**Phishing Web Sitelerinin Sınıflandırılmasında Karar Ağaçları Tabanlı Makine Öğrenmesi Uygulaması**

Bilgisayar Mühendisliği, Teknoloji Fakültesi, Gazi Üniversitesi, Ankara, Türkiye

Alper Yıldırım

alper.yildirim3@gazi.edu.tr

Bilgisayar Mühendisliği, Teknoloji Fakültesi, Gazi Üniversitesi, Ankara, Türkiye

Enes Özmekik

enes.ozmekik@gazi.edu.tr

Bilgisayar Mühendisliği, Teknoloji Fakültesi, Gazi Üniversitesi, Ankara, Türkiye

Erdem Çakır

erdem.cakir@gazi.edu.tr

**ÖZET:**

Bu projemizde, phishing web sitelerinin tespiti karar ağacı alanında makine öğrenme modeli geliştirdik. Çalışmamızdaki temel amaç, kullanıcıların karşılaşabileceği phishing web sitelerini önceden tespit edip herhangi bir veri çalınması veya zararlı dosya indirilmesi durumlarından kaçınmasıdır. Elimizde 11.000'den fazla bağlantı içeren ve 30 civarında özelliklere göre ayrılmış web sitelerinin bulunduğu bir veri seti vardı. Elimizdeki veri seti ile güvenilir ve phishing siteleri ayırt edecek bir model geliştirdik ve bu ölçümlerde en çok doğruluk oranını bize Rastgele Orman yöntemi verdi. Modeli geliştirirken, hangi yöntemin en çok doğruluk değerini verdiğini ölçüp ona göre yöntemimizi belirledik. Ayrıca, projemizin kullanıcıya sunumu ve kullanımı amacıyla Python tabanlı bir web sitesi oluşturduk. Kullanıcıdan alınan bağlantı ile sitenin güvenilirliği test edilmektedir.

**Anahtar kelimeler**:Phishing,Makine Öğrenmesi,Karar Ağacı

**Giriş:**

Phishing saldırıları, internet kullanıcılarının kişisel ve finansal bilgilerini ele geçirmek veya kurumları da hedef alarak maddi kayıplara ve bilgi güvenliği açıklarına yol açmak amacıyla sahte web siteleri kullanılarak yapılan siber saldırılardır. Bu sahte sitelere ise “phishing web siteleri” denir. Saldırılar giderek daha gelişmiş ve karmaşık hale gelmekte; bu da geleneksel yöntemlerle filtrelemeyi zorlaştırmaktadır. Bu nedenle, makine öğrenmesi tabanlı yaklaşımlar geliştirerek daha otomatik ve doğru tespitler yapmak, kullanıcı güvenliğini sağlamak ve siber saldırılarla mücadele etmek açısından ciddi bir öneme sahiptir. Bu doğrultuda yaptığımız projede, UCI veri deposundaki, 11.000’den fazla bağlantı ve her bağlantı için 30'dan fazla özellikle tanımlanmış (URL yapısı, güvenlik sertifikası durumu, domain bilgileri vb.) “Phishing Websites” veri seti kullanılmıştır. Karar ağaçları, destek vektör makineleri (SVM) ve rastgele orman gibi birçok algoritma test edilmiş; bunların sonucunda maksimum doğruluk oranını sağlayan rastgele orman yöntemi tercih edilmiştir. Geliştirdiğimiz proje, bahsettiğimiz saldırılara maruz kalmamak amacıyla, bağlantılara gitmeden önce doğruluk testi yapma imkânı sunarak güvenliği artırmaktadır. Seçtiğimiz modeli Python diliyle uyguladık ve kullanıcıdan alınan bağlantıyı test eden bir web sitesi geliştirdik.

