

VERİ MADENCİLİĞİ TEKNİKLERİ İLE BİLGİSAYAR AĞLARINDA ANOMALİ TESPİTİ

**2013
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ**

İdris BUDAK

**VERİ MADENCİLİĞİ TEKNİKLERİ İLE BİLGİSAYAR AĞLARINDA
ANOMALİ TESPİTİ**

İdris BUDAK

**Karabük Üniversitesi
Fen Bilimleri Enstitüsü**

**Bilgisayar Mühendisliği Anabilim Dalında
Yüksek Lisans Tezi
Olarak Hazırlanmıştır**

KARABÜK

Haziran 2013

İdris BUDAK tarafından hazırlanan “VERİ MADENCİLİĞİ TEKNİKLERİ İLE BİLGİSAYAR AĞLARINDA ANOMALİ TESPİTİ” başlıklı bu tezin Yüksek Lisans Tezi olarak uygun olduğunu onaylarım.

Yrd. Doç. Dr. Baha ŞEN



Tez Danışmanı, Bilgisayar Mühendisliği Anabilim Dalı

Bu çalışma, jürimiz tarafından oy birliği ile Bilgisayar Mühendisliği Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir. 20/06/2013

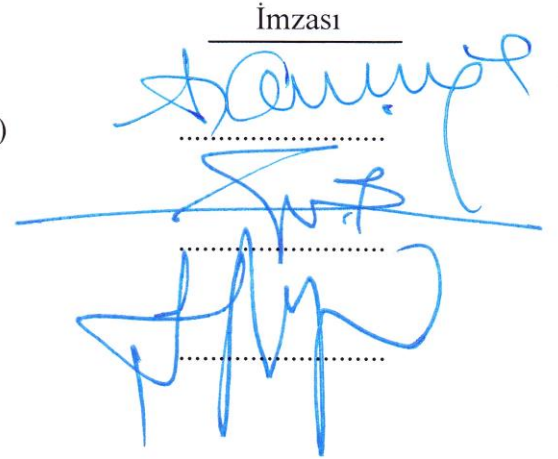
Ünvanı, Adı SOYADI (Kurumu)

Başkan: Prof. Dr. Abdullah ÇAVUŞOĞLU (YBÜ)

Üye : Yrd. Doç. Dr. Baha ŞEN (KBÜ)

Üye : Yrd. Doç. Dr. Şafak BAYIR (KBÜ)

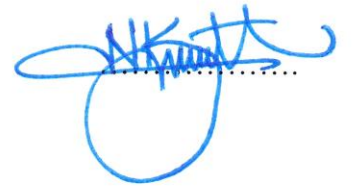
İmzası



...../...../2013

KBÜ Fen Bilimleri Enstitüsü Yönetim Kurulu, bu tez ile, Yüksek Lisans derecesini onamıştır.

Prof. Dr. Nizamettin KAHRAMAN
Fen Bilimleri Enstitüsü Müdürü





“Bu tezdeki tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak elde edildiğini ve sunulduğunu; ayrıca bu kuralların ve ilkelerin gerektirdiği şekilde, bu çalışmadan kaynaklanmayan bütün atıfları yaptığımı beyan ederim.”

İdris BUDAK

ÖZET

Yüksek Lisans Tezi

VERİ MADENCİLİĞİ TEKNİKLERİ İLE BİLGİSAYAR AĞLARINDA ANOMALİ TESPİTİ

İdris BUDAK

Karabük Üniversitesi

Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:

Yrd. Doç. Dr. Baha ŞEN

Haziran 2013, 59 sayfa

Bilgi çağının en önemli unsurları olan bilginin üretilmesi, korunması ve erişilmesinde kullanılan bilgisayar ağlarının hız, güvenlik ve sürekliliği bugün hayatın vazgeçilmezleri arasında ilk sıralarda yerini almaktadır. Bu yüzden ağdaki düzensizliklerin zamanında tespit edilip önlemlerin alınması gerekmektedir.

Bu çalışmadaki amaç bilgisayar ağlarındaki anomali tespitinde veri madenciliği tekniklerinden faydalananarak İkili Lojistik Regresyon tekniğinin uygulanabilirliğini incelemektir. Bu amaçla saldırı tespit sistemleriyle ilgili çalışmalarda en sık kullanılan veri setlerinden olan KDD Cup'99 veri seti kullanılarak bir matematiksel model oluşturulup bu modelin uygunluğu test edilmiştir.

Çalışma sonucu elde edilen model, eğitim veri seti üzerinde uygulandığında gerçekte saldırı olan kayıtların % 99,6 kadarının saldırı olduğunu bulmuştur. Test veri setinde ise bu oran % 99,9 olarak tespit edilmiştir.

Anahtar Sözcükler : Saldırı tespit sistemleri, bilgisayar ağlarında anomali tespiti, lojistik regresyon, veri madenciliği teknikleri.

Bilim Kodu : 902.1.063



ABSTRACT

M. Sc. Thesis

ANOMALY DETECTION IN COMPUTER NETWORKS WITH DATA MINING

İdris BUDAK

Karabük University

Graduate School of Natural and Applied Sciences

Department of Computer Engineering

Thesis Advisor:

Assist. Prof. Dr. Baha ŞEN

June 2013, 59 pages

The most important elements of the information age is generation, security, and access to knowledge, so the location of the first rank in indispensable of life today is computer networks with high speed, security, and continuity. Therefore, measures should be taken timely when anomalies detected in the network.

The purpose of this study is to detect if logistic regression with data mining techniques is usable in anomaly detection for computer networks. For this purpose KDD Cup'99 data set was used for mathematical model and the result model is tested for fitness.

As a result of the model, in training dataset our model correctly found % 99,6 of the real anomalies that are in the data set. In testing dataset our model found % 99,9 of the real anomalies that are in the data set.

Key Word : Intrusion detection systems, anomaly detection in computer networks, logistic regression, data mining techniques.

Science Code : 902.1.063



TEŞEKKÜR

Bu tez çalışmasının planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım, yönlendirme ve bilgilendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren sayın hocam Yrd. Doç. Dr. Baha ŞEN'e sonsuz teşekkürlerimi sunarım.

Ağ teknolojilerinin işleyişini anlamam konusunda kurum kaynaklarından faydalanmamı sağlayarak yardımlarını esirgemeyen, Karabük Üniversitesi Bilgi İşlem Daire Başkanı Recep KILINÇ'a teşekkür ederim.

Sevgili aileme manevi hiçbir yardımı esirgemedi yanımda oldukları için tüm kalbimle teşekkür ederim.

İÇİNDEKİLER

Sayfa

KABUL	ii
ÖZET	iv
ABSTRACT	vi
TEŞEKKÜR	viii
İÇİNDEKİLER	ix
ŞEKİLLER DİZİNİ	xii
ÇİZELGELER DİZİNİ	xiii
SİMGELER VE KISALTMALAR DİZİNİ	xiv
BÖLÜM 1.	1
GİRİŞ	1
BÖLÜM 2.	4
AĞ GÜVENLİĞİ	4
2.1. GÜVENLİĞİ SINIFLANDIRMAK	4
2.2. BİLGİSAYAR AĞLARINDA ALINABİLECEK 4 TEMEL ÖNLEM	5
2.2.1. Kurumsal Politika Ve Bilinçlendirme Çalışmaları	5
2.2.2. Kullanıcı Bilgisayarlarında Alınabilecek Temel Önlemler	5
2.2.3. Ağ Cihazlarında Alınabilecek Temel Önlemler	6
2.2.4. Güvenlik Yazılımlarının kullanılması	6
2.3. VPN (SANAL ÖZEL BİLGİSAYAR AĞLARI)	6
2.4. GÜVENLİK DUVARI(FIREWALL)	7
BÖLÜM 3.	10
SALDIRI TESPİT SİSTEMLERİ	10
3.1. SALDIRI TESPİT YÖNTEMİ	11
3.1.1. En Çok ve Baz Değer	12
3.1.2. İz Eşleştirme	12

BÖLÜM 4.	14
TEMEL İSTATİSTİK KAVRAMLARI.....	14
4.1. DEĞİŞKENLER	14
4.1.1. Nitel (Kategorik, Qualitative, Categorical) Değişkenler	14
4.1.2. Sayısal (Nicel) Değişkenler	14
4.2. DOĞRULUK VE DUYARLILIK.....	15
 BÖLÜM 5.	17
REGRESYON.....	17
5.1. REGRESYON NEDİR.....	17
5.2. REGRESYON ANALİZİNİN TARİHİ	18
5.3. REGRESYON ANALİZİ.....	19
5.4. EN KÜÇÜK KARELER METODU	20
 BÖLÜM 6.	29
LOJİSTİK REGRESYON.....	29
6.1. LOJİSTİK REGRESYON ANALİZİ.....	29
6.2. LOJİSTİK REGRESYON İLE DOĞRUSAL REGRESYONUN FARKI	31
6.3. LOJİSTİK REGRESYON MODELLERİ.....	32
6.4. LOGİT MODEL.....	32
6.4.1. Lojistik Regresyon Formülü	34
6.4.2. Logit Model ile Lojistik Regresyon Arasındaki Benzerlik ve Farklar ...	35
6.4.3. Lojistik Regresyon İçin Basit bir Örnek ve Çözümü.....	35
 BÖLÜM 7.	39
MATEMATİKSEL MODELİN OLUŞTURULMASI.....	39
7.1. VERİ SETİ ANALİZİ	39
7.2. VERİ SETİNİN HAZIRLANMASI.....	41
7.3. KOLON (DEĞİŞKEN) SEÇİMİ.....	43
7.4. MODELİN OLUŞTURULMASI.....	47
7.5 MODELİN UYGULAMASINA ÖRNEK	50
7.6 MODELİN TEST VERİ SETİNDE UYGULAMASI	51

	<u>Sayfa</u>
BÖLÜM 8.	53
SONUÇLAR	53
KAYNAKLAR	54
ÖZGEÇMİŞ	59



ŞEKİLLER DİZİNİ

	<u>Sayfa</u>
Şekil 2.1. VPN Erişimi	7
Şekil 2.2. VPN'nin çalışma şekli.....	7
Şekil 2.3. Firewall konumlandırılması	8
Şekil 2.4. UTM Firewall özellikleri	9
Şekil 3.1. Saldırı tespit sistemi konumlandırılması.....	13
Şekil 4.1. Değişken tiplerinin şematik gösterimi	15
Şekil 4.2. Doğruluk ve duyarlılık farkı.....	16
Şekil 5.1. En küçük kareler yöntemi	22
Şekil 5.2. En küçük kareler yöntemi ile bulunan grafik.....	27
Şekil 5.3. Türevin anlamı	27
Şekil 5.4. Genel türev formülleri.....	28
Şekil 6.1. Lojistik regresyon analizinin kullanım amacına örnek	30
Şekil 6.2. Olasılık ile tahmin edici arasındaki ilişki.....	34
Şekil 6.3. Lojistik ilişkinin biçimi	34
Şekil 7.1. Veri setinin oluşturulduğu ağın şeması	39
Şekil 7.2. Veri setinin wireshark ile analizi	40

ÇİZELGELER DİZİNİ

	<u>Sayfa</u>
Çizelge 5.1. En küçük kareler yöntemi örnek veri seti	25
Çizelge 5.2. En küçük kareler yöntemi örnek çözümü	26
Çizelge 6.1. Doğrusal ve lojistik regresyon karşılaştırması	31
Çizelge 6.2. UDP Protokolü ODDS değerleri.....	33
Çizelge 6.3. Web sunucusu bağlantıları eğitim veri seti.....	36
Çizelge 6.4. Lojistik regresyon web sunucu bağlantısı örnek çözümü.....	37
Çizelge 6.5. Lojistik regresyon web sunucu bağlantısı olasılığı için verilenler.....	38
Çizelge 7.1. Veri seti örnek görüntüsü.....	46
Çizelge 7.2. Durum işleme özeti.....	47
Çizelge 7.3. Bağımlı değişken kodlaması.....	47
Çizelge 7.4. Sınıflandırma tablosu.....	48
Çizelge 7.5. Eşitlikteki değişkenler.....	49
Çizelge 7.6. Test verisi sınıflandırma tablosu.....	51

SİMGELER VE KISALTMALAR DİZİNİ

SİMGELER

Δ : Fark, $\Delta y = y_1 - y_2$

d : Türev, $dy/dx = df(x)/dy = f'(x)$

∂ : Kısmi Türev, $y=f(a,b)$ için y nin a ya göre kısmi türevi: $\partial y/\partial a$

KISALTMALAR

BGYS : Bilgi Güvenliği Yönetim Sistemi.

ISO : Uluslararası Standartlar Organizasyonu (International Standards Organization).

UPS : Kesintisiz Güç Kaynağı (Uninterruptable Power Supply).

F/O : Fiber Optik.

VPN : Sanal Özel Bilgisayar Ağları.

LAN : Yerel Alan Ağı (Local Area Network).

UTM : Bütünleşik Tehdit Yöneticisi (Unified Threat Manager)

DARPA : ABD Savunma Bakanlığı İleri Araştırma Projeleri Birimi (The Defense Advanced Research Projects Agency).

IP : İnternet Protokolü (İnternet protokolü adresi olarak da kullanılır).

SYN : Sunucuya gönderilen bağlantı isteği paketi.

KDD Cup : Knowledge Discovery and Data Mining Cup (Bilgi keşfi ve veri madenciliği ile ilgili her yıl düzenlenen bir konferanstır).

BÖLÜM 1

GİRİŞ

Yaşadığımız bilgi çağında şimdiye kadar hiç olmadığı kadar bilgi üretilmekte işlenmekte ve bunlara erişilmektedir. Bilginin bu denli hızlı üretilip yayılmasında hiç kuşkusuz bilgisayar teknolojileri en büyük teknik faydayı sağlamaktadır. Bilgiye erişim ve paylaşım için ise en fazla verimliliği bilgisayar ağları sağladığından, ağ işleyişinin düzgün olması hayati önem taşımaktadır. Ağ trafiğindeki anormallikler ise ağın gerektiği gibi kullanımını engelleyen unsurların başında gelmektedir. Bu anormallikler altyapı sorunlarından kaynaklanabileceği gibi ağın kötüye kullanılması veya ağa yapılan saldırılardan da kaynaklanabilmektedir.

Birçok kaynakta saldırı tespit sistemleri olarak da anılan anomali tespit sistemleri ağda oluşan düzensizlikleri tespit edip ilgili kişileri veya yazılımları uyardırmayı sağlayan sistemlerdir.

Lojistik regresyonun bilgisayar ağlarındaki anomali tespitinde kullanımıyla ilgili literatürdeki örneklerden dördü aşağıda özetlenmiştir:

- 1- “Random Effects Logistic Regression Model For Anomaly Detection” adlı çalışmada “random effects logistic regression” yöntemi ile bir model oluşturulmuştur. Sonuçta modelin basitliği, değişken sayısının azlığı ve sonuçların lojistik regresyon ile yorumlanabilmesi avantajlarının elde edildiği görülmüştür [1].
- 2- “Protocol-Based Classification for Intrusion Detection” adlı çalışmada ise diskriminant analizi, lojistik regresyon analizi ve ana bileşen (principal component) analiz yöntemleri karşılaştırılmıştır. IDS modelinde özellik seçim metodu olarak lojistik regresyon kullanıldığında KDDCUP’99 data

setinde % 99,95 oranında doğruluk sağlanmıştır. Bu çalışma ile bir IDS'in analiz edeceği verilerin sadeleşmiş olması için yaklaşık 40 grup veri içinden en öncelikli hangi verilerin analize alınması gerektiği lojistik regresyon yöntemiyle elde edilerek, sonuçta bir IDS'in daha az veriyle daha performanslı ve doğru çalışması sağlanmıştır [2].

3- “Scan Detection on Very Large Networks Using Logistic Regression Modeling” adlı çalışmada ise uzmanların fikirleri ve verilerin el yordamıyla (manuel) incelenmesiyle oluşturulmuş "bayesian logistic regression" yöntemi kullanılmıştır. Çalışmada 300 TCP olayında % 0,4 oranında false pozitif hata ile % 95,5 oranında doğruluk ile anomali tespiti gerçekleştirilmiştir [3].

4- “HoneyStat: Local Worm Detection Using Honeypots” adlı çalışmada honeypot'larla benzer özelliklerde honeystat sistemi geliştirilmiştir. Amaç yerel (local) ağlarda solucan (worm) tespiti yapmaktır. Bunun için lojistik regresyon modeli kullanılmış ve yeni üretilen solucanlar (zero day worms) üzerinde başarılı sonuçlar elde edilmiştir. Zaten çalışmadaki amaç henüz imza (signature) değeri bilinmeyen solucanlar üzerinde başarılı tespit işlemi yapabilmektir [4].

Günümüzde çeşitli organizasyon ve kurumlar tarafından üretilmiş gerek ticari gerekse açık kaynak kodlu birçok saldırı tespit sistemi mevcuttur. Bu tez çalışmasının özgün tarafı ise lojistik regresyonun çözüm yöntemlerinden olan logit modelin kullanılarak ikili lojistik regresyon ile tüm ağ trafiğinin analiz edilip belli bir anda trafikte anomali olma olasılığının ne olduğunu gösteren bir çalışma olmasıdır. Bu çalışmada lojistik regresyon, hem model parametre seçiminde hem de nihai modelin oluşturulmasında kullanılmıştır.

Bu çalışmada bilgisayar ağlarındaki anomali tespitinde ikili lojistik regresyon tekniğinin uygulanabilirliğini incelemek amacıyla öncelikle “Bilgisayar Ağlarında Saldırı Tespit Sistemleri” anlatılmış, daha sonra lojistik regresyon konusunun anlaşılması için “En Küçük Kareler Yöntemi” ve “Regresyon Analiz”inden bahsedilmiş, en sonunda da bilgisayar ağlarında gerçekleştirilen saldırılardan oluşan

veri seti üzerinde çalışılarak “Lojistik Regresyon” yöntemiyle analiz edilip bir matematiksel model elde edilmiş ve modelin uygunluğu test edilmiştir.



