Web Page Phishing Detection

1st Beyza Akdeniz
Computer Science Department. TOBB ETU
Ankara, Türkiye
beyzaakdeniz@etu.edu.tr

***Abstract*—Bu proje, internet güvenliği alanında önemli bir tehdit olan phishing saldırılarını tespit etmeyi amaçlamaktadır. Phishing, kullanıcıların kimlik bilgilerini çalmak için sahte web sayfaları oluşturan saldırganlar tarafından gerçekleştirilen bir siber suç türüdür. Bu çalışmada, URL yapısı analiz edilerek phishing tespiti yapılacaktır. Bu amaçla, 11.430 URL'den oluşan dengeli bir veri seti kullanılmış ve her URL için 87 özellik çıkarılmıştır. Proje kapsamında veri ön işleme, öznitelik mühendisliği ve makine öğrenmesi modelleri kullanılarak phishing ve meşru URL'ler sınıflandırılmıştır.**

***Index Terms*—phishing, cybersecurity, data mining, machine learning, classification**

I. Introduction

A. Motivasyon

Modern hayatta iletişim, bankacılık, eğitim, e-ticaret ve sosyalleşme gibi pek çok alanda internet büyük bir önem kazanmıştır. Bu durum, kimlik avı (phishing) gibi bazı güvenlik tehditlerini de beraberinde getirir. Kimlik avı, bir saldırganın kullanıcının kişisel bilgilerini çalmaya çalıştığı ve bireylerin ve kuruluşların mali kaybına yol açtığı çevrimiçi kimlik hırsızlığıdır. Kimlik avı tehditi yeni olmasa ve ona karşı önlemler alınmış olsa da kimlik avcıları saldırılarını gerçekleştirmenin yeni, yaratıcı yollarını bulmaya devam etmektedir. Saldırı yöntemlerinin gelişmesiyle beraber tespit etme yöntemleri de araştırılmakta ve geliştirilmektedir. Bu çalışmanın amacı da phishing'in tespiti yönünde bir geliştirme yapmaktır.

B. Genel Metodoloji

Bu projede verisetinden yararlanılarak URL yapısı analizi yapılacak ve özniteliklerin birbiriyle ilişkileri incelenecektir. Bunların phishing olasılığı üzerindeki etkileri analiz edilecektir. Örneğin, URL uzunluğu, domain adı uzantısı, kullanılan karakterler gibi. Bu öznitelikler arasındaki korelasyon grafikleri çıkarılmıştır. Veri işleme ve korelasyon analizi aşamalarından sonra ise model eğitilerek sınıflandırma yapılacaktır. Temelde Random Forest üzerine kurulu çeşitli makine öğrenmesi modelleri eğitilmiş ve phishing tespiti gerçekleştirilmiştir.

C. Amaç/Hedef

Bu projenin amacı, öncelikle phishing verisetinden anlamlı öngörüler çıkarmak ve sonra phishing ve meşru URL'leri yüksek doğrulukla ayırt edebilen bir makine öğrenmesi modeli geliştirmektir. Başarım metrikleri olarak accuracy, precision, recall ve F1 skoru kullanılacaktır.

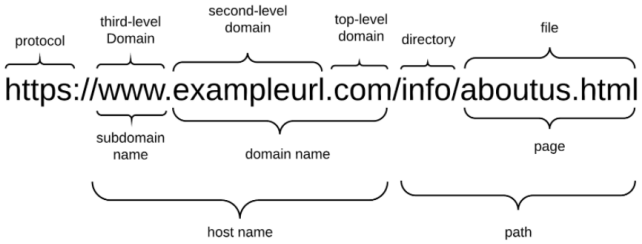

Fig. 1. URL Structure

II. Related Work

Dünya çapında internet kullanımının artmasıyla birlikte, phishing dolandırıcılıkları bireyler ve kuruluşlar için büyük bir tehdit oluşturarak, kişisel ve gizli bilgilerin yanı sıra finansal verilerin de önemli kayıplara yol açmasına neden olmaktadır. Bu sorunu ele almak amacıyla, son yıllarda sahte web sayfalarını tespit etmek ve bunları phishing olmayan sitelerden ayırt etmek için etkili yöntemler geliştirmek üzere birçok çalışma yapılmıştır.

Bazı çalışmalar yeni modeller geliştirirken, diğerleri birden fazla algoritmanın performansını karşılaştırmış ve bazıları da tespit performansını artırmak için iki veya daha fazla algoritmayı birleştirmiştir. Sonuçlar, çoğu algoritmanın yüzde 90'dan fazla doğruluk sağladığını, ancak yalnızca birkaç spesifik modelin phishing web sitelerini yüzde 100 doğrulukla tespit edebildiğini göstermektedir. Onlar da değişen saldırı teknikleriyle beraber geçerliliğini yitirebilmektedir.