**2. Materyal ve Yöntem**

**2.1 Veri Seti:**

Bu çalışmada, “Phishing websites” veri seti kullanılmıştır. Veri seti; gerçek phishing ve meşru olmak üzere iki sınıftan oluşan URL kayıtlarını içerir. Veri, ARFF formatında olup her bir satır bir URL’ye karşılık gelir. Her URL için; etiket (-1 = güvenli, 1 = oltalama) bilgisi bulunur. URL’lerin yapısal ve içeriksel özelliklerinden yola çıkarak, her bir URL’i açıklayan 30 adet özniteliğin sayısal karşılıkları bulunmaktadır.

Başlıca 10 öznitelik şunlardır:

1. **SSLfinal\_State:** SSL sertifika geçerlilik ve güvenlik düzeyi
2. **URL\_of\_Anchor:** Anchor linklerinin farklı/dahili domain dağılımı
3. **having\_Sub\_Domain:** Alt alan adı (subdomain) sayısı
4. **web\_traffic:** Site trafik hacmi ve ziyaretçi sayısı
5. **Links\_in\_tags:** Etiketlerdeki dış bağlantı sayısı ve oranı
6. **Prefix\_Suffix:** Domain adında tire (prefix/suffix) kullanımı
7. **SFH:** Form action yönlendirme değerlendirmesi
8. **Request\_URL:** Sayfa içi kaynak çağrı URL domain analizi
9. **Links\_pointing\_to\_page:** Sayfaya yönlenen backlink sayısı ve oranı
10. **Domain\_registeration\_length:** Domain kayıt süresi (gün cinsinden)

Bu 30 özniteliğe bağlı olarak, URL’lerin güvenli veya güvensiz olduğunu belirten Result etiket sütunu yer almaktadır. Veri seti toplam 11 055 kayıt içermekle birlikte, bunların 4 898’i güvenli, 6 157’si güvensiz URL’lerden oluşmaktadır.

**2.2 Kullanılan Makine Öğrenmesi Algoritmaları:**

**Rastgele Orman (Random Forest):** Random forest algoritması ,denetimli sınıflandırma algoritmalarından biridir. Hem regresyon hem de sınıflandırma problemlerinde kullanılmaktadır. Algoritma ,birden fazla karar ağacı üreterek sınıflandırma işlemi esnasında sınıflandırma değerini yükseltmeyi hedefler. Random forest algoritması birbirinden bağımsız olarak çalışan birçok karar ağacının bir araya gelerek aralarından en yüksek puan alan değerin seçilmesi işlemidir. Ağaç sayısı arttıkça kesin bir sonuç elde etme oranımız artmaktadır. Karar ağaçları algoritması ile arasındaki temel fark, random forest algoritmasında kök düğümü bulma ve düğümleri bölme işleminin rastgele olmasıdır.

**Karar ağaçaları:**Ağaç tabanlı öğrenme algoritmalarından olan Karar Ağaçları, çok sayıda kayıt içeren bir veri kümesini, bir dizi karar kuralları uygulayarak daha küçük kümelere bölmek için kullanılan bir yapıdır. Kümenin tüm elemanları aynı sınıf etiketine sahip olana kadar kümeleme işlemi derinlemesine devam eder. Kararlar yapraklarda ve veriler düğümlerde bölünür. Sınıflandırma ağacında karar değişkeni kategoriktir ve Regresyon ağacında karar değişkeni süreklidir. Karar Ağaçları, yorumlamada, kategorik ve nicel değerleri ele almada kolaylık gösterir, özniteliklerdeki eksik değerleri en olası değerle doldurabilme özelliğine sahiptir.

**Destek Vektör Makineleri (SVM):** Destek vektör makinesi, sinyal işleme, doğal dil işleme (NLP) ve konuşma ve görüntü tanıma gibi uygulamalarda sınıflandırma ve regresyon problemleri için sıklıkla kullanılan bir gözetimli makine öğrenme algoritmasıdır. SVM algoritmasının amacı, mümkün olan en iyi derecede, bir sınıfın veri noktalarını başka bir sınıfın veri noktalarından ayıran bir hiper düzlem bulmaktır. Bu hiper düzlem, 2B uzay için bir çizgi veya *n* boyutlu bir uzay için bir düzlem olabilir; burada *n,* veri kümesindeki her bir gözlem için özellik sayısıdır. Verilerdeki sınıfları ayıran birden fazla hiper düzlem olabilir. SVM algoritması tarafından türetilen en uygun hiper düzlem, iki sınıf arasındaki marjı en üst düzeye çıkarandır.