BÖLÜM 2

AĞ GÜVENLİĞİ

Ağ güvenliği ağda yer alan bilgilerin güvenliğini ve ağ üzerinden verilen hizmetlerin sürekliliğini sağlar. Ağ güvenliği tek bir yönteme dayanmaz. Uygulanacak farklı çözümler birbirini destekleyerek bir bütün oluşturur.

Bir veya daha fazla ürün tarafından uygulanan bir güvenlik modeli, üç ana amaca hizmet etmelidir:

- Confidentiality (Gizlilik): Bilginin tanılanmış/doğrulanmış kişiler haricinde okunmasına karşı korumadır.
- Integrity (Bütünlük): Bilginin yetkisi olmayanlarca modifiye edilmesine karşı korumadır.
- Availability (Erişilebilirlik): Bilginin ya da bilgisayar kaynaklarının hizmet vermesinin engellenmesine karşı korunma.

Bu üç maddenin akılda kalması için CIA kısaltması, bilgisayar güvenliğindeki üç amacı simgeleyen kolay hatırlanabilir bir kelimedir [5].

2.1. GÜVENLİĞİ SINIFLANDIRMAK

Bilgi sistemleri güvenliğini üç ana bölüme ayırabiliriz:

- Mantıksal güvenlik bilgi sisteminin iletişim ağları vasıtası ile maruz kalabileceği tehditleri kapsamaktadır.

- Fiziksel güvenlik ise bilgi sistemlerini barındıran fiziksel altyapının güvenliğini tarif etmektedir. Fiziksel güvenliğin kapsamına sunucu ve istemci donanımları, sistem odası, sistem odasının bulunduğu bina, güç hatları gibi bileşenler girer.

- Çevre güvenliği ise fiziksel güvenlikle bir düşünülebilir, ayrıldığı nokta ise bilgi sistemini barındıran bina veya kampüs alanının sınırlarında alınacak güvenlik önlemleridir [6].

2.2. BİLGİSAYAR AĞLARINDA ALINABİLECEK 4 TEMEL ÖNLEM

2.2.1. Kurumsal Politika Ve Bilinçlendirme Çalışmaları

Kurum veya kuruluşların üst düzeyde bilgi güvenliğini ve iş sürekliliğini sağlamaları için, teknik önlemlerin yanında teknik olmayan (insan faktörü, prosedürel faktörler, vb.) önlemlerin ve denetimlerin alınması, tüm bu süreçlerin devamlılığının sağlanması ve bilgi güvenliği standartlarına uygun olarak yönetilebilmesi amacıyla yönetim tarafından desteklenen, insanları, iş süreçlerini ve bilişim teknolojilerini kapsayan bilgi güvenliği standartlarına uygun olarak Bilgi Güvenliği Yönetim Sistemi (BGYS) kurmaları gerekmektedir. BGYS, ISO 27001 standardının öngördüğü bir yapıdır.

2.2.2. Kullanıcı Bilgisayarlarında Alınabilecek Temel Önlemler

- Antivirüs programı kullanılmalı.
- İşletim sistemi ve içindeki programların yamaları yapılmalı.
- Bilgisayarlar ağda paylaşım açılacaksa kuvvetli bir şifre verilmeli.
- Güvenli olmadığı bilinen sitelere (bedava mp3, program vb.) girilmemeli, internette bedava dağıtılan her programa güvenilmemeli.

2.2.3. Ağ Cihazlarında Alınabilecek Temel Önlemler

Fiziksel Altyapının Sağlıklı Çalışmasının Sağlanması:

- Aktif cihazlar nemden uzak, herkesin kolayca erişemeyeceği merkezi noktalara konumlandırılmalı.
- Cihazlarda UPS, sigortalı ve topraklı prizler kullanılmalı.
- Kablo güvenliği sağlanmalı, binalar arasında çelik zırhlı outdoor fiber optik kablo kullanılmalı ve F/O güzergahlarında (özellikle kazı çalışmaları için) yeterli ve standartlara uygun uyarı levhaları olmalıdır.

Detaylı Ağ Topolojisinin Oluşturulup Güncellemesi:

- Bağlantı ve yerleşim şemaları, alt ağlar (subnet), bilgisayar sayıları, band genişliği (bandwith), trafik analizleri ve loglanması, erişim yöntemleri, konfigürasyon yönetimi, yazılım yükseltmeleri, performans takibi, verilen sabit IP adresleri gibi unsurlar kayıt altında tutulmalı.

2.2.4. Güvenlik Yazılımlarının kullanılması

VPN, Firewall, Saldırı Tespit ve Engelleme Sistemleri kullanılmalıdır [7,8].

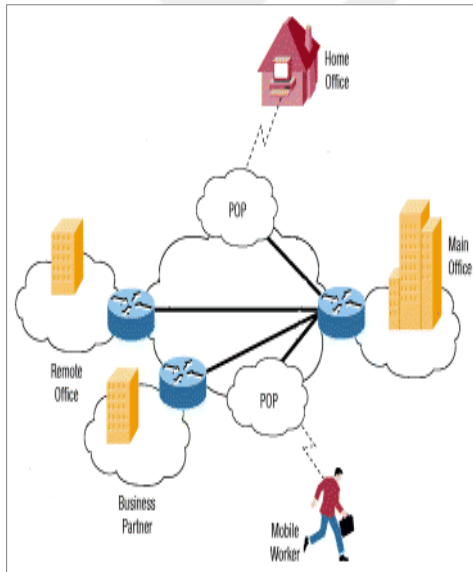
2.3. VPN (SANAL ÖZEL BİLGİSAYAR AĞLARI)

Güvenlik açısından ideal olan, istemcilerin işlerini gerçekleştirmek için gerek duydukları bütün bilgilerin, istemciler gibi aynı LAN'a bağlanmış bulunan sunucular üzerinde bulunmasıdır. Ancak kurumlar büyüdükçe merkez ve çevre ofislere bölünmekte hatta mobil hale gelmektedir.

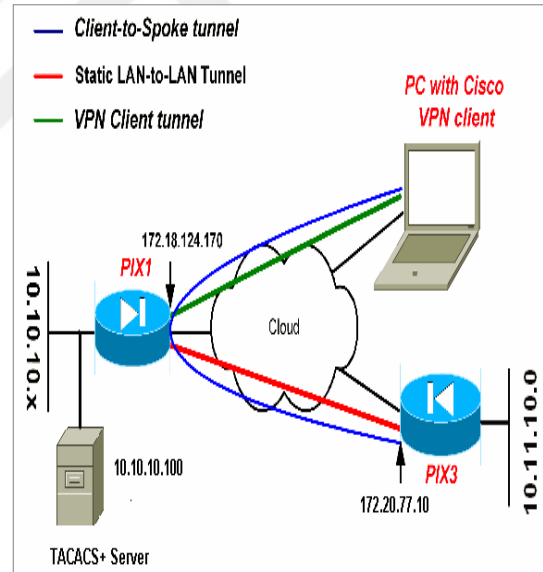
VPN uzakta yer alan bir ağdan yerel ağa erişerek bilgisayarın yerel ağdaymış gibi çalışabilmesini sağlayan bir alt yapıdır. Sanal bir ağ oluşturarak uzaktan bağlanan makine konuk gibi değil, ağa fiziksel olarak bağlanmış gibi görünür [9].

VPN platforma bağılı olmayan bir yapıdır ve çeşitli yazılımsal veya donanımsal çözümler kullanılarak farklı mimariler ile oluşturulur. Genel kullanıma açık olan ağ yapısını kullanıyor olması bir takım güvenlik endişeleri doğurabilir. Bu durum göz önünde bulundurularak bağlandığı nokta ile bilgisayar arasında adanmış ve şifreli bir tünel bağlantısı sayesinde bilgi akışı güvenliği arttırılmıştır [10].

Yalnız VPN sadece bilgi gidip gelmesi sırasındaki güvenliği kapsadığından karşılıklı yapıların güvenlik duvarları tarafından korunması gerekmektedir. Sizin bölgeniz çok güvenli olabilir ama size bağlanan evdeki bilgisayarın güvenliği de önemlidir. Çünkü dışarıdan içeri normal şartlarda sızamayan kişiler zincirin en zayıf halkasından içeri sızabilirler [11]. Ayrıca VPN sayesinde büyük şirketlerin kendi aralarında internetten faydalanmaksızın, özel ağ kurmaları çoğu yerde gereksizdir [12]. Aşağıdaki Şekil 2.1 ve Şekil 2.2’den VPN erişimi ve çalışma şekilleri görülmektedir:



Şekil 2.1. VPN erişimi [13].

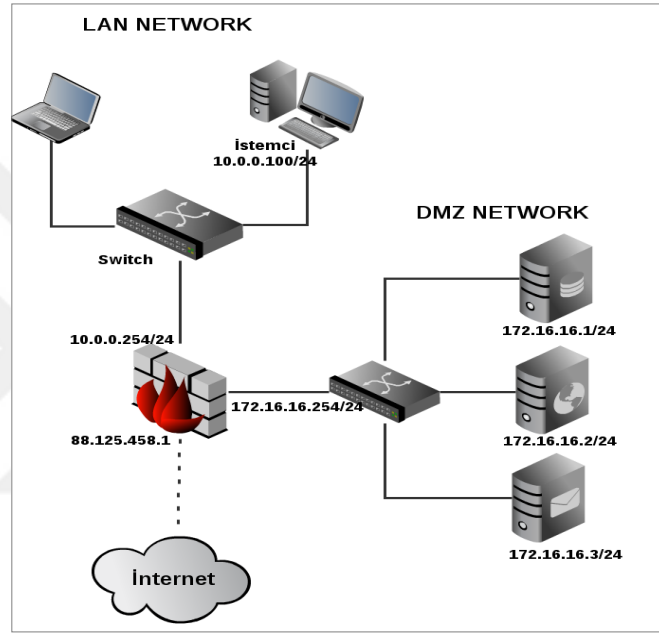


Şekil 2.2. VPN'nin çalışma şekli [14].

2.4. GÜVENLİK DUVARI (FIREWALL)

“Güvenlik duvarı” ya da “firewall” yazılımları tehlikeli ya da istenmeyen trafiğin bir bilgisayar ağına girmesini ya da çıkmasını önleyen denetim yazılımlarıdır. Firewall'lar, yerel ağınzla dış ağ arasındaki güvenlik kontrol yazılımları ya da

cihazlarıdır. Firewall ilk kurulduğunda bu nokta üzerindeki bütün geçişleri durdurur. Daha önceden belirlenen politikalar dahilinde hangi data paketinin geçip geçmeyeceği, hangi geçişlerde parola doğrulaması yapılacağı gibi bilgiler firewall kural tablolarına eklenir. Bu sayede sisteme ulaşan kişi ve bilgi trafiği kontrol altına alınmış olur. Son yıllarda bu amaca yönelik özel amaçlı donanımlar üretilmekte ve satılmakta ise de, bu işlev Linux yönlendiricilerin doğal bir özelliğidir [15]. Şekil 2.3’de bir ağ geçidinin nasıl konumlandırıldığı görülmektedir:



Şekil 2.3. Firewall konumlandırılması [16].

Günümüz firewall yazılımları basit bir paket filtreleme yazılımı değildir. Bir paketin ne yapılacağına karar vermeden önce o paketin ait olduğu TCP/IP seansına ilişkin eski paketlerin de dikkate alınmasına olanak sağlarlar. “Bağlantı izleme” (connection tracking) adı verilen bu özellik sayesinde basit paket filtresi olarak eski güvenlik duvarı yazılımlarına (ipchains gibi) göre çok büyük üstünlük sağlamaktadır.

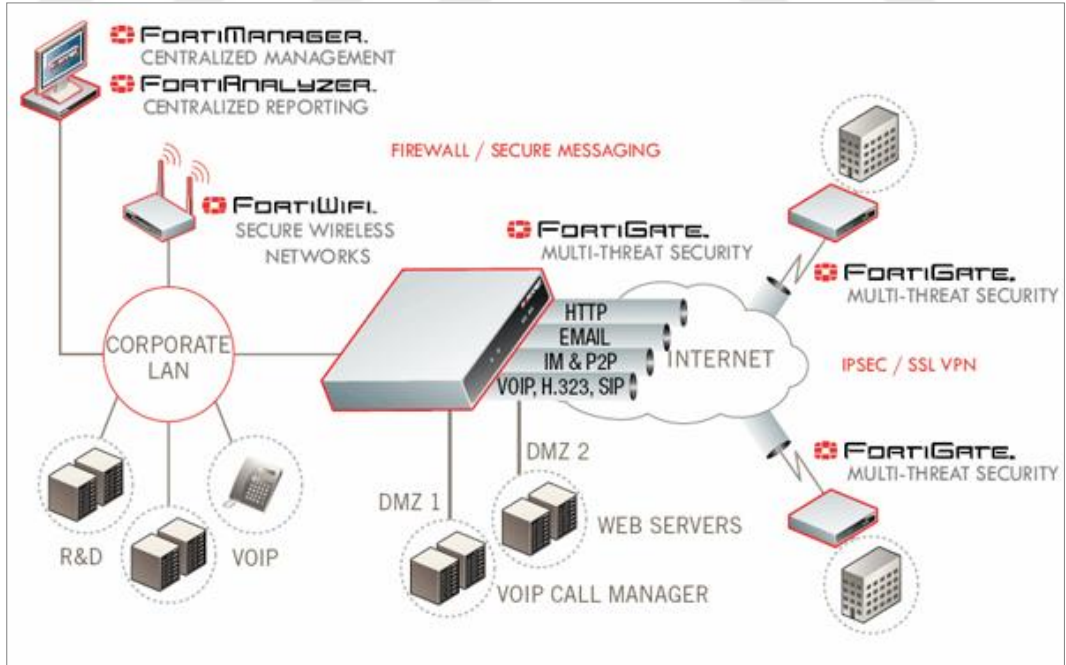
Gelen ve giden paketlerin ne yapılacağına karar verirken çıkış ve varış IP adresleri, port adresleri yanı sıra paketlerin varış/çıkış sıklıkları, içerikleri, hatta günün tarih ve saati bile dikkate alınabilmektedir.

Örneğin deneme yanılma ile şifre kırma çabasının işareti olan, bir IP adresinden beş saniye içinde üçten fazla ftp ya da ssh bağlantı isteği geldiğinde, bu paketleri reddetmek ya da daha iyisi, görmezlikten gelmek olasıdır.

Gene günümüzün sıkça kullanılan saldırı yöntemlerinden biri, http, smtp gibi servisleri aşırı yükleyerek sunucuları yanıt veremez hale getirmeye yönelik DoS (Denial of Service) saldırılarıdır.

Bu saldırılarda, sunucu yazılımının yanıt veremeyeceği sıklıkta bağlantı isteği (SYN) gönderilir. Firewall yazılımlarıyla gelen SYN paketlerinin sıklığını denetlemek, aşırı yük oluşturacak sıklıktaki paketleri durdurmak olasıdır [15].

UTM (Unified Threat Manager - Bütünleşik Tehdit Yöneticisi) Firewall: İçerik filtreleme, spam e-posta filtreleme, saldırı tespit sistemi, casus yazılım engelleme ve ağ geçidinde anti virüs görevlerini de yürüten gelişmiş cihazlardır [17]. Aşağıdaki Şekil 2.4'te UTM firewall özellikleri görülmektedir:



Şekil 2.4. UTM firewall özellikleri [18].

BÖLÜM 3

SALDIRI TESPİT SİSTEMLERİ

İnternet trafik akışlarının nasıl davranış gösterdiği halen üzerinde çalışmaların devam ettiği konulardan biridir. Bu çalışmalar “İnternet Trafik Tanımlama” olarak isimlendirilir. Trafik tanımlamanın yapılmasındaki temel amaç kapasite ihtiyaçlarını belirleyebilmek ve altyapı sorunlarını tespit edebilmektir.

Ağ trafiğini doğrudan etkileyen diğer bir faktör ise ağ üzerinde oluşan düzensizliklerdir. Ağ sistemlerine erişim, altyapı sorunlarının dışında ağı kötüye kullanma ya da ağ üzerinde düzensizlik yaratılarak kesintiye uğratılabilir. Bunların önüne geçmek için ağ trafiğini izlemek ve analiz etmek önem kazanmaktadır. Hızlı ve verimli çalışan analiz sistemlerinin varlığı, bu sorunların etkin bir şekilde ağa zarar vermesini engelleyebilmektedir. Bu şekilde çalışan sistemlere saldırı tespit sistemleri adı verilmektedir [19].

En genel anlamıyla, saldırı tespiti işini yapmak için geliştirilen sistemlere “saldırı tespit sistemleri”(STS) denir. Ancak günümüze kadar yapılan araştırmalar ve çalışmalar incelendiğinde, saldırı tespit sistemlerinin farklı tanımları olduğu görülmüştür. Yapılan bu tanımlara göre STS’ler; Saldırıyı durdurma girişiminde bulunmayan ve olası güvenlik ihlali durumlarında, sistem güvenlik çalışanlarına uyarı mesajı (alarm) veren sistemlerdir.

Bir STS, olası güvenlik açıklarını belirleyebilmek için bilgisayar veya ağ içerisinde değişik alanlardan bilgileri toplar ve analiz eder. Güvenlik duvarının statik izleme kabiliyetini tamamlayan dinamik izleme elemanıdır.

STS'lerin geliştirilmesinde günümüze kadar istatistiksel yöntemlerin dışında,

- kural tabanlı (rule based),
- eşik değeri belirleme (threshold value),
- durum geçiş diyagramları (state transition diagrams),
- yapay sinir ağları (artificial neural networks),
- veri madenciliği (data mining),
- yapay bağışıklık sistemi (artificial immune system),
- uzman sistemler (expert systems),
- örüntü eşleme (pattern recognition),
- bulanık mantık (fuzzy logic)

gibi farklı birçok teknik uygulanmıştır.

