T.C.
SAKARYA ÜNİVERSİTESİ
BİLGİSAYAR VE BİLİŞİM BİLİMLERİ FAKÜLTESİ

BSM 498 BİTİRME ÇALIŞMASI

MAKİNE ÖĞRENMESİYLE OLTALAMA SİTESİ TESPİTİ

b211210375 – Ferdi SÖNMEZ
b211210377 - Berat ÖZDİN
b211210372 – Musa KARAŞ

Fakülte Anabilim Dalı : BİLGİSAYAR MÜHENDİSLİĞİ
Tez Danışmanı : Dr. Öğr. Üyesi Hüseyin ESKİ

2021-2022 Bahar Dönemi

T.C.
SAKARYA ÜNİVERSİTESİ
BİLGİSAYAR VE BİLİŞİM BİLİMLERİ FAKÜLTESİ

MAKİNE ÖĞRENMESİYLE OLTALAMA SİTESİ TESPİTİ

BSM 498 - BİTİRME ÇALIŞMASI

b211210375 – Ferdi SÖNMEZ
b211210377 - Berat ÖZDİN
b211210372 – Musa KARAŞ

Fakülte Anabilim Dalı : BİLGİSAYAR MÜHENDİSLİĞİ

Bu tez .. / .. / … tarihinde aşağıdaki jüri tarafından oybirliği / oyçokluğu ile kabul edilmiştir.

…….	…….	…….
Jüri Başkanı	**Üye**	**Üye**

ÖNSÖZ

Projemizde bize yol gösteren Dr. Öğr. Üyesi Hüseyin ESKİ'ye teşekkürlerimizi sunarız.

İÇİNDEKİLER

ÖNSÖZ……..... iii
İÇİNDEKİLER..... iv
SİMGELER VE KISALTMALAR LİSTESİ..... vi
ŞEKİLLER LİSTESİ..... vii
TABLOLAR LİSTESİ..... viii
ÖZET..... ix

BÖLÜM 1.
GİRİŞ..... 11
1.1.Literatür Çalışması..... 11
1.2. Makine Öğrenmesi..... 12
1.2.1. Denetimli Makine Öğrenmesi..... 13
1.2.2. Denetimsiz Makine Öğrenmesi 14
1.2.3. Yarı Denetimli Öğrenmesi……..... 14
1.2.4.Takviyeli Makine Öğrenmesi..... 15
1.2.5. Özel Algoritmalar..... 15
1.3. Makine Öğrenmesi..... 16

BÖLÜM 2.
ÖN HAZIRLIK……..... 17
2.1. Veriseti için Özellik Çıkarımı…..... 17
2.2. Donanım Mimarisi..... 18
2.3. Yazılım Mimarisi..... 18
2.4. Bulut Mimarisi..... 18

BÖLÜM 3.
OLTALAMA SİTESİ TESPİT EDİLMESİ..... 19
3.1. Kullanılan Modeller..... 19
3.1.1 Random Forest Modeli..... 19
3.1.2 Gradient Boosting Modeli..... 20
3.1.3 Logistic Regression Modeli…..... 21
3.2. Modellerin Başarı Kriterleri..... 22
3.3. Modellerin Karşılaştırılması..... 24
3.4 Parametrelerin Başarıya Etkileri..... 27
BÖLÜM 4.
SONUÇLAR VE ÖNERİLER..... 27
4.1. Sonuç 27
4.2. Öneriler 29

KAYNAKLAR…….. 30

BSM 498 BİTİRME ÇALIŞMASI DEĞERLENDİRME VE SÖZLÜ SINAV TUTANAĞI…… 31

SİMGELER VE KISALTMALAR LİSTESİ

A	: Numune kesit alanı
Av	: Sıkışma katsayısı
ASTM	: Amerikan standart
Cc	: Sıkışma indisi
Cj	: Değiştirilmiş sıkışma indisi
Cr	: Yeniden yükleme indisi
Cp	: Değiştirilmiş yeniden yükleme indisi
Cv	: Konsolidasyon katsayısı
Cl	: İkincil konsolidasyon (sıkışma) katsayısı
E	: Boşluk oranı
e0	: Başlangıç boşluk oranı
Ep	: Birincil konsolidasyon sonundaki boşluk oranı
H0	: Sıkışabilir tabakanın kalınlığı
Hd	: Numune kesit yüksekliği
Ip	: Plastisite indisi
K	: Permeabilite (geçirgenlik) katsayısı
Mv	: Hacimsel sıkışma katsayısı
R0	: Başlangıç okuma değeri
R50	: %50 oturmaya karşı gelen okuma değeri
R90	: %90 oturmaya karşı gelen okuma değeri
R100	: %100 oturmaya karşı gelen okuma değeri

ŞEKİLLER LİSTESİ

Şekil 1.1.	Denetimli Makine Öğrenmesi……...	13
Şekil 1.2.	Denetimsiz Makine Öğrenmesi…….……	14
Şekil 1.3.	Yarı Denetimli Makine Öğrenmesi……..…….	14
Şekil 1.4.	Takviyeli Makine Öğrenmesi……..……	15
Şekil 3.1.	Random Forest Modeli……	19
Şekil 3.2.	Gradient Boosting Modeli……..	20
Şekil 3.3.	Lojistik Regresyon Model ……..	21
Şekil 3.4.	Eğitim verisetindeki veri türü ……..……..	24
Şekil 3.5.	Test verisetindeki veri türü ……	24
Şekil 3.6.	Gradient Boost algoritması test ve eğitim sonuçları……..	24
Şekil 3.7.	Random Forest algoritması test ve eğitim sonuçları …….	25
Şekil 3.8.	Lojistik Regresyon algoritması test ve eğitim sonuçları …	25
Şekil 3.9.	Gradient Boost eğitim ve test sonu başarım oranları…..…	25
Şekil 3.10.	Random Regresyon eğitim ve test sonu başarım oranlar…	26
Şekil 3.11.	Lojistik Regresyon eğitim ve test sonu başarım oranlar….	26
Şekil 3.12.	Kullanılan parametrelerin modellerin başarısına etkileri...	27
Şekil 4.1.	Kullanılan parametrelerin modellerin başarısına etkileri...	28
Şekil 4.2.	Kullanılan parametrelerin modellerin başarısına etkileri...	29

TABLOLAR LİSTESİ

Tablo 2.1.	Özellik Çıkarım Tablosu..……...……	17
Tablo 3.3.	Karışıklık Matrisi…….…....	22

ÖZET

Anahtar kelimeler: Oltalama Sitesi Tespiti, Oltalama Web Sitesi, Makine Öğrenmesi

Günümüzde kimlik avı yapan sitelerin sayısı gün geçtikçe artmaktadır. Bu sitelerin amacı kişilerin kişisel bilgilerini ele geçirerek çıkar sağlamaktır. Sosyal medya hesaplarındaki kullanıcı bilgileri, alışveriş sitelerindeki adres ve kart bilgileri, banka hesabına ait olan bilgiler başkalarının eline geçmesini istemeyeceğimiz kişisel bilgilerimizdir. Bu tür bilgilerin online ortamda kullanılması bu tür oltalama sitelerinden korunmak önemli hale gelmiştir.

Kullanıcı güvenliğini sağlamak üzerine çalışan firmalar çeşitli servis ve hizmetlerle kullanıcılarını bu sitelerden korumak için çeşitli çalışmalar yapmaktadır. İnternet ortamının kullanıcı sayısının hızla artmasıyla bu tür oltalama sitelerinin sayısı da artmaktadır. Bu sitelerden kullanıcıyı korumanın en kolay yolu sahte sitelerin önceden tespit edilip engellenmesi olacaktır. Yapay zekanın alt dalı olan makine öğrenmesi algoritmalarıyla oltalama sitesine ait özelliklere bakarak yapılan sınıflandırmalar başarı yüzdesini arttırmıştır.

