T.C.
KIRŞEHİR AHİ EVRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İLERİ TEKNOJİLER ANABİLİM DALI

BENZETİM VERİLERİ ÜZERİNDE DAĞITILMIŞ HİZMET REDDİ SALDIRILARININ TESPİTİ İÇİN BİR YAKLAŞIM

ÖZGÜR SEZEN

YÜKSEK LİSANS TEZİ

KIRŞEHİR / 2022

T.C.
KIRŞEHİR AHİ EVRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İLERİ TEKNOJİLER ANABİLİM DALI

BENZETİM VERİLERİ ÜZERİNDE DAĞITILMIŞ HİZMET REDDİ SALDIRILARININ TESPİTİ İÇİN BİR YAKLAŞIM

ÖZGÜR SEZEN

YÜKSEK LİSANS TEZİ

DANIŞMAN
Dr. Öğr. Üyesi Memduh KÖSE

KIRŞEHİR / 2022

TEZ BİLDİRİMİ

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

Özgür SEZEN

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, Kırşehir Ahi Evran Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

ÖNSÖZ

Yüksek Lisansa başlamamda ve yüksek lisans ders sürecinde kendisini tanıdığım günden bu yana gösterdiği sakin ve sabırlı hali ile her zaman bana örnek olmasının yanı sıra bir bilim adamının nasıl çalışması gerektiğini kendisinden öğrendiğim değerli danışmanım Dr. Öğr. Üyesi Memduh KÖSE'ye büyük bir içtenlikle teşekkür ederim.

Haziran, 2022 Özgür SEZEN

İÇİNDEKİLER **Sayfa No**

ŞEKİL LİSTESİ **Sayfa No**

TABLO LİSTESİ **Sayfa No**

SİMGE VE KISALTMA LİSTESİ

Kısaltmalar	Açıklama
ÇKA	**:** Çok Katmanlı Algılayıcı
DVM	**:** Destek Vektör Makinesi
DoS/DDoS	**:** Dağtık Hizmet Engelleme (Distributed Denial of Service)

ÖZET

YÜKSEK LİSANS TEZİ

BENZETİM VERİLERİ ÜZERİNDE DAĞITILMIŞ HİZMET REDDİ SALDIRILARININ TESPİTİ İÇİN BİR YAKLAŞIM

Özgür SEZEN

Kırşehir Ahi Evran Üniversitesi
Fen Bilimleri Enstitüsü
İleri Teknolojiler Anabilim Dalı

Dr. Öğr. Üyesi Memduh KÖSE

Veri madenciliği, verilerin toplanması, temizlenmesi, işlenmesi, analiz edilmesi ve verilerden yararlı bilgiler elde edilmesi süreçlerinin tamamına verilen isimdir. Bilgisayar endüstrisinde hızlı büyüyen alanlardan biridir. Bilgisayar biliminde ve istatistikte küçük bir ilgi alanı olarak başlamış ve hızla kendi alanını oluşturmaktadır. Büyük veri kümelerine uygulanabilmesinden dolayı, perakende, üretim, telekomünikasyon, sağlık, sigorta ve ulaşım gibi neredeyse tüm sektörlerde kullanılabilmektedir. Tahmin tabanlı modelleme üzerine yapılan çalışmalar, son zamanlarda artış göstermektedir. Özellikle sağlık, hizmet, bilişim alanında modelleme tabanlı algoritmalar oldukça sık kullanılmaktadır.

Bu çalışmada amaçlanan, Dağıtılmış Hizmet Reddi saldırılarını tespit etmek için veri madenciliği temelli bir tahmin modeli oluşturmaktır. Bu model saldırıların ilk başladığı anda sisteme vereceği zararı tahmin etmek ve saldırının başarı yüzdesi hakkında tahminde bulunabilmektir.

Bu çalışmada gerçek hayat dağıtık hizmet engelleme saldırı senaryoları baz alınarak 500 durum içeren simüle veri üretilmiştir. Veri setinde bulunan bağımsız değişkenler paketlerin yaşama süresi, atak yapılan paket sayısı, düşük ağ performansı durumu, web sitelerine erişimde yavaşlık durumu, ağ bağlantılarında kesilmeler olup olmaması, spam e-postaların sayışında artış olup olmaması, web sitesinin belli bölümlerine erişimin engellenmesi durumudur, bağımlı değişken ise yapılan atağın başarı durumudur. Bu çalışmada sınıflandırma yöntemi olarak literatürde sık kullanılan ve iyi bir performansa sahip Random Forest yöntemi ele alınmıştır.

Haziran 2022, 52 Sayfa

Anahtar Kelimeler: Veri Madenciliği, Veri Madenciliği Yöntemleri, Sınıflandırma, Dağıtık Hizmet Engelleme

ABSTRACT

M.Sc. THESIS

AN APPROACH TO DETECTING DISTRIBUTED DENTAL ATTACKS ON SIMULATION DATA

Özgür SEZEN

Kırsehir Ahi Evran University
Graduate School of Sciences and Engineering
Advanced Technologies Department

Supervisor: Dr. Öğr. Üyesi Memduh KÖSE

Data mining is the name given to all the processes of collecting, cleaning, processing, analyzing and obtaining useful information from data. It is one of the fastest growing areas in the computer industry. It started as a small field of interest in computer science and statistics and is quickly forming its own field.

Because it can be applied to large datasets, it can be used in almost all industries such as retail, manufacturing, telecommunications, healthcare, insurance and transportation. Studies on predictive modeling have been increasing recently. Modeling-based algorithms are used quite frequently, especially in the fields of health, service and informatics.

The aim of this study is to create a data mining-based prediction model to detect Distributed Denial of Service attacks. This model is to predict the damage to the system when the attacks first start and to predict the success percentage of the attack. In this study, simulated data containing 500 cases were generated based on real-life distributed denial-of-service attack scenarios.

The independent variables in the data set are the lifetime of the packets, the number of packets attacked, low network performance status, slowness in accessing websites, whether there are interruptions in network connections, whether there is an increase in the number of spam e-mails, access to certain parts of the website is blocked, dependent variable is the success of the attack. In this study, Random Forest method, which is frequently used in the literature and has good performance, is discussed as a classification method.

March 2022, 52 Pages

Keywords: Data mining, Data Mining Methods, Classification, Distributed Denial of Service

1. GİRİŞ

Modern bilim ve mühendislik, sistemleri tanımlamak için ilk ilke (prensip) modellerini temel olarak kullanır. Böyle bir yaklaşım, Newton'un hareket yasaları veya Maxwell'in elektromanyetizma denklemleri gibi temel bir yapı ile başlamakta ve daha sonra bu yapının üzerine mühendislik temelli çeşitli uygulamalar inşa edilmektedir. Bu yaklaşımın amacı belirlenen ilke modellerini tespit etmek ve onların doğrudan ölçülebilmesinin zor olduğu veya mümkün olamayacağı bazı değişkenlerin tahmini için kullanılmaktadır. Bununla birlikte, bazı alanlarda temel düşünceden kaynaklı ilkeler bilinmemekte veya araştırılan sistemler matematiksel anlamda ifade edilemeyecek kadar karmaşık bir hale dönüşmektedir. Bilgisayar kullanımının artmasıyla birlikte, bu tür sistemler tarafından üretilen büyük miktarda veri depolanmaktadır. Depolanan bu veriler, bir sistem üzerinden elde edilen değişkenlerin (yani bilinmeyen girdi-çıktı değişkenleri) kendi aralarındaki ilişkilerini tahmin etmek için modeller oluşturulmaktadır. Bu nedenle şu anda klasik modelleme ve ilk ilkelere dayalı analizlerden, geliştirilen modellere ve doğrudan verilerden ilgili analizlere doğru bir paradigma kayması söz konusudur [1, 2].

Bilgisayarları, ağları ve hayatlarımızı dolduran muazzam miktarda veri vardır. Kamu ve özel kurumlar bilişim sistemleri üzerindeki verileri toplamak ve depolamak için sistemler kurmaktadır. Ancak, sistemler üzerinde elde edilecek olan bu verilerin büyük bir kısmı kullanılamamaktadır. Çünkü verilerin kapladığı alan ve karmaşık yapıları onların analiz, yönetim ve yeni veri elde etme süreçlerini zorlaştırmaktadır [1].

Bilişim sistemleri üzerinden elde edilen verilerin büyük ve karmaşık yapılarını anlamlı bir bilgi haline dönüştürebilme ihtiyacı kurum ve kuruluşlar ile bilim ve mühendisliğin ortak amaçlarındandır. Günümüzde kurumun kendisine ait veya paydaşlarına ait verileri stratejik öneme sahip bir veri olarak kabul görmektedir. Bu verilerde gizli olan yararlı bilgileri çıkarma ve bu bilgiye göre hareket etme yeteneği, günümüzün rekabetçi dünyasında giderek daha önemli hale gelmektedir. Bu verilere, bilgisayar merkezli bir yöntem geliştirilerek, uygulanacak olan teknikler neticesinde, yeni bilgiler keşfedilmesi süreci ise veri madenciliği olarak tanımlanmaktadır[1, 3].

Veri madenciliği, veri keşfindeki ilerlemenin otomatik veya manuel yöntemlerle yapılabildiği yinelemeli bir süreçtir. Veri madenciliği, büyük kapasiteye sahip bir veri havuzundan, bilgilerin yeniden derlenmesi ile ortaya çıkacak olan yeni verileri temsil etmektedir. En verimli çalışma, uzmanların sorunları ve hedefleri tanımlaması ve bilgisayarların veri işleme yeteneklerinin birlikte uygulanmasıyla elde edilir [1,2].

Veri madenciliği, günümüz de en hızlı gelişen alanlardandır. Mevcut veri üzerinde geliştirilen algoritmalarla verinin yeniden derlenerek, mevcut veriden yeni veri üretilmesi sürecini kapsamaktadır. Veri madenciliğinin en güçlü yönlerinden biri geniş metodoloji ve teknik yelpazesine sahip olmasıdır. Veri madenciliği, büyük veri kümelerine uygulanabilmesinden dolayı, perakende, üretim, telekomünikasyon, sağlık, sigorta ve ulaşım gibi neredeyse tüm sektörlerde kullanılabilir. Sektör yetkilileri, veri madenciliğini kendilerinden ürün, hizmet, vb. temin eden paydaşlarının ürün tedarikindeki alışkanlıklarını, ihtiyaçlarını, memnuniyet ve memnuniyetsizliklerini tespit edebilmek ve yapılan tespitlerle sektörün yatırım stratejisini planlamak için kullanmaktadırlar. Bu plan çerçevesinde paydaşlarının beklentilerini tespit eden sektör yetkilileri kendi ürün yelpazesinde iyileştirmelere giderek müşteri memnuniyetini artırmayı amaçlar. Bu kapsamda veri madenciliği iş dünyasında paydaşın istek, şikayet vb. taleplerine göre kendi ürünlerinin memnuniyeti artıracak şekilde revize edilmesine olanak sağlamaktadır [1, 2, 3].

Görevleri dolandırıcılık faaliyetlerini tespit etmek ve suç eğilimlerini keşfetmek olan birçok kolluk kuvveti ve özel soruşturma birimi de veri madenciliğini başarıyla kullanmaktadır. Bu metodolojiler analistlere kritik davranış kalıplarını, iletişim etkileşimlerini, kara para aklamadaki parasal işlemleri, seri katillerin hareketlerini ve sınır geçişlerinde kaçakçıların olağan dışı aktivitelerini belirlemede yardımcı olabilir [1].

1.1. Amaç

Bu tezin amacı, veri madenciliğini, veri madenciliğinde kullanılan yöntemleri tanıtmaktır. Hangi durumda hangi yöntemi kullanmanın daha doğru sonuçlar vereceğini, hangi veri türü için hangi yöntemin daha doğru olduğunun belirlenmesidir. Veri madenciliği ile ilgili doğru bilinen yanlışlar, sık yapılan hatalar ile ilgili bilgi vermek, tüm süreçleri adım adım anlatarak sürecin doğru ilerlemesini sağlamaktır.

Bu çalışma kapsamında veri madenciliği kullanılmak sureti ile bir tahmin modeli oluşturulacak ve oluşturulan bu modelle Dağıtılmış Hizmet Reddi (DDoS) saldırıları

incelenmektedir. Yapılan saldırı hakkında (saldırıların türü, saldırının başarılı olup olmadığı ve vereceği zarar) gibi bilgiler sunulmaktadır.

1.2. Önem

Tahmin tabanlı modelleme üzerine yapılan çalışmalar, son zamanlarda artış göstermektedir. Kamu ve özel kurumlar ile bilişim alanında faaliyet gösteren tedarikçiler modelleme algoritmalarını çok sık tercih edilmektedirler. Sağlık sektöründe doktorların yoğun iş yükü altında karar verme güdülerinin etkilendiği düşünüldüğünde, tahmin modelleri ile oluşturulan hekime yardımcı bir sistemin kullanılması çok büyük önem arz etmektedir. Benzer şekilde bilişim sektöründe gün geçtikçe artan siber saldırılar, kötü amaçlı yazılım kullanım yüzdesi, bu saldırıları önceden öngörebilen bir sistemin kullanılması gerektiğinin açık bir örneğidir.