Günümüze kadar STS'lerde birçok farklı teknik kullanılmıştır. Bu teknikler, elde edilen verilerin modellenmesi, sınıflandırılması veya kural tablolarının oluşturulması için geliştirilmiştir. Kullanılan tekniklerden elde edilen veriler sayesinde, saldırı tespit yaklaşımlarının uygulanması için gerekli olan platform oluşturulmuştur. Bu tekniklerden en çok kullanılanları, veri madenciliği, kural tabanlı sistemler, açıklayıcı istatistikler, eşik değeri tespiti, durum geçiş analizi, uzman sistemler, örüntü eşleme olmuştur.

3.1. SALDIRI TESPİT YÖNTEMİ

STS'lerde, saldırı tespit yöntemi olarak anormallik tespiti ve kötüye kullanım tespiti olmak üzere iki farklı yaklaşım kullanılır. Anormallik tespitine dayanan yaklaşım, sistemdeki kullanıcı davranışlarını modellerken, kötüye kullanım (imza) tespitine dayanan yaklaşım, saldırganların davranışlarını modeller.

Bir saldırının hangi adresten veya hangi porttan geldiğini bilmeden engel olmak mümkün değildir. STS'ler saldırıları tespit ederken bu bilgileri de elde ederler. STS'ler, detaylı olarak topladığı ve depoladığı bilgilerden yararlanarak, saldırıları olabildiğince erken tespit etme özelliğine sahiptir. Saldırıları tespit etme yöntemlerine göre ikiye ayrıldığını bildiğimiz STS'lerin, hangi yöntemin seçileceğine dair dikkat edilmesi gereken noktası, anormallik tespiti yönteminin,

bütün kötü davranışları tespit etmeye çalışırken, kötüye kullanım tespiti yönteminin kötü olarak bilinen davranışları tanımaya çalışmak olduğudur. Her iki yöntemin de avantaj ve dezavantajları olduğu göz önünde bulundurularak, tasarımlarda avantajları bir araya toplayan hibrit yaklaşımlardan faydalanmanın daha gerçekçi olacağı değerlendirilmektedir [20].

Ağ trafiğinde oluşan düzensizliği inceleyerek, saldırı tespit etmek için değişik çalışmalar yapılmıştır. Bu çalışmaların birçoğunda istatistiki yöntemler ve trafik örnekleme (sampling) kullanılarak sonuç elde edilmeye çalışılmıştır. İstatistik yöntemler genelde “En Çok ve Baz Değer” ve “İz Eşleştirme” olarak iki başlık altında toplanmıştır;

3.1.1. En Çok ve Baz Değer

En çok yönteminde, meydana gelen düzensizlikleri tespit ederken iki ayrı yöntem izlenebilir. En çok oturum yönteminde; belirli bir zaman dilimindeki oturum sayısı (session) temel alınır. En çok veri yönteminde ise; belirli bir zaman diliminde aktarılan veri miktarı (byte) temel alınır.

Baz değer, geçmişte kaydedilen ağ trafiğine göre oluşturulmuş şablondur. Bu şablon normal ağ trafiğini temsil eder. Ağ trafiğinde meydana gelen ve baz değerden farklı olan her türlü trafik düzensiz olarak kabul edilir [21].

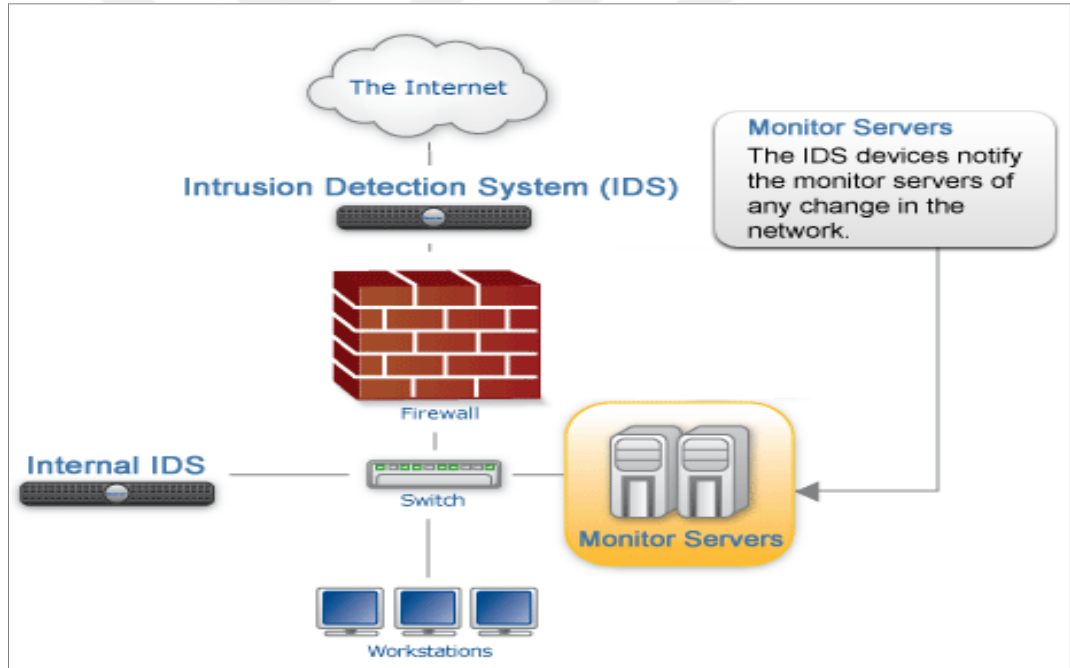
3.1.2. İz Eşleştirme

İz eşleştirme, normal olmayan trafiğin tespit edilmesinde kullanılan bir yöntemdir. Bu yöntemde motivasyon ağda bilinen servis ve IP bloklarının akış verileri ile karşılaştırılmasına dayanır. Mevcut servislere ait port ve IP bilgileri, oluşturulan bir veritabanı yardımıyla akış verileri ile karşılaştırılır. Veri tabanında bulunmayan port veya IP adreslerine gelen veya giden trafik şüpheli olarak değerlendirilebilir.

Örneğin sadece 80 portundan hizmet veren bir sunucunun 3306 portuna bir istek gönderildiğinde uyarı sistemi devreye girip şüpheli durumu bildirebilir. Bu yöntem ağ tarama tespitinde kolayca kullanılabilir [19].

STS sistemleri için, ticari çözümler kullanabileceği gibi Snort gibi açık kaynak koduna sahip çözümler de kullanılabilir. Snort için www.bleedingsnort.com adresinde bulunan “bleeding-malware.rules” dosyasındaki güncel zararlı yazılım imzaları kullanılmalı ve sistem sorumluları gözlemledikleri yeni saldırılara ait imzaları da kendileri eklemelidir [7].

Aşağıdaki Şekil 3.1’de saldırı tespit sistemlerinin ağ üzerinde konumlandırılması görülmektedir:



Şekil 3.1. Saldırı tespit sistemi konumlandırılması [22].

BÖLÜM 4

TEMEL İSTATİSTİK KAVRAMLARI

4.1. DEĞİŞKENLER

Varlıkların çeşitli değerler alabilen özelliklerine değişken denir. Değişkenler, nitel ve nicel olmak üzere genel olarak ikiye ayrılır (Şekil 4.1 değişken tiplerini şematik olarak göstermektedir) :

4.1.1. Nitel (Kategorik, Qualitative, Categorical) Değişkenler

Gözlem sonuçları sayılarla ifade edilemeyen, sayılar kullanılsa bile, bunların sembol olmaktan başka fonksiyonu olmayan değişkenlerdir. (Sınıflandırma yolu ile ölçümlenebilen değişkenlerdir. Sayısal değildirler.)

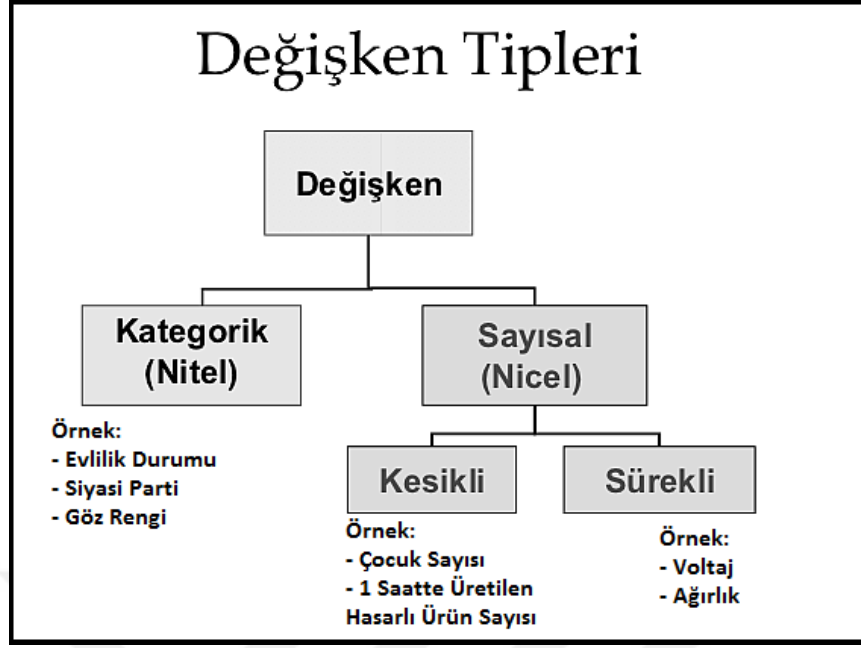
Yabancı kaynaklarda “Dummy Variable (yapay, kukla-gölge)” veya “Binary Data” adları ile rastlanılan kalitatif değişkenlere, dummy değişken de denilmektedir, çoğunlukla sayısal olarak önemi olmayan 0 veya 1 değerini alırlar.

4.1.2. Sayısal (Nicel) Değişkenler

Gözlem sonuçları sayılarla ifade edilebilen değişkenlerdir. Nicel değişkenler de kesikli ve sürekli değişken olarak iki farklı ayrıma tabi tutulur:

Kesikli (Süreksiz) değişken: Tanımlanmış olduğu aralıkta belirli değerleri alabilen değişkenlerdir (Değişkenin alabileceği değerlerin sayısı sonlu ve sayılabilir).

Kesiksiz (Sürekli) değişken: Tanımlanmış olduğu aralıkta her değeri alabilen değişkenlerdir (sonsuz sayıda değer alabilir) [23,24].



Şekil 4.1. Değişken tiplerinin şematik gösterimi [23].

4.2. DOĞRULUK VE DUYARLILIK

Doğruluk (Accuracy); analiz sonucunun gerçek değere ne kadar yakın olduğunun; Duyarlılık (Precision) ise yapılan ölçümlerin birbirine ne ölçüde yakın olduğunun ifadesidir.

Bir analizde ölçümlerin duyarlılığı gerçeğe en yakın bir şekilde her zaman hesaplanabilir. Ancak gerçek değer tam olarak bilinmezse sonucun doğruluğu ancak tahmin edilebilir.

Bir ölçümde, duyarlılığın çok iyi olması onun doğruluğunun da çok iyi olduğunu göstermez, ancak gerçeğe yakınlığı konusunda bir fikir verebilir. Örneğin, içinde tam % 20 demir bulunan bir karışımın analizi sonunda bulunan yüzde demir miktarı sonucun doğruluk derecesi hakkında bir fikir verir. Öte yandan içindeki gerçek demir miktarı bilinmeyen bir filizin analizi sonunda bulunan yüzde demir, o değer in doğruluğu konusunda bir şey söylemez.

Bu sonucun elde edilmesi sırasındaki ölçümlerin birbirine yakınlığı yani ölçümün duyarlılığı ise bulunan sonucun doğruluğu konusunda bir fikir verir.

Doğruluk, ölçümlerin aritmetik ortalamasının gerçek değere yakınlığı olarak tanımlanır ve hata olarak ifade edilir.

Duyarlılık ise ölçümlerin birbirine yakınlığı olarak tanımlanır ve sapma olarak ifade edilir. Duyarlılık, aynı değeri yeniden elde etme becerisi olarak da tanımlanır.

İyi bir duyarlılık çoğunlukla iyi bir doğruluk derecesinin göstergesidir. Ancak bu her zaman geçerli değildir. Çok iyi bir duyarlılığı olduğu halde zayıf doğruluk dereceli ölçümler de olabilir. Şekil 4.2’de doğruluk ve duyarlılık farkı gösterilmektedir.

Bir analizin sonucu yazılırken yapılan her ölçümün ayrı ayrı hesaplanması gerekmez. Bunun yerine bunların ortalamasının yazılması yeterlidir. Ancak bu ortalamanın bir anlam ifade edebilmesi için ölçüm aralığının, standart sapmasının ve kaç ölçümün yapıldığının belirtilmesi gerekir [25].



Şekil 4.2. Doğruluk ve duyarlılık farkı [25].

BÖLÜM 5

REGRESYON

5.1. REGRESYON NEDİR

Regresyon, kelime olarak bir şeyin kaynağına dönme ya da geriye gitme anlamlarına gelmektedir.

Matematiksel terim olarak Regresyon Analizi, iki ya da daha çok değişken arasındaki ilişkiyi ölçmek için kullanılan analiz metodudur. Eğer tek bir değişken kullanılarak analiz yapılıyorsa buna tek değişkenli regresyon, birden çok değişken kullanılıyorsa çok değişkenli regresyon analizi olarak isimlendirilir.

Regresyon analizi ile değişkenler arasındaki ilişkinin varlığı, eğer ilişki var ise bunun gücü hakkında bilgi edinilebilir. Regresyon terimi için öz Türkçe olarak bağlanım sözcüğü kullanılması teklif edilmiş [26] ise de Türk ekonometriciler arasında bu kullanım yaygın değildir.

Değişkenler arasındaki ilişkiye örnek olarak, mevsimlere göre yağın yağmur miktarı ile ürün çeşitliliği arasındaki ilişkiyi; eğitim düzeyi ile sahip olunan çocuk sayısı ve boşanma miktarı arasındaki ilişkiyi; öğrencilerin not ortalaması ve devamsızlık durumu ile üniversite sınavında aldığı puan arasındaki ilişkiyi verebiliriz.

Regresyon genelde nicel iki veya daha çok sayıda değişken arasındaki doğrusal ilişkinin fonksiyonunu ya da biri bağımlı diğerleri bağımsız değişkenler arasındaki doğru denklemini göstermekle birlikte, değişkenlerden biri bilindiğinde diğeri hakkında tahmin yapılmasını da sağlar.

Regresyonda, deęişkenlerden biri baęımlı dięerleri baęımsız deęişken olmalıdır. Kurulan eşitlikte dięer deęişkenlerden etkilenen baęımlı deęişken eşitlięin solunda yer alır. Baęımlı deęişkenin haricindekilere baęımsız deęişken denir ve eşitlięin saęında yer alan bu deęişkenler dięer deęişkenlerden etkilenmezler.

5.2. REGRESYON ANALİZİNİN TARİHİ

Regresyon yönteminin ilk şekli en küçük kareler prensibidir ve ilk olarak Adrien Marie Legendre tarafından 1805'te ortaya atılmıştır. Hemen takiben 1809'da C.F. Gauss aynı yöntemi açıklamıştır. En küçük kareler terimi Legendre tarafından “moindres carrés” olarak kullanılmış, ancak Gauss aynı yöntemi 1795'ten beri kullandığını iddia etmiştir. Legendre ve Gauss bu yöntemi astronomik gözlemlerden uydularının güneş etrafındaki yörüngelerini tespit etmek için kullanırken ortaya çıkartmışlardır. 1748'de Eüler'in aynı problem üzerinde uğraştığı, fakat başarı sağlayamadığı bilinmektedir. En küçük kareler kuramında sonraki gelişme Gauss'ın 1821'de yayınladığı bir makalede ortaya çıkartılmış ve bu yayında Gauss sonradan kendi adı verilen Gauss-Markov teoreminin bir şeklini açıklamıştır [27-29].

Regresyon kelimesini bilimsel bir terim olarak ilk kullanan Francis Galton'dur. İngiliz istatistikçi Galton bu terimi, ana konusu kalıtım olan bir biyolojik inceleme için kullanmıştır. İncelemesinde bir ailede çocuklarla anne-babalarının boyları arasındaki ilişkiyi araştırmıştır. Sonuçta çocukların boylarının bir nesil içinde atalarının boylarına geri döndüklerini gözlemlemiştir. Boy ortalamalarındaki bu geri dönüş için Galton ilk olarak İngilizce “reversion” terimini kullanmış, fakat daha sonra reversion ile eş anlamlı olan “regression” terimini kullanmayı tercih etmiştir. Galton bu çalışmaları yaparken istatistiksel “regresyon” yöntemini de geliştirmiştir. Bu yöntemi daha geniş genel istatistiksel alanlarda uygulayıp geliştiren ise Karl Pearson ve Udny Yüle olmuştur. Udny Yüle ve Karl Pearson çalışmalarında baęımlı ve baęımsız deęişkenlerin normal dağılım gösterdiğini varsaymıştır. R. A. Fisher ise 1922 ve 1925'deki yayınlarında bu kısıtlayıcı varsayımı biraz daha genişletmiştir. Fisher bu sayede regresyonu sadece baęımlı deęişkenin koşullu dağılımının normal olduęu durumlarda uygulanabileceğini göstermiştir [30-35].

İstatistikçiler bu metotları ve kavramları, kalıtım ya da “ortalamaya geri dönüş” konularıyla hiçbir ilgisi bulunmayan konularda kullanırken de regresyon terimini kullanmaya devam ettiklerinden günümüzde bu terimin Galton’un konusu ile hiçbir ilgisi kalmamış ve bu terim artık eğri uydurma ve doğrusal bağlantı bulunması ile aynı anlamlarda kullanılmaktadır.

5.3. REGRESYON ANALİZİ

Regresyon analizi, aralarında sebep-sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi belirlemek ve bu ilişkiyi kullanarak o konu ile ilgili tahminler (estimation) ya da kestirimler (prediction) yapabilmek amacıyla yapılır. Doğada birçok olayda sebep-sonuç ilişkisine rastlamak mümkündür [36].

