Pamukkale Univ Muh Bilim Derg, 24(2), 276-282, 2018

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Kimlik avı web sitelerinin tespitinde makine öğrenmesi algoritmalarının karşılaştırmalı analizi

Comparative analysis of machine learning algorithms in detection of phishing websites

Muhammed Ali KOŞAN[1*], *Oktay YILDIZ*[2], *Hacer KARACAN*[3]

[1]Bilgisayar Bilimleri Anabilim Dalı, Gazi Üniversitesi, Ankara, Türkiye.
ceo.muhammed@gmail.com

[2,3]Bilgisayar Mühendisliği Bölümü, Gazi Üniversitesi, Ankara, Türkiye.
oyildiz@gazi.edu.tr, hkaracan@gazi.edu.tr

Geliş Tarihi/Received: 10.01.2017, Kabul Tarihi/Accepted: 17.05.2017
doi: 10.5505/pajes.2017.27167
* Yazışılan yazar/Corresponding author
Araştırma Makalesi/Research Article

Öz

Web uygulamalarının kullanım oranındaki artış ile birlikte sayısı artan kötücül web siteleri ve saldırılar, son kullanıcıya ciddi zararlar vermektedir. Kişisel ve hassas bilgilerin çalınmasına yönelik bu saldırılardan biri Kimlik Avı saldırısıdır. Yayımlanan güvenlik raporlarında son yıllarda milyonlarca yeni kimlik avı sahteciliği yapan web sayfası tespit edildiği ifade edilmektedir. Böylesi kritik bir durumda bu web sayfalarının tespiti büyük önem arz etmektedir. Bu çalışmada, bir veri kümesi ile birlikte literatürde bulunan makine öğrenmesi sınıflandırma algoritmaları kullanılarak karşılaştırmalı analiz yapılmıştır. Analiz sonuçları, Kimlik Avı Sahteciliği çalışmalarında kullanılan sınıflandırma algoritmalarının hangi koşullarda tercih edilmesi gerektiği hakkında farklı parametreler bulunduğunu göstermektedir.

Anahtar kelimeler: Kimlik avı saldırıları, Makine öğrenmesi, Sınıflandırma algoritmaları, Değerlendirme ölçütleri

Abstract

The increasing number of malicious web sites and attacks, along with the increase in the usage rate of web applications, cause severe damage to the end user. One of these attacks aimed at stealing personal and sensitive information is the Phishing Attack. In the published security reports, it is stated that in recent years there has been millions of web pages that have made new phishing scams. In such a critical situation, the identification of these web pages is of great importance. In this study, a comparative analysis was made on a mentioned dataset using machine learning classification algorithms in the literature. The results of the analysis show that the classification algorithms used have different parameters about which conditions should be preferred in the studies on Phishing Fraud.

Keywords: Phishing attacks, Machine learning, Classification algorithms, Assessment measures

1 Giriş

Günümüzde web sitelerinin sayısı büyük boyutlara ulaşmıştır. Bunun yanı sıra artan bu web site sayısı kötücül amaçlı kişileri de harekete geçirmiştir. Saldırganlar kötü amaçlı web siteleri ile internet kullanıcılarına zarar vermektedirler. Her yıl bir çok web sitesinin yeni yayına geçtiği bir ortamda, böylesine büyük bir potansiyeli değerlendirmek isteyen saldırganlar kötü amaçlı web sitelerini farklı şablonlarda kullanıcılara çekici kılarak sunmaktadırlar. Bu siteler ile ilgili McAfee'in raporuna göre [1], 2014 yılının 3. çeyreğinde 30 milyonu aşkın yeni kötücül URL gözlemlenmiş ve 2014 yılı boyunca yaklaşık 80 milyon yeni kötücül URL tespit edilmiştir. Symantec ise yayınladığı raporda [2] bu analizi alan adına (domain) göre gerçekleştirmekle birlikte 2014 yılı içerisinde 29.927 alan adının eşsiz kötücül olarak saptandığı sunulmaktadır. McAfee tarafından yapılan 2016 raporunda [3] ise 2015 yılı boyunca 100 milyona yakın yeni kötücül URL tespit edilmiştir. Ayrıca 2015 yılının dördüncü çeyreğinde 1.5 milyona yakın yeni kimlik avı URL'i gözlemlenmiştir. Bunların 1 milyona yakını ilişkili alan adlarından oluşmaktadır. Bu raporlar, her geçen yıl kötücül web siteleri ile birlikte Kimlik Avı web sitelerinde artış olduğunu ve internet kullanıcılarına bu yöntemler ile zarar vermek isteyen saldırganların çok fazla aktivite gösterdiğini ortaya koymaktadır. Literatürde bu alanda çözüm üretmek amacıyla çıkan bir çok çalışma bulunmakla birlikte, güvenlik raporları halen kesin bir çözüm elde edilemediğini göstermektedir [1]-[3].

Bu çalışmada, Kimlik Avı web sayfalarının tespiti için makine öğrenmesi algoritmaları karşılaştırmalı olarak analiz edilmiş ve değerlendirilmiştir. Böylelikle başarım oranlarına göre karşılaştırılan makine öğrenmesi algoritmalarının, Kimlik Avı web sayfalarından korunum için yapılacak çalışmalara referans olması amaçlanmaktadır. UCI Makine Öğrenme Deposu (Machine Learning Repository) içinde bulunan Kimlik Avı Sahteciliği Yapan Web Site Veri Kümesi (Phishing Websites Data Set) [4] kullanılmıştır. C4.5, ID3, PRISM, RIPPER, Naif Bayes (Naive Bayes), k-En Yakın Komşu (k-nearest neighbor, KNN), Rastgele Ormanlar (Random Forests) algoritmaları veri kümesi ile uygulanarak, veri kümesi üzerindeki performans ve doğruluk karşılaştırmaları analiz edilmiştir. Bu çalışma beş bölümden oluşmaktadır. İkinci bölümde, konu ile ilgili literatür çalışmaları sunulmuştur. Üçüncü bölümde deney sürecinde kullanılan araç ve yöntemlerden bahsedilmiştir. Hemen ardından deneyler ve analiz, dördüncü bölüm içinde anlatılmıştır. Son olarak beşinci bölümde yapılan çalışma sonrasında elde edilen sonuçlara yer verilmiştir.