Literatür incelendiğinde Dağıtılmış Hizmet Reddi saldırıları üzerine yapılmış çalışmalar bulunsa da, hiçbir çalışma saldırının sisteme zararını öngörecek sonuçlar içermemektedir. Bu açıdan değerlendirildiğinde, yapacağımız çalışma literatürde ilk olacaktır ve belki de yeni bir bakış açısı, çalışma alanı sağlayacaktır.

1.3. İlgili Çalışmalar

Literatürde yaptığımız çalışmaya benzer bir çalışma bulunmamaktadır. Bindra ve Sood [18] yaptıkları çalışmada, saldırı türünün DDoS saldırısı olup olmadığını tahmin etmek için veri madenciliği yöntemleri kullanmışlardır. Yaptıkları çalışmada Lojistik Regresyon, K-en Yakın Komşu, Gaussian Naieve Bayes, Random Forest, Destek Vektör Makinesi ve bu yöntemlerin hiper parametre versiyonlarını kullanmışlardır. Performans ölçütü olarak doğru sınıflama oranını kullanmışlar ve en iyi sonuç veren yöntem olan Random Forest'a ait doğru sınıflama oranını ise 0,965 olarak bulmuşlardır. Lee ve diğ. [19] çalışmalarında DDoS atak tespiti için kümeleme analizi kullanmışlardır. Atak tespiti için veri setini çeşitli sayıda kümelere bölerek her kümede ayrı ayrı entropi ve gerçekleşme oranı hesaplamışlardır. Her kümenin yapısına (atak, post-atak, normal] göre ve paket sayısına göre tahminlerin değiştiğini göstermişlerdir. Alkasassbeh ve diğ. [20], Çok Katmanlı Algılayıcı, Random Forest ve Naive Bayes yöntemlerini kullanarak DDoS tahmini yapmışlar ve performans ölçütü olarak ise doğru sınıflamam oranını kullanmışlardır. En yüksek doğru sınıflama oranına 0,986 ile Çok Katmanlı Algılayıcı yöntemi ile ulaşırlarken, bu yöntemi sırasıyla

Random Forest (0,980) ve Naive Bayes (0,969) izlemektedir. Zhong ve Yue [21] yaptıkları çalışmada, DDoS atak tespiti için 4 farklı atak tipi ve 2 farklı zamanda atak tespit oranı hesaplamışlardır. SynFlood atağı için atak tespit oranları 1 dk ve 5 dk için %100 olarak bulunmuştur. Stacheldraht atağı için atak tespit oranları 1 dk için %96,5 ve 5 dk için ise %99,3 olarak bulunmuştur. Trinoo atağı için atak tespit oranları 1 dk için %98,9 ve 5 dk için ise %99,7 olarak bulunmuştur. TFN2K atağı için atak tespit oranları 1 dk için %97,2 ve 5 dk için ise %98,7 olarak bulunmuştur. Lakshminarasimman ve diğ. [22] ise çalışmalarında DDoS atak tespiti için karar ağaçları yöntemi olan J48 ve Random Forest'ı kullanmışlardır. Doğru sınıflama oranını J48 yöntemi için %99,9, Random Forest yöntemi için ise %96,9 olarak bulmuşlardır.

2. GENEL KISIMLAR

2.1. Veri Madenciliği Nedir?

Bilişim sistemleri üzerinden elde edilen büyük hacimli verilerin tahmin modelli algoritmalar oluşturularak mevcut veriden, yeni veri elde edilme süreci veri madenciliği olarak tanımlanır [2]. Daha ayrıntılı bir tanım yapılmak istenirse büyük bir veri havuzundan belirli modelleri, yapıları ve yeni verilerin keşfedilmesi sürecidir [3].

Günümüzde bilişim sistemleri üzerinde hizmet veren cihaz veya donanımlar kullanıldıkları amaç kapsamında veriler üretebilmektedir. Bu üretilen veriler, petabayt veya eksabayt mertebesine ulaşan bir veri akışına neden olmaktadır. Bu büyük hacimli veriler günümüzde bilişim cihazlarının kullanımına bağlı olarak ortaya çıkmaktadır. Bu nedenle, uygulamaya özel hedefler için mevcut verilerden işlevsel ve eyleme geçirilebilir bilgiler çıkarılmaktadır. Bu bilgilerle tahmin modelli veri madenciliği kullanılarak mevcut veriden, yeni veriler üretilir. Ham veriler genellikle yapılandırılmamış ve işleme için uygun olmayan bir formattadır. Örneğin, otomatik bir sistem üzerinden elde edilmeyen veriler, farklı kaynaklar üzerinden alındığından bu verilerin bir bilgisayar programı ile yeniden derlenmesine ihtiyaç duyulur. Bu problemin çözümü için veri madenciliği analistleri ise ham olarak toplanan verilerin belirli bir standart göre hazırlandığı yapıları kullanmaktadır. Veriler büyük bir veri tabanı sisteminde depolanmak sureti ile veri tabanına aktarılır. Bu aktarılan veriler üzerinde belirli algoritmaların uygulandığı analitik süreçler kullanılarak veri üzerindeki bilgi çıkarımı yapılmaktadır. Veri madenciliği algoritmalar üzerinden yürütülen bir süreç gibi görünse de aslında yürütülen sürecin büyük bir kısmı verilerin hazırlanması ile ilgili durumu ifade etmektedir. Bu veri işleme süreci kavramsal olarak gerçek bir madencilik sürecine benzer. "Madencilik" terimi ile ifade edilen kasıt bu benzetimdir [1, 2, 3].

Veri madenciliğinin uygulanmasında iki faktör ön plana çıkar bunlardan ilki tahmin ikincisi ise açıklamadır. Tahmin, ilgili diğer değişkenlerin varlığında bilinmeyen ve gelecekteki değerleri öngörülmek istenen bilgiye ulaşma sürecidir. Açıklama, insanlar tarafından yorumlanabilen verileri tanımlayan kalıplar bulma işidir. Bu yüzden, veri madenciliği süreçlerini iki faktörden birine koyarak değerlendirmek mümkündür [1, 2]:

1. Tahmine dayalı veri madenciliğinde; veri seti üzerinden belirtilen sistemin modeli çıkarılır veya

2. Tanımlayıcı veri madenciliğinde; mevcutta ki veri seti üzerinden verileri tanımlanması ile bilgiler ortaya çıkarılır.

2.2. Veri Madenciliği Süreci

"Süreç" kelimesi veri madenciliğinde çok önemlidir. Bazı profesyonel ortamlarda bile, veri madenciliğinin, sunulan soruna uyacak bilgisayar tabanlı bir aracı seçip uygulamaktan ve otomatik olarak bir çözüm elde etmekten ibaret olduğuna dair bir inanç vardır. Bu, dünyanın yapay olarak idealleştirilmesine dayanan bir yanılgıdır. Buradaki yanlışlığı ortaya çıkaran birkaç neden bulunmaktadır. Bunlardan ilki, her birinin birbirinden farklı olduğu ve ilgili problemle ilişkilendirmeyi bekleyen metotların tamamı olmamasındandır. Bunlardan ikincisi ise, bir sorunun bir yöntemle eşleştirilmesi düşüncesinden kaynaklanır. Çok nadiren, bir araştırma sorusu, tek bir veri madenciliği yöntemi ile sonuçlandırılabilir. Aslında, veri madenciliği yinelemeli bir süreçtir. Veriler birden fazla yöntem kullanılarak incelenir, sonuçlar karşılaştırılır ve gerekirse başka yöntemler ile tekrar analiz edilir. Bu süreç istenilen sonuçlar elde edilene kadar tekrarlanabilir çünkü her yöntem, verilerin daha farklı yönlerini araştırmak için kullanılır [1, 3].

Veri madenciliği denildiğinde istatistik, makine öğrenmesi ve diğer metot ve uygulamaların rastgele uygulandığı bir yöntem olarak düşünülmesidir. Ancak bu durum analitik tekniklerle ilgili alanda rastgele yapılan bir işlem olarak düşünülmemeli aksine en doğru karara ve açıklamaya karar verebilmek için düşünülmüş ve planlanmış bir prosedür olarak düşünülmelidir. Veri madenciliği sorunlarına uyarlanarak hazırlanan genel süreç aşağıda belirtilen adımları içermektedir [1, 2, 3]:

2.2.1. Problemi Belirleme ve Hipotez Kurma

Veri merkezli modelleme çalışmasının birçoğu, belirlenen bir uygulama sahasında gerçekleştirilmektedir. Bu nedenle, alana özgü bilgi ve deneyim genellikle anlamlı bir problem ifadesi bulmak için gereklidir. Ama çoğu uygulama araştırması sorun yerine, veri madenciliği yöntemine odaklanma düşüncesindedir. Bu kısımda, bir tek sorun için yöntem oluşturulmuş birden fazla hipotez ortaya konulabilir. İlk adım, uygulama alanı ve veri madenciliği bilgisini birlikte gerektirir. Bundan dolayı veri madenciliği alanında uzman kişi

ile yöntemin uygulamasındaki uzman kişinin beraberce hareket edecekleri bir etkileşimden bahsedilebilir. Veri madenciliğinin başarılı bir şekilde uygulandığı çalışmalarda, bu etkileşim ilk adımla sonlandırılmaz, çalışma boyunca bu etkileşim ve birliktelik devam eder. [1].

2.2.2. Verilerin Toplanması

Bu aşama, verilerin nasıl oluşturulacağı ve toplanacağı ile ilgili ilk kısımdır. Bu kısımda iki farklı durum vardır. Bunlardan ilki verinin oluşturulması sürecinin bir uzmanın kontrolünde oluşturulduğu yaklaşımdır. Bu yaklaşım planlanmış bir deney olarak ifade edilirken, ikincisi ise veri uzmanın iş sürecinde yer almadığı gözleme dayalı olarak yürütülen yaklaşımdır. Çoğu veri madenciliği uygulamasında rasgele veri üretimi gibi gözlemsel bir ortam olduğu varsayılır. Genel olarak, örneklem dağılımı, veriler toplandıktan sonra tamamen bilinmemektedir veya veri toplama prosedüründe kısmen ve dolaylı olarak verilmektedir [1, 3].

Toplanacak olan verinin evreni ne şekilde yansıtacağı hususunu bilmek bu kısımda önemlidir. Çünkü bu modelleme ile ilgili önsel veri çalışması ve çalışma sonucunda ortaya çıkacak olan sonucun yorumlanması için daha faydalı olabilmektedir. Ayrıca, bir modeli tahmin etmek için kullanılan veriler ile daha sonra bu modeli test etmek ve uygulamak için kullanılan verilerin aynı örneklemden geldiğinden emin olmak önemlidir. Söz konusu süreç bu şeklide değilse, tahminde bulunulacak olan modelin çıktılarının uygulanması başarılı şekilde kullanılamayacaktır [1, 2].

2.2.3. Verileri Önişleme

Veriler, amaçlanan kullanımın gerekliliklerini karşılıyorsa kalitelidir. Veri kalitesini oluşturan birden fazla etken vardır. Bunlardan bazıları doğru, eksiksiz, tutarlı, güncel, gerçek ve yorumlanabilmesidir. Eğer elinizdeki verilerde eksikler, hatalar, olağandışı değerler ve tutarsızlıklar varsa verileri veri madenciliği teknikleriyle analiz etmek yanlış sonuçlar doğurur [3, 4].

Veri kalitesini tanımlayan üç unsur vardır: doğruluk, eksik veri içermeme ve tutarlılık. Hatalı, eksik ve tutarsız veriler, gerçek hayatta veritabanlarında ve veri ambarlarında sık rastlanan özelliklerdir. Verinin hatalı olmasında birden fazla etken olabilir. Bunlar bazıları şunlardır.

Veri toplama için kullanılan araçların hatalı olması, verinin sisteme eklenmesi sırasında insan veya bilgisayar kaynaklı hata olması durumlarıdır. Kullanıcılara ait özel nitelikli bilgilerin (doğum tarihi, T.C. Kimlik No vb.) sistem üzerinde işlenmesini istemediklerinde zorunlu alanlara hatalı veri bilgileri gönderebilmektedirler. Örneğin, (doğum tarih bilgisinin varsayılan olarak gelen "1 ocak" parametresinin değerinin seçilmesi gibi) bu tür parametrelerden gelen ve veri alanlarının doldurulamaması sebebi ile eksik veri girişi olarak değerlendirilir. Veri aktarımında da hatalar meydana gelebilir. Verilerin aktarımının izlenmesi sırasında arabellek boyutlarının sınırlandırılmasından dolayı teknolojik anlamda kısıtlamalar olabilmektedir. Verilerin hatalı olması; isimlendirme kurallarından, veriye ait kodlardan, verinin ilgili alan parametreleri ile veri girişi sırasında ki alan parametrelerinin birbirine uymamasından kaynaklanabilmektedir. Yinelenen veriler ayrıca veri temizliği gerektirir [3, 4, 5].

