,885 | 0,668 | 0,756 | 0,718 | 0,829 | 0,636 | 0,773 | 0,925 | 0,949 | 0,824 | 0,895 | | |
| | 2e | 0,839 | 0,900 | 0,576 | 0,763 | 0,783 | 0,842 | 0,818 | 0,727 | 0,850 | 1,000 | 0,750 | 1,000 | | |
| **Ortalama** | | **0,840** | **0,862** | **0,651** | **0,701** | **0,758** | **0,797** | **0,726** | **0,744** | **0,900** | **0,929** | **0,799** | **0,862** | **0,779** | **0,816** |
| 25% | 3a | 0,897 | 0,909 | 0,667 | 0,648 | 0,779 | 0,777 | 0,726 | 0,740 | 0,954 | 0,955 | 0,841 | 0,818 | | |
| | 3b | 0,852 | 0,846 | 0,663 | 0,650 | 0,715 | 0,661 | 0,733 | 0,590 | 0,893 | 0,933 | 0,698 | 0,750 | | |
| | 3c | 0,849 | 0,840 | 0,646 | 0,645 | 0,699 | 0,667 | 0,659 | 0,605 | 0,917 | 0,924 | 0,744 | 0,743 | | |
| | 3d | 0,889 | 0,889 | 0,656 | 0,599 | 0,769 | 0,757 | 0,682 | 0,636 | 0,966 | 0,983 | 0,882 | 0,933 | | |
| | 3e | 0,902 | 0,878 | 0,672 | 0,674 | 0,800 | 0,762 | 0,727 | 0,727 | 0,967 | 0,933 | 0,889 | 0,800 | | |
| **Ortalama** | | **0,878** | **0,872** | **0,661** | **0,643** | **0,752** | **0,725** | **0,705** | **0,660** | **0,939** | **0,946** | **0,811** | **0,809** | **0,791** | **0,776** |
| 20% | 4a | 0,913 | 0,903 | 0,634 | 0,608 | 0,775 | 0,646 | 0,753 | 0,575 | 0,952 | 0,963 | 0,797 | 0,737 | | |
| | 4b | 0,873 | 0,893 | 0,643 | 0,599 | 0,655 | 0,709 | 0,600 | 0,639 | 0,941 | 0,958 | 0,720 | 0,796 | | |
| | 4c | 0,891 | 0,925 | 0,608 | 0,564 | 0,718 | 0,800 | 0,636 | 0,750 | 0,962 | 0,969 | 0,824 | 0,857 | | |
| | 4d | 0,843 | 0,863 | 0,682 | 0,664 | 0,579 | 0,696 | 0,500 | 0,727 | 0,938 | 0,900 | 0,688 | 0,667 | | |
| | 4e | 0,902 | 0,922 | 0,611 | 0,694 | 0,737 | 0,800 | 0,636 | 0,727 | 0,975 | 0,975 | 0,875 | 0,889 | | |
| **Ortalama** | | **0,884** | **0,901** | **0,636** | **0,626** | **0,693** | **0,730** | **0,625** | **0,684** | **0,954** | **0,953** | **0,781** | **0,789** | **0,762** | **0,781** |

# SONUÇ

Bu çalışmada web kullanım madenciliği ile şirketlerin sunucularında tutulan web günlük verileri kullanılmıştır. Bu veriler metin dosyaları halindedir ve gereksiz, yararlı olmayan birçok veriye sahiptir. Bu gereksiz veriler kirli veri olarak adlandırılmaktadır.

Günlük verilerinden ziyaretçinin ülke bilgisi (IP adresinden çıkarılmıştır), site içerisinde dolaşılan sayfa sayısı, sitede harcanan süre, sitenin ziyaret edildiği gün, ay ve saat dilimi ve site ziyaretinde satın alım yapılıp yapılmadığına ilişkin nitelikleri içermektedir.

Ziyaretçiler e-ticaret sitelerine her zaman kesin satın alma isteğiyle gelmemektedirler. Potansiyel müşteriler ürün kıyaslama, fiyat öğrenme gibi nedenlerle de web sitesini ziyaret edebilmektedir. Bu nedenle web sitesi ziyaretlerinde sayfa / ürün görüntüleme sayılarıyla, ürün satışları arasında büyük farklar bulunmaktadır. Bu durum MİY odaklı alanlarda, gerçek uygulamalarda kullanılan veri setlerinde sıkça gözükmektedir. Tepki oranı düşüklüğü olarak da adlandırılan bu dengesiz veri durumu, sınıflandırma yöntemlerinde doğruluk oranlarının gerçekçi olmayacak şekilde artmasına neden olmaktadır.

Çalışmada kullanılan e-ticaret sitesine ait web günlük verilerinde satış cevap oranı yüzde birden daha azdır. Bu oran literatürde karşılaşılan oranlardan da düşüktür. Çözüm olarak satın alımların (pozitif sınıf) bir, iki, üç ve dört katı oranlarında negatif sınıftan tesadüfi örnekleme yöntemiyle negatif sınıf oluşturularak veri dengesizliği giderilmeye çalışılmıştır. Genelleme yapmak için pozitif örnekten azaltmalar yapılarak model sayısı arttırılmıştır.

Destek vektör makineleri ile elde edilen sınıflandırma sonuçları, pazarlama ve MİY alanında sıklıkla kullanılan logistik regresyon analizi sonuçları ile karşılaştırılmıştır.

Düzenlenen verilerden, DVM ve LR ile kurulan modellerin hem eğitimi hem testi için rassal olarak örnek veri setleri çekilmiştir ve bu işlem her iki yöntem için de 5'er

kez yinelenerek farklı rassal veri setleri üzerinde denemeler yapılmıştır. Analizler yapıldıktan sonra ise 6 adet farklı performans ölçüm kriterine göre model değerlendirmeleri elde edilmiş ve rassal veri setleri için bulunan her bir performans ölçümünün ortalama değeri hesaplanmıştır.

Sonuçlara bakıldığında (Tablo 3.54) farklı performans ölçümleri için DVM ve LR modellerinin üstünlüklerinin de farklılık gösterdiği görülmektedir. Dört farklı tepki oranına göre (%50, %33, %25, %20) her bir performans ölçümü için elde edilen ortalama değerlerin de ortalaması alındığında ise, hem DVM hem de LR için en iyi sonucun yüzde 50 tepki oranı ile oluşturulan model olduğu ve DVM ile LR performans ölçümlerinin (0,818) birbirine eşit olduğu görülmektedir. Bir başka ifadeyle, söz konusu ikili sınıflandırma görevi için her iki sınıfa ait örnek sayılarının eşit olduğu durumda hem DVM hem de LR ile kurulan modellerde en iyi sınıflandırma sonuçlarına ulaşılmıştır.

DVM için çeşitli kernel tipleri denenmiş ve test verisi üzerinde denemelerde en iyi sınıflama sonuçlarına sigmoid kernel ile ulaşılmıştır. LR'de de logistik fonksiyon (sigmoid fonksiyon) kullanılmaktadır. Genel olarak tüm modellerde sonuçların yakın çıkmasının (özellikle tepki oranının yüzde 50 olduğu modelde aynı olması) nedeni olarak sigmoid fonksiyonun kullanılmış olması göz önünde bulundurulması gereken bir nedendir.

Yukarıda özetlenen sonuçlar ışığında, %50 tepki oranı ile kurulan modelin analiz için tercih edilmesi gerektiği açıktır. Buna göre DVM ile %50 tepki oranı için kurulan modelde tüm rassal veri setleri için vektör ağırlıkları incelendiğinde (EK5'te ayrıntılı olarak verilmiştir) "Türkiye", "sabah", "haziran" ve "ocak" niteliklerinin pozitif sınıfların (satın alım olması) belirlenmesinde; "Diğer", "kasım" ve "sayfa sayısı" niteliklerinin ise negatif sınıfların (satın alım olmaması) belirlenmesinde etkili olduğu görülmektedir.

Veri ön analizlerinde, satın alma davranışında ziyaret edilen sayfa sayısının ve ziyaret süresinin baskın çıktığı kutu grafiklerinde de görülmektedir. Modellerin çoğunda sayfa sayısı, süresi ve ülke = Türkiye, satın alım tahminini etkileyen nitelikler

olarak öne çıkmaktadır. DVM sonuçları veri ön analizi sonucunda yapılan çıkarımlarla paralellik göstermektedir.

Çalışmada web günlük verilerinden elde edilen niteliklere göre satın alma davranışı elde edilmeye çalışılmıştır. Bir ziyaretçiyi satın almaya teşvik eden dış etkenler de bulunmaktadır. Kullanıcıların demografik bilgileri, meslekleri, aylık gelirleri gibi nitelikleri bu dış etkenlerden bazılarıdır ve uygulamada yer almamaktadır. Bu durum analizin önemli kısıtlarından biridir. Çalışmada veri kaynağı olarak sadece web günlük verileri kullanılabilmiştir. Kullanıcı IP'leri gizli olduğundan web günlük verisi elde etme aşamasında da zorluklar yaşanmıştır. Şirket ismi saklanarak ve IP numaraları gizlenerek analizler gerçekleştirilmiştir. Destek vektör makinelerinin daha etkin kullanımı için e-ticaret sitelerinin ilave müşteri bilgilerini saklamaları önerilmektedir.

Çalışmanın diğer bir kısıtı çalışmaya konu olan e-ticaret sitesinin sadece tek bir ürünün (özellikli ürün) değişik varyasyonlarını satmasıdır. Bu nedenle sayfalar arası ilişki kurulamamıştır. Bir ziyaretçi bir oturumda en fazla bir ürün satın almaktadır ve ücret ya da ürün sayısı da ulaşılamayan niteliklerdir. Satın alma durumunun nadir olduğu (%1'in altında) veri seti uygulamalarında veri seti küçültülerek alt veri setleri üzerinde analizler gerçekleştirilmiştir. Veri setinin hazırlanmasında ve analizlerinde önemli bir zaman harcanmış ve deneyim kazanılmıştır. Ancak bu deneyimin demografik verilerin ve ürün çeşitliliğinin olduğu veri setleri üzerinde yapılması sonraki çalışmalara bırakılmıştır.

# KAYNAKÇA


Agosti, M., Di Nunzio, G.: "Web Log Mining: A Study of User sessions.", **Proceedings of the 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries (PersDL),** 2007.

Ahmed, S.: "Applications of Data Mining in Retail Business.", **Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04),** 2004, pp. 455-459.

Akaho, S.: "VC Dimension Theory for a Learning System with Forgetting", **Proceedings of 1993 International Joint Conference on Neural Networks,** Tokyo, Vol. 1, 25-29 October, pp.493-496.

Akbani, R., Kwek, S., Japkowicz, N.: "Applying Support Vector Machines to Imbalanced Datasets.", **Machine Learning: ECML 2004,** 2004

Akpınar, H.: "DATA Veri Madenciliği Veri Analizi", **Papatya Yayıncılık**, 2014

Alpaydın, E.: "Introduction to Machine Learning.", **MIT Press,** 2009.

Alpaydın, E.: "Yapay Öğrenme.", **İstanbul: Boğaziçi Üniversitesi Yayınları,** 2011

Arumugam, Suguna: "Predictive prefetching framework based on New Preprocessing Algorithms towards latency reduction.", **Asian J. Inform Techno**, c. 7, s. 3, 2008, pp. 87-99.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J.: "Benchmarking state of the art classification algorithms for credit scoring.", **Journal of the Operational Research Society,** c. 54, s. 6, 2003, pp. 627–635.

| | |
|---|---|
| Baglioni, M., Ferrara, U., Romei, A., Ruggieri, S., Turini, F.: | "Preprocessing and mining web log data for web personalization.", **Advances in Artificial Intelligence,** c. 2829, 2003. |
| Balaban, M., Kartal, E.: | "Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları.", **İstanbul: Çağlayan Kitabevi,** 2015. |
| Baldi, P., Frasconi, P., Smyth, P.: | "Modeling the Internet and the Web: Probabilistic Methods and Algorithms." **England: John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex,** 2003. |
| Batista, P., Silva, M.: | "Mining Web Access Logs of an On-line Newspaper.", **the proceedings of 12th International Meeting of the euro working group on decision support systems,** 2001. |
| Berendt, Spiliopoulou.: | "Analyzing navigation behaviour in web sites integrating multiple information systems.", **The International Journal on Very Large Data Bases,** c. 9, s. 1, 2000, pp. 56-75. |
| Bose, I., Chen, X.: | "Quantitative models for direct marketing: A review from systems perspective.", **European Journal of Operational Research**, c. 195, s. 1, 2009, pp. 1-16. |
| Bramer, M.: | "Principles of Data Mining.", 2013. |
| Brusilovsky, P., Kobsa, A., Nejdl, W.: | "The Adaptive Web.", 2007 |
| Bucklin, R.: | "Marketing Models for Electronic Commerce.", **Handbook of Marketing Decision Models**, s. 327, 2008. |
| Bucklin, R., Gupta, S.: | "Brand choice, purchase incidence, and segmentation: An integrated modeling approach.", **Journal of Marketing Research**, c. 2, 1992, pp. 201–215. |

| | |
|---|---|
| Burez, J., Van den Poel, D.: | "Handling class imbalance in customer churn prediction.", **Expert Systems with Applications**, s. 36, 2009, pp. 4626-4636. |
| Chapman, P., Clinton J., Kerber, R., Khabaza, T., Renartz, T., Shearer, C., Wirth, R., | **CRISP-DM 1.0 Step-by-Step Data Mining Guide**, 2000, Chicago IL: SPSS. |
| Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: | "SMOTE: synthetic minority over-sampling technique.", **Journal of Artificial Intelligence Research**, c. 16, s. 1, 2002, pp. 321-357. |
| Chitraa, V., Davamani, A.: | "A Survey on Preprocessing Methods for Web Usage Data.", **International Journal of Computer Science and Information Security,** c. 7, s. 3, 2010. |
| Clarke, B., Fokoué, E., Zhang , H.: | "Principles and Theory for Data Mining and Machine Learning.", **New York: Springer Science+Business Media,** 2009. |
| Cooley, Mobasher, Srivastava.: | "Data Preparation for Mining world wide web browsing patterns.", **Knowledge and information systems,** c. 1, s. 1, 1999, pp. 5-32. |
| Cooley, R., Mobasher, B., Srivastava, J.: | "Web Mining: Information and Pattern Discovery on the World Wide Web.", **Proceedings 9th IEEE International Conference on Tools with Artificial Intelligence**, 1997, pp. 558-567. |
| Cristianni, N., Shawe-Taylor, J.: | **An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods**, UK, Cambridge University Press, 2000. |

Das, Turkoglu.: "Creating meaningful data from web logs for improving the impresiveness of a website by using path analysis method.", **Expert Systems with Applications,** c. 36, s. 3, 2009, pp. 6635-6644.

Drummond, C., Holte, R.: "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling.", **Workshop on Learning from Imbalanced Datasets II**, ICML. Washington DC., 2003.

Dunham, H.: "Data Mining Introductory and Advanced Topics.", **Prentice Hall**, 2003.

Eirinaki , M., Vazirgannis, M.: "Web Mining for Web Personalization.", **ACM Transactions on Internet Technology**, 2003, pp. 1-27.

Etzioni, E., "The World Wide web: Quagmire or Gold Mine", **Communication of the ACM**, 1996, 39(11), pp. 65-68.

