3. ANALISIS DAN DESAIN SISTEM

3.1 Analisis Masalah

Sangat penting untuk mempertimbangkan berbagai masalah yang dihadapi oleh pengguna yang terkena dampak serangan phishing saat membuat web aplikasi untuk mengidentifikasi kemungkinan phishing pada sebuah website. Salah satu masalah utama adalah bahwa sebagian besar pengguna tidak dapat membedakan serangan phishing, yang meningkatkan kemungkinan mereka terjebak dalam jebakan penipu online. Tantangan tambahan adalah kurangnya instruksi tentang keamanan digital, yang dapat membuat pengguna kurang waspada terhadap ancaman phishing dan tidak tahu langkah pencegahan yang efektif.

Dengan memahami masalah ini, pengembangan aplikasi dapat diarahkan untuk membuat solusi yang memberdayakan pengguna untuk lebih terjaga dalam phishing untuk melindungi diri mereka dari serangan phishing.

3.2 Analisis Kebutuhan

Dalam pengembangan web aplikasi untuk menganalisis potensi phishing pada sebuah website, analisis kebutuhan mendasar dapat ditentukan untuk menanggapi permasalahan utama yang dihadapi oleh pengguna yang rentan terhadap serangan phishing.di mana aplikasi harus memberikan modul edukasi yang memperkaya pemahaman pengguna tentang tanda-tanda serangan phishing. Selain itu, penting untuk memperhatikan deteksi tanda-tanda phishing dengan mengintegrasikan fitur analisis otomatis guna memberikan peringatan dini kepada pengguna.

Untuk memastikan bahwa informasi keamanan dapat diakses dengan mudah oleh pengguna, antarmuka pengguna yang ramah pengguna sangat penting.

3.3 Desain Sistem

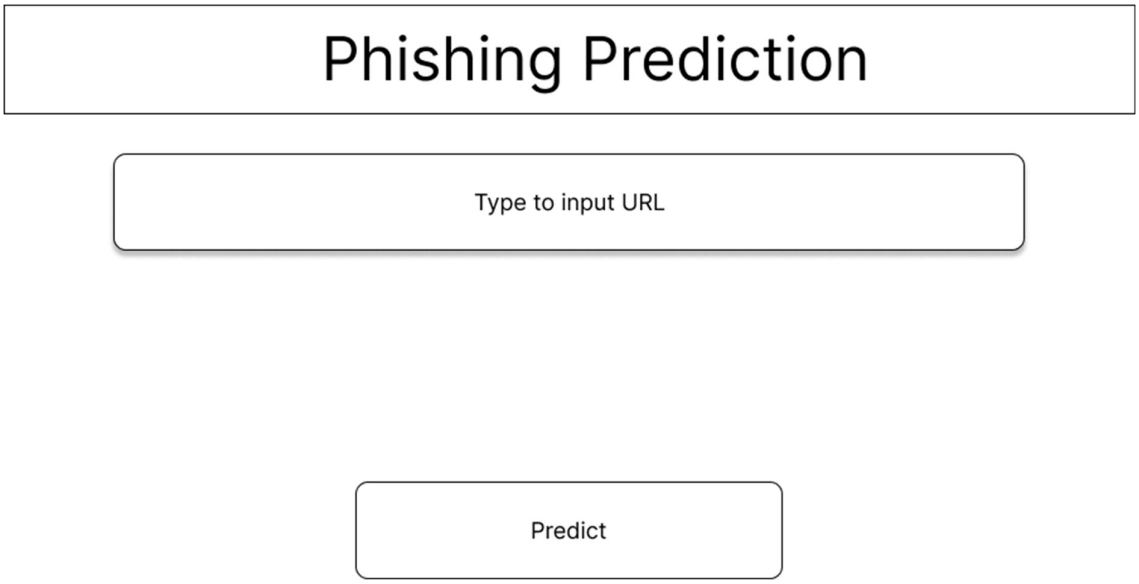
Terdapat beberapa tahapan yang dilakukan dalam mengecek data phishing pertama-tama membuat web srapping untuk mengambil data phishing dari source phishtank.org dan juga mengambil dataset phishing dari kaggle untuk digabungkan. lalu membuat fitur-fitur untuk mengambil kategori-kategori dari url untuk dijadikan menjadi dataset lalu akan dilakukan data preprocessing untuk mengubah bentuk data supaya bisa lebih efisien untuk dibaca oleh komputer lalu setelah itu akan didgunakan dua algoritma yaitu Support Vector Machine dan juga Random Forest. Lalu keduia algoritma tersebut akan dibandingkan dan memilih algoritma yang paling baik yang nantinya akan di implementasikan ke dalam sebuah web aplikasi sederhana dengan bantuan flask.

