POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Tekdüzen kaynak bulucu yoluyla kimlik avı tespiti için makine öğrenmesi algoritmalarının özellik tabanlı performans karşılaştırması

Feature-based performance comparison of machine learning algorithms for phishing detection through uniform resource locator

***Yazar(lar) (Author(s))**: Taki SAVAŞ[1], Serkan SAVAŞ[2]*

ORCID[1]: 0000-0001-7133-7071

ORCID[2]: 0000-0003-3440-6271

Bu makaleye şu şekilde atıfta bulunabilirsiniz(To cite to this article): Savaş T. ve Savaş S., "Tekdüzen kaynak bulucu yoluyla kimlik avı tespiti için makine öğrenmesi algoritmalarının özellik tabanlı performans karşılaştırması", ***Politeknik Dergisi***, 25(3): 1261-1270, (2022).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: **10.2339/politeknik.1035286**

Tekdüzen Kaynak Bulucu Yoluyla Kimlik Avı Tespiti için Makine Öğrenmesi Algoritmalarının Özellik Tabanlı Performans Karşılaştırması

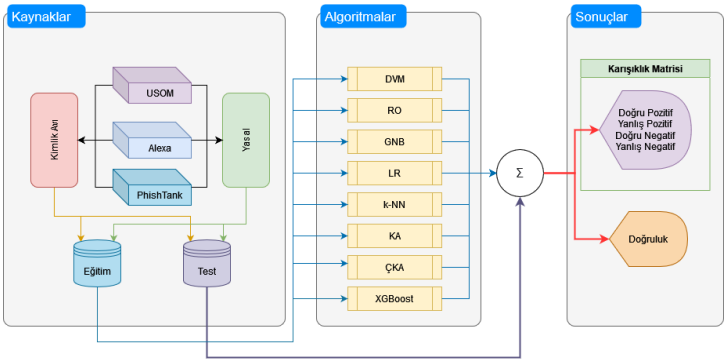
Feature-Based Performance Comparison of Machine Learning Algorithms for Phishing Detection through Uniform Resource Locator

Önemli noktalar (Highlights)

- *Sahte web siteleri üzerinden kimlik avı saldırılarının makine öğrenmesi algoritmalarıyla tespiti / Detection of phishing attacks over fake websites with machine learning algorithms*
- *Özellik kullanımı ve veri ön-işlemenin başarıma katkısı / Contribution of feature-usage and data pre-processing to success*

Grafik Özet (Graphical Abstract)

USOM, Alexa ve PhisTank sitelerinden veriler üzerinde makine öğrenmesi algoritmaları ve özellik çıkarımı ile birlikte yüksek başarımlı sınıflandırma sonucu elde edilmiştir. / High-performance classification results were obtained with machine learning algorithms and feature extraction on the data from USOM, Alexa and PhisTank websites.



Şekil. Çalışmanın blok diyagramı **/ Figure.** Block diagram of the study

Amaç (Aim)

Kimlik avı dolandırıcılığı web sitelerinin tespit edilmesi. / Detecting phishing scam websites.

Tasarım ve Yöntem (Design & Methodology)

Belirlenen özelliklerin makine öğrenmesi algoritmaları üzerinde test edilmesi. / Testing the determined features on machine learning algorithms

Özgünlük (Originality)

Kullanılan veri seti güncelliği ve çeşitliliğine ek olarak kullanılan özellikler ile yüksek başarım elde edilmesi. / Achieving high performance with the features used in addition to the up-to-dateness and diversity of the data set.

Bulgular (Findings)

Rastgele orman, karar ağaçları, çok katmanlı algılayıcı, XGBoost ve lojistik regresyon algoritmalarıyla %99.8 doğruluk. / 99.8% accuracy with random forest, decision trees, multilayer perceptron, XGBoost and logistic regression algorithms.

Sonuç (Conclusion)

Özellik çıkarımının makine öğrenmesi algoritmalarının başarımına katkısı. Güncel ve doğru özellikler kullanıldığında %99.8 oranına ulaşan doğruluk. / Contribution of feature extraction to the performance of machine learning algorithms. Up to 99.8% accuracy when using up-to-date and accurate features.

Etik Standartların Beyanı (*Declaration of Ethical Standards*)

Bu makalenin yazarları çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler. / The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Tekdüzen Kaynak Bulucu Yoluyla Kimlik Avı Tespiti için Makine Öğrenmesi Algoritmalarının Özellik Tabanlı Performans Karşılaştırması

Araştırma Makalesi / Research Article

Taki SAVAŞ[1], Serkan SAVAŞ[2*]

[1]Interprobe | Intelligence & Analytics Ankara, Türkiye

[2]Mühendislik Fakültesi, Bilgisayar Müh. Bölümü, Çankırı Karatekin Üniversitesi, Türkiye

(Geliş/Received : 11.12.2021 ; Kabul/Accepted : 14.02.2022 ; Erken Görünüm/Early View : 28.02.2022)

ÖZ

Günümüzde kimlik avı (oltalama/phishing) saldırılarına çok sık rastlanmaktadır. Bu tür saldırılar insanların kişisel bilgilerini ele geçirmek ya da insanları dolandırmak amacıyla gerçekleştirilmektedir. Kimlik avı saldırılarının birden fazla türü bulunmaktadır. Bu türlerden birisi de tekdüzen kaynak bulucu (uniform resource locater – URL) yoluyla gerçekleştirilen ve yaygın olarak rastlanılan saldırılardır. Bu çalışmanın amacı, URL adreslerinin farklı makine öğrenmesi algoritmaları kullanarak zararlı olup olmadığını sınıflandırmaktır. Çalışmada destek vektör makineleri, rastgele orman, Gauss Naive Bayes, lojistik regresyon, k-en yakın komşu, karar ağaçları, çok katmanlı algılayıcılar ve XGBoost algoritmaları olmak üzere sekiz farklı makine öğrenmesi algoritması kullanılmıştır. Eğitim ve test amaçlı kullanılmak üzere USOM, Alexa ve Phishtank üzerinden veriler elde edilmiştir. Bu verilere çeşitli veri ön-işleme adımları uygulanarak özellik çıkarımı gerçekleştirilmiştir. Araştırma sonucunda birden fazla modelde %99.8 doğruluk oranına ulaşılarak, makine öğrenmesi algoritmalarının bu alandaki başarımı kanıtlanmıştır.

Anahtar Kelimeler: Siber güvenlik, kimlik avı, makine öğrenmesi, internet alan adı, siber saldırı tespiti.

Feature-Based Performance Comparison of Machine Learning Algorithms for Phishing Detection through Uniform Resource Locator

ABSTRACT

Recently, phishing attacks are very common. Such attacks are carried out with the aim of obtaining personal information of individuals or defrauding individuals. There are multiple types of phishing attacks. One of these types is the common attacks carried out through the uniform resource locator (URL). The purpose of this study is to classify whether URL addresses are malicious or not using different machine learning algorithms. Eight different machine learning algorithms including support vector machines, random forest, Gaussian Naive Bayes, logistic regression, k-nearest neighbor, decision trees, multilayer perceptrons and XGBoost algorithms were used in the study. Data were obtained from USOM, Alexa, and Phishtank to be used for training and testing purposes. Feature extraction was performed by applying various data pre-processing steps to these data. As a result of the research, the accuracy of 99.8% in more than one model has been achieved, and the success of machine learning algorithms in this area has been proven.

Keywords: Cybersecurity, phishing, machine learning, domain, cyber-attack detection.

1. GİRİŞ (INTRODUCTION)

Hâlihazırda gerçek hayata dönüşen, insanların gerçek hayatla eşzamanlı olarak vakit geçirdiği sanal dünyada, bireylerin ve kurumların gerçek hayatta karşılaştıklarına benzer tehlikeler bulunmaktadır [1]. Zararlı yazılımlar kullanarak kişisel verilerin elde edilmesi, kurumların veya devletlerin özel bilgilerinin ifşa edilmesi, ticari şirketlerin servislerinin devre dışı bırakılması gibi pek çok siber güvenlik tehdidi, siber ortamlardaki tehlikeler arasında yer almaktadır. Bireysel kullanıcılar için bu tehlikelerin en başında dolandırıcılık gelmektedir. İnternet ortamında kimlik avı, oltalama/yemleme (Phishing) veya e-tuzak gibi terimlerle açıklanan dolandırıcılık türü bilişim araçlarıyla işlenen suçlar içerisinde en sık karşılaşılan suçlardan biridir. Federal Bureau of Investigation (FBI) 2020 raporuna göre 2020 yılı içerisinde bildirilen 780,403 suç içerisinden 241,342 tanesi (%31) kimlik avı (Phishing / Vishing / Smishing / Pharming) olarak tespit edilmiştir. Bu oran 2016 -2020 yılları arasında 11 kat artmıştır [2]. İnternet kullanımının her geçen gün daha fazla cihaz üzerinden yaygınlaşmasıyla birlikte ilerleyen yıllarda daha fazla artacağı düşünülmektedir.