Bu çalışmada oltalama sitelerinin tespiti için makine öğrenmesi algoritmaları kullanılmıştır. Analiz edilecek sitenin URL'sine ait bilgiler ve sitenin içindeki tehlike arz eden tüm bilgiler taranarak kullanıcıya bilgi verilir.

BÖLÜM 1. GİRİŞ

We are Social ve Hootsuite tarafından hazırlanan “Digital in 2018 Western Asia” istatistiklerine göre Dünya’da 4,021 milyar internet kullanıcısının 5,135 milyarı ise mobil internet kullanıcısından oluşmaktadır. Bu da demek oluyor ki; Dünya popülasyonun %53’ü internet kullanırken; bu oranın %68’i ise mobil interneti kullanmaktadır. Verilen sonuçlarda bir önceki yıla göre mobil internet kullanım oranı ise %4 oranında yani 218 milyon kişi artmıştır. Türkiye’de ise nüfusun %67‘sine tekabül eden 54,33 milyon kişinin internet kullanıcısı, 51,45 milyon kişinin de mobil kullanıcı olduğu verilmektedir ve bu rakamın son 1 yıl içinde %5 artışla 3 milyon kişi arttığını görülmektedir [1]. Bu durum bize interneti kullanan kişi sayısının hızla arttığını ve bununla birlikte bilgi güvenliğinin de önemli bir hale geldiğini gösteriyor.

Kimlik avı saldırıları genelde kişilerin e-posta hesaplarına; anket, kampanya, hediye, bağışlar gibi kullanıcıyı etkileyecek ve harekete geçirecek sahte iletiler gönderilerek gerçekleştirilir. Buradaki amaç kullanıcının kimlik, kart bilgileri, çeşitli parolaları ele geçirip kar sağlamaktır. Kimlik avı saldırılarının tarihteki ilk örneği 2004 yılında, “America Online” web sitesinin taklidini oluşturarak yapılmıştır. Bu olay sonucunda kullanıcının kredi kartı bilgilerine erişim sağlamışlardır. Bu saldırıların boyutu zamanla artmış ve daha karmaşık hale gelmiştir. Bu sebeplerden dolayı kimlik avı saldırılarının tespiti ve kullanıcıların bu saldırılardan korunması kritik önem arz etmektedir. Bunu sağlayabilmek için gelişen teknolojinin bir parçası olan makine öğrenmesinin sınıflandırma algoritmalarıyla gerçek zamanlı olarak tespit edilmesi yapay zeka çalışmalarının sunduğu önemli avantajlardan biridir.

1.1. 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 [4]. 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 [5]. Çalışmaların çoğunun (Hema vd. 2020) Naive Bayes, destek vektör, karar ağacı ve rasgele orman gibi tanıdık makine öğrenimi algoritmaları kullanılarak yapıldığı sonucuna ulaşılmıştır [6]. (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 [7].

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 [8].

Benzer bir çalışmada Sahingoz vd. (2019); kimlik avı eposta 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 [9].

Diğer bir çalışmada Kalaycı vd. (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 [10]. Biz de literatürdeki bu çalışmalar doğrultusunda web sitesi içeriğine ve adresine ait belirlenmiş toplam 15 özellik kullanılarak; bu web sitesinin kimlik avı amacıyla hazırlanmış sitenin tespiti başarım oranı en yüksek olan Rastgele Orman (RF) algoritması kullanılarak yapılmıştır. Çalışmada aşağıda listelenen katkılar sunulmaktadır:

- Şüpheli web sitesi adresine ait belirlenmiş 13 özellik ve web sitesinin içeriğine ait 8 özelliğe bakılarak sitenin oltalama amaçlı olup olmadığı tespit edilmiştir.
- Kimlik avı web sayfası tespiti için yeni bir veri seti hazırlanmıştır.

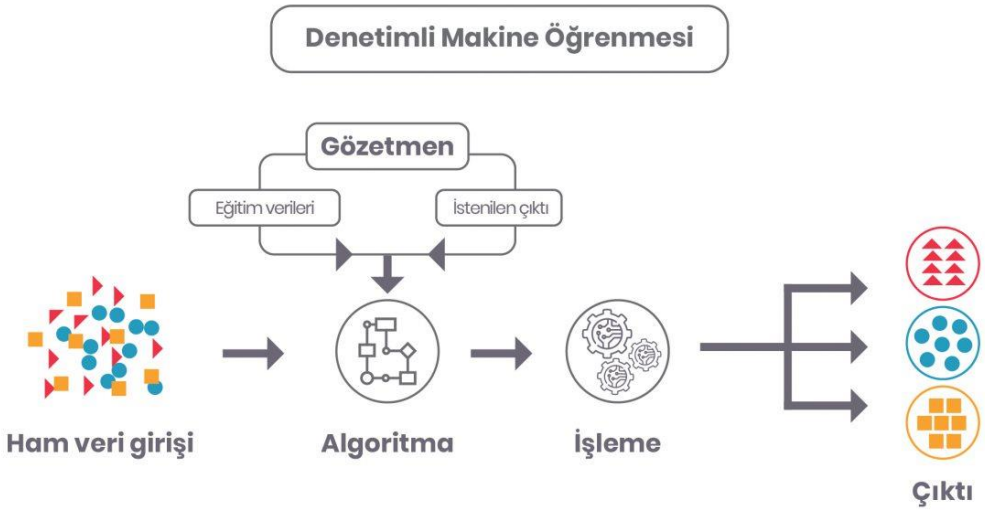
1.2. Makine Öğrenmesi

Yapay zekâ kavramı resmi olarak ilk defa 1955 yılında Honever, New Hampshire Dartmouth College'da yapılan bir konferansta John McCarthy tarafından kullanılmıştır. Yapay zekâ, kendi kendine öğrenebilme, muhakeme yapabilme ve mantıklı kararlar verme gibi insana ait becerileri makinelere uyarlamayı amaçlayan bir alandır. İlk kullanımdan sonra bu alan üzerinde birçok çalışma yapılmış ve hızlı bir ilerleme sağlanmıştır. Donanım teknolojilerinde gelişimle beraber bu alan desteklenmiştir. Günümüzde bilgisayarların da öğrenebileceğini herkes kabul etmekte ve bu teknolojiden faydalarak sorunlarına çözüm bulmak istemektedir. Gelişim süreci içerisinde birbirinden farklı yeni modeller geliştirilmiştir. Bu modellerden makine öğrenmesidir. Makine öğrenmesi, insanların öğrenme şekillerini taklit etmek için veri ve algoritmaların kullanımına odaklanıp doğruluğunu kademeli olarak artıran bir yapay zeka (AI) ve bilgisayar bilimi dalıdır [2]. Makine öğrenmesi terimini ilk kez IBM laboratuvarına katıldıktan sonra Arthur Samuel tarafından 1952 yılında dama oynamak için tasarlanmıştır. Bu tasarım makine öğrenmesinin temelini oluşturmuştur. Basit bir şekilde verilerin ayrıştırılması için bir algoritma kullanma, ayrıştırılan verileri öğrenme ve daha sonra yeni gelen verilerin ne olduğuna dair tahmin yürütme işlemlerini yapmak gibi becerileri olan eğitilmiş bir makinedir. Makine öğrenmesi algoritmaları, dış dünyadan aldığı verileri öğrenme sürecinde kullanarak makinenin kendini daha akıllı hale getirmesini sağlar. Bu algoritmaların düzenli olarak yeni verilerle sürekli beslenmesi; sınıflandırma, tahmine dayalı

modelleme ve verilerin analiziyle ilgili çeşitli görevler konusunda büyük çalışmaların ortaya konmasını sağlar. Makine öğrenmesi algoritmaları kendi içinde 4 kategoriye ayrılır ve her bir algoritma farklı problemlere özel çözüm sunmaktadır. Bunlar denetimli, denetimsiz, yarı denetimli ve takviye algoritmalarıdır. Bunun haricinde daha farklı problemlere yönelik özel algoritmalar geliştirilmiştir.