3.3.1 Flowchart Desain Pengerjaan

Gambar 3.3.1 Grafik Flowchart Pengerjaan

Pada Gambar 3.3.1 merupakan tampilan dari rencana penyusunan proses pengerjaan yang akan dilakukan mulai dari mengumpulkan data dan label menggunakan web scrapping, membuat fitur-fitur data, membuat dataset dari fitur-fitur tersebut, lalu dataset yang didapat dibagi menjadi 2 bagian data latih dan data testing. Lalu model tersebut dibandingkan dan dipilih mana yang terbaik, dan tahapan akhir yaitu implementasi model ke dalam web aplikasi.

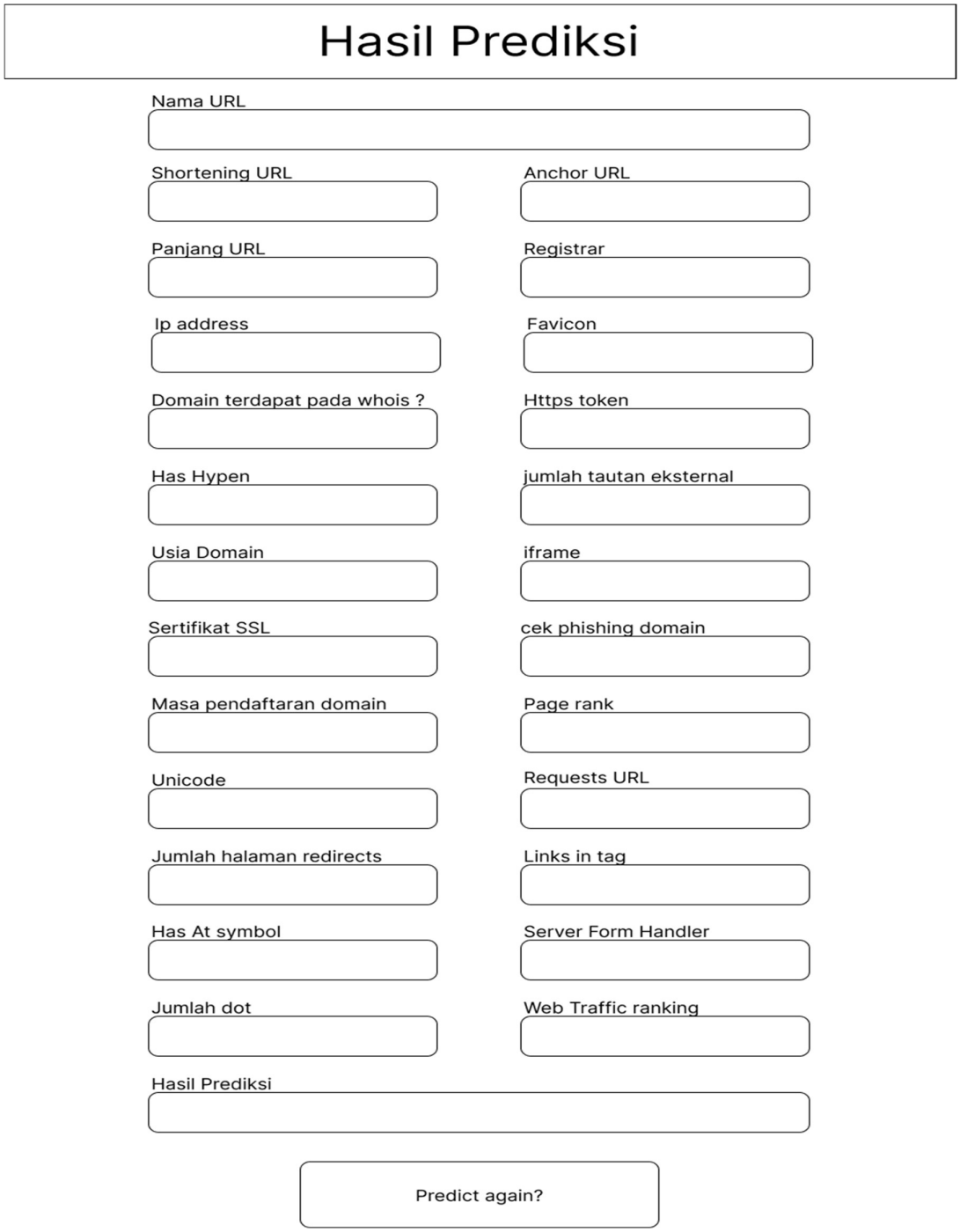
3.3.2 Desain wireframe halaman input



Gambar 3.3.2 Grafik desain wireframe input

Pada Gambar 3.3.2 merupakan kerangka tampilan website pada halaman utama yang digunakan untuk memprediksi *phishing* dengan cara memasukan url sebuah website ke dalam kolom Type to input URL.

3.3.3 Desain wireframe halaman output



Gambar 3.3.3 Gambaran desain wireframe output

Pada gambar 3.3.3 merupakan kerangka tampilan website output yang nantinya akan mengambil value dari fitur-fitur Nama URL, Shortening URL, Anchor URL, Panjang URL, Registrar, Ip address, Favicon, domain terdapat pada whois ?, Https token, Has Hypen, jumlah tautan eksternal, Usia Domain, iframe, Sertifikat SSL, cek phishing domain, Masa pendaftaran domain, Page rank,

Unicode, Requests URL, Jumlah halaman redirects, Links in tag, Has at symbol, Server form handler, Jumlah dot, Web traffic ranking, dan Hasil Prediksi untuk mendapatkan value apakah url tersebut phishing atau tidak.

3.4 Proses Preprocesing dan data cleaning

Untuk menjamin keakuratan dan keandalan model prediksi phishing yang dikembangkan dalam skripsi ini, proses preprocessing dan pembersihan data sangat penting. Tujuan dari prosedur ini adalah untuk memaksimalkan persiapan dataset sebelum digunakan dalam algoritma pembelajaran mesin, terutama Support Vector Machine (SVM) dan Random Forest.