Eksik veriler birkaç nedenden dolayı ortaya çıkabilir. Örnek olarak, pazarlama alanına ait verilerde müşterilere ait veriler her zaman bulunmayabilir. Giriş sırasında önemli görülmedikleri için bazı veriler sisteme girilmeyebilir. Veri girişi sırasındaki hatalı işlemler veya ilgili donanım cihazlarındaki arızalar sebebiyle verilerin ilgili sistemlere kayıt işlemleri yerine getirilememekte ya da kayıt edilen ancak verilerdeki tutarsızlıklar nedeniyle sistemden silinebilmektedir. Ayrıca, veri geçmişinin veya değişikliklerin kaydı gözden kaçmış olabilir. Özellikle bazı öznitelikler için eksik değerlere sahip veri kümeleri için eksik verilerin çıkarılması gerekebilir. Veri kalitesinin, verilerin amaçlanan kullanımına bağlı olduğu bilinmelidir [3, 4, 5, 6].

Yukarıda belirtilen birçok sebep nedeniyle analiz edilmeden verilerin önişleme adımlarından geçmesi gerekir. Bilinen ve en sık kullanılan veri önişleme adımları aşağıdaki gibidir [3]:

2.2.3.1. Veri Temizleme

Veriler standart biçimde olsa bile hatasız olduğu varsayılamaz. Gerçek sistemler üzerindeki verilerin ölçmeden kaynaklı hataları, sübjektif ifadeler ve otomatik olarak kayıt eden cihazların yanlış veya art niyetli kullanımlardan dolayı hatalı çalışması nedeniyle sistem üzerinde hataya neden olabilecek veriler kayıt edilebilmektedir. Gürültü veri kümesinde mevcut olan ancak hatalı kayıt edilen bir değer ifade edilmektedir. Örnek 49,62 sayısı hatalı olarak sisteme 4,962 eklenebilir veya mavi olarak ifade edilen bir nitelik hatalı olarak sisteme yeşil diye eklenebilir. Bu şekilde yapılan hatalar, gerçek sistemler üzerindeki verilerde kalıcı bir sorun olarak karşımıza çıkmaktadır. Ancak 4,962 yerine 49,6X olarak

veya mavi yerine mmavi gibi değer girilmesi veri kümesi için gürültü değiş geçersiz veri anlamına gelir. Bu tür verilerin tespiti gürültü verilere göre daha basittir. Geçersiz olarak tanımlanan bir ifadenin tespiti, düzeltilmesi veya silinmesi daha kolaydır. Veri temizlemedeki amaç da bu tür aykırı, geçersiz verilerin tespit edilerek düzeltilmesi ya da veritabanından çıkarılmasıdır [2, 5, 6,] .

2.2.3.2. Eksik Veriler

Birçok veri setinde, eksik veriler mevcuttur. Eksik veriler insan, sistem kaynaklı olabileceği gibi, bazı değişkenlerin de hastaya ya da hastalığa özgü olmasından da kaynaklanabilir (örneğin, belirli tıbbi veriler yalnızca kadın hastalar veya belirli bir yaşın üzerindeki hastalar için kaydedilmiş olabilir). Eksik değerlerin tespitinde birkaç yol bulunmaktadır. Bunlardan en çok kullanılan iki tanesi aşağıda belirtilmektedir. [2, 5, 6, 7].

2.2.3.2.1. Eksik Veri İçeren Kaydı Çıkarma

Bu yöntemde en az bir eksik değerin olduğu tüm kayıtlar silinir ve kalan kayıtlar üzerinde işlemler gerçekleştirilir. Bu strateji, herhangi bir veri kaynaklı hatadan kaçınmak için iyi bir seçenektir. Ancak eksik kayıtların silinmesi ile elde edilecek olan verilerden ortaya çıkan sonuçlarda verinin güvenilirliğinin etkilenmesinden dolayı bir dezavantaj olarak ortaya çıkmaktadır. Eksik olarak değerlendirilen kayıtların bütün veri içerisindeki oranları küçük olduğunda bu durum değerlendirilebilir. Ancak tavsiye edilen bir durum değildir. Kayıtların tümü veya büyük bir kısmı eksik değerlere sahip olduğunda ise kullanılamaz [2, 5, 6].

2.2.3.2.2. Eksik Veriyi Tahmin Etme

Bu yöntemde, veri setinde bulunan eksik olmayan değerler kullanılarak eksik değerler tahmin edilir [6].

2.2.3.3. Değişken Azaltma

Teknolojideki gelişmeler, veri tabanlarında büyük miktarda değişken ve verinin kaydedilmesine olanak sağlamaktadır (Örneğin bir süpermarket müşterisinin üç ay boyunca yaptığı tüm satın alım kayıtları veya bir hastanedeki her hasta hakkında büyük miktarda detaylı bilgi). Bazı durumlarda veri tabanları, örneklem sayısından fazla değişken içerebilir [6, 7].

Her bir örnek hakkında giderek daha fazla bilgi depolamak doğru gelse de (özellikle hangi bilgilere gerçekten ihtiyaç duyulduğu konusunda zor kararlar vermekten kaçınıldığı için) oluşturulacak modeli bozma riski taşır. Her süpermarket müşterisi hakkında 10000 bilgimiz olduğunu ve hangi müşterilerin yeni bir köpek maması markası alacağını tahmin etmek istediğimizi varsayalım. Bununla ilgili kullanılacak değişken sayısı muhtemelen çok azdır. En iyi ihtimalle, bu alakasız değişkenler, modelle şamasında kullanılacak veri madenciliği algoritmasına gereksiz bir hesaplama yükü getirecektir. En kötüsü ise, algoritmanın yanlış sonuçlar vermesine neden olabilirler [2, 6].

Daha iyi işlemciler ve daha büyük depolama alanları, daha fazla sayıda değişkeni işlemeyi mümkün kılsa da, değişkenlerin sayısı arttığında, elde edilen sonuçların yalnızca yüzeysel doğruluğa sahip olma riski her zaman vardır ve değişkenlerin yalnızca küçük bir kısmının kullanılmasına göre gerçekte daha az güvenilirdir. Bir veri kümesi işlenmeden önce değişkenlerin sayısının azaltılmasının birkaç yolu vardır. Genel olarak özelliğin azaltılması veya boyutun küçültülmesi diye ifade edilmektedir [5, 6].

2.2.4. Model Tahmini

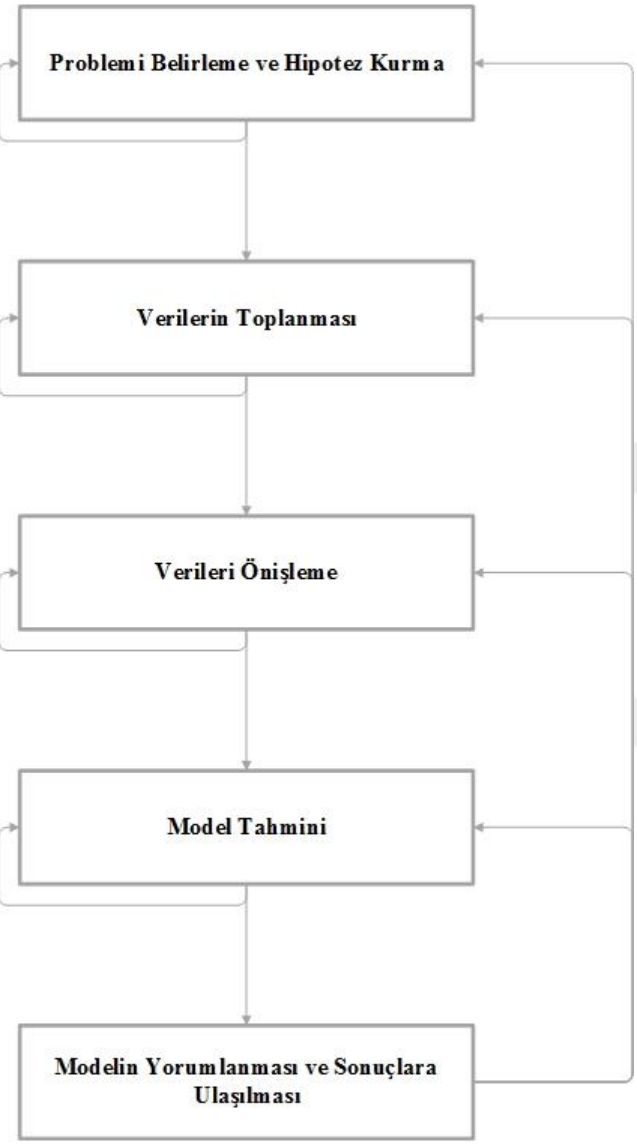
Uygun veri madenciliği yönteminin seçimi ve uygulanması bu aşamadaki ana görevdir. Ancak bu süreç sanıldığı kadar basit değildir. Genellikle, birden fazla yöntem denenerek modeller oluşturulur ve en çok amaca hizmet eden model seçilir. Doğru yöntemin seçilmesi, amaca yönelik modelin oluşturulması süreci doğru yapıldığı zaman haftalar hatta aylar alabilmektedir [2, 5, 6].

2.2.5. Modelin Yorumlanması ve Sonuçlara Ulaşılması

Veri madenciliği sonucu elde edilen modellerin karar verme sürecinde yararlı olması için yorumlanabilir olması gerekir. Hem modelin doğruluğu ve hem de yorumlanmasının doğruluğu asıl hedef olarak görülür. Genellikle basit modeller daha kolay yorumlanabilir, ancak aynı zamanda daha az doğrudur. Modern veri madenciliği yöntemleri ise daha gelişmiş modeller kullanarak daha doğru sonuçlar verir. Ancak bir diğer önemli husus ise bu modelleri yorumlanabilmesidir. Model ne kadar doğru olursa olsun düzgün yorumlanamadığı sürece hiçbir değeri yoktur [1, 5, 7].

Başarılı bir uygulama için tüm sürecin iyi anlaşılması önemlidir. 4. Adımda kullanılacak olan veri madenciliği süreci ne kadar iyi tasarlanmış olursa olsun, ortaya çıkacak modelde,

verilerin toplanması ve ön işlem süreci doğru şekilde yapılmazsa veya problem için ortaya konacak formül anlamlı olmazsa geçersiz bir durum ortaya çıkacaktır. Beş adımı özetleyen grafik ise Şekil 2.1'de verilmiştir [1, 5, 6].



Şekil 2. 1. Veri Madenciliği Süreci [1]

2.3. Veri Madenciliğinde Yapılan Analizler

Aşağıdaki liste, veri madenciliğinde en yaygın kullanılan analiz türlerini göstermektedir [8, 9]:

- Tanımlayıcı

- Değer Tahmini

- Gelecek için Tahmin

- Sınıflandırma

- Kümeleme

- Birliktelik Kuralları

2.3.1. Tanımlayıcı

Analist ve araştırmacı veri yapılarını inceleyerek içerisindeki kalıpları ve eğilimleri tanımlamaya çalışmaktadırlar. Örneğin, bir anketör, işten çıkarılmış olanların başkanlık seçimlerinde mevcut görevliyi destekleme olasılığının düşük olduğuna dair kanıtlar ortaya çıkarabilir. Tanımlanan model ve yapılarla ilgili açıklamalar, bu tip kalıp ve eğilimlerle ilgili muhtemel açıklamaları getirmektedir. Örneğin, işten çıkarılmış olanlar şu anda mali açıdan iktidarın seçilmesinden öncekine göre daha az varlıklı durumdalardır ve bu nedenle bir alternatifi tercih etme eğiliminde olmaları muhtemeldir [5, 9].

Veri madenciliği tasarımları açık olmalıdır. Bunun için veri madenciliğinde tasarlanan yapının neticeleri, yorum ve açıklamaya ilişkin net durumları ifade edebilmelidir. Bazı veri madenciliği metotları diğer yöntemlere göre yorumlanması daha şeffaf olabilmektedir. Örnek karar ağaçları neticelerinin, kullanıcı tarafından sezgisel ve daha kolay açıklanmasını sağlamaktadır. Bununla birlikte, sinir ağlarını yorumlamak, kullanılan modelin doğrusal olmaması ve karmaşıklığı nedeniyle uzman olmayanlar için zordur [5, 8, 9].