Literatürde phishing tespiti için çeşitli yöntemler önerilmiştir. Pek çok çalışmada phishing web sitelerinin sayısının meşru web sitelerinin sayısından çok daha az olduğu dengesiz veri kümeleri kullanılmıştır. Bu dengesizliği gidermek için oversampling gibi teknikler kullanılır. Bu çalışmada çok daha dengeli bir veri seti üstünde çalışılacaktır.

Bazı çalışmalar, düşük false positive oranları korurken aynı zamanda yüksek accuracy elde etmenin zorluğundan bahsetmektedir. Yüksek false positive oranları, birçok meşru web sitesinin phishing olarak işaretlenmesine yol açabilir, bu da kullanıcılar ve web sitesi sahipleri için sorun yaratır. Burada false positive oranlarını düşük tutmak için farklı modeller denenecektir.

Bazı modeller eğitim verilerine overfit edebilir ve bu da daha önce görülmeyen verilerde performansın düşmesine ne-

den olabilir. Bunu azaltmak için cross validation ve regularisation gibi teknikler kullanılır.

Bir veri kümesi üzerinde eğitilen modellerin diğer veri kümelerine ve gerçek dünya senaryolarına iyi bir şekilde genellenmesini sağlamak zorlu bir iştir. Bu yüzden pek çok farklı kaynaktan beslenen bir veri seti üstünde çalışılacaktır.

Ayrıca, overfitting gibi sorunların önüne geçmek için cross-validation ve regularization gibi teknikler kullanılmıştır.

III. Dataset, Data Features, Attributes

A. Veri Kaynağı

Bu projede kullanılan veri seti, Mendeley Data veri deposundan temin edilmiştir (https://data.mendeley.com/datasets/c2gw7fy2j4/3).

B. Veri Kümesi

Veri seti, 11.430 URL'den oluşmakta olup, her URL için 87 özellik içermektedir. Veri seti dengeli olup, phishing ve meşru URL'ler eşit sayıda bulunmaktadır.

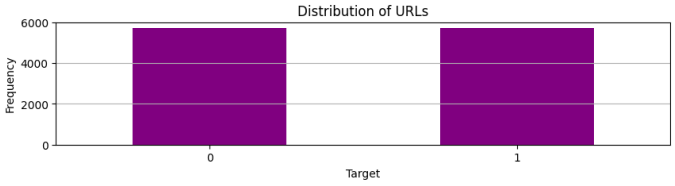

Fig. 2. Distribution of URL

C. Ön İşleme Aşamaları

Veri seti üzerinde veri temizleme, eksik değerlerin doldurulması ve aykırı değerlerin belirlenmesi gibi ön işleme adımları uygulanmasına gerek kalmamıtır. Çünkü oldukça dengeli bir veri setü üstünde çalışılmıştır. Bu aşamada, eksik veriler olup olmadığına bakılmış ve olmadığı tespit edilmiştir.

D. Öznitelik Açıklamaları

Öznitelikler, URL'lerin yapısal ve sentaktik özellikleri, içerik özellikleri ve dış kaynaklardan elde edilen bilgilerden oluşmaktadır. Öznitelikler arasında korelasyon analizi yapılmış ve birbiriyle veya hedef değerle pozitif veya negatif korale öznitelikler belirlenmiştir.

E. Öznitelik Seçimi ve Düzenlemesi

Öznitelik seçimi, Random Forest modeliyle elde edilen importance değerlerine göre yapılmıştır. Importance değeri 0.01'in üstünde olan 27 öznitelikle yeni bir dataframe oluşturulmuş ve tüm öznitelikleri içeren dataframe ile sonuçlar kıyaslanmıştır.

F. Sınıf Veri Dağılımları

Veri seti dengeli olduğundan, sınıflardaki veri sayıları eşittir. Veri seti, meşru ve phishing URL'leri eşit şekilde içermektedir.