Pertama-tama, data dikumpulkan dari situs web yang memiliki berbagai fitur phishing. Setelah itu, data dibersihkan melalui berbagai proses untuk menghilangkan suara dan outlier yang mungkin. Langkah ini mencakup pengenalan dan penanganan data yang tidak lengkap, duplikat, atau anomali lainnya yang dapat mempengaruhi kualitas hasil prediksi.

Untuk menyamakan distribusi variabel dan menyamakan skala, proses preprocessing juga mencakup proses normalisasi dan transformasi data. Proses ini sangat penting agar model pembelajaran mesin dapat memberikan hasil yang konsisten dan dapat diandalkan tanpa terpengaruh oleh perbedaan skala antar fitur.

Selain itu, untuk mengidentifikasi variabel yang paling penting dan relevan dalam memprediksi serangan phishing, seleksi fitur dilakukan. Ini membantu mengurangi kompleksitas model dan meningkatkan efisiensi pengolahan data.

Secara keseluruhan, proses pembersihan dan preprocessing data ini mendukung kesiapan dataset untuk digunakan dalam model pembelajaran mesin. Dengan memasukkan langkah-langkah ini, diharapkan hasil prediksi dari model SVM dan Random Forest menjadi lebih akurat, responsif terhadap perubahan, dan dapat diandalkan untuk menemukan serangan phishing yang mungkin.

3.4.1 Proses PreProcessing Dataset

Fitur-fiturnya :

- Shortening url mengecek apakah url tersebut mengalami penyusutan ini value nya diubah menjadi value -1 dan 1 jika mengalami penyusutan value nya 1 kalau tidak value nya 1
- Panjang url jika panjang dari url dari data yang didapat url yang panjang biasanya merupakan phising ini value nya diubah menjadi value -1,0 dan 1 jika kalau panjang url kurang dari 54

maka value nya = -1 kalau panjang url nya lebih dari atau sama dengan 54 dan kurang dari sama dengan 75 maka value nya menjadi 0 kalau melebihi 75 maka valuenya menjadi 1.