2.3.2. Değer Tahmini

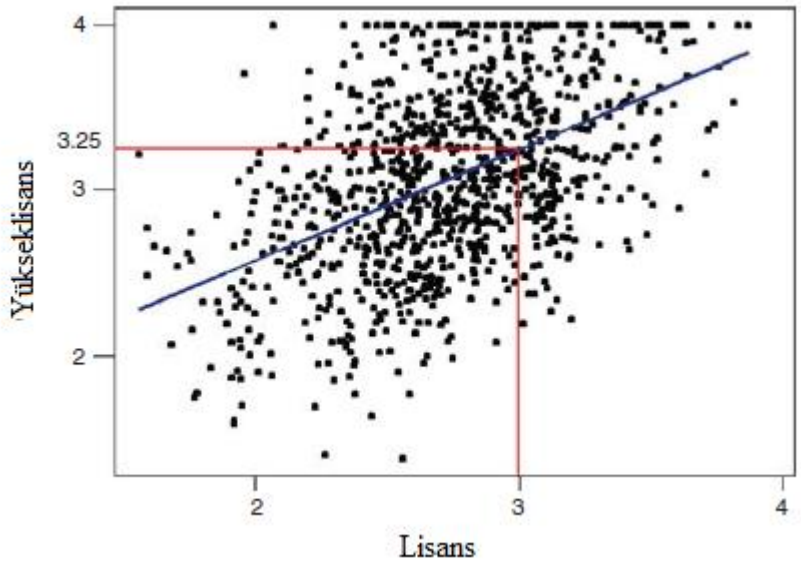
Tahmin analizinde, sayısal bir hedef değişkenin değeri bir dizi sayısal ve / veya kategorik öngörücü değişken kullanarak yaklaşık olarak hesaplanmaktadır. Tasarımlar hedef parametresinin yanında öngörücülerin belirttiği “tam” kayıt kullanılmak sureti ile oluşturulmaktadır. Yapılacak olan yeni gözlemlerde, tahmincilerin parametreleri temel alınarak hedef değişkeninin parametreleriyle alakalı tahminlerde bulunulur. Daha sonra, yeni gözlemler için, tahmin edicilerin değerlerine dayalı olarak hedef değişkenin değerine ilişkin tahminler yapılır [5, 9].

Örneğin, Bu metot kullanılarak bir hastaya ait kan basıncı (sistolik) parametresinden hastaya ait (yaş, cinsiyet, kandaki sodyum seviyesi, vücut kitle indeksi vb.) bazı bilgiler için tahminlerde bulunulabilir. Eğitim setindeki sistolik kan basıncı ile öngörücü değişkenler

arasındaki ilişki bir tahmin modeli sağlar ve daha sonra bu model yeni vakalar için kullanılabilir [8, 9].

Benzer şekilde bu analiz ile bir yüksek lisans öğrencisinin not ortalaması (GPA), öğrencinin lisans not ortalamasına göre tahmin edilebilir. 1000 öğrenci için lisans not ortalamalarına karşı lisansüstü not ortalamalarının dağılım grafiği Şekil 2'de verilmiştir. Bu değişkenler arasındaki ilişkiye en iyi yaklaşacak olan doğrunun bulunmasında, basit doğrusal regresyon en küçük kareler kriterleri bulmamızı sağlayacaktır. Basit doğrusal regresyon, en küçük kareler kriterine göre bu iki değişken arasındaki ilişkiye en iyi yaklaşan doğruyu bulmamızı sağlar. Şekil 2.2'de mavi ile gösterilen regresyon çizgisi, öğrencinin lisans not ortalamasına göre bir öğrencinin lisansüstü not ortalamasını tahmin etmek için kullanılabilir [9].

Burada, regresyon çizgisine ait denklem y = 1,24 + 0,67x olarak bulunur. Bu bize tahmini yüksek lisans not ortalamasının 1,24 artı öğrencinin lisans not ortalamasının 0,67 katına eşit olduğunu söylemektedir. Örneğin, örneğin öğrencinin lisans not ortalaması 3,0 ise, tahmini yüksek lisans not ortalaması y = 1,24 + 0,67 * (3) = 3,25 olarak bulunur. Bu noktanın (x = 3,0, y = 3,25), tam olarak regresyon çizgisinde bulunduğu şekilde kırmızı ile gösterilmiştir [9].



Şekil 2. 2. Regresyona Dayalı Tahmin Örneği [9]

2.3.3. Gelecek için Tahmin

Geleceğe dair sonuçlar için tahminde bulunmak, sınıflandırma ve değer tahminine benzeyen özellikleri üzerinde taşımaktadır. İş ve araştırma için örnekleri aşağıdaki gibidir [9]:

- bir hisse senedinin 3 ay sonraki fiyatını tahmin etmek;

- hız sınırı artırılırsa gelecek yıl trafik ölümlerinde yüzde artışı tahmin etmek;

- ilaç keşfinde belirli bir molekülün bir ilaç şirketi için karlı yeni bir ilaca yol açıp açmayacağını tahmin etmek.

Sınıflandırma ve değer tahmininde kullanılacak metot ve yöntemlerden birisinin koşullarının uygunluğu durumunda geleceğe yönelik tahmin içinde kullanılabilmektedir. Sınıflandırma ve değer tahmini için kullanılan yöntem ve tekniklerden herhangi biri, uygun koşullar altında gelecek için tahmin için de kullanılabilir. Geleneksel istatistiksel değer tahmini ve güven aralığı tahminleri, basit doğrusal regresyon, korelasyon ve çoklu doğrusal regresyon yöntemlerinin yanı sıra veri madenciliği yöntemleri olan k-en yakın komşu, karar ağaçları ve sinir ağları gibi yöntemler gelecek için tahminde kullanılabilir [8, 9].

2.3.4. Sınıflandırma

Sınıflandırmada kullanılan hedef değişkeninin sayısal olmayıp, kategorik olması nedeniyle değer tahminine benzemektedir. Sınıflandırmada, örneğin üç sınıfa veya kategoriye ayrılabilen gelir grubu gibi kategorik bir hedef değişken vardır: yüksek gelir, orta gelir ve düşük gelir. Veri madenciliği yöntemi ile geniş kapsamlı bir veri kümesi incelenmektedir. Veri kümesi içerisinde hedef değişkene ait bilgilerin dışında girdi ve tahmin parametrelerine ait verilerde yer almaktadır. Sınıflandırma için örnek bir veri kümesi, Tablo 2.1'de verilmiştir [5,9].

Tablo 2. 1. Sınıflandırma için Örnek Veri Seti

Kişi	Yaş	Cinsiyet	Meslek	Gelir
1	47	Kadın	Bilgisayar Mühendisi	Yüksek
2	28	Erkek	Pazarlama Danışmanı	Orta
3	35	Erkek	İşsiz	Düşük
…	…	…	…	…

Araştırmacı, yukarıdaki veri tabanında bulunmayan yeni bireylerin gelir grubunu, o kişiyle ilişkili yaş, cinsiyet ve meslek gibi diğer özelliklere göre sınıflandırmak isteyebilir. Bu durum veri madenciliği metot ve yöntemleri için çok makul bir sınıflandırma işidir. Algoritma kabaca şu şekilde ilerleyecektir. İlk olarak, hem yordayıcı değişkenleri hem de

(önceden sınıflandırılmış) hedef değişkeni olan geliri içeren veri seti kullanılarak, algoritmaya (yazılıma) hangi değişken kombinasyonlarının hangi gelir gruplarıyla ilişkilendirildiği öğretilen bu sete, eğitim veri seti denilmektedir. Daha sonra algoritma, hedef değişken olan gelir grubu hakkında hiçbir bilginin bulunmadığı yeni kayıtlara bakacaktır. Bu veri seti ise test veri seti olarak adlandırılır. Eğitim veri setinde yer alan sınıflandırmalar neticesinde, yeni kayıtlara algoritma tarafından parametre (değer) atanmaktadır. Örneğin, 63 yaşındaki kadın bir profesörün gelir grubu bazında "yüksek" parametresi ile sınıflandırılması [8, 9].

İş ve araştırma için sınıflandırma örnekleri aşağıdaki gibidir [9]:

• belirli bir kredi kartı işleminin hileli olup olmadığının belirlenmesi,

• belirli bir hastalığın mevcut olup olmadığını teşhis etmek,

• bir vasiyetin gerçekte ölen kişi tarafından mı yoksa başka biri tarafından hileli olarak mı yazıldığını belirlemek,

• belirli mali veya kişisel davranışların olası bir terör tehdidini gösterip göstermediğini belirlemek.

2.3.5. Kümeleme

Kümeleme; Gözlem, vaka veya kayıt içerisinde yer alan benzer nitelikteki verilerin sınıflandırılarak gruplandırılmasını ifade etmektedir. Küme denildiğinde benzer nitelikteki nesnelerin, diğer kümeler içindeki verilerden farklı olarak kayıt koleksiyonu şeklinde tanımlanmasıdır. Kümeleme veri setinde hedef değişkene ait parametre olmadığından bu yönü ile sınıflandırmadan ayrılmaktadır. Kümeleme hedef değişkenine ait parametreyi sınıflandırmak için veya tahminde bulunmak için kullanmamaktadır. Kümelemeye ait algoritmalar veri kümesinin tümünü eşit alt basamaklara veya kümelere ayırmaya çalışmaktadır. Bu yöntemle küme içerisinde yer alan verilerin nispeten benzerliği maksimum seviyeye çıkarılırken, küme dışarısında yer alan verilerin benzerliği minimum seviyeye indirilmektedir [9, 10].

2.3.6. Birliktelik Kuralları

Veri madenciliği için birliktelik kuralları, hangi özniteliklerin birlikte yer aldığını (çalıştığını) bulma işidir. Sepeti analizi olarak bilindiği iş dünyasında en yaygın kullanım alanı, iki veya daha fazla değişken arasındaki ilişkiyi ölçmek için kuralları ortaya çıkarmaktır. Örneğin, belirli bir süpermarket, bir Perşembe gecesi alışveriş yapan 1000 müşteriden 200'ünün çocuk bezi satın aldığını ve çocuk bezi satın alan 200 kişiden 50'sinin de bira aldığını varsayalım. Bu durumda, birliktelik kurallına göre "Çocuk bezi alan kişiler bira da alır" şeklinde kural oluşturulur. Bu kurala göre de marketler reyonlarda çocuk bezi ile birayı yakın reyonlara koyarak daha çok satış yapmayı amaçlar [9, 11, 12].

İş ve araştırma için birliktelik kuralına ait örnekler aşağıdaki gibidir [9]:

• Kendileri iyi okuyucu olan ebeveynlerin çocuklarının okuma oranını incelemek;

• Telekomünikasyon ağlarındaki bozulmanın tahmin edilmesi;

• Bir süpermarkette hangi öğelerin birlikte satın alındığını ve hangi öğelerin asla birlikte satın alınmadığını öğrenmek;

• Yeni bir ilacın tehlikeli yan etkiler göstereceği vakaların oranının belirlenmesi

2.4. Veri Madenciliğinde Yapılan Hatalar

Veri madenciliği, verilerdeki önemli bilgileri çıkarmaya yönelik bir yaklaşımdır. Verilerin sağlayabileceği birçok ipucunun peşinden gitmek çok fazla hazırlık ve sabır gerektirir. Bu tür bilgi keşfi için çok fazla alan bilgisi, araç ve beceri gereklidir. Veri madenciliğini uygularken yapılan en yaygın hatalar ve kaçınılması gerekenler aşağıda verilmiştir [4, 9].

Hata 1: Veri madenciliğinde hatalı problemin seçilmesi: hedefin doğru olarak belirlenememesi veya hedef olmaksızın hareket edilmesi süreci, veri madenciliği açısından zaman kabından başka bir şey değildir. Alakasız bir soruya doğru cevabı almak ilginç olabilir, ancak anlamsız olacaktır [4].

Hata 2: Çok fazla veriye/değişkene sahip olmak her zaman iyi sonuçlar elde edileceği anlamına gelmez. Başlangıçta çok daha fazla veriye ihtiyaç duyulacağı düşünülürken daha az veri ile süreçler tamamlanabilmektedir. Bu kadar çok değişkenle uğraşmak yanlış sonuçlara, zaman kaybına neden olabilir. Önemli olan probleme odaklanıp, problemi açıklayacak değişkenleri seçmektir [4].

Hata 3: Net hedefler olmadan, çok fazla zaman boşa harcanır. Bir sonraki aşamayı düşünmeden, plan yapmadan aynı madencilik algoritmalarını kullanarak aynı testleri tekrar tekrar ve körü körüne yapmak, zaman ve enerji israfına yol açacaktır. Bu, veri madenciliği prosedürünü ve sonuçlarını takip etme konusunda yetersiz bilgiye, tecrübeye sahip olmaktan kaynaklanabilir [4].

Hata 4: veri madenciliği için kullanılacak olan cihazların birbirleri ile olan uyumsuzlukları: Verinin toplanması, hazırlanması ve görselleştirilmesinde kullanılacak olan cihazlar birbiri ile uyumlu ve bir arada çalışabilecek şekilde olmalıdır [4].

Hata 5: Birden çok farklı tipte veri formatına sahip kaynaktan gelen verileri kullanmadan önce, verilerin hepsinin aynı formata dönüştürülmesi gerektiğini unutmayınız. Bunu sağlayan, bu verilerle çalışabilen araçlar kullanın [4].

Hata 6: Tek tek kayıtlara / tahminlere değil, yalnızca toplu sonuçlara bakmak. Tüm verilerden elde edilen genel düzeydeki doğru sonuçlar, bireysel bir kayıt düzeyinde yanlış sonuçlar verebilir. Bu yanılgıya düşmeyin. Çünkü eğer her bireysel kayıt genel sonuçlar ile uyumlu olsa sınıflama başarınız %100 olur, bu da gerçek hayatta imkansızdır [4].