2 Literatür İncelemesi

Kazemian ve Ahmed [5], yaptıkları çalışmada 100.000 web sayfasını bir örümcek (crawler/spider) aracılığı ile indirerek, web sayfalarının özellik vektörüne dönüşümünü gerçekleştirmişlerdir. Web Uygulama Sınıflandırıcısı (Web

Application Classifier-WAC) adı verilen araç ile bu vektörler giriş olarak alınıp Makine Öğrenmesi algoritmaları uygulanmıştır. Uygulanan Makine Öğrenmesi algoritmaları sırası ile; k-En Yakın Komşu, Doğrusal Destek Vektör Makineleri (Support Vector Machines, SVM), Radyal Tabanlı Fonksiyon (Radial Basis Function-RBF) Çekirdekli Destek Vektör Makineleri ve Naif Bayes'dir. Web sayfaları 50, 100, 500, 5.000 ve 100.000 sayıları ile 4 algoritma bazında test edilmiş ve en iyi sonuç Radyal Tabanlı Fonksiyon Çekirdekli Destek Vektör Makineleri (RBF-SVM) algoritması ile alınmıştır.

Li ve diğ. [6], Minimum Kapsayan Top-tabanlı Destek Vektör Makinesi (Minimum Enclosing Ball-based Support Vector Machine - BVM) adını verdikleri çalışmada ilk olarak veri kümesi oluştururken 12 özelliğe göre özellik vektörleri çıkartılmıştır. Sonrasında Destek Vektör Makineleri ve önerilen yöntem olarak Top-tabanlı Destek Vektör Makinesi'nin karşılaştırmalı analizi gerçeklenmiştir. Yapılan deneyler sonucunda Top-tabanlı Destek Vektör Makinesi'ninDestek Vektör Makineleri'ne göre doğruluk oranı ve performans konusunda daha iyi sonuçlar verdiği gözlemlenmiştir.

Moghimi ve Varjani [7], internet bankacılığında kimlik avı (phishing) saldırılarını tespit etmek için kural bazlı yeni bir yöntem sunmaktadır. Önerdikleri özellik kümesi, sayfa kaynağı kimliğini değerlendirmek için 4 özellik ve sayfa kaynak elemanlarının erişim protokollerini tanımlamak için 4 özellik içermektedir. Önerilen özellik kümesindeki sayfaların URL'leri ile içerikleri arasındaki ilişkiyi belirlemek için string eşleştirme algoritması kullanılmıştır. Yapılan deneylerde, internet bankacılığındaki kimlik avı sahteciliği sayfalarını tespit etmek için önerilen yöntemin %99.14 doğruluk oranına (true-positive) sahip olduğu gözlemlenmiştir.

URL tabanlı bir yaklaşım uygulanan çalışmada [8], 9.661 kimlik avı sahteciliği yapan web sitesi ile birlikte 1.000 adet iyicil web sitesinden oluşturulan bir veri kümesi kullanılmıştır. URL'in birincil adı, alt alan adı, URL yolu ile birlikte Alexa ve Google gibi web site sıralama hizmetlerindeki sıra bilgisi üzerinden özellikler sınıflandırılmıştır. Her özellik ağırlıklandırılarak sisteme dahil edilmekle birlikte elde edilen değerler için kıyaslama yapılacak sınır limitleri belirlenmiştir. Yapılan çalışmada, gerçek zamanlı bir kimlik avı sahteciliği tespiti yapan model tasarımı sunulmuş ve 9.661 veri kümesi ile eğitilen sistem üzerinde 1.000 kimlik avı sahteciliği içeren web sitesi ile birlikte 1.000 adet iyicil web sitesi test edilmiştir. Yapılan testler ile birlikte %97 doğruluk oranı elde edilmiştir.

Mohammad ve diğ. [9], ilk olarak kimlik avı sitelerinin özelliklerini değerlendirmek için 2500 adet kimlik avı sahteciliği adresini Phishtank [10] arşivinden almışlardır. Sonrasında, javascript ile yazdıkları bir program ile tüm özellik tanımları, web sayfaları üzerinde bazı parametreler temel alınarak elde edilmiştir. Elde edilen özellikler, URL'in bazı zararlı veriler içerip içermediği ve web sayfasında bulunan zararlı olabilecek kod ve yapılara göre oluşturulmuştur. Daha sonra, bu veri kümesinden 17 özellik kullanılarak kendi kendini yapılandıran yapay sinir ağları ile kimlik avı saldırılarını tahmin eden bir sistem [11] tasarlamışlardır. Oluşturulan sistem, gürültülü veri, hata toleransı ve yüksek tahmin doğruluğu için yüksek kabul edilebilirlik sağlamaktadır. Yapılan deneylerde veri kümesi olarak, 600 iyicil ve 800 kimlik avı sahteciliği yapan web sitesi toplanmıştır. Yapay sinir ağında giriş olarak 17 nöron, ilk katmanda 3 nöron ve ikinci katmanda ise 1 nöron kullanılmıştır. Geliştirilen model, kimlik avı sahteciliğinde değişen web sitelerindeki farklı özelliklerin farklı önemlerde olmasından kaynaklı ortaya çıkan problemi çözmek amacıyla ortaya çıkmıştır. Bir diğer çalışmalarında [12], C4.5, RIPPER, PRISM ve CBA (Apriori algoritmasının bir uygulaması olan kural tabanlı sınıflandırma algoritması) algoritmalarını veri kümesi üzerinde test etmişlerdir. Phishing Websites [4] veri kümesinden 450 kimlik avı sahteciliği yapan URL ile birlikte 450 iyicil URL'den oluşan bir veri kümesi kullanılarak, algoritmaların sınıflandırma performansları karşılaştırılmıştır. Yapılan deneysel çalışmalar sonucunda C4.5 algoritması diğer kural tabanlı sınıflandırma algoritmalarına nazaran daha iyi sonuçlar vermiştir. C4.5 algoritması %5.76 ortalama hata oranı göstermek ile birlikte, RIPPER %5.94 ve en yüksek hata oranı ile PRISM %21.24 ortalama hata oranına sahip olarak gözlemlenmiştir.