- Apakah url di kolom domain menggunakan ipaddress kalau iya maka value nya menjadi 1 dan kalau tidak valuenya menjadi -1.
- Apakah domain terdapat pada database whois jika iya maka value nya menjadi -1 jika tidak valuenya 1.
- Apakah domain mempunyai tanda – karena website asli jarang menggunakan tanda – pada domainnya jika iya menjadi 1 dan jika tidak menjadi -1.
- Usia Domain jika usia domain kurang dari atau sama dengan 365 maka value nya menjadi 1 kalau melebihi 365 maka valuenya -1.
- Mengecek sertifikat ssl apakah valid atau tidak jika usia dari https nya melebihi atau sama dengan 365 maka valuenya menjadi -1 dan jika kurang dari 365 maka valuenya menjadi 1.
- Lamanya masa pendaftaran sebuah domain jika lamanya masa pendaftarasebuah domainnya tersebut lebih dari atau sama dengan 180 maka valuenya menjadi -1 jika kurang dari 180 valuenya menjadi 1.
- Mengecek karakter url apakah menggunakan unicode jika iya maka value nya menjadi 1 dan jika tidak value nya -1.
- Mengecek Jumlah halaman redirect jika jumlah halaman redirect suatu url nya adalah kurang dari sama dengan 1 maka value nya dijadikan -1 jika lebih dari maka value nya dijadikan 1.
- Mengecek url mempunyai simbol @jika url menggunakan simbol @ maka value nya menjadi 1 dan jika tidak value nya menjadi -1.
- Jumlah dot pada domain berdasarkan sub domainnya karena dalam phishing sering menggunakan domain yang rumit dengan menggunakan banyak sub domain jika jumlah dot pada domain kurang dari sama dengan 1 maka value nya menjadi -1 jika jumlah dot pada domain melebihi 1 maka valuenya menjadi 1.
- Anchor url apakah url terhubung ke link lain dalam domain yang berbeda jika persentase external anchor url kurang dari 31 maka valuenya menjadi -1 kalau valuenya lebih dari sama dengan 31 dan kurang dari 67 valuenya adalah 0 dan jika persentase anchor nya melebihi sama dengan 67 valuenya adalah 1.
- Mengecek apakah registrar sudah terakreditasi oleh ICANN mengecek apakah registrar dari domainnya terdaftar di dataset ICANN atau tidak jika iya maka valuenya -1 jika tidak valuenya 1.

- Favicon untuk mendeteksi apakah favicon yang digunakan berasal pada internal domain atau berasal pada external domain mengecek apakah link favicon yang ada pada web sama dengan domain yang ada pada url nya jika iya valuenya -1 jika tidak valuenya 1.
- Mengecek HTTPS Token jika url memiliki kaat https pada domainya valuenya menjadi 1 jika tidak aluenya menjadi 1.
- Mengecek jumlah tautan eksternal yang mengarah pada halaman jika jumlah tautan eksternal nya jumlahnya lebih dari sama dengan 2 maka valuenya menjadi -1 jika tidak valuenya menjadi 1.
- Apakah url menggunakan Iframe jika iya valuenya menjadinya 1 dan jika tidak valuenya menjadi -1.
- Apakah domain dari url terdaftar pada data phishing domain jika iya valuenya menjadi 1 dan jika tidak valuenya menjadi -1.
- Mengambil pagerank pada api openpagerank dalam page rank nya jika pagerank nya didapat nilai kurang dari sama dengan 2 maka valuenya menjadi 1 kalau pagerank yang didapat lebih dari 2 maka valuenya menjadi -1.
- Requests URL jika persentase external request Url kurang dari 22 maka valuenya menjadi -1 jika persentasenya lebih dari sama dengan 22 dan kurang dari sama dengan 61 valuenya menjadi 0 dan kalau melebihi 61 valuenya menjadi 1.
- Link In Tags jika persentase external Links in Tags kurang dari 17 maka valuenya menjadi -1 kalau persentasenya lebih dari sama dengan 17 dan kurang dari sama dengan 81 maka valuenya menjadi 0, jika persentasenya melebihi 81 maka valuenya menjadi 1.
- Server Form Handler. Jika server form handler berisi about blank, ataupun kosong maka valuenya menjadi 1, kalau sfh nya merujuk pada domain yang berbeda maka valuenya menjadi 0 dan jika kalau bukan tadi valuenya menjadi -1.
- Web traffic ranking. Jika web traffic ranking yang didapat lebih dari sama dengan 200000 maka valuenya menjadi 1 kalau kurang dari 200000 maka valuenya menajdi -1.

