Avrupa Bilim ve Teknoloji Dergisi
Sayı 31 (Ek Sayı 1), S. 182-187, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 31 (Supp. 1), pp. 182-187, December 2021

Research Article

Machine Learning Detection of Phishing Websites Using Website Data

Murathan Ok [1*], İlker Kara[2]

[1*] Hacettepe Üniversitesi, Bilişim Enstitüsü, Ankara, Türkiye, (ORCID: 0000-0003-2584-0199), murathanok@gmail.com
[2] Çankırı Karatekin Üniversitesi, Eldivan Tıbbi Hizmetler ve Sağlık Meslek Yüksek Okulu, Çankırı, Türkiye (ORCID: 0000-0003-3700-4825), karaikab@gmail.com

(First received 11 Ekim 2021 and in final form 23 December 2021)

(**DOI:** 10.31590/ejosat.1008335)

ATIF/REFERENCE: Ok, M. & Kara, İ. (2021). Machine Learning Detection of Phishing Websites Using Website Data. *European Journal of Science and Technology*, (31), 182-187.

Abstract

Attackers are designing fake websites to collect sensitive data such as credit card, email, social media account information of their victims. These attacks keep getting more sophisticated day by day. This leads to highly convincing website designs that can easily trick users and steal their information. In order to prevent this increasingly dangerous problem from spreading, systems with machine learning capabilities have been developed to predict if a web page or web site is created exclusively for phishing or not. In this study, using the determined 6 features of the selected sample website address 12.275. It is tried to determine whether this website address is prepared for phishing purposes or not by using a random forest algorithm. The selected data set in this study have been collected from open-source datas which is published on the official website of the Computer Emergency Response Team of Turkey. The data set was created by categorizing and labeling internet urls and domain names according to 6 determined features. Tags (Phishing=1, Suspect=0, Legitimate=-1) were determined using scripts and tabulation programs developed on python programming language. As a conclusion of the study, used method has shown 95 % success performance.

Keywords: Website Features, Phishing, Feature Extraction, Machine Learning.

İnternet Sayfası Verileri Kullanarak Kimlik Avı Web Sitelerinin Makine Öğrenme Tekniğiyle Tespiti

Öz

Saldırganlar kurbanlarının kredi kartı, e-posta, sosyal medya hesap bilgileri gibi hassas verileri ele geçirmek için sahte web siteleri tasarlamaktadırlar. Bu saldırılar son zamanlarda daha karmaşık hale gelmiş dikkatli kullanıcıları kolayca kandırabilen oldukça ikna edici tasarımlar geliştirmişlerdir. Giderek daha tehlikeli hale gelen bu soruna çözüm bulmak için yapılan çalışmalar sahte web sayfalarının verileri ile kimlik avı amacı ile hazırlanmış web sayfalarının makine öğrenmesi yöntemi kullanarak tahminini yapabilecek sistemler tasarlanmıştır. Bu çalışmada seçilen örnek web sitesi adresine ait belirlenmiş 6 özellik kullanılarak; bu web site adreslerinin kimlik avı amacı ile hazırlanıp hazırlanmadığı rastgele orman (random forest) algoritması kullanarak tespit edilmeye çalışılmaktadır. Çalışmada seçilen veri seti, Uluslararası Siber Olaylara Müdahale Merkezinin resmi web sitesinde yer alan açık kaynak verileri kullanılmıştır. Toplamda 12.275 adet web sitesi çalışma için değerlendirilmiştir. Veri seti, internet URL ve alan adlarının belirlenen 6 özelliğin kategorilendirilmesi ve etiketlenmesi ile oluşturulmuştur. Etiketler (Kimlik avı=1, Şüpheli=0, Meşru=-1) python dilinde geliştirilmiş betikler ve tablolama programlarından yararlanılarak belirlenmiştir. Çalışma sonucunda kullanılan yöntem 95% başarı performansı göstermiştir.

Anahtar Kelimeler: Web Sitesi Özellikleri, Kimlik Avı, Özellik Çıkarma, Makine Öğrenmesi.