Facca, F. M., Lanzi, P. L.: "Mining interesting knowledge from weblogs: a survey.", **Data & Knowledge Engineering**, 2005, pp. 225-241.

Fayyad, U., Mannila, H., Smyth, P., Uthurusamy, R.: "Advances in Knowledge Discovery and Data Mining.", Cambridge: MA: AAAI / MIT Press, 1996.

Frawley, W., Pitatetsky-Shapiro, G., Matheus, C.: "Knowledge Discovery in Databases: An Overview.", **AI Magazine,** c. 13, 1992, pp. 213-228.

Giraud-Carrieer, C., Povel, O.: "Characterising Data Mining software.", **Intelligent Data Analysis**, 2003, pp. 181-192.

Gündüz, Ş., Adalı, E.: "Web Kullanıcılarının Davranışları İçin Örüntü Bulma ve Modelleme", itüdergisi/d mühendislik, c.3, s. 6, 2004, pp. 15-24.

Han, J., Kamber, M., Pei, J.: "Data Mining: Concepts and Techniques", Elsevier, 3rd Edition, 2012.

Hand, D., Mannila, H., Smyth, P.: "Principles of Data Mining.", **London: The MIT Press**, 2003.

Haykin, S.: Neural Networks: A Comprehensive Foundation, **Prentice Hall, New Jersey, USA**, 2 . Edition, 1999.

He, H., Garcia, E.: "Learning from Imbalanced Data.", **IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING**, c. 21, s. 9, 2009, pp. 1263-1284.

He, L., Kong, F., Shen, Z.: "Multiclass SVM Based Land Cover Classification with Multisource Data.", **Proceedings of International Conference of Machine Learning and Cybernetics**, 2005, pp. 3541-3545.

Hughes, A.: "Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-Based Marketing Program.", **Irwin Professional,** McGraw-Hill, 2005.

Huiying, Wei.: "An intelligent algorithm of data preprocessing in web usage mining.", **Intelligent Control and Automation**, c.4, 2004, pp. 3119-3123.

Izenman, A. J.: "Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning", **Springer Science+Business Media, LLC**, 2008.

Japkowicz, N.: "The Class Imbalance Problem: Significance and Strategies.", **In Proceedings of the 2000 International Conference on Artificial Intelligence: Special Track on Inductive Learning**, c.1, 2000.

Japkowicz, N.: "Concept learning in the presence of between class and within-class imbalances.", **In Proceedings of the Fourteenth Conference of the Canadian Society for Computational Studies of Intelligence,** 2001, pp. 67-77.

| | |
|---|---|
| Jiawei, H., Kamber, M., Pei, J.: | "Data Mining Concepts and Techniques.", Morgan Kaufmann, 2011. |
| Joshi, K., Joshi, A., Yesha, Y.: | "On Using a Warehouse to Analyze Web Logs.", **Distrubuted and Parallel Databases,** 2003. |
| Khoshgoftaar, T., Van Hulse, J., Napolitano, A.: | "Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data.", **IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS**, c. 41, s. 3, 2011, pp. 552-568. |
| Kim, G., Chae, B., Olson, D.: | "A support vector machine (SVM) approach to imbalanced datasets of customer responses: comparison with other customer response models.", **Service Business**, c. 7, 2013, pp. 167-182. |
| Kohavi, R., Mason, L., Parekh, R., Zheng, Z.: | "Lessons and Challenges from mining retail e-commerce data.", **Machine Learning**, c. 57, s. 1-2, 2004, pp. 83-113. |
| Kosala, R., Blockeel, H.: | "Web Mining Research: A Survey.", **SIGKDD Explorations: Newsletter of the Special Interest Group on Knowledge Discovery & Data Mining**, 2000, pp. 1-15. |
| Kracklauer, A., Mills, D., Dirk, S.: | "Customer Management as the Origin of Collaborative Customer Relationship Management.", **Collaborative Customer Relationship Management Taking CRM to the Next Level,** 2004, pp. 3-6. |
| Kreps, J.: | "I Logs Event Data, Stream Processing, and Data Integration.", **Oreily**, 2015. |
| Levin, N., Zahavi, J.: | "Continuous predictive modeling, a comparative analysis.", **Journal of Interactive Marketing**, s. 12, 1998, pp. 5–22. |

Li, S., Liechty, J. C., Montgomery, A. L.: "Modeling Category Viewership of Web Users with Multivariate Count Models.", Carniege Mellon University, 2002.

Ling, C., Li, C.: "Data Mining for Direct Marketing: Problems and Solutions.", **In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)**, 1998, pp. 73-79.

Liu, B.: "Web Data Mining Exploring Hyperlinks, Contents, and Usage Data.", 2007.

Markov, Z., Larose, D.: "Data Mining the Web Uncovering Patterns in Web Content, Structure, and Usage.", New Jersey: Wiley-Interscience, 2007.

McCarty, J., Hastak, M.: "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression.", **Journal of Business Research**, c. 60, s. 6, 2007, pp. 656-662.

Mitra, S., Pal, S., Mitra, P.: "Data Mining in Soft Computing Framework: A Survey.", **IEEE TRANSACTIONS ON NEURAL NETWORKS**, c. 13, s. 1, 2002, pp. 3-14.

Nelson, T.: "A File Structure for the Complexi the Changing and the Indeterminate.", **ACM 20th National Conference**, 1965.

Neslin, S., Gupta, S., Kamakura, W., Lu, J., Mason, C.: "Detection defection: Measuring and understanding the predictive accuracy of customer churn models." **Journal of Marketing Research**, c. 43, s. 2, 2006, pp. 204-211.

Newton, R. R., Rudestam, K. E.: "Your Statistical Consultant: Answer to your Data Anaysis Questions.", **CA: Sage Publishing**, 1999.

Ngai, E., Xiu, L., Chau, D.: "Application of data mining techniques in customer relationship management: A literature review and classification.", **Expert Systems with Applications**, c. 36, s. 2, 2009, pp. 2592-2602.

| | |
|---|---|
| Nielsen, J.: | "User Interface Directions for the Web.", **Communications of ACM**, 1999, pp. 65-72. |
| Nisbet, R., Elder, J., Miner, G.: | "Handbook of Statistical Analysis & Data Mining Applications.", **Elseiver**, 2009. |
| Öztemel, E.: | "Yapay Sinir Ağları", İstanbul, **Papatya Yayıncılık**, 2003. |
| Parvatiyar, A., Sheth, J.: | "Customer Relationship Management: Emerging Practice, Process, and Discipline.", **Journal of Economic and Social Research**, c. 3, s. 2, 2001, pp. 1-34. |
| Prabhu, S., Venkatesan, N.: | "Data Mining and Warehousing", Prabhu, S., Venkatesan, N., **New Age International (P) Limited, Publishers**, 2007. |
| Pöyhönen, S.: | "Support Vector Machine Based Classification in Condition Monitoring of Induction Motors", **Helsinki University of Technology Control Engineering Laboratory, Finland**, 2004. |
| Prasad, G. S., Reddy, N., Acharya, U.: | "Knowledge Discovery from Web Usage Data: A Survey of Web Usage Pre-Processing Techniques.", **BAIP 2010**, CCIS 70, 2010, pp. 505-507. |
| Provost, F., Fawcett, T.: | "Robust Classification for Imprecise Environments.", **Machine Learning**, c. 42, s. 3, 2001, pp. 203-231. |
| Ren, J., Shen, Y., Ma, S.: | "Appliying Multi-class SVMs into Scene Image Classification.", **Proceedings of the 17th International Conference on Innovations in Applied Artificial Intelligence**, 2004, pp. 924-934. |
| Sen, A., Dacin, P., Pattichis, C.: | "Current Trends In Web Data Analysis.", **Communications of the ACM**, 2006, pp. 85-91. |

| | |
|---|---|
| Shahabi, Zarkesh, Adibi, Shah.: | "Knowledge Discovery from Users Web-page Navigation.", **Research Issues in Data Engineering,** 1997, pp. 20-29. |
| Shaw, M., Subramaniam, C., Tan, G., Welge, M.: | "Knowledge management and data mining for marketing.", **Decision Support Systems**, c. 31, 2001, pp. 127-137. |
| Shin, H., Cho, S.: | "How to Deal with Large Dataset, Class Imbalance and Binary Output in SVM based Response Model.", **Proceedings of the Korean Data Mining Conference**, 2003. |
| Spiliopoulou, M.: | "The Laborious Way From Data Mining to Web Log Mining.", **Computer Systems Science and Engineering**, 1999. |
| Spiliopoulou, Mobasher, Berendt , Nakagawa.: | "A framework for evaluation of session reconstruction heuristic in web usage analysis.", **Informs journal on computing,** c. 15, s. 2, 2003, pp. 171-190. |
| Srivastava, J., Cooley, R., Deshpande, M., Tan, P.: | "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data.", **SIGKDD Explorations**, 2000, pp. 12-23. |
| Swift, R.: | "Accelerating Customer Relationships: Using CRM and Relationship Technologies.", **N.J.: Prentice Hall PTR**, 2001. |
| Tan, Kumar.: | "Modeling of web robot navigational patterns.", **Army High Performance Computing Research Center,** 2000. |
| Tang, Y., Zhang, Y.-Q., Chawla, N., Krasser, S.: | "SVMs Modeling for Highly Imbalanced Classification.", **Journal of LATEX Class Files**, 2002. |
| Thakare, S. B., Gawali, S.: | "A Effective and Complete Preprocessing for Web Usage Mining.", **International Journal on Computer Science and Engineering**, c. 2, s. 3, 2010, pp. 848-851. |

Tolun S.: "Destek Vektör Makineleri: Banka Başarısızlığının Tahmini Üzerine Bir Uygulama", **Doktora Tezi**, İ.Ü. S.B.E., 2008.

Vapnik, V.: "Statistical Learning Theory.", **Wiley**, 1998.

Velasquez, J. D., Palade, V.: "Adaptive Web Sites A Knowledge Extraction from Web Data Approach.", **Amsterdam: IOS Press**, 2008.

Velasquez, J. D., Yasuda, H., Aoki, T., Weber, R., Vera, E.: "Using Self Organizing Feature Maps to Acquire Knowledge About User Behavior in a Web Site.", **7th Int. Conf. Knowledge-Based Intelligent Information & Engineering Systems**, 2003, pp. 951-958.

Velasquez, J., Yasuda, H., Aoki, T., Weber, R.: "Using the KDD Process to Support the Web Site Reconfiguration.", **IEE/WIC Int. Conf. on Web Intelligence**, 2013, pp. 511-515.

Viaene, S., Baesens, B., Van Gestel, T., Suykens, J., Van den Poel, D., Vantihienen, J., Dedene, G.: "Knowledge Discovery in a Direct Marketing Case using Least Squares Support Vector Machines.", **International Journal of Intelligent Systems**, c. 16, s. 9, 2001, pp. 1023-1036.

Wang, X., Zhang, Y.: "Statistical Learning Theory and State of Art in DVM", **Proceedings of the Second IEEE International Conference on Cognitive Informatics**, IEEE, 2003, pp. 55-59.

Weiss, G.: "Learning with rare cases and small disjuncts.", **In Proceedings of the Twelfth International Conference on Machine Learning**, 1995, pp. 558-565.

Weiss, G.: "Mining with rarity: a unifying framework.", **ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets**, c. 6, s. 1, 2004, pp. 7-19.

Witten, I.H., Frank, E.: "Data Mining: Practical Machine Learning Tools and Techniques", **Elsevier**, 2nd Edition, 2005.

Xu, G., Zhang, Y., Li, L.: "Web Mining and Social Networking Techniques and Applications.", **Springer Science+Business Media**, 2011.

Zaiane, O., Xin, M., Han, J.: "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs.", **Research and Technology Advances in Digital Libraries,** 1998, pp. 19-29.

# EKLER

### EK:1 Veritabanı veri aktarımı

```
DELIMITER $$
DROP PROCEDURE IF EXISTS RepeatLoopProc$$
CREATE PROCEDURE RepeatLoopProc()
BEGIN DECLARE x INT;
SET x = 2;

        REPEAT
                LOAD DATA INFILE CONCAT("/importmysql/",x,".txt" )
                INTO TABLE data6
                FIELDS TERMINATED BY " "
                (v1, v2, v3, v4, v5, v6, v7, v8, v9, v10, v11, v12, v13, v14);
                SET x = x + 1;
        UNTIL x > 2525
        END REPEAT;
END$$
DELIMITER ;
call RepeatLoopProc();

delete
from data6
where v1 like "#%"
```

**EK:2 Filtrelenen parametrelere ait SQL komutları**

```
delete
from loganaliz
where
v6 LIKE "%.png"
or v6 LIKE "%.jpg"
or v6 LIKE "%.gif"
or v6 LIKE "%.jpeg"
or v6 LIKE "%.swf"
or v6 LIKE "%.ico"

SELECT max(LENGTH(v5)) FROM `loganaliz`
insert into user (ip) select distinct(clientip) as ip from loganaliz

UPDATE loganaliz2, serverip
SET loganaliz2.server = serverip.id
where loganaliz2.server = serverip.ip

UPDATE loganaliz2, user
SET loganaliz2.clientip = user.id
where loganaliz2.clientip = user.ip

SELECT distinct(filename) FROM `loganaliz2` WHERE parameter = "-"


// sessionları belirleyen sql
SELECT clientip, count(id) FROM `loganaliz2` WHERE `filename` LIKE
'/Product.aspx%' group by clientip order by count(id) DESC
// sessionlist tablosunu oluşturuyor:
insert into sessionlist (clientip)
SELECT clientip FROM `loganaliz2` WHERE `filename` LIKE '/Product.aspx%'
group by clientip order by count(id) DESC
// session sorgulama:

select *
from loganaliz2 la2, sessionlist sl
where la2.clientip = sl.clientip
and sl.id =
order by la2.id asc
//cluster tablosu satın alındı onayı:
7
```

```
UPDATE loganaliz2 la2, cluster clu
SET clu.satinalma = 1
WHERE
la2.clientip = clu.ip
and
la2.parameter LIKE '%Step=Approved%'
update cluster set satinalma = 1
where ip in
(select clientip as ip
 from loganaliz2
 where parameter LIKE '%Step=Approved%')
CREATE VIEW satinalanlar AS
select clientip as ip
 from loganaliz2
 where parameter LIKE '%Step=Approved%'

update cluster set satinalma = 1
where ip in
(select clientip as ip
 from satinalanlar)

delete
from loganaliz2
where filename = "/DXR.axd"

delete
from loganaliz2
where filename = "/iisstart.htm"

delete
from loganaliz2
where filename like "/Scripts/%"

delete
FROM `loganaliz2`
WHERE filename like "/%.css"

delete
FROM `loganaliz2`
WHERE filename like "/robots.txt"
```

```
delete
FROM `loganaliz2`
WHERE filename like "/undefined"

delete
FROM `loganaliz2`
WHERE filename = "/no-text%"

delete
FROM `loganaliz2`
WHERE filename like "%phpMyAdmin%"
delete FROM `loganaliz2` WHERE parameter = "-" and filename like "%phpm%"

SELECT  clientip, count(id)  FROM `loganaliz2` WHERE `filename` LIKE
'/Product.aspx%' group by clientip order by count(id) DESC

SELECT *
FROM `data1`
WHERE `v6` NOT LIKE '%jpg%'
AND `v6` NOT LIKE '%png%'
AND `v6` NOT LIKE '%gif%'
AND `v6` NOT LIKE '%swf%'
AND `v6` NOT LIKE '%axd%'
AND `v6` NOT LIKE '%css%'
AND `v6` NOT LIKE '%js%'
AND `v10`
LIKE '178.2.185.126'
ORDER BY id asc
LIMIT 0,250
```