1.2.1. Denetimli Makine Öğrenmesi (Supervised Algorithms)

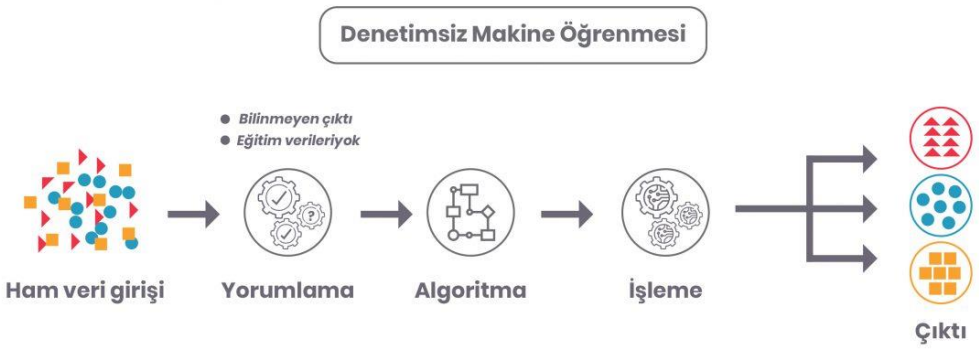
Sürecin belirli kısmında kullanıcının denetimini gerekli kılan algoritmalardır. Geliştirici eğitim sürecinden önce verilerin etiketlenmesi ve algoritmanın izleyeceği yol haritasını kesin ve net bir şekilde belirlemelidir. Bu sayede makine etiketlenmiş verileri kullanarak gelecek hakkında tahminler yürütür.



Şekil 1.1 [3]

1.2.2. Denetimsiz Makine Öğrenmesi (Unsupervised Algorithms)

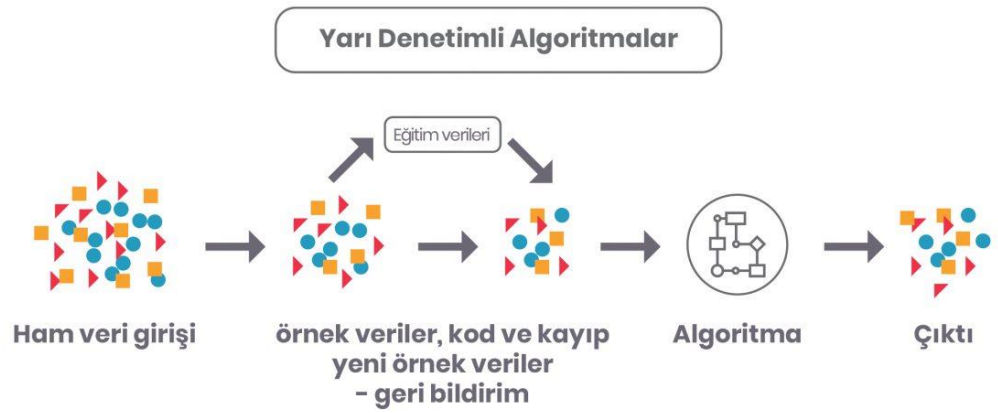
Denetimli öğrenmenin aksine herhangi bir sınırlandırmanın olmadığı algoritmalardır. Öğrenme sürecinden önce bilgiler sınıflandırmaya ve etiketlendirmeye tabi tutulmaz. Makinenin bu karmaşık bilgilerden mantıklı çıkarımlar yapmaya çalışır.



Şekil 1.2 [3]

1.2.3. Yarı Denetimli Makine Öğrenmesi (Semi supervised algorithms)

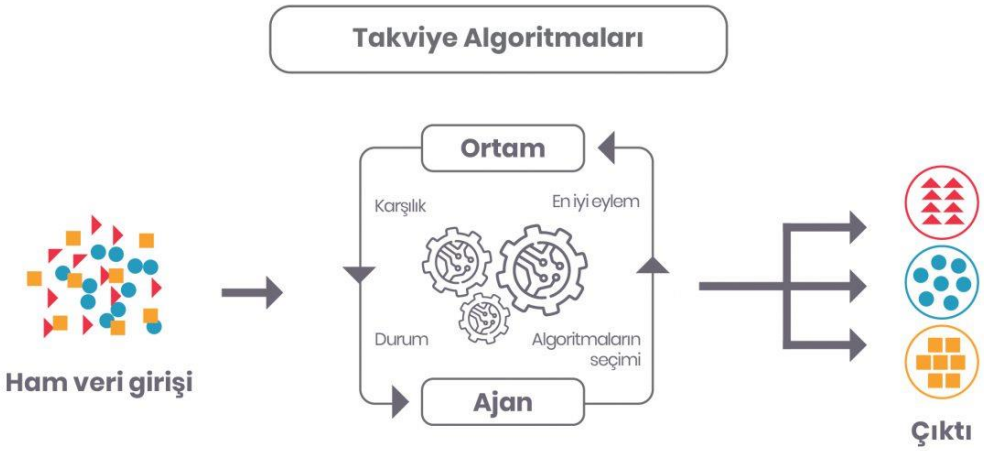
Bu öğrenme algoritmalarında hem etiketli hem etiketsiz veriler eğitim amacıyla kullanılabilir. Az miktarda etiketli ve büyük miktarda etiketsiz veri öğrenme başarısını artırmaktadır.



Şekil 1.3 [3]

1.2.4. Takviyeli Makine Öğrenmesi (Reinforcement Algorithms)

Bu makine öğrenmesi algoritmalarında ortamın keşfi önemli bir yer tutmaktadır. Makine çeşitli faaliyetler yaparak ortamdan geri bildirim alır. Bu bildirimleri gözlemler ve öğrenmesine devam eder. Bu süreç algoritmanın doğru stratejiyi seçip doğru sonuçları elde edene kadar devam eder.



Şekil 1.4 [3]

1.2.5. Özel Algoritmalar

Bu algoritmalar farklı görevleri yerine getirme konusunda daha başarılı performans gösterirler. Diğer farklı algoritmalarla senkronize olarak çalışabilirler. Örnek olarak şu 5 temel algoritma verilebilir. Bunlar regresyon, sınıflandırma, topluluk, ilişkilendirme ve kümelemedir.

Regresyon; iki veya daha fazla değişken arasındaki ilişkiyi anlamamızı sağlar. Doğrusal, lojistik, ridge, lasso ve polinom alt başlıklarıdır.

Sınıflandırma; önceden etiketlenmiş olan veri setine göre kategorilere ayırma algoritmalarıdır. KNN, karar ağaçları, naive bayes, svm(support vector machine) alt başlıklarıdır.

Topluluk; iki veya daha fazla makine öğrenmesi algoritmalarını birleştirerek daha başarılı sonuçlar üretilmesini amaçlar. Bagging, boosting, stacking gibi 3 temel türe sahiptir.

Kümeleme; benzer özelliklere sahip verileri gruplamak için kullanılan algoritmalardır. Bunlar centroid tabanlı, yoğunluğa dayalı, dağıtım tabanlı ve hiyerarşik kümeleme gibi alt başlıkları vardır.

İlişkilendirme; belirli verilerin kümelenmesi sonucunda oluşan kümenin içerisindeki verilerin ilişkilerini analiz eder. Alışveriş sitelerinde bolca kullanılmaktadır.