Kimlik avı dolandırıcılığı; güvenilir bir şirket şeklinde (bu genelde rüştü ispat olmuş ulusal bir firma olacağı gibi uluslararası bir firma şeklinde de olabilir) veya kişi kılığına girerek, hile yoluyla karşısındaki kişinin banka

*Sorumlu Yazar (Corresponding Author)
e-posta : serkansavas@karatekin.edu.tr

şifresi, kredi kartı bilgileri, mail şifresi, kullanıcı şifresi gibi tamamen kişisel bilgilerini elde etmeyi amaçlar [3, 4]. Kimlik avı saldırılarının %96'sı e-posta yoluyla gerçekleştirilen saldırılar olup %3'ü ise kötü amaçlı siteler vasıtasıyla gerçekleştirilmektedir. Bu saldırıların %1'i ise telefon yoluyla gerçekleştirilmektedir [5, 6]. Internet yoluyla gerçekleştirilen bu dolandırıcılık faaliyetleri içerisinde e-posta vasıtasıyla yapılanlardan pek çoğu günümüzde artık e-posta hizmeti veren firmaların kullanmış olduğu yapay zekâ algoritmaları sayesinde engellenerek sahte mail kutusuna gönderilmektedir. Yine de buralardan da mağduriyet yaşayan kişiler de bulunmaktadır. Diğer taraftan, web siteleri üzerinden gerçekleştirilen kimlik avı dolandırıcılığının anti-virüs yazılımları, güvenli köprü metni aktarım protokolü (secure hypertext transfer protocol – https) ve web tarayıcılarının uyarı sistemlerine rağmen yine de kullanıcı dikkatine ve deneyimine bağlı olarak önlenebilir olması, özellikle bilişim okuryazarlığı yeterli seviyede olmayan kullanıcılar için büyük risk taşımaktadır.

Literatürde internet siteleri üzerinden kimlik avı dolandırıcılığının tespit edilebilmesine yönelik farklı çalışmalar gerçekleştirilmiştir. Yapay zekâ [7] çalışmalarının özellikle son 20 yılda hız kazanmasıyla birlikte makine öğrenmesi algoritmaları bu konuda önemli başarımlar elde etmeye başlamıştır. Korkmaz and Büyükgöze [3] tarafından Rastgele Orman (RO), Destek Vektör Makineleri (DVM), J48, k-En Yakın Komşu (k-Nearest Neighbor – kNN) ve Naive Bayes (NB) algoritmalarıyla gerçekleştirilen sınıflandırma çalışmasında %85 ile %95.64 arasında doğruluk oranları elde edilmiştir. En yüksek başarımı ise %95.64 ile RO algoritması yakalamıştır. Bir başka çalışmada ise DVM ve NB sınıflandırıcısı 33,000 Uniform Resource Locator (URL - Tekdüzen Kaynak Bulucu) üzerinde kimlik avı veya yasal site sınıflandırması için kullanılmış ve %90'ın üzerinde başarım elde edilmiştir [8]. Kadı [9] ise çalışmasında kNN, DVM ve yapay sinir ağları (YSA) algoritmalarını test etmiş ve YSA ile %98.71 başarıma ulaşmıştır. NB, DVM, RO, J48, AdaBoost ve sinir ağları (SA) algoritmaların başarımlarının karışıklık matrisi ve f-score üzerinden değerlendirildiği bir çalışmada kimlik avı saldırıları %73 ile %83 oranlarında tespit edilirken, web sitesi benzerlikleri ise %81 ile %92.5 oranları arasında yakalanabilmiştir [10]. Bu çalışmalara ek olarak kimlik avı dolandırıcılığını tespit etmek için alan adından özellik çıkarımı yöntemini kullanan [11-13], web sitesi kaynak kodu üzerinden kimlik avı karakteristiklerini çıkaran [14], sözcüksel özelliklere göre sınıflandırma yapan [15], tanımlayıcı [16] veya kural tabanlı yaklaşım kullanılan [17], doğal dil işleme teknikleri ile %97.98 başarım oranına ulaşılan [18] çalışmalar da bulunmaktadır. Ayrıca NB, kNN, RO algoritmalarına ek olarak C4.5, ID3, PRISM ve RIPPER algoritmaları da kullanılmış ve bunlarla da %95 ile %96.5 arasında değişen oranlarda başarımlar elde edilmiştir [19]. Belirtilen bu araştırmalarda elde edilen doğruluk oranlarının %90 başarımın üzerinde olmasına rağmen, sürekli gelişmekte olan makine öğrenmesi çalışmaları nedeniyle hala geliştirilebilir düzeyde olduğu görülmektedir. Özellikle topluluk öğrenme modelleri [20], özellik seçimi ve hibrid makine öğrenmesi algoritmaları [21] ile birlikte her geçen gün çeşitlenen çalışmalar gerçekleştirilmektedir. Bu çalışmada gerçekleştirilen özellik çıkarımı yoluyla makine öğrenmesi algoritmalarının performansının arttırılması, araştırmanın önceki çalışmalardan farkını ve özgünlüğünü ortaya koymaktadır. Araştırmanın sonuçları da kullanılan yöntemin makine öğrenmesi algoritmalarının başarımlarını arttırdığını kanıtlar nitelikte olmuştur.

Web sitelerindeki aykırı davranışların tespitinde de farklı araştırmalarda makine öğrenmesi yaklaşımları kullanılırken [22] özellikle son yıllarda derin öğrenme çalışmalarının artmasıyla birlikte derin öğrenme algoritmaları da bu konuda kullanılmaya başlanmıştır. Uzun Kısa Süreli Bellek (UKSB) algoritması ikili sınıflandırmada %97.25 başarım elde ederken evrişimli sinir ağları (ESA) algoritması ise %98.86 doğruluk elde etmiştir. Çoklu sınıflandırmada ise UKSB ve ESA algoritmaları sırasıyla %91.13 ve %95.37 oranlarında doğruluk elde etmiştir [23]. Derin öğrenme algoritmalarının bu oranları da yine makine öğrenmesi algoritmalarına benzer şekilde geliştirilebilir olduğunun göstergesi olmuştur. Awadh and Akbaş [24] tarafından gerçekleştirilen başka bir çalışmada ise TF.IDF ve C4.5 algoritmaları, kötü amaçlı verilerin etkin bir şekilde algılanmasını sağlamak için birlikte kullanılmış ve MLP ve NB algoritmalarından daha yüksek başarım elde edilmiştir.

Buradan yola çıkarak bu çalışmada önemli web sitesi veri tabanlarından elde edilen veriler üzerinde, özellik çıkarımı yöntemi kullanılarak, makine öğrenmesi algoritmalarının başarımlarının karşılaştırmalı analizi için kapsamlı bir araştırma gerçekleştirilmiştir. Bu çalışmanın ikinci bölümünde kullanılan materyal ve metotlar belirtilirken üçüncü bölümde elde edilen sonuçlar açıklanmıştır. Son bölümde ise tartışma, sonuç ve gelecek çalışma ile ilgili bilgiler sunulmuştur.

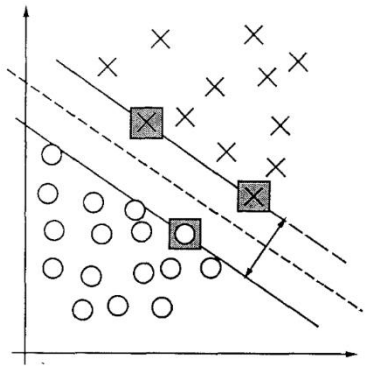
2. MATERYAL VE METOD (MATERIAL and METHOD)

Makine öğrenmesi algoritmaları günümüzde sanayiden [25] eğitime [26, 27], sağlıktan [28, 29] ticarete [30, 31] kadar pek çok farklı alanda kullanılmaktadır. Bu çalışmada da literatür araştırması sırasında tespit edilen farklı algoritmaların farklı sayılarda veri üzerindeki başarımlarının doğruluk ve güvenirliğinin bir standarda bağlanabilmesi, kapsamlı bir analiz gerçekleştirilerek sonuçların açıklanması amacıyla sekiz farklı makine öğrenmesi algoritması kullanılmıştır:

- Destek vektör makineleri,
- Rastgele orman,
- Gaussian Naive Bayes (GNB),
- Lojistik regresyon (LR),

- K - en yakın komşu,
- Karar ağaçları (KA),
- Çok katmanlı algılayıcı (ÇKA),
- XGBoost (eXtreme Gradient Boosting).

