AICOMS
Applied Information Technology and Computer Science
e-ISSN: 2964-7703
Vol. 4 No. 2, Desember-2025, pp. 18-27

Analisis Komparatif Support Vector Machine dan Random Forest untuk Deteksi Email Phishing

Indah Purnama Sari [1*], Oris Krianto Sulaiman [2], Dicky Apdilah [3], Pastima Simanjuntak [4]

[1] Universitas Muhammadiyah Sumatera Utara ; indahpurnama@umsu.ac.id
[2] Universitas Islam Negeri Ar-Raniry Banda Aceh ; oris.ks@ar-raniry.ac.id
[3] Universitas Asahan; dickyapdi1404@gmail.com
[4] Universitas Putera Batam ; P.lastria@gmail.com
* Korespondensi: indahpurnama@umsu.ac.id

Info Artikel:

Dikirim: 22 Mei 2025

Direvisi: 31 Mei 2025

Diterima: 07 September 2025

Abstrack: Information and communication technology has rapidly advanced, bringing significant changes to daily life. With these advancements, access to information has become faster and easier; however, this convenience also introduces challenges, particularly concerning personal data security. One common cybercrime is email phishing, where attackers use malicious links to encrypt user data or devices and demand a ransom to restore access. Phishing emails often resemble official messages from trusted sources, making recipients unaware of the potential threat. To minimize such risks, technology can be utilized to automatically classify phishing emails. This study focuses on developing a machine learning model for automatic phishing email classification. The dataset used consists of 18,650 emails, including 11,322 non-phishing and 7,328 phishing emails. The proposed models employ two algorithms: Support Vector Machine (SVM) and Random Forest. To optimize performance, hyperparameter tuning was conducted using GridSearchCV. The experimental results demonstrate that the SVM algorithm achieved an accuracy of 97.27%, while the Random Forest algorithm achieved 96.51%. These findings indicate that the developed models can effectively support efforts to anticipate and mitigate phishing email threats..

Keywords : Email Phishing; Machine Learning; Support Vector Machine; Random Forest

Intisari: Teknologi informasi dan komunikasi kini telah berkembang dengan sangat pesat, membawa perubahan signifikan dalam kehidupan sehari-hari kita. Dengan semakin majunya teknologi informasi dan komunikasi, akses terhadap informasi menjadi sangat mudah dan cepat. Namun, kemudahan ini juga membawa tantangan tersendiri, terutama dalam hal keamanan data pribadi. Sebagai pengguna teknologi, kita dituntut untuk bijak dan waspada dalam menjaga data pribadi kita agar tidak disalahgunakan oleh pihak yang tidak bertanggung jawab. Salah satu contoh kejahatan siber yang sering terjadi adalah email phishing. Dalam serangan ini, pelaku menggunakan tautan berisi virus untuk mengenkripsi data atau perangkat pengguna, kemudian meminta tebusan untuk mengembalikan akses data tersebut. Phishing email biasanya tampak seperti email resmi dari sumber tepercaya, sehingga sering kali penerima tidak menyadari bahaya yang mengintai. Oleh karena itu, untuk meminimalisir kerugian yang dapat terjadi, kita juga dapat memanfaatkan teknologi sehingga dapat melakukan proses klasifikasi email phishing secara otomatis. Oleh karena itu, pada penelitian ini akan melakukan proses Pembangunan model machine learning yang Dimana dapat melakukan proses klasifikasi email phishing secara otomatis. Sehingga dengan adanya model yang dibangun pada penelitian ini, diharapkan dapat membantu dalam mengantisipasi terkena email phishing. Pada penelitian ini, Pembangunan model machine learning akan menggunakan data dengan total sebanyak 18650 data yang dimana terdiri dari 11322 data email tidak phishing dan sebanyak 7328 data email phising. Model yang akan dibangun pada penelitian ini yaitu model dengan menggunakan algoritma Support Vector Machine dan Random

Forest. Dalam proses Pembangunan model, untuk menemukan parameter yang optimal dilakukan proses hyperparameter tuning dengan menggunakan gridsearch CV, sehingga dapat menghasilkan parameter yang optimal. Setelah dilakukan proses pengujian model untuk melakukan proses klasifikasi email phishing, didapatkan hasil bahwa dengan menggunakan algoritma Support Vector Machine menghasilkan akurasi pengujian sebesar 97.27%, sedangkan dengan menggunakan algoritma Random Forest menghasilkan akurasi sebesar 96.51%.

Kata Kunci: Email Phishing; Machine Learning; Support Vector Machine

1. Pendahuluan

Perkembangan teknologi informasi dan komunikasi telah membawa perubahan besar dalam berbagai aspek kehidupan, termasuk dalam cara individu dan organisasi berinteraksi serta bertukar informasi. Salah satu sarana komunikasi yang paling umum digunakan saat ini adalah surat elektronik (email). Namun, seiring dengan meningkatnya penggunaan email, ancaman keamanan siber seperti phishing juga semakin marak terjadi. Phishing merupakan upaya penipuan yang dilakukan oleh pihak tidak bertanggung jawab untuk memperoleh informasi sensitif seperti kata sandi, nomor kartu kredit, atau data pribadi lainnya melalui penyamaran sebagai entitas terpercaya.

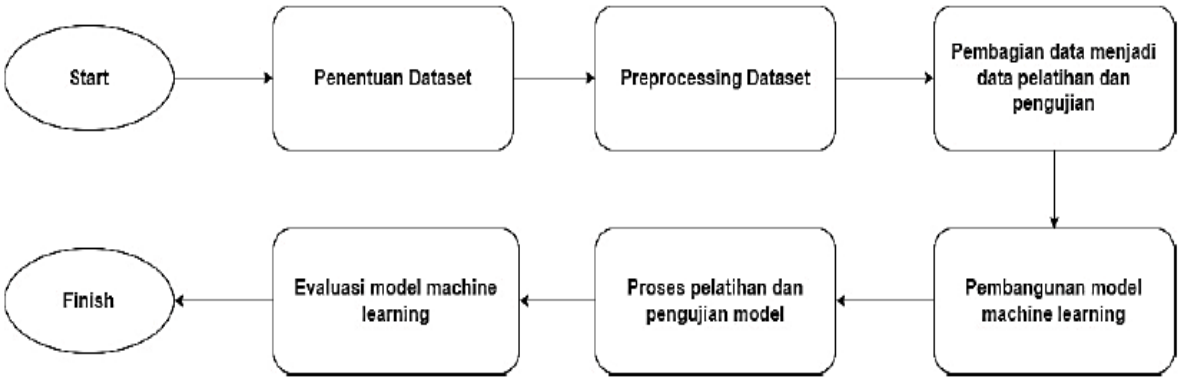
Serangan phishing tidak hanya merugikan individu, tetapi juga dapat berdampak besar terhadap institusi, baik dari segi kerugian finansial maupun reputasi. Oleh karena itu, deteksi dini terhadap email phishing menjadi hal yang sangat penting untuk mencegah dampak yang lebih luas. Dalam beberapa tahun terakhir, pendekatan berbasis machine learning telah banyak digunakan dalam mendeteksi email phishing karena kemampuannya dalam mengklasifikasikan data berdasarkan pola-pola tertentu.

Dua metode klasifikasi yang populer dalam bidang machine learning adalah Support Vector Machine (SVM) dan Random Forest (RF). SVM dikenal efektif dalam menangani data berdimensi tinggi dan bekerja dengan baik pada masalah klasifikasi biner, sementara RF memiliki keunggulan dalam menangani data yang kompleks dan mengurangi risiko overfitting melalui pendekatan ensemble. Meskipun keduanya banyak digunakan dalam deteksi phishing, kinerja masing-masing metode dapat bervariasi tergantung pada karakteristik dataset dan parameter yang digunakan. Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk melakukan analisis komparatif antara algoritma Support Vector Machine dan Random Forest dalam mendeteksi email phishing. Penelitian ini diharapkan dapat memberikan gambaran mengenai kelebihan dan kekurangan masing-masing metode serta membantu dalam memilih pendekatan yang paling efektif untuk sistem deteksi phishing berbasis machine learning.

2. Metode Penelitian

2.1. Alur Penelitian

Dalam proses penelitian yang dilakukan yaitu membangun model klasifikasi email phishing dengan menggunakan algoritma Support Vector Machine dan Random Forest, maka diperlukan tahapan tahapan sehingga model yang dibangun merupakan model yang prediktif dan dapat melakukan proses prediksi dengan akurat. Untuk alur yang dilakukan dalam penelitian ini dalam membangun model klasifikasi diberikan pada gambar 1.



Gambar 1. Alur Proses Penelitian

Gambar 1 menunjukkan alur proses yang dilakukan dalam membangun model klasifikasi email pishing. Untuk penjelasan tahap alur penelitian diberikan dibawah ini.

1. Hal pertama yang dilakukan yaitu menyiapkan dataset yang akan digunakan dalam penelitian. Dataset yang digunakan pada penelitian ini merupakan dataset yang didapatkan dari website kagge.com yang berjudul email phishing dataset, yang Dimana terdiri dari 11.650 data email dengan 10322 data email safe dan 1328 data email phishing.
2. Langkah selanjutnya yang dilakukan yaitu melakukan pra pemrosesan dataset. Hal hal yang dilakukan pada tahapan ini yaitu pertama, menghilangkan stop words atau kata kata yang kurang berpengaruh dalam proses pembentukan suatu kalimat, sehingga dengan dilakukannya penghilangan stop words ini, akan mempercepat proses pelatihan sehingga dapat efisien. Selanjutnya, setelah menghilangkan stop words, maka akan dilakukan konversi data dari teks menjadi bentuk vector dengan menggunakan TF-IDF Vectorizer. Hal ini dilakukan sehingga model dapat melakukan pemahaman dan prediksi dengan baik karena algoritma machine learning memerlukan data dalam bentuk numerik untuk melakukan perhitungan. Selain itu, dengan mengolah data berbentuk vector maka menciptakan representasi yang konsisten dan seragam, yang penting untuk memastikan bahwa algoritma dapat membandingkan dan menganalisis data secara efektif.
3. Lalu setelah melakukan pra pemrosesan pada data, maka akan dilakukan proses pembagian dataset menjadi data pelatihan dan data pengujian. Pada penelitian ini, data akan dibagi dengan persentase 70% data pelatihan dan 30% data pengujian. Data pelatihan berguna untuk model dapat mempelajari pola dari data sehingga dapat memiliki knowledge unuk melakukan proses klasifikasi. Sedangkan data pengujian berguna untuk menguji performa model yang sebelumnya sudah dilatih untuk dapat melakukan prediksi dengan baik sehingga dapat dilakukan analisis performa model.
4. Selanjutnya, akan melakukan proses pembagunan machine learning. Pada penelitian ini, proses Pembangunan machine learning akan menggunakan 2 algoritma yaitu algoritman Support Vector Machine dan Random Forest.
5. Setelah melakukan pemrosesan data dan Pembangunan model, maka selanjutnya dapat melakukan proses pelatihan dan pengujian pada model yang telah dibangun.
6. Lalu, hasil dari proses pengujian tersebut dapat dilakukan analisis dengan menggunakan confusion matrix sehingga dapat diambil kesimpulan performa dari masing masing model dengan algoritma yang dibangun untuk dapat mana model yang memiliki performa paling optimal.

2.2. Prosedur Pengambilan Data

Untuk dapat mendukung berlangsungnya penelitian ini sehingga dapat dibangun model yang dapat melakukan proses klasifikasi email phishing secara akurat dan prediktif, maka diperlukan sebuah data yang akan digunakan algoritma machine learning untuk dapat belajar dan memahami pola yang ada dari data tersebut. Dalam penelitian ini, data yang akan digunakan sebagai data utama dalam proses Pembangunan model baik untuk melatih model mengenali pola dari data ataupun untuk menguji performa model dalam melakukan prediksi data sehingga dapat dilakukan analisis performa merupakan data public yang didapatkan dari website kaggle.com. Data tersebut merupakan data yang berisi banyak email dengan label email tersebut merupakan email phishing atau email yang aman. Yang Dimana pada dataset tersebut terdiri dari total 18.650 data yang Dimana terdiri dari 11.322 data email safe dan 7328 data email phishing. Sehingga dengan menggunakan data tersebut diharapkan dapat membangun model yang lebih akurat dan prediktif dalam melakukan klasifikasi email phishing sehingga dapat membantu mengurangi kerugian yang terjadi akibat adanya kejadian email phishing.

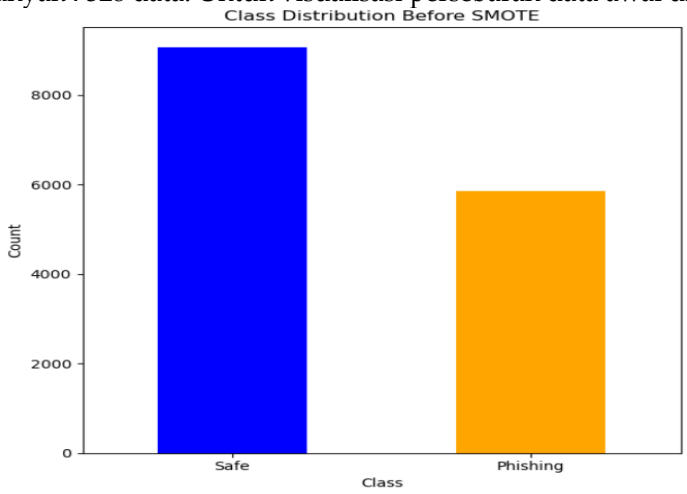
3. Pembahasan

3.1. Analisis Data

Dalam penelitian ini, data yang digunakan merupakan data yang didapatkan dari website kaggle.com dengan tautan https://www.kaggle.com/datasets/subhajournal/phishingemails. Yang Dimana data tersebut terdiri dari total 18.650 data email dengan persebaran sebanyak 11322 data sebagai data email safe dan 7328 data email phishing.

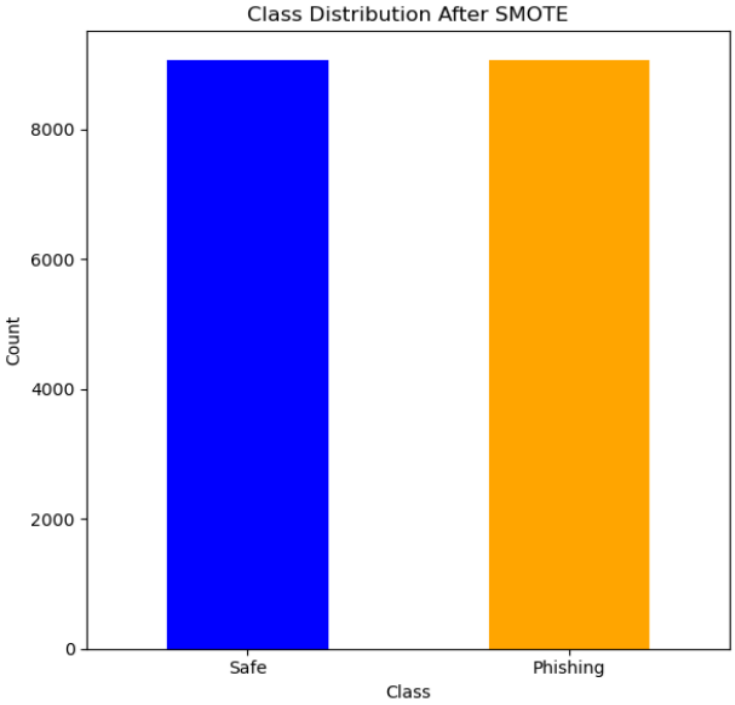
Setelah dilakukan proses encoding data, untuk dapat memudahkan model machine learning dalam menganalisis dan memprediksi pola yang ada pada dataset, maka selanjutnya dapat melakukan proses konversi data teks menjadi data berbentuk vector. Tujuan dilakukan proses konversi data teks menjadi vector yaitu agar algoritma yang digunakan dapat belajar data dengan baik. Dikarenakan, model machine learning hanya dapat belajar dari data yang berbentuk numerik atau vector. Selanjutnya, setelah melakukan proses konversi data menjadi bentuk vector, maka selanjutnya

dapat dilihat pada persebaran data yang diberikan, terjadi distribusi yang tidak merata, yang Dimana data pada kelas email safe lebih dominan dengan memiliki sebanyak 11322 data apabila dibandingkan dengan data pada kelas email phising yang hanya terdapat sebanyak 7328 data. Untuk visualisasi persebaran data awal diberikan pada gambar 2.



Gambar 2. Persebaran Dataset Awal

Gambar 3. menunjukkan perbedaan banyaknya jumlah data pada kelas safe dan phishing, yang Dimana ditunjukkan pada gambar 4.1, data pada kelas safe lebih dominan. Oleh karena itu, diperlukan suatu pemrosesan data lebih lanjut untuk dapat melakukan proses penyetaraan distribusi dari data. Pada penelitian ini, proses penyetaraan distribusi data akan menggunakan metode SMOTE (Synthetic Minority Oversampling Technique), yang dimana SMOTE merupakan metode yang sering kali dan optimal untuk digunakan dalam mengatasi masalah ketidakseimbangan kelas pada dataset. Setelah dilakukan proses pengolahan data dengan menggunakan SMOTE, maka dataset tersebut persebaran datanya menjadi seimbang yaitu pada kelas safe terdiri dari 11322 data dan pada kelas phishing terdiri dari 11322 data juga. Untuk visualisasi persebaran data setelah dilakukan proses SMOTE diberikan pada gambar 3.



Gambar 3. Visualisasi Dataset Hasil Smote

Gambar 3. menunjukkan hasil persebaran data setelah dilakukan proses SMOTE. Synthetic Minority Over-sampling Technique (SMOTE) adalah metode yang digunakan untuk mengatasi ketidakseimbangan kelas dalam dataset dengan membuat sampel sintetis dari data minoritas. Dapat dilihat pada gambar 4.2, data hasil SMOTE mendapatkan jumlah yang merata pada setiap kelas. Dengan persebaran data yang lebih seimbang, model machine

learning dapat belajar dari representasi yang lebih baik dari masing-masing kelas, sehingga mengurangi bias terhadap kelas mayoritas. Hal ini diharapkan dapat membantu model untuk lebih memahami pola dari data dengan baik dan memberikan hasil pengujian yang lebih optimal.

3.2. Analisis Parameter Pengujian

Setelah dilakukan proses preprocessing data berupa label encoding dan SMOTE, serta proses ekstraksi fitur dari data dengan menggunakan TF-IDF Vectorizer, maka selanjutnya dapat melakukan proses pengembangan model klasifikasi atau prediksi email phishing untuk mengolah dataset yang telah diproses. Pada penelitian ini, proses pelatihan dan pengujian akan menggunakan model yang dibangun menggunakan algoritma Random Forest dan Support Vector Machine. Algoritma Support Vector Machine (SVM) dipilih karena terkenal dengan kemampuannya untuk bekerja dengan baik pada ruang fitur tinggi yang dihasilkan oleh TF-IDF Vectorizer. SVM efektif dalam menemukan hyperplane optimal yang memaksimalkan margin antara kelas-kelas, sehingga sangat cocok untuk masalah klasifikasi biner seperti identifikasi email phishing. Algoritma Random Forest (RF) dipilih karena kemampuannya yang baik dalam menangani dataset yang besar dan kompleks. Random Forest adalah ensemble learning method yang menggabungkan banyak pohon keputusan untuk meningkatkan akurasi dan mengurangi overfitting. Algoritma ini juga dikenal tangguh terhadap outlier dan dapat memberikan estimasi feature importance yang berguna untuk memahami faktor-faktor yang paling mempengaruhi klasifikasi email phishing. Sehingga, dengan menggunakan kedua algoritma ini, diharapkan model yang dihasilkan akan memiliki performa yang baik dalam mendeteksi email phishing dan memberikan hasil yang optimal. Dalam Pembangunan model SVM dan Random Forest, diperlukan parameter parameter optimal yang diatur agar model dapat bekerja secara optimal. Pada penelitian ini, akan melakukan beberapa pengujian parameter baik pada algoritma Support Vector Machine dan Random Forest.

4. Metode dan Hasil Penelitian

4.1. Analisis Performa Model

Setelah dilakukan pemrosesan dataset dan penentuan parameter pengujian yang akan digunakan, langkah berikutnya adalah melatih model. Proses pelatihan model ini bertujuan untuk melatih algoritma machine learning agar dapat mengenali dan mempelajari pola-pola yang terdapat dalam data. Dengan memahami pola-pola tersebut, model dapat lebih efektif dalam mendeteksi email phishing. Setelah model dilatih untuk mengenali pola dari data, langkah selanjutnya adalah menguji model untuk mengidentifikasi email phishing menggunakan model yang telah dilatih. Proses pengujian ini akan menghasilkan metrik performa model yang kemudian dianalisis untuk menilai efektivitas dan akurasi model dalam mendeteksi email phishing. Untuk hasil akurasi proses pengujian model diberikan pada tabel 1.

Tabel 1. Hasil Akurasi Pengujian

Model	Akurasi
Random Forest 1	96.51 %
Random Forest 2	96.25 %
Random Forest 3	96.49 %
Random Forest 4	96.27 %
Random Forest 5	96.41 %
Random Forest 6	96.25 %
Support Vector Machine 1	97.24 %
Support Vector Machine 2	97.24 %
Support Vector Machine 3	84.13 %
Support Vector Machine 4	84.13 %
Support Vector Machine 5	97.27 %
Support Vector Machine 6	97.18 %

Tabel 1. menunjukkan hasil akurasi yang didapatkan setelah melakukan proses pengujian model. Hasil dari tabel menunjukkan performa akurasi yang bervariasi di antara model Random Forest dan Support Vector Machine (SVM) yang telah diuji. Model Random Forest menunjukkan akurasi antara 96.25% hingga 96.51%. Model terbaik adalah

Random Forest 1 dengan akurasi 96.51%, yang menggunakan 500 pohon keputusan dengan kriteria entropy. Random Forest 3 dan 5 juga menunjukkan akurasi yang tinggi dengan 96.49% dan 96.41% berturut-turut, menggunakan 400 dan 300 pohon keputusan dengan kriteria entropy. Di sisi lain, model SVM menunjukkan variasi akurasi yang lebih besar. SVM dengan kernel linear (Support Vector Machine 1 dan 2) mencapai akurasi 97.24%, yang merupakan yang tertinggi di antara semua model yang diuji. Namun, model SVM dengan kernel polynomial (Support Vector Machine 3 dan 4) menunjukkan akurasi yang lebih rendah sekitar 84.13%. SVM dengan kernel rbf (Support Vector Machine 5 dan 6) memiliki akurasi antara 97.18% dan 97.27%, yang juga cukup tinggi. Analisis ini menunjukkan bahwa SVM dengan kernel linear dan kernel rbf cenderung memberikan performa yang lebih baik dalam kasus identifikasi email phishing dibandingkan dengan kernel polynomial. Di sisi lain, Random Forest menunjukkan stabilitas yang baik dengan variasi akurasi yang lebih konsisten di sekitar nilai 96%. Setelah mendapatkan hasil akurasi pengujian model dengan menggunakan model yang dibangun, proses analisis juga dilakukan dengan menggunakan perhitungan metrik performa yaitu confusion matrix, sehingga dengan confusion matrix ini diharapkan dapat lebih memberikan Gambaran mengenai kinerja model dalam melakukan proses identifikasi email phishing. Untuk hasil confusion matrix hasil pengujian model model yang telah dibangun diberikan pada tabel 2. berikut.

Tabel 2. Hasil Presisi, Recall, dan F1 Sccore Pengujian

Model	Presisi	Recall	F1-Sccore
Random Forest 1	97 %	97 %	97 %
Random Forest 2	96 %	96 %	96 %
Random Forest 3	97 %	96 %	96 %
Random Forest 4	96 %	96 %	96 %
Random Forest 5	96 %	96 %	96 %
Random Forest 6	96 %	96 %	96 %
Support Vector Machine 1	97 %	97 %	97 %
Support Vector Machine 2	97 %	97 %	97 %
Support Vector Machine 3	86 %	84 %	83 %
Support Vector Machine 4	86 %	84 %	83 %
Support Vector Machine 5	97 %	97 %	97 %
Support Vector Machine 6	97 %	97 %	97 %

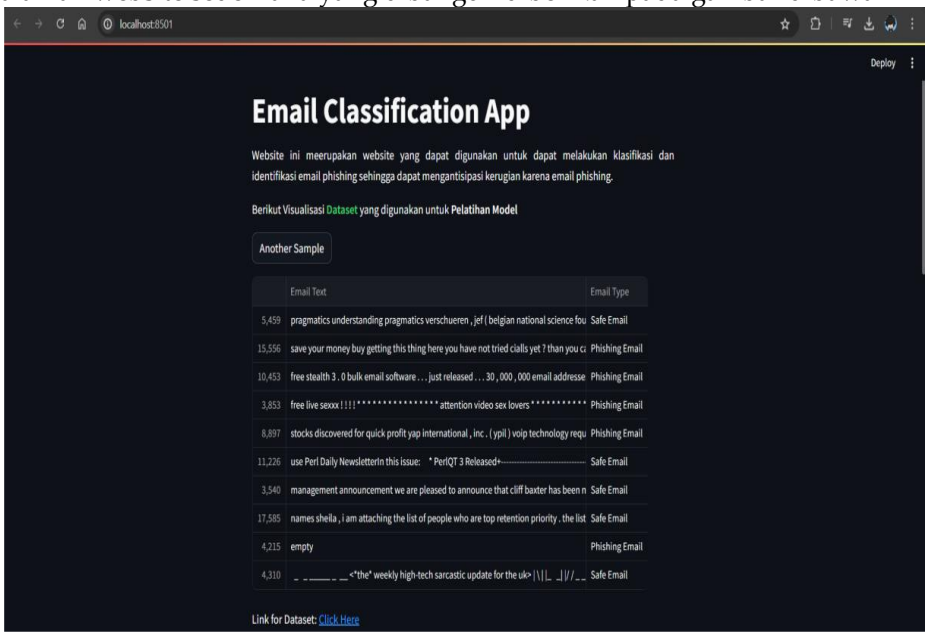
Tabel 2. menunjukkan hasil presisi, recall dan juga f1-score yang didapatkan selama proses pengujian model Support Vector Machine dan Random Forest yang dibangun. Dapat dilihat pada tabel 4.6, menunjukkan performa yang detail dalam metrik presisi, recall, dan F1-score untuk model Random Forest dan Support Vector Machine (SVM). Model Random Forest menunjukkan stabilitas yang konsisten dengan nilai presisi, recall, dan F1-score yang tinggi, antara 96% hingga 97% untuk semua konfigurasi. Random Forest 1 mencapai performa tertinggi dengan presisi 97%, recall 97%, dan F1-score 97%, menggunakan 500 pohon keputusan dan kriteria entropy.

Model ini menonjol dalam kemampuannya untuk mengklasifikasikan email phishing dengan akurasi yang sangat baik. Di sisi lain, SVM juga menunjukkan hasil yang kuat dalam beberapa konfigurasi. SVM 1 dan SVM 2, dengan kernel linear, mencapai presisi, recall, dan F1-score sekitar 97%. Hal yang sama juga terjadi pada SVM 5 dan SVM 6 dengan kernel rbf, yang menunjukkan performa yang konsisten dengan nilai presisi, recall, dan F1-score sekitar 97%. Namun, SVM dengan kernel polynomial (SVM 3 dan SVM 4) menunjukkan penurunan dalam performa dengan presisi sekitar 86%, recall 84%, dan F1-score 83%. Hal ini menunjukkan bahwa kernel polynomial mungkin kurang cocok untuk dataset ini yang mungkin memerlukan pengenalan pola yang lebih kompleks.

Analisis ini menggarisbawahi pentingnya memilih model yang sesuai dengan karakteristik data dan tujuan aplikasi. Meskipun SVM dapat memberikan performa yang tinggi dalam beberapa kasus, terutama dengan kernel linear dan rbf, pemilihan yang tidak tepat dari kernel seperti polynomial dapat mengurangi performa model secara signifikan. Sebaliknya, Random Forest menunjukkan stabilitas dan konsistensi yang baik dalam mengklasifikasikan email phishing dengan tingkat akurasi yang tinggi. Dengan demikian, pemilihan model terbaik harus mempertimbangkan tidak hanya akurasi tetapi juga interpretasi model, biaya komputasi, dan kebutuhan spesifik aplikasi untuk memastikan bahwa model yang dipilih dapat memberikan hasil yang optimal dan dapat diandalkan.

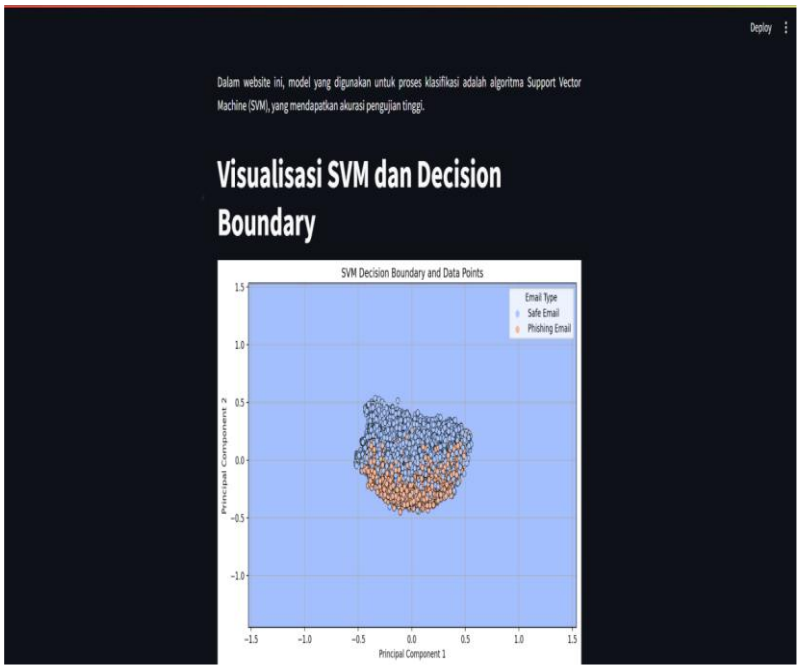
4.2. Implementasi Sistem

Setelah melalui proses pelatihan dan pengujian model machine learning, langkah selanjutnya adalah melakukan implementasi sistem menggunakan model yang telah terbukti memiliki akurasi tertinggi pada sebuah platform web sederhana. Platform ini dirancang untuk secara otomatis mengidentifikasi email phishing berdasarkan teks yang dimasukkan pengguna. Pengembangan web ini menggunakan Streamlit, sebuah library Python yang memungkinkan pembuatan antarmuka pengguna interaktif dengan mudah tanpa memerlukan keahlian mendalam dalam pengembangan web. Penggunaan Streamlit dalam penelitian ini bertujuan untuk mempermudah implementasi model machine learning yang telah dilatih untuk deteksi email phishing ke dalam sebuah antarmuka yang ramah pengguna. Melalui aplikasi web ini, pengguna dapat memasukkan teks email untuk diuji apakah termasuk phishing atau tidak, serta melihat hasil prediksi secara langsung. Streamlit tidak hanya menyediakan kemudahan dalam integrasi model, tetapi juga memungkinkan visualisasi yang intuitif dari hasil prediksi, meningkatkan interaksi dan pengalaman pengguna dalam proses identifikasi. Dengan demikian, aplikasi ini tidak hanya efektif dalam memproses data dengan akurasi tinggi, tetapi juga mudah digunakan oleh pengguna akhir tanpa kompleksitas teknis yang berlebihan. Untuk Visualisasi tampilan halaman website sederhana yang dibangun diberikan pada gambar dibawah ini.



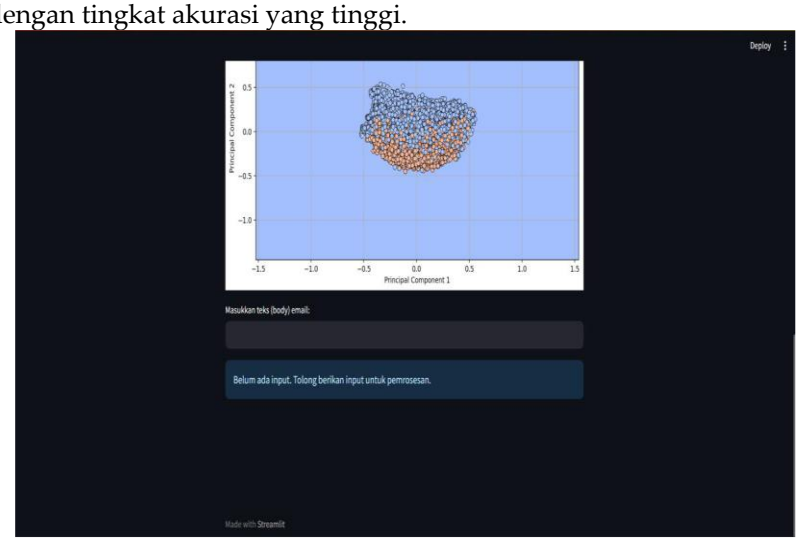
Gambar 4. Tampilan Web 1

Gambar 4. menunjukkan antarmuka dari aplikasi klasifikasi email yang dikembangkan menggunakan Streamlit. Aplikasi ini dirancang untuk mengklasifikasikan dan mengidentifikasi email phishing, sehingga pengguna dapat mengantisipasi kerugian akibat email phishing. Pada antarmuka tersebut, terlihat judul "Email Classification App" di bagian atas yang menjelaskan tujuan dari aplikasi ini. Di bawahnya terdapat deskripsi singkat dalam bahasa Indonesia yang menyatakan bahwa website ini dapat digunakan untuk klasifikasi dan identifikasi email phishing. Bagian utama dari tampilan ini menunjukkan sebuah tabel yang memuat contoh data yang digunakan untuk melatih model klasifikasi. Tabel tersebut terdiri dari dua kolom, yaitu "Email Text" dan "Email Type". Kolom "Email Text" berisi cuplikan teks email, sementara kolom "Email Type" menunjukkan jenis email tersebut, apakah "Safe Email" atau "Phishing Email". Beberapa contoh email yang ditampilkan dalam tabel termasuk email yang berisi kata-kata yang mencurigakan, yang kemudian diklasifikasikan sebagai "Phishing Email", serta email yang terlihat aman dan diklasifikasikan sebagai "Safe Email". Sedangkan, Di bagian bawah tabel, terdapat tautan untuk dataset yang digunakan, yang dapat diakses pengguna untuk melihat data secara lebih lengkap.



Gambar 5. Tampilan Website 2

Gambar 5 merupakan bagian dari antarmuka aplikasi klasifikasi email phishing yang dijelaskan sebelumnya, yang menunjukkan hasil dari penerapan algoritma Support Vector Machine (SVM) untuk proses klasifikasi. Di bagian atas, terdapat deskripsi yang menyatakan bahwa model yang digunakan dalam website ini adalah SVM. Deskripsi tersebut juga mencantumkan kinerja model yang mencakup akurasi pengujian sebesar 97.27%, dengan nilai presisi sebesar 97%, recall sebesar 97%, dan f1-score sebesar 97%. Pada bagian ini juga, ditunjukkan illustrasi proses SVM. Ilustrasi ini menunjukkan bagaimana SVM bekerja dengan menampilkan beberapa elemen penting, seperti "Support Vectors", "Optimal Hyperplane", dan "Maximize Margin". Support vectors adalah titik data yang paling dekat dengan hyperplane dan memainkan peran penting dalam menentukan posisi hyperplane tersebut. Optimal hyperplane adalah garis yang memisahkan dua kelas data (dalam kasus ini, biru dan hijau) dengan margin maksimum. Margin adalah jarak antara hyperplane dan support vectors terdekat dari setiap kelas, dan tujuan dari SVM adalah untuk memaksimalkan margin ini agar menghasilkan pemisahan yang optimal antara kelas-kelas tersebut. Ilustrasi ini membantu menjelaskan bagaimana SVM dapat digunakan untuk memisahkan data dari dua kelas yang berbeda, yaitu email phishing dan email yang aman, dalam konteks aplikasi klasifikasi email phishing. Penjelasan ini memberikan wawasan mendalam tentang kinerja dan mekanisme dari algoritma yang digunakan, memperlihatkan bahwa model ini sangat efektif dalam mengklasifikasikan email dengan tingkat akurasi yang tinggi.



Gambar 6. Tampilan Website 3

Gambar 6 merupakan kelanjutan dari antarmuka aplikasi klasifikasi email phishing yang menampilkan proses klasifikasi menggunakan algoritma Support Vector Machine (SVM). Di bagian atas, terdapat bagian akhir dari ilustrasi

konsep SVM yang telah dijelaskan sebelumnya, yang diikuti dengan teks "Visualisasi SVM". Di bawahnya, terdapat judul "Classification Process Using SVM" yang menandakan bahwa ini adalah bagian interaktif dari aplikasi di mana pengguna dapat memasukkan teks email untuk diklasifikasikan. Terdapat sebuah kotak teks dengan label "Masukkan teks (body) email:" di mana pengguna dapat memasukkan isi email yang ingin mereka klasifikasikan. Di bawah kotak teks tersebut, terdapat pesan yang berbunyi "Belum ada input. Tolong berikan input untuk pemrosesan." yang menunjukkan bahwa belum ada teks yang dimasukkan untuk diproses. Bagian ini memungkinkan pengguna untuk menguji sendiri klasifikasi email phishing dengan memasukkan teks email dan melihat hasil klasifikasi yang diberikan oleh model SVM yang sudah dilatih. Hal ini menambah aspek interaktif dan fungsionalitas dari aplikasi, membuatnya lebih bermanfaat bagi pengguna yang ingin memeriksa email mereka terhadap potensi ancaman phishing.

5. Kesimpulan

Berdasarkan hasil analisis dan penelitian yang telah dilakukan, yang dimana berkaitan dengan perhitungan performa algoritma Support Vector Machine dan Random Forest dalam melakukan proses klasifikasi berdasarkan data yang diolah dengan menggunakan metode SMOTE untuk mengatasi persebaran data tidak merata serta ekstraksi fitur dengan menggunakan TF-IDF Vectorizer, mendapatkan Kesimpulan beberapa hal berikut ini: Algoritma SMOTE yang diimplementasikan dalam dataset berhasil untuk mengatasi ketidakseimbangan data, yang Dimana dapat dilihat pada persebaran data awalnya terdiri dari 11322 data safe dan 7328 data phishing, setelah dilakukan proses implementasi SMOTE, maka persebaran data menjadi 11322 data safe dan 11322 data phishing. Berdasarkan hasil yang didapatkan, algoritma Random Forest mendapatkan akurasi yang paling optimal dengan menggunakan nilai n_estimators sebesar 500 dan criterion yaitu entropy, mendapatkan akurasi sebesar 96.51%. Sedangkan, berdasarkan hasil pengujian dengan menggunakan algoritma Support Vector Machine dengan parameter nilai C sebesar 1.0 dan kernel yaitu rbf, mendapattkan hasil akurasi pengujian sebesar 97.27%. Berdasarkan hasil tersebut, dapat diketahui algoritma SVM dapat melakukan proses klasifikasi dan identifikasi yang paling optimal untuk email phishing. Model yang dibangun dengan memiliki nilai akurasi pengujian terbaik dapat diimplementasikan pada website sederhana dengan menggunakan streamlit, sehingga dapat digunakan secara efisien untuk pengguna dapat melakukan proses prediksi email phishing. Sehingga, dapat disimpulkan bahwa algoritma Random Forest dan Support Vector Machine yang dibangun dapat bekerja dengan baik dan optimal yang Dimana ditunjukkan pada hasil akurasi pengujian yang didapatkan rata rata 94.61%.

Daftar Pustaka

[1] Anggarda, M., Kustiawan, I., Nurjanah, D., & Hakim, N. (2023). Pengembangan Sistem Prediksi Waktu Penyiraman Optimal pada Perkebunan: Pendekatan Machine Learning untuk Peningkatan Produktivitas Pertanian. JURNAL BUDIDAYA PERTANIAN, 19(2), 124-136. https://doi.org/10.30598/jbdp.2023.19.2.124

[2] Erlangga, F., & Sari, I.P. (2024). Perancangan Sistem Untuk Merekomendasikan Produk Skincare Menggunakan Metode NLP. Portal Riset dan Inovasi Sistem Perangkat Lunak 2 (4), 1-11

[3] Avcı, C., Budak, M., Yağmur, N., Balçık, F. (2023). Comparison between random forest and support vector machine algorithms for LULC classification. International Journal of Engineering and Geosciences, 8(1), 1-10. https://doi.org/10.26833/ijeg.987605

[4] Azzahrah., A & Sari., I.P. (2024). Perbandingan Sistem Prediksi Menggunakan Metode Monte Carlo dengan Metode K-NN pada Nilai Peserta Didik Uji Kompetensi Kejuruan. sudo Jurnal Teknik Informatika 3 (3), 127-135

[5] Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4. 5, Random Forest, SVM dan Naive Bayes. Jurnal Media Informatika Budidarma, 5(2), 640-651. http://dx.doi.org/10.30865/mib.v5i2.2937

[6] Sari, I.P., & Batubara, I.H. (2021). Perancangan Sistem Informasi Laporan Keuangan Pada Apotek Menggunakan Algoritma K-NN. Seminar Nasional Teknologi Edukasi dan Humaniora (SiNTESa)

[7] Hasibuan., W.R, Sari.., I.P, & Basri., M. (2025). Klasifikasi Kerusakan (Cacat) pada Biji Kopi Arabika Menggunakan Algoritma KNN (K-Nearest Neighbor). Blend Sains Jurnal Teknik 3 (4), 452-459

[8] Sari, I.P., Al-Khowarizmi, A., & Batubara, I.H. (2021). Cluster Analysis Using K-Means Algorithm and Fuzzy C-Means Clustering For Grouping Students' Abilities In Online Learning Process. Journal of Computer Science, Information Technology and Telecommunication Engineering, 2(1), 139-144

[9] Apdilah, D., & Sari, I.P. (2021). Optimization Of The Fuzzy C-Means Cluster Center For Credit Data Grouping Using Genetic Algorithms. Al'adzkiya International of Computer Science and Information Technology (AIoCSIT) Journal, 2(2), 156-163

[10] Badillo, S., Banfai, B., Birzele, F., Davydov, I.I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B. and Zhang, J.D. (2020), An Introduction to Machine Learning. Clin. Pharmacol. Ther., 107: 871-885. https://doi.org/10.1002/cpt.1796

[11] Sari, I.P., Ramadhani, F., & Satria, A. (2024). Classification of Tuberculosis Based on Thorax X-ray Images Using Multi-Scale Convolutional Neural Network. 2024 7th International Conference of Computer and Informatics Engineering (IC2IE)

[12] CASUARINA, Indah Putri; HAYATI, Memi Nor; PRANGGA, Surya. (2022). Klasifikasi Status Pembayaran Kredit Barang Elektronik dan Furniture Menggunakan Support Vector Machine. EKSPONENSIAL, [S.l.], v. 13, n. 1, p. 71-78, june 2022. ISSN 2798-3455. Available at: <https://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/view/887>. Date accessed: 28 may 2024. doi: https://doi.org/10.30872/eksponensial.v13i1.887.

[13] Rolly Junius Lontaan Muhammad Fairuzabadi, Indah Purnama Sari Imam Ekowicaksono, Fatimah Nur Arifah Rahman Indra Kesuma, Nizirwan Anwar Andika Setiawan Deep Learning untuk Pemula: Memahami Algoritma, Tools, dan Masa Depan AI

[14] Chairunisa, G., Najib, M. K., Nurdiati, S., Imni, S. F., Sanjaya, W., Andriani, R. D., Henriyansah, Putri, R. S. P., & Ekaputri, D. (2024). Life Expectancy Prediction Using Decision Tree, Random Forest, Gradient Boosting, and XGBoost Regressions. JURNAL SINTAK, 2(2), 71–82. https://doi.org/10.62375/jsintak.v2i2.249

[15] Ernianti Hasibuan, & Elmo Allistair Heriyanto. (2022). ANALISIS SENTIMEN PADA ULASAN APLIKASI AMAZON SHOPPING DI GOOGLE PLAY STORE MENGGUNAKAN NAIVE BAYES CLASSIFIER. Jurnal Teknik Dan Science, 1(3), 13–24. https://doi.org/10.56127/jts.v1i3.434