1.3. Amaç ve Önem

Günümüzde internetin kullanımı ve internete bağlı cihaz sayısı sürekli artmaktadır. İnternet ortamında veriler yoğun bir şekilde paylaşılabilir, depolanabilir ve etkileşime girilebilir bir duruma gelmiştir. İnternete olan ihtiyaç arttıkça kişisel verinin korunması son derece önemli bir hal almıştır. Bu önemli konu üzerine aşağıdaki amacı açıklanan çalışma yapılmıştır.

Bu çalışmanın amacı kimlik avı yapmak amacıyla hazırlanmış olan sitelerin makine öğrenmesi algoritmalarıyla kullanıcı siteyi ziyaret etmeden veya URL adresine tıklamadan önce kullanıcıyı uyararak kimlik avı saldırısının önüne geçmektir.

BÖLÜM 2. ÖN HAZIRLIK

Projede modelin eğitimi için veri seti önemli bir yer tutmaktadır. Ön hazırlık aşamasında veri seti oluşturulmuştur.

2.1. Veri Seti için Özellik Çıkarımı

Makine öğrenmesinde en önemli noktası veri setinin yeterliliği sağlaması ve projeye uygun olmasıdır. Projede veri seti oluştururken oltalama sitelerinin belli başlı ayırt edici noktaları göz önüne alınarak aşağıda tabloda bulunan özellikler belirlenmiştir.

Tablo 2.1. Özellik çıkarım tablosu

Özellik Numarası	Özellik İsmi	Özellik Numarası	Özellik İsmi
1	URL uzunluğu	12	Redirect kontrolü
2	Kısa link kontrolü	13	Mouse-over kontrolü
3	@ sembolü kontrolü	14	Sağ tık kontrolü
4	URL'de subdomain kontrolü	15	Pop-up kontrolü
5	SSL varlığı	16	Iframe kontrolü
6	Domain tarih kontrolü	17	Domain yaşı kontrolü
7	Port yönlendirme kontrolü	18	DNS kaydı
8	https kontrolü	19	Pagerank kontrolü
9	Site içerik kontrolü	20	Google index kaydı
10	SFH kontrolü	21	İstatistiksel rapor
11	Email gönderim kontrolü	-	-

2.2. Donanım Mimarisi

Yapılan bu çalışmada makine öğrenmesi algoritmalarını içermesi sebebiyle yüksek donanıma ihtiyaç vardır. GPU'lar aynı anda büyük veri kümesindeki paralel yürütülen işlemleri çok hızlı bir şekilde hesaplayabilmesi sebebiyle bu projede tercih edilmiştir.

2.3. Yazılım Mimarisi

Günümüzde makine öğrenmesi projelerinde bu alanda çokça kütüphane bulunduran ve kullanıcıya hızlı geliştirme imkanı sunan python programlama dili yaygın şekilde kullanılmaktadır. Python veri bilimi alanında kullanılan en yaygın dillerden biridir. Makine öğrenmesinde kullanılan scikit-learn ve doğal dil işleme yazılımı nltk gibi daha özel kütüphane desteği de bulunur.

2.4. Bulut Mimarisi

Makine öğrenmesi, derin öğrenme, yapay zeka gibi teknolojiler yüksek donanım gerektirir. Bu amaçla hem donanım performansı hem ortak proje geliştirme imkanı sunması sebebiyle Google Colab platformu bu çalışmada tercih edilmiştir. Colab bir bulut depolama servisi olan Google Drive ürününü altyapı olarak kullanan bir bulut mimari sistemine sahip çalışma ortamıdır.

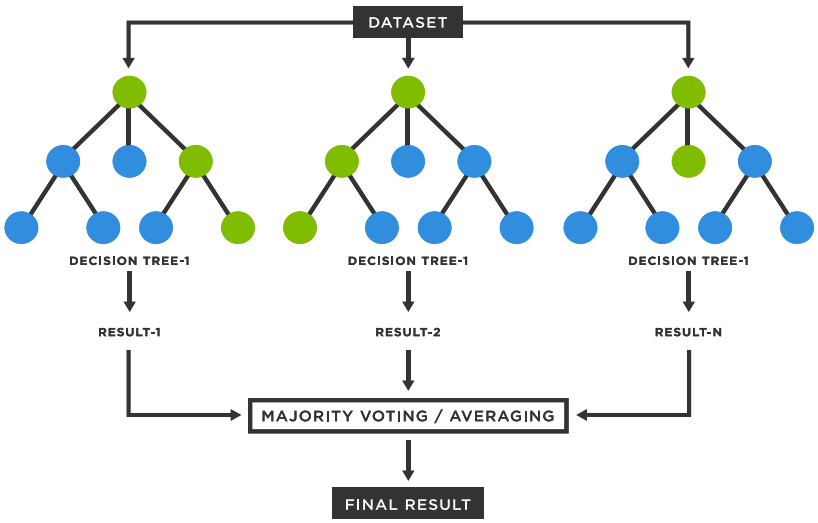
BÖLÜM 3. OLTALAMA SİTESİ TESPİT EDİLMESİ

Son yapılan çalışmalarda birçok model ortaya konmuştur. Bu modellerin başarım oranları probleme ve kullanılan verisetine göre değişmektedir. Bu sebeple yaptığımız çalışmada birden fazla model kullanılıp başarı oranları karşılaştırılıp en iyi başarım oranını veren model uygulamanın gerçeklemesinde kullanılmıştır.

3.1. Kullanılan Modeller

3.1.1. Random Forest Modeli

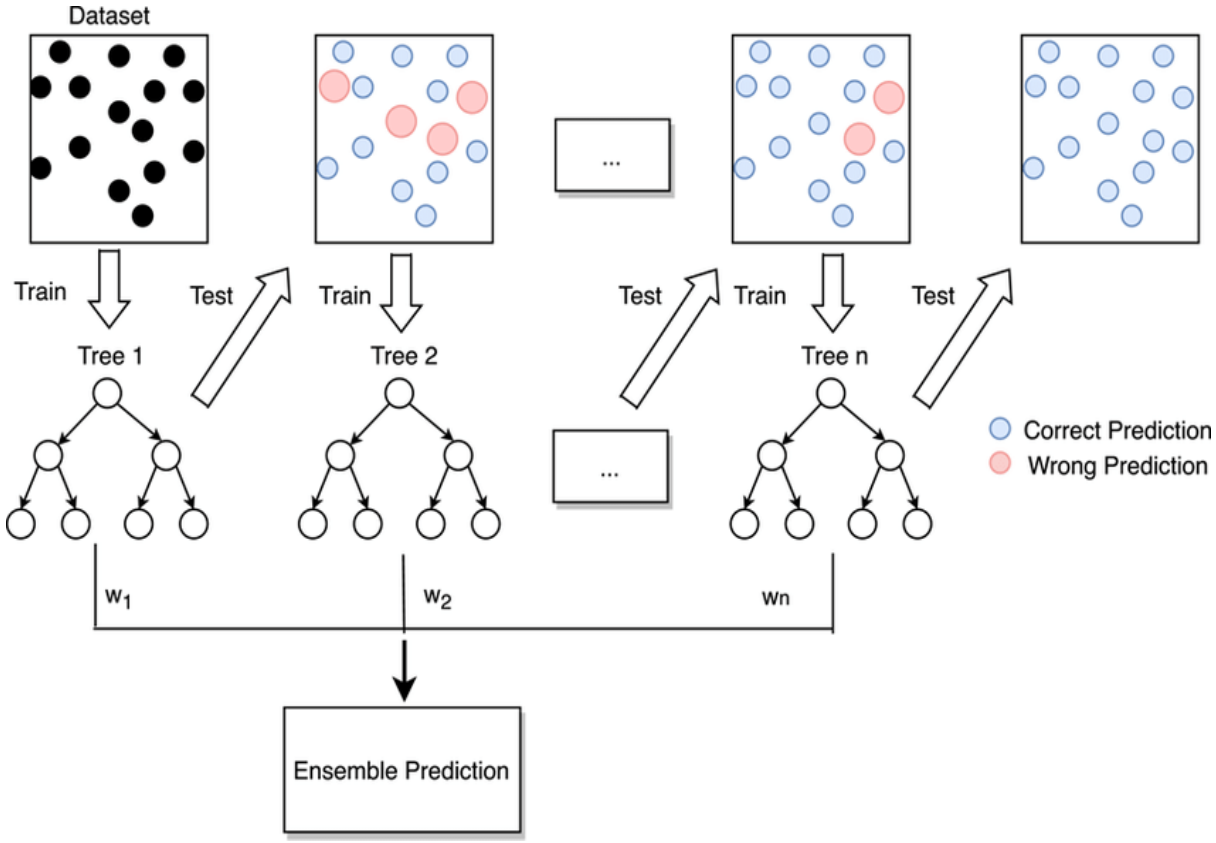
Random forest algoritması, denetimli sınıflandırma algoritmalarından biridir. Hem regresyon hem de sınıflandırma problemlerinde uygulanabilir olmasından dolayı popüler makine öğrenmesi modellerinden biridir. Algoritmanın çalışma mantığı birden fazla karar ağacı üreterek sınıflandırma işlemi sırasında sınıflandırma değerini yükseltmeyi amaçlamaktadır. Bu modelin en büyük avantajı karar ağaçlarının en büyük problemlerinden biri olan overfitting sorununu azaltmasıdır.