**K-En Yakın Komşu (K-NN):** K-en yakın komşular (KNN) algoritması, yakınlığı kullanarak bireysel bir veri noktasının gruplandırılması hakkında sınıflandırmalar veya tahminler yapan, parametrik olmayan, denetlenen bir öğrenme sınıflandırıcısıdır. Günümüzde makine öğreniminde kullanılan popüler ve en basit sınıflandırma ve regresyon sınıflandırıcılarından biridir.

**2.3 Literatür Taraması**

Geçtiğimiz yıllarla birlikte internetin geniş kullanım alanlarına ulaşması, phishing saldırılarının artmasına neden olmuştur. Bu saldırılar, kullanıcıların bilgilerini çalmaya yönelik olarak hazırlanan sahte web siteleri aracılığıyla gerçekleştirilmektedir. Bu nedenle konu, bilimsel araştırmalarda da oldukça fazla analiz edilip değerlendirilmektedir. Özellikle bu tür saldırıların tespitinde genellikle makine öğrenmesi yöntemleri tercih edilmekte ve bu konuda birçok araştırma yapılmaktadır. Makine öğrenmesi tabanlı modeller, verilerdeki örüntüleri tanıyarak kullanıcıları tehdit edebilecek bağlantıları daha hızlı ve etkili şekilde ayırt etme potansiyeline sahiptir.

Birçok çalışmada ise UCI veri deposunda bulunan "Phishing Websites" veri seti tercih edilmiştir. Bu veri seti, farklı özelliklerde URL’lerden oluşmakta olup, veriler "güvenli" ya da "phishing" olarak etiketlenmiştir. Bu özellikler arasında URL uzunluğu, IP adresi içerip içermemesi, alan adının süresi gibi belirleyici faktörler yer almaktadır. Söz konusu veriler, çeşitli algoritmalarla işlenerek modellerin başarı düzeyleri karşılaştırılmıştır.

Alternatif olarak, Mao ve arkadaşları , kimlik avı saldırı sayfalarını belirlemek için kullanılabilecek sayfa tasarımı karşılaştırılabilirliğini ölçen öğrenme-tabanlı bir teknik geliştirmiştir. İkili ML algoritmaları destek vektör makinesi ve etkili sayfa düzeni yönleri için ikili karar ağaçları kullanılarak bir kimlik avı sınıflandırıcısı oluşturulmuş; metodolojilerini doğrulamak için Phishtank.com ve Alexa.com’dan gerçek web sitesi sayfa testleri uygulanmıştır.

Tyagi ve diğerleri , Irvine’daki Kaliforniya Üniversitesi’nin makine öğrenimi havuzundan 2.456 benzersiz URL ve 6.157 kimlik avı ile 4.898 normal URL olmak üzere toplam 11.000’den fazla URL içeren bir veri kümesi kullanmıştır. URL’lerden 30 özellik çıkarılmış ve bunlar saldırıları tahmin etmek için kullanılmıştır. Makine öğrenimi teknikleri olarak Karar Ağacı (DT), Rastgele Orman (RF), gradyan artırma, genelleştirilmiş doğrusal modeller ve PCA uygulanmıştır.

Chen ve Chen, modellerin tespit kapsamını artırmak için SMOTE yaklaşımını kullanmıştır. Sampling, RF ve XGBoost gibi makine öğrenmesi modelleri eğitilmiş; önerdikleri XGBoost yaklaşımı en yüksek doğruluğu sağlamıştır. Çalışmalarında 24.000’den fazla kimlik avı ve 4.000’den fazla yasal web sitesi içeren Phishtank veritabanı kullanılmıştır.