Sebe-sonuç ilişkisine örnek olarak: Gelir-Harcama, Yaş-Boy, Gübre-Verim, Yem Miktarı-Süt Miktarı, Çalışma Süresi-Alınan Not.

Gerçek yaşamın çeşitli alanlarında herhangi bir uygulama ile toplanan veriler tablo şekline getirilerek incelenir ve toplanan veriyi modelleyen bir fonksiyon bulunmaya çalışılır. Çoğu zaman bu veri tablosuna tam olarak uyan bir fonksiyon bulmak mümkün olmaz; veri tablosuna en iyi uyan fonksiyon belirlenmeye çalışılır. Bir veri tablosuna en iyi uyan fonksiyonu bulma sürecine regresyon analizi denir [37].

Regresyon analizi ile;

- Bağımlı ve bağımsız değişkenler arasında bir ilişki var mıdır?
- Eğer bir ilişki varsa bu ilişkinin gücü nedir?
- Değişkenler arasında ne tür bir ilişki vardır?
- Bağımlı değişkene ait ileriye dönük değerleri tahmin etmek mümkün müdür ve nasıl tahmin edilmelidir?
- Belirli koşulların kontrol edilmesi durumunda özel bir değişken veya değişkenler grubunun diğer değişken veya değişkenler üzerindeki etkisi nedir ve nasıl değişir?

gibi sorulara cevap aranmaya çalışılır [38].

Bu analiz tekniğinde iki (basit regresyon) veya daha fazla değişken (çoklu regresyon) arasındaki ilişkiyi açıklamak için matematiksel bir model kullanılır ve bu model regresyon modeli olarak adlandırılır [36].

Regresyon analizi yaparken en çok kullanılan yöntemlerden biri en küçük kareler yöntemidir.

Basit doğrusal regresyon analizinde bir bağımlı ve bir bağımsız değişken söz konusu iken, çoklu doğrusal regresyon analizinde ise bir bağımlı değişken varken iki ya da daha fazla bağımsız değişken vardır ve her iki analizde de değişkenler arasında doğrusal bir ilişki vardır. Ayrıca bağımlı ve bağımsız iki değişken arasında eğrisel bir ilişki var ise değişkenler arasındaki ilişki eğrisel regresyon modeli ile açıklanır.

Korelasyon analizinde, değişkenlerin bağımlı ve bağımsız değişken olarak belirlenmesi hesaplamaların sonucu açısından önemli değilken, regresyon analizinde ise değişkenlerin hangisinin bağımlı hangisinin bağımsız değişken olduğunu tespit etmek çok önemlidir [36].

İki ya da daha çok sayıda, değişken arasındaki ilişkiye “korelasyon” denir. Korelasyon ilişkisinde ilişkinin miktarı bir sayı ile belirtilir. Bu sayıya "korelasyon katsayısı" denir [39].

5.4. EN KÜÇÜK KARELER METODU

Regresyon analizi yaparken en çok kullanılan yöntemlerden biri “En Küçük Kareler” yöntemidir. Büyük matematikçi C. F. Gauss’un 18 yaşındayken (1795) geliştirdiği bu yöntem, ilk kez 1801 de Ceres asteroidinin yörüngesinin belirlenmesinde kullanılmış ve ilk kez Gauss’un toplu eserlerinin yayınlandığı ciltlerden ikincisinde 1809 yılında yayınlanmıştır.

Fransız matematikçi A. Legendre 1805 ve Amerikalı matematikçi R. Adrain de 1808 yıllarında aynı yöntemi Gauss’dan habersiz ve bağımsız olarak keşfetmişlerdir.

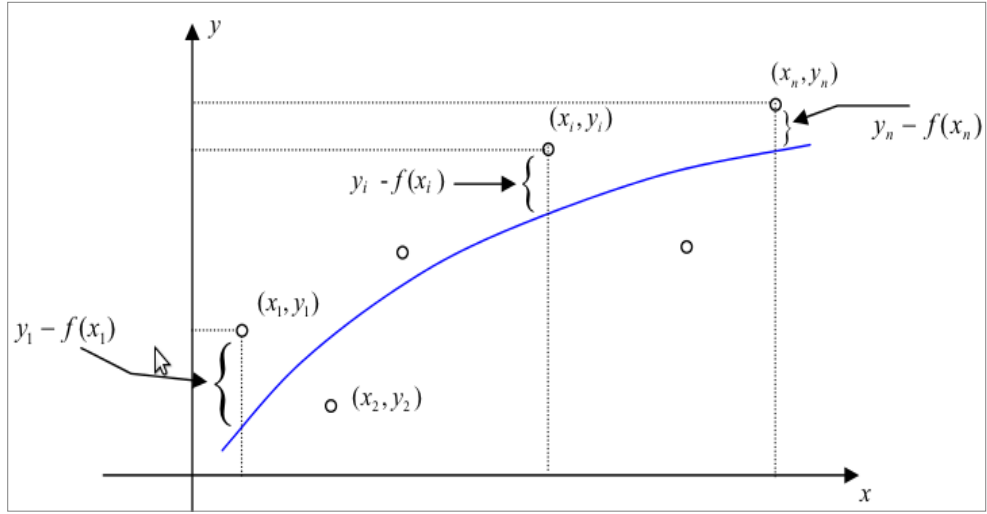
En küçük kareler yöntemi, tıp, finans, mühendislik, ziraat, biyoloji ve sosyoloji gibi çeşitli bilim dallarında çeşitli değişkenler arasındaki ilişkiler belirlenirken kullanılan en önemli araçlar arasındadır.

Belli ölçümler sonucunda $i = 1, 2, \dots, n$ için (x_i, y_i) verileri elde edilmiş olsun. Burada, her bir y_i nin x_i ye bağlı olarak değiştiği varsayılmaktadır. Düzlemde noktalar olarak düşünüldüğünde, pratikte bu noktalar düzgün bir eğri üzerinde, başka bir deyimle, bilinen bir fonksiyonun grafiği üzerinde bulunmazlar. Hatta bazı durumlarda, (x_i, y_i) ler arasında ne tür bir bağıntı bulunduğu dahi bilinmeyebilir.

Ancak, yapılan ölçümlerin doğası gereği, her $i = 1, 2, \dots, n$ için $y_i = f(x_i)$ olacak biçimde bir fonksiyonun var olduğu, ölçümlerde yapılan hata nedeniyle bu eşitliklerin bazıları veya hepsinin sağlanmadığı kabul edilebilir. Bu düşünceyle, ölçülen y_i değeri $f(x_i)$ için yaklaşık değer kabul edilerek bu yaklaşımdaki hatanın minimum olduğu f fonksiyonu belirlenmeye çalışılır.

Bu amacı gerçekleştirmek için f fonksiyonunun bir takım parametrelere bağlı bir ifadesi bulunduğu varsayıp eldeki veriler yardımıyla bu parametreler belirlenmeye çalışılır. Örneğin, f fonksiyonu “ $y = f(x) = mx + b$ ” ifadesinde olduğu gibi bir doğrusal fonksiyon veya “ $y = f(x) = ax^2 + bx + c$ ” ifadesinde olduğu gibi bir karesel fonksiyon olabilir ki bu durumda belirlenmesi gereken parametreler a, b, c ve m ’dir.

y_i değeri $f(x_i)$ için yaklaşık değer, $f(x_i) \approx y_i$ kabul edilince yapılan hata “ $y_i - f(x_i)$ ”dır ve amaç, bu hatalar minimum olacak şekilde bir f fonksiyonu bulmaktır (aşağıdaki Şekil 5.1’den görülmektedir).



Şekil 5.1. En küçük kareler yöntemi.

Tanım1: $y_i - f(x_i)$ farklarından her birine bir artık denir.

Tanım2: Bir veri tablosuna en iyi uyan doğrusal fonksiyonun grafiği olan doğruya regresyon doğrusu veya en küçük kareler doğrusu denir.

En küçük kareler yönteminde aranan fonksiyon, ya da onun parametreleri, tüm artıkların kareleri toplamı olan:

$$\sum_{i=1}^n (y_i - f(x_i))^2 = (y_1 - f(x_1))^2 + \dots + (y_n - f(x_n))^2 \quad (5.1.1)$$

ifadesini minimum yapacak şekilde belirlenir. Bu, yönteme neden en küçük kareler yöntemi dendiğini açıklar. Sözü edilen kareler toplamının minimum olması için her bir hatanın küçük olması gerektiğine dikkat edilmelidir [37].

Formül biraz daha açıklanacak olursa:

Gerçek değer (y) ile teorik değer (y_t) arasındaki fark Δy ile gösterilirse,

$$\Delta y = y - y_t \quad (5.1.2)$$

olur.

N tane deneysel nokta için:

$$\sum_{i=1}^N (\Delta y)^2 = \sum_{i=1}^N (y - y_t)^2 \rightarrow \text{Minimum} \quad (5.2)$$

olmalıdır. Burada:

$$y_t = a + bx \quad (5.3)$$

dir. Eşitlik-5.2 ile verilen ifadenin minimum olma şartını sağlayan a ve b sabitleri, matematiksel kural gereğince aynı eşitliğin a ve b'ye göre türevlerinin sıfıra eşitlenmesi ile bulunur. Yani;

$$\frac{\partial}{\partial a} \left(\sum_{i=1}^N (y - y_t) \right)^2 = 0 \quad (5.4)$$

$$\frac{\partial}{\partial b} \left(\sum_{i=1}^N (y - y_t) \right)^2 = 0 \quad (5.5)$$

$$\frac{\partial}{\partial a} \left(\sum_{i=1}^N [y - (a + bx)] \right)^2 = 2Na - 2 \sum_{i=1}^N y + 2b \sum_{i=1}^N x = 0 \quad (5.6)$$

$$\frac{\partial}{\partial b} \left(\sum_{i=1}^N [y - (a + bx)] \right)^2 = 2b \sum_{i=1}^N x^2 - 2 \sum_{i=1}^N xy + 2a \sum_{i=1}^N x = 0 \quad (5.7)$$

Eşitlik-5.6 ve 5.7 den,

$$\sum_{i=1}^N y = Na + b \sum_{i=1}^N x \quad (5.8)$$

$$\sum_{i=1}^N xy = a \sum_{i=1}^N x + b \sum_{i=1}^N x^2 \quad (5.9)$$

bulunur.

Bu eşitliklerden a ve b çekilirse,

$$b = \frac{N \sum_{i=1}^N xy - \sum_{i=1}^N x \sum_{i=1}^N y}{N \sum_{i=1}^N x^2 - \left(\sum_{i=1}^N x \right)^2} \quad (5.10)$$

$$a = \frac{1}{N} \left(\sum_{i=1}^N y - b \sum_{i=1}^N x \right) \quad (5.11)$$

olur.

En küçük kareler metodu lineer fonksiyonlara uygulandığı gibi eksponansiyel fonksiyonlara da uygulanabilir.

$$y = k \cdot e^{mx} \quad (5.12)$$

fonksiyonu “ $\ln y = \ln k + mx$ ” şeklinde yazılabilir. Dikkat edilirse “ $\ln y = f(x)$ ” fonksiyonu eğimi m olan bir doğrudur. Eşitlik 5.10 ve 5.11’den k ve m sabitlerinin değerleri,

$$m = \frac{N \sum_{i=1}^N x \ln y - \sum_{i=1}^N x \sum_{i=1}^N \ln y}{N \sum_{i=1}^N x^2 - \left(\sum_{i=1}^N x \right)^2} \quad (5.13)$$

$$k = \exp \left[\frac{1}{N} \left(\sum_{i=1}^N \ln y - m \sum_{i=1}^N x \right) \right] \quad (5.14)$$

bulunur [40].

Örnek-1:

Bir üretici, ürettiği ürünün çeşitli üretim seviyeleri için maliyetini belirliyor ve aşağıdaki tabloyu (Çizelge 5.1) oluşturuyor, buna göre 3 ürünün maliyeti yaklaşık kaç TL'dir?

Çizelge 5.1. En küçük kareler yöntemi örnek veri seti.

Ürün sayısı (x yüz)	2	5	6	9
Maliyet (y bin YTL)	4	6	7	8

Çözüm:

Bu üretici için gider fonksiyonu yukarıdaki tabloya en iyi uyan doğrusal fonksiyon olarak belirlenir.

Çizelge 5.1'deki veri tablosu, düzlemde şu (x,y) noktalarını verir: (2,4) , (5,6) , (6,7) ve (9,8). Bu noktaların hepsini üzerinde bulunduran bir doğru yoktur. Amacımız, bu noktalara en iyi uyan doğruyu, yani regresyon doğrusunu bulmaktır. Regresyon doğrusunun denklemi:

$$y_t = a + bx$$

a ve b belirlenerek bulunacaktır.

Artıklar hesaplanır ve veri tablosu aşağıdaki gibi genişletilir:

Çizelge 5.2. En küçük kareler yöntemi örnek çözümü.

Ürün sayısı (x adet)	2	5	6	9
Maliyet (y TL)	4	6	7	8
$y_t = a + bx$	$a+2b$	$a+5b$	$a+6b$	$a+9b$
$(y_t - y)^2 = (a + bx - y)^2$	$(a + 2b - 4)^2$	$(a + 5b - 6)^2$	$(a + 6b - 7)^2$	$(a + 9b - 8)^2$
$F_a(a,b) = 0$ (a ya göre kısmi türevleri)	$2*(a + 2b - 4)$	$2*(a + 5b - 6)$	$2*(a + 6b - 7)$	$2*(a + 9b - 8)$
	$2a+4b-8$	$2a+10b-12$	$2a+12b-14$	$2a+18b-16$
	$8a + 44b - 50 = 0$ (1. Denklem)			
$F_b(a,b) = 0$ (b ye göre kısmi türevleri)	$2*(a+2b-4)*2$	$2*(a + 5b - 6)*5$	$2*(a + 6b - 7)*6$	$2*(a + 9b - 8)*9$
	$4a+8b-16$	$10a+50b-60$	$12a+72b-84$	$18a+162b-144$
	$44a + 292b - 304 = 0$ (2. Denklem)			

Yukarıdaki tablodan (Çizelge 5.2) elde edilen 1. ve 2. denklemler ortak olarak çözümlürse:

“b = 0,58” ve “a = 3,06” bulunur. Bu durumda bulunacak denklem:

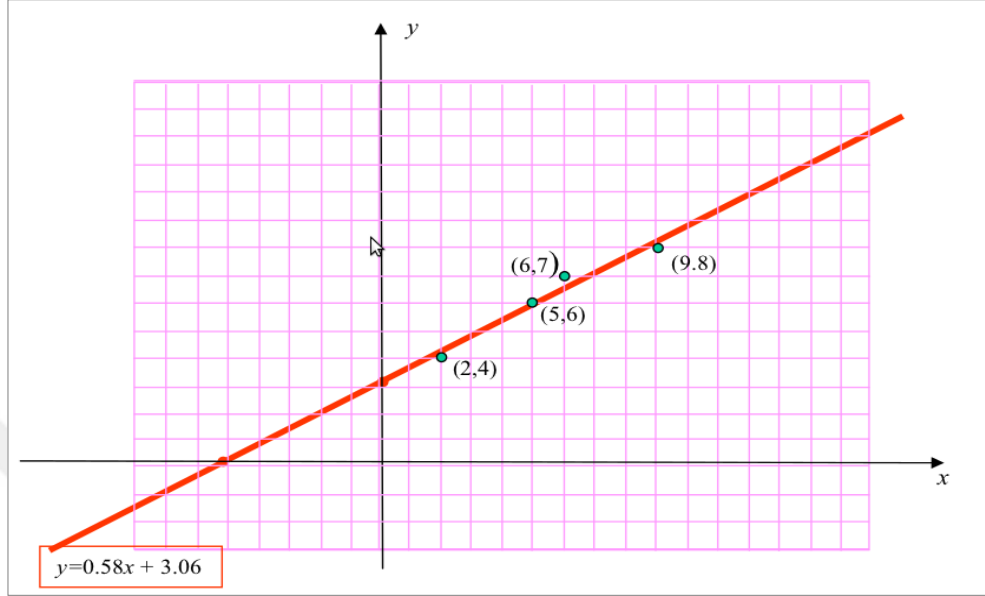
$$y_t = a + bx = 3,06 + 0,58*x$$

Soruda 3 ürün için maliyet sorulduğundan x yerine 3 yazılırsa:

$$3,06 + 1,74 = 4,8 \text{ TL}$$

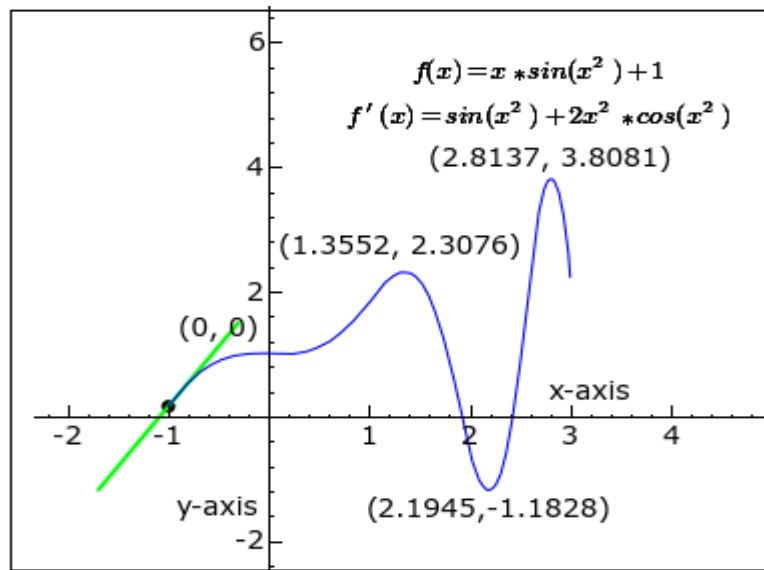
olur.