[*] Sorumlu Yazar: murathanok@gmail.com

1. 1. Giriş

Kimlik avı saldırıları genel olarak kurbanların e-posta hesaplarına; ödül, kampanya, hediye gibi cezbedici ve/veya bir panik durumunu oluşturmayı amaçlayan sahte iletiler gönderilmesi ile gerçekleşir. Burada amaç kredi kartı ve kimlik bilgisi gibi hassas verilerin çalınmasıdır. Dolandırıcılar bu bilgileri farklı amaçlar için kullanabilir. İlk kimlik avı vakası 2004 yılında, "America Online" web sitesinin taklidini oluşturan Kaliforniyalı bir genç ile başlamıştır. Sahte bir web sitesi ile kullanıcılardan hassas bilgiler elde edinilebilmiş ve hesaplarından para çekmek için kredi kartı bilgilerine erişim sağlanabilmiştir (CNN, Phishing scams reel in your identity).

E-posta ve web sitesi ile yapılan kimlik avı faaliyetleri dışında, 'vishing' (sesli kimlik avı), 'smishing' (SMS kimlik avı) ve siber suçluların sürekli olarak geliştirdikleri diğer kimlik avı teknikleri de bulunmaktadır.

Kimlik avı saldırıları keşfedildiği günden bugüne benzer yöntemler kullanmakta ve kurbanları tuzağa düşürmeyi yüksek oranda başarabilmektedir. Uluslararası operasyonları ile kimlik avı saldırılarına karşı koruma ve farkındalık sağlamayı hedefleyen Keepnet firmasının (Keepnetlabs 2020 phishing statistics) raporuna göre kullanıcıların %97'si hedefli bir kimlik avı e-postasını tanımlayamamaktadır. SANS Enstitüsü'ne göre, kurumsal ağlara yapılan tüm saldırıların %95'i hedefli kimlik avının sonucu meydana gelmektedir. Verizon firmasının 2020 Veri İhlali Araştırma Raporunda (Verizon 2020 Summary of Findings), yalnızca kimlik avı içeren saldırıların 2020 yılı içinde %32'lik bir artış gösterdiği ve bunun tüm veri ihlallerinin neredeyse üçte birine karşılık geldiği ifade edilmektedir.

Kimlik avı saldırılarının bu düzeyde başarılı olmasının nedenlerinden en önemlileri,

- Kişilere yeterli farkındalık eğitimini verilmemiş olması,
- Kullanılan sosyal mühendislik metotları ile kullanıcıların tuzağa düşmesini kolaylaşması,
- İnsan zaaflarına yönelik içerikler bulunması,
- E-posta güvenlik sistemlerinin kimlik avı amaçlı gönderileri yeterli seviyede tespit edememesi,
- Yaygınlaşan IoT cihazları ve bu cihazlarla toplanan verinin kontrolsüz ve yeterli seviyede güvenlik altına alınmadan saklanması, veri sahibini hedef haline getirilmesi,
- Saldırıların başarılı olması nedeni ile yaygınlaşarak devam etmesidir (Bhardwaj vd., 2020).

Kimlik avı saldırılarının sahip olduğu başarı yüzdesi nedeni ile bu saldırıları durdurmak, tespit etmek ve önlemek için çalışmalar hem akademik hem de sektörel düzeyde devam etmektedir. Kimlik avı saldırılarını önlemek için çok aşamalı doğrulama mekanizmaları kullanılmaktadır. Kişi erişim sağladığında bildiği, sahip olduğu ve kendisinde olan birimlerle yetkilendirilmektedir (Ometov vd., 2018). Bu üç unsur aynı anda sağlayamayan saldırgan başarısız olacaktır. Kimlik avı saldırılarının tespiti için kişi farkındalıklarını arttırmak ya da yazılım tabanlı tespit çözümleri kullanılmaktadır. Yazılım tabanlı çözümlerde makine öğrenmesi, istihbarat veri tabanları ve kara listeler yer almaktadır (Apandi vd., 2020).

Yapılan çalışmada, internet sayfalarının verileri ile kimlik avı amacı ile hazırlanmış sayfaların tahminini yapabilecek bir sistem tasarlamak amaçlanmıştır. Sistem tahminini makine öğrenmesi kullanarak yapmaktadır. Geliştirme sürecinde kullanılacak makine öğrenme algoritması seçilirken, kullanılan veri setinin birçok makine öğrenme algoritması ile karşılaştırılması yapılmıştır. En yüksek doğruluk oranına sahip olduğu tespit edilen rasgele orman (random forest) algoritması tahmin için seçilmiştir. Makine öğrenmesi pyton dili ve spyder tümleşik geliştirme ortamında kodlanmıştır.