Phishing Websites [4] veri kümesini kullanan bir diğer çalışmada [13] ise yapay sinir ağları ile kimlik avı sahteciliği yapan web sitelerinin tespitine yönelik deneyler yapılmıştır. Elde edilen sonuçların, literatürde ortaya konan sonuçlara göre daha az hata oranına sahip olduğu söylenmektedir. Ek olarak yapılan çalışmalar ile kıyaslandığında performans açısından daha iyi sonuçlar sunduğu anlatılmaktadır. Aynı veri kümesi ile yapılan özellik çıkartım ve sınıflandırma algoritmaları ile yapılan deneylerde [14] ise nitelik seçim algoritmalarından Korelasyon-bazlı Öznitelik Seçme (Correlation Feature Selection, CFS), Bilgi Kazancı (Information Gain, IG), Tutarlılık Alt Kümesi (Consistency Subset, CS) ve Temel Bileşenler Analizi (Principal Component Analysis, PCA) kullanılmakla birlikte sınıflandırma algoritmalarından J48, Naif Bayes, Destek Vektör Makineleri, Rastgele Orman (Random Forest) ve AdaBoost kullanılmıştır. Her sınıflandırma algoritması tüm nitelik seçim algoritmaları ile birlikte kullanılarak algoritmaların hepsi karşılaştırılmıştır. Elde edilen sonuçlar, en yüksek doğruluk değerlerine Rastgele Orman sınıflandırma algoritmasının sahip olduğunu göstermektedir.

3 Materyal ve yöntem

Bu bölümde, uygulamada kullanılan araç ve yöntemlerden bahsedilmiştir.

3.1 Veri kümesi

Çalışmada, veri kümesi olarak Phishing Websites [4] kullanılmıştır. Veri kümesinde 30 öznitelik (1 öznitelik ise sınıflandırma için ek olarak bulunmaktadır), 2456 kimlik avı sahteciliği barındıran örnekten türetilmekte ve toplam 11055 kayıttan oluşmaktadır. Veri kümesi 4898 kimlik avı sahteciliği olarak sınıflandırılmış örnek ve 6157 iyicil olarak sınıflandırılmış örnek bulundurmaktadır. 26 mart 2015 tarihinde UCI havuzuna bağışlanmış yeni bir veri kümesidir. Dört temel başlık altında özellikler kategorilendirilmiştir. Bu kategoriler sırası ile adres alanı-tabanlı özellikler, anormal tabanlı özellikler, HTML ve JavaScript tabanlı özellikler, alan adı tabanlı özelliklerdir. Ayrıca her nitelik kendi içinde İyicil (1), Şüpheli (0) ve Kimlik Avı (-1) özelliklerinden en az ikisini barındıracak şekilde ayırt edilmiştir. Nitelikler sadece 1, 0 ve -1 değerlerini almaktadır. Söz konusu niteliklerin tamamının seçilmesinde, uygulanacak algoritmaların belirlenen veri kümesindeki bütün niteliklerin dahil edildiği bir veri kümesi üzerinde nasıl hareket ettiğini gözlemleme ihtiyacı etkili olmuştur. Veri kümesinin içeriğinde bulunan özniteliklere Tablo 1'de yer verilmiştir. Ayrıca Korelasyon-bazlı Öznitelik Seçme (CFS) yöntemi ile belirlenen önem derecesi yüksek öznitelikler aşağıda açıklanmıştır.

Tablo 1: Phishing websites [4] veri kümesi öznitelikleri.

Having_IP_Address	Submitting_to_email
URL_Length	Abnormal_URL
Shortining_Service	Redirect
having_At_Symbol	on_mouseover
double_slash_redirecting	RightClick
Prefix_Suffix	popUpWidnow
having_Sub_Domain	Iframe
SSLfinal_State	age_of_domain
Domain_registeration_length	DNSRecord
Favicon	web_traffic
port	Page_Rank
HTTPS_token	Google_Index
Request_URL	Links_pointing_to_page
URL_of_Anchor	Statistical_report
Links_in_tags	Result
SFH	

- ✓ Prefix_Suffix: Alan adı ön ek veya son ek ile ayrılmışmı (- işareti ile ayrılmışmı) (Alan adı '-' işareti içeriyor ise -1, içermiyor ise 1),
- ✓ having_Sub_Domain: URL'de alt alan adı veya çoklu alt alan adı kullanılmışmı (Alan adı kısmında nokta işareti sayısı 1 ise 1, nokta işareti sayısı 2 ise 0, diğer durumlarda ise -1),
- ✓ SSLfinal_State: HTTPS protokolünün varlığı (https kullanılıyor ve sağlayıcı güvenilir ve sertifika yaşı 1 yıl ve üstü ise 1, https kullanılıyor ve sağlayıcı güvenilir değil ise 0, diğer durumlarda -1),
- ✓ Request_URL: Web sayfası farklı alan adlarından farklı nesneleri çekiyormu (Harici web sitelerinden çekilen nesne istek URL'lerinin yüzdesi 22'den az ise 1, yüzdesi 22 ve 61 arasında ise 0, diğer durumlarda ise -1),
- ✓ URL_of_Anchor: HTML çapa etiketi (<a> etiketi) kullanımının URL'de varlığı (Çapa etiketlerindeki URL varlığının yüzdesi 31 değerinin altında ise 1, yüzdesi 31 ve 67 arasında ise 0, diğer durumlarda -1),
- ✓ Links_in_tags: Meta, Script ve Link etiketlerindeki bağlantıların oranı (Meta,Script ve Link etiketlerindeki bağlantıların yüzdesi 17 değerinden az ise 1, yüzdesi 17 ve 81 arasında ise 0, diğer durumlarda -1),
- ✓ SFH: Sunucu Form Tutucusu/İşleyicisi (Server Form Handler) içinde boş string değişkeni barındırılıyormu (SFH 'about:blank' veya string değeri boş ise -1, farklı bir alan adına referans veriyor ise 0, diğer durumlarda 1),
- ✓ web_traffic: Alexa web site trafik durumu (Alexa sırası 100.000 altında ise 1, üstünde ise 0, hiç liste de yok ise -1),
- ✓ Google_Index: Google tarafından indeksleniyormu (Google tarafından indeksleniyor ise 1, diğer durumda -1),
- ✓ Result: Son parametre ise kimlik avı olarak işaretlenip işaretlenmediğini belirten sınıf alanıdır. (Kimlik avı sahteciliği olarak işaretlenmiş ise -1, iyicil olarak işaretlenmiş ise 1 değeri alır),

Veri kümesinin işlenmesi ve sınıflandırma algoritmalarının test edilmesi için Weka makine öğrenmesi algoritmaları aracı [15] kullanılmıştır.