Bulunan regresyon doğrusu veri noktaları ile birlikte aşağıda Şekil 5.2'deki grafikte gösterilmiştir:



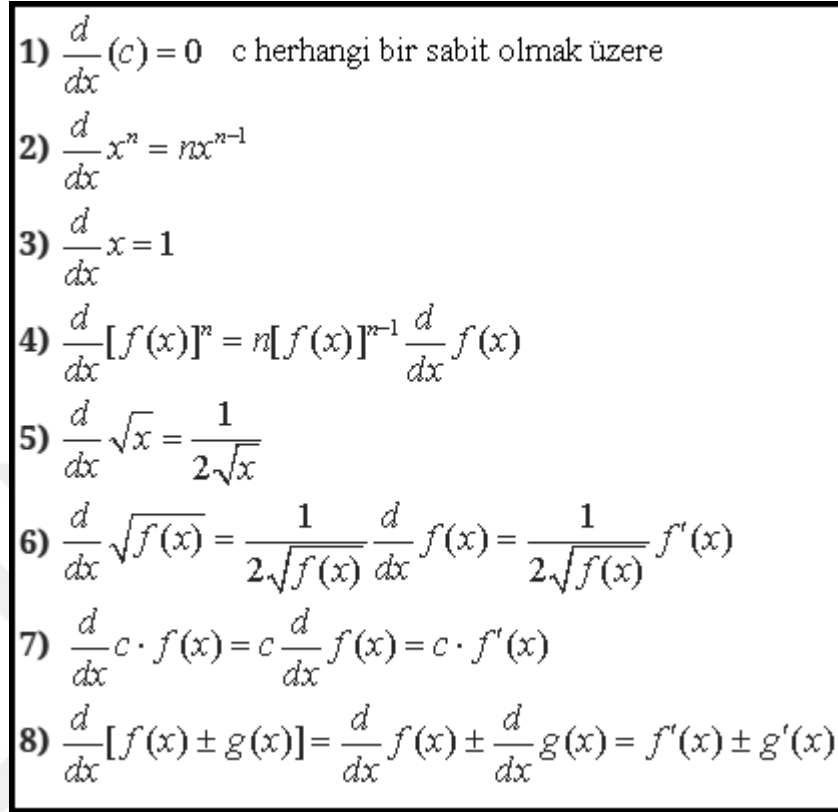
Şekil 5.2. En küçük kareler yöntemi ile bulunan grafik [41].

Konunun daha iyi anlaşılması için türevin anlamına da değinmek gerekir. Türev, herhangi bir eğrinin, herhangi bir noktasından geçen teğetinin, eğimini hesaplama işlemidir (Şekil 5.3). Yani türev ile eğim aynı şeydir (Slope: Eğim).



Şekil 5.3. Türevin anlamı [42].

Aşağıdaki Şekil 5.4’te genel türev formülleri görülmektedir:



- 1) $\frac{d}{dx}(c) = 0$ c herhangi bir sabit olmak üzere
- 2) $\frac{d}{dx}x^n = nx^{n-1}$
- 3) $\frac{d}{dx}x = 1$
- 4) $\frac{d}{dx}[f(x)]^n = n[f(x)]^{n-1} \frac{d}{dx}f(x)$
- 5) $\frac{d}{dx}\sqrt{x} = \frac{1}{2\sqrt{x}}$
- 6) $\frac{d}{dx}\sqrt{f(x)} = \frac{1}{2\sqrt{f(x)}} \frac{d}{dx}f(x) = \frac{1}{2\sqrt{f(x)}} f'(x)$
- 7) $\frac{d}{dx}c \cdot f(x) = c \frac{d}{dx}f(x) = c \cdot f'(x)$
- 8) $\frac{d}{dx}[f(x) \pm g(x)] = \frac{d}{dx}f(x) \pm \frac{d}{dx}g(x) = f'(x) \pm g'(x)$

Şekil 5.4. Genel türev formülleri.

“f(x,y)”nin “x”e göre **kısmi türevini** alırken y bir sabitmiş gibi düşünülür. Kısmi türev işareti: ∂ şeklindedir. Örneğin $f(x, y) = 3x^2y + 5x - 2y^2 + 1$ fonksiyonunun “x”e göre kısmi türevi:

$$\partial f(x, y) / \partial x = 6xy + 5$$

“y”ye göre kısmi türevi:

$$\partial f(x, y) / \partial y = 3x^2 - 4y$$

olur.

BÖLÜM 6

LOJİSTİK REGRESYON

6.1. LOJİSTİK REGRESYON ANALİZİ

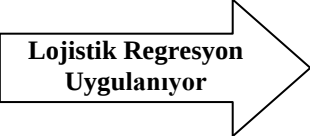
Lojistik regresyon analizinin kullanım amacı istatistikte kullanılan diğer model yapılandırma teknikleriyle aynıdır. En az değişkeni kullanarak en iyi uyuma sahip olacak şekilde sonuç değişkeni (bağımlı ya da cevap değişkeni) ile bağımsız değişkenler kümesi (açıklayıcı değişkenler) arasındaki ilişkiyi tanımlayabilen ve genel olarak kabul edilebilir modeli kurmaktır [43].

Son yıllarda yoğun bir şekilde kullanılan lojistik regresyon analizi, gözlemlerin gruplara atanmasında sık kullanılan üç yöntemden (diğerleri kümeleme analizi ve diskriminant analizi) birisidir.

Kümeleme analizinde gözlemlerin atanacağı küme sayısı tam bilinmezken, diskriminant ve lojistik regresyon analizinde grup sayısı bilinmekte, mevcut veriler kullanılarak bir ayırımsama modeli elde edilmekte ve kurulan bu model yardımıyla veri kümesine eklenen yeni gözlemlerin gruplara atanması mümkün olmaktadır [43].

Lojistik regresyon analizi, son dönemlerde özellikle sosyal bilimler alanında kullanımı yaygınlaşan bir yöntemdir. Neden sonuç ilişkilerinin ortaya konulması amacıyla yapılan çoğu sosyo-ekonomik araştırmada, incelenen değişkenlerden bazıları olumlu-olumsuz, başarılı-başarısız, evet-hayır, memnun - memnun değil şeklinde iki düzeyli verilerden oluşmaktadır. Bu türden bağımlı değişkenin iki düzeyli ya da çok düzeyli kategorik verilerden oluşması durumunda; bağımlı değişken ile bağımsız değişken(ler) arasındaki neden-sonuç ilişkisinin incelenmesinde, Lojistik Regresyon Analizi önemli bir yere sahiptir [44,45].

Amaçlarından birisi sınıflandırma, diğeri ise bağımlı ve bağımsız değişkenler arasındaki ilişkileri araştırmak olan lojistik regresyon analizinde, bağımlı değişkeni kategorik veri oluşturmakta ve kesikli değerler almaktadır. Bağımsız değişkenlerin ise hepsinin veya bazılarının sürekli ya da kategorik değişkenler olmasına ilişkin bir zorunluluk bulunmamaktadır [46]. Aşağıdaki Şekil 6.1’de lojistik regresyonun kullanım örneği görülmektedir.

İsim	Yaş	Eğitim	Gelir	Müşteri Mi?	Lojistik Regresyon Uygulanıyor 	İsim	Müşteri Olma İhtimali
Ali	34	8	3	E		Can	0,67
Veli	23	11	2	E	Bir banka kendisine müşteri olma ihtimali yüksek olan kişileri belirleyerek reklam çalışmalarını ona göre yapabilir.	Kaan	0,78
Ela	56	5	6	H		Helin	0,34
Gül	54	20	4	E		Arda	0,97
...	...	...	...	...		...	...

Şekil 6.1. Lojistik regresyon analizinin kullanım amacına örnek.

Bazı nicel değişkenler de regresyon modellerinde nitel olarak kullanılabilir. Örneğin işletme büyüklüğü; işçi sayısı, üretim değeri, katma değer gibi ölçütlerle sayısal olarak ifade edilebildiği gibi; küçük, orta ve büyük ölçekli işletmeler olmak üzere üç şıklı bir nitel değişken olarak da modellere dahil edilebilir [47].

Gölge(nitel) değişkenlerin regresyon modelinde nicel olarak kullanıldığı modellere varyans çözümlemesi(varçöz) modeli denilmektedir [48].

Regresyon çözümlemesinde açıklayıcı (bağımsız) değişkenler gibi, açıklanan (bağımlı) değişkenler de iki değerli (niteliksel) olabilmektedir. Örneğin bir ailenin ev ya da otomobil sahibi olup olmadığına, gelir düzeyine, ailedeki kişi sayısına... bağlı olabilir. Burada bağımlı değişken, ev ya da otomobil sahipliği olup tıpkı yapay açıklayıcı değişkenlerdeki gibi 0 ve 1 değerini alır. İki değerli (kesikli) bağımlı değişkenleri ele almada çok yaygın şekilde kullanılan dört yaklaşım vardır. Bunlar; **Doğrusal Olasılık Modeli, Logit Model, Probit Model ve Tobit Model’dir** [49].

6.2. LOJİSTİK REGRESYON İLE DOĞRUSAL REGRESYONUN FARKI

Lojistik regresyonu, doğrusal regresyondan ayıran en belirgin özellik, lojistik regresyonda bağımlı değişkeninin kategorik değişken olmasıdır. Lojistik regresyon ve doğrusal regresyon arasındaki bu fark, hem parametrik model seçimine, hem de varsayımlara yansımaktadır. Lojistik regresyonda da, doğrusal regresyon analizinde olduğu gibi bazı değişken değerlerine dayanarak kestirim yapılmaya çalışılır, ancak iki yöntem arasında üç önemli fark vardır :

- 1- Doğrusal regresyon analizinde tahmin edilecek olan bağımlı değişken sürekli iken, lojistik regresyonda bağımlı değişken kesikli bir değer olmalıdır.
- 2- Doğrusal regresyon analizinde bağımlı değişkenin değeri, lojistik regresyonda ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı kestirilir.
- 3- Doğrusal regresyon analizinde bağımsız değişkenlerin çoklu normal dağılım göstermesi koşulu aranırken, lojistik regresyonun uygulanabilmesi için bağımsız değişkenlerin dağılımına ilişkin hiçbir ön koşul yoktur [43].

Bilinen doğrusal regresyon analizinde bağımlı değişken ve bağımsız değişken(ler) sayısal (ölçümle belirtilen sürekli ya da kesikli sayısal) olarak belirtilir. Örneğin, yaş ile kan basıncı arasında bir ilişki aranacaksa; hem yaş, hem de kan basıncı sayısal olarak belirtilmelidir. Nitelik olarak belirtilemezler. Bağımlı değişken nitelik olarak belirtilirse, bağımsız değişken ya da değişkenlerle arasındaki ilişki lojistik regresyon yöntemiyle aranır [50]. Aşağıdaki Çizelge 6.1’de iki yöntem karşılaştırılmıştır:

Çizelge 6.1. Doğrusal ve lojistik regresyon karşılaştırması [50].

Değişkenler	Doğrusal Regresyon Analizi	Lojistik Regresyon Analizi
Bağımlı	Sürekli Sayısal Kesikli Sayısal	Nitelik
Bağımsız	Sürekli Sayısal Kesikli Sayısal	Sürekli Sayısal Kesikli Sayısal Nitelik (Her bağımsız değişken başka bir ölçüm biçimine de sahip olabilir)

Lojistik regresyon analizi sonucunda elde edilen modelin uygun olup olmadığı “model ki-kare” testi ile her bir bağımsız değişkenin modelde varlığının anlamlı olup olmadığı ise wald istatistiği ile test edilir.

6.3. LOJİSTİK REGRESYON MODELLERİ

Log-linear, logit ve probit modeller iki şıklı bağımlı değişkenleri açıklamada regresyon gibi genel doğrusal modellerin temelini oluşturmaktadır. Bu modeller bağlantı fonksiyonu olarak “Sıradan En Küçük Kareler” tahmini yerine “Maksimum Benzerlik (En Çok Olabilirlik)” tahminini kullanarak standart regresyondan ayrılır [47].

Fonksiyon, log-linear analizde bağımlı değişken “y”nin logaritmasında, logit analizde log odd’lar (bahis oranları) kullanılırken, probit analizinde kümülatif normal dağılım (cumulative normal distribution) kullanılır [47,51].

Lojistik Regresyon genel olarak üçe ayrılır:

- 1- İkili (Binary) lojistik regresyon: Bağımlı değişken iki düzeyli olduğunda kullanılır(Var-Yok, Evet-Hayır). Bu çalışmada ikili lojistik regresyon tekniği kullanılmıştır.
- 2- Sıralı (Ordinal) lojistik regresyon: Bağımlı değişken sıralı nitel veri tipinde (hafif-orta-şiddetli vb.) olduğunda kullanılır.
- 3- Multinomial lojistik regresyon: Bağımlı değişken ikiden çok düzeyli sıralı olmayan nitel veri tipinde olduğunda kullanılır.

6.4. LOGİT MODEL

Odds başarı ya da görülme olasılığının “p”, başarısızlık ya da görülmemeye olasılığına “1-p” oranıdır. Başka bir deyişle: İncelenen bir olayın olasılığının kendi dışında kalan diğer olayların olasılığına oranına “odds değeri” denilir. **Odds ratio (OR)** ise iki odds’un birbirine oranıdır. İki değişken arasındaki ilişkinin özet bir ölçüsüdür (bahis oranı da denir).

Çizelge 6.2. UDP Protokolü ODDS değerleri.

Protokol udp mi?	Saldırı		Toplam
	Var	Yok	
Evet	3	127	130
Hayır	45	180	225
Toplam	48	307	355

Örneğin yukarıdaki Çizelge 6.2'ye bakıldığında:

Protokol tipi udp olan bağlantılarda saldırı olma odds'u:

$$(3/130) / (127/130) = 3/127 = 0,024$$

Udp olmayan bağlantılarda ise:

$$45/180 = 0,25$$

$$\text{Odds ratio} = 0,25 / 0,024 = 10,4$$

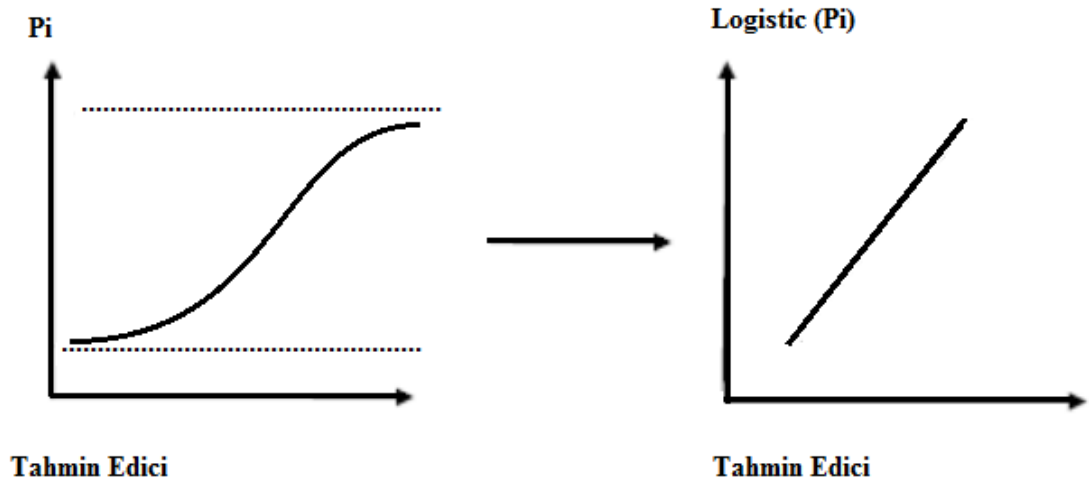
Bu sonuç udp protokolü kullanmayan bağlantıların saldırı olma olasılığının, udp kullananlardan yaklaşık 10 kat daha fazla olduğunu göstermektedir.

Logit ismi, odd değerinin doğal logaritmasını ifade etmektedir. Yani π olasılığı göstermek üzere, logit;

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) \quad (6.1)$$

Logit model, bağımsız değişken değeri sonsuza gittiği zaman, bağımlı değişkenin 1'e asimptot olduğu matematiksel bir fonksiyondur [24].

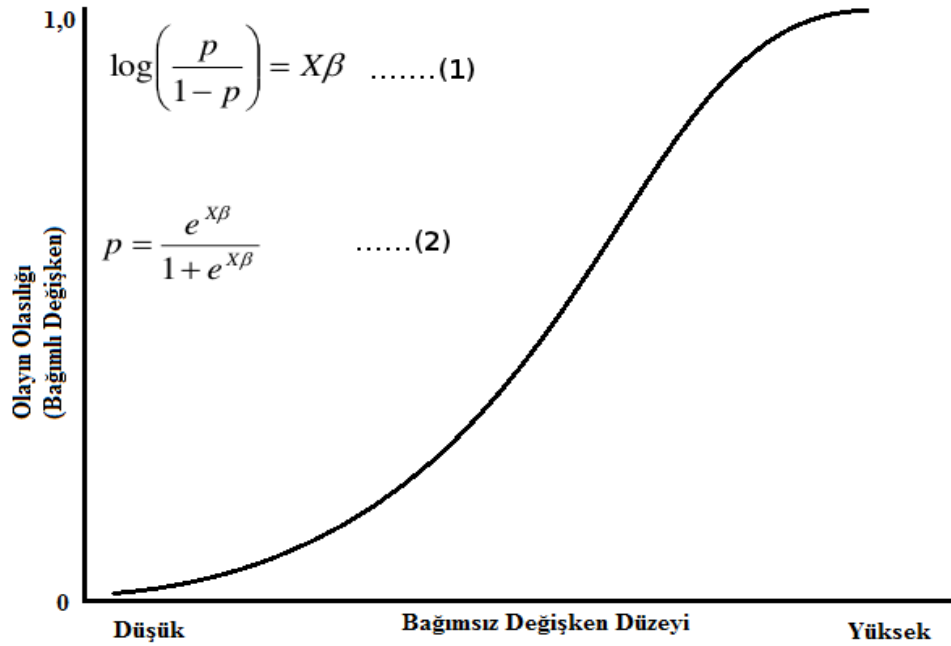
Logit modellerinde olasılıklar 0 ile 1 arasında sınırlandırılmışlardır. Bunu yaparken lojistik regresyon modeli olasılıklara bir dönüşüm uygulamaktadır; çünkü olasılıklar ve tahmin edici değişken arasındaki ilişki doğrusal değildir ve S şeklinde bir eğridir. Aşağıda Şekil 6.2'den görüleceği üzere, lojistik regresyon varsayımı altında olasılıkların lojistik dönüşümü ok ile gösterilmekte olup bu dönüşüm, tahmin edici değişkenler ile olasılıkların doğrusal bir ilişki içerisinde sonuçlanmasını sağlamaktadır [52].