Destek vektör makineleri, sınıflandırma problemleri için kullanılan bir makine öğrenmesi algoritmasıdır. Makine kavramsal olarak şu fikri uygular: girdi vektörleri, çok yüksek boyutlu bir özellik uzayına doğrusal olmayan bir şekilde eşlenir. Bu öznitelik uzayında doğrusal bir karar yüzeyi oluşturulur. Karar yüzeyinin özel özellikleri, öğrenme makinesinin yüksek genelleme kabiliyetini sağlar (Şekil 1) [32].



Şekil 1. İki boyutlu düzlem üzerinde DVM ile sınıflandırılmış iki sınıf (Two classes classified by SVM on a two-dimensional plane)

Şekil 1'de görüldüğü gibi DVM ile temel olarak iki sınıf arasına bir doğru çizilerek bu yolla her iki sınıfın birbirine mümkün olan en fazla uzaklıkta olması amaçlanır. Denetimli bir makine öğrenmesi algoritmasıdır.

Rastgele orman algoritması da DVM gibi denetimli bir makine öğrenmesi algoritmasıdır. Algoritmanın temeli, özellik uzayının rastgele seçilmiş alt uzaylarında birden çok ağaç oluşturmaya dayanmaktadır Farklı alt uzaylardaki ağaçlar, sınıflandırmalarını tamamlayıcı yollarla genelleştirir ve birleşik sınıflandırmaları monoton olarak geliştirilebilir [33]. Düğümlerin dağıtılması için hesaplama Eşitlik (1)'deki gibi gerçekleştirilir [34].

Bir x noktası için, T_j $(j = 1,2,\ldots,t)$ ağacından aşağı indiğinde x'in atandığı uçbirim düğümü $v_j(x)$ olsun. Buna göre, x'in c $(c = 1,2,\ldots,n)$ sınıfına ait olduğu sonsal olasılık $P(c|v_j\ (x))$ ile gösterilsin:

$$P\left(c\middle|v_j\ (x)\right) = \frac{P(c,v_j(x))}{\sum_{l=1}^{n} P(c_l,v_j(x))} \quad (1)$$

$v_j(x)$ 'e atanan tüm noktalar üzerindeki c sınıfı puanların kesri ile tahmin edilebilir. Bu bağlamda, ağaçlar tamamen bölündüğünden, çoğu terminal düğümünün yalnızca tek bir sınıf (anormal duruşlar hariç) ve dolayısıyla tahmin değeri içerir. $\hat{P}(c|v_j\ (x))$ neredeyse her zaman 1'dir. Diskriminant fonksiyonu ise Eşitlik (2)'de gösterildiği gibi tanımlanır [33].

$$g_c(x) = \frac{1}{t}\sum_{j=1}^{t} \hat{P}(c|v_j(x)) \quad (2)$$

Burada karar kuralı, x 'i $g_c(x)$ 'in maksimum olduğu c sınıfına atamaktır.

Naive Bayes bir olasılık sınıflandırıcı algoritmasıdır. Bu algoritmada veri içerisinde bulunan değerlerin frekansları ve bunların aralarındaki kombinasyonlar sayılarak bir olasılık kümesi oluşturulur [35]. Algoritma matematikçi Thomas Bayes teoremini (Eşitlik 3) kullanır ve sınıf değişkeninin değeri göz önüne alındığında tüm değişkenlerin bağımsız olduğunu varsayar [36].

$$p(c_j|d) = \frac{p(d|c_j)\,p(c_j)}{p(d)} \quad (3)$$

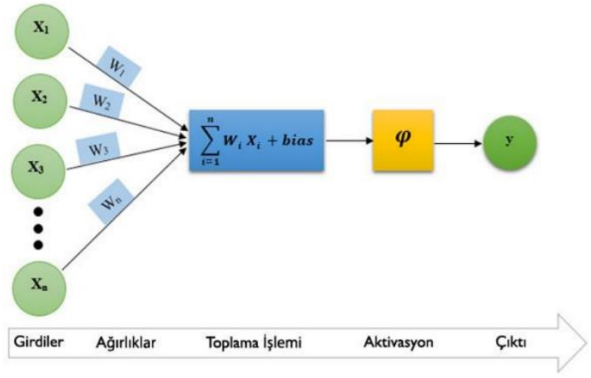
Denklemde d bilinen durum, c_j ise farklı koşullu durumları ifade etmektedir.

Lojistik Regresyon ise doğrusal sınıflandırma problemlerinde kullanılan ve bir veya daha fazla bağımsız değişken kullanılarak sonuca ulaşmayı amaçlayan bir istatistiksel olasılık hesaplama yöntemdir.

K en yakın komşu algoritması [37] denetimli öğrenme yöntemleri içerisinde etkili sınıflandırma algoritmalarından bir tanesidir. Öğrenme kümesi içerisinde bulunan normal davranış benzerlikleri hesaplanarak, en yakın görülen "k" verinin ortalamasının oluşturduğu eşik değerine göre sınıflandırma işlemi gerçekleştirilir. Burada sınıflandırma işlemi için öncelikle, sınıf özelliklerinin belirlenmiş olması gerekmektedir [38]. Ayrıca "k" komşu sayısı, benzerlik ölçümü, eşik değeri ve öğrenme kümesindeki normal davranışların yeterliliği, bu yöntemin performansını etkileyen faktörlerdendir [39]. K en yakın komşu algoritmasının performansı için önemli parametreler komşu sayısı "k", ağırlıklandırma yöntemi ve uzaklık ölçütüdür. Başlıca uzaklık ölçütleri olarak, Minkowski, Öklid ve Manhattan, uzaklıklarından bahsedilebilir [36].

Karar ağaçları, bir nesnenin sınıfını yordayıcı değişkenlerin değerlerinden tahmin etmek için kullanılan deneysel bir kuraldır [40]. Tahminsel bir model olup ağaç yapısında ilerlediğinden bu isim verilmiştir. Ağacın dalları ve yaprakları, sınıflandırma probleminin çözümünü içeren parçalardır [36].

Çok katmanlı algılayıcılar, "perceptrons" olarak adlandırılan basit sinir hücreleri ağıdır. Temel fikir olan tek perceptron ilk olarak Rosenblatt [41] tarafından tanıtılmıştır. Çok katmanlı algılayıcılar hata üzerine dayandırılmış bir öğrenme algoritmasıdır. Temel iki işlevi olan öğrenme ve karar verme aşamalarını ağırlıklandırma, aktivasyon fonksiyonu ve bias sayesinde yapar. Ağırlık, her girdinin bir sonraki aşamaya gitmeden önce çarpıldığı katsayıdır. Tüm girdiler, kendilerine ait ağırlıklarla çarpılarak toplanır. Daha sonra aktivasyon fonksiyonuna bu değerin gönderilmesi sonucu ortaya çıkan cevap, sistemin kararı olur. Bias ise kullanıcıdan kullanıcıya, mekanizmanın çalışma şekline veya amacına göre değişebilen, kullanıcı tarafından eklenen bir parametredir [42].



Şekil 2. Algılayıcının çalışma mantığı (Working logic of the perceptrons) [36]

XGBoost, Gradient Boosting algoritmasının çeşitli düzenlemeler ile optimize edilmiş yüksek performanslı halidir. Yüksek tahmin gücü, aşırı öğrenmenin önüne geçilmesi, boş verilerin yönetimi ve hızlı çalışma özellikleriyle birlikte Chen and Guestrin [43] tarafından geliştirilmiştir.