3.2 Sınıflandırma algoritmaları

Bu çalışmada, Weka aracını kullanarak veri seti üzerinde sınıflandırma algoritmalarının karşılaştırmalı analizi yapılmıştır. Kullanılacak sınıflandırma algoritmaları C4.5, ID3, PRISM, RIPPER, Naif Bayes, k-En Yakın Komşu (KNN) ve Rastgele Ormanlar (RF)'dır. Söz konusu algoritmaların özelliklerine aşağıda kısaca yer verilmiştir.

C4.5: Karar ağacı oluşturarak sınıflandırma amaçlı kullanılan bir algoritmadır. J48 olarak Weka aracı üzerinde açık kaynak uygulanmıştır. Ağaç oluşturma işleminde tüm özellikler kontrol edilerek bilgi kazanç (information gain) değerleri hesaplanır. En iyi bilgi kazancı değeri ağaçta karar noktası olarak işlenir. Sonrasında bu işlem ağaçların alt dalları oluşturulacak şekilde tüm ağaç yapısı oluşana kadar devam ettirilir. Ağaç oluşumu sonrasında budama işlemi gerçekleştirilebilir. Budama işlemi ile gereksiz görülen dallar ağaç yapısından çıkartılır. Böylelikle karar işlemlerinde gereksiz işlemlerden kurtulmuş olunur [16].

ID3: Karar ağacı oluşturarak sınıflandırma amaçlı kullanılan ve C4.5'in öncülü olan bir algoritmadır. C4.5 algoritmasından en büyük farkı budama işleminin yapılmamasıdır. Algoritmanın temelinde, ağacın oluşturulmasında ağaca dahil edilmemiş niteliklerin entropi değerleri hesaplanarak en küçük entropi değerine sahip niteliğin karar ağacına dahil edilmesi bulunmaktadır [17].

PRISM: Kural tabanlı olarak geliştirilmiştir. Eğitim kümesindeki örneklerin sınıfları temel alınarak her sınıf için sınıf değeri dışındaki niteliklerin hangi örnekler tarafından içerildiklerine göre katkı değerleri hesaplanır. Hesaplanan katkı değeri, niteliğin ayrıt edici değerinin sınıfta bulunma sayısı ile kaç örnekte bulunma sayısına oranı üzerinden hesaplanır. Bu katkı değeri en yüksek olan niteliğe göre nitelik değerinin bulunduğu örnekler, örnek kümesinden çıkarılarak diğer kuralların oluşturulması için aynı adımlar tekrarlanır. ID3'den farkı, niteliğin değerine bağımlı bir algoritmadır. ID3 ise niteliğe bağlı olarak oluşturulur. Ayrıca ID3'e göre daha az kural bulunmakta ve daha genel ağaçlar oluşturulmaktadır [18].

RIPPER: IREP algoritmasının geliştirilmesi ile elde edilmiş kural tabanlı sınıflandırma algoritmasıdır. Temel olarak inşa sürecinde kuralların oluşturulup budama işleminin yapılması ile algoritma işletilir. Sonrasında kurallar optimize edilerek yapı daha güvenilir bir hale getirilir. C4.5 kurallarının oluşturulmasında, başlangıçta elde edilen model RIPPER'a göre daha büyük boyutlarda oluşturulmaktadır. Bu bir dezavantaj olarak görülmektedir. Yapılan test sonuçları, hata oranında C4.5 ile hemen hemen yakın değerler saptandığını göstermiştir. Fakat geliştirildiği IREP'e göre büyük gelişim gösterdiği gözlemlenmektedir [19].

Naif Bayes: En yaygın kullanılan sınıflandırma algoritmalarından biri olarak Naif Bayes, olasılık tabanlı bir algoritmadır. Elde bulunan veri kümesi üzerinden her bir niteliğin sistem üzerinde koşullu olasılığı, bu algoritma yardımı ile hesaplanabilmektedir. Bir koşullu olasılığın hesaplanması için Gauss Dağılımı gibi yöntemler yardımı ile istenen niteliğe bağlı olasılıkların her birinin hesaplanarak çarpma işleminin uygulanması gerekir. Son değer ise normalleştirilerek işlemeye uygun hale getirilir [20]. C4.5 algoritması ile karşılaştırıldığı bir çalışmada elde edilen başarı ve başarısızlık oranları birbiri ile uyumlu halde gözlemlenmektedir [21]. Ayrıca 18 veri kümesi

üzerinde yapılan deneylerde elde edilen doğruluk oranları arasında sadece 0.2'lik bir fark ile C4.5 az ara önde doğruluk oranı elde edilmiştir [22].

k-En Yakın Komşu (KNN): Örnek tabanlı bir sınıflandırma algoritmasıdır. Eğitim kümesi üzerinde test edilmek istenen veri için ilk olarak test verisinin diğer eğitim örneklerine olan öklit uzaklıkları hesaplanır. Hesaplanan değerler, mesafeye bağlı olarak en küçükten büyüğe doğru örnekler sıralanır. Test örneğinin sınıflandırılması için en yakın kaç komşuya bakılacağı değeri, *k* parametresi olarak verilir. Sonrasında bakılacak komşu değerleri için, en küçükten büyüğe doğru sıralanan örnek listesinin en başından *k* tanesi alınır. Alınan *k* örneğin bulundukları sınıflar temel alınarak en fazla bulunan sınıf, test örneğinin sınıfı olarak kabul edilir. Bakılacak en yakın komşu değerine göre test örneğinin sınıfı da değişebilir [23]. Üç farklı veri kümesi üzerinde Naif Bayes ve C4.5 algoritmaları ile birlikte uygulanan KNN algoritması, diğer iki algoritmaya göre daha iyi doğruluk sonuçları vermiştir [24].