Şekil 6.2. Olasılık ile tahmin edici arasındaki ilişki [52].

6.4.1. Lojistik Regresyon Formülü

Lojistik regresyon analizi bağımlı değişkenin 0 ile 1 değerleri arasında aldığı ve tahmin edilen olasılıkların 0-1 aralığında sınırlı olacağını kabul eder. 0-1 aralığında sınırlandırılmış ilişkiyi tanımlamak için lojistik regresyon yöntemi lojistik eğrisini kullanmaktadır (Şekil 6.3). Lojistik regresyon fonksiyonu bağımlı ve bağımsız değişkenler arasında aşağıdaki lojistik fonksiyonunu kullanmaktadır:



Şekil 6.3. Lojistik ilişkinin biçimi [53].

Yukarıdaki Şekil 6.3'te geçen formülde p olayın olma olasılığı (bağımlı değişkenin tahmin edilen değerini) vektörünü, β model parametreleri vektörünü, X ise sabit terimi de içerisinde barındıran bağımsız değişkenler matrisini temsil etmektedir. Bu fonksiyondan hareketle bağımlı değişkenin tahmin edilen olasılık değerleri (p) vektörü aşağıdaki Eşitlik 6.2'deki formülle hesaplanmaktadır [53].

$$P = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k}} \quad (6.2)$$

Yukarıdaki Eşitlik 6.2'de:

P : İncelenen olayın gözlenme olasılığını,

β_0 : Bağımsız değişkenler sıfır değerini aldığı anda bağımlı değişkenin değerini (yani sabiti),

$\beta_1 \beta_2 \dots \beta_k$: Bağımsız değişkenlerin regresyon katsayılarını,

$X_1 X_2 \dots X_k$: Bağımsız değişkenleri,

k: Bağımsız değişken sayısını,

e: 2.71 sayısını göstermektedir [54].

6.4.2. Logit Model ile Lojistik Regresyon Arasındaki Benzerlik ve Farklar

Lojistik regresyon, bir ya da daha fazla açıklayıcı değişken ile ikili bir yanıt değişkeni arasındaki ilişkiyi göstermek için kullanılan bir tekniktir. Açıklayıcı değişkenler kategorik, sürekli ya da her ikisi de olabilir. Logit modellerde ise açıklayıcı değişkenler sadece kategorik değişkenlerden oluşabilir [55].

6.4.3. Lojistik Regresyon İçin Basit bir Örnek ve Çözümü

Aşağıdaki Çizelge 6.3'te okulumuzdaki web sunucusuna yapılan(hedef IP'ler sunucunun, kaynaklar farklı olabilir) bağlantılarla ilgili ağ trafiğinden elde edilen logların incelenmesi sonucu özetlenen veriler görülmektedir:

Çizelge 6.3. Web sunucusu bağlantıları eğitim veri seti.

TOPLAM BYTE	PROTOKOL	PORT	ANOMALİ	NORMAL
Gönderilen veri > 150 byte ise 1 değilse 0	TCP = 1 UDP = 0	Login portu ise 1 değilse 0	Saldırı Bağlantısı Sayısı	Normal Bağlantı Sayısı
0	0	0	9	168
0	1	0	35	275
1	0	0	8	37
1	1	1	19	58

Buna göre yapılan bir bağlantının aşağıdaki gibi gözlemlendiği bir anda saldırı olma ihtimali nedir?

Toplam oturum sayısı: 350

Protokol: TCP

PORT: 21 (FTP login)

Gönderilen veri: 200 byte

Çözüm:

Denklemden bağımlı değişken saldırı olmasıdır. Bağımsız değişkenler:

“TOPLAM BYTE” kısaca X_1 denilmiştir.

“PROTOKOL” kısaca X_2 denilmiştir.

“PORT” kısaca X_3 denilmiştir.

Sabitimize de kısaca β_0 denilmiştir.

İlk aşamada bulunacak regresyon eşitliği aşağıdaki gibi olacaktır:

$$g(x) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$$

Şimdi ilk satırdan başlayarak Çizelge 6.4'te görüldüğü üzere denklemleri yazılacak olursa:

Çizelge 6.4. Lojistik regresyon web sunucu bağlantısı örnek çözümü.

TOPLAM BYTE	PROTOKOL	PORT	ANOMALİ	NORMAL	LOGIT
Gönderilen veri > 150 byte ise 1 değilse 0	TCP = 1 UDP = 0	Login portu ise 1 değilse 0	Saldırı Bağlantı Sayısı	Normal Bağlantı Sayısı	$\ln(\text{Anomali/Normal})$
$X_1 = 0$	$X_2 = 0$	$X_3 = 0$	9	168	$\ln(9/168) = -2,9267$
$X_1 = 0$	$X_2 = 1$	$X_3 = 0$	35	275	$\ln(35/275) = -2,0614$
$X_1 = 1$	$X_2 = 0$	$X_3 = 0$	8	37	$\ln(8/37) = -1,5315$
$X_1 = 1$	$X_2 = 1$	$X_3 = 1$	19	58	$\ln(19/58) = -1,1160$

Yukarıdaki tablodan ilk satırdan başlayarak her satır için bir denklem olmak üzere aşağıdaki denklemler elde edilir:

$$\beta_0 + \beta_1*0 + \beta_2*0 + \beta_3*0 = -2,9267$$

$$\beta_0 + \beta_1*0 + \beta_2*1 + \beta_3*0 = -2,0614$$

$$\beta_0 + \beta_1*1 + \beta_2*0 + \beta_3*0 = -1,5315$$

$$\beta_0 + \beta_1*1 + \beta_2*1 + \beta_3*1 = -1,1160$$

Yukarıdaki denklemler ortak olarak çözüldüğünde katsayılar aşağıdaki gibi bulunur:

$$\beta_0 = -2,9267$$

$$\beta_1 = 1,3952$$

$$\beta_2 = 0,8683$$

$$\beta_3 = -0,4498$$

Regresyon eşitliği:

$$g(x) = \beta_0 + \beta_1*X_1 + \beta_2*X_2 + \beta_3*X_3$$

$$g(x) = -2,9267 + 1,3952*X_1 + 0,8683*X_2 + -0,4498*X_3$$

Soruda olasılığı sorulan veriler aşağıdaki Çizelge 6.5'te görülmektedir:

Çizelge 6.5. Lojistik regresyon web sunucu bağlantısı olasılığı için verilenler.

TOPLAM BYTE	PROTOKOL	PORT
Gönderilen veri > 150 byte ise 1 değilse 0	TCP = 1 UDP = 0	Login portu = 1 değilse 0
350 > 150	TCP	21 (FTP login)
$X_1 = 1$	$X_2 = 1$	$X_3 = 1$

Bulunan g(x) eşitliğinde yukarıdaki veriler yerine koyulursa:

$$g(x) = -2,9267 + 1,3952*1 + 0,8683*1 + -0,4498*1 = -1,113$$

İstenen olayın olasılığı (Y=1);

$$P(Y = 1 | X_1, X_2, \dots, X_p) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}}$$

Olasılık hesabı aşağıdaki gibi olur:

$$P = 1/(1 + e^{-g(x)}) = 1/(1 + e^{1,113}) = 0,2473$$

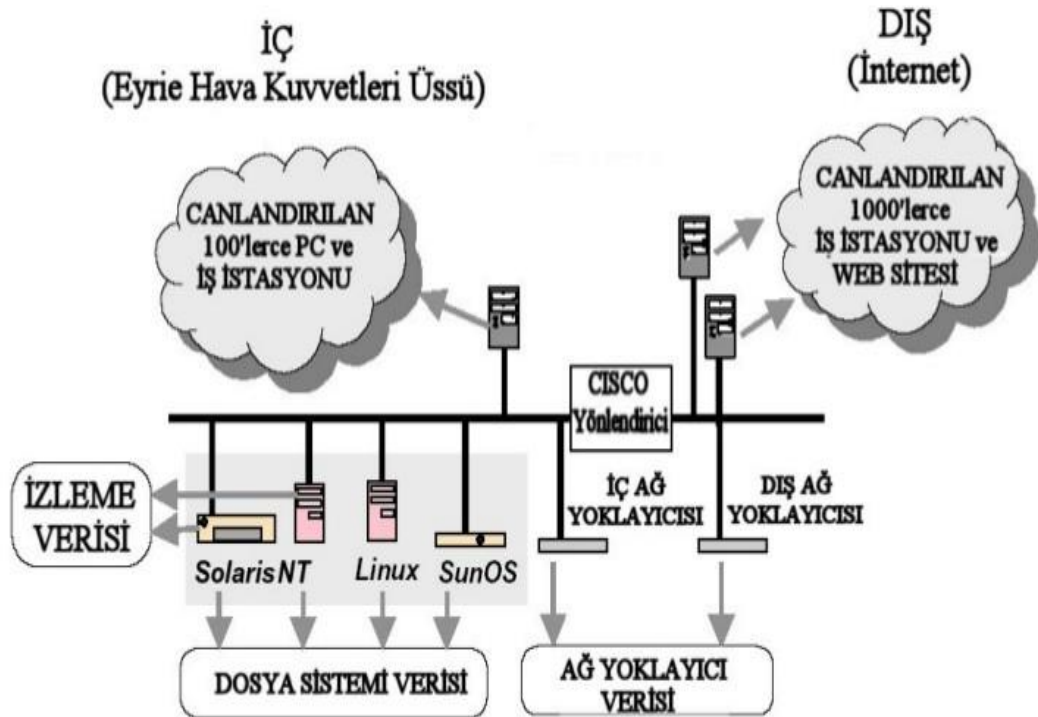
Yani verilen bağlantı bilgilerinin bir saldırı olma olasılığı yaklaşık olarak ¼ tür (Bu da düşük bir ihtimali göstermektedir).

BÖLÜM 7

MATEMATİKSEL MODELİN OLUŞTURULMASI

7.1. VERİ SETİ ANALİZİ

Saldırı Tespit Sistemleriyle ilgili çalışmalarda en sık kullanılan veri seti DARPA 1998 ve 1999 veri setleridir. Model oluşturma çalışmasında yine bu verilerden türetilen KDD Cup'99 veri seti kullanılmıştır. Veri setini oluşturan kaynak aşağıdaki Şekil 7.1'de de görüldüğü gibi saldırının hedefi olan bir iç ağ ve saldırıyı gerçekleştiren bir dış ağ olmak üzere iki farklı ağdan oluşmaktadır:



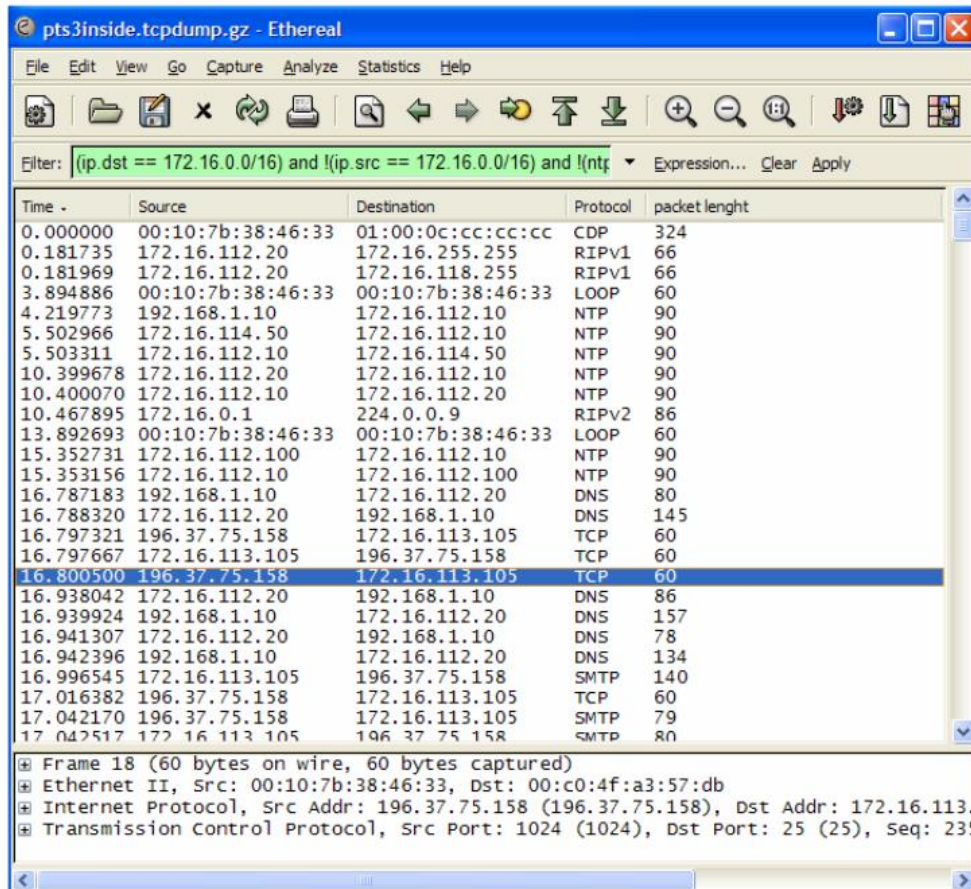
Şekil 7.1. Veri setinin oluşturulduğu ağın şeması [56].

Hava kuvvetlerini temsil eden iç ağ 172.16.0.0/16 IP adres uzayına sahiptir. Bu ağ içerisinde dört “kurban” makine bulunmaktadır.

Bunların üzerinde SunOS, Solaris, Linux, ve Windows NT çalışmaktadır (1998 veri setlerinde sadece UNIX makineler kullanılmıştır). Şekil 7.1’de görülen trafik oluşturucular yüzlerce sunucuyu ve çeşitli uygulamaları çalıştıran internet kullanıcılarını temsil etmektedir.

Protokoller (HTTP, SMTP, telnet ...) ve trafik yoğunlukları, 1998’de gerçek Hava Kuvvetleri ağından toplanan trafiğe benzer olacak şekilde tasarlanmıştır. Ağ üzerinde 2 noktadan veri toplanmıştır: dört kurban makine ile yönlendirici arasındaki iç ağ dinleyicisi ve yönlendirici ile internet arasındaki dış ağ dinleyicisinden. Bu saldırı yazılımları internet ortamından ve hacker sitelerinden toplanmış saldırılardır [56].

DARPA verileri ile çalışırken matlab ya da SQL sunucularla birlikte tcpdump çıktıları Şekil 7.2’de görüldüğü üzere wireshark (eski adı ethereal) programıyla da incelenebilir (bu çalışmada SQL kullanılmıştır) :



Time	Source	Destination	Protocol	packet length
0.000000	00:10:7b:38:46:33	01:00:0c:cc:cc:cc	CDP	324
0.181735	172.16.112.20	172.16.255.255	RIPv1	66
0.181969	172.16.112.20	172.16.118.255	RIPv1	66
3.894886	00:10:7b:38:46:33	00:10:7b:38:46:33	LOOP	60
4.219773	192.168.1.10	172.16.112.10	NTP	90
5.502966	172.16.114.50	172.16.112.10	NTP	90
5.503311	172.16.112.10	172.16.114.50	NTP	90
10.399678	172.16.112.20	172.16.112.10	NTP	90
10.400070	172.16.112.10	172.16.112.20	NTP	90
10.467895	172.16.0.1	224.0.0.9	RIPv2	86
13.892693	00:10:7b:38:46:33	00:10:7b:38:46:33	LOOP	60
15.352731	172.16.112.100	172.16.112.10	NTP	90
15.353156	172.16.112.10	172.16.112.100	NTP	90
16.787183	192.168.1.10	172.16.112.20	DNS	80
16.788320	172.16.112.20	192.168.1.10	DNS	145
16.797321	196.37.75.158	172.16.113.105	TCP	60
16.797667	172.16.113.105	196.37.75.158	TCP	60
16.800500	196.37.75.158	172.16.113.105	TCP	60
16.938042	172.16.112.20	192.168.1.10	DNS	86
16.939924	192.168.1.10	172.16.112.20	DNS	157
16.941307	172.16.112.20	192.168.1.10	DNS	78
16.942396	192.168.1.10	172.16.112.20	DNS	134
16.996545	172.16.113.105	196.37.75.158	SMTP	140
17.016382	196.37.75.158	172.16.113.105	TCP	60
17.042170	196.37.75.158	172.16.113.105	SMTP	79
17.042517	172.16.113.105	196.37.75.158	SMTP	80

Frame 18 (60 bytes on wire, 60 bytes captured)
Ethernet II, Src: 00:10:7b:38:46:33, Dst: 00:c0:4f:a3:57:db
Internet Protocol, Src Addr: 196.37.75.158 (196.37.75.158), Dst Addr: 172.16.113.105
Transmission Control Protocol, Src Port: 1024 (1024), Dst Port: 25 (25), Seq: 235

Şekil 7.2. Veri setinin wireshark ile analizi [57].

Ağ dinleyicisi iki yönlü paketleri yakaladığı için kurban makinelere gelen paketler için varış IP adresi 172.16.x.x olan paketler olarak süzölmelidir. Örnek bir filtre: (ip.dst == 172.16.0.0/16) and !(ip.src == 172.16.0.0/16) and !(ntp) and !(rip) and !(loop) and !(arp) and !(nbns) [57].