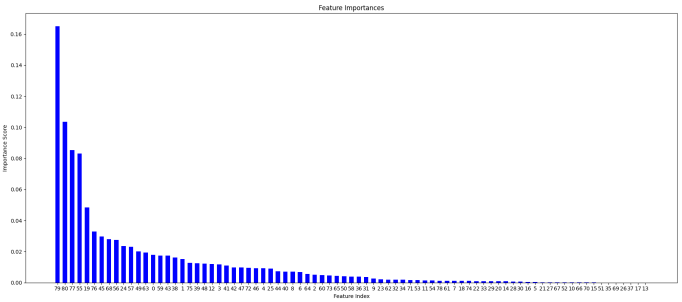

Fig. 3. Importance Values of Features

G. Veri Görselleştirmesi

Öznitelikler önce kategorik ve numerik öznitelikler olarak ikiye ayrılmış ve grafikler bu iki ayrı grup için ayrı ayrı çıkarılmıştır. Bir özniteliğin içerdiği unique değerlerin oranı 0.002'den küçükse onun kategorik veri olduğu varsayılmıştır. Özniteliklerin birbirleriyle korelasyonları görselleştirilmiştir. Korelasyon analizi sonucunda, phishing tespitinde belirleyici olan öznitelikler belirlenmiş ve bu öznitelikler görselleştirilmiştir.

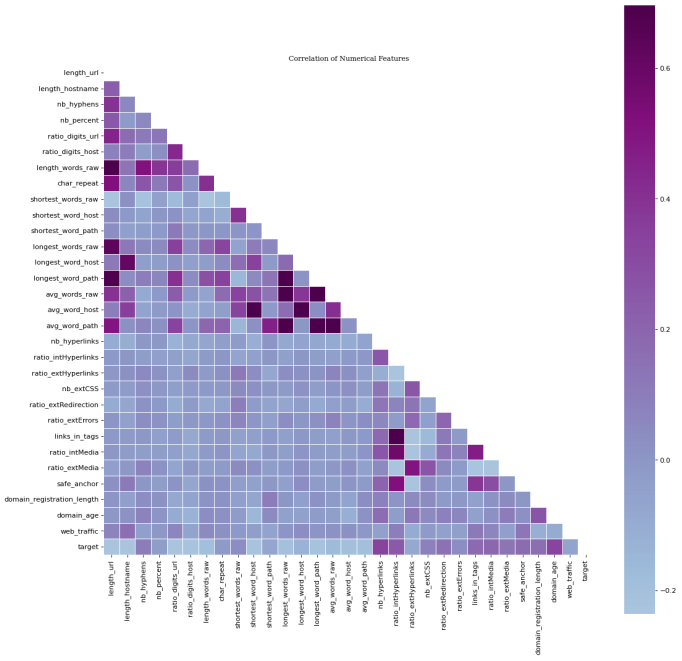

Fig. 4. Correlation of Categorical Features

IV. Methodology

Bu projede Random Forest (RF) makine öğrenmesi modeli kullanılmıştır. Bu model, phishing ve meşru URL'leri ayırt etmek için eğitilmiştir. Aynı zamanda feature extraction'da kullanılmıştır.

Random Forest, bir makine öğrenmesi algoritmasıdır ve temel olarak birden fazla karar ağacının birleşiminden oluşur. İlk olarak, veri seti birden fazla alt veri setine bölünür. Bu bölme işlemi, her bir karar ağacının eğitiminde kullanılacak