Hata 7: Zamanın Verimli Kullanılmaması: Veri hazırlama süreci için (toplama, seçme vb.) yeterli vakit ayırılmamasından dolayı veri hazırlamada sıkıntılar ortaya çıkmaktadır. Benzer şekilde modeli test etmek, kullanıcıları eğitmek ve sistemi dağıtmak için yeterli zamanın sağlanmaması projeyi başarısız hale getirebilir [4].

Hata 8: Modeli, kullanacak kişinin onları ölçme yönteminden farklı bir şekilde oluşturmak: Bu, iş hedeflerini kaybetmekten, sorunu tam olarak anlayamamaktan ve yanlış planlama yaparak veri madenciliği yapmaya başlamaktan kaynaklanır. Önce tüm hedefler, planlar belirlenmeli, daha sonra çözüme gidilmelidir [9].

Hata 9: Veriler hakkında söylenen her şeye inanmak: Eğer veriler hakkında söylenenleri tamamen doğru kabul edip, veri önişleme adımlarını atlayarak analizlere başlarsanız, elde edeceğiniz sonuçlar yanlış olur. Çünkü gerçek hayatta, insan etkileşiminin olduğu yerde hata olmaması imkansızdır [9].

2.5. Veri Madenciliği Yöntemleri

Literatürde sık kullanılan veri madenciliği yöntemlerine ait özet bilgi Tablo 2.2'de verilmiştir [4].

Tablo 2. 2. Önemli Veri Madenciliği Yöntemleri

Denetimli Öğrenme	Karar Ağaçları
	Çok Katmanlı Algılayıcı
	Destek Vektör Makinesi
	Naive Bayes
	Lojistik Regresyon
Denetimsiz Öğrenme	Kümeleme Analizi
	Birliktelik Kuralları

2.5.1. Random Forest

Random Forest ilk olarak, 1995 yılında Ho tarafından geliştirildi ve daha sonra Breiman tarafından bagging fikri kullanılarak önemli ölçüde geliştirildi. Random Forest, adından da anlaşılacağı gibi, bir topluluk olarak çalışan çok sayıda bireysel karar ağacından oluşur. Random Forest'daki karar ağaçlarının her biri, bir sınıfla ilgili bir tahminde bulunur ve oyu en fazla alan sınıf kurulan modelin tahmincisi olmaktadır [10, 13].

2.5.1.1. Random Forest Modellerinin Avantaj ve Dezavantajları

Random Forest kullanmanın birçok avantajı vardır [1, 10].

1. Random forest ile hazırlanan yöntemler kolay ve okunabilir olmalarının yanında üretilmeleri çok hızlı olabilmektedir.

2. Pek çok istatistiksel yaklaşımın aksine, kullandığı yaklaşım öznitelik değerlerinin dağılımı veya özniteliklerin bağımsızlığı hakkındaki varsayımlara bağlı değildir.

3. Bu yöntem ile oluşturulan modeller diğer birçok istatistiksel yönteme göre daha iyidir.

Random Forest kullanmanın avantajılarının yanı sıra dezavantajları da vardır [1, 13].

1. Bazı durumlarda eğitim için çok fazla örneklem gerektirebilir.

2. Çok fazla değişken ile model oluşturulması gerektiğinde, sınıflandırma, çok sayıda kural ve normalden daha büyük bir hatayla son derece karmaşık hale gelebilir.

3. Model oluşturma süreci diğer yöntemlere göre daha uzun sürebilir.

2.5.2. Çok Katmanlı Algılayıcı

Çok Katmanlı Algılayıcı (ÇKA) giriş, çıkış ve çok sayıda algılayıcının yer aldığı yapay bir sinir ağıdır. Sinyalin alındığı giriş katmanı ile bu girişle ilgili karar veya tahminde bulunulan bir çıkış katmanı ve bu katmanlar arasında ÇKA'nın hesaplandığı random sayıdaki gizli katmanlardan oluşmaktadır [4, 5].

ÇKA yapısı gereği denetimli öğrenme sorunlarında uygulanan bir metottur. Girdi ve Çıktı katmanları üzerinden bilgi alırlar ve bu katmanlar arasındaki bağımlılıkları (korelasyon) modelleyerek öğrenmektedirler. Öğrenme aşamasında hatanın minimuma indirilmesi için modelin değişkenlerini ve bu değişkenlerin ağırlık ve önyargılarını ayarlamayı içermektedir. Geri yayılım ise modelleme üzerindeki hatalara göre ağırlığının ve önyargısının yeniden yapılandırılması için kullanılmaktadır [4, 5].

2.5.2.1. Çok Katmanlı Algılayıcı Modellerinin Avantaj ve Dezavantajları

Çok Katmanlı Algılayıcı kullanmanın birçok avantajı vardır [4, 5].

1. Çok Katmanlı Algılayıcı kullanımına çok az kısıtlama getirir. Kullanıcının veya analistin çok fazla çalışmasına gerek kalmadan, oldukça doğrusal olmayan ilişkileri kendi başına halleder (tanımlar / modeller).

2. Örneklerden öğrendiği için Çok Katmanlı Algılayıcıyı programlamaya gerek yoktur. Programlamaya ihtiyaç duyulmaz çünkü kullanıldıkça daha verimli hale dönüşmektedir.

3. Çok Katmanlı Algılayıcı, sınıflandırma, kümeleme, ilişkilendirmeler vb. dahil olmak üzere çeşitli problem türlerini için kullanılabilir.

4. Çok Katmanlı Algılayıcı veri üzerindeki problemlere karşın müsamahalıdır. Verileri katı normal ve/veya bağımsızlığa göre yapılacak olan önermelere uyması için kısıtlama getirmemektedir.

5. Çok Katmanlı Algılayıcı sayısal değişkenler ile sözel değişkenlere ait parametreleri işleyebilmektedir.

6. Çok Katmanlı Algılayıcı diğer yöntemlerden çok daha hızlı çalışır.

7. En önemlisi, Çok Katmanlı Algılayıcı genellikle yeterince eğitildikten sonra diğer yöntemlere kıyasla daha iyi sonuçlar (tahmin ve / veya kümeleme) sağlar.

Dezavantajları ise yorumlamanın, açıklamanın veya hesaplamanın kolay olmamasından kaynaklanmaktadır [4, 5].

1. Açıklanabilirlikten yoksun kara kutu çözümleri olarak kabul edilir.

2. Çok Katmanlı Algılayıcının yapısı çok karmaşıktır. Uzmanlık ve kapsamlı denemeler gerektirir.

3. Çok Katmanlı Algılayıcı ile çok fazla parametrenin (nominal değişken) yer aldığı verileri incelemek zor olabilmektedir.

4. Çok Katmanlı Algılayıcıyı eğitmek için büyük veri kümeleri gerekir.

2.5.3. Destek Vektör Makinesi

Destek vektör makinesi (DVM), iki kategoriden oluşan sınıflandırma sorunları için sınıflandırma algoritmalarının kullanıldığı denetimli makine öğrenmesi yöntemidir. Her kategori için bir eğitim veri seti verdikten sonra, yeni değeri sınıflara ayırabilir. DVM, sınıflandırma ve regresyon sorunlarında kullanılabilir ama en çok sınıflandırmaya ait sorunlarda tercih edilmektedir. DVM algoritmasında, her bir veri öğesi n boyutlu uzayda bir nokta olarak çizilmektedir (n= sahip olduğunuz özellik sayısı), her özelliğin değeri belirli bir koordinatın değerine denk gelmektedir. Ardından, iki sınıfı çok iyi ayıran hiper düzlem bulunarak sınıflandırma yapılmaktadır [10, 14].

2.5.4. Naive Bayes

Naive Bayes, parametreler arasında bağımsızlık varsayımı olduğunu düşünen ve sonuçlar için Bayes teoremini kullanan basit yapıdaki denetimli makine öğrenme algoritması olarak tanımlanır. Algoritma üzerindeki girdi parametrelerinin bağımsız değişkenler olarak varsayıldığının ifade edilmesidir. Algoritmanın birçok karmaşık sorunu etkili bir şekilde çözebildiği bilinmektedir. Örnek olarak Naive Bayes ile sınıflandırıcı oluşturmak sinir ağı

tarzındaki algoritma yapılarına göre kurulumları kolay olmaktadır. Bu yöntem yetersiz veya hatalı olarak etiketlenen veriler üzerinde bile iyi sonuç vermektedir [10, 14].

2.5.4.1. Naive Bayes Yönteminin Avantaj ve Dezavantajları

Naive Bayes kullanmanın birçok avantajı vardır [10].

1. Eğitim veriseti için çok fazla örneklem gerektirmez.

2. Az örneklemde bile diğer yöntemlerden daha iyi sonuçlar verebilir.

3. Çok fazla değişken içeren verisetlerinde bile kısa sürede sonuçlar vermektedir.

Ancak bu yöntemi kullanmanın da bazı dezavantajları mevcuttur [10, 14].

1. Bağımsız değişkenler kategorik olduğu durumda iyi sonuçlar verir. Sayısal değişkenler için tahmin gücü düşüktür.

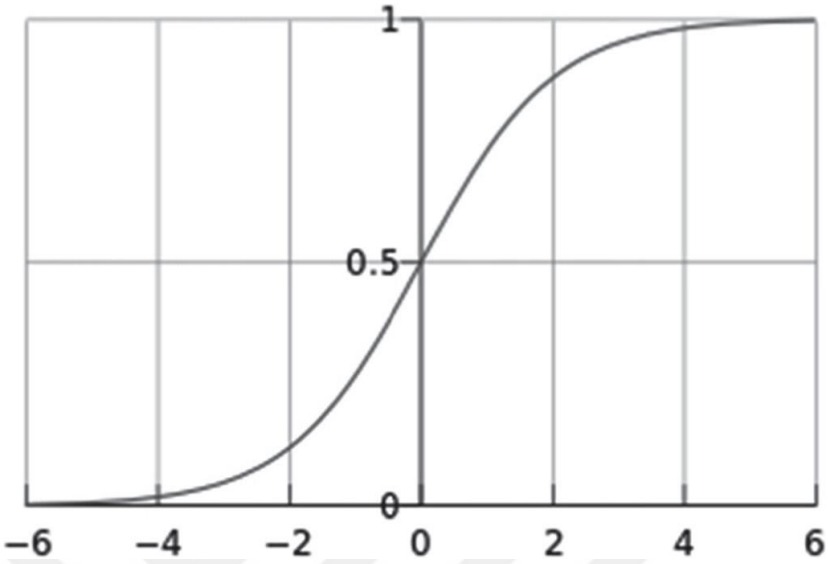
2. Kategorik bir değişkenin test veri setinde eğitim veri setinde yer almayan bir kategori varsa, model bu kategoriye 0 olasılık atayacak ve bir tahmin yapamayacaktır. Buna Sıfır Frekans sorunu denmektedir.

2.5.5. Lojistik Regresyon

Regresyon yöntemi bağımlı ve bağımsız parametreler için her zaman sayısal veriler üzerinde çalışmaktadır. Aynı zamanda, lojistik regresyon yöntemleri, kredi onaylanması veya onaylanmaması (evet veya hayır) durumlarına karşın, ikili parametreye sahip kategori bağımlı değişkeni ile birlikte çalışmaktadır. Bu regresyon'un amacı kategorik olarak bağımlı değişkenle bağımsız (bir veya fazla) değişkeler arasındaki ilişkileri ölçmektedir. Örneğin, lojistik regresyon, hastanın gözlenen özelliklerine (yaş, cinsiyet, vücut kitle indeksi, kan testleri, vb. sonuçlar) dayalı olarak, bir hastanın belirli bir hastalığa (örneğin, diyabet) sahip olup olmadığını tahmin etmek için kullanılır [4, 13].

Lojistik regresyon yöntemleri, bağımlı değişkene ait tahmin edilebilen değerlerini olasılık hesaplamasında kullanmaktadır. Lojistik regresyon, bağımlı değişken olasılıklarının algoritmasını almaktadır (logit olarak isimlendirilir). Bu nedenle, lojistik regresyonda bağımlı değişken olarak logit dönüşümü kullanılır. Yatay eksende bağımsız değişken ve dikey eksende logit bağımlı değişken ile genel lojistik fonksiyon Şekil 2.3'te verilmiştir [4, 14].

Veri madenciliği alanında işlem yapan popüler platformların hepsi düzenli çoklu regresyon yöntemlerinin yanında lojistik regresyon modelleri içinde destek sağlamaktadır [4].



Şekil 2. 3. Genel Lojit Fonksiyonu [4]

2.5.5.1. Lojistik Regresyon Modellerinin Avantaj ve Dezavantajları

Lojistik Regresyon yöntemleri çok sayıda avantajı kullanıma sundukları için popüler olmaktadırlar [4].

1. Lojistik Regresyon yöntemleri en küçük kısımda meydana gelen hata gibi temel istatistik ilkeler üzerine kurulduğundan anlaşılması kolay olmaktadır.