Çalışma 8 bölümden oluşmaktadır. 2. bölümde benzer çalışmalardan, 3. bölümde veri kümesinin tanımından, 4. bölümde veri ön işleme ve dönüşümünden, 5. bölümde seçilen algoritmadan, 6. bölümde sonuçlardan ve 7. bölümde gelecekte yapılacak çalışmalardan bahsedilmiştir.

2. Literatür Çalışması

Saldırı başarı oranının yüksek ve e-posta üzerinden dağıtıldığından ve kolay şekilde yayılım gösterebildiğinden kimlik avı saldırılarını önlemeye yönelik literatürde pek çok çalışma bulunmaktadır.

Literatürde, bu çalışma ele alındığında problemin çözümü için farklı veri setleri ve makine öğrenmesi algoritmaları kullanılmıştır. Bu çalışmada (Al-Ahmadi vd. 2020) kimlik avı web sayfası tanımlama sorunu bir görüntü sınıflandırma görevi olarak ele alınmaktadır. İnternet sayfası ekran görüntülerinden kompakt görsel özellikleri çıkarılmakta ve rasgele ağaç algoritması ile sınıflandırılmaktadır. (Awasthi vd., 2021) çalışmalarında alan adı özellikleriyle kimlik avı web sitesi URL'lerini tespit etmeye odaklanıldığı tespit edilmiştir. Çalışmaların çoğunun (Hema vd., 2020) Naive Bayes, destek vektör, karar ağacı ve rasgele orman gibi tanıdık makine öğrenimi algoritmaları kullanılarak yapıldığı sonucuna ulaşılmıştır. (Hossain vd., 2020) çalışmalarında internet sitesi arasında ayrım yapabilen özelliklerden oluşan veri kümelerine göre performans gösterecek şekilde rasgele orman makine öğrenmesi algoritması ile en iyi şekilde sonuç alınmıştır.

Chiew vd. (2019) Machine Learning Repository (UCI) verilerini kullanarak farklı makine öğrenme algoritmaları kullanarak kimlik avı web sayfası tespitini amaçlamışlardır. Çalışma sonucunda Random Forest algoritmasının %94,6 doğruluk oranında başarı gösterdiğini tespit etmişlerdir.

Benzer bir çalışmada Sahingoz vd. (2019); kimlik avı e-posta tespiti için 7 farklı sınıflandırma algoritması kullanarak 73.575 adet e-postası veri setini Random Forest algoritması kullanarak analiz etmiştir. Analizler sonucunda %97,98 doğruluk oranıyla kimlik avı yapan e-postaları tespit edebilmişlerdir.

Diğer bir çalışmada Kalaycı (2018), kimlik avı web sayfası tespiti için makine öğrenmesi yöntemleri kullanmıştır. Bu amaçla 1.353 örnekten oluşan bir veri setinde web sitesi adresine ait belirlenmiş 9 özellik kullanılmıştır. Çalışma sonucunda Rastgele Orman (RF) algoritmasının en yüksek başarı oranına ulaştığını vurgulamıştır. Biz de literatürdeki bu çalışmalar doğrultusunda web sitesi adresine ait belirlenmiş 6 özellik kullanılarak; bu web sitesi adresinin kimlik avı amacıyla hazırlanmış olup olmadığı öncelikle farklı algoritma çıktılarına göre karşılaştırılmıştır. Başarım oranı en yüksek olan Rastgele

Orman (RF) algoritması kullanarak da tespit çalışması yapılmıştır. Çalışma aşağıda listelenen 3 katkı sunmaktadır:

- Bu çalışma kimlik avı web sayfalarını farklı algoritmalarda karşılaştırılarak en yüksek başarım sağlanacak algoritmayı tespit etmeye odaklanmıştır.
- Şüpheli web sitesi adresine ait belirlenmiş 6 özelliğin öncelikle tespiti sağlanarak incelenmiştir.
- Kimlik avı web sayfası tespiti için yeni bir veri seti oluşturulmuştur. Bu veri setini ticari bir kaygı duymadan yayınlanmıştır.

3. Araştırma Sonuçları ve Tartışma

3.1. Veri Seti

Öğrenme ve test için kullanılan veri seti Uluslararası Siber Olaylara Müdahale Merkezinin resmi web sitesi ve açık kaynak verileri kullanarak temin edilmiştir. 12.275 adet web sitesi verisine ulaşılmıştır. Veri seti, internet URL ve alan adlarının belirlenen 6 özelliğinin kategorilendirilmesi ve etiketlenmesi ile oluşturulmuştur.