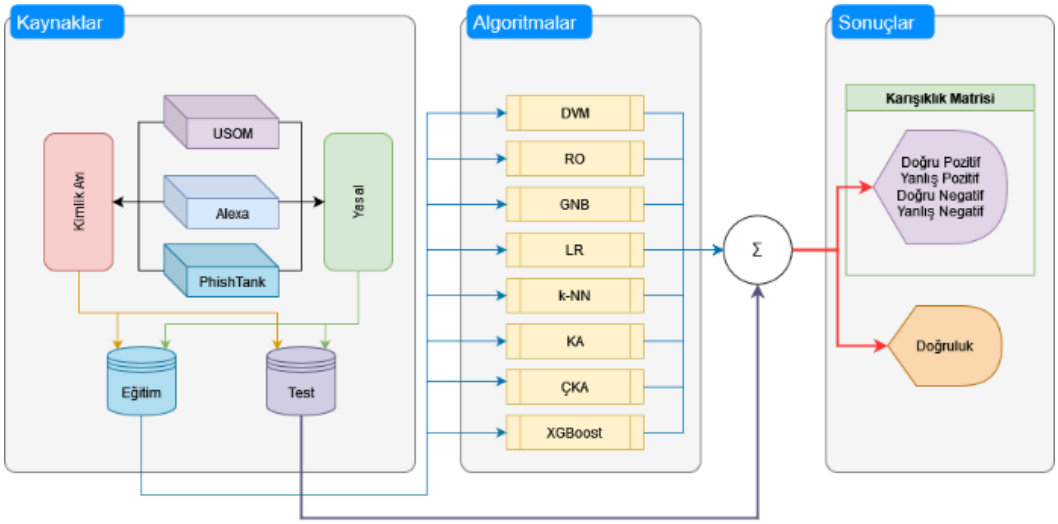
Yapay zekâ çalışmalarında, makine öğrenmesi algoritmalarının etkililiğini ve performansını etkileyen unsurların en başında özellik çıkarımı yer almaktadır. Özellik çıkarımı ve kullanılan özellikler, algoritmanın doğruluğuna doğrudan etki etmektedir. Bu nedenle günümüzde sıklıkla özellik çıkarımı algoritmalarını kullanarak [44] veya belirli kurallara göre özelliklerin çıkarılıp kullanıldığı [45] farklı makine öğrenmesi çalışmaları gerçekleştirilmektedir. Bu çalışmada da performansları test edilecek olan algoritmaların sınıflandırma işlemlerini gerçekleştirebilmeleri için Mohammad, Thabtah [46] tarafından düzenlenen web sitelerinin belirli kimlik avı özellikleri kullanılmıştır. Bu özellikler ve açıklamaları Çizelge 1'de gösterilmiştir [46].

Çizelge 1'de gösterilen özelliklerin kullanım amaçları şöyle açıklanabilir. Bir URL'nin up/down durumu ve SSL protokolü kontrol edilir. URL ile bağlantı kurularak SSL sertifikası çekilir. URL içerisinde Internet Protokol (IP) adresi geçiyor ise kullanıcının bilgileri çalınma ihtimali bulunmaktadır. Kimlik avcıları ya da dolandırıcılar, adres çubuğundaki şüpheli kısmı gizlemek için uzun URL kullanabilirler. Bu yüzden dolayı URL uzunluğu da kontrol edilmektedir. URL kısaltma, "World Wide Web" de, bir URL'in uzunluğunun önemli ölçüde daha küçük hale getirilebileceği ve gerekli web sayfasına yönlendirebileceği bir yöntemdir. Böylelikle gerçek domain name gizlenmiş olur. URL'de "@" sembolünün kullanılması, tarayıcının "@" sembolünden önceki her şeyi görmezden gelmesine neden olur ve gerçek adres genellikle "@" sembolünü takip eder. URL yolu içinde "//" ifadesinin bulunması da, kullanıcının başka bir web sitesine yönlendirileceği anlamına gelir. "-" sembolü de meşru URL'lerde nadiren kullanılır. Kimlik avcıları, kullanıcıların meşru bir web sayfasıyla uğraştıklarını hissetmeleri için alt alan adına (-) ile ayrılmış önekler veya sonekler ekleme eğilimindedir. Bir diğer kimlik avı özelliği ise alt etki alanı sayısıdır. Kimlik avcıları kullanıcıyı yanılgıya düşürmek için domain name alırken başında https veya http ifadesi ekleyerek de alan adı alabilirler. Ayrıca, kimlik avı web siteleri kısa bir süre içerisinde yaşadığı için Alexa veri tabanı tarafından tanınmayabilir.

Belirtilen algoritmaların deneysel testlerinin gerçekleştirilebilmesi için uluslararası farklı veri tabanlarından web site bilgileri çekilerek veri tabanı oluşturulmuştur. Algoritmaların eğitimi için 5,000 kimlik avı sitesi USOM [47] üzerinden elde edilmiş, Alexa [48] üzerinden de 5,000 adet yasal web site URL adresi elde edilmiştir. Modellerin eğitimi sonrasında test işlemi için 4,000 adet Alexa, 1,935 adet USOM ve 2,000 adet PhishTank [49] verisi kullanılmıştır. Modellerin kodlanması, eğitim ve test işlemlerinde Python programlama dili kullanılmıştır. Çalışmanın blok diyagramı Şekil 3'te gösterilmiştir.

Çizelge 1. Sınıflandırma işleminde kullanılan özellikler (Features used in classifications)

Özellik	Açıklama
Up veya Down	Up ise 1, Down ise -1 değeri döner.
IP kontrol	IP adresi varsa -1, yoksa 1 değeri döner.
URL uzunluğu kontrolü	URL uzunluğu 54 den küçükse 1, 54 ile 75 arasında ise 0, 75 den büyük ise -1 değeri döner.
Tiny URL kontrolü	Eğer URL kısaltılmış ise -1, değilse 1 değeri döner.
"@" Sembolü Kontrolü	Eğer "@" sembolü içeriyorsa -1, içermiyorsa 1 değeri döner.
"//" Kontrolü	Eğer URL "//" ifadesini içeriyorsa -1, içermiyorsa 1 değeri döner.
"-" Prefix Suffix Kontrolü	Eğer URL "-" sembolü içeriyorsa -1, içermiyorsa 1 değeri döner.
"." ile Alt Etki Alanı Kontrolü	"www." ve "ccTLD" ülke bölümleri kaldırılarak nokta sayıları belirlenir. Eğer bire eşit ve küçük ise 1, ikiye eşit ise 0, ikiden büyük ise -1 değeri döner.
URL içinde https ve http Kontrolü	Eğer URL bu ifadeleri içeriyorsa -1, içermiyorsa 1 değeri döner.
SSL Sertifika Kontrolü	Eğer sitenin SSL sertifikası yok ise (http) -1 değeri dönmektedir. Eğer var ise; güvenilir bir sertifika ve sertifika günü 360 günden fazla ise 1, güvenilir ve 360 günden düşük ise 0, güvenilmez ve 360 günden fazla ise 0, hem güvenilmez hem de 360 günden az ise -1 değeri döner.
URL Yaşının Kontrolü	Eğer, domain oluşturulalı 2 yıldan fazla olduysa 1, olmadıysa -1 değeri döner.
Web Site Trafiği	Eğer Website rankı 100,000'den düşük ise 1, 100,000'den büyük ise 0, Alexa veritabanında yok ise -1 değeri döner.



Şekil 3. Çalışmanın blok diyagramı (Block diagram of the study)

Çalışmada kullanılan verilerin elde edilmesi için web scraping uygulaması yazılmıştır. Bu uygulama yardımıyla çekilen veriler metin formatında URL adresi olarak gelmektedir. Makine öğrenmesi algoritmalarıyla eğitebilmek için gelen bu metin verilerinden özellik çıkarımı yapılması gerekmektedir. Veri ön-işleme aşamasında, kaynaklardan çekilen veriler Tablo 1'de belirtilen 12 farklı özellik için yazılan fonksiyonlarla sayısal hale getirilmiştir.