## EK: 3 Temel İstatistik Tablolar

### Saat aralıklarına göre İstek, Sayfa, Ziyaret sayıları

| Saat Dilimi | İstek Sayısı | Sayfalar | Ziyaretler | Ortalama Ziyaret Süresi |
|---|---|---|---|---|
| 00:00 - 00:59 | 377.525 | 36.272 | 12.592 | 06:11 |
| 01:00 - 01:59 | 211.101 | 25.098 | 10.121 | 07:00 |
| 02:00 - 02:59 | 121.793 | 21.821 | 10.893 | 06:12 |
| 03:00 - 03:59 | 82.345 | 18.374 | 9.107 | 08:14 |
| 04:00 - 04:59 | 90.691 | 19.925 | 10.936 | 06:08 |
| 05:00 - 05:59 | 269.821 | 31.594 | 12.268 | 06:44 |
| 06:00 - 06:59 | 636.117 | 61.359 | 17.032 | 05:57 |
| 07:00 - 07:59 | 1.179.547 | 103.507 | 19.677 | 06:03 |
| 08:00 - 08:59 | 1.343.446 | 133.099 | 22.832 | 05:46 |
| 09:00 - 09:59 | 1.770.665 | 182.957 | 24.479 | 05:27 |
| 10:00 - 10:59 | 1.606.917 | 158.916 | 24.627 | 05:50 |
| 11:00 - 11:59 | 1.736.141 | 155.369 | 24.460 | 05:28 |
| 12:00 - 12:59 | 1.638.862 | 151.943 | 25.059 | 04:45 |
| 13:00 - 13:59 | 1.687.131 | 151.307 | 24.319 | 05:16 |
| 14:00 - 14:59 | 1.503.873 | 144.678 | 24.884 | 04:36 |
| 15:00 - 15:59 | 1.283.618 | 120.599 | 21.661 | 04:52 |
| 16:00 - 16:59 | 1.114.102 | 99.711 | 21.824 | 04:21 |
| 17:00 - 17:59 | 1.111.955 | 89.179 | 20.726 | 04:38 |
| 18:00 - 18:59 | 1.041.245 | 90.100 | 21.514 | 03:59 |
| 19:00 - 19:59 | 1.138.569 | 99.667 | 20.712 | 04:07 |
| 20:00 - 20:59 | 1.175.290 | 107.538 | 22.795 | 03:44 |
| 21:00 - 21:59 | 1.094.224 | 91.885 | 19.669 | 03:46 |
| 22:00 - 22:59 | 824.561 | 69.615 | 18.084 | 03:20 |
| 23:00 - 23:59 | 496.334 | 44.775 | 13.033 | 26:01 |
| **Ortalama** | **980.661** | **92.053** | **18.887** | **06:11** |
| **Toplam** | **23.535.873** | **2.209.288** | **453.304** | |

.

**Aylara göre istek, sayfa, ziyaret sayıları**

| Ay | İstek Sayısı | Sayfalar | Ziyaret Sayısı | Ortalama Ziyaret Süresi |
|---|---|---|---|---|
| May.11 | 180.760 | 13.947 | 962 | 12:42 |
| Haz 2011 | 780.593 | 49.165 | 9.193 | 04:21 |
| Tem 2011 | 871.423 | 53.489 | 13.896 | 06:03 |
| Ağu 2011 | 587.219 | 46.655 | 11.751 | 05:58 |
| Eyl 2011 | 492.975 | 56.943 | 10.213 | 06:38 |
| Eki 2011 | 735.403 | 52.686 | 12.178 | 05:06 |
| Kas 2011 | 658.263 | 44.388 | 11.709 | 05:21 |
| Ara 2011 | 1.208.750 | 150.781 | 16.192 | 04:45 |
| Oca 2012 | 1.165.913 | 176.648 | 16.162 | 05:32 |
| Şub 2012 | 1.925.408 | 183.439 | 20.108 | 05:26 |
| Mar.12 | 919.668 | 142.740 | 15.964 | 04:24 |
| Nis 2012 | 679.012 | 78.778 | 12.898 | 04:05 |
| May.12 | 888.435 | 96.612 | 13.829 | 05:02 |
| Haz 2012 | 562.585 | 63.230 | 11.806 | 04:01 |
| Tem 2012 | 650.031 | 83.190 | 11.517 | 05:01 |
| Ağu 2012 | 970.611 | 109.910 | 13.564 | 05:05 |
| Eyl 2012 | 988.906 | 137.974 | 13.479 | 07:18 |
| Eki 2012 | 1.505.383 | 79.482 | 19.576 | 05:56 |
| Kas 2012 | 930.761 | 61.514 | 15.479 | 07:43 |
| Ara 2012 | 1.111.556 | 60.550 | 17.961 | 05:20 |
| Oca 2013 | 904.482 | 52.249 | 17.578 | 04:32 |
| Şub 2013 | 862.857 | 51.786 | 17.086 | 05:33 |
| Mar.13 | 846.654 | 67.715 | 29.855 | 06:04 |
| Nis 2013 | 865.057 | 81.691 | 36.733 | 07:11 |
| May.13 | 867.002 | 74.432 | 30.257 | 07:51 |
| Haz 2013 | 601.393 | 57.101 | 24.451 | 07:28 |
| Tem 2013 | 773.826 | 81.776 | 28.721 | 10:09 |
| Ağu 2013 | 947 | 417 | 186 | 08:52 |
| **Ortalama** | **840.566** | **78.903** | **16.189** | **06:11** |
| **Toplam** | **23.535.873** | **2.209.288** | **453.304** | |

## Ülkelere göre istek, ziyaret, Sayfa sayıları

| | Ülke / Bölge | Ziyaretler | Toplam Ziyaret Yüzdesi | İstek Sayısı | Toplam İstek Yüzdesi | Ziyaret Başına Düşen Sayfa Görüntüleme | Ziyaret Başına Düşen Kalma Süresi | Tek Sayfalık Ziyaret Oranı |
|---|---|---|---|---|---|---|---|---|
| 1 | United States | 189.846 | 41,88% | 835.622 | 3,55% | 1,47 | 08:35 | 75,03% |
| 2 | Turkey | 185.160 | 40,85% | 21.607.502 | 91,81% | 9,51 | 04:53 | 52,07% |
| 3 | China | 36.590 | 8,07% | 105.515 | 0,45% | 1,39 | 00:35 | 94,27% |
| 4 | Japan | 12.408 | 2,74% | 38.460 | 0,16% | 1,07 | 01:19 | 89,40% |
| 5 | Russian Federation | 9.049 | 2,00% | 79.812 | 0,34% | 2,36 | 20:29 | 75,30% |
| 6 | Germany | 3.101 | 0,68% | 195.430 | 0,83% | 8,22 | 04:07 | 59,31% |
| 7 | United Kingdom | 2.983 | 0,66% | 123.662 | 0,53% | 2,78 | 02:39 | 66,55% |
| 8 | Belgium | 1.541 | 0,34% | 24.891 | 0,11% | 1,02 | 00:53 | 65,78% |
| 9 | Netherlands | 1.481 | 0,33% | 53.145 | 0,23% | 3,65 | 01:22 | 66,19% |
| 10 | Ukraine | 1.368 | 0,30% | 4.799 | 0,02% | 1,39 | 01:38 | 87,26% |
| 11 | France | 1.301 | 0,29% | 40.020 | 0,17% | 4,77 | 01:10 | 75,30% |
| 12 | Korea, Republic Of | 738 | 0,16% | 5.204 | 0,02% | 1,87 | 04:39 | 84,40% |
| 13 | Taiwan | 683 | 0,15% | 3.324 | 0,01% | 0,43 | 02:11 | 78,66% |
| 14 | Iceland | 491 | 0,11% | 59.127 | 0,25% | 3,22 | 03:07 | 55,08% |
| 15 | Finland | 416 | 0,09% | 5.115 | 0,02% | 1,09 | 00:20 | 90,72% |
| 16 | Canada | 396 | 0,09% | 14.598 | 0,06% | 5,86 | 01:06 | 76,63% |
| 17 | Azerbaijan | 317 | 0,07% | 23.992 | 0,10% | 4,15 | 02:23 | 63,01% |
| 18 | Italy | 304 | 0,07% | 17.779 | 0,08% | 4,94 | 02:13 | 71,88% |
| 19 | Brazil | 255 | 0,06% | 10.902 | 0,05% | 2,27 | 00:34 | 85,17% |
| 20 | Ireland | 246 | 0,05% | 5.164 | 0,02% | 2,34 | 01:14 | 72,16% |
| 21 | Switzerland | 219 | 0,05% | 25.128 | 0,11% | 9,15 | 03:08 | 50,54% |
| 22 | Austria | 214 | 0,05% | 26.408 | 0,11% | 19,92 | 03:57 | 58,43% |
| 23 | Singapore | 212 | 0,05% | 2.474 | 0,01% | 0,41 | 07:12 | 80,00% |
| 24 | Romania | 198 | 0,04% | 6.133 | 0,03% | 2,20 | 00:29 | 71,93% |
| 25 | Poland | 180 | 0,04% | 12.093 | 0,05% | 3,77 | 01:51 | 60,71% |
| 26 | Sweden | 170 | 0,04% | 12.879 | 0,05% | 5,04 | 01:10 | 68,31% |
| 27 | India | 162 | 0,04% | 12.191 | 0,05% | 4,63 | 03:04 | 70,13% |
| 28 | Bulgaria | 149 | 0,03% | 11.866 | 0,05% | 3,68 | 02:03 | 71,65% |
| 29 | Greece | 147 | 0,03% | 12.032 | 0,05% | 5,45 | 02:42 | 75,74% |
| 30 | United Arab Emirates | 145 | 0,03% | 6.992 | 0,03% | 6,88 | 01:20 | 53,95% |
| 31 | Australia | 140 | 0,03% | 7.070 | 0,03% | 10,28 | 02:50 | 56,90% |
| 32 | Saudi Arabia | 134 | 0,03% | 11.098 | 0,05% | 3,78 | 02:18 | 84,17% |
| 33 | Egypt | 120 | 0,03% | 7.157 | 0,03% | 1,56 | 03:59 | 97,37% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 34 | Belarus | 107 | 0,02% | 7.031 | 0,03% | 8,21 | 02:47 | 50,57% |
| 35 | Czech Republic | 106 | 0,02% | 6.005 | 0,03% | 13,58 | 00:34 | 61,62% |
| 36 | Israel | 103 | 0,02% | 4.222 | 0,02% | 8,60 | 00:58 | 76,92% |
| 37 | Morocco | 95 | 0,02% | 5.280 | 0,02% | 1,83 | 02:18 | 96,67% |
| 38 | Tunisia | 94 | 0,02% | 5.240 | 0,02% | 1,40 | 01:51 | 100,00 % |
| 39 | Algeria | 91 | 0,02% | 6.330 | 0,03% | 2,18 | 03:41 | 93,26% |
| 40 | Thailand | 91 | 0,02% | 2.446 | 0,01% | 3,62 | 02:00 | 80,00% |
| 41 | Norway | 89 | 0,02% | 3.129 | 0,01% | 3,17 | 01:07 | 77,50% |
| 42 | Philippines | 89 | 0,02% | 2.276 | 0,01% | 1,85 | 01:33 | 89,53% |
| 43 | Spain | 88 | 0,02% | 7.826 | 0,03% | 4,08 | 02:43 | 63,29% |
| 44 | Denmark | 85 | 0,02% | 8.372 | 0,04% | 8,44 | 02:26 | 63,38% |
| 45 | Hong Kong | 80 | 0,02% | 2.302 | 0,01% | 4,38 | 03:11 | 65,22% |
| 46 | Iraq | 75 | 0,02% | 3.544 | 0,02% | 2,24 | 03:42 | 79,37% |
| 47 | Lithuania | 74 | 0,02% | 814 | 0,00% | 2,24 | 00:02 | 57,53% |
| 48 | Lebanon | 72 | 0,02% | 5.059 | 0,02% | 4,64 | 03:54 | 85,48% |
| 49 | Unknown | 66 | 0,01% | 3.638 | 0,02% | 1,92 | 00:56 | 80,33% |
| 50 | Mexico | 64 | 0,01% | 2.397 | 0,01% | 2,00 | 00:37 | 82,76% |
| | **Toplam** | **453.304** | **100,00 %** | **23.535.873** | **100,00 %** | **4,87** | **06:11** | **65,29%** |
| | **Alttoplam** | **452.333** | **99,79%** | **23.483.495** | **99,78%** | **4,87** | **06:12** | **65,75%** |

## Şehirlere göre ziyaret, istek, sayfa sayıları

| | Şehir | Ziyaretler | İstek Sayısı | Ziyaret Başına Sayfa Görüntüleme | Ziyaret Başına Kalış Süresi | Tek Sayfalık Ziyaret Oranı |
|---|---|---|---|---|---|---|
| 1 | Ankara, Turkey | 104.174 | 12.152.946 | 8,52 | 03:23 | 54,47% |
| 2 | Istanbul, Turkey | 60.498 | 6.940.293 | 11,44 | 07:52 | 47,63% |
| 3 | Redmond, Washington, United States | 33.367 | 87.409 | 1,30 | 03:48 | 76,57% |
| 4 | Mountain View, California, United States | 27.989 | 140.570 | 1,97 | 17:45 | 72,41% |
| 5 | Bei Jing, Beijing, China | 24.820 | 28.922 | 0,95 | 00:29 | 99,27% |
| 6 | New York City, New York, United States | 19.235 | 56.290 | 1,08 | 03:03 | 79,02% |
| 7 | Bristow, Virginia, United States | 13.445 | 38.841 | 0,51 | 00:34 | 86,32% |
| 8 | Tokyo, Japan | 10.388 | 34.192 | 1,11 | 01:26 | 88,45% |
| 9 | Holtsville, New York, United States | 10.026 | 26.095 | 1,92 | 13:25 | 74,66% |
| 10 | Atlanta, Georgia, United States | 8.884 | 40.269 | 0,39 | 00:40 | 85,06% |
| 11 | Brooklyn, New York, United States | 8.786 | 30.794 | 2,49 | 20:22 | 72,41% |
| 12 | Moscow, Russian Federation | 7.961 | 73.577 | 2,47 | 22:50 | 76,24% |
| 13 | Washington, Indiana, United States | 6.614 | 9.677 | 0,93 | 00:31 | 51,16% |
| 14 | Seattle, Washington, United States | 6.436 | 12.868 | 0,99 | 00:30 | 48,25% |
| 15 | Chicago, Illinois, United States | 6.324 | 23.548 | 1,08 | 01:14 | 89,71% |
| 16 | Colorado Springs, Colorado, United States | 6.220 | 9.005 | 0,45 | 00:59 | 93,95% |
| 17 | Menlo Park, California, United States | 5.752 | 9.670 | 0,38 | 00:54 | 95,35% |
| 18 | Santa Fe, New Mexico, United States | 5.505 | 20.101 | 1,99 | 19:00 | 72,19% |
| 19 | Piscataway, New Jersey, United States | 5.201 | 19.137 | 3,24 | 24:25:00 | 54,49% |