Untuk dataset gabungan atau mix(categorical atau biner + numeric) data nya sama dengan yang biner tetapi untuk fitur calculate_url_length, Usia Domain url, Lamanya Masa pendaftaran sebuah domain, Jumlah halaman redirects, dot pada domain, anchor url, jumlah external link,page rank, Requests URL, links tag, Web Traffic. Untuk fitur-fitur tersebut digunakan datasetnya semestinya.

Pada dataset juga dilakukan feature selection untuk mengeathui fitur apa yang paling berdampak pada dataset untuk memprediksi apakah website tersebut merupakan phishing atau tidak.

Dataset yang akan dipakai akan dibedakan menjadi 2 yaitu training data dan data testing nanti akan dibuat menjadi 80 percent data training dan 20 percent data testing untuk mengecek keakuratan algoritma tersebut. lalu nanti hasil yang didapat akan dibuat untuk menghasilkan precision, recall, accuracy, f-1 score dengan rumus matrix of confusion.

3.4.2 Proses Pengolahan Dataset

Dataset yang digunakan berasal dari dataset yang sudah ada pada Kaggle yang didapat dari (Akram, 2023) pada url https://www.kaggle.com/datasets/akrammostafavi/farah-phishing, Openphish dari https://openphish.com/feed.txt, dan Phishtank dari https://phishtank.org/. URL dan label dari dataset ini diambil. Setelah data diambil, proses preprocessing dataset dilakukan, seperti yang ditunjukkan pada 3.4.2, dan data akan diseleksi. Jika data tersebut ada yang kosong fiturnya setelah fitur-fitur dataset diambil, baris dari dataset tersebut dibuang karena data fiturnya tidak lengkap.

Lalu dataset yang sudah lengkap dan fitur-fitur nya tidak kosong akan dilakukan konversi nilai numeric menjadi biner untuk membuat dataset full categorical untuk beberapa fitur calculate_url_length, Usia Domain url, Lamanya Masa pendaftaran sebuah domain, Jumlah halaman redirects, dot pada domain, anchor url, jumlah external link,page rank, Requests URL, links tag, Web Traffic.

Tabel 3.4.2

Tabel dataset program yang dipakai

URL	https://tevaera.org/	https://index.mufgrsz.net	http://4k.com/gaming/	http://www.orafaq.com/
Label	1 (phishing)	1 (phishing)	0 (legitimate)	0 (legitimate)
calculate_url_length	20	26	21	22
apakah url terganti link nya setelah di requests	-1	1	1	-1
seperated_url_type	-1	-1	-1	-1

is_domain_registered	1	-1	-1	-1
has_hyphen_in_domain	-1	-1	-1	-1
Usia Domain url	0	364	125	183
Usia Sertifikat SSL	86	88	60	0
Lamanya Masa pendaftaran sebuah domain	3	1	10102	8582
URL Unicode	-1	-1	-1	-1
Jumlah halaman redirects	0	0	1	0
url menggunakan simbol @	-1	-1	-1	-1
dot pada domain	0	1	0	1
Percentage anchor url	83	0	0	11
check_registrar	1	-1	-1	-1
Favicon	-1	1	-1	-1
https token	-1	-1	-1	-1
jumlah link	14	0	0	13
iframes	1	-1	-1	-1
hasil pencarian	-1	-1	-1	-1
page rank	0	0	4.64	4.99
Percentage Requests URL	0	0	0	50
Percentage links tag	100	0	0	40
Server Form Handler	-1	-1	1	0
Web Traffic	2000000	2000000	206575	96532

Pada tabel diatas ini beberapa data yang ada pada dataset disini ada 2 varian data yang berlabel 1 phishing dan data yang berlabel 0 legitimate disini akan dibahas satu-satu mengenai fiturnya.