Şekil 3.1 [11]

3.1.2. Gradient Boosting Modeli

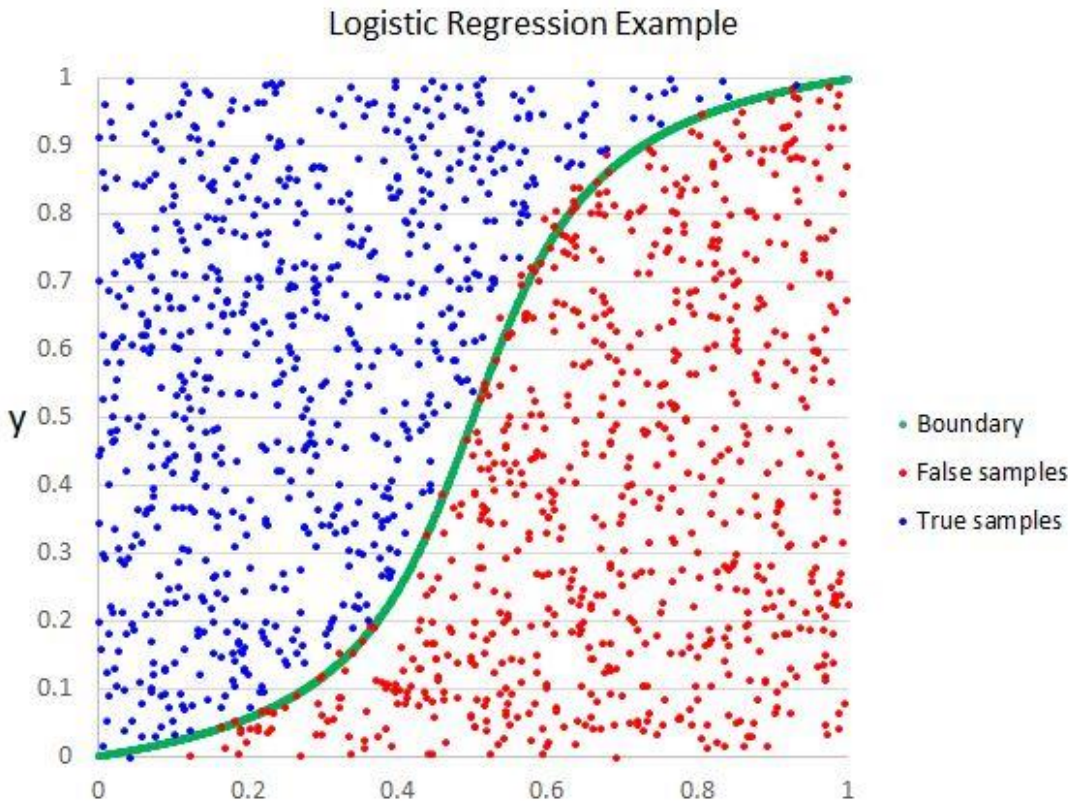
Gradient Boost algoritması regresyon ve sınıflandırma problemleri için karar ağaçlarına benzer tahmin modelleri oluşturan bir makine öğrenmesi tekniğidir. Gradient Boost algoritmasının çalışma mantığı her ağaçtan sonra bir optimizasyon yapmak için düğüm oluşturmak yerine yaprak oluşturarak başlar. Sonrasında tahminde oluşan hatalar dikkate alınarak yeni ağaçlar oluşturulur. Bu yapraklar tüm ağırlıklar için bir ilk tahmin skorudur. Bu skor ortalama bir değerdir. Ardından bu model bir ağaç oluşturur Karar verilen ağaç sayısına ya da modelden daha fazla gelişme kaydedilemeyinceye kadar devam eder.



Şekil 3.2 [12]

3.1.3. Logictic Regression Modeli

Doğrusal sınıflandırma problemlerinde yaygın biçimde kullanılan algoritmalardan biridir. İsminden bir regresyon algoritması olduğu düşünülmesine rağmen bu bir sınıflandırma algoritmasıdır. Lojistik regresyon sınıflandırma yapmak için sigmoid fonksiyonunu kullanır. Bu fonksiyon "S" şeklinde olan bir eğriyi temsil eder. Bu fonksiyonun temel mantığı verileri 0 ve 1 arasında sıkıştırmaktır. Böylece sınıflandırma yapmaya olanak sağlamaktadır. Veri seti doğrusal olması bu modelin çalışması için kritik öneme sahiptir. Aksi halde yeterli performansı veremez.



Şekil 3.3 [13]

3.2. Modellerin Başarı Kriterleri

Test kümesini daha önce elde ettiğimiz modele girecek ve bir dizi gösterge elde edeceğiz. Bu göstergeler, modelin performansını değerlendirmek için kullanılır. Aşağıda bu göstergeler açıklanmıştır.

True Positive(TP): Doğru algılanan oltalama sitesi sayısını ifade eder.
False Negative(FN): Oltalama sitesi olan fakat güvenli olarak algılanan sitelerin sayısını ifade eder.
False Positive(FP): Güvenli olan fakat oltalama sitesi olarak tespit edilen site sayısını ifade eder.
True Negative(TN): Doğru algılanan güvenli olan sitelerin sayısını ifade eder.

Aşağıda bu sınıflandırmanın karışıklık matrisi gösterilmiştir.