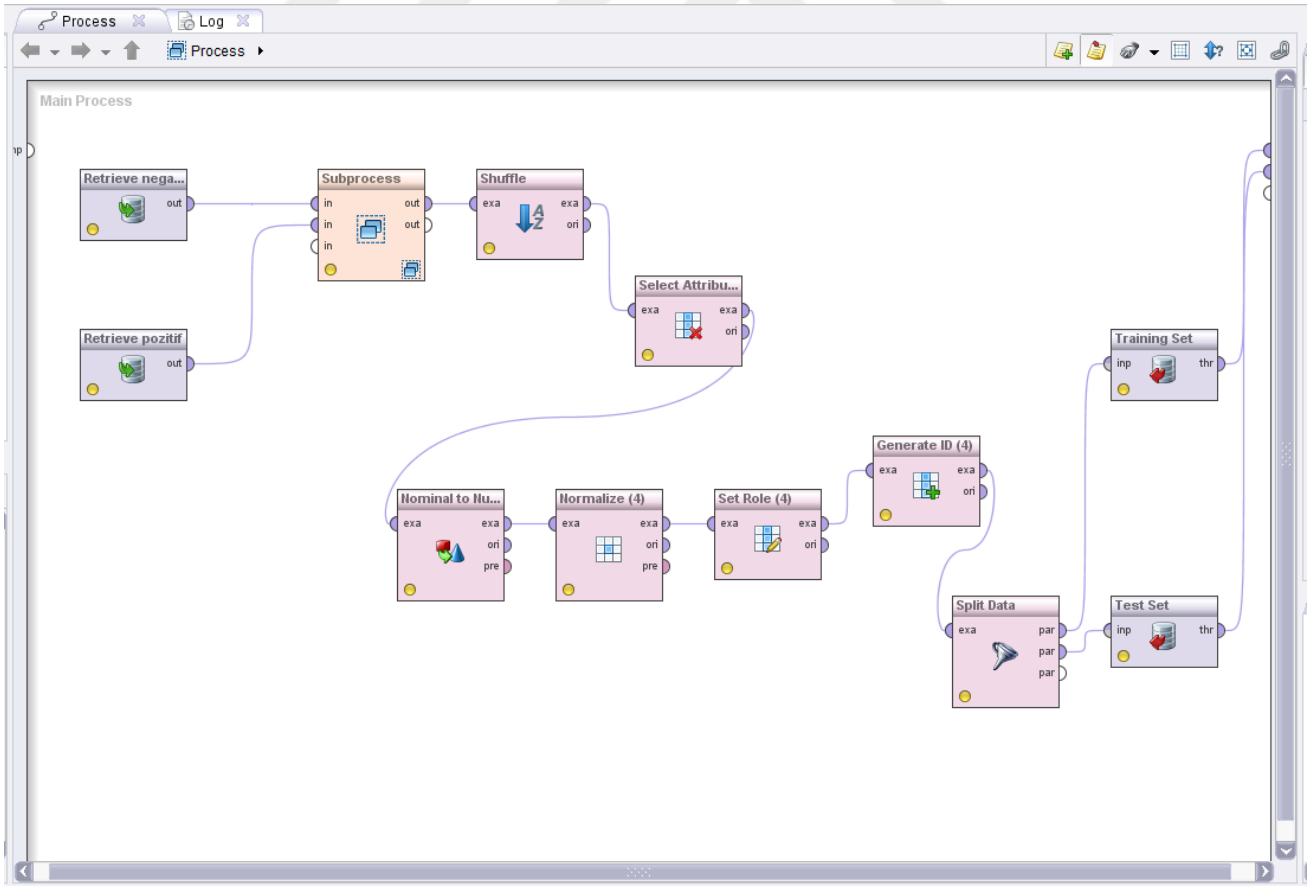
| | | | | | | |
|---|---|---|---|---|---|---|
| 20 | New Port Richey, Florida, United States | 3.430 | 6.835 | 0,02 | 00:07 | 100,00% |
| 21 | Zheng Zhou, Henan, China | 2.624 | 4.336 | 1,33 | 01:09 | 64,47% |
| 22 | Shang Hai, Shanghai, China | 2.509 | 5.281 | 0,97 | 00:28 | 71,55% |
| 23 | Izmir, Turkey | 2.364 | 333.841 | 11,35 | 04:42 | 46,65% |
| 24 | Ashburn, Virginia, United States | 2.225 | 29.828 | 1,35 | 00:56 | 84,56% |
| 25 | Cayce, South Carolina, United States | 2.173 | 2.658 | 0,44 | 00:52 | 92,58% |
| 26 | Naha-shi, Japan | 1.992 | 2.167 | 0,83 | 00:41 | 94,62% |
| 27 | San Francisco, California, United States | 1.933 | 79.877 | 1,02 | 01:28 | 94,37% |
| 28 | Slough, United Kingdom | 1.711 | 64.460 | 1,61 | 03:25 | 56,83% |
| 29 | Council Bluffs, Iowa, United States | 1.499 | 5.349 | 0,10 | 00:15 | 97,79% |
| 30 | Brussels, Belgium | 1.444 | 15.976 | 0,63 | 00:48 | 69,66% |
| 31 | Manitou Springs, Colorado, United States | 1.420 | 1.748 | 0,28 | 00:57 | 98,96% |
| 32 | Kiev, Ukraine | 1.161 | 2.511 | 1,25 | 01:51 | 89,00% |
| 33 | Sunnyvale, California, United States | 1.088 | 5.881 | 1,23 | 02:09 | 65,71% |
| 34 | Antalya, Turkey | 992 | 111.087 | 8,64 | 03:38 | 57,01% |
| 35 | Bursa, Turkey | 939 | 101.591 | 7,24 | 03:04 | 54,32% |
| 36 | Palo Alto, California, United States | 926 | 15.979 | 5,33 | 01:28:17 | 65,08% |
| 37 | Broomfield, Colorado, United States | 909 | 2.627 | 0,39 | 00:57 | 81,62% |
| 38 | Gaziantep, Turkey | 844 | 110.946 | 8,15 | 03:27 | 51,31% |
| 39 | Sisli, Turkey | 840 | 120.799 | 12,44 | 05:41 | 37,25% |
| 40 | Nuremberg, Germany | 822 | 29.358 | 12,02 | 09:50 | 58,63% |
| 41 | The Dalles, Oregon, United States | 817 | 2.009 | 0,09 | 00:11 | 95,65% |
| 42 | Izmit, Turkey | 801 | 137.633 | 12,71 | 05:20 | 42,54% |
| 43 | Xi An, Shaanxi, China | 781 | 4.806 | 0,75 | 00:11 | 53,18% |
| 44 | Shen Zhen, Guangdong, China | 780 | 9.533 | 2,14 | 01:55 | 95,39% |
| 45 | Adana, Turkey | 762 | 92.143 | 7,35 | 02:58 | 61,23% |

| | | | | | | |
|---|---|---|---|---|---|---|
| 46 | Vries, Netherlands | 750 | 4.424 | 0,38 | 00:08 | 79,90% |
| 47 | Taipei, Taiwan | 674 | 3.139 | 0,34 | 02:13 | 79,35% |
| 48 | Seongnam, Korea, Republic Of | 651 | 1.907 | 1,06 | 05:10 | 86,82% |
| 49 | Bend, Oregon, United States | 595 | 697 | 0,86 | 00:29 | 95,32% |
| 50 | Colorado City, Colorado, United States | 567 | 934 | 0,74 | 00:32 | 95,00% |
| | **Alttoplam** | **412.648** | **21.064.564** | **4,78** | **05:58** | **65,44%** |
| | **Toplam** | **453.304** | **23.535.873** | **4,87** | **06:11** | **65,29%** |

## Akıllı telefon ve tabletlerin ziyaret bilgileri

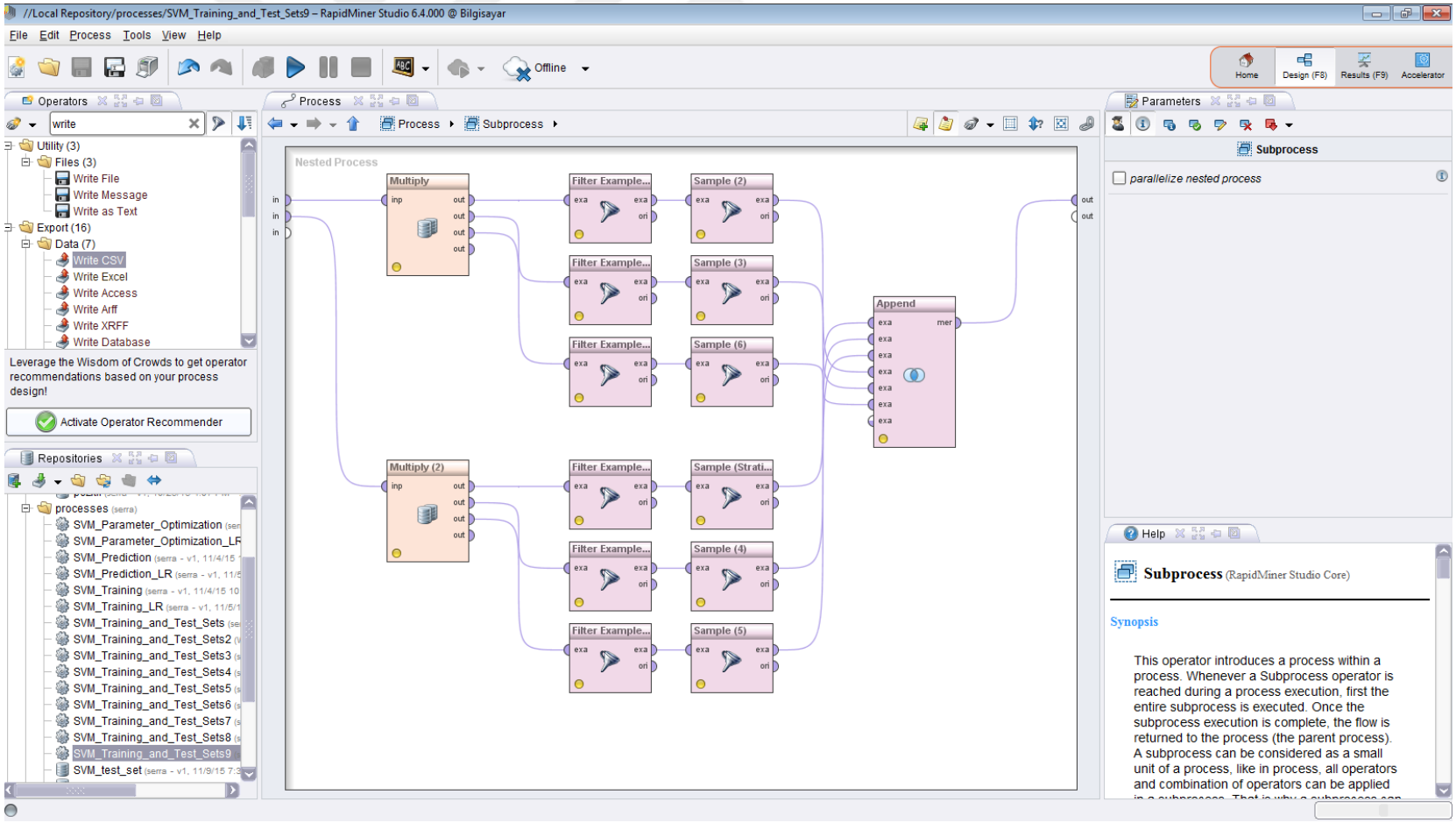
| | Mobil Aletler | Ziyaretler | Toplam Ziyaretler Yüzdesi | İstek | Toplam İstek Yüzdesi |
|---|---|---|---|---|---|
| 1 | iPhone | 15.537 | 3,43% | 900.456 | 3,83% |
| 2 | iPad | 5.707 | 1,26% | 762.406 | 3,24% |
| 3 | Samsung SGH-E250 | 500 | 0,11% | 914 | 0,00% |
| 4 | BlackBerry 9800 | 355 | 0,08% | 14.819 | 0,06% |
| 5 | BlackBerry 9700 | 275 | 0,06% | 17.679 | 0,08% |
| 6 | Samsung Galaxy S | 106 | 0,02% | 6.836 | 0,03% |
| 7 | iPod Touch | 101 | 0,02% | 11.313 | 0,05% |
| 8 | Samsung Galaxy Pad | 47 | 0,01% | 6.754 | 0,03% |
| 9 | Nokia E71 | 41 | 0,01% | 4.031 | 0,02% |
| 10 | BlackBerry 9000 | 37 | 0,01% | 1.972 | 0,01% |
| 11 | Samsung Galaxy Nexus Prime | 26 | 0,01% | 1.879 | 0,01% |
| 12 | SonyEricsson X10 | 16 | 0,00% | 2.208 | 0,01% |
| 13 | Nokia N97 | 14 | 0,00% | 2.157 | 0,01% |
| 14 | Nokia N95 | 8 | 0,00% | 537 | 0,00% |
| 15 | Motorola Xoom | 6 | 0,00% | 12 | 0,00% |
| 16 | Dell Streak | 6 | 0,00% | 814 | 0,00% |
| 17 | HTC Droid Incredible | 4 | 0,00% | 373 | 0,00% |
| 18 | Nexus One | 2 | 0,00% | 96 | 0,00% |
| 19 | Nokia N900 | 2 | 0,00% | 577 | 0,00% |
| 20 | Nexus S | 2 | 0,00% | 112 | 0,00% |
| 21 | Samsung i7500 | 2 | 0,00% | 472 | 0,00% |
| 22 | SonyEricsson W995 | 1 | 0,00% | 22 | 0,00% |
| 23 | Nokia E51 | 1 | 0,00% | 59 | 0,00% |
| 24 | HTC EVO 4G | 1 | 0,00% | 64 | 0,00% |
| 25 | HTC HD2 | 1 | 0,00% | 33 | 0,00% |
| 26 | Nokia 6630 | 1 | 0,00% | 1 | 0,00% |
| 27 | Kindle 3 | 1 | 0,00% | 77 | 0,00% |
| 28 | HTC Vision | 1 | 0,00% | 118 | 0,00% |
| | **Alttoplam** | **22.801** | **5,03%** | **1.736.791** | **7,38%** |
| | **Toplam** | **453.304** | **100,00%** | **23.535.873** | **100,00%** |