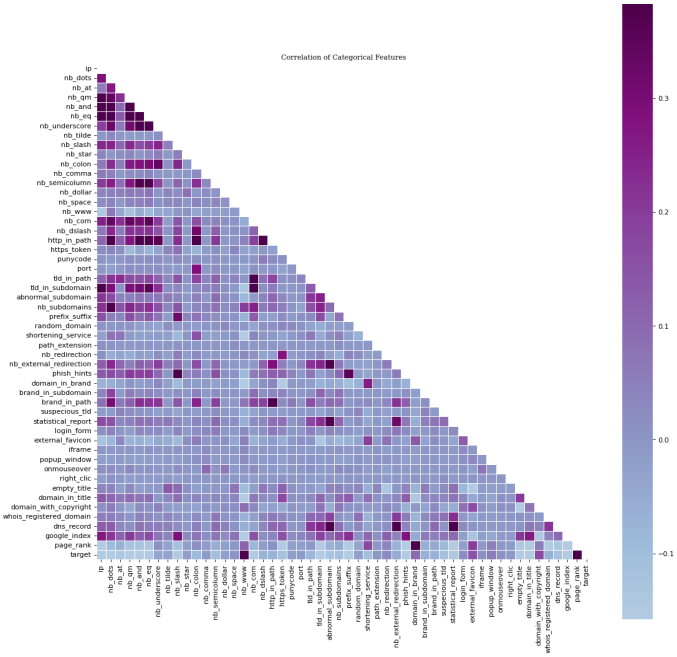

Fig. 5. Correlation of Numerical Features

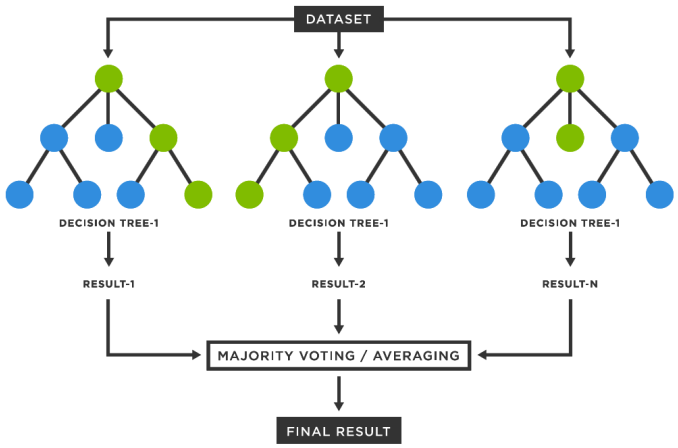

Fig. 6. Random Forest

veri alt kümesini oluşturur. Bu işlem genellikle ”bootstrap aggregating” veya ”bagging” olarak bilinir, yani veriler rastgele seçilir ve her ağaç için tekrar örneklenir. Her alt veri seti kullanılarak bir karar ağacı oluşturulur. Karar ağaçları, verileri belirli özelliklere göre bölerek sınıflandırma yapar veya tahminler üretir. Bu ağaçlar genellikle derin ve karmaşık olabilir, çünkü ağaçların her biri sadece alt veri setinde eğitim aldığından genel veri setinin tamamı hakkında bilgiye sahip değildir. Her karar ağacının oluşturulmasında, her düğümde bir özelliği seçmek için rastgele bir alt küme seçilir. Bu, ağaçların çeşitlenmesini sağlar ve modelin genelleme yeteneğini artırır. Eğitim tamamlandıktan sonra, tüm karar ağaçlarının sonuçları birleştirilir. Sınıflandırma problemlerinde, ağaçların çoğunluk oyuna başvurularak en yaygın sınıf seçilir. Regresyon problemlerinde ise, ağaçların tahminlerinin ortalaması alınır. Random Forest modeli, tüm karar ağaçlarının tahminlerinin birleşimini kullanarak final tahminini yapar. Bu, modelin genel performansını artırır ve aşırı öğrenme (overfitting) riskini azaltır.

Model seçiminde, accuracy, precision, recall gibi metrikler göz önünde bulundurulmuştur. Random Forests daha yüksek performans gösterdiği için tercih edilmiştir. Aynı zamanda gürültüden daha az etkilenmesi ve overfit direncinin iyi olması nedeniyle kimlik avı sitelerini tespit etmede etkilidir.

V. Results and Discussion

A. Performans Metrikleri

Modellerin performansı, doğruluk, precision, recall ve f1 score gibi metriklerle değerlendirilmiştir.

B. Modeller Arası Karşılaştırma

Çeşitli modeller arasında yapılan karşılaştırmalar sonucunda, Grid Search yöntemiyle hyperparameter tuning yapılan Random Forests modelinin en yüksek performansı gösterdiği belirlenmiştir. Literatür araştırmalarında da bu modellerin doğruluk ve precision değerleri diğer modellere göre daha yüksek bulunmuştur.

```
Base Accuracy: 0.9706911636045494
Base Precision: 0.9707824331609772
Base Recall: 0.9706278186538136
Base F1 Score: 0.9706826306042465
```

Fig. 7. Evaluation Metrics of RF

```
Final Accuracy: 0.9619422572178478
Final Precision: 0.9619533594659153
Final Recall: 0.961920470230499
Final F1 Score: 0.9619352572856368
```

Fig. 8. Evaluation Metrics of RF with selected features

```
Tuned Accuracy: 0.9706911636045494
Tuned Precision: 0.9707375452524027
Tuned Recall: 0.9706492540112827
Tuned F1 Score: 0.9706842915906397
```

Fig. 9. Evaluation Metrics of Tuned RF

C. En İyi Modelin Seçimi

Hyperparameter tuning yapılan Random Forests modeli, yüksek doğruluk ve düşük hatalı pozitif oranı ile en iyi performansı göstermiştir. Ancak tuning yapılmayan Random Forest modeli de hayli yakın birer accuracy, precision, recall ve f1 score değerine ulaşmıştır. Bu modelin phishing tespitinde daha başarılı olmasının nedeni, veri setindeki özniteliklerin birbirleriyle olan korelasyonlarını daha iyi değerlendirmesi ve denenen hiperparametreler arasından validation sonucunda en iyi çıkan değerleri kullanmasıdır. Öznitelik seçimi yapılan Random Forest modeli nispeten daha düşük değerlere sahiptir.