Patil ve arkadaşları kimlik avı web sitelerini tespit etmek için üç strateji önermiştir. İlki, çeşitli URL özniteliklerinin değerlendirilmesini gerektirmiş; ikincisi, web sitesinin nerede barındırıldığını ve kim tarafından yönetildiğini belirleyerek web sitesinin geçerliliğini tespit etmiş; üçüncü yöntem ise görsel, görünüme dayalı analiz yoluyla web sitesinin gerçekliğini belirlemiştir. URL’lerin ve web sitelerinin sayısız yönünü değerlendirmek için makine öğrenimi metodolojileri ve algoritmaları kullanılmıştır.

Joshi ve diğerleri RF algoritmasına dayalı bir ikili sınıflandırıcı ve ReliefF algoritmasına dayalı bir özellik seçim algoritması kullanmıştır. Özellik seçim algoritması için kaynak olarak Mendeley alanından elde edilen verileri kullanmışlardır. Daha sonra seçilen özellikleri, kimlik avı saldırılarını tahmin etmek üzere bir RF sınıflandırıcısında test etmişlerdir.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Veri Seti** | **Algoritma** | **Doğruluk** |
| Tyagi ve diğerleri | URL’ler | DT, RF, GB | %98,40 |
| Chen ve Chen | Web Siteleri | ELM, SVM, LR, C4.5, LC-ELM, KNN, XGB | %99,20 |
| Mao ve diğerleri | Web Siteleri | SVM, RF, DT, AB | %93 |
| Patil ve diğerleri | URL’ler | LR, DT, RF | %96,58 |
| Joshi ve diğerleri | Web Siteleri | RF | %97,63 |

**3.Bulgular ve Tartışma**

URL içeren veri seti, toplam 11055 adet örneğin %30’unu oluşturan 3317 örnek ile Destek vektör makinesi (SVM), Karar Ağaçları (KA), Rastgele orman (RF) ve K-en yakın komşu (KNN) modelleri ile ayrı ayrı sınıflandırılmıştır. Her bir model için confusion matrix oluşturulmuş, doğruluk, duyarlılık, belirleyicilik, kesinlik değerleri belirlenmiş ve aşağıdaki tablolarda gösterilmiştir.

**metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir. metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir. metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Doğruluk** | **Duyarlılık** | **Belirleyicilik** | **Kesinlik** |
| KA(Gini) | %95,62 | %95,39 | %95,79 | %94,39 |
| KA(Entropy) | %96,14 | %95,96 | %96,98 | %95,02 |
| RF | %96,68 | %97,54 | %98,20 | %94,67 |
| SVM | %94,24 | %94,21 | %95,71 | %92,29 |
| KNN | %93,39 | %92,78 | %94,60 | %91,80 |

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

metin, ekran görüntüsü, yazılım, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Bu çalışmada beş farklı sınıflandırma algoritmasının performansı karşılaştırılmış olup, elde edilen sonuçlar arasında Random Forest modeli en yüksek doğruluk oranını sergilemiştir. Spesifik olarak, Random Forest %96,68’lik bir başarı düzeyiyle en üst sırada yer alırken, Karar Ağacı (Entropy) %96,14 ve Karar Ağacı (Gini) %95,62 oranlarıyla kendisini takip etmiştir. Destek Vektör Makineleri (SVM) %94,24 ve K-En Yakın Komşu (KNN) %93,39 doğruluk oranlarıyla diğer modellerin gerisinde kalmıştır. Bu bulgular, Random Forest yaklaşımının veri içindeki karmaşık ilişkileri daha etkin yakalayabildiğini ve tek ağaç temelli yöntemlere kıyasla genel sınıflandırma başarısını anlamlı biçimde artırdığını göstermektedir.

|  |  |
| --- | --- |
| **Model** | **Doğruluk** |
| Tyagi ve diğerleri | %98,40 |
| Chen ve Chen | %99,20 |
| Mao ve diğerleri | %93 |
| Patil ve diğerleri | %96,58 |
| Joshi ve diğerleri | %97,63 |
| Bizim modelimiz | %96,68 |