2. Lojistik Regresyon yöntemleri basit cebirsel denklemlerle anlaşılması ve kullanımı kolay yapılar sağlamaktadır.

3. Lojistik Regresyon modelinin gücü (veya uyum iyiliği), korelasyon katsayıları ve iyi anlaşılmış diğer ilgili istatistiksel parametreler açısından ölçülür.

4. Lojistik Regresyon yöntemleri farklı modelleme tekniklerinin tahminleriyle eşleşebilmekte ve onlara göre çok daha iyi sonuçlar verebilmektedir.

5. Lojistik Regresyon yöntemleri kişilerin model içerisine dahil etmek istedikleri tüm parametreleri içerebilmektedir.

6. Lojistik Regresyon modellemesine ait araçlar yaygın olarak kullanılmakta olup istatiksel ve veri madenciliği paketlerinin tümünde bulunmaktadırlar.

Ancak regresyon modelleri bazı durumlarda yetersiz kalabilir [4].

1. Lojistik Regresyon yöntemleri veri kalitesinde düşme olması durumunda ortaya çıkabilecek problemleri karşılayamamaktadır. Bu yöntem veri üzerindeki eksik parametreleri gidermek üzere hazırlanmamışsa, performansına olumsuz olarak etki edebilmektedir.

2. Lojistik Regresyon modelleri, bağımsız değişkenler arasında güçlü ilişkiler (korelasyonlar) olması durumunda düşük performans gösterebilir. Bağımsız değişkenler kendi aralarında güçlü korelasyonlara sahipse, birbirlerinin öngörü gücünü etkilerler ve regresyon katsayıları bundan etkilenir.

3. Lojistik Regresyon yöntemlerinde kullanılmakta olan paketlerin bazıları bunu yapmaya çalışsa da, birbiri ile yüksek seviyede ilişkili olan parametreler arasında otomatik seçme yapılamamaktadır. Yönteme birçok parametre dahil edilmesi halinde ise bu model kullanışlı ve güvenilir olmaktan çıkabilmektedir. Yönteme dahil edilecek olan tüm parametreler, yöntemin öngörü gücüne katkısından bağımsız olarak lojistik regresyon denklemine de dahil edilmektedirler.

4. Lojistik Regresyon modelleri otomatik olarak doğrusal olmayışla ilgilenmez. Kullanıcının, uyumunu iyileştirmek için regresyon modeline eklenmesi gerekebilecek ek terimleri bilmesi gerekir.

2.6. Dağıtık Hizmet Engelleme (DDOS)

DoS ve DDoS saldırıları, kötü niyetli bir kişinin veya grubun bir web sitesi için önemli bir kesinti süresine, finansal kayba ve itibar kaybına neden olabildiği saldırılar olarak tanımlanmaktadır. Bu tip olaylar her gün dünya çapında haber manşetlerine çıkmaktadır. Bilgi güvenliği araştırmacıları, dünya çapında meydana gelen DoS ve DDoS saldırılarının sayısı veya doğası ile ilgili verileri toplamak için henüz standart bir strateji geliştirmemiş olsa da, her gün 7000'den fazla bu tür saldırının meydana geldiği tahmin edilmektedir. Web sitesi üzerinden hizmet vermekte olan kuruluşlar, düzenli olarak erişim sağlanan verilerine, DoS ve DDoS ataklarına karşın tedbir mekanizmalarını kurmaları gerekmektedir. Bunun yapılmaması durumunda ise büyük mali kayıplara ve kamu itibarının zedelenmesine neden olabilir [15, 16, 17].

DoS ataklarında saldırgan tarafından çok sayıda istek talebi kurumun bilişim ağına yönlendirmekte ya da saldırgan tarafından veri göndermesi yöntemi ile bu saldırlar yapılmaktadır. Bu ataklar bilişim cihazlarının isteklere cevap veremeyecekleri hale gelene kadar devam etmektedir. Böylece kurum/kuruluşlardan hizmet talep eden gerçek kullanıcılar sistemlere erişemeyeceklerdir. Daha basit bir ifadeyle, DoS saldırısı, bir saldırganın bir makinenin kaynaklarını, başka bir makinenin normal şekilde çalışmasını engellemek için kullanmasıdır. Web sayfalarını barındıran web sunucuları saldırgan tarafından tek makine ile yapacağı saldırıları performans kaybı yaşamadan cevap verebilmektedir. Ancak saldırganlar genellikle hedef sistemin tüm kaynaklarını tüketmek için, birden fazla makine kullanan DDoS saldırıları gerçekleştirir. Bu senaryoda, saldırganları manuel olarak tespit etmek ve engellemek genellikle zordur. Bu nedenle bilişim altyapısında kullanılmak üzere saldırı tespit ve önleme sitemleri gerekmektedir. Aslında saldırganlar, saldıran makinelerini neredeyse hiçbir zaman yasal olarak kontrol etmezler. Daha ziyade, bu tür makinelere yetkisiz erişim elde etmek için dünyanın dört bir yanına yayılmış binlerce bilgisayara özel kötü amaçlı yazılımlar bulaştırırlar. Bir saldırganın kontrolü altında bir ordu olarak hareket eden, güvenliği ihlal edilmiş yüzlerce veya binlerce makineden oluşan bu bilgisayar topluluğuna "botnet" denir [15, 17].

2.6.1. Saldırı Başlatma

DDoS saldırılarını başlatmak, teknik bilgisi olamayan bir kişi için bile zor olmamaktadır. Herkes tarafından kullanılabilecek türde kiralık DDoS hizmetleri bulunduğundan, kullanıcılar büyük çaplı bir saldırı için kendi botnetlerine ihtiyaç duymamaktadırlar. Böyle bir hizmeti kullanan herkes, saldırının boyutuna ve süresine bağlı olarak, seçtiği bir hedefe saati 40 ile 1500 TL arasında değişen bir ücrete güçlü bir DDoS saldırısı başlatabilir [15].

2.6.2. İş Üzerine Etkisi

DDoS saldırılarını incelenmek üzerine yapılan anketlerde, katılımcıların %70'i en az bir defa saldırıya maruz kaldıklarını ifade etmektedirler. Saldırganların hedefinde geçmişte sektörlere özgü hedefler varken günümüzde finans kuruluşları, hükümetler, online satış firmaları vb. farklı iş kolları ve kamu veya özel nitelikli kurumlar yer almaktadır. [15].

Bir DDoS saldırısının ticari etkisi büyüktür ve saldırının kapsamına bağlı olarak bir kurbanı belirli bir süre etkileyebilir. Yapılan araştırma/anket raporlarına göre, 2011'de gerçekleştirilen DDoS saldırıları, ortalama 24 saat olmak üzere birkaç saatten birkaç güne

kadar sürmektedir. DDoS saldırısına maruz kalan kurum veya kuruluşlara vereceği zarar kurumların potansiyel durumuna göre maddi veya itibar kaybı şeklinde olabilmektedir[15].

2.6.3. Finansal Kayıp

Web sitesi kesinti yaşadığında bir kuruluşun maliyeti, söz konusu kuruluşun ait olduğu sektöre bağlı olarak önemli ölçüde değişir. Neustar anketi, internete bağlı olarak çalışan kuruluşların (özellikle çevrimiçi perakende veya oyun siteleri), kesinti durumunda günlük ortalama 2.000.000 dolar (saatte yaklaşık 100.000 dolar) gelir kaybı yaşadığını ortaya koydu. Finansal hizmetler ile ilgili hizmet veren kuruluşlarda ise, kesinti durumunda bu maliyet saatte yaklaşık 10.000 dolardır. Hesaplamada kullanılan parametreler birden fazla değişkenin hesaplanması ile ortaya çıkmaktadır. Bunlar, saldırının kendisi, müşterilerin gelir kaybı ve potansiyel müşterilerin web sitesine erişememesinden kaynaklanan gelir kaybı, müşteri destek çağrılarını yanıtlamak için harcanan zaman ve olası ek mali cezalardır. 2000 yılında Yahoo ve Amazon gibi büyük web sitelerini hedef alan DDoS saldırılarının kümülatif olarak 1,2 milyar doların üzerinde hasara mal olduğu tahmin ediliyordu. Sony'nin web sayfalarına saldırganlar tarafından yapılan saldırı ile firmanın uğradığı zararın boyutu çok yüksek olduğu tahmin edilmektedir. Sony tarafından DDoS saldırısı ve veri kaybıyla ilgili temizlik için 170 milyon dolardan fazla para harcanmıştır. Ancak bazı analistler, tehlikeye atılan 77 milyon kullanıcı hesabının her biri için Sony'ye yüzlerce dolarlık nihai bir maliyet tahmin etmektedir. Analist tahminlerinden bağımsız olarak bir şey çok açıktır: DDoS saldırılarına karşı yeterince korunmayan bir kuruluşun sonrasında yapacağı maliyet aşırı derecede yüksek olacaktır [15].

2.6.4. Müşteri Zayiatı

Ankete katılan şirketler tarafından özetlenen en önemli ticari etki, müşterileriyle ilgilidir. Kurum tarafından verilen hizmetlerin 7/24 kesintisiz bir şekilde kullanıcılara ulaştırılması çok önemlidir. Bilişim sistemlerinin saldır sonucunda hizmet veremeyecek hale gelmesi firmalara iş kaybı olarak yansımaktadır. Bu kaybın büyüklüğü ise saldırının süresi ve firmanın mali büyüklüğü ile ilgili olmaktadır. Bu süreç hizmet alamayan müşterilerin şikayetleri, mali tazminat talepleri ve hatta rakipler için artan iş ile sonuçlanabilir. Buda saldırıya maruz kalan şirketin hizmet verdiği sunucularına müşterileri tarafından erişememesine yol açtığından, şirketin maddi ve prestij kaybına uğramasını sağlar [15].

2.6.5. İtibar Kaybı

İşletmeler, başarılarını göstererek manşetlere çıkmak isterler. Bir şirketin müşterilerini ve verilerini tehlikeye atan bir siber saldırının kurbanı olduğu herkes tarafından bilindiğinde, ortaya çıkan kötü tanıtım hem itibar hem de gelecekteki satışlar üzerinde yıkıcı etkilere sahip olabilir. Bilgisayar korsanlarının kurbanı olan herhangi bir şirket, ne yapılmaması gerektiğinin bir örneği haline gelir ve ortaya çıkan sonuç, genellikle kurumsal yeniden markalaşmaya kadar giden sonuçlar doğurur [15].

2.6.6. Yasal İşlemler

Saldırınca hasar gördüklerini kanıtlayabilen müşteriler, genellikle şirketin böyle bir saldırı olasılığına karşı yeterince önlem almadığını iddia ederek bir dava açabilir ve mali tazminat arayışına girebilir. Bir örnek olarak, 2011'de büyük bir DDoS saldırısına maruz kalan büyük bir borsa, normal hizmet verememelerini telafi etmek için alım satım şirketlerine yüklü miktarda cezalar ödemek zorunda kalmıştır [15].

2.6.7. Sonuç

Bir kuruluşun kendisini DoS ve DDoS saldırılarına karşı koruma yeteneği, başarısı için çok önemlidir. Dos veya DDoS saldırılarından korunmak için bilişim altyapısı üzerinde gerekli önlem ve tedbirlerin alınmaması kurum ve kuruluşların itibar, prestij, mali kayıplarına yol açmakta ve hukuki süreçlerle uğraşmak zorunda kalmaktadır. Bu tür kayıplar göz önünde bulundurulduğunda, şirketlerin yedek sunucular bulundurmak, saldırılar için ek önlem planları oluşturmak ve saldırı tespiti için karar destek sistemleri oluşturmak için zaman kaybetmeden harekete geçmesi gerekmektedir [15, 16, 17].

3. MALZEME VE YÖNTEM

3.1. Yöntem

Verilerin analizinde SPSS 11.5 ve Weka 3.7 programlarından faydalanılmıştır. Tanımlayıcı olarak nicel değişkenler için ortalama±standart sapma ve ortanca (minimum-maksimum), nitel değişkenler için ise saldırı sayısı (yüzde) kullanılmıştır. İki nitel değişken arasındaki ilişki incelenmek istendiğinde Ki-kare testi kullanılmıştır. İstatistiksel anlamlılık düzeyi 0.05 olarak alınmıştır. WEKA programında sınıflandırma yöntemlerinden Lojistik Regresyon, Çok Katmanlı Algılayıcı, Naive Bayes, Destek Vektör Makinesi ve Random Forest kullanılmıştır. Veri seti 10-fold çağraz geçerlilik kullanılarak değerlendirilmiştir. Performans ölçütü olarak ise Doğru Sınıflama Oranı, F-ölçütü, Matthews korelasyon katsayısı (MCC), Duyarlık, Precision Recall Eğrisi (PRC) ve ROC Alanı kullanılmıştır.