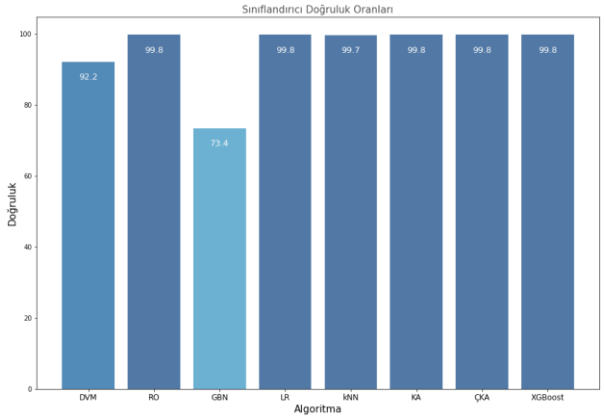
Özellik-1 için yazılan fonksiyonda "requests" modülü kullanılmıştır. Bu modül ile fonksiyona giren URL üzerinden istek atılmış ve geri dönen cevabın durum kodu 200 ise "1" değilse "-1" değeri kaydedilmiştir. Özellik-2 için yazılan fonksiyonda, URL içerisinde IP geçip geçmediğini kontrol edilmiş olup "regex" modülü kullanılmıştır. Gerekli regex komutları yazılarak URL içerisinde hexadecimal ya da normal formatta IP bulunma durumu kontrol edilmiş, eğer içeriyorsa "-1" içermiyorsa "1" değeri kaydedilmiştir. Özellik-3 için URL uzunluğunun 54'ten büyük olup olmadığını kontrol eden fonksiyonda "len" modülü kullanılarak URL uzunluğu hesaplanmıştır. Yapılan araştırmalara göre uzunluğu 54'ten daha fazla olan URL adreslerinin büyük çoğunluğu phishing amaçlı kullanılmaktadır. Bu yüzden URL uzunluğu 54'ten küçük olanlar "1", 54 ile 75 arasında olanlar "0" ve 75'ten büyük olanlar ise "-1" olarak kaydedilmiştir. Özellik-4'te URL kısaltmak için herhangi bir yapı kullanılıp kullanılmadığını kontrol eden fonksiyon yazılmış ve en çok kullanılan URL kısaltma isimlerinin geçip geçmediği kontrol edilmiştir. Eğer kısaltma geçiyorsa "-1", geçmiyorsa "1" değeri kaydedilmiştir. Özellik-5, 6 ve 7 için yazılan fonksiyonlarda URL içerisinde sırasıyla "@", "//" ve "-" sembollerinin kullanılıp kullanılmadığı kontrol edilmiştir. Eğer URL bu sembolleri içeriyorsa "-1", içermiyorsa "1" değeri kaydedilmiştir. Özellik-8'de URL içerisinde çoklu alt etki alanı içerip içermediği kontrol edilmiştir. Bu fonksiyonda URL içerisindeki noktaların sayısını kontrol etmek için URL içerisinden "www" ve "ccTLD" yani ülke kodlarına ait olan noktalar çıkarıldıktan sonra toplam nokta sayısı hesaplanır. Eğer nokta sayısı birden küçük ve eşitse "1", ikiye eşitse "0" ve ikiden fazla ise "-1" değeri kaydedilmiştir. Özellik-9, URL içerisinde http veya https geçip geçmediğini kontrol eden fonksiyondur. Eğer URL adı içerisinde "http" veya "https" ifadesi geçiyorsa "-1", geçmiyorsa "1" değeri kaydedilmiştir. Özellik-10 için URL'nin SSL sertifikasının olup olmadığı, varsa ne kadar sürelik sertifika olduğu ve güvenilir olup olmadığını kontrol eden fonksiyon yazılmıştır. Yazılan fonksiyonda "SSL" modülü kullanarak URL'nin SSL sertifikası varsa sertifikaya ait bilgiler bir değişkene çekilmiştir. Çekilen bilgiler sayesinde sertifikanın süresinin değeri hesaplanmıştır. Güvenilir sertifika sağlayıcıların listesi çekilerek şu değerlendirmeler yapılmıştır:

- Sertifika güvenilir ve süresi 1 yıldan uzunsa "1",
- Sertifika güvenilir ancak süresi 1 yıldan az ise "0",
- Sertifika güvensiz ancak süresi 1 yıldan fazlaysa "0",
- Hem güvensiz hem de süresi 1 yıldan az ise "-1",

değerleri kayıt edilmiştir. Özellik-11 için yazılan fonksiyonda URL yaşı hesaplanmıştır. "whois" modülü yardımı ile URL bilgileri çekilmiş, URL başlama ve sona erme tarihleri üzerinden URL'nin yaşı hesaplanmıştır. Eğer URL yaşı iki yıldan küçükse "-1", büyükse "1" değeri kaydedilmiştir. Özellik-12 için yazılan fonksiyonla Alexa üzerinden "BeautifulSoup" modülü kullanılarak URL sıralama bilgisi çekilmiştir. Eğer sıralaması 100.000'den küçükse "1", değilse "0" ve Alexa veri tabanında yoksa "-1" değeri kaydedilmiştir.

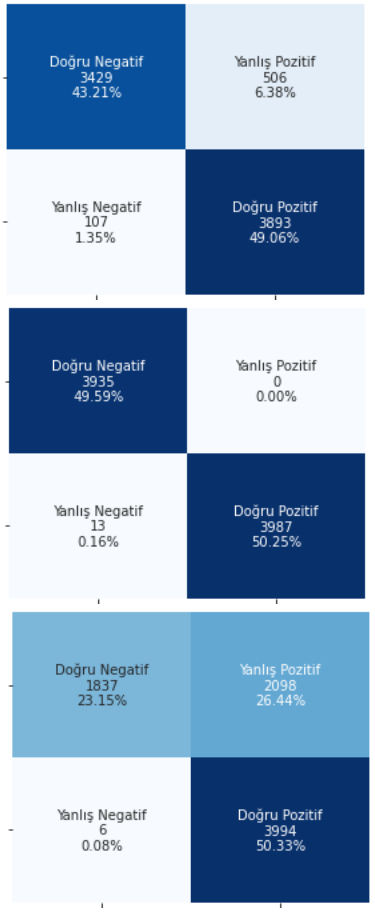
3. DENEYSEL SONUÇLAR (EXPERIMENTAL RESULTS)

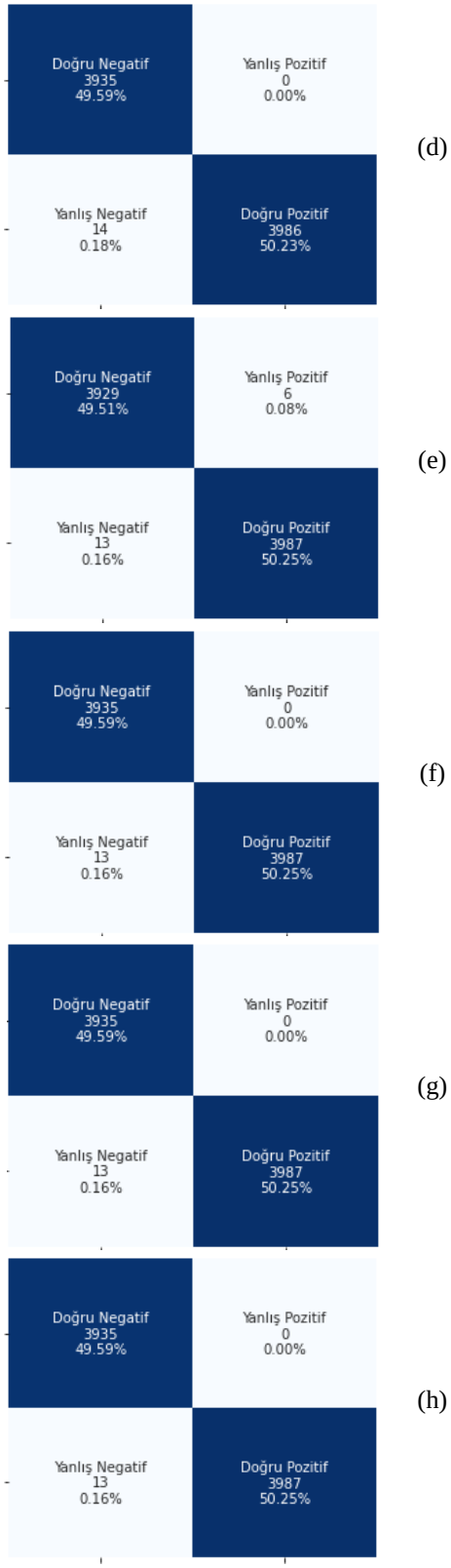
Çalışmada ilk olarak makine öğrenmesi algoritmalarının eldeki veri seti üzerindeki doğruluk oranları incelenmiştir. Her bir algoritmanın eğitim işleminden sonra test verileri üzerindeki doğruluk oranları Şekil 4'te gösterilmiştir.



Şekil 4. Algoritmaların doğruluk oranları (Accuracy rates of the algorithms)

Şekil 4'te görüldüğü gibi algoritmalar içerisinde RO, LR, KA, ÇKA ve XGBoost algoritmaları %99.8 doğruluk oranı ile yüksek başarım oranlarına ulaşmışlardır. Doğruluk oranlarında en düşük oranı %73.4 ile GNB algoritması elde etmiştir. Sonrasında ise DVM %92.2 doğruluk oranı ile en başarısız ikinci algoritma olmuştur. K-NN algoritması %99.7 doğruluk oranı ile diğer algoritmalara çok yakın bir sonuç üretmiştir. Doğruluk oranlarındaki bu sonuçların derinlemesine ve ayrıntılı incelenebilmesi çalışmada karışıklık matrisi de oluşturulmuş ve hassasiyet, hatırlama ve f1-puan değerleri de incelenmiştir. Modellerin karmaşıklık sonuçlar Şekil 5'te sırasıyla gösterilmiştir.