3.1.1. IP Adres Kullanma

URL'deki alan adına alternatif olarak "http://88.18.221.19/phish.html" örneğindeki gibi bir IP adresi kullanılması, kişisel bilgilerin çalmaya çalıştığına dair bir göstergedir. Bazen, IP adresi aşağıdaki bağlantıda gösterildiği gibi onaltılık koda dönüştürülmüş de olabilir. "http: //0x58.0xCC.0xCA.0x62/2/phish.html ". URL içeriğinde IP adresi kullanılmış ise sayfa kimlik avı kategorisinde sınıflandırılmıştır.

3.1.2. Şüpheli Kısmı Gizlemek İçin Uzun URL Kullanımı

Kimlik avcıları, adres çubuğundaki şüpheli kısmı gizlemek için uzun URL'ler kullanabilmektedir. Örneğin: http://bimcelltr.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/? URL'nin uzunluğu 30 karakterden fazla veya buna eşitse, URL kimlik avı sayfası olarak sınıflandırılmıştır.

3.1.3. Alan Adında "-" Sembolü ile Ön ya da Son Ek Varlığı

Kısa çizgi simgesi, meşru URL'lerde nadiren kullanılır. Kimlik avcıları, kullanıcıların meşru bir web sayfasıyla uğraştıklarını hissetmeleri için alan adına (-) ile ayrılmış ön ekler veya son ekler ekleme eğilimindedir. Örneğin: http://www.ödeme-vakifbank.com/. Bu tarz kullanımlara kimlik avı saldırılarında sıkça rastlanmaktadır.

3.1.4. Alt Alan Adı ve Çoklu Alt Alan Adı

Alan adlarında ülke ve sektör tanımı için kullanılan uzantılar dışındaki nokta işareti bulunuyorsa, URL bir alt alana sahip demektir. Bir alan adına sahip URL'ler "Şüpheli" olarak sınıflandırılmıştır. Nokta sayısı ikiden büyükse, birden çok alt etki alanına sahip olacağından kimlik avı sayfası olarak sınıflandırılmıştır.

3.1.5. Alan Adı Kayıt Süresi

Bir kimlik avı internet sitesinin kısa bir süre önce internette yayına başladığı gerçeğinden yola çıkarak, güvenilir alan adlarının düzenli olarak birkaç yıl önceden yayına alındığı tespit edilmiştir. Alan adı bir yıldan daha kısa sürede kayıt edilmiş ise kimlik avı olarak sınıflandırılmıştır.

3.1.6. CPR Puanı

Bir web sayfasının internette hangi sırada ve ne kadar önemli olduğunu ölçülebilmektedir. Web sayfasının teknik yönlerinin arama motoru optimizasyonlarına göre daha yüksek sıralamalarda sonuç vermesi ve organik trafiğe ne kadar iyi ulaşılacağının ölçüsüdür. Kimlik avı amacı ile hazırlanmış internet sayfalarının "0,2" seviyesinde bir CPR değerine ulaşabileceğini görülmüştür. CPR puanı "0,2" olan internet sayfaları kimlik avı olarak sınıflandırılmıştır.

3.1.7. Google İndexi

Bir internet sitesi Google tarafından indekslendiğinde, arama sonuçlarında görüntülenir. Genellikle, kimlik avı web sayfalarına yalnızca kısa bir süre için erişilebilir olduğundan birçok kimlik avı amaçlı internet sayfası Google indexinde bulunmamaktadır. Google indexinde bulunmayan sayfalar kimlik avı olarak sınıflandırılmıştır.

4. Veri Dönüşümü ve Ön Işleme

Çalışmada kullanılan veri seti Uluslararası Siber Olaylara Müdahale Merkezinin resmî web sitesinde yer alan açık kaynak verileri kullanarak oluşturulmuştur. 23.09.2021 ve 19.08.2021 tarihleri arasında kayda geçmiş 39.441 zararlı bağlantı bilgisine ulaşılmıştır. Bu bilgilerden 20.614 kaydın kimlik avı kategorisinde olduğu saptanmış ve kayıtlar ayrıştırılmıştır. Veri seti için nitelik toplama aşamasında 20.614 kayıttan 8.707 tane kimlik avı verisinin tüm niteliklerine ulaşılabilmiştir. Ayrıca 3.567 adet kimlik avı olmayan web sayfasının verisi de veri setine eklenmiştir.

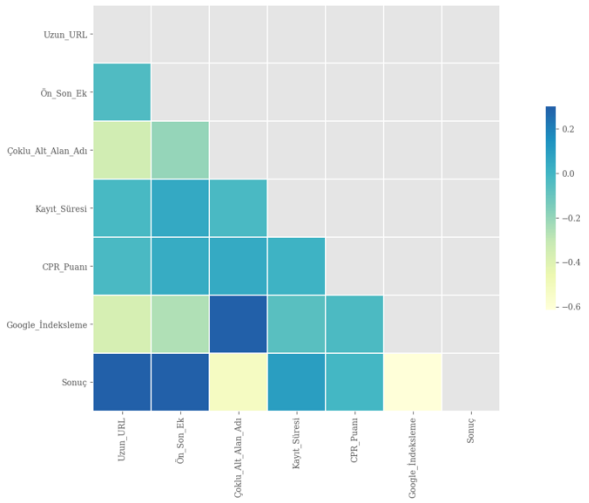
Veriler, kimlik avı kategorisinde değerlendirilen bir özellik ise "1", değil ise "-1" olarak nitelendirilmiştir. Veri tanımlarına göre kimlik avı şüphesi olan veriler "0" ile nitelendirilmiştir. Veri kümesinde makine öğrenmesi çalışmalarına başlamadan önce 12.275 adet girdi bulunmaktadır. Veri kümesi kimlik avı ve meşru siteler için yakın dağılım bir olmasına dikkat edilmiştir. Dengeli bir veri seti üzerinde hazırlanmaya çalışılmıştır.

Tablo 1. Veri Dağılımı

Toplam Veri	Kimlik Avı Olan Veri	Kimlik Avı Olmayan Veri	Veri Oranı
12274	8707	3567	0,7093

Veri kümesi içerisinde eksik/kayıp (missing value) kontrollü yapılmıştır. Veri kümesi dâhil hiçbir tanımda eksik veri tespit edilmemiştir. Veri seti üzerinde eksik/kayıp veri çözümle stratejilerinden hiçbiri uygulanmamıştır.

Şekil 1. Örnek Korelasyon Tablosu



Veri algoritma öğrenmesinde ve kullanılmak üzere öğrenme ve test için ayrıştırılmıştır.

Tablo 2. Öğrenmesi Verisi Dağılımı

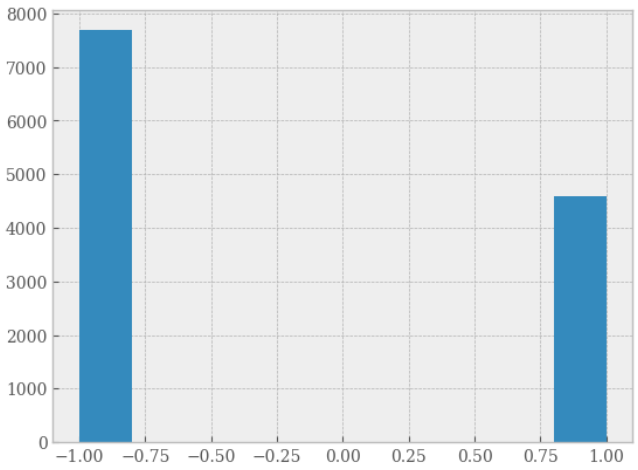
Toplam Veri	Kimlik Avı Olan Veri	Kimlik Avı Olmayan Veri	Veri Oranı
8591	6094	2497	0,7093

Tablo 3. Test Verisi Dağılımı

Toplam Veri	Kimlik Avı Olan Veri	Kimlik Avı Olmayan Veri	Veri Oranı
3683	2613	1070	0,7094

Veri kümesi tanımlarında öğrenme algoritmasını etkileyecek, sadece bir değere yakın verilerin tespiti için tanımlar değerlerine göre grafiğe dönüştürülmüştür. Dönüşen verilerde değerlerin kabul edilemez varyanslar gözlemlenmemiştir. Veri üzerinde manipülasyona gerek duyulmamıştır.