3.2. Veri Toplama Araçları

Çalışmada gerçek hayat dağıtık hizmet engelleme saldırı senaryoları baz alınarak 500 durum içeren simüle veri üretilmiştir. Veri setinde bulunan bağımsız değişkenler paketlerin yaşama süresi, atak yapılan paket sayısı, düşük ağ performansı durumu, web sitelerine erişimde yavaşlık durumu, ağ bağlantılarında kesilmeler olup olmaması, spam e-postaların sayışında artış olup olmaması, web sitesinin belli bölümlerine erişimin engellenmesi durumudur, bağımlı değişken ise yapılan atağın başarı durumudur.

4. BULGULAR

Bu tez çalışmasında, paketlerin yaşama süresine ait ortalama 124,34±38,60 iken, bu değişken kategorize edildiğinde <120 sn olan saldırı yüzdesi %49,2 ve ≥120 sn olan saldırı yüzdesi ise %50,8 olarak bulunmuştur. Atak yapılan paket sayısına ait ortalama 69717,29±34057,39 iken, bu değişken kategorize edildiğinde <60000 olan saldırı yüzdesi %50,0 ve ≥60000 olan saldırı yüzdesi ise %50,0 olarak bulunmuştur. Saldırıların %37,0'sinde düşük ağ performansı, %41,0'inde Web Sitelerine Erişimde Yavaşlık, %33,0'ünde Ağ Bağlantılarında Kesilmeler, %49,0'unda Spam E-postaların Sayısında Artış, %27,0'sinde Web Sitesinin Belli Bölümlerine Erişimin Engellenmesi mevcuttur. Ayrıca saldırıların %57,0'si başarısız iken %43,0'ü ise başarılı olmuştur (Tablo 4.1).

Tablo 4. 1. Tanımlayıcılar.

Değişkenler		
Paketlerin Yaşama Süresi	Ort.±SS	124,34±38,60
	Ortanca (Min.-Maks.)	120,50 (59,00-240,00)
Paketlerin Yaşama Süresi, n(%)	<120 sn	246 (49,2)
	≥120 sn	254 (50,8)
Atak Yapılan Paket Sayısı	Ort.±SS	69717,29±34057,39
	Ortanca (Min.-Maks.)	59950,50 (1603,00-275000,00)
Atak Yapılan Paket Sayısı, n(%)	<60000	250 (50,0)
	≥60000	250 (50,0)
Düşük Ağ Performansı, n(%)	Yok	315 (63,0)
	Var	185 (37,0)
Web Sitelerine Erişimde Yavaşlık, n(%)	Yok	295 (59,0)
	Var	205 (41,0)
Ağ Bağlantılarında Kesilmeler, n(%)	Yok	335 (67,0)
	Var	165 (33,0)
Spam E-postaların Sayısında Artış, n(%)	Yok	255 (51,0)
	Var	245 (49,0)
Web Sitesinin Belli Bölümlerine Erişimin Engellenmesi, n(%)	Yok	365 (73,0)
	Var	135 (27,0)
Atak Başarısı, n(%)	Başarısız	285 (57,0)
	Başarılı	215 (43,0)

Ort.:Ortalama, SS:Standart Sapma, Min.:Minimum, Maks.:Maksimum

Tablo 4.2'de atak durumu ile nitel değişkenlerin ilişkisine bakılmış ve tüm değişkenler bakımından anlamlı fark bulunmuştur ($p<0,05$). Paketlerin yaşama süresi <120 sn olan atakların %87,0'si başarısızken, paketlerin yaşama süresi ≥120 sn olan atakların ise %72,0'si başarılıdır. Atak yapılan paket sayısı <60000 olan atakların %84,4'ü başarısızken, Atak yapılan paket sayısı ≥60000 olan atakların ise %70,4'ü başarılıdır. Düşük ağ performansı görülen atakların %49,7'si başarısızken %50,3'ü ise başarılıdır. Web sitelerine erişimde yavaşlık görülen atakların %47,8'i başarısızken %52,2'si ise başarılıdır. Ağ bağlantılarında kesilmeler görülen atakların %46,1'i başarısızken %53,9'u ise başarılıdır. Spam e-posta sayısında artış görülen atakların %45,7'si başarısızken %54,3'ü ise başarılıdır. Web sitelerinin belli bölümlerine erişimin engellendiği atakların %37,0'si başarısızken %63,0'ü ise başarılıdır.

Tablo 4. 2. Atak Durumu ile Nitel Değişkenlerin İlişkisi

Değişkenler		**Başarısız**		**Başarılı**		
		n	**%**	**n**	**%**	**p değeri**
Paketlerin Yaşama Süresi	<120 sn	214	87,0	32	13,0	<0,001
	≥120 sn	71	28,0	183	72,0	
Atak Yapılan Paket Sayısı	<60000	211	84,4	39	15,6	<0,001
	≥60000	74	29,6	176	70,4	
Düşük Ağ Performansı	Yok	193	61,3	122	38,7	0,012
	Var	92	49,7	93	50,3	
Web Sitelerine Erişimde Yavaşlık	Yok	187	63,4	108	36,6	0,001
	Var	98	47,8	107	52,2	
Ağ Bağlantılarında Kesilmeler	Yok	209	62,4	126	37,6	0,001
	Var	76	46,1	89	53,9	
Spam E-postaların Sayısında Artış	Yok	173	67,8	82	32,2	<0,001
	Var	112	45,7	133	54,3	
Web Sitesinin Belli Bölümlerine Erişimin Engellenmesi	Yok	235	64,4	130	35,6	<0,001
	Var	50	37,0	85	63,0	

Tablo 4.3'teki veri madenciliği yöntemlerinin performans ölçüt değerleri verilmiştir. Bu ölçütlerden en sık kullanılanları ise Doğru Sınıflama Oranı, F-ölçütü ve MCC'dir. Bu 3 ölçüt beraber değerlendirildiğinde en iyi performansa sahip yöntem Random Forest olarak bulunmuştur. Bu yöntemi ise sırasıyla Çok Katmanlı Algılayıcı, Naive Bayes, Lojistik Regresyon ve Destek Vektör Makinesi izlemektir.

Tablo 4. 3. Veri Madenciliği Performans Ölçütleri

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Duyarlık	MCC	PRC Alanı	ROC Alanı
Lojistik Regresyon	**Başarısız**	0,909	0,878	0,849	0,705	0,924	0,908
	Başarılı	0,786	0,824	0,867	0,705	0,886	0,908
	Genel	0,856	0,855	0,857	0,705	0,907	0,908
Naive Bayes	**Başarısız**	0,926	0,895	0,866	0,747	0,928	0,909
	Başarılı	0,809	0,849	0,892	0,747	0,890	0,909
	Genel	0,876	0,875	0,877	0,747	0,912	0,909
Çok Katmanlı Algılayıcı	**Başarısız**	0,940	0,912	0,884	0,788	0,949	0,941
	Başarılı	0,837	0,874	0,914	0,788	0,932	0,941
	Genel	0,896	0,895	0,897	0,788	0,942	0,941
Destek Vektör Makinesi	**Başarısız**	0,877	0,862	0,847	0,672	0,813	0,834
	Başarılı	0,791	0,810	0,829	0,672	0,746	0,834
	Genel	0,840	0,839	0,840	0,672	0,784	0,834
Random Forest	**Başarısız**	0,944	0,913	0,885	0,792	0,951	0,951
	Başarılı	0,837	0,876	0,918	0,792	0,946	0,951
	Genel	0,898	0,897	0,899	0,792	0,949	0,951

MCC: Matthews korelasyon katsayısı, PRC:Precision Recall Eğrisi

Şekil 4.1'deki ağaç diyagramı Random Forest yönteminin saldırıyı başarılı ya da başarısız olarak sınıflarken kullandığı algoritmayı göstermektedir. Algoritmaya ait detaylı açıklama ise aşağıdaki gibidir:

- Eğer Paketlerin Yaşama Süresi 120 sn'den azsa ve Atak Yapılan Paket Sayısı 60000'den azsa yöntem saldırıyı başarısız olarak sınıflamaktadır.

- Eğer Paketlerin Yaşama Süresi 120 sn'den azsa ve Atak Yapılan Paket Sayısı 60000'den fazla ise ve Spam E-postaların Sayısında Artış yoksa yöntem saldırıyı başarısız olarak sınıflamaktadır.

- Eğer Paketlerin Yaşama Süresi 120 sn'den azsa ve Atak Yapılan Paket Sayısı 60000'den fazla ise ve Spam E-postaların Sayısında Artış varsa ve Ağ Bağlantılarında Kesilmeler yoksa yöntem saldırıyı başarısız olarak sınıflamaktadır.

- Eğer Paketlerin Yaşama Süresi 120 sn'den azsa ve Atak Yapılan Paket Sayısı 60000'den fazla ise ve Spam E-postaların Sayısında Artış varsa ve Ağ Bağlantılarında Kesilmeler varsa yöntem saldırıyı başarılı olarak sınıflamaktadır.

- Eğer Paketlerin Yaşama Süresi 120 sn'den fazla ise ve Atak Yapılan Paket Sayısı 60000'den azsa ve Web Sitesinin Belli Bölümlerine Erişim Engellenmesi yoksa yöntem saldırıyı başarısız olarak sınıflamaktadır.

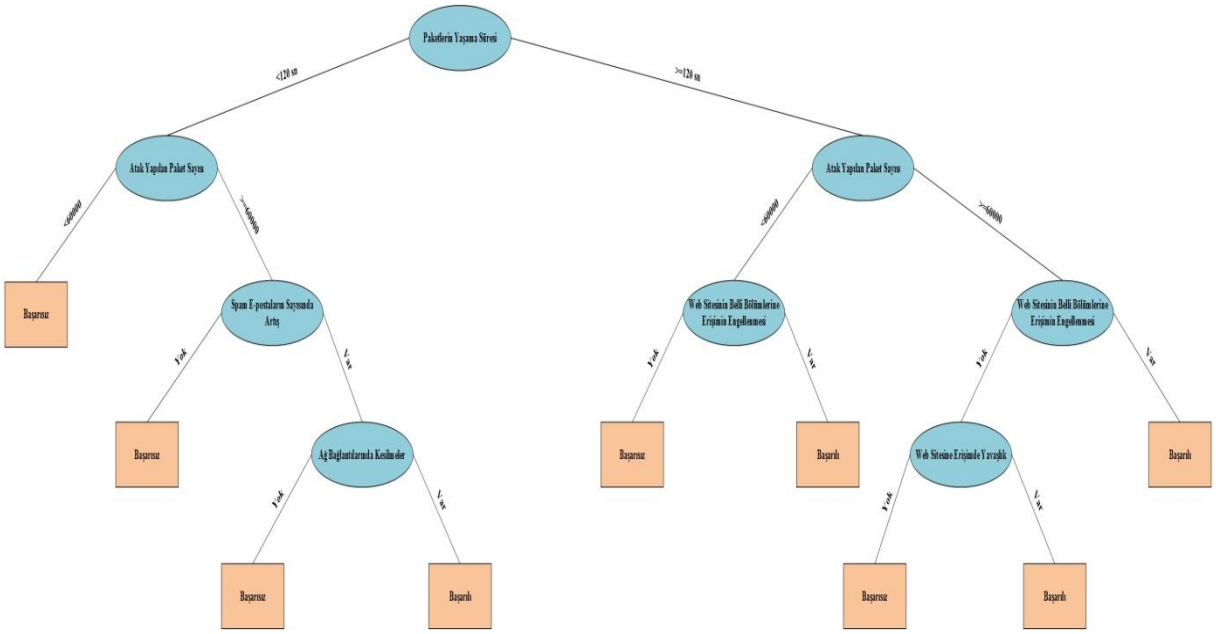
- Eğer Paketlerin Yaşama Süresi 120 sn'den fazla ise ve Atak Yapılan Paket Sayısı 60000'den azsa ve Web Sitesinin Belli Bölümlerine Erişim Engellenmesi varsa yöntem saldırıyı başarılı olarak sınıflamaktadır.

- Eğer Paketlerin Yaşama Süresi 120 sn'den fazla ise ve Atak Yapılan Paket Sayısı 60000'den fazla ise ve Web Sitesinin Belli Bölümlerine Erişim Engellenmesi yoksa ve Web Sitesine Erişimde Yavaşlık yoksa yöntem saldırıyı başarısız olarak sınıflamaktadır.

- Eğer Paketlerin Yaşama Süresi 120 sn'den fazla ise ve Atak Yapılan Paket Sayısı 60000'den fazla ise ve Web Sitesinin Belli Bölümlerine Erişim Engellenmesi yoksa ve Web Sitesine Erişimde Yavaşlık varsa yöntem saldırıyı başarılı olarak sınıflamaktadır.