Şekil 5. Algoritmaların karışıklık matrisleri (a) DVM, (b) RO, (c) GNB, (d) LR, (e) kNN, (f) KA, (g) ÇKA, (h) XGBoost. (Confusion matrices of the algorithms (a) SVM, (b) RF, (c) GNB, (d) LR, (e) kNN, (f) DT, (g) MLP, (h) XGBoost)

Şekil 5'te görüldüğü gibi %99.8 başarıma ulaşan modeller içerisinde RO, KA, ÇKA ve XGBoost

algoritmaları karmaşıklık matrislerinde de aynı sayılarda "doğru negatif, yanlış negatif, doğru pozitif ve yanlış pozitif" değerleri elde etmişlerdir. Bu algoritmalar içerisinde sadece LR, bir tane siteyi yanlış negatif olarak fazla işaretlemiştir. Doğruluk oranları diğer algoritmalar göre daha düşük olan DVM, GNB ve k-NN algoritmalarının karışıklık matrisleri ise diğer algoritmalara göre belirgin şekilde yanlış tahminler gerçekleştirmiştir.

Karışıklık matrisinde doğruluk hesaplaması, sınıflandırıcının ne sıklıkla doğru olduğunu verir. Hassasiyet, tüm sınıfların ne kadar doğru tahmin edildiğinin bir ölçüsüdür. Pozitif tahmin değeri olarak da bilinir. Hatırlama (duyarlılık), doğru sınıflandırılan toplam pozitif örnek sayısının toplam pozitif örnek sayısına bölümü olarak tanımlanabilir. F1 puanı, kesinlik ve hatırlamanın harmonik ortalamasıdır. Sınıflandırıcının ne kadar iyi performans gösterdiğinin bir ölçüsüdür ve genellikle sınıflandırıcıları karşılaştırmak için kullanılır [50]. Karışıklık matrisi hesaplamaları Eşitlik 4'te gösterildiği gibidir [50].

$$
\begin{aligned}
Doğruluk &= \frac{(DP + DN)}{DP + DN + YP + YN} \\
Hassasiyet &= \frac{DP}{DP + YP} \\
Hatırlama &= \frac{DP}{DP + YN} \\
F1\ Puanı &= \frac{2DP}{2DP + YP + YN}
\end{aligned}
\tag{4}
$$

Eşitlikte doğru pozitif için DP, yanlış pozitif için YP, doğru negatif için DN ve yanlış negatif içinse YN kısaltmaları kullanılmıştır.

Bu hesaplamalar kullanılarak her bir algoritma için araştırmada ortaya çıkan doğruluk, hassasiyet, hatırlama ve F1-puanı sonuçları Çizelge 2'de gösterilmiştir.

Çizelge 2. Algoritmaların karışıklık metrikleri (Confusion metrics of the algorithms)

Algoritma	Doğruluk (Accuracy)	Hassasiyet (Precision)	Hatırlama (Recall)	F1-Puanı (Score)
DVM	%92.2	%88.50	%97.33	%92.70
RO	%99.8	%100.00	%99.68	%99.84
GNB	%73.4	%65.56	%99.85	%79.15
LR	%99.8	%100.00	%99.65	%99.82
kNN	%99.7	%99.85	%99.68	%99.76
KA	%99.8	%100.00	%99.68	%99.84
ÇKA	%99.8	%100.00	%99.68	%99.84
XGBoost	%99.8	%100.00	%99.68	%99.84

Çizelge 2'deki karışıklık matrisi metrikleri incelendiğinde, en başarılı sonuçların hassasiyet değerlerinde elde edildiği görülmektedir. Algoritmaların hatırlama sonuçları, yine %99'un üzerinde gerçekleşmiştir. Burada ilginç bir analiz sonucu da ortaya çıkmıştır. Doğruluk oranı düşük olan GNB algoritmasının hassasiyet oranı, diğer algoritmalardan yüksek çıkmıştır. Ancak bu algoritmanın hassasiyet oranı %65.56 gibi çok düşük bir değer olduğu için, bu sonuç F1-puanı ve doğruluk oranını da etkilemiştir. Karışıklık matrisi metrikleri incelendiğinde RO, KA, ÇKA ve XGBoost algoritmalarının bu alandaki en başarılı algoritmalar olduğu ortaya çıkmıştır. Sadece pozitif sınıflar üzerine odaklanılan bir araştırma gerçekleştirileceği zaman, yüksek sınıflama oranı nedeniyle GNB algoritması da tercih edilebilir bir algoritmadır.

4. TARTIŞMA VE SONUÇ (DISCUSSION AND CONCLUSION)

4.1. Tartışma (Discussion)

Makine öğrenmesi tekniklerinin özellikle son yirmi yılda pek çok farklı disiplinde başarılı sonuçlar üretmesi, siber güvenlik alanında da bu algoritmaların sıklıkla kullanılmasının yolunu açmıştır. Siber saldırı çalışmalarında da kullanılan bu algoritmalar, kimlik avı saldırılarının tespiti ve sınıflandırılması için de kullanılmıştır. Literatürde gerek farklı özellik türlerinin gerekse de farklı algoritmaların kullanıldığı çalışmalar Çizelge 3'te gösterilmiştir.

Çizelge 3'te görüldüğü gibi makine öğrenmesi algoritmaları farklı veri setleri üzerinde farklı kişiler tarafından uygulanarak test edilmiştir. Bu algoritmalar içerisinde ön plana çıkan rastgele orman algoritması %93.75 [13], %95.64 [3], %97.3 [19], %89.93 [58], %97.9 [54], %97.98 [18], %94.6 [57] ve %98.11 [53] gibi yüksek oranlarda başarımlar sağlamıştır. Ancak bu çalışmada aynı algoritma %99.8 oranı ile gerçekleştirilen çalışmalar içerisinde en başarılı sonuca ulaşmıştır. Buradaki başarım artışı, gerçek ve güncel veri seti elde etme ve kullanılan özellik çıkarımı kurallarının uygunluğundan kaynaklanmaktadır. Benzer şekilde bu çalışmada %99.8 oranında başarım sağlayan lojistik regresyon algoritması, farklı araştırmalarda %91.90 [13], %99.2 [52], %86.5 [54] ve %98.4 [12] oranları elde etmiştir. Ayrıca tabloda görüldüğü gibi farklı çalışmalarda kendi yaklaşımlarını öneren araştırmacılar olmuştur ve %88.4 [22] ile %99.7 [11] oranlarında başarımlara ulaşılmıştır.

Çizelge 3 incelendiğinde bu çalışmada %99.8 başarım oranına ulaşan diğer algoritmalar olan karar ağaçları, çok katmanlı algılayıcılar ve XGBoost algoritmalarının da farklı araştırmalarda daha düşük başarım oranları elde ettiği görülmektedir. Buradan yola çıkarak özellik çıkarımının makine öğrenmesi algoritmalarının başarımını ne kadar etkilediği anlaşılabilir. Doğru yöntemler uygulandığında makine öğrenmesi algoritmalarının başarımları daha fazla artırılabilir. Bir diğer öne çıkan konu, elde edilen veri setinin önemidir. Araştırmalarda kullanılan veri setlerinin tercih edilen özellik çıkarımı yöntemine uyarlanması, algoritmaların öğrenimine katkı sağlamakta ve başarıyı arttırmaktadır. Bu araştırmada kullanılan ve Mohammad, Thabtah [46] tarafından düzenlenen kimlik avı saldırıları özellikleri, makine öğrenmesi algoritmalarının başarımlarını önemli ölçüde etkilemiştir.

Çizelge 3. Daha önce gerçekleştirilen çalışmalar (Previous studies)