VI. Conclusions

A. Çalışma Özeti

Bu çalışmada, phishing tespiti için URL yapısı ve öznitelikleri kullanılarak çeşitli makine öğrenmesi modelleri geliştirilmiştir. Veri ön işleme, keşifsel veri analizi(EDA) ve model eğitimi gibi adımlar detaylı olarak gerçekleştirilmiştir.

B. Öğrenilenler ve Katkılar

Çalışma sonucunda, impotance değerlerinin sonucuna göre phishing tespiti için etkili öznitelikler belirlenmiş ve bu özniteliklerle yüksek doğrulukta bir model geliştirilmiştir. Ancak değerlendirme ölçütleri yine de tüm özniteliklerin kullanıldığı modelden daha aşağıda kalmıştır. Bu URL'lerde phishing tespitinde detaylı ve çok sayıda öznitelik kullanılmasının önemini gösterir. Aynı zamanda, öznitelik seçimi için daha etkili yöntemler kullanılmasına ihtiyaç duyulduğunu da göstermektedir. Veri setinin dengeli olması, model performansını olumlu yönde etkilemiştir. En etkili öznitelikler ise burada listelenmiştir.

C. Yapılamayanlar ve Sebepleri

Bazı özniteliklerin karmaşıklığı ve modelin eğitim süresi gibi sebeplerden dolayı, Recursive Feature Elimination (RFE) gibi daha ileri düzeydeki derin öğrenme teknikleri bu çalışmada kullanılmamıştır.

D. Gelecek Çalışmalar

Gelecek çalışmalarda, derin öğrenme teknikleri kullanılarak daha karmaşık modellerin geliştirilmesi pyapılabilir. Ayrıca, modelin gerçek zamanlı phishing tespitinde kullanılabilmesi için optimizasyon çalışmaları yapılabilir. Daha iyi sistem kaynaklarına sahip olunması halinde daha komplike ve maliyetli öznitelik seçimi algoritmaları kullanılarak daha etkili bir öznitelik mühendisliği yapılabilir.

```
Selected Features:
length_url: 0.0178976036892644
length_hostname: 0.01515640146131702
nb_dots: 0.011633051495568127
nb_slash: 0.01201682333117625
nb_www: 0.0483306365278267
ratio_digits_url: 0.023449031990853308
length_words_raw: 0.016091394025377743
char_repeat: 0.012439116595863973
shortest_word_host: 0.010898776399953089
longest_words_raw: 0.01732362789553734
longest_word_path: 0.02975351445610565
avg_word_path: 0.012259012514339007
phish_hints: 0.02008444718863991
nb_hyperlinks: 0.0832040927500219
ratio_intHyperlinks: 0.0274170348786581
ratio_extHyperlinks: 0.023060488657283097
ratio_extRedirection: 0.017516690292189933
links_in_tags: 0.01933826393256184
safe_anchor: 0.028107160870949586
domain_registration_length: 0.012747695743907467
domain_age: 0.03303341832927692
web_traffic: 0.08544989155074419
google_index: 0.16513232672370404
page_rank: 0.10355853776337935
```

Fig. 10. Selected Features

References

[1] Jibat D, Jamjoom S, Abu Al-Haija Q, et al. A systematic review: Detecting phishing websites using data mining models. Intelligent and Converged Networks, 2023, 4(4): 326-341. https://doi.org/10.23919/ICN.2023.0027

[2] Tang, L.; Mahmoud, Q.H. A Survey of Machine Learning-Based Solutions for Phishing Website Detection. Mach. Learn. Knowl. Extr. 2021, 3, 672-694. https://doi.org/10.3390/make3030034

[3] Harinahalli Lokesh, G., BoreGowda, G. (2020). Phishing website detection based on effective machine learning approach. Journal of Cyber Security Technology, 5(1), 1–14. https://doi.org/10.1080/23742917.2020.1813396

[4] Opara, C., Chen, Y., Wei, B. (2024). Look before you leap: Detecting phishing web pages by exploiting raw URL and HTML characteristics. Expert Systems with Applications, 236, Article 121183. https://doi.org/10.1016/j.eswa.2023.121183

[5] Aljofey, A., Jiang, Q., Rasool, A. et al. An effective detection approach for phishing websites using URL and HTML features. Sci Rep 12, 8842 (2022). https://doi.org/10.1038/s41598-022-10841-5