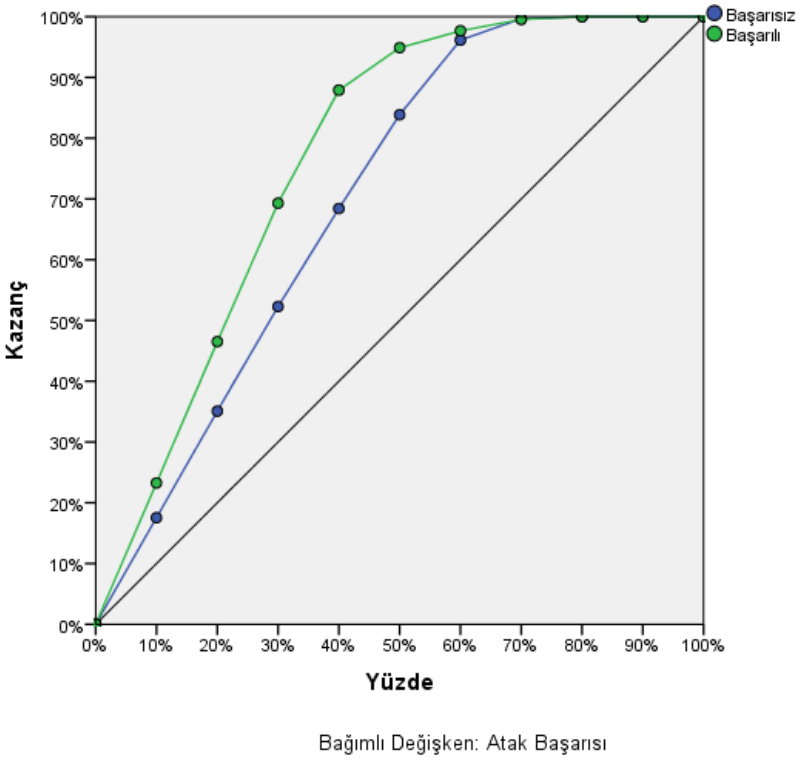
- Eğer Paketlerin Yaşama Süresi 120 sn'den fazla ise ve Atak Yapılan Paket Sayısı 60000'den fazla ise ve Web Sitesinin Belli Bölümlerine Erişim Engellenmesi varsa yöntem saldırıyı başarılı olarak sınıflamaktadır.

Random Forest yönteminde kullanılan 100 ağaçtan birine ait örnek ağaç diyagramı ise Şekil 4.1'de verilmiştir.



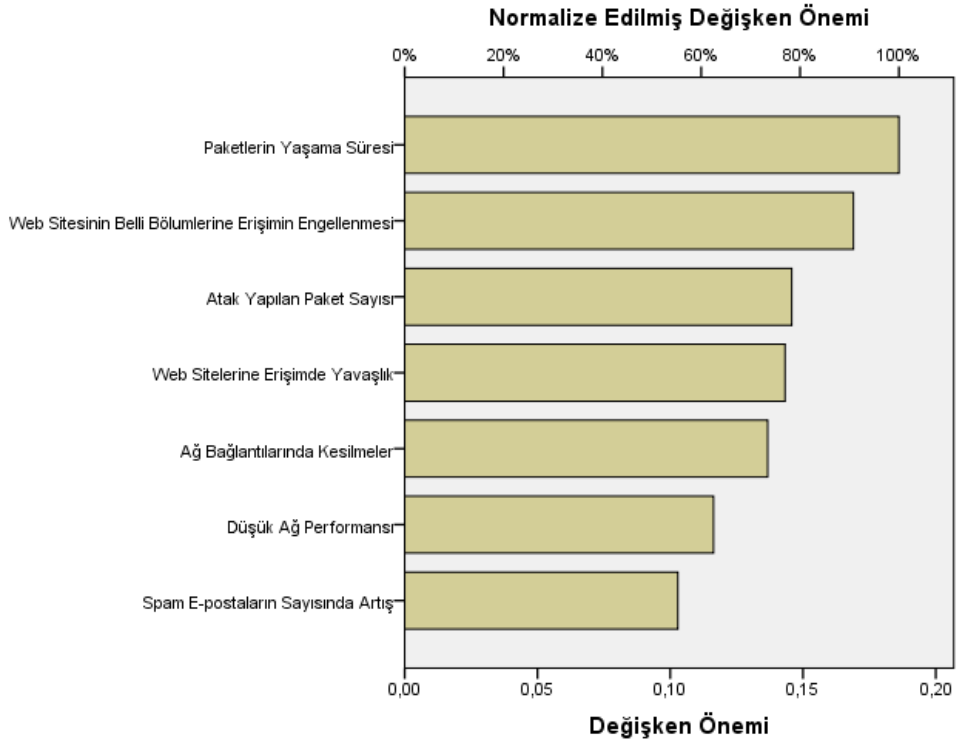
Şekil 4. 1. Random Forest Yöntemindeki 100 Ağaçtan Rastgele Birine ait Ağaç Diyagramı

Çok Katmanlı Algılayıcı yöntemine ait kazanç grafiği Şekil 4.2'de, veri setinin yüzde kaçı kullanıldığı zaman başarılı/başarısız kategorilerine ait ne kadarlık bilgi/kazanç elde edebildiğimizi göstermektedir. Örneğin, veri setindeki verinin sadece %10'u kullanıldığı zaman başarısız olarak sınıfladığımız kategoriye ait yaklaşık %20'lik, başarılı olarak sınıfladığımız kategoriye ait ise yaklaşık %25'lik bilgi/kazanç elde edilebilmektedir. Benzer şekilde verinin %50'si kullanıldığı zaman ise başarısız olarak sınıfladığımız kategoriye ait yaklaşık %85'lik, başarılı olarak sınıfladığımız kategoriye ait ise yaklaşık %95'lik bilgi/kazanç elde edilebilmektedir.



Şekil 4. 2. Çok Katmanlı Algılayıcı Yöntemine ait Kazanç Grafiği

Değişken önemi grafiği ise Şekil 4.3'de verilmiştir. Bu grafikte ise bağımlı değişkeni sınıflandırmak için kullandığımız değişkenlerin veri setine kattığı değere/öneme ait grafik verilmiştir. Veri seti için en önemli değişken olarak Paketlerin Yaşama Süresi bulunmuştur. Bu değişkeni sırasıyla Web Sitesinin Belli Bölümlerine Erişimin Engellenmesi, Atak Yapılan Paket Sayısı, Web Sitelerine Erişimde Yavaşlık, Ağ Bağlantılarında Kesilmeler, Düşük Ağ Performansı ve Spam E-postaların Sayısında Artış değişkenleri takip etmektedir.



Şekil 4. 3. Çok Katmanlı Algılayıcı Yöntemine ait Değişken Önemi Grafiği

5. TARTIŞMA VE SONUÇ

Veri madenciliğine duyulan ihtiyaç gün geçtikçe artmaktadır. Son zamanlarda yapılan ve başarılı olan veri madenciliği temelli uygulamalar görüldükçe bu alana ilgi artmaktadır. Özellikle sağlık, ticaret, bilişim vb. alanlarda birkaç adım ileriyi görmenin önemi düşünüldüğünde, veri madenciliği ile tahmin yapan karar destek sistemleri vb. modelleme tabanlı yazılımlar oldukça popüler hale gelmektedir.

Veri madenciliği birçok alan için çok faydalı olsa da, analizler ve çıkan sonuçların yorumlaması oldukça zordur. Gerçek hayatta elde edilen gürültülü, eksik veriler düşünüldüğünde, veri önişleme adımlarını uygulamadan analizlere başlamanın yanlış sonuçlar doğuracağı, bizi yanlış yere yönelteceği de bilinmelidir. Bu yüzden bu tür zaman, emek ve uzmanlık bilgisi gerektiren çalışmalarda, alan uzmanı ile veri madenciliği konusundaki uzman kişinin birlikte çalışması şarttır.

Son zamanlarda hızla artan siber saldırılar düşünüldüğünde, bilişim alanında bu tür bir sisteme çok ihtiyaç duyulmaktadır. Saldırıların verebileceği zararları saldırının başında tahmin edip, yöneticilere aksiyon planı yapacak kadar zaman tanıyacak bir model, şirketler için çok gereklidir. Çünkü bu tür saldırılar sadece para değil, prestij, güven vb. kayıplara yol açmakta ve şirketlere geri dönüşü imkansız zararlar vermektedir.

Bu tez çalışmasında, sınıflamalar için iyi bir performansa sahip olan Random Forest yöntemi alınmıştır. Bu yönteme göre başarısız olarak sınıflanan bir saldırının başarısız olma ihtimali %94,4 olarak bulunmuştur. Başka bir deyişle, başarısız olarak sınıflanan 100 adet saldırıdan 94,4'ü doğru sınıflanmıştır. Bu yönteme göre başarılı olarak sınıflanan bir saldırının başarılı olma ihtimali %83,7 olarak bulunmuştur. Başka bir deyişle, başarılı olarak sınıflanan 100 adet saldırıdan 83,7'si doğru sınıflanmıştır. Genel sınıflama performansına bakıldığında ise başarılı ya da başarısız olarak yapılan 100 tahminden 89,8'i doğrudur.

Yaptığımız modelleme tabanlı sistem, saldırıların başında, saldırının sistemde başarılı olup/olamayacağını yani sisteme büyük zararlar verip veremeyeceğini tahmin etmektedir. Bu sayede yapılan saldırıların zararlarını minimuma indirgemek amaçlanmaktadır. Gelişen sistemler, teknoloji düşünüldüğünde bir saldırının gerçekleşmesini engellemek neredeyse imkansızdır, yapılabilecek en doğru hamle saldırılardan en az zararla kurtulmayı planlamaktır.

KAYNAKLAR

[1]. Kantardzic, M., 2011, Data mining: concepts, models, methods, and algorithms, John Wiley & Sons, 3rd ed.

[2]. Aggarwal, C.C., 2015, Data mining: the textbook, Springer.

[3]. Han, J., Kamber, M., Pei, J., 2011, Data mining concepts and techniques, The Morgan Kaufmann Series in Data Management Systems, 5(4), 83-124.

[4]. Maheshwari, A., 2014, Business intelligence and data mining, Business Expert Press.

[5]. Nisbet, R., Elder, J., Miner, G., 2009, Handbook of statistical analysis and data mining applications, Academic Press.

[6]. Bramer, M., 2007, Principles of data mining, London: Springer. Vol. 180.

[7]. Brown, M.S., 2014, Data mining for dummies, John Wiley & Sons.

[8]. Sammut, C., & Webb, G. I. (Eds.). (2011). Encyclopedia of machine learning. Springer Science & Business Media.

[9]. Larose, D.T., 2015, Data mining and predictive analytics, John Wiley & Sons.

[10]. Yang, X.S., 2019, Introduction to algorithms for data mining and machine learning, Academic press.

[11]. Dean, J., 2014, Big data, data mining, and machine learning: value creation for business leaders and practitioners, John Wiley & Sons.

[12]. Provost, F., Fawcett, T., 2013, Data Science for Business: What you need to know about data mining and data-analytic thinking, O'Reilly Media Inc.

[13]. Tan, P.N., Steinbach, M., Kumar, V., 2016, Introduction to data mining, Pearson Education India.

[14]. Hastie, T., Tibshirani, R., Friedman, J., 2017, The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2nd ed.

[15]. Kenig, R., Manor, D., Gadot, Z., Trauner, D., 2013, DdoS survival handbook, Radware.

[16]. Yalda M.K., 2019, Detection of Sources Being Used on DdoS Attacks, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Bilişim Enstitüsü.

[17]. Özer E., 2015, Derin Paket Analizi ile DdoS Ataklarının Tespiti ve DLP Uygulaması, Yüksek Lisans Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü.

[18]. Bindra, N., Sood, M., 2019, Detecting DdoS attacks using machine learning techniques and contemporary intrusion detection dataset, Automatic Control and Computer Sciences, 53(5), 419-428.

[19]. Lee, K., Kim, J., Kwon, K. H., Han, Y., Kim, S., 2008, DdoS attack detection method using cluster analysis. Expert systems with applications, 34(3), 1659-1665.

[20]. Alkasassbeh, M., Al-Naymat, G., Hassanat, A., Almseidin, M., 2016, Detecting distributed denial of service attacks using data mining techniques, International Journal of Advanced Computer Science and Applications, 7(1), 436-445.

[21]. Zhong, R., Yue, G., 2010, DDoS detection system based on data mining, In Proceedings of the 2nd International Symposium on Networking and Network Security, Jinggangshan, China, Vol. 2, p. 62.

[22]. Lakshminarasimman, S., Ruswin, S., Sundarakantham, K., 2017, Detecting DDoS attacks using decision tree algorithm. In 2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN) p. 1-6.

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Özgür SEZEN
Doğum Yeri	
Doğum Tarihi	
Uyruğu	☑ T.C. ❒ Diğer:

Eğitim Bilgileri	
Lisans	
Üniversite	Lefke Avrupa Üniversitesi
Fakülte	Mühendislik-Mimarlık
Bölümü	Bilgisayar Mühendisliği
Mezuniyet Yılı	2004

Yüksek Lisans	
Üniversite	Kırşehir Ahi Evran Üniversitesi
Enstitü Adı	Fen Bilimleri
Anabilim Dalı	İleri Teknolojiler
Programı	İleri Teknolojiler Ana Bilim Dalı Tezli Yüksek Lisans Programı
Mezuniyet Tarihi	2022

Doktora	
Üniversite	
Enstitü Adı	
Anabilim Dalı	
Programı	Program Adı
Mezuniyet Tarihi	

Makale ve Bildiriler