**EK:4 RapidMiner ile kurulan SVM model ve model değerleme görüntüleri**

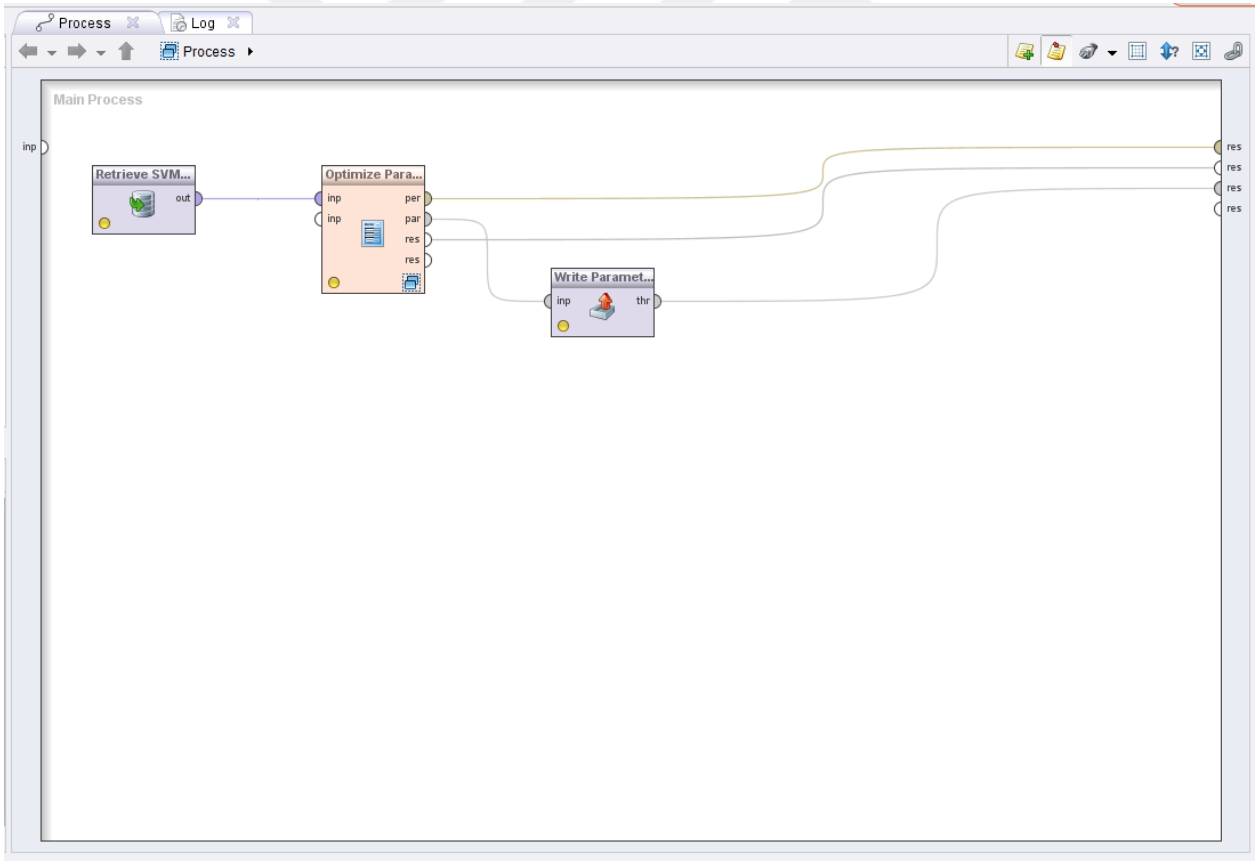



**Yeniden Örnekleme ve Eğitim-Test verisi ayrımı:**

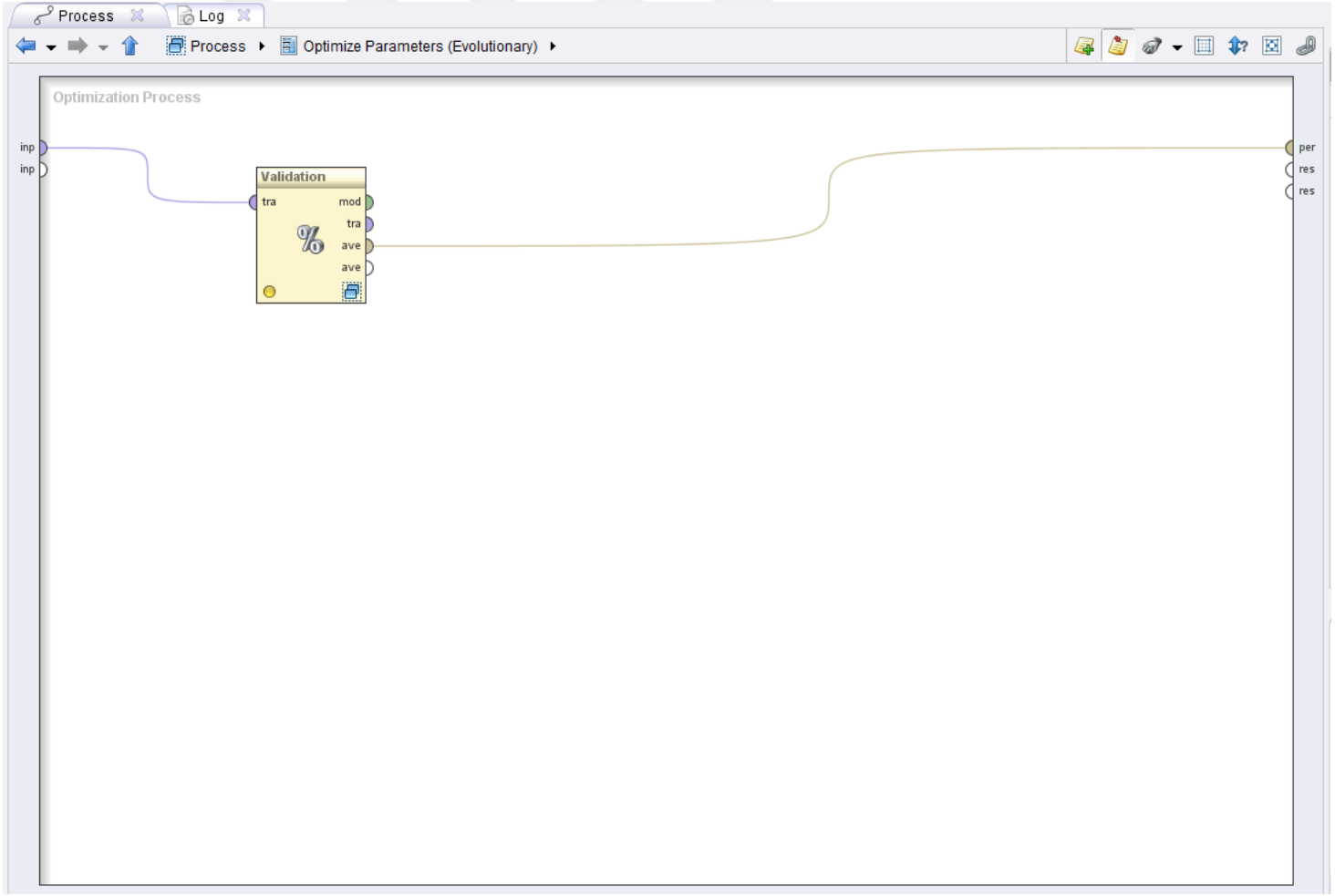
Negatif örnekler ve Pozitif örnekler iki farklı operatörle tanımlanır. Subprocess operatörüyle sonraki sayfada istenilen oranlarda örnekler çekilir. Ardından "Select Attributes" operatörüyle analizde kullanılacak nitelikler seçilir. Normalize operatörüyle birimler normalize edilir. Set Role ile satınalım niteliği "Label" edilir. Split Data ile istenilen oranda "Training" ve "Test" data üretilir.

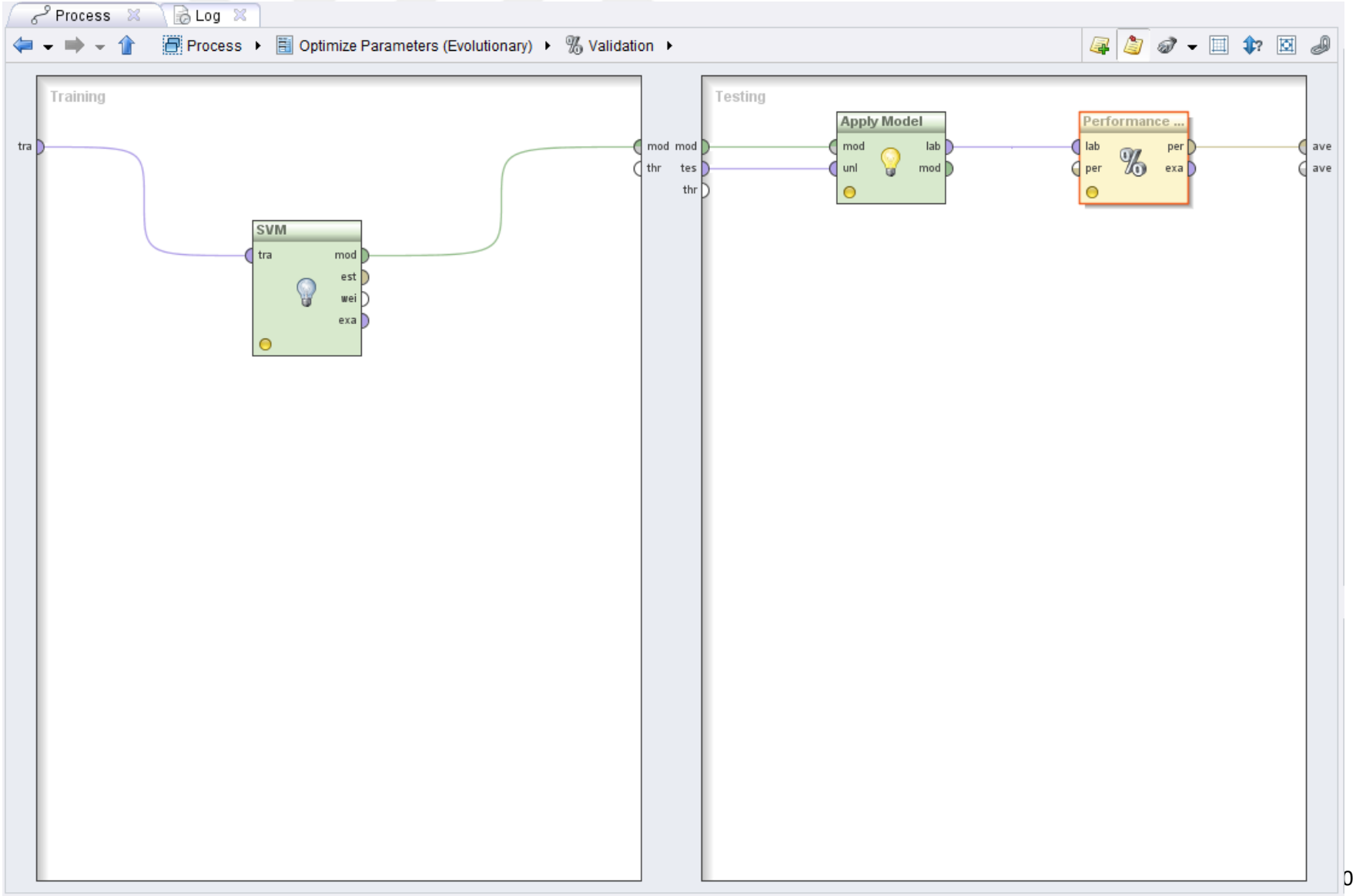



Optimum Parametrelerin Belirlenmesi Aşaması:

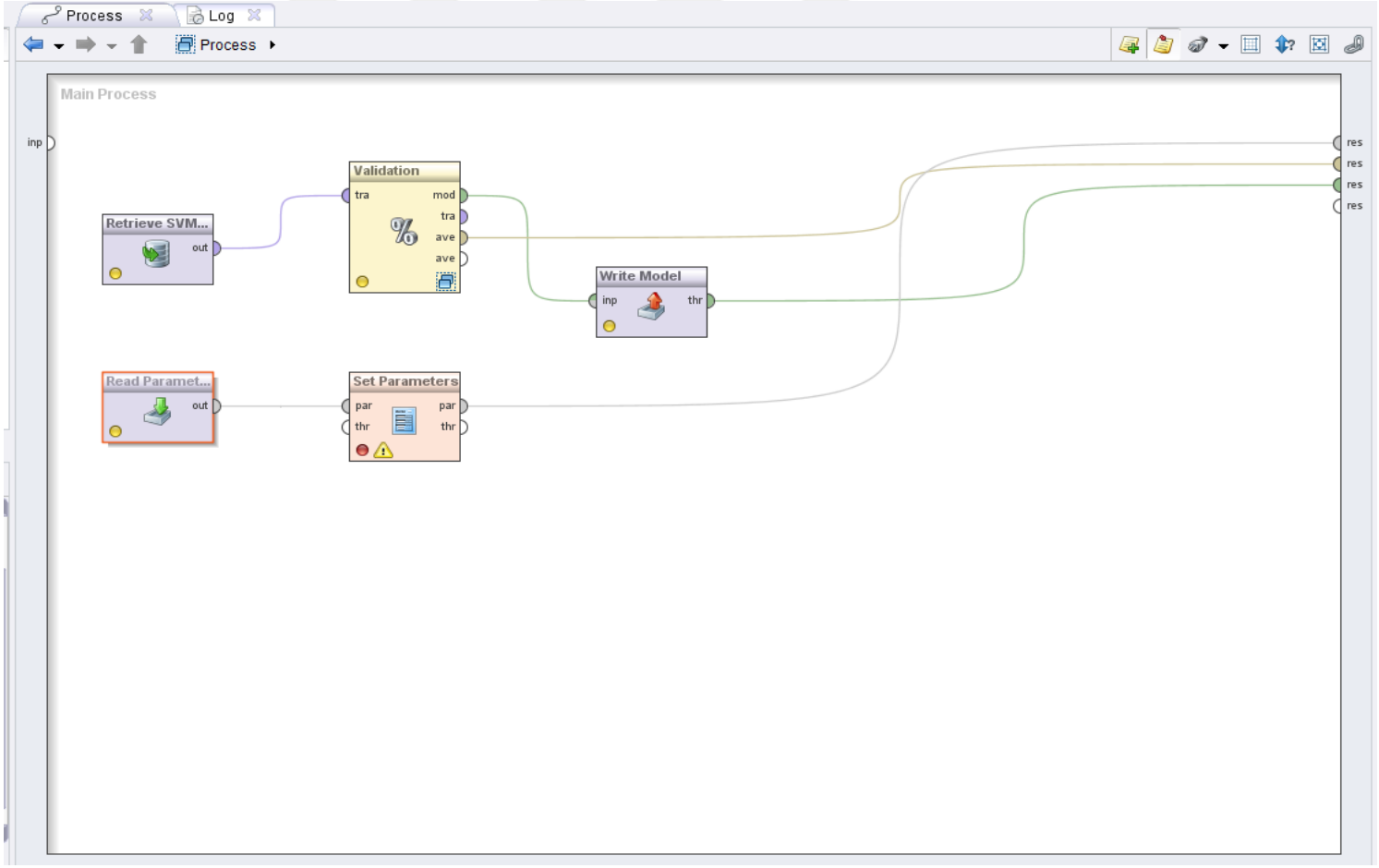


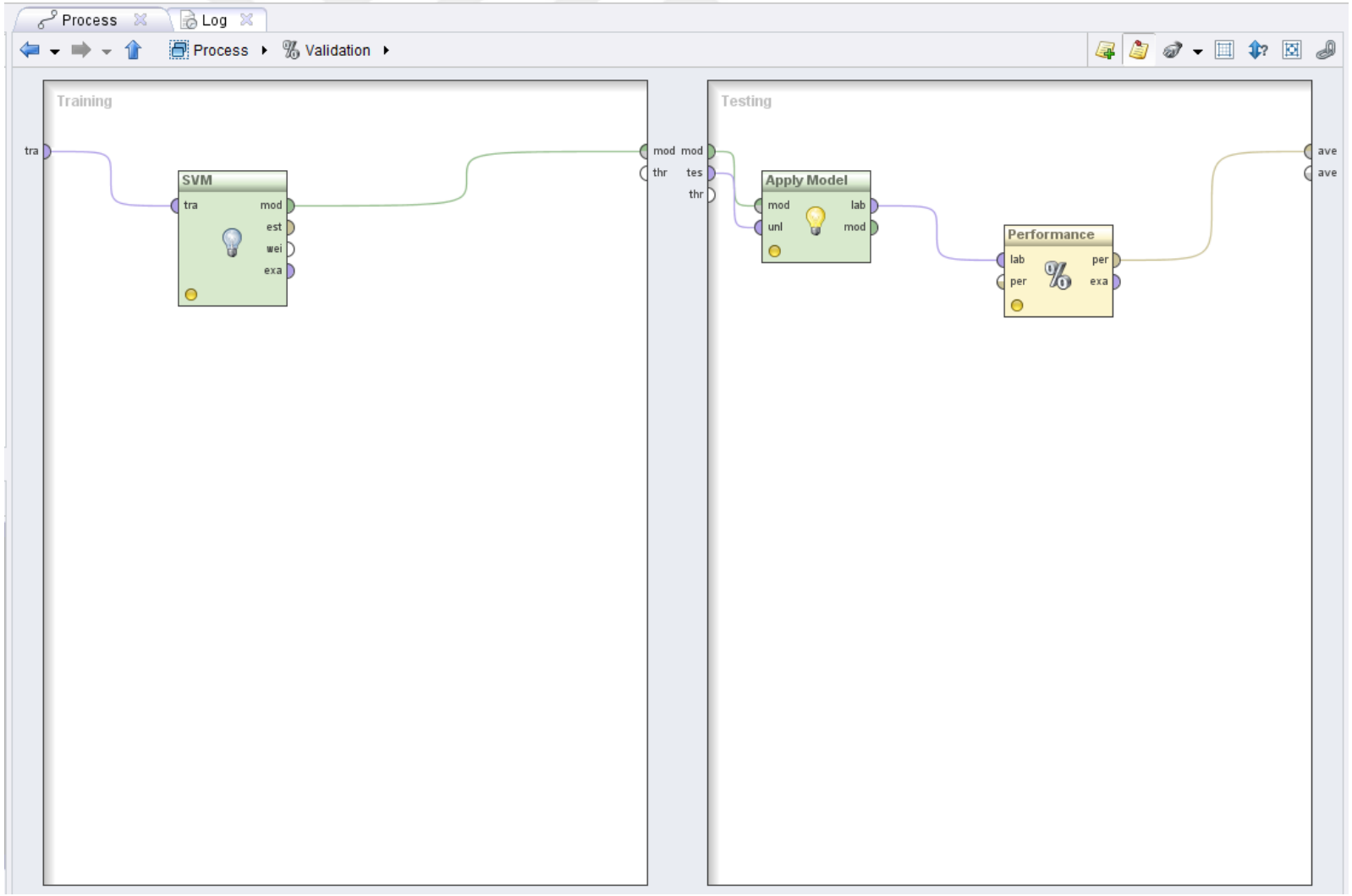

Optimize Parameter operatörü ile Validation operatörüne geçiş gerçekleştirilir.


Process
Log
Process
Optimize Parameters (Evolutionary)
Optimization Process
inp
inp
Validation
tra
mod
tra
ave
ave
per
res
res

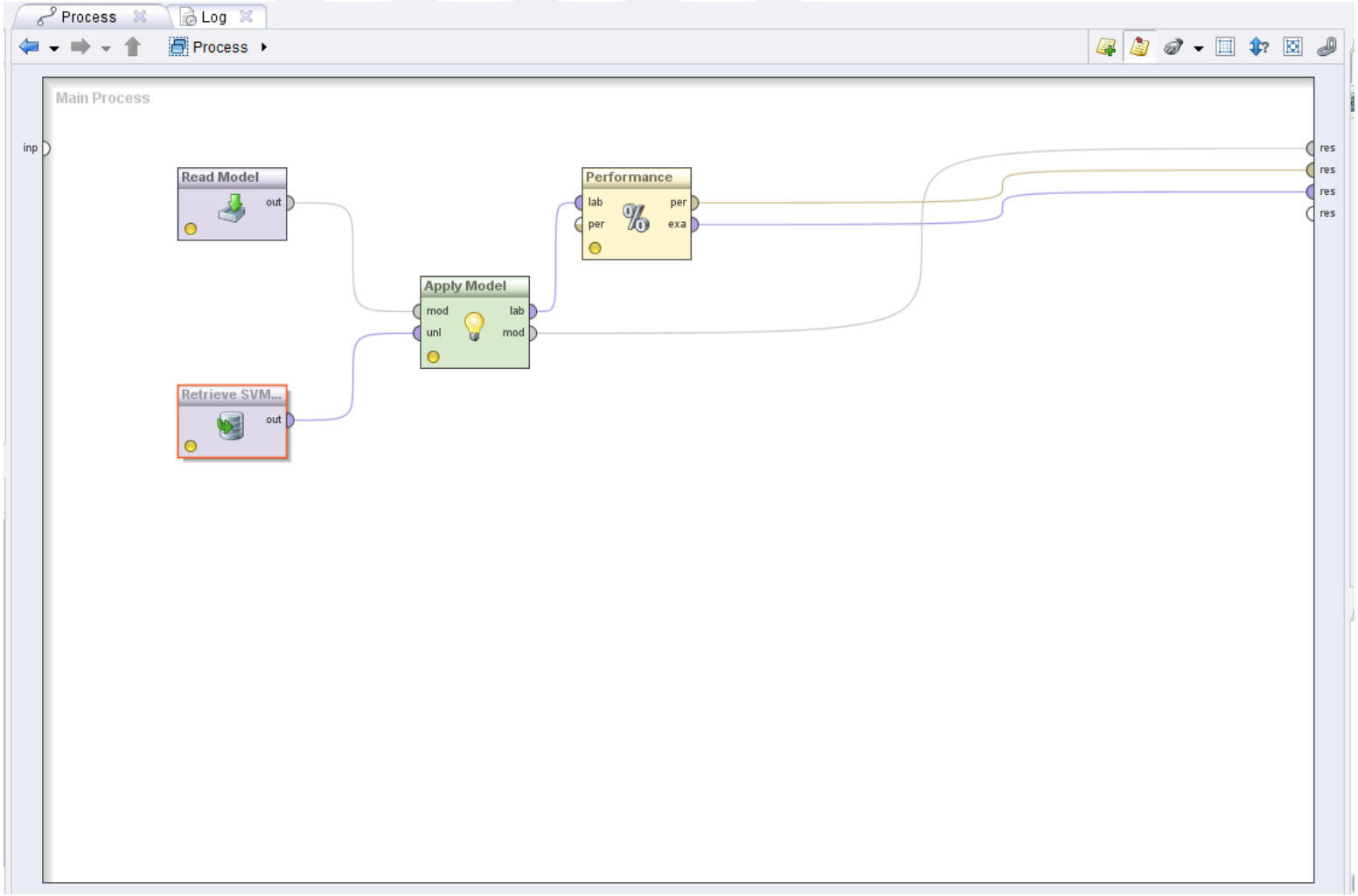

Process
Log
Process
Optimize Parameters (Evolutionary)
Validation
Training
tra
SVM
tra
mod
est
wei
exa
mod mod
thr tes
thr
Testing
Apply Model
mod
unl
lab
mod
Performance ...
lab
per
per
exa
ave
ave

Eğitim aşaması:




Process
Log
Process
Validation
Training
tra
SVM
tra
mod
est
wei
exa
mod
thr
mod
tes
thr
Testing
Apply Model
mod
unl
lab
mod
Performance
lab
per
per
exa
ave
ave

Test Aşaması:



## EK5: Model Parametreleri ve Performans Çıktıları

| Model Parametreleri | | | |
|---|---|---|---|
| **DVM Model 1a** | **DVM Model 2a** | **DVM Model 3a** | **DVM Model 4a** |
| w[ulke = Turkiye] = 386,241 | w[ulke = Turkiye] = 1173,633 | w[ulke = Turkiye] = 3139,305 | w[ulke = Turkiye] = 4091,640 |
| w[ulke = Diger] = -386,241 | w[ulke = Diger] = -1173,633 | w[ulke = Diger] = -3139,305 | w[ulke = Diger] = -4091,640 |
| w[saatdilimi = Ogle] = -167,948 | w[saatdilimi = Ogle] = 13,500 | w[saatdilimi = Ogle] = 660,294 | w[saatdilimi = Ogle] = 785,624 |
| w[saatdilimi = GeceYarisi] = 91,294 | w[saatdilimi = GeceYarisi] = -207,001 | w[saatdilimi = GeceYarisi] = -164,447 | w[saatdilimi = GeceYarisi] = -761,445 |
| w[saatdilimi = Sabah] = 260,798 | w[saatdilimi = Sabah] = 213,629 | w[saatdilimi = Sabah] = 550,810 | w[saatdilimi = Sabah] = 1141,138 |
| w[saatdilimi = Aksam] = -161,219 | w[saatdilimi = Aksam] = -162,591 | w[saatdilimi = Aksam] = -1276,787 | w[saatdilimi = Aksam] = -1727,347 |
| w[ay = Mayis] = -343,459 | w[ay = Mayis] = -67,771 | w[ay = Mayis] = -312,214 | w[ay = Mayis] = -2027,181 |
| w[ay = Haziran] = 371,156 | w[ay = Haziran] = -91,728 | w[ay = Haziran] = 1343,452 | w[ay = Haziran] = 1720,085 |
| w[ay = Temmuz] = 77,947 | w[ay = Temmuz] = -530,094 | w[ay = Temmuz] = -610,930 | w[ay = Temmuz] = -192,736 |
| w[ay = Agustos] = 154,213 | w[ay = Agustos] = -199,028 | w[ay = Agustos] = 40,296 | w[ay = Agustos] = -315,279 |
| w[ay = Eylul] = 146,642 | w[ay = Eylul] = 582,319 | w[ay = Eylul] = -148,098 | w[ay = Eylul] = -974,134 |
| w[ay = Ekim] = -186,364 | w[ay = Ekim] = -90,195 | w[ay = Ekim] = 374,829 | w[ay = Ekim] = -127,059 |
| w[ay = Kasim] = -144,598 | w[ay = Kasim] = -138,131 | w[ay = Kasim] = -416,053 | w[ay = Kasim] = -1306,313 |
| w[ay = Aralik] = 200,346 | w[ay = Aralik] = -339,388 | w[ay = Aralik] = -548,398 | w[ay = Aralik] = 1190,939 |
| w[ay = Ocak] = 16,910 | w[ay = Ocak] = 675,053 | w[ay = Ocak] = 1078,702 | w[ay = Ocak] = 109,544 |
| w[ay = Subat] = 23,173 | w[ay = Subat] = 471,381 | w[ay = Subat] = 206,367 | w[ay = Subat] = 537,577 |
| w[ay = Mart] = -286,755 | w[ay = Mart] = -158,643 | w[ay = Mart] = -352,636 | w[ay = Mart] = 1603,452 |
| w[ay = Nisan] = -4,444 | w[ay = Nisan] = -64,855 | w[ay = Nisan] = -732,171 | w[ay = Nisan] = -401,562 |
| w[gun = Cuma] = -337,498 | w[gun = Cuma] = 147,067 | w[gun = Cuma] = -1052,279 | w[gun = Cuma] = 1514,538 |
| w[gun = Pazar] = -139,926 | w[gun = Pazar] = -903,965 | w[gun = Pazar] = -516,390 | w[gun = Pazar] = -1650,090 |
| w[gun = Pazartesi] = 231,284 | w[gun = Pazartesi] = -258,383 | w[gun = Pazartesi] = 595,978 | w[gun = Pazartesi] = 1913,492 |
| w[gun = Sali] = -171,387 | w[gun = Sali] = -288,617 | w[gun = Sali] = 345,152 | w[gun = Sali] = -685,835 |
| w[gun = Persembe] = 186,464 | w[gun = Persembe] = 212,783 | w[gun = Persembe] = 94,375 | w[gun = Persembe] = -1212,756 |
| w[gun = Carsamba] = 584,708 | w[gun = Carsamba] = 1075,390 | w[gun = Carsamba] = 1185,377 | w[gun = Carsamba] = 1366,554 |
| w[gun = Cumartesi] = -459,015 | w[gun = Cumartesi] = -129,471 | w[gun = Cumartesi] = -888,281 | w[gun = Cumartesi] = -1777,597 |
| w[sayfasayisi] = -663,848 | w[sayfasayisi] = 649,785 | w[sayfasayisi] = 3233,743 | w[sayfasayisi] = 7681,074 |
| w[sure] = -665,648 | w[sure] = 1834,817 | w[sure] = 5383,462 | w[sure] = 12095,604 |