Tahmin Gerçek	0(Güvenli)	1(Oltalama)
0(Güvenli)	TN	FP
1(Oltalama)	FN	TP

Tablo 3.1

Doğruluk (Accuracy)
Doğruluk, modelin veri kümesindeki gerçek sınıflandırmalarla karşılaştırıldığında elde edilen doğru tahminlerin oranını gösterir. Aşağıdaki formülle hesaplanır.

$$Doğruluk = \frac{TP + TN}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

Tutturma (Precision)

Kesinlik, doğru tahmin edilen pozitif örneklerin toplam pozitif tahminlere oranıdır. Aşağıdaki formülle hesaplanır.

$$Tutturma = \frac{TP}{\mathrm{TP} + \mathrm{FP}}$$

Bulma (Recall)

Hatırlama, doğru tahmin edilen oltalama sitelerinin toplam oltalama sitelerine oranıdır. Aşağıdaki formülle hesaplanır.

$$Bulma = \frac{TP}{\mathrm{TP} + \mathrm{FP}}$$

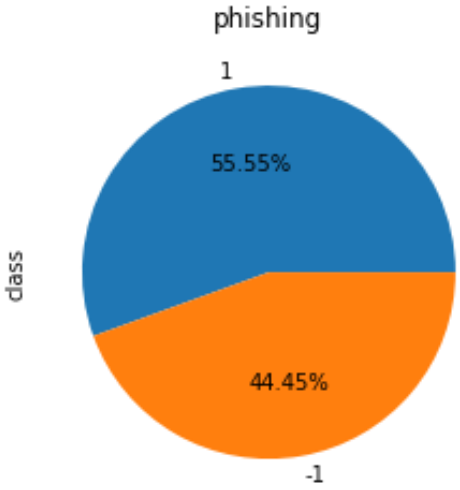
F- Değeri (F1 Skoru)

F- skoru, hassasiyet ve hatırlamanın harmonik ortalamasıdır. F-skoru 0 ile 1 arasında bir değer alır. Burada 1 mükemmel hassasiyete eşittir ve her ikisi de 1'e eşittir. F1 skoru, hassasiyet ve hatırlama arasında bir ilişki metriği sağlamak için faydalıdır. Çünkü bu iki değer bir şekilde ters benzetmeye sahiptir. Aşağıdaki formülle ifade edilir.

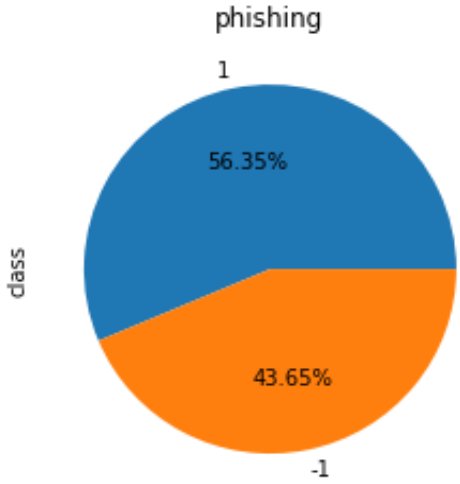
$$F - değeri = 2 * \frac{tutturma * bulma}{\mathrm{tutturma} + \mathrm{bulma}}$$

3.3. Kullanılan Modellerin Karşılaştırmaları

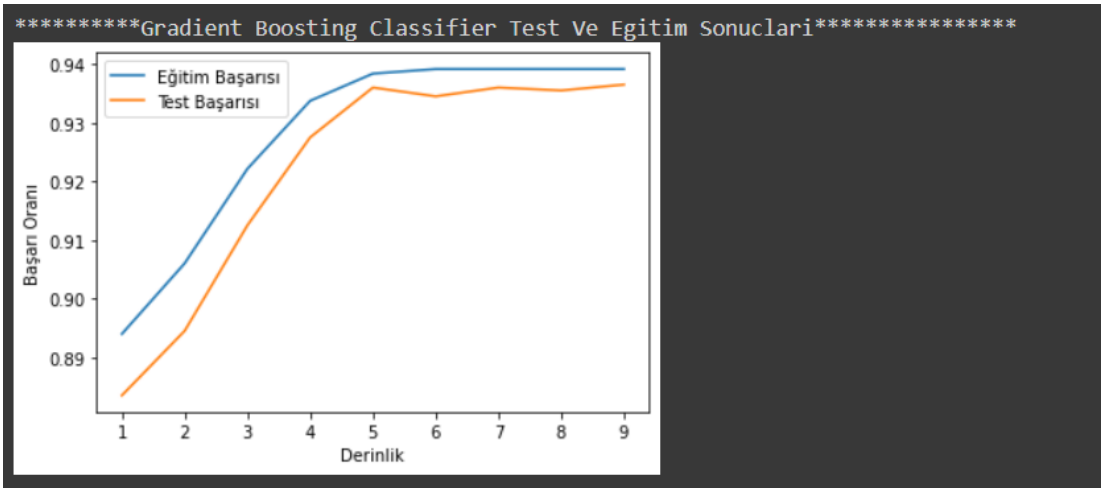
Projede lojistik regresyon, random forest ve gradient boosting modelleri kullanılmış ve başarım oranları karşılaştırılarak gösterilmiştir. Projede yaklaşık olarak 12000 adet veri kullanılmıştır. Bu verilerin 10000 adedi bu modellerin eğitim aşamasında 2000 adedi de test aşamasında kullanılarak aşağıdaki sonuçlar elde edilmiştir.



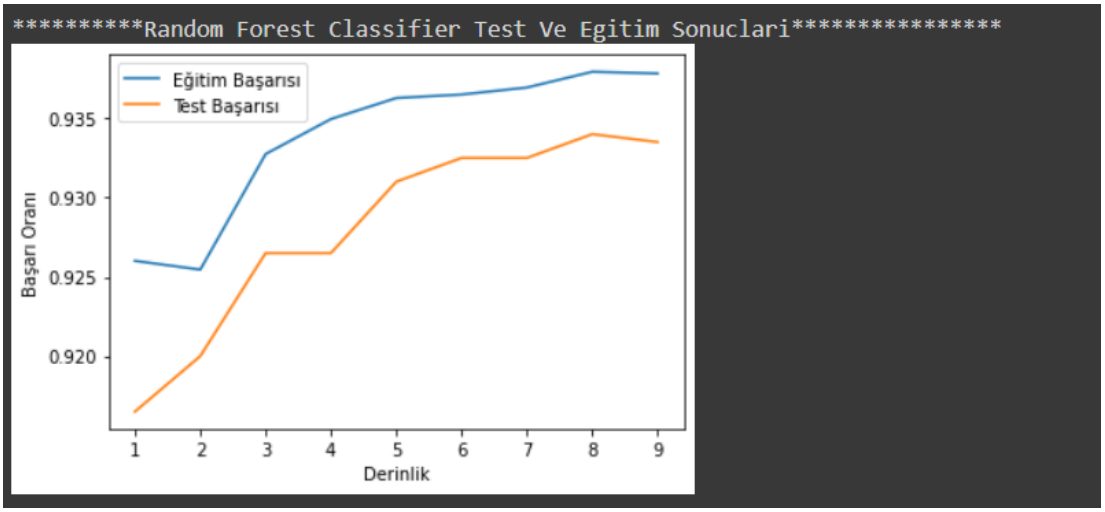
Şekil 3.4. Eğitim verisetindeki veri türü