**4.Sonuç ve Öneriler**

Bu çalışmada, kullanıcıların işlerinde ve günlük hayatlarında karşılaşabilecekleri zararlı web sitelerine yönlendirilerek kişisel bilgilerinin çalınması gibi tehditlere karşı korunabilmeleri amacıyla, bağlantı (URL) tabanlı phishing tespiti yapan bir sistem geliştirilmiştir. Bu doğrultuda, 30 farklı URL özelliğine dayalı olarak sınıflandırma yapan çeşitli makine öğrenmesi algoritmaları karşılaştırılmış ve en yüksek başarı oranını sağlayan model belirlenmiştir. Kullanılan veri seti, “güvenli” ve “güvensiz” olmak üzere sınıflandırılmış toplam 11.055 URL içermektedir.

Çalışmada hazır veri seti doğrudan kullanılmış ve bu veriler üzerinde çeşitli makine öğrenmesi modelleri eğitilmiştir. Yapılan testler sonucunda en başarılı performans, Rastgele Orman (Random Forest) algoritmasıyla elde edilmiştir. Öte yandan, Destek Vektör Makinesi gibi bazı modeller daha düşük doğruluk oranlarıyla daha zayıf sonuçlar vermiştir. Ayrıca sistemin kullanılabilirliğini artırmak amacıyla bir web sitesi geliştirilmiş, böylece kullanıcıların doğrudan URL girerek analiz sonuçlarına kolayca ulaşmaları sağlanmıştır.

Bu çalışmanın en önemli katkılarından biri, phishing tespitine yönelik hem etkili hem de kullanıcı dostu bir çözüm sunmasıdır. Geliştirilen sistem, bireysel kullanıcıların internet güvenliğini artırmak ve zararlı web sitelerine karşı farkındalık kazanmalarını sağlamak açısından faydalıdır. Gelecek çalışmalarda, veri setine daha güncel URL verilerinin eklenmesi ve farklı derin öğrenme yöntemlerinin test edilmesiyle sistemin doğruluğu daha da artırılabilir.

**Kaynakça:**

* <https://ece-akdagli.medium.com/makine-%C3%B6%C4%9Frenmesinde-random-forest-algoritmas%C4%B1-a79b044bbb31>
* <https://www.mathworks.com/discovery/support-vector-machine.html>
* <https://www.ibm.com/think/topics/knn>
* Patil, V.; Thakkar, P.; Shah, C.; Bhat, T.; Godse, S.P. Detection and Prevention of Phishing Websites Using Machine Learning
* Approach. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation
* (ICCUBEA), Pune, India, 19–18 August 2018; pp. 1–5
* Joshi, A.; Pattanshetti, P.T.R. Phishing Attack Detection Using Feature Selection Techniques; Social Science Research Network:
* Rochester, NY, USA, 2019
* Mao, J.; Bian, J.; Tian, W.; Zhu, S.; Wei, T.; Li, A.; Liang, Z. Phishing Page Detection via Learning Classifiers from Page Layout
* Feature. EURASIP J. Wirel. Commun. Netw. 2019, 2019, 43. Available online: https://jwcn-eurasipjournals.springeropen.com/

articles/10.1186/s13638-019-1361-0 (accessed on 27 September 2021).

* A Novel Machine Learning Approach to Detect Phishing Websites. Available online: https://ieeexplore.ieee.org/abstract/

document/8474040/ (accessed on 27 September 2021).

* Chen, Y.H.; Chen, J.L. AI@ntiPhish—Machine Learning Mechanisms for Cyber-Phishing Attack. IEICE Trans. Inf. Syst. 2019, 102,
* 878–887. Available online: https://www.jstage.jst.go.jp/article/transinf/E102.D/5/E102.D\_2018NTI0001/\_article/-char/ja/

(accessed on 27 September 2021).