Şekil 2. Örnek Tanım Değer Dağılımı(Google İndeksleme)



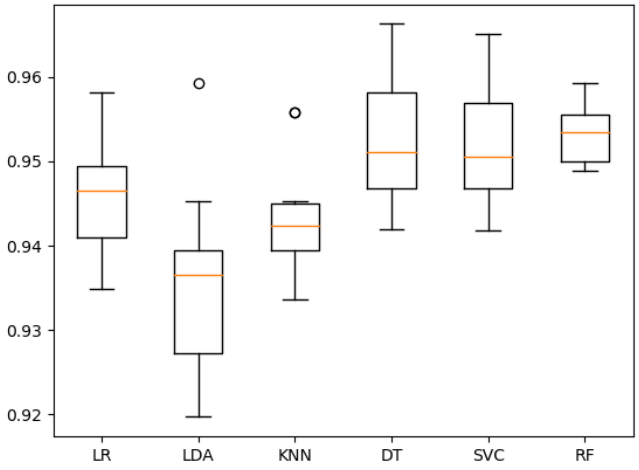
5. Algoritma Seçme

Elde edilen öğrenme veri kümesi ile Lojistik Regresyon (LR), Doğrusal Diskriminant Analizi (LDA), En Yakın Komşu(KNN), Kara Ağacı (DT), Destek Vektör Makineleri (SVM), Rasgele Orman (RF) algoritmaları karşılaştırılmıştır. Veri kümesinde kimlik avı ya da meşru siteler yakın oranda ayrıştırıldığı için burada sınıflandırma yaparken "doğruluk" (accuracy) kullanılmıştır. Algoritmalar çapraz doğrulama ile başarı oranı en yüksek algoritma üzerine gidilmiştir.

Tablo 4. Test Verisi Dağılımı

Algoritma	Çapraş Doğrulama Skoru	Standart Sapma
LR	0.9456405501258901	0.00591633200
LDA	0.9355138479031865	0.00650729457
KNN	0.9413347050383084	0.00538125036
DT	0.9522759022118743	0.00259795316
SVM	0.9512281722933643	0.00621256462
RF	0.952391233722284	0.00626242280

Şekil 3. Algoritmaların karşılaştırılması



Karar ağaçları ve destek vektör makineleri algoritmalarında yapılan karşılaştırmalarda belirli aralıklarda yüksek başarım sağlandığı görülmüştür ancak ortalama başarımda rastgele orman algoritması daha istikrarlı sonuçlar vermiştir ve ortalama başarım daha yüksektir.

Alınan sonuçlar neticesinde rastgele orman algoritması ile makine öğrenmesi çalışmasına devam edilmesine karar verilmiştir.

5.1. Rasgele Orman Algoritması (Random Forest)

Rastgele orman algoritması hem sınıflandırma hem de regresyon görevlerini yerine getirebilmektedir. Rastgele orman iki veya daha fazla algoritmayı birleştirerek tahmin sonucuna ulaşmaktadır. (Hatwell vd., 2020) Temelde rastgele seçilmiş bir noktadan bir karar ağacı oluşturmaktadır. Bu işlem "N" ağaç sayısı kadar tekrarlanır. Yapılan çalışmada ağaç sayısı 100 olarak

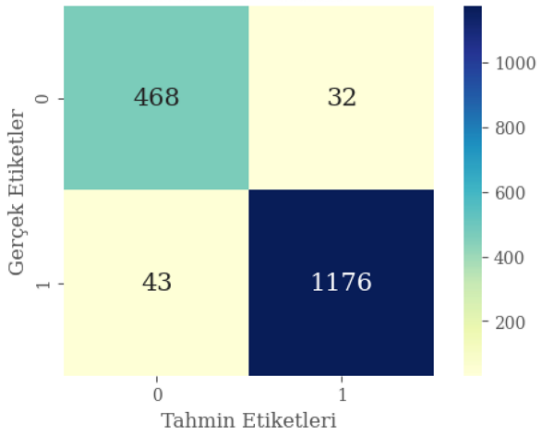
atanmıştır. Ağaç derinliği ve en yüksek yaprak düğümü değeri sınırlandırılmamıştır.

Veri seti üzerinden oluşturulmuş öğrenme ve test verileri kullanılarak önce algoritmayla öğrenmenin tamamlanması sağlanmıştır. Bu aşamada ayrıştırılmış öğrenme veri seti kullanılmıştır. Daha sonra öğrenmenin gerçekleştiğini test etmek için algoritmaya daha önce sokulmamış test verisi ile sonuçlar alınmıştır. Asıl amaç algoritmanın daha önce karşılaşmadığı veri ile vereceği sonuçları saptamaktır.