7.2 VERİ SETİNİN HAZIRLANMASI

Bu çalışmada “kddcup.data_10_percent_corrected” dosya ismi ile internetten indirilebilen yaklaşık 75 Mb büyüklüğünde ve içinde yaklaşık 500 bin kayıt bulunan gerçek kddCupp-99 veri seti’nin % 10’una karşılık gelen veri seti kullanılmıştır. Veri setinin ilk 250 bin kaydı model oluşturmak için kalanı ise test için kullanılmıştır.

Veritabanında kullanılan tablo aşağıdaki SQL komutu ile oluşturuldu:

```
create table tbl_kdd( transmission int NOT NULL AUTO_INCREMENT, duration long, protocol_type varchar(35), service varchar(35), flag varchar(35), src_bytes double, dst_bytes double, land varchar(35), wrong_fragment double, urgent double, hot double, num_failed_logins double, logged_in varchar(60), num_compromised double, root_shell double, su_attempted double, num_root double, num_file_creations double, num_shells double, num_access_files double, num_outbound_cmds double, is_host_login varchar(60), is_guest_login varchar(60), countk double, srv_count double, serror_rate double, srv_serror_rate double, rerror_rate double, srv_rerror_rate double, same_srv_rate double, diff_srv_rate double, srv_diff_host_rate double, dst_host_count double, dst_host_srv_count double, dst_host_same_srv_rate double, dst_host_diff_srv_rate double, dst_host_same_src_port_rate double, dst_host_srv_diff_host_rate double, dst_host_serror_rate double, dst_host_srv_serror_rate double, dst_host_rerror_rate double, dst_host_srv_rerror_rate double, label varchar(60), CONSTRAINT transmissionPk PRIMARY KEY (transmission) );
```

Verileri veritabanına aktarma işleminde verileri tabloya alma işlemi için MySQL komut satırında aşağıdaki komut kullanılabilir:

load data local infile 'xaa' into table tbl_kdd fields terminated by ',' enclosed by "
lines terminated by '\n'

(duration, protocol_type, service, flag, src_bytes, dst_bytes, land, wrong_fragment,
urgent, hot, num_failed_logins, logged_in, num_compromised, root_shell,
su_attempted, num_root, num_file_creations, num_shells, num_access_files,
num_outbound_cmds, is_host_login, is_guest_login, countk, srv_count, serror_rate,
srv_serror_rate, error_rate, srv_error_rate, same_srv_rate, diff_srv_rate,
srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_same_srv_rate,
dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate,
dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_error_rate,
dst_host_srv_error_rate, label)

Fakat yukarıdaki “load data local infile 'kddcup.data.corrected' ...” komutu eski MySQL versiyonlarında desteklenmediğinden bunun yerine phpmyadmin web arayüzünden de verileri tablolara alma işlemi yapılabilir. Bunun için ise veri miktarı çok fazla olduğundan bazı ayar dosyalarında değişiklikler yapılması gerekir:

- /etc/php5/apache2/php.ini dosyasındaki değerler aşağıdaki gibi yapıldı:

upload_max_filesize = 3000, memory_limit = 4000, post_max_size = 3000,
max_execution_time = 0, max_input_time = -1

- /etc/phpmyadmin/config.inc.php dosyasına aşağıdaki satır eklenmiştir (varsayılan olarak 300 saniye iken 0 yapılarak sınırlar kaldırılmıştır)

\$cfg['ExecTimeLimit'] = 0;

Ayrıca veri setinin sadece % 10’unda değil de tümünde yapılacak veri madenciliği çalışmaları için tüm dosya boyutu 750 Mb olduğu için incelemelerde dosya boyutunun fazlalığından kaynaklanan sorunlarla karşılaşmamak için dosya her birinde 1 milyon satır bulunan beş parçaya bölünebilir, bunun için linux makinede kullanılan komut:

split -l 1000000 kddcup.data.corrected

750 Mb’lık veriyi tabloya aldıktan sonra (ki ubuntu-apache-MySQL kurulu 4Gb Ram’e sahip, 4 çekirdekli, 2 Ghz işlemcili bir PC için bu işlem ortalama 2 gün sürmektedir) KDD Cup’99 veri seti toplam 4 898 431 satır (her satır bir bağlantı - connection veya transmission-) içermektedir:

```
mysql> select count(*) from tbl_kdd ;
+-----+
| count(*) |
+-----+
| 4898431 |
+-----+
1 row in set (0,00 sec)
```

Bu çalışmada “kddcup.data_10_percent_corrected” dosyası ikiye bölünerek ilk yarısı eğitim, ikinci yarısı test verisi olarak kullanıldığından eğitim seti toplam 250 000 satır içermektedir:

```
mysql> select count(*) from yuzdeon_ilk250bin ;
+-----+
| count(*) |
+-----+
| 250000 |
+-----+
1 row in set (0,37 sec)
```

7.3 KOLON (DEĞİŞKEN) SEÇİMİ

Veri setinde toplam 41 adet değişken bulunmaktadır. Aşağıdaki prensiplere uyarak bu sayı 9’a indirilmiştir:

- 1- Paketlerin sadece başlık bilgisine değil içeriğine de bakılarak anlaşılacak alanlar da alınmıştır (örneğin bu yüzden hot, su_attempted gibi alanlar alınmıştır).
- 2- Parametrelerin birbirlerinden bağımsız olanları seçilmiştir. Örneğin “root_shell”, “su_attempted”, “num_root” alanlarının tümü birden alınmak yerine “su_attempted” alanı alınmıştır.
- 3- Parametrelerin bağımlı değişkeni etkilemeyecek olanları seçilmemiştir. Örneğin “src_bytes” ve “dst_bytes” alanları bu yüzden alınmamıştır.

Verileri incelemek ve ikili hale getirmek için öncelikle veriler SQL sunucusuna alınıp aşağıdaki kurallara uygun olarak ikili hale getirilmiştir. İkili hale getirmekle hem performans artışı hem de çalışmalarda kolaylık olması sağlanmıştır, ayrıca anlaşılabilirlik da arttırılmıştır, yoksa ikili (binary) lojistik regresyonda bağımlı değişkenin iki düzeyli olması yeterli olup, bağımsız değişkenler için böyle bir şart bulunmamaktadır:

protocol_type: Bağlantının(connection, veri setindeki her satır bir bağlantıdır) protokol tipini gösterir; tcp, udp, icmp değerlerinden birini alabilir. İkili hale dönüşüm:

Tcp=1; udp veya icmp=0.

service: Hedefteki ağ servisini gösterir(network service on the destination). İkili hale dönüşüm:

Smtpt, ftp, pop_3, ldap, login, imap4, auth, IRC, telnet, sql_net, exec, shell, klogin, kshell = 1; diğerleri = 0.

flag: Bağlantı durumunu gösterir. SF bağlantının normal bir şekilde sonlandığını, OTH ise bağlantı takip işinin bağlantının ortasında başladığını gösterir. İkili hale dönüşüm:

SF veya OTH = 0; diğerleri 1.

land: Hedef ve kaynak IP/port bilgilerinin aynı olup olmadığını gösterir. İkili hale dönüşüm:

Hedef ve kaynak ip/port bilgileri aynı ise = 1; değilse 0.

wrong_fragment: Hatalı fragment sayısını gösterir. İkili hale dönüşüm:

Sıfır ise=0; değilse=1.

hot: Bir bağlantıda çalıştırılan kritik komut sayısını gösterir. Örneğin sistem klasörüne girmek, programlar oluşturup çalıştırmak gibi. İkili hale dönüşüm:

Sıfırdan büyük ise 1; değilse 0.

num_failed_logins: Hatalı login işlemlerinin sayısını gösterir. İkili hale dönüşüm: Sıfır 0; değilse 1.

su_attempted: Root kullanıcısı haklarını elde etmek için, “su root” komutunun denenip denenmediğini gösterir. İkili hale dönüşüm: “su root” komutu denenmişse 1 diğer durumda 0.

num_access_files: Kontrol ya da erişim izinlerini tutan kritik dosyalarda yapılan işlem sayısını gösterir. İkili hale dönüşüm: Sıfırdan büyük ise 1; değilse 0.

Yukarıdaki değişimler için çalıştırılan SQL komutları:

```
update yuzdeon_ilk250bin set protocol_type = '1' where protocol_type = 'tcp';
update yuzdeon_ilk250bin set protocol_type = '0' where protocol_type != '1';
update yuzdeon_ilk250bin set service = '1' where service in('smtp', 'ftp', 'pop_3',
'ldap', 'login', 'imap4', 'auth', 'IRC', 'telnet', 'sql_net', 'exec', 'shell', 'klogin', 'kshell' );
update yuzdeon_ilk250bin set service = '0' where service != '1';
update yuzdeon_ilk250bin set flag = '0' where (flag = 'SF') or (flag = 'OTH') ;
update yuzdeon_ilk250bin set flag = '1' where flag != '0';
update yuzdeon_ilk250bin set wrong_fragment = 1 where wrong_fragment > 0;
update yuzdeon_ilk250bin set hot = 1 where hot != 0;
update yuzdeon_ilk250bin set num_failed_logins = 1 where num_failed_logins
!= 0;
update yuzdeon_ilk250bin set su_attempted = 1 where su_attempted > 0;
update yuzdeon_ilk250bin set num_access_files = 1 where num_access_files > 0;
update yuzdeon_ilk250bin set label = '0' where label = 'normal.';
update yuzdeon_ilk250bin set label = '1' where label != '0';
```

İkili hale getirirken veri analizinin dikkatli yapılması ve sonuca etkisi yüksek oranda olan değişkenlerin özellikle seçilmesi gerekmektedir. Örneğin **protocol_type** değeri için aşağıda yapılan analizlere göre en az ağırlık icmp protokolünde olduğundan icmp protokolü ayrı tutulmayıp, udp ile birlikte alınmıştır (tcp=1; udp veya icmp=0).
Select count(*) from tbl_kdd ; => Sonuç: **4898431**

Select count(*) from tbl_kdd where protocol_type = 'udp'; => Sonuç: **194288**

Select count(*) from tbl_kdd where protocol_type = 'udp' and label != 'normal.';

=> Sonuç: **2940**

Select count(*) from tbl_kdd where protocol_type = 'icmp'; => Sonuç: **2833545**

Select count(*) from tbl_kdd where protocol_type = 'icmp' and label != 'normal.';

Sonuç: **2820782**

Select count(*) from tbl_kdd where protocol_type = 'tcp'; => Sonuç: **1870598**

Select count(*) from tbl_kdd where protocol_type = 'tcp' and label != 'normal.';

Sonuç: **1101928**

Verilerin örnek görüntüsü aşağıda Çizelge 7.1’de görülmektedir (en son kolondaki label bu kaydın-satırın- saldırı olup olmadığını tutmaktadır, saldırılar için 1, normal kayıtlar için 0 değeri verilmiştir) :

Select protocol_type, service, flag, land, wrong_fragment, hot, num_failed_logins, su_attempted, num_access_files, label **from yuzdeon_ilk250bin**

Çizelge 7.1. Veri seti örnek görüntüsü.

protocol_type	service	flag	land	wrong_fragment	hot	num_failed_logins	su_attempted	num_access_files	label
1	1	0	0	1	1	1	0	1	1
1	0	0	1	0	0	0	1	1	1
0	0	1	1	1	0	0	0	1	0
0	1	1	0	0	1	1	0	0	1
1	0	0	1	0	1	0	1	1	0

7.4. MODELİN OLUŞTURULMASI

SPSS yazılımını kullanarak veri seti ikili lojistik regresyon ile analiz edildiğinde aşağıdaki sonuçlar elde edilmiştir:

Çizelge 7.2. Durum işleme özeti.

Ağırlıklandırılmamış Durumlar		N	Yüzde
Seçilen Durumlar	Analize Alınanlar	250000	100,0
	Eksik Durumlar	0	,0
	Toplam	250000	100,0
Seçilmeyen Durumlar		0	,0
Toplam		250000	100,0

Yukarıdaki Çizelge 7.2 veri setindeki tüm kayıtların analiz için kullanıldığını göstermektedir. Veri setinde 250 bin kayıt vardı, çizelgede görüldüğü üzere tümü incelemeye alınmıştır.

Çizelge 7.3. Bağımlı değişken kodlaması.

Orjinal Değer	İçsel Değer
0	0
1	1

Çizelge 7.3 SPSS'in, bağımlı değişken için, veri setindeki "0"ları ikili (binary) "0" olarak , "1"leri ise ikili "1" olarak aldığını göstermektedir, ki böyle olması için data seti önceden buna göre hazırlanmıştır. Çünkü incelenmek istenen konu saldırı olma durumu olduğundan saldırı olma durumu 1, olmaması durumu ise 0 olarak kodlanmıştır. Bu tamamen bir tercih meselesidir ve tamamen tersi de seçilebilirdi. Fakat sonuçların yorumunun kolay olması için genelde asıl ilgilenilen cevap için "1" kullanılması işleri daha çok kolaylaştırmaktadır.

“Block 1: Method = Enter”

SPSS çıktısındaki yukarıdaki ifade ise metot olarak giriş (enter) seçildiğini yani adımsal (stepwise) ya da hiyerarşik metotların kullanılmadığını gösterir.

Çizelge 7.4. Sınıflandırma tablosu.

Gerçek		Tahmin		
		label		Doğruluk Yüzdesi
		0	1	
label	0	59805	11420	84
	1	736	178039	99,6
Toplam				95,1
Yüzde				
Eşik değeri 0,5				

Çizelge 7.4’e bakıldığında tüm veri setinde doğru sınıflandırılan yani saldırı olup olmadığının doğru olarak tahmin oranının % 95,1 olduğu görülmektedir.

Saldırı olmayan birbirinden farklı toplam $59805+11420=71225$ adet kayıt olduğu ve bunların 59805 tanesini yani % 84’ünü modelin doğru tahmin ettiği görülmektedir.

Çizelge 7.5. Eşitlikteki değişkenler.

		B	S.E.	Wald	df	Sig.	Exp(B)
A d i m 1 ^a	protocol_type	-7,133	0,039	3,380E4	1	0,000	0,001
	service	-1,372	0,068	401,313	1	0,000	0,254
	flag	6,882	0,041	2,811E4	1	0,000	974,877
	land	34,951	3,446E7	0,000	1	1,000	1,509E15
	wrong_fragment	34,700	6,372E6	0,000	1	1,000	1,175E15
	hot	6,454	0,071	8,375E3	1	0,000	635,420
	num_failed_logins	3,803	0,889	18,313	1	0,000	44,854
	su_attempted	2,610	1,434	3,315	1	0,069	13,602
	num_access_files	-,267	0,486	0,302	1	0,583	0,765
	Constant	2,843	0,012	5,856E4	1	0,000	17,175
a. 1. Adımda ele alınan değişkenler: protocol_type, service, flag, land, wrong_fragment, hot, num_failed_logins, su_attempted, num_access_files.							

Yukarıdaki Çizelge 7.5 ise kullanılan değişkenlerin hangilerinin hangi katsayılarla nihai modelde bulunacağını göstermektedir. Örneğin tabloda “service” değişkeninin katsayısının (coefficent değerinin) -1,372 olması demek değişkendeki 1 birimlik artışın(yani 0 yerine 1 olmasının başka bir deyişle “http” yerine “telnet” olmasının) sonucun log odds (yani logit) değerini -1,372 oranında düşürdüğünü gösterir.

“land” ve “wrong_fragment” hariç diğer değişkenlerin içinde mutlak değer olarak en büyük katsayıya -7,133 ile ”protocol_type” ; en küçüğüne ise -0,267 ile “num_access_files” değişkenlerinin sahip olduğu görülmektedir. Bu da bir kaydın saldırı olup olmadığını belirlemede en büyük belirleyici role sahip parametrenin ”protocol_type” olduğunu; en az ağırlığın ise “num_access_files” değişkeninde olduğunu göstermektedir.

Parametrelerin S.E.(Standart Error) değeri ise tahminin ne kadar stabil olduğunun ölçüsüdür ve ne kadar düşüğe o kadar tutarlı sonuçlar alınır. Örneğin “land” ve “wrong_fragment” hariç diğer tüm değişkenlerin ortalaması yaklaşık 0,4 iken bu iki değişkenin ortalaması yaklaşık 5 civarındadır, bu da bu iki değişkenin stabiliteyi ne kadar fazla bozduğunu göstermektedir.

Wald istatistiği değişkenlerin hangilerinin model için anlamlı olduğunu hangilerinin gereksiz olduğunu göstermektedir. Değişkenler içinde wald değeri sıfıra çok yakın olan “land” ve “wrong_fragment” değişkenlerinin modelde gereksiz olduğu sonucu çıkmaktadır.

Tablodaki “Sig.” kolonu ise değişkenin anlamlılık düzeyini göstermekte olup SPSS’te varsayılan olarak $p < 0,05$ olarak çalışmaktadır. Sig değeri sıfıra ne kadar çok yakınsa parametrenin modeldeki anlamlılık düzeyi o kadar fazla demektir. Buna göre Sig değeri 1 olan “land” ve “wrong_fragment” değişkenlerinin modelde anlamlı olmadıkları sonucu çıkmaktadır.

Sig, wald ve SE değerlerinden anlaşıldığı üzere “land” ve “wrong_fragment” değişkenleri nihai modelde bulunmayacaklardır. Buna göre nihai model aşağıdaki gibi olacaktır:

Regresyon eşitliği aşağıdaki gibi olmak üzere:

$$g(x) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_k \cdot X_k$$

$$g(x) = 2,843 + \text{protocol_type}*(-7,133) + \text{service}*(-1,372) + \text{flag}*6,882 + \text{hot}*6,454 + \text{num_failed_logins}*3,803 + \text{su_attempted}*2,610 + \text{num_access_files}*(-0,267)$$

$$P = 1/(1 + e^{-g(x)})$$

7.5 MODELİN UYGULAMASINA ÖRNEK

Örneğin aşağıdaki gibi bir kayıt için modelin ürettiği sonuca bakılırsa:

protocol_type=tcp, service=telnet, flag=S0, hot=0, num_failed_logins=0, su_attempted=0, num_access_files=0, label=neptune.