a. Fitur *Calculate_url_length* digunakan untuk menghitung panjang url redirect terakhir jika ada redirect; jika tidak, panjang url input digunakan.
b. apakah url terganti link nya setelah di *requests* fitur ini digunakan untuk mengecek apakah url mengalami redirect atau tidak jika iya valuenya menjadi 1 dan jika tidak value nya menjadi -1.
c. *seperated_url_type* digunakan untuk mengetahui apakah domain pada url menggunakan alamat IP. Jika itu benar, maka nilainya menjadi 1 dan jika tidak, nilainya menjadi -1.

d. *is_domain_registered* digunakan untuk menentukan apakah domain terdaftar pada whois. Jika terdaftar, maka nilainya menjadi -1 dan jika tidak, nilainya menjadi 1.
e. Fitur *has_hyphen_in_domain* menentukan apakah url menggunakan tanda "-" di dalam domain. Jika iya, maka nilainya menjadi 1 dan jika tidak, nilainya menjadi -1.
f. Usia Domain URL fitur ini digunakan untuk mengetahui berapa lama usia domain yang ada pada URL tersebut. Nilai dihitung dengan satuan hari.
g. Usia Sertifikat SSL fitur ini digunakan untuk mengetahui usia sertifikat SSL https dari nilai url dalam satuan hari.
h. Lamanya masa pendaftaran suatu domain fitur ini digunakan untuk mengetahui berapa lama URL telah ada. Nilai dihitung dengan satuan hari.
i. Url Unicode fitur ini digunakan untuk mengetahui apakah url yang dimasukan mempunyai karakter yang bersifat *unicode*.
j. Jumlah halaman *redirects* fitur ini digunakan untuk menghitung berapa kali website tersebut mengalami *redirects*.
k. Url menggunakan simbol @ fitur ini digunakan untuk mengecek apakah url tersebut memiliki karakter @ pada urlnya jika itu benar, maka nilainya menjadi 1 dan jika tidak, nilainya menjadi -1.
l. Dot pada domain fitur ini digunakan untuk mengecek berapa jumlah dot yang ada pada domain suatu url.
m. *Percentage anchor url* fitur ini digunakan untuk mengetahui berapa persentase *url external* (yang beda dengan domain url) dibagi dengan total *anchor* yang didapat berdasarkan *tag a*.
n. *check_registrar* fitur ini digunakan untuk mengetahui apakah registrar suatu domain terdaftar dalam *ICANN* jika iya, maka nilainya menjadi -1 dan jika tidak nilainya menjadi 1.
o. Favicon fitur ini digunakan untuk mengethaui apakah url tersebut memiliki *favicon* yang domainnya sama dengan url dari website tersebut jika iya maka nilainya menjadi -1 jika *favicon domain* nya berbeda dengan url dan favicon tidak bisa di ambil nilainya menjadi 1.
p. https token fitur ini digunakan untuk mengetahui apakah url tersebut memiliki unsur https yang tertulis pada domain nya jika iya, maka nilainya menjadi 1 jika tidak nilainya menjadi -1.
q. jumlah link fitur ini digunakan untuk menghitung seluruh link tag a yang ada pada webiste tersebut.

r. *iframes* fitur ini digunakan untuk mencari seluruh elemen *iframe* yang ada pada website tersebut jika webiste memiliki *tag iframe*, maka nilainya menjadi 1 dan jika tidak nilainya menjadi -1.

s. hasil pencarian fitur ini digunakan untuk mencocokan nama domain website tersebut apakah rawan dijadikan sebagai phishing dengan cara mencocokan nama domain tersebut dengan *list phishing-domain-database* jika domain tidak terdaftar pada *list phishing-domain-database* dalam fitur ini dimanfaatkan Phishing database yang disusun oleh Mitchell (2023) dan dapat diakses melalui Github (https:// github.com/mitchellkrogza/Phishing.Database). maka nilainya menjadi -1 dan kalau jika terdaftar maka nilainya menjadi 1.

t. *page rank* fitur ini digunakan untuk menghitung skor domain dari url menggunakan bantuan *api open page rank*.

u. *Percentage Requests URL* fitur ini digunakan untuk menghitung persentase eksternal dibagi oleh jumlah link berdasarkan *tag img, video, audio, source*.

v. *Percentage links tag* fitur ini digunakan untuk menghitung persentase eksternal dibagi oleh jumlah link berdasarkan *tag meta, script, link.*

w. *Server Form Handler* fitur ini digunakan untuk mengecek apakah website memiliki *tag form* jika ada *form* akan di cek apakah action pada form itu sama dengan domain url atau tidak jika iya atau tidak ditemukan *tag form* maka nilainya menjadi -1 jika tidak sama dengan domain nilainya menjadi 0 , dan jika action nya kosong nilainya menjadi 1.