| Model Parametreleri | | | |
|---|---|---|---|
| **DVM Model 1b** | **DVM Model 2b** | **DVM Model 3b** | **DVM Model 4b** |
| Total number of Support Vectors: 234 | Total number of Support Vectors: 271 | Total number of Support Vectors: 293 | Total number of Support Vectors: 333 |
| Bias (offset): -0,308 | Bias (offset): 0,503 | Bias (offset): 0,773 | Bias (offset): -1,071 |
| w[ulke = Turkiye] = 263,841 | w[ulke = Turkiye] = 1167,113 | w[ulke = Turkiye] = 2044,862 | w[ulke = Turkiye] = 2916,074 |
| w[ulke = Diger] = -263,841 | w[ulke = Diger] = -1167,113 | w[ulke = Diger] = -2044,862 | w[ulke = Diger] = -2916,074 |
| w[saatdilimi = Ogle] = -92,552 | w[saatdilimi = Ogle] = 71,151 | w[saatdilimi = Ogle] = -385,398 | w[saatdilimi = Ogle] = 120,041 |
| w[saatdilimi = GeceYarisi] = -197,733 | w[saatdilimi = GeceYarisi] = -418,033 | w[saatdilimi = GeceYarisi] = -706,953 | w[saatdilimi = GeceYarisi] = -1110,427 |
| w[saatdilimi = Sabah] = 273,280 | w[saatdilimi = Sabah] = 422,062 | w[saatdilimi = Sabah] = 1427,513 | w[saatdilimi = Sabah] = 2033,573 |
| w[saatdilimi = Aksam] = -111,602 | w[saatdilimi = Aksam] = -361,506 | w[saatdilimi = Aksam] = -820,747 | w[saatdilimi = Aksam] = -1847,227 |
| w[ay = Mayis] = 13,093 | w[ay = Mayis] = -78,305 | w[ay = Mayis] = -524,948 | w[ay = Mayis] = 584,025 |
| w[ay = Haziran] = 70,776 | w[ay = Haziran] = 744,558 | w[ay = Haziran] = 1303,644 | w[ay = Haziran] = 669,695 |
| w[ay = Temmuz] = -194,240 | w[ay = Temmuz] = -281,814 | w[ay = Temmuz] = -96,302 | w[ay = Temmuz] = -644,851 |
| w[ay = Agustos] = 265,523 | w[ay = Agustos] = -62,857 | w[ay = Agustos] = 87,132 | w[ay = Agustos] = 479,343 |
| w[ay = Eylul] = 80,077 | w[ay = Eylul] = 155,712 | w[ay = Eylul] = 654,468 | w[ay = Eylul] = 519,544 |
| w[ay = Ekim] = -160,496 | w[ay = Ekim] = 30,001 | w[ay = Ekim] = 391,694 | w[ay = Ekim] = -335,016 |
| w[ay = Kasim] = -44,248 | w[ay = Kasim] = -751,372 | w[ay = Kasim] = -1179,747 | w[ay = Kasim] = -769,303 |
| w[ay = Aralik] = -167,628 | w[ay = Aralik] = 513,678 | w[ay = Aralik] = 727,356 | w[ay = Aralik] = -75,781 |
| w[ay = Ocak] = 57,234 | w[ay = Ocak] = -146,855 | w[ay = Ocak] = -54,298 | w[ay = Ocak] = 1380,415 |
| w[ay = Subat] = 195,895 | w[ay = Subat] = 82,070 | w[ay = Subat] = -67,873 | w[ay = Subat] = 203,228 |
| w[ay = Mart] = -123,982 | w[ay = Mart] = -116,547 | w[ay = Mart] = -495,357 | w[ay = Mart] = -532,382 |
| w[ay = Nisan] = 43,390 | w[ay = Nisan] = -255,511 | w[ay = Nisan] = -955,903 | w[ay = Nisan] = -1542,054 |
| w[gun = Cuma] = -96,412 | w[gun = Cuma] = -52,520 | w[gun = Cuma] = -272,900 | w[gun = Cuma] = 471,350 |
| w[gun = Pazar] = -102,762 | w[gun = Pazar] = -277,901 | w[gun = Pazar] = -1087,220 | w[gun = Pazar] = -1872,627 |
| w[gun = Pazartesi] = 269,549 | w[gun = Pazartesi] = -9,006 | w[gun = Pazartesi] = 902,297 | w[gun = Pazartesi] = 1531,266 |
| w[gun = Sali] = -56,596 | w[gun = Sali] = -72,264 | w[gun = Sali] = -163,362 | w[gun = Sali] = -1096,614 |
| w[gun = Persembe] = -47,363 | w[gun = Persembe] = 8,821 | w[gun = Persembe] = -436,392 | w[gun = Persembe] = -114,867 |
| w[gun = Carsamba] = -58,892 | w[gun = Carsamba] = 591,198 | w[gun = Carsamba] = 1711,642 | w[gun = Carsamba] = 1511,614 |
| w[gun = Cumartesi] = 74,408 | w[gun = Cumartesi] = -304,799 | w[gun = Cumartesi] = -1003,048 | w[gun = Cumartesi] = -810,535 |
| w[sayfasayisi] = -298,438 | w[sayfasayisi] = 549,331 | w[sayfasayisi] = 1706,352 | w[sayfasayisi] = 5360,566 |
| w[sure] = -453,928 | w[sure] = 1134,523 | w[sure] = 3488,440 | w[sure] = 8539,169 |

| Model Parametreleri | | | |
|---|---|---|---|
| **DVM Model 1c** | **DVM Model 2c** | **DVM Model 3c** | **DVM Model 4c** |
| Total number of Support Vectors: 154 | Total number of Support Vectors: 203 | Total number of Support Vectors: 222 | Total number of Support Vectors: 257 |
| Bias (offset): -0,182 | Bias (offset): -0,456 | Bias (offset): -0,817 | Bias (offset): -0,990 |
| w[ulke = Turkiye] = 108,594 | w[ulke = Turkiye] = 413,817 | w[ulke = Turkiye] = 758,910 | w[ulke = Turkiye] = 1538,579 |
| w[ulke = Diger] = -108,594 | w[ulke = Diger] = -413,817 | w[ulke = Diger] = -758,910 | w[ulke = Diger] = -1538,579 |
| w[saatdilimi = Ogle] = 33,824 | w[saatdilimi = Ogle] = -4,741 | w[saatdilimi = Ogle] = 203,942 | w[saatdilimi = Ogle] = 268,354 |
| w[saatdilimi = Sabah] = 8,807 | w[saatdilimi = Sabah] = 160,647 | w[saatdilimi = Sabah] = 88,722 | w[saatdilimi = Sabah] = 822,750 |
| w[saatdilimi = GeceYarisi] = -114,034 | w[saatdilimi = GeceYarisi] = -102,297 | w[saatdilimi = GeceYarisi] = -69,222 | w[saatdilimi = GeceYarisi] = -1004,380 |
| w[saatdilimi = Aksam] = 7,383 | w[saatdilimi = Aksam] = -131,334 | w[saatdilimi = Aksam] = -297,654 | w[saatdilimi = Aksam] = -730,843 |
| w[ay = Mayis] = -16,856 | w[ay = Mayis] = -2,759 | w[ay = Mayis] = -265,791 | w[ay = Mayis] = -314,495 |
| w[ay = Haziran] = 48,917 | w[ay = Haziran] = 331,603 | w[ay = Haziran] = 215,819 | w[ay = Haziran] = 389,376 |
| w[ay = Temmuz] = -190,382 | w[ay = Temmuz] = -275,952 | w[ay = Temmuz] = 70,209 | w[ay = Temmuz] = -436,511 |
| w[ay = Agustos] = -40,215 | w[ay = Agustos] = -153,404 | w[ay = Agustos] = 181,639 | w[ay = Agustos] = 170,423 |
| w[ay = Eylul] = 61,047 | w[ay = Eylul] = 364,676 | w[ay = Eylul] = 366,693 | w[ay = Eylul] = 614,192 |
| w[ay = Ekim] = 62,530 | w[ay = Ekim] = 214,502 | w[ay = Ekim] = 6,132 | w[ay = Ekim] = 499,486 |
| w[ay = Kasim] = -132,134 | w[ay = Kasim] = -74,441 | w[ay = Kasim] = -248,095 | w[ay = Kasim] = -549,327 |
| w[ay = Aralik] = 168,532 | w[ay = Aralik] = -99,032 | w[ay = Aralik] = 13,452 | w[ay = Aralik] = 86,897 |
| w[ay = Ocak] = 188,583 | w[ay = Ocak] = 277,227 | w[ay = Ocak] = 154,272 | w[ay = Ocak] = 507,091 |
| w[ay = Subat] = -14,211 | w[ay = Subat] = -15,685 | w[ay = Subat] = 113,254 | w[ay = Subat] = -149,002 |
| w[ay = Mart] = -87,683 | w[ay = Mart] = -35,998 | w[ay = Mart] = -175,528 | w[ay = Mart] = -124,316 |
| w[ay = Nisan] = -59,249 | w[ay = Nisan] = -529,979 | w[ay = Nisan] = -498,976 | w[ay = Nisan] = -692,310 |
| w[gun = Cuma] = -69,494 | w[gun = Cuma] = -15,519 | w[gun = Cuma] = 71,916 | w[gun = Cuma] = 45,682 |
| w[gun = Pazartesi] = 61,321 | w[gun = Pazartesi] = 430,482 | w[gun = Pazartesi] = 973,736 | w[gun = Pazartesi] = 908,792 |
| w[gun = Sali] = -14,975 | w[gun = Sali] = -289,249 | w[gun = Sali] = -625,649 | w[gun = Sali] = -295,356 |
| w[gun = Carsamba] = 36,352 | w[gun = Carsamba] = -102,915 | w[gun = Carsamba] = 203,671 | w[gun = Carsamba] = 300,732 |
| w[gun = Cumartesi] = -18,018 | w[gun = Cumartesi] = -211,204 | w[gun = Cumartesi] = -37,894 | w[gun = Cumartesi] = -194,076 |
| w[gun = Persembe] = 90,808 | w[gun = Persembe] = 140,846 | w[gun = Persembe] = -14,952 | w[gun = Persembe] = 371,716 |
| w[gun = Pazar] = -105,730 | w[gun = Pazar] = -36,186 | w[gun = Pazar] = -741,099 | w[gun = Pazar] = -1309,382 |
| w[sayfasayisi] = -259,288 | w[sayfasayisi] = 396,860 | w[sayfasayisi] = 1031,187 | w[sayfasayisi] = 2353,143 |
| w[sure] = -343,750 | w[sure] = 589,360 | w[sure] = 1996,191 | w[sure] = 4161,522 |

| Model Parametreleri | | | |
|---|---|---|---|
| **DVM Model 1d** | **DVM Model 2d** | **DVM Model 3d** | **DVM Model 4d** |
| Total number of Support Vectors: 91 | Total number of Support Vectors: 99 | Total number of Support Vectors: 126 | Total number of Support Vectors: 128 |
| Bias (offset): -0,296 | Bias (offset): -0,519 | Bias (offset): 0,913 | Bias (offset): -1,103 |
| w[ulke = Turkiye] = 63,626 | w[ulke = Turkiye] = 113,551 | w[ulke = Turkiye] = 148,771 | w[ulke = Turkiye] = 271,570 |
| w[ulke = Diger] = -63,626 | w[ulke = Diger] = -113,551 | w[ulke = Diger] = -148,771 | w[ulke = Diger] = -271,570 |
| w[saatdilimi = Ogle] = -24,270 | w[saatdilimi = Ogle] = -79,533 | w[saatdilimi = Ogle] = 144,993 | w[saatdilimi = Ogle] = 37,711 |
| w[saatdilimi = Sabah] = -12,800 | w[saatdilimi = Sabah] = 50,604 | w[saatdilimi = Sabah] = -73,379 | w[saatdilimi = Sabah] = 64,200 |
| w[saatdilimi = GeceYarisi] = -45,990 | w[saatdilimi = GeceYarisi] = -44,519 | w[saatdilimi = GeceYarisi] = -44,656 | w[saatdilimi = GeceYarisi] = -228,951 |
| w[saatdilimi = Aksam] = 59,880 | w[saatdilimi = Aksam] = 57,551 | w[saatdilimi = Aksam] = -57,251 | w[saatdilimi = Aksam] = -2,689 |
| w[ay = Mayis] = -9,188 | w[ay = Mayis] = 17,721 | w[ay = Mayis] = -48,979 | w[ay = Mayis] = 225,989 |
| w[ay = Haziran] = 84,188 | w[ay = Haziran] = -81,773 | w[ay = Haziran] = 126,494 | w[ay = Haziran] = 170,236 |
| w[ay = Temmuz] = 26,024 | w[ay = Temmuz] = -102,343 | w[ay = Temmuz] = -50,334 | w[ay = Temmuz] = -283,817 |
| w[ay = Agustos] = -36,983 | w[ay = Agustos] = -27,644 | w[ay = Agustos] = 17,208 | w[ay = Agustos] = -275,315 |
| w[ay = Eylul] = -24,615 | w[ay = Eylul] = -21,468 | w[ay = Eylul] = -98,814 | w[ay = Eylul] = -253,231 |
| w[ay = Ekim] = 18,690 | w[ay = Ekim] = 30,671 | w[ay = Ekim] = -17,664 | w[ay = Ekim] = 41,973 |
| w[ay = Kasim] = -51,750 | w[ay = Kasim] = -38,266 | w[ay = Kasim] = -12,798 | w[ay = Kasim] = -46,795 |
| w[ay = Aralik] = 41,496 | w[ay = Aralik] = 182,894 | w[ay = Aralik] = 52,520 | w[ay = Aralik] = 383,118 |
| w[ay = Ocak] = 6,114 | w[ay = Ocak] = 100,333 | w[ay = Ocak] = 113,826 | w[ay = Ocak] = 157,681 |
| w[ay = Subat] = -54,453 | w[ay = Subat] = -26,892 | w[ay = Subat] = -83,427 | w[ay = Subat] = 172,018 |
| w[ay = Mart] = 40,079 | w[ay = Mart] = 13,498 | w[ay = Mart] = 39,509 | w[ay = Mart] = -75,333 |
| w[ay = Nisan] = -58,231 | w[ay = Nisan] = -43,586 | w[ay = Nisan] = -50,027 | w[ay = Nisan] = -228,223 |
| w[gun = Cuma] = -74,417 | w[gun = Cuma] = -130,819 | w[gun = Cuma] = 28,451 | w[gun = Cuma] = 163,382 |
| w[gun = Pazartesi] = 2,624 | w[gun = Pazartesi] = 87,505 | w[gun = Pazartesi] = 36,365 | w[gun = Pazartesi] = 46,004 |
| w[gun = Carsamba] = -9,745 | w[gun = Carsamba] = -1,742 | w[gun = Carsamba] = 85,825 | w[gun = Carsamba] = 189,276 |
| w[gun = Persembe] = 6,100 | w[gun = Persembe] = 76,439 | w[gun = Persembe] = 5,199 | w[gun = Persembe] = -29,430 |
| w[gun = Cumartesi] = 6,073 | w[gun = Cumartesi] = 84,538 | w[gun = Cumartesi] = 89,069 | w[gun = Cumartesi] = -124,523 |
| w[gun = Pazar] = 46,162 | w[gun = Pazar] = -49,964 | w[gun = Pazar] = -168,387 | w[gun = Pazar] = -176,041 |
| w[gun = Sali] = 46,580 | w[gun = Sali] = -46,839 | w[gun = Sali] = -92,311 | w[gun = Sali] = -132,368 |
| w[sayfasayisi] = -81,967 | w[sayfasayisi] = -13,571 | w[sayfasayisi] = 457,279 | w[sayfasayisi] = 482,491 |
| w[sure] = -52,076 | w[sure] = 91,647 | w[sure] = 664,885 | w[sure] = 893,077 |