Kaydın label yani saldırı olup olmadığı ile ilgili bilgi alınacak alanda “neptune” yazmaktadır. Yani bu bir “neptune” saldırısıdır.

Her parametre iki kategorili hale çevrilip $g(x)$ fonksiyonunda yerine koyulursa:

$$g(x) = 2,843 + 1*(-7,133) + 1*(-1,372) + 1*6,882 + 0*6,454 + 0*3,803 + 0*2,610 + 0*(-0,267) = 1,22$$

$$P = 1/(1 + e^{-g(x)}) = 1/(1 + e^{-1,22}) = 0,7721$$

Sonuçta bu kaydın yaklaşık % 77 ihtimalle saldırı olduğu söylenebilir. Modelde eşik(cutoff) değeri % 50 olduğundan ve $77 > 50$ olduğundan model, sonuç için % 100 doğrulukla bu bir saldırı demektir.

7.6 MODELİN TEST VERİ SETİNDE UYGULAMASI

Aşağıdaki Çizelge 7.6 ise test için kullanılan veri setinde (KddCupp-99 veri seti'nin % 5'ini içeren yaklaşık 250bin kayıt) modelin uygulanması ile elde edilen sonuçları göstermektedir. Model test verileri üzerinde gerçekte saldırı olan kayıtları % 99,99 oranında doğru saptayabilirken, gerçekte saldırı olmayan kayıtlarda ise % 42 oranında doğru sonuç üretebilmektedir.

Çizelge 7.6. Test verisi sınıflandırma tablosu.

Gerçek	Tahmin		
	label		Doğruluk Yüzdesi
	0	1	
label 0	10912	15141	42
label 1	11	217957	99,9
Toplam Yüzde			93,8
Tahmin $\geq 0,5$ ise saldırı kabul edildi.			

Yukarıdaki tablodaki verileri elde etmede kullanılan SQL komutları aşağıda görülebilir ("ihtimal" adındaki yeni açılan kolonda modelin ürettiği sonuç depolanmaktadır):

```
UPDATE yuzdeon_son250bin SET ihtimal = 1/(1+exp( -1* (
2.843 + protocol_type*(-7,133) + service*(-1,372 ) + flag*6,882 + hot*6,454 +
num_failed_logins*3,803 + su_attempted*2,610 + num_access_files*(-0,267) )) )
```

```
SELECT count(*) FROM yuzdeon_son250bin WHERE ihtimal >= 0,5
AND label = 1
```

Sonuç: **217957**

```
SELECT count(*) FROM yuzdeon_son250bin WHERE ihtimal >= 0,5
AND label = 0
```

Sonuç: **15141**

```
SELECT count(*) FROM yuzdeon_son250bin WHERE ihtimal < 0,5
AND label = 0
```

Sonuç: **10912**

```
SELECT count(*) FROM yuzdeon_son250bin WHERE ihtimal < 0,5
AND label = 1
```

Sonuç: **11**

```
SELECT count(*) FROM yuzdeon_son250bin
```

Sonuç: **244021**

Toplam hata oranı: $(15141+11)/244021 = 0,0621$

Toplam doğru oranı: $(10912 + 217957) / 244021 = 0,9379$

Gerçekte saldırı olmayan kayıtlarda doğruluk oranı:

$10912/(10912 + 15141) = 0,4188$

Gerçekte saldırı olan kayıtlarda doğruluk oranı:

$217957/(217957 + 11) = 0,9999$

BÖLÜM 8

SONUÇLAR

Sonuç olarak denilebilir ki modelin analiz ettiği bir kayıt gerçekte bir saldırı ise bunun % 99'un üzerinde bir ihtimalle saldırı olduğu bulunabilmektedir. Bu oran çok yüksek bir başarı oranıdır. Fakat gerçekte saldırı olmayan bir kaydın analizinde modelin başarı oranı biraz düşük (modelin bulunduğu eğitim veri setinde % 84. Test veri setinde ise % 42). Bu da modelin sürekli saldırılara maruz kalan, fakat güvenlik seviyesi çok yüksek olması gereken, hiçbir saldırıya tahammülü olmayan, yanlış alarmlarla (false-pozitif) uğraşacak yeterli elemanı olan, kritik öneme sahip ağ işletim merkezleri için son derece uygun bir model olduğunu göstermektedir.

İleriki çalışmalarda günümüz internet trafiği verilerinin özelliklerini de dikkate alıp bu modeli geliştirerek bunu kullanan bir yazılım üretilip gerçek hayatta kullanılabilir, ya da bu çalışmadaki ile aynı veri seti ve parametreleri kullanarak yapay sinir ağları ile de bir model oluşturup iki modelin karşılaştırılması sağlanabilir.

KAYNAKLAR

1. Mok M. S., Sohn S. Y., Ju Y. H., "Random effects logistic regression model for anomaly detection", Department of Information and Industrial Engineering, Yonsei University, **Expert Systems with Applications** 37 (2010): 7162–7166 (2010).
2. Yu K. M., Wu M. F., Wong W. T., "Protocol-Based Classification for Intrusion Detection" Chung Hua University, **Wseas Transactions On Computer Research** 3 (3): 135-141 (2008).
3. Gates C., McNutt J. J., Kadane J. B., and Marc I. Kellner M. I., "Scan Detection on Very Large Networks Using Logistic Regression Modeling", Carnegie Mellon University, **Proceedings of the 11th IEEE Symposium on Computers and Communications (ISCC'06)** , Pittsburgh, 14 (2006).
4. Dagon D. et al., "HoneyStat: Local Worm Detection Using Honeypots", **Recent Advances in Intrusion Detection**, 3224: 39-58 (2004).
5. İnternet: Anonim, "Ağ Güvenliği", <http://tr.docdat.com/docs/index-99198.html> (2013).
6. Tan H., Aktaş Z. A., "Bir Kuruluşun Bilgi Sistemi Güvenliği İçin Bir Yaklaşım", **4. Ağ Ve Bilgi Güvenliği Sempozyumu**, Ankara, 34-40 (2011).
7. Karaaslan E., Akın G., Demir H., "Kurumsal Ağlarda Zararlı Yazılımlarla Mücadele Yöntemleri ", **Akademik Bilişim 2008 Bildiriler**, Çanakkale, 225-230 (2008).
8. İnternet: Karaaslan E., "Kampüs Ağ Yönetimi", Ege Üniversitesi BİTAM Kampüs Network Yönetim Grubu, <http://www.karaaslan.net/bildiri/KampusAgYonetimi.pdf> (2006).
9. İnternet: Anonim, "OpenVpn Kurulumu", <http://www.turkcenet.org/el-dosyalar-mainmenu-38/openvpn-kurulumu-mainmenu-72.html?showall=1> (2013).
10. İnternet: Gönen S., "Özel Sanal Ağ ve Servis Kalitesi (VPN- Virtual Private Network QoS- Quality of Service)" <http://okul.selyam.net/docs/index-23409.html> (2013).
11. İnternet: Anonim, "Ağ Güvenliği Nasıl Sağlanır", <http://www.redbilisim.com/sayfa.aspx?id=42> (2013).

12. İnternet: Kaya S., "Vpn bağlantısı kurma ip adresini farklı gösterme(Virtual Private Network)", <http://secotech.wordpress.com/> (2013).
13. İnternet: Anonim, "How VPN Works" <http://www.alliancedatacom.com/how-vpn-works.asp> (2011).
14. İnternet: Cisco Support Community, "Cisco PIX 500 Series Security Appliances", http://www.cisco.com/en/US/products/hw/vpndevc/ps2030/products_configuration_example09186a008046f307.shtml (2013).
15. Kömür Y. S., Ayfer C. U., "Linux ile Ağ Yönetimi", Bilkent Üniversitesi, Bilgisayar Merkezi, *Akademik Bilişim 2007 Bildiriler Kitabı*, Kütahya, 253-260 (2007).
16. İnternet: Anonim, <http://cehturkiye.com/fw-dmz.png> (2011).
17. İnternet: Altun H., "FortiGate 110C Firewall Nasıl Kurulur Ayarları Nasıl Yapılır" <http://www.mshowto.org/fortigate-110c-firewall-nasil-kurulur-ayarlar-nasil-yapilir.html> (2013).
18. İnternet: Virtual Graffiti Inc., <http://www.avfirewalls.com/images/FortiGate/deployment-enterprise.gif> (2011).
19. İnternet: Seral D., "Kuzey Kıbrıs Geniş Bant Kullanıcılarının Davranışları", Bilişim Sistemleri Mühendisliği Bölümü, Uluslararası Kıbrıs Üniversitesi, http://www.emo.org.tr/ekler/8dd5e44f08b1423_ek.pdf (2010).
20. Güven E. N., Sağiroğlu Ş., "Saldırı Tespit Sistemleri Üzerine Bir İnceleme", **3. Uluslararası Katılımlı Bilgi Güvenliği Ve Kriptoloji Konferansı Bildiriler Kitabı**, Ankara, 273-278 (2008).
21. Seral D., Çalışkan B., "Pasif Ağ Verileri Üzerinden Düzensizlik Tespiti", **3. Uluslararası Katılımlı Bilgi Güvenliği Ve Kriptoloji Konferansı**, Ankara, 152-156 (2008).
22. İnternet: Interactive Systems "Intrusion Detection" <http://www.interactivesys.net/intrusion-detection.html> (2011).
23. İnternet: Şen A., "Parametrik Olmayan İstatistiksel Teknikler", www.deu.edu.tr/userweb/ali.sen/dosyalar/HAFTA1.pdf (2011).
24. İnalı M. E., Topuz D., Uçan O., "Doğrusal Olasılık ve Logit Modelleri ile Parametre Tahmini", *Hacettepe Üniversitesi Sosyoekonomi Dergisi* 2006-1 49-51 (2006).
25. İnternet: Demir M., "Analitik Verilerin Değerlendirilmesi", <http://web.adu.edu.tr/user/mdemir/K214/K21402analitikveridegerlendirme090301.pdf> (2009).

26. İnternet: Caymaz C., "Regresyon analizi", <http://www.makaleler.com/egitim-makaleleri/regresyon-analizi.htm> (2013).
27. İnternet: Legendre, A. M., "Nouvelles méthodes pour la détermination des orbites des comètes - Sur la Méthode des moindres carrés" www.bibnum.education.fr/math/C3%A9matiques/alg/C3%A9bre/legendre-et-la-m%C3%A9thode-des-moindres-carr%C3%A9s# (2013).
28. İnternet: School of Mathematics and Statistics University of St Andrews, "Johann Carl Friedrich Gauss", [www-history.mcs.st-and.ac.uk/Biographies/Gauss.html](http://www.history.mcs.st-and.ac.uk/Biographies/Gauss.html) (2013).
29. İnternet: Xavier University, "Carl Friedrich Gauss - Theoria Combinationis Observationum Erroribus Minimis Obnoxiae", www.cs.xu.edu/math/Sources/Gauss/Gauss.html (2013).
30. İnternet: Yalta A. T., Bağlanım Çözümlemesi Temel Kavramlar, http://www.acikders.org.tr/pluginfile.php/2261/mod_resource/content/2/ekonometri1-tuba-06-baglanim-cozumlemesi-temel-kavramlar.pdf (2013).
31. İnternet: Yalta A. T., Ekonometri 1 Ders Notları, <http://yalta.etu.edu.tr/files/ekonometri1-kitap-surumu-%28s2,0%29.pdf> (2013).
32. İnternet: Anonim, "George Udny Yule Presentation", <http://www.aprender-mat.info/turquia/historyDetail.htm?id=Yule> (2013).
33. İnternet: Otrar M., "İstatistik Biliminin Kısa Tarihi", www.mustafaotrar.net/teknikler/001_ist_tarihi.htm (2013).
34. Fisher, R.A., "The goodness of fit of regression formula, and the distribution of regression coefficients", *J. Royal Statist. Soc.*, 85: 597-612 (1922).
35. Fisher, R.A., "Statistical Methods For Research Workers", 5, *Cosmo Publications*, New Delhi, 96-109 (2007).
36. İnternet: Gültekin F., "Regresyon Analizi" www.fikretgultekin.com (2012).
37. İnternet: Golayoglu A., "En Küçük Kareler Yöntemi", http://www.baskent.edu.tr/~afet/dersler/genel_matematik_2/dersnotlari_listesi/ DERS_%207.pdf (2011).
38. İnternet: İstatistik Analiz, "Regresyon Analizi", http://istatistikanaliz.com/regresyon_analizi.asp (2012).
39. İnternet: Anonim, "Korelasyon", http://www.akademikdestek.net/info/korelasyon_regresyon.doc (2012).
40. İnternet: Tan M., Karadağ M., "Atom Fiziği Laboratuvarı Deney Klavuzu", <http://gef-ortaogretim-fenmatematik-fizik.gazi.edu.tr/posts/download?id=28993> (2013).

41. İnternet: Dariga A. İ., En Küçük Kareler Yöntemi www.matemasuk.com (2013).
42. İnternet: Anonim, "Differentiation And The Derivative", <http://en.wikipedia.org/wiki/Derivative> (2011).
43. Coşkun S., Kartal M., Bircan H., Coşkun A., “Lojistik Regresyon Analizinin İncelenmesi ve Diş Hekimliğinde Bir Uygulaması” *Cumhuriyet Üniversitesi Diş Hekimliği Fakültesi Dergisi*, 7 (1): 42-44 (2004).
44. Girginer N., Cankuş B., “Tramvay Yolcu Memnuniyetinin Lojistik Regresyon Analiziyle Ölçülmesi: Estram Örneği” *Celal Bayar Üniversitesi İ.İ.B.F Yönetim Ve Ekonomi Dergisi* 15 (2008-1): 185-187 (2008).
45. Agresti, A., “An Introduction to Categorical Data Analysis”, 423, *John Wiley and Sons Inc.*, New Jersey, 103 (1996).
46. Işığışok, E., “Bebeklerin Doğum Ağırlıklarını ve Boylarını Etkileyen Faktörlerin Lojistik Regresyon Analizi İle Araştırılması”, *VI. Ulusal Ekonometri ve İstatistik Sempozyumu Bildiri Kitabı*, Gazi Üniversitesi, Ankara, 3 (2003).
47. Altıntaş T., “Yapay Bağımlı Değişkenli Tahmin Modelleri ve Bir Uygulama”, Yüksek Lisans Tezi, *Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü*, Samsun, 13-17 (2006).
48. Aydın N., “Gölge Değişkenlerin Kullanımı Ve Çeşitli Uygulamalar”, Yüksek Lisans Tezi, *İstanbul Üniversitesi Sosyal Bilimler Enstitüsü*, İstanbul, 3 (2005).
49. Altıntaş T., Sağlam V., Uslu V. R., "Logit Ve Probit Analizle Bir Marka Bağımlılığı Araştırması", TÜİK 16. İstatistik Araştırma Sempozyumu Bildiriler Kitabı, *Türkiye İstatistik Kurumu* 143-144 (2007).
50. İnternet: Sumbüloğlu K., "Lojistik Regresyon Analizi" http://78.189.53.61/-/bs/ess/k_sumbuloglu.pdf (2011).
51. Kaptan Ö. B., “Bireysel Kredilerin Risk Ve Tüketici Davranışı Açısından Analizi”, Doktora Tezi, *Ankara Üniversitesi Sosyal Bilimler Enstitüsü*, Ankara, 174 (2011).
52. İnternet: Büyükkara G., "Kredi Riski", http://www.acikders.org.tr/pluginfile.php/3496/mod_resource/content/2/Kredi_Riski.pdf (2012).
53. Polat F., “Yemeklik Yağ Sektöründe Tüketici Davranışlarını Etkileyen Faktörlerin Analizi”, Doktora Tezi, *Ankara Üniversitesi Fen Bilimleri Enstitüsü*, Ankara, 28-31 (2011).
54. Özdamar, K., “Paket Programlar ile İstatistiksel Veri Analizi”, *Kaan Kitabevi*, Eskişehir, 1: 475-477 (2002).

55. Karahan S., “Multinomial Logit Modeller Ve Bir Uygulama”, Yüksek Lisans Tezi, *Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü*, Ankara, 5-7, 25-27 (2005).
56. İnternet: Aydın M. A., Örencik B., “Bilgisayar Ağlarında Saldırı Tespiti için İstatistiksel Yöntem Kullanımı Ve Bir Karma Saldırı Tespit Sistemi Tasarımı”, http://www.emo.org.tr/ekler/3eaa6722798a773_ek.pdf (2013).
57. İnternet: Erol M., “Saldırı Tespit Sistemlerinde İstatistiksel Anormallik Belirleme Kullanımı” <http://web.itu.edu.tr/~orencik/SizmaBelirlemedeAnormallikTespitiKullanimi.pdf> (2005).



ÖZGEÇMİŞ

İdris BUDAK 1981 yılında Konya ili Akşehir ilçesinde doğdu; orta öğrenimini Akşehir Anadolu Lisesi'nde tamamladı. Liseye Konya Meram Fen Lisesi'nde başlayıp, Akşehir Selçuklu Lisesi'nde bitirdi. 1999 yılında Ege Üniversitesi Bilgisayar Mühendisliği Bölümü'nde öğrenime başlayıp 2003 yılında mezun oldu. 2004-2008 yılları arasında bilgisayar yazılım üreticisi olarak çalıştı. 2009 yılında Karabük Üniversitesi Bilgi İşlem Daire Başkanlığı'nda sistem-ağ sorumlusu olarak göreve başladı ve halen aynı yerde ağ sorumlusu olarak çalışmaya devam etmektedir. 2013 yılında Karabük Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda yüksek lisans programını tamamladı.

ADRES BİLGİLERİ

Adres : Karabük Üniversitesi
Bilgi İşlem Daire Başkanlığı
Balıklarkayası Mevkii / KARABÜK

ÇALIŞMANIN BİLDİRİ OLARAK SUNULDUĞU YERLER

1. Budak İ., Şen B., Yıldırım M. Z., "Lojistik Regresyon ile Bilgisayar Ağlarında Anomali Tespiti", *Akademik Bilişim 2013*, Antalya (2013).