Yazar(lar)	Yöntem	Sonuç(lar)	Veri seti
Shirazi, Bezawada [11]	Önerilen yaklaşım	%99.7	1,000 kimlik avı, 1,000 yasal
Arslan [52]	LR (+Doc2Vec)	%99.2	165,372 örnek
Moghimi and Varjani [17]	DVM	%99.14	1448 kimlik avı, 686 yasal
Uçar, İncetaş [23]	LSTM, CNN	LSTM: %97.25, CNN: %98.86	36,697 örnek
Jain and Gupta [12]	LR	%98.4	2,544 örnek
Almseidin, Zuraiq [53]	RO	%98.11	5,000 kimlik avı, 5,000 yasal
Özker [54]	DVM, NB, RO, LR, kNN, KA, ÇKA, XGBoost	DVM: %92.7, NB: %83.4, RO: 97.9, LR: %86.5, kNN: %94.3, KA: %96.3, ÇKA: %93.5, XGBoost: %97.8	8,353 kimlik avı, 5,438 yasal
Koşan, Yıldız [19]	C4.5, ID3, PRISM, RIPPER, NB, kNN, RO	C4.5: %95.9, ID3: %96.5, PRISM: %95.8, RIPPER: %95, NB: %93, kNN: %96, RO: %97.3	4,898 kimlik avı, 6,157 yasal
Sahingoz, Buber [18]	Doğal Dil İşleme tabanlı: KA, AdaBoost, Kstar, kNN, RO, Sıralı Minimum Optimizasyon (SMO), NB	KA: %97.2, AdaBoost: %93.24, Kstar: %95.27, kNN: %95.86, RO: %97.98, SMO: %94.92, NB: %95.86	37,175 kimlik avı, 36,400 yasal
Zhang, Yan [13]	SMO, NB, RO, LR	SMO: %95.83, NB: %92.94, RO: %93.75, LR: %91.90	3,000 örnek
Korkmaz and Büyükgöze [3]	RF, J48, SVM, KNN ve NB	RO: %95.64, J48: %93.57, DVM: %88.62, kNN: %90.24, NB: %85.07	548 yasal, 702 kimlik avı, 103 şüpheli
İncir [55]	Çok Katmanlı YSA	%95.42	5,000 kimlik avı, 5,000 yasal
Abu-Nimeh, Nappa [56]	LR, CART, BART, DVM, RO, NN	Hassasiyet, Hatırlama ve F1-Puanı: LR: %95.11, %82.96, %88.59, CART: %92.32, %87.07, %89.59, DVM: %92.08, %82.74, %87.07, NN: %94.15, %78.28, %85.45, BART: %94.18, %81.08, %87.09, RO: %91.71, %88.88, %90.24	2,889 Örnek
Chiew, Tan [57]	RO (+HEFS)	%94.6	5,000 kimlik avı, 5,000 yasal
Sanglerdsinlapachai and Rungsawang [10]	CANTANIA	F-Ölçümü: %92.5	100 kimlik avı, 100 yasal
Kalaycı [58]	AdaBoost, ÇKA, DVM, KA, kNN, NB, RO	AdaBoost: %84.45, ÇKA: %86.84, DVM: %84.92, KA: %87.46, kNN: %86.29, NB: %81.92, RO: %89.93	1,353 örnek
Pan and Ding [22]	Önerilen yaklaşım	%88.4	279 kimlik avı, 100 yasal

4.2. Sonuç (Conclusion)

Giderek artan internet kullanımı ve hemen hemen her şirketin internet üzerinden hizmetler sunmaya başlamasıyla birlikte, insanlar internet alan adları üzerinden kimlik avı saldırılarına daha fazla maruz kalmaktadır. Dolandırıcıların yeni yöntemlerle kişilerin özel bilgilerini elde etmesine imkân sağlayan kimlik avı aldatmacası son yıllarda katlanarak artmıştır. Bu dolandırıcılık faaliyetlerinin önüne geçmek, sahte alan adlarını tespit edebilmek için kullanılan makine öğrenmesi algoritmalarının başarımı da farklı araştırmalarda kanıtlanmıştır. Bu nedenle bu çalışmada farklı veri tabanlarından elde edilen kimlik avı alan adları ile Mohammad, Thabtah [46] tarafından düzenlenen özellikler kullanılarak güvenilir yasal alan adlarının sınıflandırılması gerçekleştirilmiş ve sekiz algoritmanın karşılaştırılması yapılmıştır.

Çalışma sonucunda algoritmalar içerisinde rastgele orman, karar ağaçları, çok katmanlı algılayıcı, XGBoost ve lojistik regresyon algoritmaları olmak üzere beş tanesi hem doğruluk oranında %99.8'e ulaşarak yüksek başarım sağlamış, hem de karmaşıklık matrisinde bu orana ait performansları doğrulamıştır. Bu çalışma özellikle kimlik avı dolandırıcılığına maruz kalan kişilerin mağduriyetinin engellenmesi açısından önem arz etmektedir. Ayrıca belirlenen modellerin uygun özelliklerle eğitilmesi durumunda ne kadar yüksek başarım sağlayabileceğinin de bir göstergesi olmuştur. Ek olarak bu çalışma, veri tabanından elde edilen veriler üzerinde derin öğrenme algoritmalarının başarımının değerlendirilmesi ve bu başarımların makine öğrenmesi algoritmalarıyla karşılaştırılması planlanan farklı bir çalışmanın da temelini oluşturmaktadır.

4.3. Sınırlılıklar (Limitations)

Bu çalışmanın çeşitli sınırlılıkları bulunmaktadır. Bunlardan ilki özelliklerin yapılandırılması için harcanan zamandır. Çalışmada kullanılan özelliklerin veri seti üzerinde normalizasyon işlemlerinde kullanılması için yoğun veri ön-işleme aşaması gerçekleştirilmiştir. Bir diğer sınırlılık ise saldırıların çeşitlenmesidir. Gelişen teknolojiyle birlikte kimlik avı saldırılarının çeşitliliği de artmaktadır. Bunları tespit edebilen algoritmaların da, özelliklerin çeşitlenmesiyle birlikte güncellenmesi

gerekmektedir. Son olarak güncel veri seti elde etme zorluğu da bu çalışmanın sınırlılıklarından bir tanesidir. Sınıflandırma işleminde kullanılacak algoritmayı eğitmek için kategorik verilere ihtiyaç duyulmaktadır. Bu verileri elde etmek için de yine yoğun insan çabası harcanmaktadır.

ETİK STANDARTLARIN BEYANI
(DECLARATION OF ETHICAL STANDARDS)

Bu makalenin yazarları çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

YAZARLARIN KATKILARI (AUTHORS' CONTRIBUTIONS)

Taki SAVAŞ: Veri setlerinin elde edilmesi, normalizasyon ve algoritmaların uygulanması.

Serkan SAVAŞ: Literatür araştırması, karışıklık matrisi, sonuçların karşılaştırılması ve yorumlanması.

ÇIKAR ÇATIŞMASI (CONFLICT OF INTEREST)

Bu çalışmada herhangi bir çıkar çatışması yoktur. / There is no conflict of interest in this study.

KAYNAKLAR (REFERENCES)

[1] Savaş, S. and Topaloğlu, N., "Data analysis through social media according to the classified crime", ***Turkish Journal of Electrical Engineering & Computer Sciences***, 27(1): 407-420, (2019).

[2] FBI, "Internet Crime Report", (2020).

[3] Korkmaz, A. and Büyükgöze, S., "Sahte Web Sitelerinin Sınıflandırma Algoritmaları İle Tespit Edilmesi", ***Avrupa Bilim ve Teknoloji Dergisi***, (16): 826-833, (2019).

[4] Sönmez, Ü., "Bilişim Sistemleri Aracılığıyla Dolandırıcılık Suçu", ***Dicle Üniversitesi Adalet Meslek Yüksekokulu Dicle Adalet Dergisi***, 1(2): 47-68, (2017).

[5] Bassett, G., et al., "Data Breach Investigations Report (DBIR 2021)", (2021).

[6] Rosenthal, M. Must-Know Phishing Statistics: Updated 2021. 2021 [cited 2021; Available from: https://www.tessian.com/blog/phishing-statistics-2020/.

[7] McCarthy, J., et al., "A proposal for the Dartmouth summer conference on artificial intelligence", ***Dartmouth Workshop***, (1955).

[8] Jain, A.K. and Gupta, B. B., "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning. in Cyber Security", Singapore: ***Springer Singapore***, (2018).

[9] Kadı, C., "Zararlı Web Sayfalarının Tespiti ve Sınıflandırılması için Yeni Bir Sistem Önerisi", Yüksek Lisans Tezi, ***Fen Bilimleri Enstitüsü, Gazi University***: Ankara, (2018).

[10] Sanglerdsinlapachai, N. and Rungsawang, A., "Using domain top-page similarity feature in machine learning-based web phishing detection", ***2010 Third International Conference on Knowledge Discovery and Data Mining***, IEEE, (2010).

[11] Shirazi, H., Bezawada, B., and Ray, I., ""Kn0w Thy Doma1n Name": Unbiased Phishing Detection Using Domain Name Based Features", ***in Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies***, Association for Computing Machinery: Indianapolis, Indiana, USA, 69–75, (2018).

[12] Jain, A.K. and Gupta, B. B., "A machine learning based approach for phishing detection using hyperlinks information", ***Journal of Ambient Intelligence and Humanized Computing***, 10(5): 2015-2028, (2019).

[13] Zhang, D., et al., "A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites", ***Information & Management***, 51(7): 845-853, (2014).

[14] Alkhozae, M.G. and Batarfi, O. A., "Phishing websites detection based on phishing characteristics in the webpage source code", ***International Journal of Information and Communication Technology Research***, 1(6), (2011).

[15] Hong, J., et al., "Phishing url detection with lexical features and blacklisted domains", ***in Adaptive Autonomous Secure Cyber System***, Springer, 253-267, (2020).