Rastgele Ormanlar (RF): Karar ağacı türünde yapılandırılıp kullanılan sınıflandırma algoritmasıdır. Algoritmanın temelinde, rastgele özellikler alınarak oluşturulan alt kümelerden tahminlerle verileri bölüp sınıflandırma işlemini gerçekleştirme işlemi bulunmaktadır [25].

3.3 Performans ölçütleri

Sunulan sınıflandırma algoritmaları, k çapraz doğrulama (cross-validation) kullanılarak test edilecektir. Elde edilen sonuçlar ile birlikte Karmaşıklık Matrisi (Confusion Matrix), Doğru Pozitif Oranı (DP Oranı, True Positive Rate-TP Rate), Yanlış Pozitif Oranı (YP Oranı, False Positive Rate-FP Rate), F-Ölçütü (F-Measure), ROC Alanı (ROC Area) ve Doğruluk Oranına (Accuracy) göre algoritmalar karşılaştırılacaktır [26]-[29].

Karmaşıklık Matrisi: Test sonucu elde edilen bilgiler üzerinden, gerçek ve tahmin edilen örneklerin sayılarını temel alan bir matris ile algoritmanın başarısını ölçmeye yarar bir yöntemdir [29]. Tablo 2'de matrisin içeriği gösterilmektedir.

Tablo 2: Karmaşıklık matrisi [29].

		Tahmin Edilen Sınıf	
		Sınıf = 1	Sınıf=0
Gerçek Sınıf	Sınıf=1	DP (Doğru Pozitif)	YN (Yanlış Negatif)
	Sınıf=0	YP (Yanlış Pozitif)	DN (Doğru Negatif)

DP Oranı: Karmaşıklık matrisinden elde edilen bilgilerden yola çıkarak algoritmanın seçilen sınıfa ait doğru tahmin oranını hesaplamak için kullanılan bir yöntemdir. Hesaplama işlemi 1 numaralı formül de gösterilmektedir [29].

$$DPOranı = \frac{DP}{DP + YN} \quad (1)$$

YP Oranı: DP Oranına benzer şekilde karmaşıklık matrisinden elde edilmektedir. Seçilen sınıfın yanlış tahmin oranını hesaplamak için kullanılmaktadır. 2 numaralı formül üzerinden hesaplama işlemi görülmektedir [29].

$$YPOranı = \frac{YP}{YP + DN} \quad (2)$$

F-Ölçütü: Kesinlik (Precision) ve Hassasiyet (Recall) değerlerinin harmonik ortalaması olarak hesaplanmaktadır. 3, 4 ve 5 numaralı formüller ile sırasıyla Kesinlik (K), Hassasiyet (H) ve F-Ölçütü (Fm) değerlerinin nasıl hesaplandığı formüller ile verilmektedir [29].

$$K = \frac{DP}{DP + YP} \quad (3)$$

$$H = \frac{DP}{DP + YN} \quad (4)$$

$$Fm = 2 \times \frac{K \times H}{K + H} \quad (5)$$

ROC Alanı: DP oranı ile YP oranından elde edilen eğri grafiği üzerinden hesaplanan, algoritmaların doğruluğunu ölçmekte kullanılan ölçütlerden biridir. ROC Alan değeri 0 ile 1 arasında değer almak ile birlikte 1 değerine yakınsaması, yapılan testin başarısındaki artışı göstermektedir [26].

Doğruluk Oranı: Tahmin edilmesi istenen verilerin algoritmalar tarafından doğru tahmin edilme oranını ölçmek amacıyla kullanılan yöntemdir [28]. Doğru tahmin edilen örneklerin tüm örneklere oranından elde edilir. 6 numaralı formül ile nasıl hesaplandığı gösterilmiştir.

$$Dogruluk = \frac{DP + DN}{DP + YP + DN + YN} \quad (6)$$

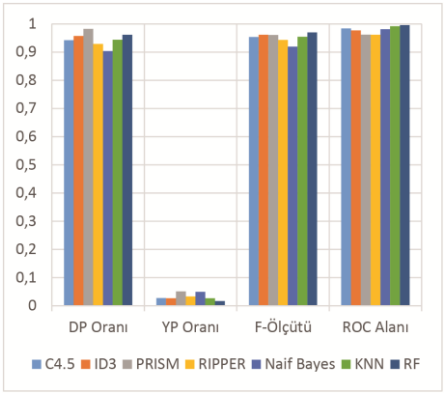
4 Deneysel çalışma ve analiz

Bu bölümde, deneysel ortam oluşturularak elde edilen sonuçlar üzerinden karşılaştırmalı analiz yapılmıştır. Yapılan analizler sonrasında, kimlik avı tespiti için uygulanan sınıflandırma algoritmalarının başarı ve başarısızlık oranları üzerinden değerlendirmeler sunulmuştur.

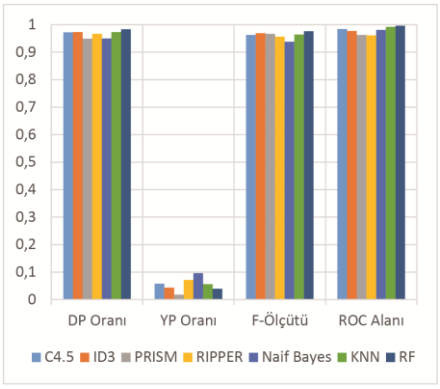
Sınıflandırma algoritmalarının değerlendirme ölçütlerine göre karşılaştırması Tablo 3'te gösterilmiştir. Ayrıca, "Kimlik Avı Sahteciliği" ve "İyicil" sınıf değerlerine göre sınıflandırma algoritmalarının başarı ölçümlerinin karşılaştırması Şekil 1 ve Şekil 2'de ortaya konmuştur. Ek olarak, algoritmaların doğruluk oranları Şekil 3'te sunulmuştur.

Tablo 3: Sınıflandırma algoritmalarının DP Oranı, YP Oranı, F-Ölçütü ve ROC Alanı değerlerine göre karşılaştırma.

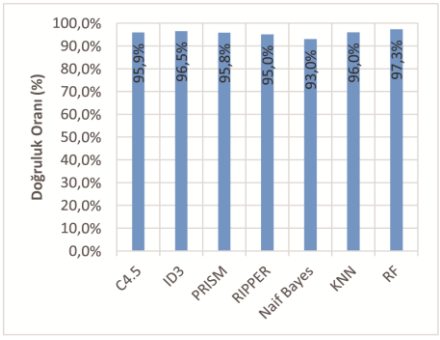
Algoritma	Sınıf	DP Oranı	YP Oranı	F-Ölçütü	ROC Alanı	Doğruluk (%)
C4.5	-1	0.942	0.028	0.953	0.984	95.9
	1	0.972	0.058	0.963	0.984	
ID3	-1	0.957	0.027	0.961	0.977	96.5
	1	0.973	0.043	0.969	0.977	
PRISM	-1	0.982	0.051	0.96	0.962	95.8
	1	0.949	0.018	0.967	0.963	
RIPPER	-1	0.929	0.033	0.943	0.961	95.0
	1	0.967	0.071	0.956	0.961	
Naif Bayes	-1	0.904	0.05	0.919	0.981	93.0
	1	0.95	0.096	0.938	0.981	
KNN	-1	0.944	0.027	0.954	0.992	96.0
	1	0.973	0.056	0.964	0.992	
RF	-1	0.961	0.017	0.97	0.996	97.3
	1	0.983	0.039	0.976	0.996	



Şekil 1: Sınıflandırma algoritmalarının "Kimlik Avı Sahteciliği (Sınıf Değeri: -1)" website tespitine göre başarı ölçümü.



Şekil 2: Sınıflandırma algoritmalarının "İyicil (Sınıf Değeri: 1)" website tespitine göre başarı ölçümü.

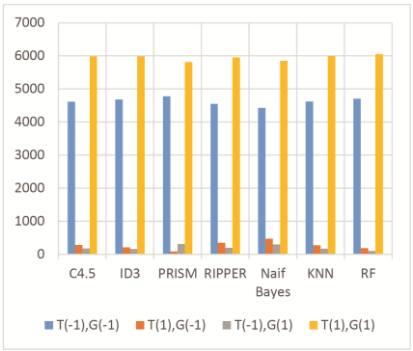


Şekil 3: Sınıflandırma algoritmalarının doğruluk ölçümü.

Deneyler, örnek sayısı 11055 olan Phishing Websites veri kümesi ile Weka makine öğrenmesi aracı kullanılarak gerçekleştirilmiştir. Veri kümesinde niteliklerin aldıkları değerler 1, 0 ve -1 olduğundan ve sınıf değeri 1 (İyicil) ve -1 (Kimlik Avı Sahteciliği) aldığından dolayı C4.5, ID3, PRISM, RIPPER, Naif Bayes, KNN ve RF algoritmalarının kullanılmasında herhangi bir engel bulunmamaktadır. Sınıflandırma algoritmaları, veri kümesi üzerinde çapraz-doğrulama (cross-validation) kullanılarak uygulanmıştır. Çapraz doğrulama için genel kabul gören 10 bölütte çapraz doğrulama değeri temel alınmıştır. Ayrıca RF algoritmasının uygulanmasında ağaç sayısı 100 ve KNN algoritmasının uygulanmasında *k* değeri 3 olarak alınmıştır. Bu değerler, farklı değerler denenerek doğruluk oranı ve performans ölçümünde optimal denge temel alınarak belirlenmiştir.

Algoritmaların uygulanması sonrasında, kimlik avı web sayfalarının tespitine yönelik uygunluklarını kıyaslamak amacıyla Karmaşıklık Matrisi, DP Oranı, YP Oranı, F-Ölçütü ve ROC Alanı yöntemleri kullanılmıştır. Bununla birlikte, algoritmaların veri kümesi üzerinde işletim sürelerine göre performans karşılaştırmaları gerçekleştirilmiştir. Tablo 3'te uygulama sonrasında algoritmaların DP Oranı, YP Oranı, F-Ölçütü değeri ve ROC Alan değeri ile elde edilen sonuçlar ve Doğruluk Oranları sunulmuştur. Temel amaç, Kimlik Avı tespitini yapmak olduğundan, Sınıf parametresi üzerinden sadece -1 baz alınmıştır. En temel algoritma değerlendirme kriteri olarak, DP Oranı (Doğruluk) ve YP Oranı (Hata Oranı) değerleri göz önüne alındığında, PRISM algoritması DP Oranında en başarılı sonucu vermektedir. En kötü sonuç ise Naif Bayes algoritması ile elde edilmiştir. Ayrıca tahmin edilen sınıf değerlerindeki hata oranı en düşük algoritma RF olarak gözlemlenmiş ve RF algoritmasına en yakın hata oranına sahip algoritmalar; C4.5, ID3 ve KNN olarak elde edilmiştir. Elde edilen sonuçlardaki hata oranı en yüksek olan algoritmalar ise PRISM ve Naif Bayes algoritmalarıdır. Kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması olan F-Ölçüt değerleri temel alındığında, en başarılı algoritmaların RF, ID3 ve PRISM algoritmaları olduğu anlaşılmaktadır. Diğer algoritmalarla kıyaslandığında düşük başarıya sahip olarak değerlendirilen algoritma ise Naif Bayes algoritmasıdır. Algoritmaların tahmin işlemlerinde; DP Oranı ve YP Oranından elde edilen eğrinin altında kalan alanı temel alan ROC Alanı değeri, 1 değerine ne kadar yakınsarsa o kadar daha iyi sonuçlar verdiği bilinmektedir. Buradaki ana amaç, DP Oranı ile YP Oranını birlikte değerlendirerek doğruluk oranını arttırmaktır. Bu bağlamda, RF ve KNN algoritması diğerleri ile kıyaslandığında ROC Alanı değerine göre en iyi algoritmalar olarak gözlemlenmektedir. ROC Alanına göre daha az başarılı algoritmalar ise PRISM ve RIPPER algoritmalarıdır. Bakılan başarı ölçütlerine ek olarak, algoritmaların işletiminde elde edilen doğruluk oranları kıyaslandığında RF ve ID3 algoritmaları en fazla örneği doğru şekilde sınıflandırmış algoritmalar olarak tespit edilmiştir. Doğruluk oranı en düşük olan algoritma ise Naif Bayes algoritmasıdır.

Veri kümesinde bulunan Kimlik Avı ve İyicil olarak sınıflandırılan örneklerin algoritmanın uygulanması sonrasında, 11055 örneğin ne kadarının doğru şekilde tahmin edildiğini gösteren Karmaşıklık Matrisi sonuçları Tablo 4'te gösterilmiştir. Kimlik Avı web sayfalarının tespiti amaçlandığından dolayı Tablo 4'teki algoritmalar karşılaştırıldığında Tablo 3'te verilen analizlerin gerçekliği daha açık şekilde görülmektedir. Ek olarak, tahmini (T) ve gerçek (G) sınıf değerlerine göre tespit edilen örnek sayısının, sınıflandırma algoritmalarına göre karşılaştırması Şekil 4'te sunulmaktadır.



Şekil 4: Sınıflandırma algoritmalarının karmaşıklık matrisine göre doğruluk ölçümü (T: Tahmini, G: Gerçek).

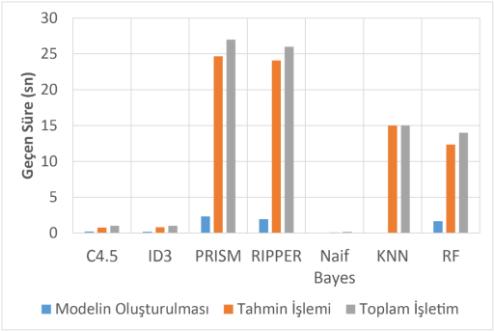
Tablo 4: Sınıflandırma algoritmalarının Karmaşıklık Matrisi.

Algoritma	Tahmini Sınıf -1	Tahmini Sınıf 1	Gerçek Durum Sınıfları
C4.5	4615	283	-1
	173	5984	1
ID3	4682	211	-1
	166	5986	1
PRISM	4777	86	-1
	313	5814	1
RIPPER	4549	349	-1
	202	5955	1
Naif Bayes	4427	471	-1
	305	5852	1
KNN	4624	274	-1
	167	5990	1
RF	4709	189	-1
	105	6052	1

Algoritmaların doğruluk ve hassasiyetleri, Tablo 3 ve Tablo 4 üzerinden yorumlanabilmekle birlikte performans açısından bu bilgiler ile herhangi bir yorum yapılamamaktadır. Bu sebeple, Intel(R) Core(TM) i7-3610QM 2.30 Ghz işlemci, 8 GB Ram özelliklerine sahip Windows 10 işletim sistemi kurulu bir bilgisayar üzerinde testler yapılmıştır. Yapılan testler ile algoritmaların veri kümesi üzerinde modelin oluşturulması, tahmin işlemi ve toplam işletim süreleri karşılaştırmalı olarak Tablo 5'te ve grafiksel olarak Şekil 5'te sunulmuştur.

Tablo 5: Sınıflandırma algoritmalarının veri kümesi üzerinde modelin oluşturulma, tahmin ve toplam işletim süreleri.

Algoritma	Modelin Oluşturulma Süresi (sn.)	Tahmin Süresi (sn.)	Toplam İşletim Süresi (sn.)
C4.5	0.22	0.78	1.0
ID3	0.18	0.82	1.0
PRISM	2.33	24.67	27.0
RIPPER	1.93	24.07	26.0
Naif Bayes	~0	~0	~0
KNN	~0	15.0	15.0
RF	1.66	12.34	14.0



Şekil 5: Sınıflandırma algoritmalarının veri kümesi üzerinde modelin oluşturulması, tahmin işlemi ve toplam işletim süreleri.

C4.5 ve ID3 algoritmaları, birbiri ardına gelen algoritmalar olması sebebi ile genel olarak birbirine yakın sonuçlar vermektedir. Bu bağlamda yapılan performans analizinde, C4.5 ve ID3 algoritmalarından elde edilen hata oranı düşük olmakla birlikte performansı en yüksek olan algoritmalar arasında gözlemlenmektedir. Bunun aksine, diğer algoritmalar ile kıyaslandığında daha düşük başarı oranları elde eden Naif Bayes algoritması; C4.5 ve ID3'den daha düşük modelin oluşturulma süresi ve tahmin süresine sahip olarak en hızlı algoritma olarak belirlenmiştir. Ayrıca elde edilen deney sonuçlarına göre; başarı oranlarında çok iyi sonuçlar elde edilen PRISM ve başarı oranlarında gözle görülür bir fark görünmeyen RIPPER algoritmalarının en düşük performansa sahip oldukları gözlemlenmiştir.

5 Sonuçlar

Web uygulamalarındaki artışla birlikte yükselen kimlik avı sahteciliğine dair web sayfalarının tespiti, verdiği zararlar göz önüne alındığında büyük önem kazanmaktadır. Yapılan araştırmalar sonucunda 2015 yılının 4. çeyreğinde 1.5 milyon yeni kimlik avı sayfasının tespit edilmesi problemin önemini daha iyi göstermektedir. Tespit amacı ile kullanılan makine öğrenmesi tekniklerinden sınıflandırma algoritmalarını, seçilen bir veri kümesi üzerinde test ederek bu problemin çözümünde hangi algoritmanın daha başarılı veya başarısız olduğu konusunda literatüre bir katkı sunulması amaçlanmıştır. Yapılan testler sonucunda RF ve PRISM algoritmalarının doğruluk oranında yüksek başarı gösterdiği gözlenmiş, fakat PRISM algoritmasının hata oranında ve RF ile PRISM algoritmalarının modelin oluşturulma ve tahmin süreleri açısından diğer algoritmaların gerisinde kaldıkları sonucu elde edilmiştir. Tespit etme, hata oranı ve işlem süreleri baz alındığında, test edilen algoritmalar arasında doğruluk oranının en yüksek olduğu algoritmalar RF ve ID3 olarak elde edilmiştir. Modelin oluşturulması ve tahmin sürelerinde en iyi sonucu elde eden Naif Bayes algoritması ise diğer algoritmalar ile kıyaslandığında daha düşük başarı oranına sahiptir. Literatürde incelenen Yapay Sinir Ağları ile yapılan çalışmalarda en büyük başarı faktörü otomatize çalışma mantığı olmasına rağmen çalışmamızda test edilen algoritmaların doğruluk oranları ve performans değerleri dengesi karşılaştırıldığında bizim sonuçlarımızın daha iyi değerlerde olduğu gözlemlenmiştir. Veri kümesinin tüm özniteliklerinden alt küme elde ederek sınıflandırma algoritmaları ile test edilen diğer literatür çalışmalarında ise yüksek başarım oranları elde edilmesine rağmen performans açısından bir değerlendirme sunulmamıştır. Yaptığımız çalışma da test edilen algoritmalar ve karşılaştırma ölçütlerinin çeşitliliği ile birlikte kullanılan verisetinde bulunan örnek miktarı, incelenen diğer çalışmalarda tespit edilmemiştir. Bu nedenle, literatüre bu açıdan da katkı sağlayacağı düşünülmektedir.

Gelecek çalışmalarda dinamik çalışacak şekilde geliştirilen bir sistem ile birlikte ID3 veya RF algoritmaları kullanılarak, gerçek zamanlı ve sürekli veri kümesi genişleyen bir uygulama gerçekleştirilebilir. Buna ek olarak, elde edilen dinamik veri kümesi kullanılarak yeni bir özellik seçme veya boyut indirgeme yöntemi geliştirilerek test edilebilir. Son olarak, özniteliklerin değerli alt kümeleri ile hibrit veya yeni algoritmalar geliştirilerek yeni deneyler yapılabilir.

6 Kaynaklar

[1] McAfee Inc. "McAfee Labs Threats Report-February 2015". http://www.mcafee.com/us/resources/reports/rp-quarterly-threat-q4-2014.pdf (26.03.2016).

[2] Symantec Corp. "Internet Security Threat Report-ISTR 20 April 2015". https://www4.symantec.com/mktginfo/whitepaper/ISTR/21347932_GA-internet-security-threat-report-volume-20-2015-social_v2.pdf (26.03.2016).

[3] McAfee Inc. "McAfee Labs Threats Report-March 2016". http://www.mcafee.com/us/resources/reports/rp-quarterly-threats-mar-2016.pdf (01.05.2016).
[4] UCI Machine Learning Repository. "Phishing Websites Dataset". https://archive.ics.uci.edu/ml/datasets/Phishing+Websites (26.03.2016).
[5] Kazemian HB, Ahmed S. "Comparisons of machine learning techniques for detecting malicious webpages". *Expert Systems with Applications,* 42(3), 1166-1177, 2015.
[6] Li Y, Yang L, Ding J. "A minimum enclosing ball-based support vector machine approach for detection of phishing websites". *Optik*, 127(1). 345-351. 2016.
[7] Moghimi M, Varjani AY. "New rule-based phishing detection method". *Expert Systems with Applications*, 53, 231-242. 2016.
[8] Nguyen LAT, To BL, Nguyen HK, Nguyen MH. "Detecting phishing web sites: A heuristic URL-based approach". *2013 International Conference on Advanced Technologies for Communications (ATC 2013)*, Hochiminh, Vietnamese, 16-18 October 2013.
[9] Mohammad RM, Thabtah F, McCluskey L, Ieee. "An assessment of features related to phishing websites using an automated technique". *2012 International Conference for Internet Technology and Secured Transactions*, London, UK, 10-12 December 2012.
[10] Phishtank. "Phishtank Archive". https://www.phishtank.com/phish_archive.php (01.05.2016).
[11] Mohammad RM, Thabtah F, McCluskey L. "Predicting phishing websites based on self-structuring neural network". *Neural Computing & Applications*, 25(2), 443-458. 2014.
[12] Mohammad RM, Thabtah F, McCluskey L. "Intelligent rule-based phishing websites classification". *Iet Information Security*, 8(3), 153-160, 2014.
[13] Selvan K, Vanitha M. "A Machine Learning Approach for Detection of Phished Websites Using Neural Networks". *International Journal of Recent Technology and Engineering (IJRTE)*, 4(6), 19-23, 2016.
[14] Singh P, Jain N, Maini A. "Investigating the effect of feature selection and dimensionality reduction on phishing website classification problem". *1st International Conference on Next Generation Computing Technologies (NGCT 2015)*, Dehradun, India, 4-5 September 2015.
[15] The University of Maikato. "Weka Machine Learning Algorithms Collection Tool". http://www.cs.waikato.ac.nz/ml/weka/ (01.05.2016).
[16] Quinlan JR. *C4.5: Programs for Machine Learning*. 1st ed. Massachusetts, USA, Morgan Kaufmann Publishers Inc., 1993.
[17] Quinlan JR. "Induction of decision trees". *Machine Learning*, 1(1), 81-106, 1986.
[18] Cendrowska J. "PRISM: An algorithm for inducing modular rules". *International Journal of Man-Machine Studies.* 27(4), 349-370, 1987.
[19] Cohen WW. "Fast effective rule induction". *Twelfth International Conference on Machine Learning (ML95)*, California, USA, 9-12 July 1995.
[20] John GH, Langley P. "Estimating continuous distributions in Bayesian classifiers". *Eleventh conference on Uncertainty in artificial intelligence*, Montreal, Canada, 18-20 August 1995.
[21] Dimitoglou G, Adams JA, Jim CM. "Comparison of the C4.5 and a Naïve Bayes classifier for the prediction of lung cancer survivability". *Journal of Computing*, 4(8), 1-9, 2012.
[22] Huang J, Lu J, Ling CX. "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy". *Third IEEE International Conference on Data Mining (ICDM),* Melbourne, USA, 22 November 2003.
[23] Aha DW, Kibler D, Albert MK. "Instance-based learning algorithms". *Machine learning*, 6(1), 37-66. 1991.
[24] Tan S. "An effective refinement strategy for KNN text classifier". *Expert Systems with Applications*, 30(2), 290-298, 2006.
[25] Breiman L. "Random forests". *Machine Learning*. 45(1), 5-32. 2001.
[26] Davis J, Goadrich M. "The relationship between Precision-Recall and ROC curves". *23rd international Conference on Machine Learning*, Pennsylvania, USA, 25-29 June 2006.
[27] Fawcett T. "An introduction to ROC analysis". *Pattern recognition letters*, 27(8), 861-874, 2006.
[28] Ferri C, Hernández-Orallo J, Modroiu R. "An experimental comparison of performance measures for classification". *Pattern Recognition Letters*, 30(1), 27-38. 2009.
[29] Powers DM. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". *Journal of Machine Learning Technologies*, 2(1), 37-63. 2011.