| Model Parametreleri | | | |
|---|---|---|---|
| **DVM Model 1e** | **DVM Model 2e** | **DVM Model 3e** | **DVM Model 4e** |
| Total number of Support Vectors: 49 | Total number of Support Vectors: 63 | Total number of Support Vectors: 66 | Total number of Support Vectors: 69 |
| Bias (offset): 0,085 | Bias (offset): -0,447 | Bias (offset): -0,863 | Bias (offset): -1,027 |
| w[ulke = Turkiye] = 6,044 | w[ulke = Turkiye] = 10,762 | w[ulke = Turkiye] = 13,857 | w[ulke = Turkiye] = 39,999 |
| w[ulke = Diger] = -6,044 | w[ulke = Diger] = -10,762 | w[ulke = Diger] = -13,857 | w[ulke = Diger] = -39,999 |
| w[saatdilimi = Ogle] = -15,606 | w[saatdilimi = Ogle] = -50,006 | w[saatdilimi = Ogle] = -12,937 | w[saatdilimi = Ogle] = -48,135 |
| w[saatdilimi = Sabah] = -2,242 | w[saatdilimi = Sabah] = 0,847 | w[saatdilimi = Sabah] = 28,162 | w[saatdilimi = Sabah] = 49,251 |
| w[saatdilimi = GeceYarisi] = -4,659 | w[saatdilimi = GeceYarisi] = 9,369 | w[saatdilimi = GeceYarisi] = -60,820 | w[saatdilimi = GeceYarisi] = -23,383 |
| w[saatdilimi = Aksam] = 19,233 | w[saatdilimi = Aksam] = 48,395 | w[saatdilimi = Aksam] = 16,092 | w[saatdilimi = Aksam] = 13,386 |
| w[ay = Mayis] = -12,808 | w[ay = Mayis] = -5,251 | w[ay = Mayis] = 10,571 | w[ay = Mayis] = 55,299 |
| w[ay = Haziran] = -4,380 | w[ay = Haziran] = -0,582 | w[ay = Haziran] = 0,322 | w[ay = Haziran] = 18,522 |
| w[ay = Temmuz] = 28,452 | w[ay = Temmuz] = -4,976 | w[ay = Temmuz] = -29,862 | w[ay = Temmuz] = -31,623 |
| w[ay = Agustos] = -1,031 | w[ay = Agustos] = 8,707 | w[ay = Agustos] = 40,242 | w[ay = Agustos] = 6,223 |
| w[ay = Eylul] = -13,690 | w[ay = Eylul] = 0,788 | w[ay = Eylul] = 23,670 | w[ay = Eylul] = -54,782 |
| w[ay = Kasim] = -19,661 | w[ay = Kasim] = -3,494 | w[ay = Kasim] = -3,586 | w[ay = Kasim] = -58,730 |
| w[ay = Ocak] = 7,542 | w[ay = Ocak] = 15,619 | w[ay = Ocak] = -16,753 | w[ay = Ocak] = 42,867 |
| w[ay = Subat] = 3,645 | w[ay = Subat] = -16,397 | w[ay = Subat] = -9,929 | w[ay = Subat] = 21,948 |
| w[ay = Mart] = 20,558 | w[ay = Mart] = 22,827 | w[ay = Mart] = 38,598 | w[ay = Mart] = -45,614 |
| w[ay = Ekim] = 13,131 | w[ay = Ekim] = 21,118 | w[ay = Ekim] = 17,837 | w[ay = Ekim] = -1,214 |
| w[ay = Aralik] = -18,772 | w[ay = Aralik] = 0,099 | w[ay = Aralik] = -0,378 | w[ay = Aralik] = 94,983 |
| w[ay = Nisan] = -11,547 | w[ay = Nisan] = -40,756 | w[ay = Nisan] = -70,494 | w[ay = Nisan] = -50,562 |
| w[gun = Cuma] = 1,318 | w[gun = Cuma] = -11,887 | w[gun = Cuma] = -16,624 | w[gun = Cuma] = -32,877 |
| w[gun = Pazartesi] = -29,398 | w[gun = Pazartesi] = 18,130 | w[gun = Pazartesi] = 13,006 | w[gun = Pazartesi] = -29,299 |
| w[gun = Carsamba] = -7,939 | w[gun = Carsamba] = -17,417 | w[gun = Carsamba] = -2,340 | w[gun = Carsamba] = 93,584 |
| w[gun = Persembe] = 5,962 | w[gun = Persembe] = 41,447 | w[gun = Persembe] = -34,548 | w[gun = Persembe] = 47,029 |
| w[gun = Cumartesi] = 42,066 | w[gun = Cumartesi] = -21,678 | w[gun = Cumartesi] = -8,787 | w[gun = Cumartesi] = 1,431 |
| w[gun = Pazar] = -12,449 | w[gun = Pazar] = -6,639 | w[gun = Pazar] = -6,385 | w[gun = Pazar] = -54,164 |
| w[gun = Sali] = 13,131 | w[gun = Sali] = 0,690 | w[gun = Sali] = 50,783 | w[gun = Sali] = -31,675 |
| w[sayfasayisi] = -7,690 | w[sayfasayisi] = -6,230 | w[sayfasayisi] = 68,960 | w[sayfasayisi] = 128,201 |
| w[sure] = 5,882 | w[sure] = 52,658 | w[sure] = 149,225 | w[sure] = 266,616 |

**EK5: Model Parametreleri ve Performans Çıktıları**

**LR Model 1a:**

0.15 + [saatdilimi = Aksam] * -0.07 + [ay = Haziran] * 0.09 + [ay = Ekim] * 0.08 + [ay = Aralik] * 0.08 + [ay = Ocak] * 0.04 + [ay = Subat] * -0.05 + [ay = Mart] * -0.05 + [gun = Pazartesi] * 0.09 + [gun = Carsamba] * 0.12 + [gun = Cumartesi] * 0.06 + [ulke = Turkiye] * 0.11 + [sayfasayisi] * 1.42 + [sure] * 1.11

**LR Model 2a:**

-0.28 + [saatdilimi = Sabah] * 0.08 + [ay = Subat] * -0.06 + [ay = Mart] * -0.07 + [gun = Pazartesi] * 0.09 + [gun = Carsamba] * 0.16 + [gun = Cuma] * 0.07 + [sayfasayisi] * 1.26 + [sure] * 1

**LR Model 3a:**

-1.01 + [saatdilimi = Ogle] * 0.01 + [saatdilimi = Sabah] * -0.03 + [saatdilimi = GeceYarisi] * 0.02 + [saatdilimi = Aksam] * -0.07 + [ay = Haziran] * 0.09 + [ay = Temmuz] * -0.05 + [ay = Agustos] * -0.09 + [ay = Ekim] * 0.03 + [ay = Kasim] * 0.02 + [ay = Aralik] * 0.02 + [ay = Subat] * -0.02 + [ay = Mart] * 0.02 + [ay = Nisan] * -0.06 + [gun = Pazartesi] * 0.06 + [gun = Carsamba] * 0.11 + [gun = Cumartesi] * -0.02 + [gun = Persembe] * -0.05 + [gun = Sali] * 0.03 + [ulke = Turkiye] * 0.15 + [sayfasayisi] * 0.53 + [sure] * 0.92

**LR Model 4a:**

-1.25 + [gun = Cuma] * 0.07 + [gun = Carsamba] * 0.12 + [ulke = Turkiye] * 0.11 + [sayfasayisi] * 0.31 + [sure] * 0.77

**LR Model 1b: 300:300**

-0.16 + [ulke = Turkiye] * -0.44 + [saatdilimi = Ogle] * -0.11 + [saatdilimi = GeceYarisi] * 0.08 + [saatdilimi = Sabah] * -0.16 + [saatdilimi = Aksam] * 0.02 + [ay = Mayis] * -0.1 + [ay = Haziran] * 0.02 + [ay = Agustos] * -0.02 + [ay = Eylul] * 0.16 + [ay = Ekim] * -0.1 + [ay = Aralik] * -0.02 + [ay = Ocak] * -0.13 + [ay = Subat] * 0.04 + [ay = Mart] * 0.14 + [ay = Nisan] * 0.12 + [gun = Pazar] * 0.14 + [gun = Pazartesi] * -0.06 + [gun = Sali] * 0.09 + [gun = Carsamba] * -0.1 + [gun = Cumartesi] * 0.13 + [sayfasayisi] * -0.86 + [sure] * -0.73

**LR Model 2b:300:600**

0.45 +  [ulke = Turkiye] * -0.35 + [saatdilimi = Ogle] * -0.08 + [saatdilimi = GeceYarisi] * 0.09 + [saatdilimi = Sabah] * -0.11 + [saatdilimi = Aksam] * 0.06 + [ay = Temmuz] * 0.06 + [ay = Eylul] * 0.03 + [ay = Ekim] * -0.05 + [ay = Aralik] * -0.04 + [ay = Ocak] * -0.07 + [ay = Mart] * 0.07 + [ay = Nisan] * 0.13 + [gun = Cuma] * 0.02 + [gun = Pazar] * 0.09 + [gun = Pazartesi] * -0.04 + [gun = Sali] * 0.03 + [gun = Persembe] * -0.05 + [gun = Carsamba] * -0.06 + [gun = Cumartesi] * 0.11 + [sayfasayisi] * -0.96 + [sure] * -0.56

**LR Model 3b: 300:900**

0.87 +  [ulke = Turkiye] * -0.51 + [saatdilimi = Ogle] * -0.04 + [saatdilimi = Sabah] * -0.11 + [saatdilimi = Aksam] * 0.07 + [ay = Haziran] * -0.08 + [ay = Eylul] * 0.08 + [ay = Ekim] * -0.11 + [ay = Aralik] * -0.04 + [ay = Ocak] * -0.07 + [gun = Pazar] * 0.12 + [gun = Pazartesi] * -0.05 + [gun = Persembe] * -0.05 + [gun = Cumartesi] * 0.12 + [sayfasayisi] * -0.72 + [sure] * -0.51

**LR Model 4b: 300:1200**

1.07 +  [ulke = Turkiye] * -0.45 + [saatdilimi = Ogle] * -0.03 + [saatdilimi = GeceYarisi] * 0.06 + [saatdilimi = Sabah] * -0.09 + [saatdilimi = Aksam] * 0.06 + [ay = Mayis] * -0.04 + [ay = Haziran] * 0.03 + [ay = Eylul] * 0.06 + [ay = Ekim] * -0.08 + [ay = Aralik] * -0.07 + [ay = Ocak] * -0.07 + [ay = Nisan] * 0.07 + [gun = Pazar] * 0.15 + [gun = Pazartesi] * -0.02 + [gun = Sali] * 0.04 + [gun = Persembe] * 0.01 + [gun = Carsamba] * -0.02 + [gun = Cumartesi] * 0.09 + [sayfasayisi] * -0.58 + [sure] * -0.55

**LR Model 1c: 200:200**

-0.24 + [ulke = Diger] * 0.27 + [gun = Cumartesi] * 0.15 + [sayfasayisi] * -0.83 + [sure] * -0.67

**LR Model 2c: 200:400**

0.39 +  [ulke = Diger] * 0.23 + [sayfasayisi] * -0.44 + [sure] * -0.59

**LR Model 3c: 200:600**

0.84 + [ulke = Turkiye] * -0.38 + [saatdilimi = Ogle] * -0.03 + [saatdilimi = Sabah] * -0.08 + [saatdilimi = GeceYarisi] * 0.02 + [saatdilimi = Aksam] * 0.01 + [ay = Mayis] * 0.01 + [ay = Haziran] * -0.05 + [ay = Temmuz] * 0.02 + [ay = Agustos] * -0.08 + [ay = Eylul] * 0.14 + [ay = Ekim] * -0.14 + [ay = Kasim] * -0.03 + [ay = Aralik] * -0.08 + [ay = Subat] * -0.02 + [ay = Mart] * 0.03 + [ay = Nisan] * 0.25 + [gun = Cuma] * 0.01 + [gun = Pazartesi] * -0.11 + [gun = Sali] * 0.06 + [gun = Carsamba] * -0.06 + [gun = Persembe] * -0.08 + [gun = Pazar] * 0.1 + [gun = Cumartesi] * 0.12 + [sayfasayisi] * -0.89 + [sure] * -0.52

**LR Model 4c: 200:800**

0.88 + [ulke = Turkiye] * -0.34 + [ay = Ekim] * -0.12 + [sayfasayisi] * -0.5 + [sure] * -0.37

**LR Model 1d: 100:100**

-0.65 + [ulke = Turkiye] * -0.5 + [saatdilimi = Ogle] * -0.09 + [saatdilimi = GeceYarisi] * 0.2 + [ay = Mayis] * -0.31 + [ay = Temmuz] * 0.21 + [ay = Eylul] * 0.12 + [ay = Aralik] * -0.1 + [ay = Ocak] * -0.21 + [ay = Subat] * -0.15 + [gun = Pazartesi] * -0.08 + [gun = Persembe] * -0.07 + [gun = Cumartesi] * 0.17 + [gun = Pazar] * 0.1 + [gun = Sali] * 0.07 + [sayfasayisi] * -1.96 + [sure] * -0.28

**LR Model 2d: 100:200**

0.43 + [ulke = Turkiye] * -0.15 + [ay = Temmuz] * 0.2 + [ay = Eylul] * 0.64 + [ay = Aralik] * -0.11 + [ay = Subat] * -0.15 + [ay = Nisan] * 0.45 + [gun = Cumartesi] * 0.18 + [gun = Sali] * 0.14 + [sayfasayisi] * -1.41 + [sure] * -0.4

**LR Model 3d: 100:300**

0.86 + [ulke = Turkiye] * -0.29 + [saatdilimi = GeceYarisi] * 0.12 + [ay = Haziran] * 0.05 + [ay = Temmuz] * 0.15 + [ay = Eylul] * 0.56 + [ay = Kasim] * -0.08 + [ay = Aralik] * -0.08 + [ay = Ocak] * -0.07 + [ay = Subat] * -0.09 + [ay = Nisan] * 0.13 + [gun = Persembe] * -0.13 + [gun = Cumartesi] * 0.09 + [gun = Pazar] * 0.07 + [sayfasayisi] * -0.63 + [sure] * -0.51

**LR Model 4d: 100:400**

1.16 + [ulke = Turkiye] * -0.33 + [saatdilimi = GeceYarisi] * 0.33 + [ay = Mayis] * 0.09 + [ay = Temmuz] * 0.19 + [ay = Eylul] * 0.65 + [ay = Ekim] * -0.06 + [ay = Kasim] * -0.05 + [ay = Aralik] * -0.17 + [ay = Ocak] * -0.07 + [ay = Subat] * -0.07 + [ay = Nisan] * 0.27 + [gun = Pazartesi] * -0.09 + [gun = Persembe] * -0.09 + [gun = Cumartesi] * 0.1 + [gun = Pazar] * 0.23 + [gun = Sali] * 0.03 + [sayfasayisi] * -1.02 + [sure] * -0.46

**LR Model 1e: 50:50**

-0.71 + [ulke = Turkiye] * -0.29 + [saatdilimi = Sabah] * 0.33 + [ay = Temmuz] * 0.4 + [gun = Persembe] * -0.24 + [sayfasayisi] * -2.98

**LR Model 2e: 50:100**

0.43 + [saatdilimi = Ogle] * 0.27 + [ay = Temmuz] * 0.27 + [sayfasayisi] * -1.64 + [sure] * -0.47

**LR Model 3e: 50:150**

1.13 + [ulke = Turkiye] * -0.19 + [ay = Temmuz] * 0.44 + [ay = Agustos] * -0.28 + [ay = Eylul] * 0.39 + [ay = Kasim] * -0.17 + [ay = Subat] * -0.16 + [ay = Mart] * 0.13 + [ay = Aralik] * -0.12 + [ay = Nisan] * 0.22 + [gun = Cuma] * 0.23 + [gun = Carsamba] * -0.16 + [gun = Cumartesi] * 0.2 + [sayfasayisi] * -0.83 + [sure] * -0.97

**LR Model 4e: 50:200**

1.12 + [ulke = Turkiye] * -0.15 + [ay = Temmuz] * 0.17 + [ay = Agustos] * -0.2 + [ay = Eylul] * 0.2 + [ay = Subat] * -0.22 + [ay = Nisan] * 0.23 + [gun = Carsamba] * -0.2 + [gun = Cumartesi] * 0.19 + [sayfasayisi] * -0.75 + [sure] * -0.58

# ÖZGEÇMİŞ

1981 yılında İstanbul'da dünyaya gelen Serra Çelik, ilk ve orta öğretimini İstanbul'da tamamlamıştır. 2001 yılında İstanbul Üniversitesi Fen Fakültesi Astronomi ve Uzay Bilimleri'nden mezun olmuştur. 2009 yılında İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Sayısal Yöntemler Anabilim Dalı'nda "Çağrı Merkezlerinde Performans Değerleme" isimli tez ile yüksek lisans derecesini almıştır. Aynı yıl İstanbul Üniversitesi Sosyal Bilimler Enstitüsü'nde Doktora Programı'na kabul edilmiştir. 2012 yılında İstanbul Üniversitesi Enformatik Bölümü'nde Araştırma Görevlisi olarak çalışmaya başlayan Serra Çelik, daha öncesinde özel sektörde müşteri ilişkileri yönetimi, çağrı merkezi operasyonları, monitoring, telesatış, e-ticaret alanlarında çalışmış, ayrıca web site editörlüğü ve web analistliği görevlerinde bulunmuştur.