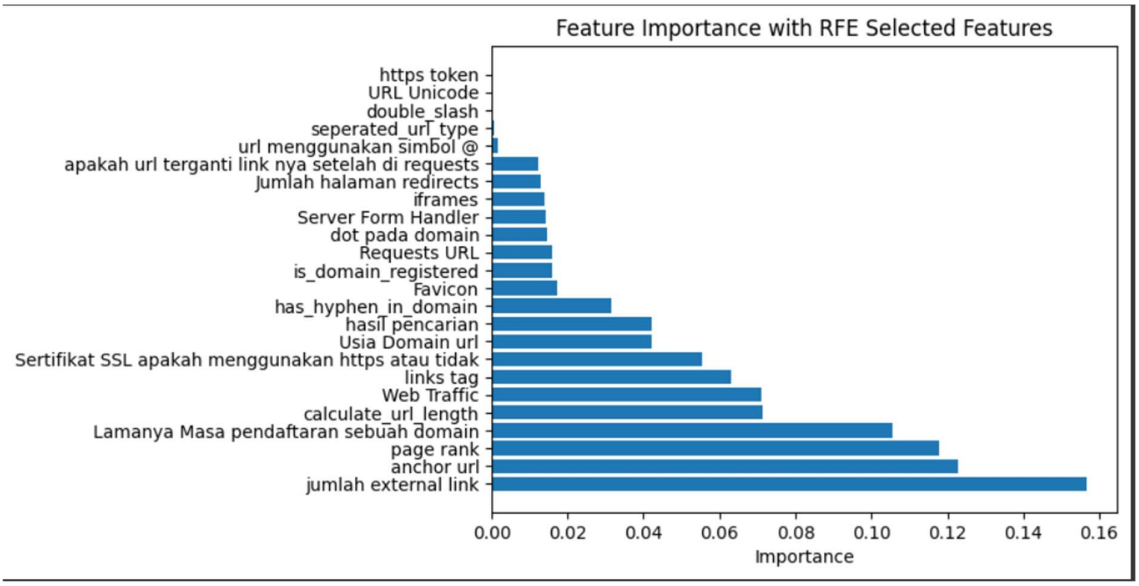
x. *Web Traffic* fitur ini digunakan untuk berapa peringkat domain url tersebut dibantu dengan api simillarweb.

3.4.3 Pembersihan Data

Karena web phishing seringkali menghapus isi URL, data fitur yang diambil menjadi kosong. Oleh karena itu, dataset dengan fitur yang tidak memiliki nilai akan dihapus barisnya supaya tidak mengganggu performa keakuratan model yang akan dilakukan.

3.4.4 Pemilihan Fitur

Dikarenakan fitur yang di test di dapet fitur *importance* nya seperti ini dan dilihat https token dan Url Unicode dinilai kurang memberi dampak terhadap prediksi maka fitur *Https token* dan *Url unicode* dihapus.



Gambar 3.4.4 Feature importance menggunakan RFE

Pada gambar 3.4.4 nilai importance dari fitur-fitur dataset menggunakan RFE selected features ini memberikan pandagan tentang faktor-faktor yang paling berpengaruh dalam model prediksi.

3.4.5 Pemecahan data

Dataset yang diapkai dipecah menjadi 2 bagian 20 persen untuk dijadikan dataset pengujian dan 80 persen untuk dijadikan dataset pelatihan.

3.4.6 Proses Training dan Testing

Dataset yang tadi sudah dilakukan data cleaning dan preprocessing lalu akan dicoba untuk mendapatkan nilai akurasinya dari yang dataset biner dan juga dataset campuran tadi menggunakan algoritma svm dan random forest lalu dibandingkan dan juga menggunakan *k fold cross validation* supaya bisa mendapatkan nilai rata-ratanya akurasi dari modelnya kemudian dilakukan pengujian testing untuk mendapatkan confusion matriksnya.