İlk aşama olarak öğrenme verisi ile bir doğrulama aşaması gerçekleştirilmiştir. Öğrenme veri setinden bir doğrulama seti ayrıştırılmış ve öğrenme sağlanmıştır. Veri içerisinden 6.872 kayıt öğrenme, 1.719 kayıt doğrulama için kullanılmıştır.

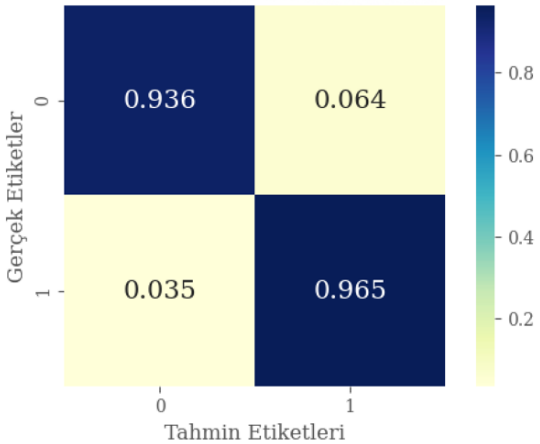
Bu model ile 0.9621 oranında başarılı tahmin yapıldığı sonucuna ulaşılmıştır. Doğrulama setinde 1176 kimlik avı, 468 meşru internet sayfası başarı ile tespit edilmiştir.

Şekil 4. Hata Matrisi (Değer)



Doğrulama setinde kimlik avı sayfalarında % 96, meşru internet sayfalarının tespitinde ise %93 başarı elde edilmiştir.

Şekil 5. Hata Matrisi (Oran)



Tanımların özellik sınıfını tahmin etmede aldığı rol, o tanımın önemini ifade etmektedir. Veri seti ve rasgele orman modelindeki özellik önemi aşağıdaki kod çıktısında paylaşılmaktadır.

Tablo 5. Özellik Önemi

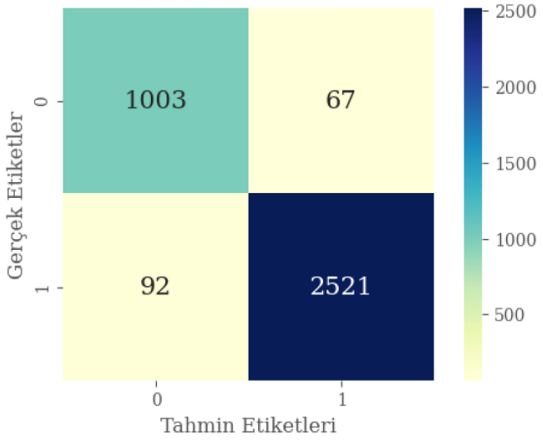
Etiket	Önem
Google_İndeksleme	0.3108731
CPR_Puanı	0.286435
Uzun_URL	0.233696
Ön_Son_Ek	0.148223
Çoklu_Alt_Alan_Adı	0.013059
Kayıt_Süresi	0.007714

6. Sonuç

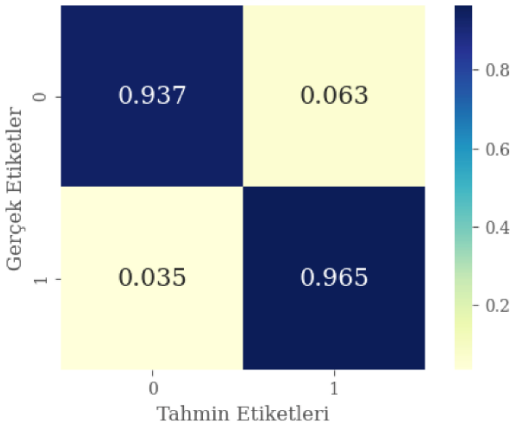
Gerçekleştirilen çalışmada internet sayfalarının verilerinden, sayfanın kimlik avı için hazırlanıp hazırlanmadığı tahmin edilmeye çalışılmıştır. Çalışmada rasgele orman modeli kullanılmıştır. Model başarılı bir veri seti ile uygulanmış ve tahmin oranları tatmin edici seviyelerde sonuç vermiştir. Model öğretilmemiş verilerde de yüksek tahmin oranı ile çalışmaktadır.

Modelin başarısı veri dönüşümünde az kayıp yaşanması, doğru makine öğrenmesi algoritmasının seçilmesi, veri seti içerisindeki tanımların tutarlılığı şeklinde aktarılabilir. Ayrıştırılmış test verisi sonuçlarına göre alınan son tahmin değerleri aşağıda paylaşılmıştır.

Şekil 6. Hata Matrisi (Oran) - Test Verisi



Şekil 7. Hata Matrisi (Oran) - Test Verisi



Test veri setinde 3.683 kayıttan 2521 kimlik avı, 1003 meşru internet sayfası başarı ile tespit edilmiştir. Kimlik avı sayfalarında % 96, meşru internet sayfalarının tespitinde ise %93 başarı elde edilmiştir. Toplan doğru tahmin oranı ise %95 olarak saptanmıştır.

7. Gelecek Çalışmalar

Model üzerinde elde edilen başarı oranı, modelin kullanıcılara gönderilmiş e-postalar içerisindeki bağlantıların kimlik avı tespitini yaptıktan sonra engellemesi ya da sayfayla ilgili bilgilendirme yapacak bir mekanizmanın (Xue vd., 2020) içerisinde kullanılabileceğini göstermiştir. Bir tespit mekanizması öğesi olarak makine öğrenmesi alt yapısı ile kullanıcıların güvenli bağlantılara erişmesi için gerekli altyapıda önemli bir rol üstlenebilecektir. Kullanıcıya ulaşan e-posta içerisindeki bağlantı geliştirilmiş bir kod yardımı ile bu modelde kullanılan veri seti isterlerine dönüştürülecek, model gelen veriye göre karar verip kullanıcının ilgili bağlantı ile iletişimini yönetebilecektir.

Kaynakça

CNN | Phishing scams reel in your identity, CNN. [Çevrimiçi]. https://edition.cnn.com/2003/TECH/internet/07/21/phishing.scam/index.html [Erişim: 27-Eylül-2021].

Keepnetlabs | 2020 phishing statistics, Keepnetlabs. [Çevrimiçi]. https://www.keepnetlabs.com/phishing-statistics-you-need-to-know-to-protect-your-organization/#easy-footnote-bottom-3-3791 [Erişim: 19-Eylül-2021].

Verizon | 2020 Summary of Findings, Verizon. [Çevrimiçi]. https://enterprise.verizon.com/resources/reports/dbir/2020/summary-of-findings/ [Erişim: 6-Ekim-2021].

Bhardwaj, A., Sapra, V., Kumar, A., Kumar, N., & Arthi, S. (2020). Why is phishing still successful?. Computer Fraud & Security, 2020(9), 15-19.

Ometov, A., Bezzateev, S., Mäkitalo, N., Andreev, S., Mikkonen, T., & Koucheryavy, Y. (2018). Multi-factor authentication: A survey. Cryptography, 2(1), 1.

Apandi, Siti & Sallim, Jamaludin & Sidek, Roslina. (2020). Types of anti-phishing solutions for phishing attack. IOP Conference Series: Materials Science and Engineering. 769. 012072. 10.1088/1757-899X/769/1/012072. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

Al-Ahmadi, S. (2020). A Deep Learning Technique for Web Phishing Detection Combined URL Features and Visual Similarity. International Journal of Computer Networks & Communications (IJCNC) Vol, 12.

Awasthi, A., & Goel, N. (2021). Phishing Website Prediction: A Machine Learning Approach. In Progress in Advanced Computing and Intelligent Engineering (pp. 143-152). Springer, Singapore.

Hema, R., Ramya, V., Sahithya, K., & Sekharan, R. (2020). Detecting of Phishing Websites using Deep Learning. Journal of Critical Reviews, 7(11), 3606-3613.

Hossain, S., Sarma, D., & Chakma, R. J. (2020). Machine Learning-Based Phishing Attack Detection. Machine Learning, 11(9).

Hatwell, J., Gaber, M. M., & Azad, R. M. A. (2020). CHIRPS: Explaining random forest classification. Artificial Intelligence Review, 53, 5747-5788.

Xue, M., Yuan, C., Wu, H., Zhang, Y., & Liu, W. (2020). Machine learning security: Threats, countermeasures, and evaluations. IEEE Access, 8, 74720-74742.