[16] Christou, O., et al. "Phishing url detection through top-level domain analysis: A descriptive approach", ***in 6th ICISSP***, arXiv (2020).

[17] Moghimi, M. and Varjani, A. Y., "New rule-based phishing detection method", ***Expert Systems with Applications***, 53: 231-242, (2016).

[18] Sahingoz, O.K., et al., "Machine learning based phishing detection from URLs", ***Expert Systems with Applications***, 117: 345-357, (2019).

[19] Koşan, M.A., Yıldız, O., and Karacan, H., "Kimlik avı web sitelerinin tespitinde makine öğrenmesi algoritmalarının karşılaştırmalı analizi", ***Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi***, 24(2): 276-282, (2018).

[20] Buyrukoğlu, S. "Improvement of Machine Learning Models' Performances based on Ensemble Learning for the detection of Alzheimer Disease", ***in 2021 6th International Conference on Computer Science and Engineering (UBMK)***, IEEE, (2021).

[21] Dalmaz, H., Erdal, E., and Ünver, H. M., "Machine Learning Approaches in Detecting Network Attacks", ***in 2021 6th International Conference on Computer Science and Engineering (UBMK)***, IEEE, (2021).

[22] Pan, Y. and Ding, X., "Anomaly based web phishing page detection", ***in 2006 22nd Annual Computer Security Applications Conference (ACSAC'06)***, IEEE, (2006).

[23] Uçar, E., İncetaş, M., and Ucar, M., "A Deep learning approach for detection of malicious URLs", ***in 6th International Management Information Systems Conference***, (2019).

[24] Awadh, K. and Akbaş, A., "Intrusion Detection Model Based on TF. IDF and C4. 5 Algorithms", ***Politeknik Dergisi***, 24(4): 1691-1698, (2021).

[25] Calp, M. H., "The role of artificial intelligence within the scope of digital transformation in enterprises, in Advanced MIS and digital transformation for increased creativity and innovation in business", ***IGI Global***, 122-146, (2020).

[26] Güler, O. and Yücedağ, İ., "Mesleki ortaöğretim öğrencilerinin alan seçimi problemine bulanık mantık temelli yaklaşım", ***Hacettepe Üniversitesi Eğitim Fakültesi Dergisi***, 32(1): 111-122, (2017).

[27] Çetin, G. and Karakış, R., "A wiki application for artificial neural network course in engineering education", ***in 2012 15th International Conference on Interactive Collaborative Learning (ICL)***, (2012).

[28] Akbaş, A., "Machine Learning based Heart Failure Risk Analysis in Python", ***in Programming Solutions for Engineering Problems***, A. Akbaş, S. Buyrukoğlu, and A. Gökçe, Editors, Nobel Akademik Yayıncılık: Ankara. 89-110, (2021).

[29] Yılmaz, Y. and Buyrukoğlu, S, "Hybrid Machine Learning Model Coupled with School Closure For Forecasting COVID-19 Cases in the Most Affected Countries", ***Hittite Journal of Science and Engineering***, 8(2): 123-131, (2021).

[30] Kaynar, O., et al., "Makine öğrenmesi yöntemleriyle müşteri kaybı analizi", ***Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi***, 18(1): 1-14, (2017).

[31] Calp, M. H., "İşletmeler için Personel Yemek Talep Miktarının Yapay Sinir Ağları Kullanılarak Tahmin Edilmesi", ***Politeknik Dergisi***, 22(3):675-686, (2019).

[32] Cortes, C. and Vapnik, V., "Support-vector networks", ***Machine Learning***, 20(3):273-297, (1995).

[33] Ho, T. K, "Random decision forests", ***in Proceedings of 3rd International Conference on Document Analysis and Recognition***, (1995).

[34] Ho, T.K., "Recognition of handwritten digits by combining independent learning vector quantizations", ***in Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93)***, IEEE, (1993).

[35] Patil, T.R. and Sherekar, S. S., "Performance analysis of naive bayes and J48 classification algorithm for data classification", ***Journal of Computer Science and Applications***, 6(2): 256-261 (2013).

[36] Savaş, S., "Karotis Arter Intima Media Kalınlığının Derin Öğrenme ile Sınıflandırılması", Doktora Tezi, ***Fen Bilimleri Enstitüsü, Gazi University***: Ankara, (2019).

[37] Fix, E. and Hodges, J. L., "Discriminatory analysis. Nonparametric discrimination: Consistency properties", ***International Statistical Review/Revue Internationale de Statistique***, 57(3): 238-247, (1989).

[38] Tekerek A., "Support Vector Machine Based Spam SMS Detection", ***Politeknik Dergisi***, 22(3): 779-784, (2019).

[39] Kırmızıgül Çalışkan, S. and Soğukpınar, İ., "KxKNN: K-Means ve K En Yakin Komşu Yöntemleri İle Ağlarda Nüfuz Tespiti" ***EMO Yayınları***, 120-24, (2008).

[40] SPSS. "AnswerTree Algorithm Summary", [cited 2021, from: https://s2.smu.edu/~mhd/8331f03/AT.pdf, (1999).

[41] Rosenblatt, F., "The perceptron: a probabilistic model for information storage and organization in the brain", ***Psychological review***, 65(6): 386, (1958).

[42] Bulut, F., "Çok Katmanlı Algılayıcılar ile Doğru Meslek Tercihi", ***Anadolu University Journal of Science and Technology A-Applied Sciences and Engineering***, 17(1): 97-109, (2016).

[43] Chen, T. and Guestrin, C., "Xgboost: A scalable tree boosting system", ***in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining***, (2016).

[44] Buyrukoğlu, S., "New hybrid data mining model for prediction of Salmonella presence in agricultural waters based on ensemble feature selection and machine learning algorithms", ***Journal of Food Safety***, 41(4): e12903, (2021).

[45] Al-Helli, S. and Akbaş, A., "Guided Feature Selection and Dimensionality Reduction Method for IDS Improvement in DDoS Attacks", ***in International Conference on Engineering Technologies (ICENTE'20)***, Konya: Selçuk University, (2020).

[46] Mohammad, R.M., Thabtah, F., and McCluskey, L., "Phishing websites features. School of Computing and Engineering", ***University of Huddersfield***, (2015).

[47] USOM. "Zararlı Bağlantılar", Available from: https://www.usom.gov.tr/adres. (2021).

[48] Alexa. "Site Info", Available from: https://www.alexa.com/siteinfo. (2021).

[49] PhishTank. "Join the fight against phishing", Available from: http://data.phishtank.com/data/online-valid.csv. (2021).

[50] Savaş, S., Topaloğlu, N., Kazcı, Ö., and Koşar, P. N., "Classification of Carotid Artery Intima Media Thickness Ultrasound Images with Deep Learning", ***Journal of Medical Systems***, 43(8): 273, (2019).

[51] Savaş, S., Topaloğlu, N., Kazcı, Ö., and Koşar, P. N., "Performance Comparison of Carotid Artery Intima Media Thickness Classification by Deep Learning Methods", ***in International Congress on Human-Computer Interaction, Optimization, and Robotic Applications***, SETSCI Conference Proceedings: Urgup, Nevşehir, Turkey. 125-131, (2019). doi: https://doi.org/10.36287/setsci.4.5.025

[52] Arslan, R. S., "Kötücül Web Sayfalarının Tespitinde Doc2Vec Modeli ve Makine Öğrenmesi Yaklaşımı" ***Avrupa Bilim ve Teknoloji Dergisi***, (27): 792-801, (2021).

[53] Almseidin, M., et al., "Phishing detection based on machine learning and feature selection methods". ***International Association of Online Engineering***, (2019).

[54] Özker, U., "İçerik tabanlı oltalama saldırısı tespit sistemi", Yüksek Lisans Tezi, ***Lisansüstü Eğitim Enstitüsü, İstanbul Kültür Üniversitesi***, (2021).

[55] İncir, R., "Derin öğrenme yöntemi kullanarak web tabanlı kimlik avı saldırılarının sınıflandırılması", Yüksek Lisans Tezi, ***Fen Bilimleri Enstitüsü, Fırat Üniversitesi***, (2020).

[56] Abu-Nimeh, S., et al., "A comparison of machine learning techniques for phishing detection", ***in Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit***, Association for Computing Machinery: Pittsburgh, Pennsylvania, USA. 60–69, (2007).

[57] Chiew, K.L., et al., "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system", ***Information Sciences***, 484: 153-166, (2019).

[58] Kalaycı, T. E., "Kimlik hırsızı web sitelerinin sınıflandırılması için makine öğrenmesi yöntemlerinin karşılaştırılması", ***Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi***, 24(5): 870-878, (2018).