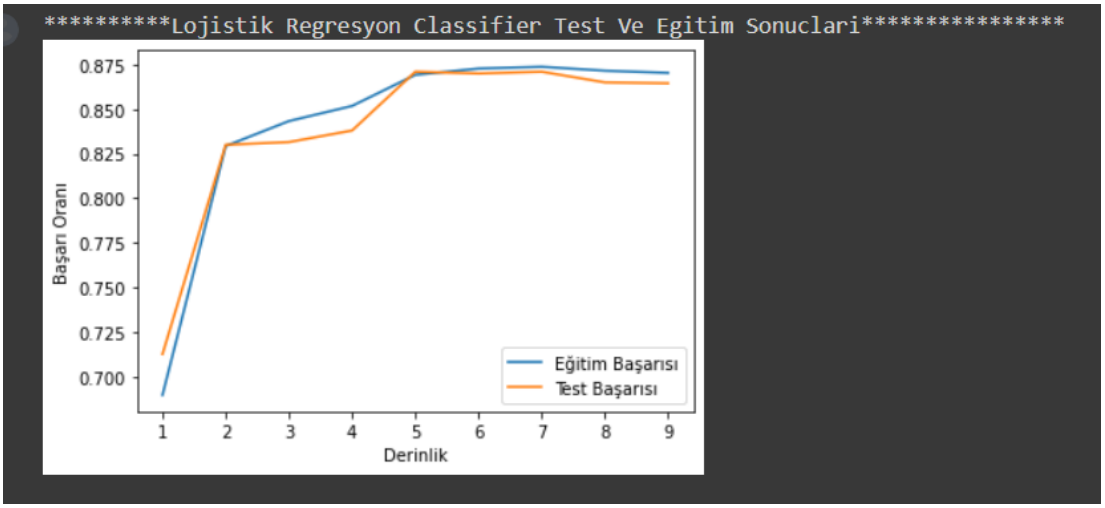
Şekil 3.5. Test verisetindeki veri türü



Şekil 3.6



Şekil 3.7



Şekil 3.8

```
**********Gradient Boosting Classifier**********
Gradient Boosting Classifier Egitim Accuracy  => 0.934
Gradient Boosting Classifier Test Accuracy => 0.927

Gradient Boosting Classifier Egitim f1_score  =>  0.941
Gradient Boosting Classifier Test f1_score =>  0.936

Gradient Boosting Classifier Egitim Recall  =>  0.944
Gradient Boosting Classifier Test Recall => 0.940

Gradient Boosting Classifier Egitim Precision  => 0.937
Gradient Boosting Classifier Test Precision  => 0.932
**********
```

Şekil 3.9 Modellerin eğitim işlemi sonrası eğitim ve test verisiyle test edilen başarım oranları

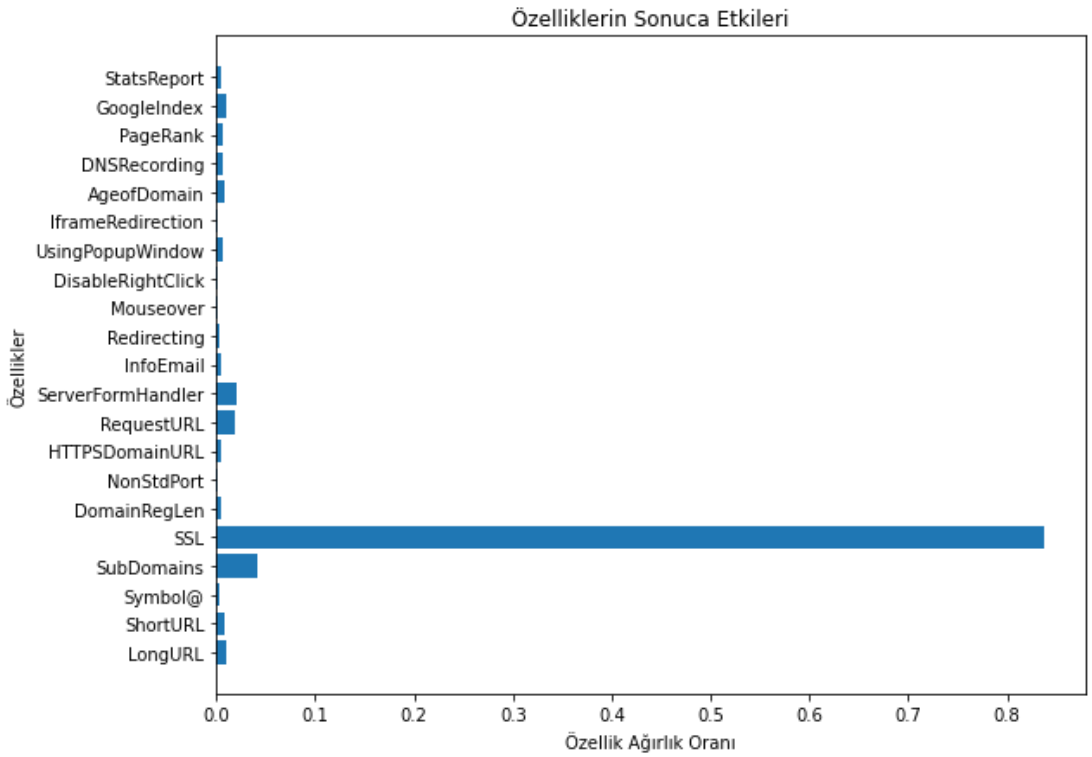
```
**********RANDOM FOREST**********
Random Forest Egitim Accuracy  => 0.938
Random Forest Test Accuracy => 0.933

Random Forest Egitim f1_score  =>  0.944
Random Forest Test f1_score  => 0.940

Random Forest Egitim Recall => 0.945
Random Forest Test Recall => 0.937

Random Forest Egitim Precision  => 0.943
Random Forest Test Precision  => 0.944
**********
```

Şekil 3.10. Modellerin eğitim işlemi sonrası eğitim

```
**********
**********Logistic Regression**********
Logistic Regression Egitim Accuracy  => 0.870
Logistic Regression Test Accuracy => 0.864

Logistic Regression Egitim f1_score  => 0.886
Logistic Regression Test f1_score  => 0.883

Logistic Regression Egitim Recall  => 0.910
Logistic Regression Test Recall  => 0.912

Logistic Regression Egitim Precision  => 0.863
Logistic Regression Test Precision  => 0.855
```

Şekil 3.11. Modellerin eğitim işlemi sonrası eğitim

3.4. Parametrelerin Başarıya Etkileri



Şekil 3.12. Kullanılan parametrelerin modellerin başarısına etkileri

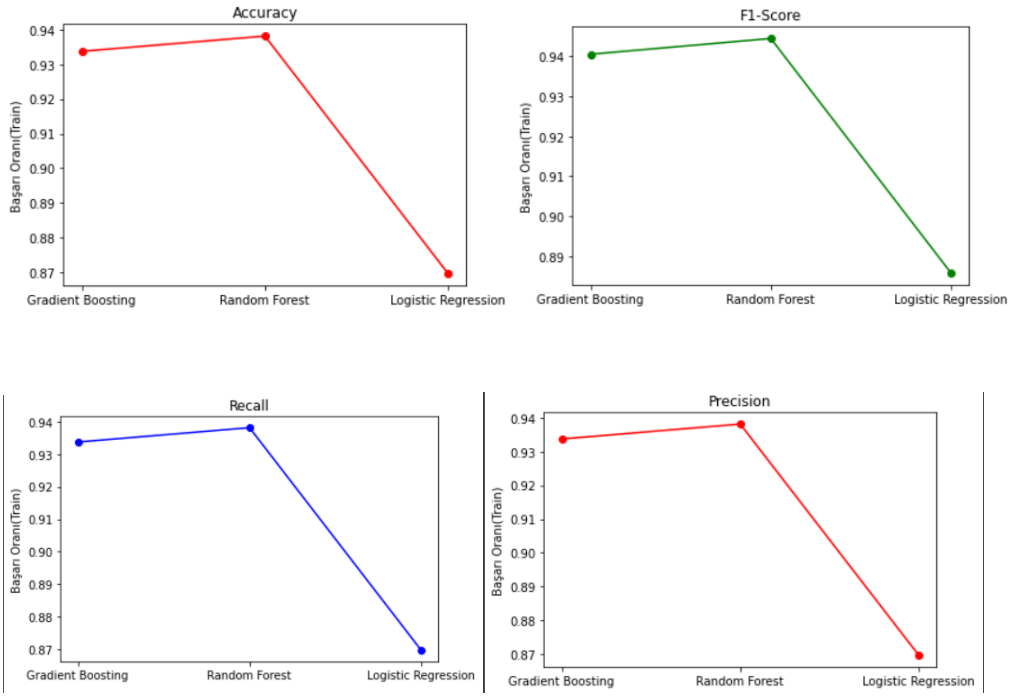
BÖLÜM 4. SONUÇ VE ÖNERİLER

4.1. Sonuç

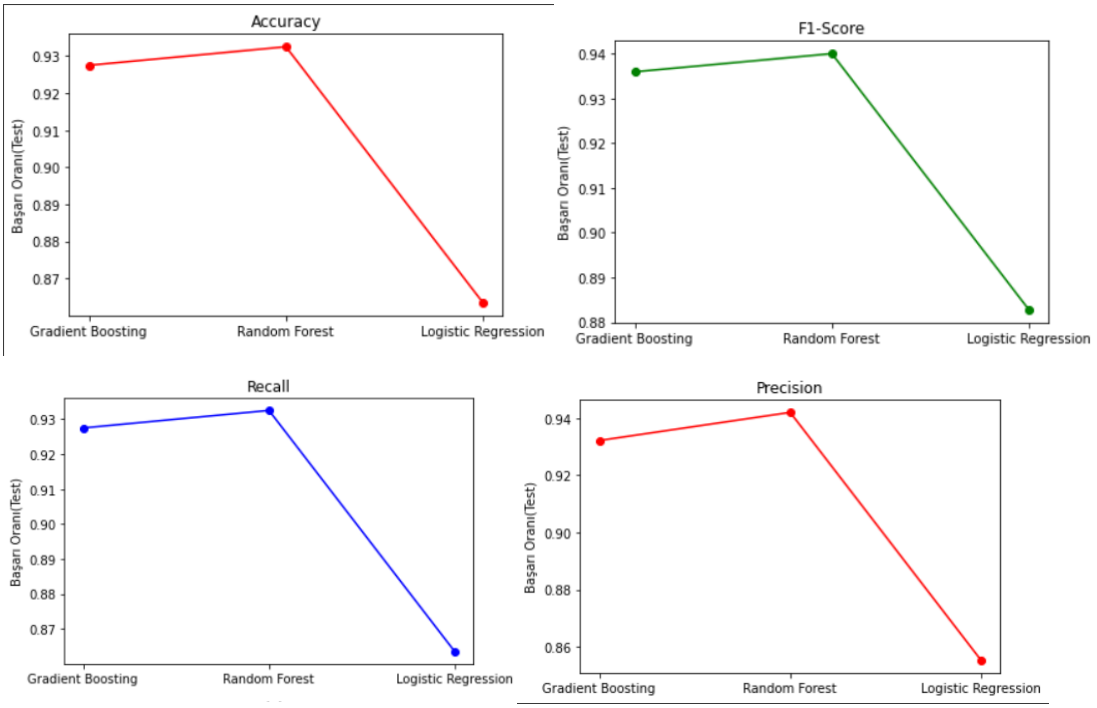
Gerçekleştirilen çalışmada web sayfalarının belirlenen özelliklerinden, sayfanın oltalama amacıyla hazırlanıp hazırlanmadığı makine öğrenmesi algoritmalarıyla saptanmaya çalışılmıştır. Çalışmada lojistik regresyon, rastgele orman ve gradient boosting modelleri kullanılmış ve başarımları karşılaştırılmıştır. Modeller başarılı bir veri seti ile eğitilmiş ve tahmin oranları gayet başarılı sonuçlar alınmıştır. Model öğretilmemiş verilerde de bu sonuçlara yakın başarı oranlarını yakalamıştır.

Çalışmada oltalama sitesi tespiti için 21 özellik belirlenip verisetleri buna göre düzenlenmiştir. Çalışmada parametrelerin başarıya etkileri incelendiğinde bu 21 özellik arasından gözle görülür şekilde öne çıkan özelliğin SSL varlığı olduğu görülmüştür.

Çalışmada elde edilen tüm performans ölçütleri birbirlerini teyit ederek aynı sonuçları vermektedir. Bu doğrultuda elde edilen sonuçlarının tutarlı olduğu görülmektedir. Bu çalışmanın benzer çalışmalar için geçerli sonuçlar açısından yol gösterici olacağı düşünülmektedir.



Şekil 4.1



Şekil 4.2

4.2. Öneriler

Kullanıcı deneyimini arttırmak için gelen mailler içerisindeki linklerin ayıklanarak kullanıcıya arkaplanda çalışarak sonuçları bildiren bir masaüstü uygulamasına evrilebilir. Bir diğer alternatifi kullanıcının internet kullanımı sırasında gitmek istediği web sayfalarının adreslerinin tarayıcı tarafından o anda analiz edilerek kullanıcıya bildirim veren bir tarayıcı eklentisi geliştirilmesidir.

KAYNAKLAR

[1] https://wearesocial.com/blog/2018/01/globaldigital-report-2018, Erişim tarihi: 25.06.2018

[2] https://www.ibm.com/tr-tr/cloud/learn/machine-learning

[3] https://www.turhost.com/blog/makine-ogrenmesi-machine-learning-nedir/

[4] 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.

[5] 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

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

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

[8] https://archive.ics.uci.edu/ml/index.php

[9] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029

[10] https://dergipark.org.tr/tr/pub/pajes/issue/39683/469468

[11] https://www.tibco.com/sites/tibco/files/media_entity/2021-05/random-forest-diagram.svg

[12] https://www.researchgate.net/figure/Flow-diagram-of-gradient-boosting-machine-learning-method-The-ensemble-classifiers_fig1_351542039

[13] https://helloacm.com/wp-content/uploads/2016/03/logistic-regression-example.jpg

BSM 498 BİTİRME ÇALIŞMASI
DEĞERLENDİRME VE SÖZLÜ SINAV TUTANAĞI

KONU :
ÖĞRENCİLER (Öğrenci No/AD/SOYAD):

Değerlendirme Konusu	İstenenler	Not Aralığı	Not
Yazılı Çalışma			
Çalışma klavuza uygun olarak hazırlanmış mı?	x	0-5	
Teknik Yönden			
Problemin tanımı yapılmış mı?	x	0-5	
Geliştirilecek yazılımın/donanımın mimarisini içeren blok şeması (yazılımlar için veri akış şeması (dfd) da olabilir) çizilerek açıklanmış mı?			
Blok şemadaki birimler arasındaki bilgi akışına ait model/gösterim var mı?			
Yazılımın gereksinim listesi oluşturulmuş mu?			
Kullanılan/kullanılması düşünülen araçlar/teknolojiler anlatılmış mı?			
Donanımların programlanması/konfigürasyonu için yazılım gereksinimleri belirtilmiş mi?			
UML ile modelleme yapılmış mı?			
Veritabanları kullanılmış ise kavramsal model çıkarılmış mı? (Varlık ilişki modeli, noSQL kavramsal modelleri v.b.)			
Projeye yönelik iş-zaman çizelgesi çıkarılarak maliyet analizi yapılmış mı?			
Donanım bileşenlerinin maliyet analizi (prototip-adetli seri üretim vb.) çıkarılmış mı?			
Donanım için gerekli enerji analizi (minimum-uyku-aktif-maksimum) yapılmış mı?			
Grup çalışmalarında grup üyelerinin görev tanımları verilmiş mi (iş-zaman çizelgesinde belirtilebilir)?			
Sürüm denetim sistemi (Version Control System; Git, Subversion v.s.) kullanılmış mı?			
Sistemin genel testi için uygulanan metotlar ve iyileştirme süreçlerinin dökümü verilmiş mi?			
Yazılımın sızma testi yapılmış mı?			
Performans testi yapılmış mı?			
Tasarımın uygulamasında ortaya çıkan uyumsuzluklar ve aksaklıklar belirtilerek çözüm yöntemleri tartışılmış mı?			
Yapılan işlerin zorluk derecesi?	x	0-25	
Sözlü Sınav			
Yapılan sunum başarılı mı?	x	0-5	
Soruları yanıtlama yetkinliği?	x	0-20	
Devam Durumu			
Öğrenci dönem içerisindeki raporlarını düzenli olarak hazırladı mı?	x	0-5	
Diğer Maddeler			
Toplam			

DANIŞMAN (JÜRİ ADINA):
DANIŞMAN İMZASI: