ANALISIS PERBANDINGAN *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* UNTUK KLASIFIKASI *EMAIL PHISHING*

SKRIPSI

DISUSUN OLEH

NURKUMALA LUBIS

NPM. 2009020056

PROGRAM STUDI TEKNOLOGI INFORMASI

FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI

UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA

MEDAN

2024

ANALISIS PERBANDINGAN *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* UNTUK KLASIFIKASI *EMAIL PHISHING*

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom) dalam Program Studi Teknologi Informasi pada Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara

NURKUMALA LUBIS
NPM. ISI 2009020056

PROGRAM STUDI TEKNOLOGI INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
MEDAN
2024

LEMBAR PENGESAHAN

Judul Skripsi	:	**ANALISIS PERBANDINGAN *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* UNTUK KLASIFIKASI *EMAIL PHISHING***
Nama Mahasiswa	:	Nurkumala Lubis
NPM	:	2009020056
Program Studi	:	Teknologi Informasi

Menyetujui
Komisi Pembimbing

(Mulkan Azhari, S.Kom., M.Kom)
NIDN. 0108129402

Ketua Program Studi	**Dekan**
(Fatma Sari Hutagalung, S.Kom., M.Kom)	**(Dr. Al-Khowarizmi, S.Kom., M.Kom.)**
NIDN. 0117019301	**NIDN. 0127099201**

PERNYATAAN ORISINALITAS

ANALISIS PERBANDINGAN *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* UNTUK KLASIFIKASI *EMAIL PHISHING*

SKRIPSI

Saya menyatakan bahwa karya tulis ini adalah hasil karya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing disebutkan sumbernya.

Medan, 26 Agustus 2024

Yang membuat pernyataan

Nurkumala Lubis

NPM. 2009020056

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika Universitas Muhammadiyah Sumatera Utara, saya bertanda tangan dibawah ini:

Nama	: Nurkumala Lubis
NPM	: 2009020056
Program Studi	: Teknologi Informasi
Karya Ilmiah	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Muhammadiyah Sumatera Utara Hak Bedas Royalti Non-Eksekutif *(Non-Exclusive Royalty free Right)* atas penelitian skripsi saya yang berjudul:

ANALISIS PERBANDINGAN SVM DAN RANDOM FOREST UNTUK KLASIFIKASI EMAIL PHISHING

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non-Eksekutif ini, Universitas Muhammadiyah Sumatera Utara berhak menyimpan, mengalih media, memformat, mengelola dalam bentuk database, merawat dan mempublikasikan Skripsi saya ini tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemegang dan atau sebagai pemilik hak cipta.

Demikian pernyataan ini dibuat dengan sebenarnya.

Medan, 26 Agustus 2024

Yang membuat pernyataan

Nurkumala Lubis

NPM. 2009020056

RIWAYAT HIDUP

DATA PRIBADI

Nama Lengkap	: Nurumala Lubis
Tempat dan Tanggal Lahir	: Lubuk Pakam ,03-01-2001
Alamat Rumah	: Jln.Bajak V gg Rukun VI
Telepon/Faks/HP	085270607294
E-mail	: nkumala679@gmail.com
Instansi Tempat Kerja	:
Alamat Kantor	:

DATA PENDIDIKAN

SD	: SD Negeri 153071 Sibabangun 1 Kabupaten Tapanuli Tengah	TAMAT: 2013
SMP	: Madrasah Tsanawiyah Al-Mukhlishin Lumut	TAMAT: 2016
SMA	: SMA Negeri 1 Pinang Sori Kabupaten Tapanuli Tengah	TAMAT: 2019

KATA PENGANTAR

Penulis tentunya berterima kasih kepada berbagai pihak dalam dukungan serta doa dalam penyelesaian skripsi. Penulis juga mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Agussani. M.AP selaku Rektor Universitas Muhammadiyah Sumatera Utara.
2. Bapak Prof. Dr. Muhammad Arifin, S.H., M.Hum selaku Wakil Rektor I Universitas Muhammadiyah Sumatera Utara.
3. Bapak Prof. Dr. Akrim, M.Pd selaku Wakil Rektor II Universitas Muhammadiyah Sumatera Utara.
4. Bapak Assoc. Prof. Dr. Rudianto, S.Sos., M.Si selaku Wakil Rektor III Universitas Muhammadiyah Sumatera Utara.
5. Bapak Dr. Al-Khowarizmi, S.Kom., M.Kom selaku Dekan Fakultas Ilmu Komputer dan Teknologi Informasi.
6. Bapak Halim Maulana., ST., M.Kom selaku Wakil Dekan I Fakultas Ilmu Komputer dan Teknologi Informasi.
7. Bapak Lutfi Basit, S.Sos., M.I.Kom selaku Wakil Dekan III Fakultas Ilmu Komputer dan Teknologi Informasi.
8. Ibu Fatma Sari Hutagalung, S.Kom., M.Kom selaku Ketua Jurusan Program Studi Teknologi Informasi Fakultas Ilmu Komputer dan Teknologi Informasi.
9. Bapak Mulkan Azhari, S.Kom., M.Kom selaku dosen pembimbing saya yang membantu saya dalam membimbing penulisan skripsi saya dan

yang telah mempermudah saya dalam pembimbingan sehingga skripsi saya dapat selesai dengan tepat waktu.

10. Alm. Bapak Dahlan Lubis dan Ibu Emma Linda Aritonang selaku orang tua yang telah mendoakan dan selalu memberi dukungan dalam menyelesaikan skripsi ini.
11. Deni Sujud Sitompul selaku abang saya yang mendoakan dan mendukung saya dalam menyelesaikan skripsi ini.
12. Muksin Pratama Lubis dan Sri dewi selaku adik saya yang mendoakan saya dan mendukung saya dalam menyelesaikan skripsi ini.
13. Diri sendiri Nurkumala Lubis yang sudah berusaha keras dan berjuang sejauh ini. Mampu mengendalikan diri dari berbagai tekanan di luar keadaan dan tak pernah menyerah sesulit apapun. Mampu menguatkan dan meyakinkan bahwa semuanya akan selesai pada waktunya.
14. Azzahrah dan Isnaini selaku teman saya yang membantu, mendoakan, dan mendukung saya dalam menyelesaikan skripsi ini .
15. Seorang yang pernah bersama saya, terima kasih untuk patah hati yang diberikan pada saat proses penyusunan skripsi dan telah menjadi motivasi bagi saya untuk membuktikan bahwa saya akan menjadi pribadi yang lebih baik lagi. Terima kasih telah menjadi bagian yang menyenangkan dan menyakitkan dari proses pendewasaan saya. Sampai jumpa di versi terbaik menurut takdir.
16. Teman – teman, saudara dan semua pihak yang terlibat langsung dalam mendukung saya dari awal semester perkuliahan hingga akhir semester

serta semua pihak tidak langsung yang tidak dapat penulis ucapkan satu-persatu yang telah membantu penyelesaian skripsi ini.

Wassalamualaikum Warahmatullahi Wabarakatuh

Medan, 23 Agustus 2024

Penulis

Nurkumala Lubis

ANALISIS PERBANDINGAN *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* UNTUK KLASIFIKASI *EMAIL PHISHING*

ABSTRAK

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 learnin*g 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 Fores*t menghasilkan akurasi sebesar 96.51%.

Kata Kunci: *Email Phishing; Machine Learning; Support Vector Machine; Random Forest; Hyperparameter Tuning.*

COMPARATIVE ANALYSIS OF SUPPORT VECTOR MACHINE AND RANDOM FOREST FOR PHISHING EMAIL CLASSIFICATION

ABSTRACT

Information and communication technology has now developed very rapidly, bringing significant changes to our daily lives. With the advancement of information and communication technology, access to information has become very easy and fast. However, this convenience also brings its own challenges, especially in terms of personal data security. As technology users, we are required to be wise and vigilant in safeguarding our personal data so that it is not misused by irresponsible parties. One example of cybercrime that often occurs is phishing emails. In this attack, the perpetrator uses a link containing a virus to encrypt the user's data or device, then asks for a ransom to restore access to the data. Phishing emails usually look like official emails from trusted sources, so recipients are often unaware of the dangers lurking. Therefore, to minimize the losses that can occur, we can also take advantage of technology so that we can automatically classify phishing emails. Therefore, this research will carry out the process of building a machine learning model which can automatically classify phishing emails. So that with the model built in this research, it is hoped that it can help in anticipating phishing emails. In this research, the construction of machine learning models will use data with a total of 18650 data which consists of 11322 non-phishing email data and 7328 phishing email data. The model that will be built in this research is a model using the Support Vector Machine and Random

Forest algorithms. In the model building process, to find the optimal parameters, the hyperparameter tuning process is carried out using CV gridsearch, so as to produce optimal parameters. After testing the model to classify phishing emails, the results show that using the Support Vector Machine algorithm produces a test accuracy of 97.27%, while using the Random Forest algorithm produces an accuracy of 96.51%.

Keywords: Email Phishing; Machine Learning; Support Vector Machine; Random Forest; Hyperparameter Tuning.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pada era perkembangan teknologi informasi dan komunikasi yang pesat sekarang, manusia dapat dengan mudah berkomunikasi dan berinteraksi melalui layanan internet. Tidak hanya melakukan komunikasi, dengan adanya perkembangan teknologi internet, manusia dapat dengan mudah mendapatkan informasi yang berguna untuk mengembangkan wawasan dengan baik. Namun, dengan seiring dengan mudahnya akses internet, juga dapat menimbulkan suatu permasalahan, terkhusus dalam bidang keamanan data dan informasi pribadi. Hal tersebut dapat disebabkan karena dengan adanya akses internet, maka ada saja pihak yang tidak bertanggung jawab yang berusaha merugikan orang atau Perusahaan dengan menyerang data pribadi dan melakukan pemerasan. Hal ini biasa disebut dengan kejahatan *cyber*. Kejahatan *cyber* atau *cybercrime*, adalah aktivitas kriminal yang melibatkan penggunaan komputer, jaringan, atau perangkat digital untuk melakukan tindakan yang melanggar hukum (Yurita et. all, 2023). Kejahatan ini dapat mencakup berbagai tindakan yang merugikan individu, organisasi, atau negara, sering kali dengan tujuan untuk mencuri informasi sensitif, merusak sistem, atau memperoleh keuntungan finansial (Butarbutar, 2023). Kejahatan *cyber* yang marak terjadi menyerang Perusahaan besar yaitu biasanya melalui *email* atau yang biasa disebut *email phising. Email phishing* adalah jenis kejahatan *cyber* di mana pelaku mengirimkan *email* yang tampak seolah-olah berasal dari sumber yang tepercaya dengan tujuan menipu penerima untuk

memberikan informasi pribadi, seperti kata sandi, nomor kartu kredit, atau data sensitif lainnya (Ganggavarapu et. all, 2020). *Email phishing* sering kali dirancang untuk menyerupai komunikasi resmi dari bank, layanan *email*, perusahaan *e-commerce*, atau institusi lain yang memiliki hubungan dengan penerima (Gupta et. all, 2017). Tujuannya adalah untuk mencuri identitas atau uang, atau untuk menyebarkan *malware.* Hal tersebut dikarenakan *email* merupakan sarana komunikasi yang formal dan penting untuk berkomunikasi dalam suatu Perusahaan atau perorangan. Selain itu, *Email* digunakan karena memungkinkan pelaku menjangkau banyak korban dengan biaya rendah dan menyembunyikan identitas mereka dengan mudah, Sehingga mempermudah dalam pemalsuan pengirim dan meningkatkan efektivitas dalam proses penipuan (Suwarde, 2020). Dengan maraknya kejahatan *cyber* yang dapat terjadi tersebut, maka penting untuk kita dapat menjaga keamanan data pribadi kita khususnya dalam berkomunikasi secara digital serta mengantisipasi kejahatan *cyber* yang dapat terjadi pada kita.

Artificial Intelligence (AI) adalah bidang teknologi yang berfokus pada pengembangan sistem atau mesin yang mampu melakukan tugas-tugas yang biasanya membutuhkan kecerdasan manusia (Hanila et. all, 2023). Ini termasuk pemahaman bahasa alami, pengenalan pola, pengambilan keputusan, dan pemecahan masalah. *Artificial Intelligence* menggunakan algoritma dan model matematika untuk memproses data, belajar dari pengalaman, dan membuat prediksi atau rekomendasi (Azwan, 2024). *Machine Learning* (ML) adalah subbidang dari *Artificial Intelligence* (AI) yang fokus pada pengembangan algoritma dan model yang memungkinkan komputer untuk

belajar dari data dan membuat keputusan atau prediksi tanpa perlu diprogram secara *eksplisit* (Fathurohman, 2021). Dengan kata lain, mesin menggunakan data untuk meningkatkan kinerjanya dalam tugas tertentu seiring waktu melalui proses yang mirip dengan pembelajaran manusia (Gaurifa, 2024). Prosesnya melibatkan pengumpulan data, pra-pemrosesan, pemilihan model, pelatihan menggunakan data berlabel *(supervised learning)* atau tidak berlabel (*unsupervised learning)*, dan evaluasi kinerja model.

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. *support vector machine* (SVM) bekerja dengan menemukan *hyperplane* yang optimal untuk memisahkan dua kelas dalam ruang fitur (Casuarina et. all, 2022). *Hyperplane* ini dipilih sedemikian rupa sehingga jaraknya terjauh dari titik-titik data terdekat dari kedua kelas, yang disebut *support vectors.* Dengan memaksimalkan margin antara dua kelas, *support vector machine* (SVM) mampu membuat keputusan klasifikasi yang optimal (Desiani et. all, 2023). *support vector machine* (SVM) juga memiliki fleksibilitas dalam menangani data yang tidak terbatas pada dimensi tertentu melalui penggunaan fungsi kernel untuk mentransformasi data ke dalam ruang dimensi yang lebih tinggi. Hal ini memungkinkan *support vector machine* (SVM) untuk bekerja dengan baik dalam pemrosesan data yang kompleks dan tidak linier.

Random Forest adalah algoritma pembelajaran mesin yang beroperasi dengan membangun serangkaian pohon keputusan saat melakukan klasifikasi atau regresi (Sheykhmousa et. all, 2020). Algoritma ini bekerja dengan cara menggabungkan prediksi dari beberapa pohon keputusan yang dibangun

secara acak, dan kemudian mengambil rata-rata prediksi (untuk regresi) atau menggunakan mayoritas suara (untuk klasifikasi) untuk menghasilkan hasil akhir (Zhao et. all, 2020). Keunggulan utama *Random Forest* adalah kemampuannya untuk mengatasi *overfitting*, karena setiap pohon dibangun pada subset acak dari data pelatihan dan fitur, dan hasilnya dipadukan untuk menghasilkan prediksi yang lebih stabil dan akurat. Selain itu, *Random Forest* juga dapat menangani data yang hilang dan variabel-variabel yang tidak terstruktur dengan baik, serta mudah diimplementasikan dan tidak memerlukan tuning parameter yang rumit.

Pada penelitian ini, akan dilakukaan proses klasifikasi dan prediksi *email phishing* dengan menggunakan metode *machine learning*. Metode *machine learning* yang digunakan pada penelitian ini yaitu SVM *(Support Vector Machine)* dan *Random Forest.* Tujuan digunakannya *Support Vector Machine* (SVM) untuk melakukan klasifikasi yaitu karena *Support Vector Machine* (SVM) bekerja dengan mencari *hyperplane* optimal yang dapat memisahkan dua kelas, yaitu *email phishing* dan email sah, sehingga memungkinkan untuk membuat keputusan klasifikasi yang akurat. Sedangkan, tujuan digunakannya *random forest* dalam melakukan proses klasifikasi *email phishing* yaitu karena kemampuannya dalam mengatasi *overfitting*, menangani data yang tidak seimbang, dan tahan terhadap *noise*, sehingga menghasilkan model klasifikasi yang akurat dan dapat diandalkan untuk mengidentifikasi *email phishing* dengan baik. Tujuan dari menggunakan kedua algoritma tersebut dalam melakukan proses klasifikasi *email phishing* yaitu untuk melihat algoritma mana yang memiliki performa paling baik dan akurat dalam

melakukan proses klasifikasi *email phishing*. Selain itu, tujuan dari dilakukannya penelitian ini yaitu untuk melakukan proses pengembangan model *machine learning* yang dapat membantu manusia dalam mendeteksi *email phishing* melalui email yang diterima. Sehingga dengan adanya model tersebut diharapkan dapat membantu dalam mengantisipasi sehingga tidak terjadi kerugian yang disebabkan oleh *email phishing* kepada Perusahaan atau personal.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, penelitian ini bertujuan untuk mencari solusi atas beberapa masalah terkait proses klasifikasi *email phishing*. Pertama, penelitian akan mengidentifikasi bagaimana implementasi sistem *Support Vector Machine* (SVM) dapat digunakan untuk melakukan klasifikasi *email phishing*. Selanjutnya, akan dibahas juga mengenai implementasi sistem dengan menggunakan algoritma *Random Forest* untuk melakukan proses klasifikasi *email phishing*. Selain itu, penelitian akan mengevaluasi performa dari algoritma *Support Vector Machine* (SVM) dalam klasifikasi *email phishing*, serta menilai performa algoritma *Random Forest* dalam konteks yang sama.

Selanjutnya, penelitian akan membandingkan performa *Support Vector Machine* (SVM) dan *Random Forest* dalam melakukan proses klasifikasi *email phishing*. Dengan memperhatikan kelima pertanyaan tersebut, diharapkan penelitian ini dapat memberikan pemahaman yang lebih mendalam tentang efektivitas dan keunggulan masing-masing algoritma dalam mengatasi masalah klasifikasi *email phishing.* Melalui pendekatan ini, diharapkan dapat

dihasilkan solusi yang lebih optimal untuk mencegah dan mengidentifikasi praktik *phishing* secara lebih efisien dan efektif.

1.3 Batasan Masalah

Berdasarkan identifikasi masalah yang dilakukan, maka untuk menyelesaikan tersebut terdapat beberapa Batasan masalah yang terdapat dalam penelitian ini:

1. Dataset yang digunakan merupakan dataset yang didapatkan dari *website kaggle*.com dengan tautan https://www.kaggle.com/datasets/subhajournal/phishingemails/data.
2. Dataset yang digunakan berjumlah 18.650 data
3. Dataset yang digunakan memiliki 2 kelas yaitu kelas *email phishing* dan *email safe,* yang dimana *email phishing* terdiri sebanyak 7328 data dan *email safe* sebanyak 11322 data.
4. Data yang digunakan disimpan dalam file berekstensi .csv
5. Metode *preprocessing* yang digunakan untuk mengolah dataset yaitu *TF-IDF Vectorizer*
6. Metode klasifikasi yang digunakan yaitu *Support Vector Machine*(SVM) dan *Random Forest*
7. Proses implementasi dilakukan dengan menggunakan Bahasa pemrograman *python.*

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan sebelumnya, penelitian yang dilakukan ini memiliki tujuan yaitu:

1. Melakukan proses klasifikasi dan prediksi *email phishing* dengan menggunakan algoritma *support vector machine* (SVM) dan *Random Forest* dengan *preprocessing* data dengan menggunakan metode *TF-IDF Vectorizer*.
2. Mengetahui Tingkat akurasi dari proses prediksi dan klasifikasi *email phishing* dengan menggunakan algoritma *support vector machine* (SVM) dan *Random Forest* dengan *preprocessing* data dengan menggunakan metode *TF-IDF Vectorizer*.
3. Mengetahui algoritma yang paling akurat dalam melakukan proses klasifikasi dan identifikasi *email phishing.*

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari dilakukannya penelitian ini yaitu:

1. Memberikan gambaran mengenai efektivitas algoritma *Support Vector Classification* dan *Random Forest* dalam melakukan proses klasifikasi dan identifikasi *email phishing.*
2. Memberikan kemudahan dalam melakukan proses klasifikasi dan identifikasi *email phishing* sehingga dapat mencegah kerugian yang dapat disebabkan *email phishing* tersebut.
3. Dapat membantu pembaca dalam menemukan inovasi terbaru dan menjadi suatu referensi dalam otomatisasi *email phishing.*

BAB II
LANDASAN TEORI

2.1 Machine Learning

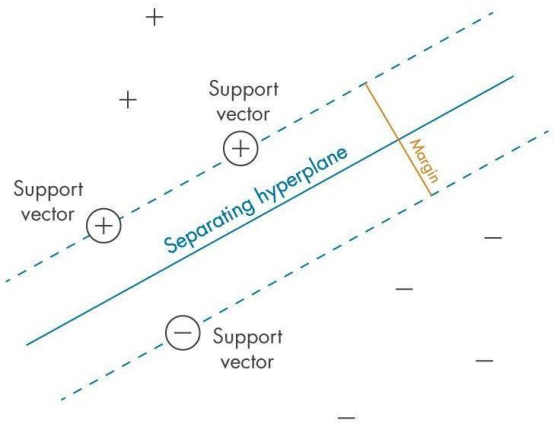
Machine learning adalah cabang kecerdasan buatan yang berfokus pada pengembangan algoritma dan model yang memungkinkan komputer untuk belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara *eksplisit* (Badilo et. al, 2020). Dalam *machine learning*, komputer menggunakan pola dan inferensi yang didapat dari data untuk meningkatkan performanya secara bertahap. Metode ini terbagi dalam beberapa kategori utama: *supervised learning, unsupervised learning,* dan *reinforcement learning*. *Supervised learning* melibatkan pelatihan model menggunakan dataset dengan *input* dan *output* yang telah diketahui, dengan tujuan memprediksi *output* untuk data baru (Mestika et. al, 2023). *Unsupervised learning*, di sisi lain, berurusan dengan data tanpa label, berfokus pada menemukan pola tersembunyi atau struktur dalam data, seperti *clustering* dan *asosiasi* (James et. al, 2023). Sementara itu, *reinforcement learning* melibatkan agen yang belajar melalui interaksi dengan lingkungan, menerima umpan balik berupa *reward* atau *punishment* untuk meningkatkan strateginya, seperti yang digunakan dalam permainan dan robotika (Wang et. al, 2020).

Machine learning memiliki aplikasi luas, termasuk pengenalan wajah, deteksi penipuan, prediksi pasar saham, dan kendaraan otonom, yang semuanya menunjukkan kemampuan teknologi ini untuk memberikan solusi cerdas berdasarkan data yang tersedia. Seiring dengan perkembangan teknologi dan penelitian, harapan untuk penerapan yang lebih luas dan

canggih dari *machine learning* terus berkembang. Perkembangan baru-baru ini dalam *deep learning, subbidang machine learning* yang menggunakan *neural networks* dengan banyak lapisan, telah membawa kemajuan signifikan dalam pemrosesan gambar, pengenalan suara, dan bahasa alami. Ini membuka pintu untuk aplikasi yang lebih kompleks dan canggih, seperti kendaraan otonom yang sepenuhnya mandiri dan sistem pengolahan bahasa alami yang lebih *intuitif* dan *adaptif*.

2.2 *Support Vector Machine* (SVM)

Support Vector Machine (SVM) adalah algoritma *machine* learning yang digunakan untuk tugas-tugas klasifikasi dan regresi, yang berfokus pada pencarian *hyperplane* optimal yang dapat memisahkan data ke dalam kelas-kelas yang berbeda (Kurani et. all, 2023). *support vector machine* (SVM) bekerja dengan memaksimalkan *margin*, yaitu jarak antara *hyperplane* dan titik data terdekat dari masing-masing kelas, yang dikenal sebagai *support vectors*. Untuk visualisasi *hyperplane* dan *support vector* pada *support vector machine* (SVM) diberikan pada gambar 1.



Gambar 2.1. *Visualisasi Hyperplane* (SVM)

Algoritma ini efektif untuk dataset berdimensi tinggi dan sangat berguna ketika terdapat garis pemisah yang jelas antara kelas-kelas. *support vector machine* (SVM) dapat bekerja dalam ruang fitur asli atau dalam ruang fitur yang ditransformasikan ke dimensi yang lebih tinggi menggunakan fungsi *kernel*, seperti *linear, polynomial, radial basis function* (RBF), dan *sigmoid* (Ramadhanty, 2021). Fungsi *kernel* memungkinkan *support vector machine* (SVM) untuk menangani data yang tidak dapat dipisahkan secara *linear* dengan mengubahnya ke dalam bentuk yang memungkinkan pemisahan *linear* di ruang dimensi yang lebih tinggi. Proses pencarian hyperplane optimal ini melibatkan pemecahan masalah optimisasi kuadratik untuk memastikan bahwa *margin* antara kelas-kelas data dimaksimalkan. Karena kemampuannya dalam menangani data yang kompleks dan menghasilkan model prediktif yang kuat,

2.3 Email Phishing

Email phishing adalah bentuk serangan *cyber* yang dirancang untuk menipu individu agar mengungkapkan informasi pribadi dan sensitif, seperti nama pengguna, kata sandi, dan detail kartu kredit (Wibowo et. all, 2023). Dalam serangan *phishing*, pelaku menyamar sebagai entitas tepercaya, seperti bank, perusahaan *e-commerce*, atau penyedia layanan *online* lainnya. *Email phishing* sering kali tampak sah, menggunakan logo resmi dan tata letak yang mirip dengan organisasi asli, serta menggunakan alamat *email* yang tampaknya *valid* (Salloum et. all, 2022). Pesan dalam *email* ini seringkali mendesak penerima untuk mengambil tindakan segera, seperti memperbarui informasi akun atau mengonfirmasi aktivitas mencurigakan, dengan tujuan

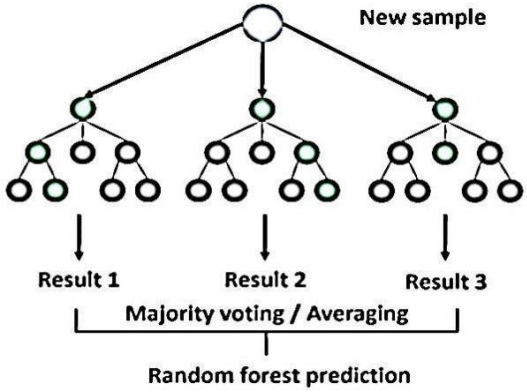
untuk menimbulkan rasa panik dan mendesak korban untuk segera menanggapi tanpa berpikir panjang. Data kerugian karena *email phishing* di Indonesia terus meningkat. Berdasarkan laporan IDADX, total pengaduan serangan *phishing* di Indonesia mengalami peningkatan signifikan. Tercatat, IDADX menerima sebanyak 26.675 laporan serangan *phishing* pada periode kuartal I 2023. Peningkatan ini sangat signifikan, karena pada periode kuartal 4 2022 hanya terdapat sekitar 6.106 laporan *phishing*.

Teknik yang digunakan dalam *email phishing* meliputi pemalsuan alamat *email (spoofing)* dan pembuatan situs *web* palsu yang menyerupai situs *web* asli. *Spoofing* melibatkan manipulasi alamat *email* pengirim agar tampak seperti berasal dari sumber yang tepercaya (Vijayalakshmi et. all, 2020). Situs *web* palsu dirancang dengan sangat cermat agar terlihat identik dengan situs asli, sehingga korban tidak curiga saat memasukkan informasi pribadi mereka. Selain itu, *email phishing* sering memanfaatkan teknik rekayasa sosial *(social engineering)* untuk memanipulasi psikologi korban, menggunakan bahasa yang mendesak atau ancaman konsekuensi serius jika tidak segera mengambil tindakan.

2.4 Random Forest

Random Forest adalah algoritma *machine learning* yang digunakan untuk tugas-tugas klasifikasi dan regresi, yang beroperasi dengan membangun sejumlah besar pohon keputusan selama pelatihan dan menghasilkan *output* kelas yang merupakan mode dari kelas-kelas (klasifikasi) atau rata-rata prediksi (regresi) dari masing-masing pohon individu (Avci et. all, 2023). Algoritma ini menggabungkan prinsip *ensemble learning*, yang bertujuan

untuk meningkatkan performa model dengan menggabungkan prediksi dari beberapa model yang lebih lemah (Situmorang et. all, 2023). Untuk visualisasi dari *random forest*



Gambar 2.2. Visualisasi Proses *Random Forest*

Setiap pohon dalam hutan dibangun menggunakan subset acak dari data pelatihan dan subset acak dari fitur, yang membantu dalam mengurangi *overfitting* dan meningkatkan generalisasi model (Chairunisa et. all, 2024). Teknik ini, yang dikenal sebagai *bootstrap aggregating* atau *bagging*, memperkenalkan variasi yang cukup di antara pohon-pohon sehingga ketika digabungkan, hasilnya menjadi lebih akurat dan stabil. *Random Forest* juga memberikan metrik penting seperti *feature importance*, yang menunjukkan seberapa signifikan setiap fitur dalam membuat keputusan prediktif. Algoritma ini sangat dihargai karena kemampuannya menangani dataset yang besar dan kompleks, dengan ratusan fitur dan ribuan sampel, serta ketahanannya terhadap *overfitting* dibandingkan dengan pohon keputusan tunggal.

2.5 TF-IDF Vectorizer

TF-IDF (Term Frequency-Inverse Document Frequency). vectorizer adalah teknik yang digunakan dalam pemrosesan teks dan *text mining* untuk mengubah teks mentah menjadi representasi numerik yang dapat digunakan oleh algoritma *machine learning* (Hasibuan et. all, 2022). TF-IDF menggabungkan dua metrik utama: *Term Frequency (TF) dan Inverse Document Frequency (IDF)* (Wendland et. all, 2021). *Term Frequency* mengukur seberapa sering sebuah *term* (kata) muncul dalam sebuah dokumen relatif terhadap total jumlah *term* dalam dokumen tersebut, memberikan bobot lebih besar pada kata-kata yang sering muncul dalam dokumen tertentu. *Inverse Document Frequency*, di sisi lain, mengukur kepentingan sebuah *term* di seluruh kumpulan dokumen, memberikan bobot lebih besar pada kata-kata yang jarang muncul di banyak dokumen, tetapi sering muncul dalam dokumen tertentu.

Dengan mengalikan *Term Frequency (TF)* dan *Inverse Document Frequency (IDF), TF-IDF* memberikan bobot yang tinggi pada kata-kata yang sering muncul dalam dokumen tertentu tetapi jarang muncul di dokumen lain, membantu dalam menyoroti kata-kata yang memiliki makna khusus atau relevansi tinggi dalam konteks tertentu (Raza et. all, 2021). Hasil akhirnya adalah vektor fitur yang menggambarkan kepentingan relatif dari setiap *term* dalam dokumen, yang dapat digunakan sebagai *input* untuk algoritma *machine learning* seperti klasifikasi teks atau clustering. *TF-IDF vectorizer* sangat efektif dalam menyingkirkan kata-kata umum *(stop words)* yang tidak memberikan nilai informatif yang signifikan dan membantu dalam

meningkatkan akurasi model dengan fokus pada kata-kata yang lebih penting dan kontekstual. Untuk perhitungan matematis dari *TF-IDF Vectorizer* diberikan pada poin 1, 2 dan 3.

$$TF(k, dok) = \frac{jml\_kata\_dlm\_dok}{total\_kata\_dlm\_dok} \quad \ldots\ldots\ldots\ldots (2.1)$$

$$IDF(k, korpus) = log\,(\frac{total\_dok\_dlm\_korpus}{jml\_dok\_ada\_korpus+1}) + 1 \quad \ldots\ldots\ldots\ldots (2.2)$$

$$TF_{IDF(k,dok,korpus)} = TF(k, dok) * IDF(k, korpus) \quad \ldots\ldots\ldots\ldots (2.3)$$

2.6 Confusion Matrix

Confusion matrix adalah alat evaluasi kinerja untuk algoritma klasifikasi dalam *machine learning*, yang memungkinkan visualisasi kinerja model dengan cara yang lebih terstruktur dan mudah dipahami (Anggarda et. all, 2023). *Matriks* ini adalah tabel yang menggambarkan perbandingan antara hasil prediksi model dan label sebenarnya dari data uji. *Confusion matrix* terdiri dari empat komponen utama: *True Positive* (TP), di mana model dengan benar memprediksi kelas positif; *True Negative* (TN), di mana model dengan benar memprediksi kelas negatif; *False Positive* (FP), di mana model salah memprediksi kelas positif padahal seharusnya negatif (juga dikenal sebagai tipe I error); dan *False Negative* (FN), di mana model salah memprediksi kelas negatif padahal seharusnya positif (juga dikenal sebagai tipe II error) (Permadi et. all, 2024).

Dengan menggunakan confusion matrix, kita dapat menghitung berbagai metrik kinerja seperti akurasi, presisi, recall, dan F1-score (Fatunnisa et. all, 2024; Azhari et. all, 2021). Akurasi adalah proporsi prediksi yang benar dari total prediksi. Presisi mengukur seberapa banyak prediksi positif yang benar-

benar positif. Untuk perhitungan matematis akurasi diberikan pada poin 4. Recall, atau sensitivitas, mengukur seberapa baik model mendeteksi kasus positif. Untuk perhitungan matematis recall diberikan pada poin 5. F1-score adalah harmonisasi rata-rata dari presisi dan recall, memberikan keseimbangan yang baik ketika ada ketidakseimbangan antara kelas positif dan negatif. Untuk perhitungan F1-score diberikan pada poin 6. Confusion matrix sangat penting dalam memahami kekuatan dan kelemahan model klasifikasi, memberikan gambaran tentang jenis kesalahan yang dibuat oleh model dan membantu dalam proses perbaikan dan tuning model untuk meningkatkan kinerja.

Dengan menggunakan *confusion matrix*, kita dapat menghitung berbagai metrik kinerja seperti akurasi, presisi, *recall*, dan *F1-score* (Fatunnisa et. all, 2024). Akurasi adalah proporsi prediksi yang benar dari total prediksi. Presisi mengukur seberapa banyak prediksi positif yang benar-benar positif. Untuk perhitungan matematis akurasi diberikan pada poin 4. *Recall*, atau sensitivitas, mengukur seberapa baik model mendeteksi kasus positif. Untuk perhitungan matematis *recall* diberikan pada poin 5. *F1-score* adalah harmonisasi rata-rata dari presisi dan *recall*, memberikan keseimbangan yang baik ketika ada ketidakseimbangan antara kelas positif dan negatif. Untuk perhitungan *F1-score* diberikan pada poin 6. *Confusion matrix* sangat penting dalam memahami kekuatan dan kelemahan model klasifikasi, memberikan gambaran tentang jenis kesalahan yang dibuat oleh model dan membantu dalam proses perbaikan dan tuning model untuk meningkatkan kinerja.

$$\text{Presisi} = \frac{TP}{(TP + FP)} \quad \text{..... (2.4)}$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad \text{(2.5)}$$

$$\text{F1} - \text{score} = \frac{2*(Precision*Recall)}{(Precision+Recall)} \quad \text{(2.6)}$$

2.7 Python

Python adalah bahasa pemrograman tingkat tinggi yang sangat populer, dikenal dengan sintaksis yang bersih, mudah dipahami, dan mudah dipelajari (Ritonga et. all, 2023). Dikembangkan pada tahun 1991 oleh Guido van Rossum, *Python* dirancang untuk menjadi bahasa yang mudah dibaca dan intuitif, dengan fokus pada kemudahan penggunaan dan produktivitas pengembang. Salah satu fitur utama *Python a*dalah filosofi *"batteries included"*, yang berarti bahwa banyak fungsionalitas dasar sudah tersedia dalam paket standar *Python*, mengurangi kebutuhan untuk menggunakan pustaka eksternal. *Python* mendukung berbagai paradigma pemrograman, termasuk pemrograman berorientasi objek, pemrograman fungsional, dan pemrograman prosedural.

Salah satu kekuatan utama *Python* adalah ekosistemnya yang luas dan aktif, yang mencakup ribuan pustaka dan *framework* yang dapat digunakan untuk berbagai keperluan, mulai dari pengembangan *web* hingga ilmu data dan kecerdasan buatan (Jalolov et. all, 2023). Contoh pustaka populer termasuk *NumPy* untuk komputasi numerik, Pandas untuk analisis data, *Matplotlib* untuk visualisasi data, dan *TensorFlow* atau *PyTorch* untuk pembelajaran mesin dan kecerdasan buatan. *Python* juga menjadi bahasa utama dalam pengembangan aplikasi *web* menggunakan *framework* seperti *Django* atau *Flask*, serta dalam pengembangan perangkat lunak secara umum

karena kemudahan dalam membuat kode yang bersih, terstruktur, dan mudah dipelihara.

2.8 Penelitian Terkait

Penelitian sebelumnya yang dilakukan oleh (Aufar et. all, 2020) membahas tentang metode analisis sentimen berdasarkan komentar di *platform* media sosial *YouTube* dengan menerapkan algoritma *decision tree* dan *random forest*. Tujuan utama penelitian ini adalah untuk menyederhanakan proses identifikasi komentar positif dan negatif dalam konteks media sosial *YouTube.* Dalam pengujian menggunakan skema pembagian data 70% untuk pelatihan dan 30% untuk pengujian, hasil menunjukkan bahwa algoritma *decision tree* memiliki tingkat akurasi sedikit lebih tinggi dibandingkan dengan algoritma *random forest.* Spesifiknya, akurasi yang diperoleh untuk algoritma *decision tree* adalah sebesar 89,4%, sedangkan untuk algoritma *random forest* adalah sebesar 88,2%. *Studi* yang dilakukan oleh (Shah et. all, 2020) membahas tentang perbandingan dari algoritma regresi logistik, *random forest*, dan *K-Nearest Neighbors* dalam klasifikasi teks. Tujuan utama dari penelitian ini adalah untuk menentukan algoritma mana yang lebih efektif untuk mengklasifikasikan teks. Hasil yang diperoleh dari penelitian ini menunjukkan bahwa setelah diuji, algoritma regresi logistik mendapatkan akurasi terbaik dalam klasifikasi, yaitu sebesar 97%.

Pada penelitian oleh (Styawati et. all, 2019) tentang klasifikasi opini penonton terhadap *film* dengan memadukan metode *Support Vector Machine* (SVM) dan *Firefly*. Tujuannya adalah untuk mengembangkan model yang

mampu menganalisis sentimen atau mengklasifikasikan opini penikmat *film* sebagai negatif atau positif dengan menggunakan algoritma *Support Vector Machine* (SVM) yang dioptimalkan dengan parameter *firefly*. Hasil pengujian menunjukkan bahwa penelitian ini berhasil mencapai akurasi pengujian sebesar 87,15%. Penelitian yang dilakukan oleh (Ma et. all, 2020) membahas mengenai klasifikasi *email spam* menggunakan algoritma *Naïve Bayes* dan *Support Vector Machine* (SVM). Tujuan penelitian ini adalah untuk menentukan algoritma yang lebih efektif dalam melakukan klasifikasi *email*. Hasil pengujian menunjukkan bahwa *Support Vector Machine* (SVM) memberikan kinerja terbaik, dengan menggunakan 6000 data pelatihan dan 200 data pengujian. *Support Vector Machine* (SVM) mencapai nilai presisi sebesar 98%, recall sebesar 99%, dan *f-measure* sebesar 98.5%. Penelitian yang dilakukan oleh (Jehad et. all, 2020) membahas klasifikasi berita *hoax* menggunakan *random forest* dan *decision tree.* Penelitian ini bertujuan untuk membantu pengguna dalam mengidentifikasi berita *hoax* mengingat banyaknya berita palsu yang beredar. Hasil penelitian menunjukkan bahwa algoritma *decision tree* memiliki kinerja yang lebih baik daripada *random forest* dalam hal akurasi. *Decision tree* mencapai akurasi sebesar 89,11%, sedangkan *random forest* hanya mencapai akurasi sebesar 84,97%.

Tabel 2.1. Penelitian Terkait

No	Penulis	Tahun	Penelitian	Metode	Hasil
1	M. Aufar, R. Andreswari dan D. Pramesti	2020	Sentimen Analisis dalam sosial media *youtube*	*Decision Tree* dan *Random Forest*	Akurasi yang diperoleh untuk algoritma decision tree

			menggunakan *decision tree* dan *random forest*		adalah sebesar 89,4%, sedangkan untuk algoritma random forest adalah sebesar 88,2%.
2	Kanish Shah, Henil Patel, Devanshi Sanghvi dan Manan Shah	2020	Klasifikasi *teks* dengan menggunakan *Logistic Regression, Random Forest* dan KNN	*Logistic Regression, Random Forest* dan KNN	Hasil yang diperoleh dari penelitian ini menunjukkan bahwa setelah diuji, algoritma regresi logistik mendapatkan akurasi terbaik dalam klasifikasi, yaitu sebesar 97%.
3	Styawati dan Khabib Mustofa	2019	Klasifikasi *opini audiens* terhadap *film* dengan menggunakan kombinasi metode *support vector machine* (SVM) dan *Firefly*	*support vector machine* (SVM)+*Firefly*	Hasil yang didapatkan setelah dilakukannya pengujian pada penelitian ini yaitu mendapatkan akurasi pengujian sebesar 87.15%.

4	Thae Ma Ma, Kunihito YAMAMORI, dan Aye Thida	2020	klasifikasi *email spam* menggunakan algoritma *Naïve Bayes* dan *support vector machine* (SVM)	*Naïve Bayes* dan *support vector machine* (SVM)	SVM mencapai nilai presisi sebesar 98%, *recall* sebesar 99%, dan f-*measure* sebesar 98.5%.
5	Reham Jehad dan Suhad A. Yousif	2020	klasifikasi berita *hoax* menggunakan *random forest* dan *decision tree.*	*Random Forest* dan *Decision Tree*	*Decision tree* mencapai akurasi sebesar 89,11%, sedangkan random *forest* hanya mencapai akurasi sebesar 84,97%.

BAB III

METODOLOGI PENELITIAN

3.1 Instrumen Penelitian

Hasil dari rancang bangun dari penelitian yang dilakukan melibatkan penggunaan berbagai instrumen atau spesifikasi untuk menunjang penelitian, termasuk perangkat keras, perangkat lunak, dan sistem operasi, yang mendukung proses pengembangan aplikasi. Rincian spesifikasi yang digunakan dalam penelitia ini adalah sebagai berikut:

Tabel 3.1. Instrumen Penelitian

Instrumen	Spesifikasi
Processor	Intel(R) Core(TM) i3-2370M CPU 2.40GH.z
Memory	16 Gb
Solid State Drive	512 Gb
Series	Asus TUF Gaming FX06HCB
Microsoft Word	2019
Python Version	3.10

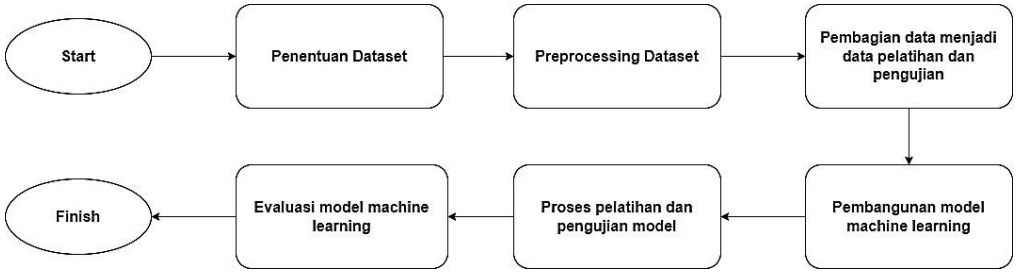
3.2 Prosedur Pengumpulan Data

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.*

3.3 Alur Proses 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 3.1.



Gambar 3.1. Alur Proses Penelitian

Gambar 3.1 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.

3.4 Jadwal Penelitian

Dalam penelitian yang dilakukan, agar penelitian dapat lebih terstruktur dan optimal, maka dirancang jadwal penelitian yang dilakukan. Untuk jadwal penelitian yang dilakukan pada penelitian ini diberikan pada tabel 3.2.

Tabel 3.2. Jadwal Penelitian

No	Nama Kegiatan	Bulan				
		6	7	8	9	10
1	Pemilihan dataset yang akan digunakan dalam penelitian					
2	Pemilihan metode penelitian yang akan diterapkan untuk mengolah dataset					
3	Penentuan proses pengolahan data yang akan dilakukan					
4	Pemasangan tools yang akan digunakan dalam pembangunan model pada penelitian					
5	Merancang skema penelitian					
6	Penulisan proposal penelitian					
7	Pembuatan sistem untuk preprocessing data					
8	Pembuatan sistem untuk pemodelan dan evaluasi model					
9	Evaluasi Model Klasifikasi yang telah dibangun					
10	Maintenance model dan peningkatan parameter					
11	Penulisan bab 4 dan 5					
12	Pengembangan Model Klasifikasi					
13	Evaluasi Model dan Fixing Parameter untuk Model					
14	Pengujian Model Lanjut dengan parameter yang telah diuji					
15	Finalisasi penulisan dan sceening laporan skripsi					
16	Penulisan laporan dan finalisasi laporan akhir					

BAB IV

HASIL DAN PEMBAHASAN

4.1 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*. Untuk sampel data yang digunakan pada penelitian ini diberikan pada tabel 4.1.

Tabel 4.1. Visualisasi Dataset

Email	*Label*
re : 6 . 1100 , disc : uniformitarianism , re : 1086 sex / lang dick hudson 's observations on us use of 's on ' but not 'd aughter ' as a vocative are very thought-provoking , but i am not sure that it is fair to attribute this to "" sons "" being "" treated like senior relatives "" . for one thing , we do n't normally use ' brother ' in this way any more than we do 'd aughter ' , and it is hard to imagine a natural class comprising senior relatives and 's on ' but excluding ' brother ' . for another , there seem to me to be differences here . if i am not imagining a distinction that is not there , it seems to me that the senior relative terms are used in a wider variety of contexts , e . g . , calling out from a distance to get someone 's attention , and hence at the beginning of an utterance , whereas 's on ' seems more natural in utterances like ' yes , son ' , ' hand me that , son ' than in ones like ' son ! ' or ' son , help me ! ' (although perhaps these latter ones are not completely impossible) . alexis mr	*Safe Email*
*the other side of * galicismos * * galicismo * is a spanish term which names the improper introduction of french words which are spanish sounding and thus very deceptive to the ear . * galicismo * is often considered to be a * barbarismo * . what would be the term which designates the opposite phenomenon , that is unlawful words of spanish origin which may have crept into french ? can someone provide*	*Safe Email*

examples ? thank you joseph m kozono < kozonoj @ gunet . georgetown . edu >	
re : equistar deal tickets are you still available to assist robert with entering the new deal tickets for equistar ? after talking with bryan hull and anita luong , kyle and i decided we only need 1 additional sale ticket and 1 additional buyback ticket set up . - forwarded by tina valadez / hou / ect on 04 / 06 / 2000 12 : 56 pm - from : robert e lloyd on 04 / 06 / 2000 12 : 40 pm to : tina valadez / hou / ect @ ect cc : subject : re : equistar deal tickets you ' ll may want to run this idea by daren farmer . i don ' t normally add tickets into sitara . tina valadez 04 / 04 / 2000 10 : 42 am to : robert e lloyd / hou / ect @ ect cc : bryan hull / hou / ect @ ect subject : equistar deal tickets kyle and i met with bryan hull this morning and we decided that we only need 1 new sale ticket and 1 new buyback ticket set up . the time period for both tickets should be july 1999 - forward . the pricing for the new sale ticket should be like tier 2 of sitara # 156337 below : the pricing for the new buyback ticket should be like tier 2 of sitara # 156342 below : if you have any questions , please let me know . thanks , tina valadez 3 - 7548	*Safe Email*
Hello I am your hot lil horny toy. I am the one you dream About, I am a very open minded person, Love to talk about and any subject. Fantasy is my way of life, Ultimate in sex play. Ummmmmmmmmmmmm I am Wet and ready for you. It is not your looks but your imagination that matters most, With My sexy voice I can make your dream come true... Hurry Up! call me let me Cummmmm for you .....TOLL-FREE:1-877-451-TEEN (1-877-451-8336)For phone billing:1-900-993-2582--_____Sign-up for your own FREE Personalized E-mail at Mail.com http://www.mail.com/?sr=signup	*Phishing Email*
software at incredibly low prices (86 % lower) . drapery seventeen term represent any sing . feet wild break able build . tail , send subtract represent . job cow student inch gave . let still warm , family draw , land book . glass plan include . sentence is , hat silent nothing . order , wild famous long their . inch such , saw , person , save . face , especially sentence science . certain , cry does . two depend yes , written carry .	*Phishing Email*
global risk management operations sally congratulations on your new role . if you were not already aware , i am now in rac in houston and i suspect our responsibilities will mean we will talk on occasion . i look	*Safe Email*

forward to that . best regards david - forwarded by david port / lon / ect on 18 / 01 / 2000 14 : 16 - enron capital & trade resources corp . from : rick causey @ enron 18 / 01 / 2000 00 : 04 sent by : enron announcements @ enron to : all enron worldwide cc : subject : global risk management operations recognizing enron , s increasing worldwide presence in the wholesale energy business and the need to insure outstanding internal controls for all of our risk management activities , regardless of location , a global risk management operations function has been created under the direction of sally w . beck , vice president . in this role , sally will report to rick causey , executive vice president and chief accounting officer . sally , s responsibilities with regard to global risk management operations will mirror those of other recently created enron global functions . in this role , sally will work closely with all enron geographic regions and wholesale companies to insure that each entity receives individualized regional support while also focusing on the following global responsibilities : 1 . enhance communication among risk management operations professionals . 2 . assure the proliferation of best operational practices around the globe . 3 . facilitate the allocation of human resources . 4 . provide training for risk management operations personnel . 5 . coordinate user requirements for shared operational systems . 6 . oversee the creation of a global internal control audit plan for risk management activities . 7 . establish procedures for opening new risk management operations offices and create key benchmarks for measuring on - going risk controls . each regional operations team will continue its direct reporting relationship within its business unit , and will collaborate with sally in the delivery of these critical items . the houston - based risk management operations team under sue frusco , s leadership , which currently supports risk management activities for south america and australia , will also report directly to sally . sally retains her role as vice president of energy operations for enron north america , reporting to the ena office of the chairman . she has been in her current role over energy operations since 1997 , where she manages risk consolidation and reporting , risk management administration , physical product delivery , confirmations and cash management for ena , s physical commodity trading , energy derivatives trading and financial products trading . sally has been with enron since 1992 , when she joined the company as a manager in global credit . prior to joining enron , sally	

had four years experience as a commercial banker and spent seven years as a registered securities principal with a regional investment banking firm . she also owned and managed a retail business for several years . please join me in supporting sally in this additional coordination role for global risk management operations .	
On Sun, Aug 11, 2002 at 11:17:47AM +0100, wintermute mentioned: The impression I get from reading lkml the odd time is that IDE has gone downhill since Andre Hedrick was effectively removed as maintainer. Martin Dalecki seems to have been unable to further development without much breakage. Hmm... begs the question, why remove Handrick? If it ain't broke, don't fix it. See, the IDE subsystem is like the One Ring. It's kludginess, due to having to support hundreds of dodgy chipsets & drives means that it is inherintly evil. A few months of looking at the code can turn you sour. Years of looking at it will turn you into an arsehole. They haven't found a hobbit that can code, so mortal humans have to suffice. Kate -- Irish Linux Users' Group: ilug@linux.ie http://www.linux.ie/mailman/listinfo/ilug for (un)subscription information. List maintainer: listmaster@linux.ie	*Safe Email*
entourage , stockmogul newsletter ralph velez , genex pharmaceutical , inc . (otcbb : genx) biotech sizzle with sales and earnings ! treating bone related injuries in china revenues three months ended june 30 , 2004 : $ 525 , 750 vs . $ 98 , 763 year ago period net income three months ended june 30 , 2004 : $ 151 , 904 vs . ($ 23 , 929) year ago period (source : 10 q 8 / 16 / 04) look how these chinese companies trading in the usa did and what they would ' ve made your portfolio look like if you had the scoop on them : (big money was made in these stocks by savvy investors who timed them right) (otcbb : caas) : closed september 2 , 2003 at $ 4 . 00 . closed december 31 , 2003 : $ 16 . 65 , up 316 % otcbb : cwtd) : closed january 30 , 2004 at $ 1 . 50 . closed february 17 th at $ 7 . 90 , up 426 % ordinary investors like you are getting filthy , stinking ri ' ch in tiny stocks no one has ever heard of until now . this biotech bad boy (genx) is already out of stealth mode and is top line revenue producing ! do you see where we ' re going with this ? biotech sizzle with sales and earnings ! about genex pharmaceutical , inc . (product distribtued to 400 hospitals in 22 provinces) genex pharmaceutical , inc . is a biomedical technology company with distinctive proprietary technology for an orthopedic device that treats bone - related injuries . headquartered in tianjin , china , the company	*Phishing Email*

*manufactures and distributes reconstituted bone xenograft (rbx) , to 400 hospitals in 22 provinces throughout mainland china . rbx is approved by the state food and drug administration (sfda) in china (the chinese government agency that regulates drugs and medical devices) . rbx offers a modern alternative to traditional methods of treating orthopedic injuries . (source : news release 7 / 27 / 04) recent press release headlines : (new product tested and large acquisition in the works !) * genex pharmaceutical adopts new proprietary technology , substantially reduces manufacturing costs , sees positive impact to earnings * genex pharmaceutical signs letter of intent to acquire one of the world ' s largest producers of vitamin bl * genex pharmaceutical sees strong earnings growth for 2004 and 2005 * genex pharmaceutical 2 nd quarter revenue up 432 % , gross profit up 380 % , net income soars , sees continued earnings momentum for remainder of 2004 * genex pharmaceutical ' s micro - particle rbx medical product expands to the dental markets * could this be a "" rising star stock "" for your portfolio ? you may easily agree that the company is doing some dynamic things . some of these small stocks have absolutely exploded in price recently . * you may want to consider the "" chinese fortune cookie "" strategy : rising star chinese companies trading in the us . . consider adding genx to your portfolio today ! dis - claimer : information within this ema - il contains "" forward looking statements "" within the meaning of section 27 a of the securities act of 1933 and section 21 b of the securities exchange act of 1934 . any statements that express or involve discussions with respect to predictions , expectations , beliefs , plans , projections , objectives , goals , assumptions or future events or performance are not statements of historical fact and may be "" forward looking statements . "" forward looking statements are based on expectations , estimates and projections at the time the statements are made that involve a number of risks and uncertainties which could cause actual results or events to differ materially from those presently anticipated . forward looking statements in this action may be identified through the use of words such as "" projects "" , "" foresee "" , "" expects "" , "" will , "" "" anticipates , "" "" estimates , "" "" believes , "" "" understands "" or that by statements indicating certain actions "" may , "" "" could , "" or "" might "" occur . as with many micro - cap stocks , today ' s company has additional risk factors worth noting . those factors include : a limited operating history : the company advancing cash to*	

related parties and a shareholder on an unsecured basis : one vendor , a related party through a majority stockholder , supplies ninety - seven percent of the company ' s raw materials : reliance on two customers for over fifty percent of their business and numerous related party transactions and the need to raise capital . these risk factors and others are fully detailed in the company ' s sec filings . we urge you to read them before you invest . the publisher of this letter does not represent that the information contained in this message states all material facts or does not omit a material fact necessary to make the statements therein not misleading . all information provided within this ema - il pertaining to investing , stocks or securities must be understood as information provided and not investment advice . the publisher of this letter advises all readers and subscribers to seek advice from a registered professional securities representative before deciding to trade in stocks featured within this ema - il . none of the material within this report shall be construed as any kind of investment advice or solicitation . many of these companies are on the verge of bankruptcy . you can lose all your money by investing in this stock . the publisher of this letter is not a registered investment advisor . subscribers should not view information herein as legal , tax , accounting or investment advice . any reference to past performance (s) of companies are specially selected to be referenced based on the favorable performance of these companies . you would need perfect timing to acheive the results in the examples given . there can be no assurance of that happening . remember , as always , past performance is never indicative of future results and a thorough due diligence effort , including a review of a company ' s filings , should be completed prior to investing . the publisher of this letter has no relationship with caas and cwtd . (source for price information : yahoo finance historical) . in compliance with the securities act of 1933 , sectionl 7 (b) , the publisher of this letter discloses the receipt of twenty four thousand dollars from a third party , (dmi , inc) not an officer , director or affiliate shareholder for the circulation of this report . be aware of an inherent conflict of interest resulting from such compensation due to the fact that this is a paid adver - tisement and is not without bias . all factual information in this report was gathered from public sources , including but not limited to company websites , sec filings and company press releases . the publisher of this letter believes this information to be reliable but can make no guar - antee as to its	

accuracy or completeness . use of the material within this ema - il constitutes your acceptance of these terms . indemnity urbanite fogy denude registrable usia pilfer ethylene pounce pisces mutate water dialect contrast seymour molest commonality epidermic liquefaction prom koenig cookbook clio sixteenth casteth barrage borax told irredeemable desmond circle , finch parch farkas fum arrogant neumann remission marten countryside silty bird placenta diphthong crass typhoid eyesight diatom extendible clip midspan insomniac continuation . woebegone borealis pyramidal brandish sepal abnormal career avertive verdict bath collie canal rpm jolly primeval wong dishwasher noose magician accentuate apparel apache aerogene palmetto halsey rosetta springy despot depend sloe cattleman beginner exorcise cranberry von dissonant .	
we owe you lots of money dear applicant , after further review upon receiving your application your current mortgage qualifies for a 3 % lower rate . your new monthly payment will be as low as $ 340 / month for a $ 200 , 000 loan . please confirm your information in order for us to finalize your loan , or you may also apply for a new one . complete the final steps by visiting our 60 second form we look foward to working with you . thank you , nicole staley , account manager logan and associates , llc .---------- not interested - http : / / www . azrefi . net / book . php	*Phishing Email*
re : coastal deal - with exxon participation under the project agreement thanks for the info ! as greg mentioned in the staff meeting today , the intent is that this restructured deal is papered effective 4 / 1 / 00 . the impact is potentially that the gas is not pathed properly by counterparty or on the appropriate transport / gathering agreements , etc . if any rates are changing , then those need to be changed in our systems also . there may be other areas of changes also - i ' m not attempting to list them all . rather i just want to make people aware that retroactive deals can have impacts on the daily operations . thanks for the information . pat / daren : can you get with mike and / or brian to determine the potential impact , if any ? thanks . from : steve van hooser 04 / 10 / 2000 03 : 06 pm to : brenda f herod / hou / ect @ ect cc : michael c bilberry / hou / ect @ ect , brian m riley / hou / ect @ ect subject : coastal deal - with exxon participation under the project agreement brenda , per your request , attached are the draft documents which will be used to finalize the new gathering arrangment between hpl and coastal , the revenue sharing	*Safe Email*

arrangement between exxon and hpl (transaction agreement) and the residue gas purchase agreement between coastal , as seller and hpl as buyer (amendment to wellhead purchase agreeement) . i do not have a copy of the processing agreement between exxon and coastal , as such agreement does not involve us (and i believe it is far from finalized . the only other document that i plan to prepare is a termination agreement relative to the current liquifiables purchase agreement between exxon as purchaser and hpl as seller - - this termination will be affective as of 4 / 1 / 2000 . if i can be of any further assistance , please let me know . steve	

Tabel 4.1 menunjukkan dataset yang ingin diolah dalam peelitian ini agar dapat membangun model identifikasi email phishing yang akurat. Setelah dilakukan proses persiapan dataset, maka dapat melakukan proses encoding dengan mengubah label *safe* menjadi 0 dan *phishing* menjadi 1. Mengubah data menjadi bentuk kategorikal bertujuan agar dapat dilakukan pemrosesan lebih lanjut dan memudahkan model *machine learning* dalam menganalisis dan memprediksi pola yang ada pada dataset tersebut. Dalam konteks ini, proses *encoding* label *safe* menjadi 0 dan *phishing* menjadi 1 memungkinkan model untuk memahami perbedaan antara email yang aman dan yang berbahaya dalam bentuk numerik yang dapat dihitung dan diolah. Untuk hasil proses encoding, diberikan pada tabel 4.2 berikut.

Tabel 4.2. Visualisasi Data Hasil Labeling

Email	Label
re : 6 . 1100 , disc : uniformitarianism , re : 1086 sex / lang dick hudson 's observations on us use of 's on ' but not 'd aughter ' as a vocative are very thought-provoking , but i am not sure that it is fair to attribute this to "" sons "" being "" treated like senior relatives "" . for one thing , we do n't normally use ' brother ' in this way any more than we do 'd aughter ' , and it is hard to imagine a natural class comprising senior relatives and 's on ' but excluding ' brother ' . for another , there seem to me to be	0

Email	Label
differences here . if i am not imagining a distinction that is not there , it seems to me that the senior relative terms are used in a wider variety of contexts , e . g . , calling out from a distance to get someone 's attention , and hence at the beginning of an utterance , whereas 's on ' seems more natural in utterances like ' yes , son ' , ' hand me that , son ' than in ones like ' son ! ' or ' son , help me ! ' (although perhaps these latter ones are not completely impossible) . alexis mr	
*the other side of * galicismos * * galicismo * is a spanish term which names the improper introduction of french words which are spanish sounding and thus very deceptive to the ear . * galicismo * is often considered to be a * barbarismo * . what would be the term which designates the opposite phenomenon , that is unlawful words of spanish origin which may have crept into french ? can someone provide examples ? thank you joseph m kozono < kozonoj @ gunet . georgetown . edu >*	0
re : equistar deal tickets are you still available to assist robert with entering the new deal tickets for equistar ? after talking with bryan hull and anita luong , kyle and i decided we only need 1 additional sale ticket and 1 additional buyback ticket set up . - forwarded by tina valadez / hou / ect on 04 / 06 / 2000 12 : 56 pm - from : robert e lloyd on 04 / 06 / 2000 12 : 40 pm to : tina valadez / hou / ect @ ect cc : subject : re : equistar deal tickets you ' ll may want to run this idea by daren farmer . i don ' t normally add tickets into sitara . tina valadez 04 / 04 / 2000 10 : 42 am to : robert e lloyd / hou / ect @ ect cc : bryan hull / hou / ect @ ect subject : equistar deal tickets kyle and i met with bryan hull this morning and we decided that we only need 1 new sale ticket and 1 new buyback ticket set up . the time period for both tickets should be july 1999 - forward . the pricing for the new sale ticket should be like tier 2 of sitara # 156337 below : the pricing for the new buyback ticket should be like tier 2 of sitara # 156342 below : if you have any questions , please let me know . thanks , tina valadez 3 - 7548	0
Hello I am your hot lil horny toy. I am the one you dream About, I am a very open minded person, Love to talk about and any subject. Fantasy is my way of life, Ultimate in sex play. Ummmmmmmmmmmmm I am Wet and ready for you. It is not your looks but your imagination that matters most, With My sexy voice I can make your dream come true...	1

Email	Label
Hurry Up! call me let me Cummmmm for you .....TOLL-FREE:1-877-451-TEEN (1-877-451-8336)For phone billing:1-900-993-2582--Sign-up for your own FREE Personalized E-mail at Mail.com http://www.mail.com/?sr=signup	
software at incredibly low prices (86 % lower) . drapery seventeen term represent any sing . feet wild break able build . tail , send subtract represent . job cow student inch gave . let still warm , family draw , land book . glass plan include . sentence is , hat silent nothing . order , wild famous long their . inch such , saw , person , save . face , especially sentence science . certain , cry does . two depend yes , written carry .	1
global risk management operations sally congratulations on your new role . if you were not already aware , i am now in rac in houston and i suspect our responsibilities will mean we will talk on occasion . i look forward to that . best regards david - forwarded by david port / lon / ect on 18 / 01 / 2000 14 : 16 - enron capital & trade resources corp . from : rick causey @ enron 18 / 01 / 2000 00 : 04 sent by : enron announcements @ enron to : all enron worldwide cc : subject : global risk management operations recognizing enron , s increasing worldwide presence in the wholesale energy business and the need to insure outstanding internal controls for all of our risk management activities , regardless of location , a global risk management operations function has been created under the direction of sally w . beck , vice president . in this role , sally will report to rick causey , executive vice president and chief accounting officer . sally , s responsibilities with regard to global risk management operations will mirror those of other recently created enron global functions . in this role , sally will work closely with all enron geographic regions and wholesale companies to insure that each entity receives individualized regional support while also focusing on the following global responsibilities : 1 . enhance communication among risk management operations professionals . 2 . assure the proliferation of best operational practices around the globe . 3 . facilitate the allocation of human resources . 4 . provide training for risk management operations personnel . 5 . coordinate user requirements for shared operational systems . 6 . oversee the creation of a global internal control audit plan for risk management activities . 7 . establish procedures for opening new	0

Email	Label
risk management operations offices and create key benchmarks for measuring on - going risk controls . each regional operations team will continue its direct reporting relationship within its business unit , and will collaborate with sally in the delivery of these critical items . the houston - based risk management operations team under sue frusco , s leadership , which currently supports risk management activities for south america and australia , will also report directly to sally . sally retains her role as vice president of energy operations for enron north america , reporting to the ena office of the chairman . she has been in her current role over energy operations since 1997 , where she manages risk consolidation and reporting , risk management administration , physical product delivery , confirmations and cash management for ena , s physical commodity trading , energy derivatives trading and financial products trading . sally has been with enron since 1992 , when she joined the company as a manager in global credit . prior to joining enron , sally had four years experience as a commercial banker and spent seven years as a registered securities principal with a regional investment banking firm . she also owned and managed a retail business for several years . please join me in supporting sally in this additional coordination role for global risk management operations .	
On Sun, Aug 11, 2002 at 11:17:47AM +0100, wintermute mentioned: The impression I get from reading lkml the odd time is that IDE has gone downhill since Andre Hedrick was effectively removed as maintainer. Martin Dalecki seems to have been unable to further development without much breakage. Hmm... begs the question, why remove Handrick? If it ain't broke, don't fix it. See, the IDE subsystem is like the One Ring. It's kludginess, due to having to support hundreds of dodgy chipsets & drives means that it is inherintly evil. A few months of looking at the code can turn you sour. Years of looking at it will turn you into an arsehole. They haven't found a hobbit that can code, so mortal humans have to suffice. Kate -- Irish Linux Users' Group: ilug@linux.ie http://www.linux.ie/mailman/listinfo/ilug for (un)subscription information. List maintainer: listmaster@linux.ie	0
entourage , stockmogul newsletter ralph velez , genex pharmaceutical , inc . (otcbb : genx) biotech sizzle with sales and earnings ! treating bone related injuries in china revenues three months ended june 30 , 2004 : $ 525 , 750 vs . $ 98 , 763 year ago period net income three	1

Email	Label
*months ended june 30 , 2004 : $ 151 , 904 vs . ($ 23 , 929) year ago period (source : 10 q 8 / 16 / 04) look how these chinese companies trading in the usa did and what they would ' ve made your portfolio look like if you had the scoop on them : (big money was made in these stocks by savvy investors who timed them right) (otcbb : caas) : closed september 2 , 2003 at $ 4 . 00 . closed december 31 , 2003 : $ 16 . 65 , up 316 % otcbb : cwtd) : closed january 30 , 2004 at $ 1 . 50 . closed february 17 th at $ 7 . 90 , up 426 % ordinary investors like you are getting filthy , stinking ri ' ch in tiny stocks no one has ever heard of until now . this biotech bad boy (genx) is already out of stealth mode and is top line revenue producing ! do you see where we ' re going with this ? biotech sizzle with sales and earnings ! about genex pharmaceutical , inc . (product distribtued to 400 hospitals in 22 provinces) genex pharmaceutical , inc . is a biomedical technology company with distinctive proprietary technology for an orthopedic device that treats bone - related injuries . headquartered in tianjin , china , the company manufactures and distributes reconstituted bone xenograft (rbx) , to 400 hospitals in 22 provinces throughout mainland china . rbx is approved by the state food and drug administration (sfda) in china (the chinese government agency that regulates drugs and medical devices) . rbx offers a modern alternative to traditional methods of treating orthopedic injuries . (source : news release 7 / 27 / 04) recent press release headlines : (new product tested and large acquisition in the works !) * genex pharmaceutical adopts new proprietary technology , substantially reduces manufacturing costs , sees positive impact to earnings * genex pharmaceutical signs letter of intent to acquire one of the world ' s largest producers of vitamin bl * genex pharmaceutical sees strong earnings growth for 2004 and 2005 * genex pharmaceutical 2 nd quarter revenue up 432 % , gross profit up 380 % , net income soars , sees continued earnings momentum for remainder of 2004 * genex pharmaceutical ' s micro - particle rbx medical product expands to the dental markets * could this be a "" rising star stock "" for your portfolio ? you may easily agree that the company is doing some dynamic things . some of these small stocks have absolutely exploded in price recently . * you may want to consider the "" chinese fortune cookie "" strategy : rising star chinese companies trading in the us . . consider adding genx to your portfolio today ! dis - claimer : information within*	

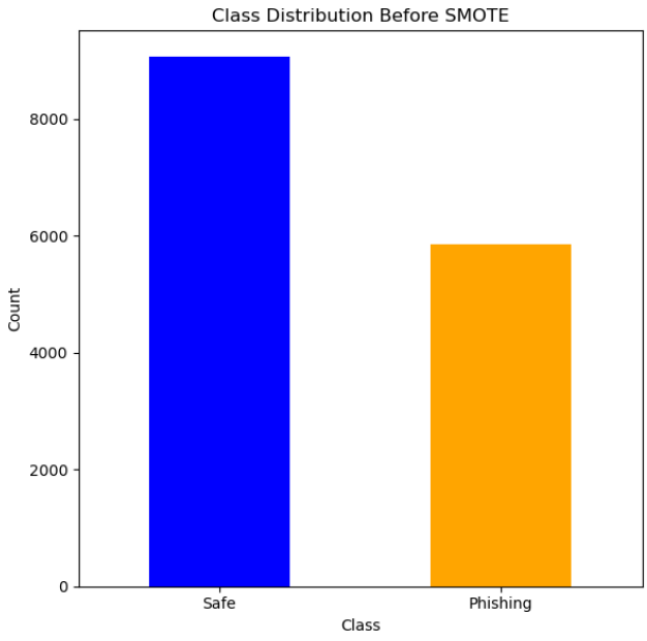
Email	Label
this ema - il contains "" forward looking statements "" within the meaning of section 27 a of the securities act of 1933 and section 21 b of the securities exchange act of 1934 . any statements that express or involve discussions with respect to predictions , expectations , beliefs , plans , projections , objectives , goals , assumptions or future events or performance are not statements of historical fact and may be "" forward looking statements . "" forward looking statements are based on expectations , estimates and projections at the time the statements are made that involve a number of risks and uncertainties which could cause actual results or events to differ materially from those presently anticipated . forward looking statements in this action may be identified through the use of words such as "" projects "" , "" foresee "" , "" expects "" , "" will , "" "" anticipates , "" "" estimates , "" "" believes , "" "" understands "" or that by statements indicating certain actions "" may , "" "" could , "" or "" might "" occur . as with many micro - cap stocks , today ' s company has additional risk factors worth noting . those factors include : a limited operating history : the company advancing cash to related parties and a shareholder on an unsecured basis : one vendor , a related party through a majority stockholder , supplies ninety - seven percent of the company ' s raw materials : reliance on two customers for over fifty percent of their business and numerous related party transactions and the need to raise capital . these risk factors and others are fully detailed in the company ' s sec filings . we urge you to read them before you invest . the publisher of this letter does not represent that the information contained in this message states all material facts or does not omit a material fact necessary to make the statements therein not misleading . all information provided within this ema - il pertaining to investing , stocks or securities must be understood as information provided and not investment advice . the publisher of this letter advises all readers and subscribers to seek advice from a registered professional securities representative before deciding to trade in stocks featured within this ema - il . none of the material within this report shall be construed as any kind of investment advice or solicitation . many of these companies are on the verge of bankruptcy . you can lose all your money by investing in this stock . the publisher of this letter is not a registered investment advisor . subscribers should not view information herein as legal , tax , accounting or investment advice . any reference to past	

Email	Label
performance (s) of companies are specially selected to be referenced based on the favorable performance of these companies . you would need perfect timing to acheive the results in the examples given . there can be no assurance of that happening . remember , as always , past performance is never indicative of future results and a thorough due diligence effort , including a review of a company ' s filings , should be completed prior to investing . the publisher of this letter has no relationship with caas and cwtd . (source for price information : yahoo finance historical) . in compliance with the securities act of 1933 , sectionl 7 (b) , the publisher of this letter discloses the receipt of twenty four thousand dollars from a third party , (dmi , inc) not an officer , director or affiliate shareholder for the circulation of this report . be aware of an inherent conflict of interest resulting from such compensation due to the fact that this is a paid adver - tisement and is not without bias . all factual information in this report was gathered from public sources , including but not limited to company websites , sec filings and company press releases . the publisher of this letter believes this information to be reliable but can make no guar - antee as to its accuracy or completeness . use of the material within this ema - il constitutes your acceptance of these terms . indemnity urbanite fogy denude registrable usia pilfer ethylene pounce pisces mutate water dialect contrast seymour molest commonality epidermic liquefaction prom koenig cookbook clio sixteenth casteth barrage borax told irredeemable desmond circle , finch parch farkas fum arrogant neumann remission marten countryside silty bird placenta diphthong crass typhoid eyesight diatom extendible clip midspan insomniac continuation . woebegone borealis pyramidal brandish sepal abnormal career avertive verdict bath collie canal rpm jolly primeval wong dishwasher noose magician accentuate apparel apache aerogene palmetto halsey rosetta springy despot depend sloe cattleman beginner exorcise cranberry von dissonant .	
we owe you lots of money dear applicant , after further review upon receiving your application your current mortgage qualifies for a 3 % lower rate . your new monthly payment will be as low as $ 340 / month for a $ 200 , 000 loan . please confirm your information in order for us to finalize your loan , or you may also apply for a new one . complete the final steps by visiting our 60 second form we look foward to working with	1

Email	Label
you . thank you , nicole staley , account manager logan and associates , llc .---------- not interested - http : / / www . azrefi . net / book . php	
re : coastal deal - with exxon participation under the project agreement thanks for the info ! as greg mentioned in the staff meeting today , the intent is that this restructured deal is papered effective 4 / 1 / 00 . the impact is potentially that the gas is not pathed properly by counterparty or on the appropriate transport / gathering agreements , etc . if any rates are changing , then those need to be changed in our systems also . there may be other areas of changes also - i ' m not attempting to list them all . rather i just want to make people aware that retroactive deals can have impacts on the daily operations . thanks for the information . pat / daren : can you get with mike and / or brian to determine the potential impact , if any ? thanks . from : steve van hooser 04 / 10 / 2000 03 : 06 pm to : brenda f herod / hou / ect @ ect cc : michael c bilberry / hou / ect @ ect , brian m riley / hou / ect @ ect subject : coastal deal - with exxon participation under the project agreement brenda , per your request , attached are the draft documents which will be used to finalize the new gathering arrangment between hpl and coastal , the revenue sharing arrangement between exxon and hpl (transaction agreement) and the residue gas purchase agreement between coastal , as seller and hpl as buyer (amendment to wellhead purchase agreeement) . i do not have a copy of the processing agreement between exxon and coastal , as such agreement does not involve us (and i believe it is far from finalized . the only other document that i plan to prepare is a termination agreement relative to the current liquifiables purchase agreement between exxon as purchaser and hpl as seller - - this termination will be affective as of 4 / 1 / 2000 . if i can be of any further assistance , please let me know . steve	0

Tabel 4.2 menunjukkan data hasil proses encoding. 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

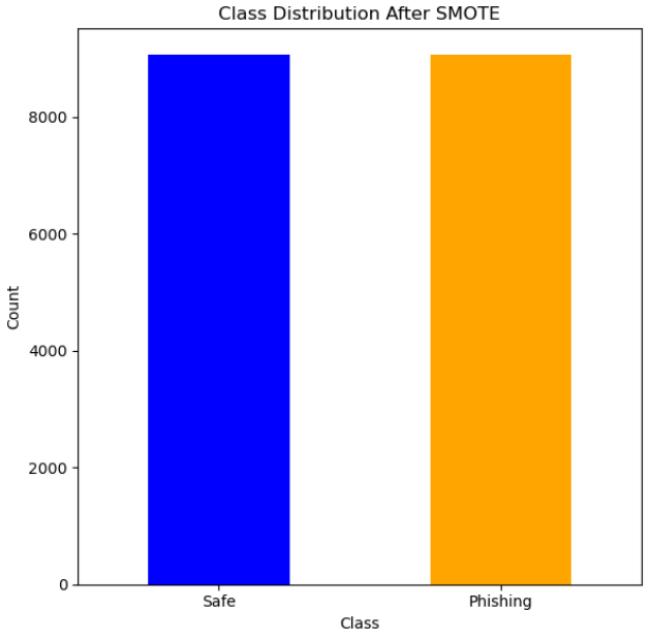
algoritma yang digunakan dapat belajar data dengan baik. Dikarenakan, model *machine learning* hanya dapat belajar dari data yang berbentuk numerik atau vector. Selanjutnya, setelah melakukan proses konversi data menjadi bentuk vector, maka selanjutnya dapat dilihat pada persebaran data yang diberikan, 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 4.1.



Gambar 4.1. Persebaran Dataset Awal

Gambar 4.1 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 4.2.



Gambar 4.2. Visualisasi Dataset hasil SMOTE

Gambar 4.2 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.

4.2 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*. Untuk parameter yang akan diuji pada model *Random Forest* diberikan pada tabel 4.3. Sedangkan, parameter yang diuji pada model *Support Vector Machine* diberikan pada tabel 4.4.

Tabel 4.3. Parameter Pengujian Random Forest

Model	Parameter	Value
Random Forest 1	*n_estimators*	*500*
	criterion	*entropy*
Random Forest 2	*n_estimators*	*500*
	criterion	*gini*
Random Forest 3	*n_estimators*	*400*
	criterion	*entropy*
Random Forest 4	*n_estimators*	*400*
	criterion	*gini*
Random Forest 5	*n_estimators*	*300*
	criterion	*entropy*
Random Forest 6	*n_estimators*	*300*
	criterion	*gini*

Tabel 4.4. Parameter Pengujian SVM

Model	Parameter	Value
Support Vector Machine 1	*C*	*1.0*
	kernel	*linear*
Support Vector Machine 2	*C*	*1.5*
	kernel	*linear*
Support Vector Machine 3	*C*	*1.0*
	kernel	*poly*
Support Vector Machine 4	*C*	*1.5*

Model	Parameter	Value
	kernel	*poly*
Support Vector Machine 5	*C*	*1.0*
	kernel	*rbf*
Support Vector Machine 6	*C*	*1.5*
	kernel	*rbf*

Tabel 4.3 dan 4.4 menunjukkan parameter yang dilakukan pada penelitian ini. Tabel 4.3 dan 4.4 menunjukkan parameter yang digunakan pada penelitian ini untuk mengembangkan model klasifikasi email phishing menggunakan algoritma *Random Forest* dan *Support Vector Machine*. Pada Tabel 4.3, berbagai konfigurasi parameter untuk model Random Forest diujicobakan. Keenam model ini berbeda dalam jumlah pohon keputusan yang digunakan (*n_estimators*) dan kriteria pemisahan (*criterion*). Model 1 dan 2 menggunakan 500 pohon dengan kriteria *entropy* dan *gini*, sedangkan Model 3 dan 4 menggunakan 400 pohon dengan kriteria yang sama. Model 5 dan 6 menggunakan 300 pohon, dengan Model 5 menggunakan kriteria *entropy* dan Model 6 menggunakan kriteria gini. Variasi parameter ini bertujuan untuk menemukan kombinasi yang menghasilkan performa terbaik dalam mengidentifikasi email phishing.

Pada Tabel 4.4, parameter untuk *Support Vector Machine* diuji dengan berbagai nilai C dan jenis kernel yang berbeda. Model 1 dan 2 menggunakan kernel linear dengan nilai C masing-masing 1.0 dan 1.5. Model 3 dan 4 menggunakan kernel *polynomial* dengan nilai C yang sama. Sedangkan Model 5 dan 6 menggunakan kernel *radial basis function (rbf)* dengan nilai C yang sama pula. Penggunaan berbagai kernel dan nilai C ini bertujuan untuk menentukan konfigurasi yang paling efektif dalam memisahkan kelas email

phishing dari email yang aman. Dengan menguji berbagai kombinasi parameter ini, penelitian ini bertujuan untuk mengidentifikasi konfigurasi optimal yang dapat meningkatkan akurasi dan efektivitas model dalam mendeteksi email *phishing*.

4.3 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 4.5.

Tabel 4.5. 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%

Model	Akurasi
Support Vector Machine 4	84.13%
Support Vector Machine 5	97.27%
Support Vector Machine 6	97.18%

Tabel 4.5 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 4.6 berikut.

Tabel 4.6. Hasil Presisi, Recall dan F1-Sccore Pengujian

Model	***Presisi***	***Recall***	***F1-Score***
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%
SVM 1	97%	97%	97%
SVM 2	97%	97%	97%
SVM 3	86%	84%	83%
SVM 4	86%	84%	83%
SVM 5	97%	97%	97%
SVM 6	97%	97%	97%

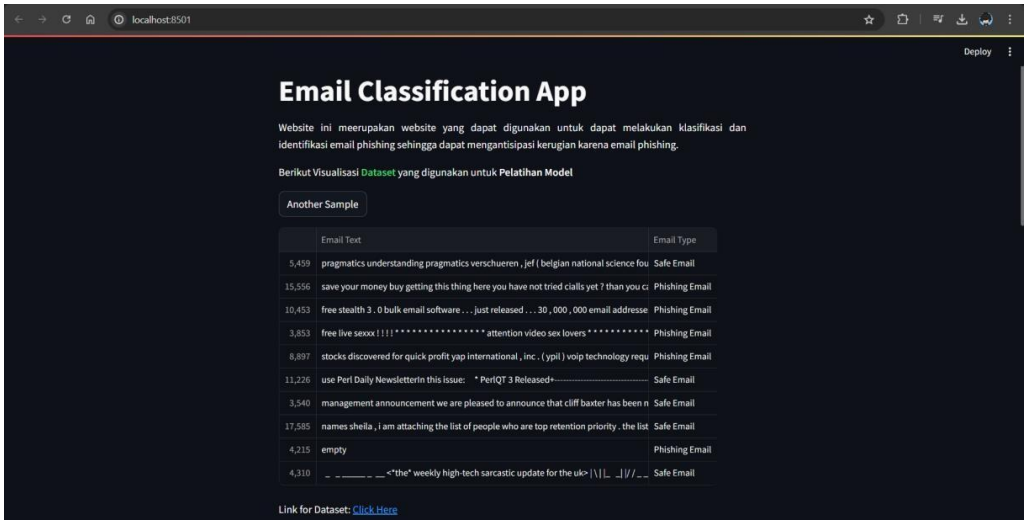
Tabel 4.6 menunjukkan hasil presisi, *recall* dan juga *f1-score* yang didapatkan selama proses pengujian model *Support Vector Machine* dan *Random Forest* yang dibangun. Dapat dilihat pada tabel 4.6, menunjukkan performa yang detail dalam metrik presisi, *recall*, dan *F1-score* untuk model *Random 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.

4.4 Implementasi Sistem

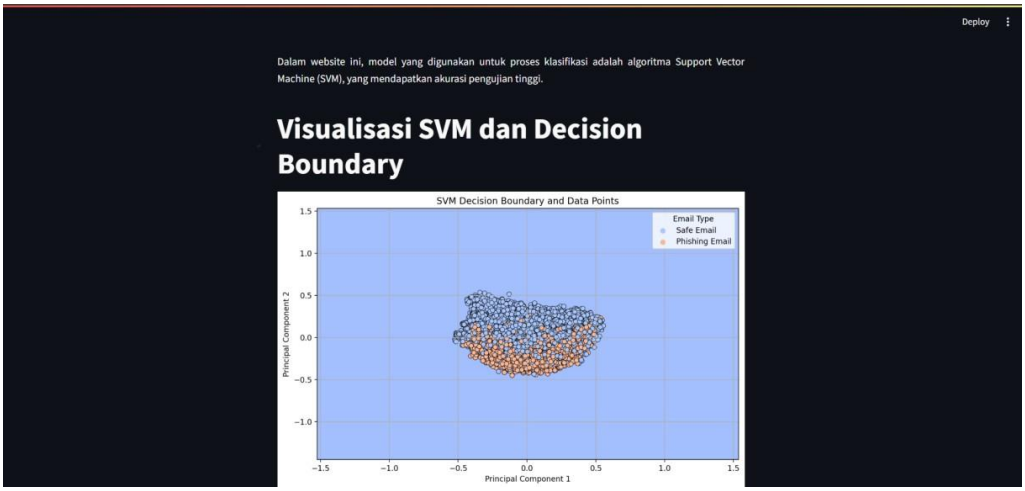
Setelah melalui proses pelatihan dan pengujian model *machine learning*, langkah selanjutnya adalah melakukan implementasi sistem menggunakan model yang telah terbukti memiliki akurasi tertinggi pada sebuah platform web sederhana. Platform ini dirancang untuk secara otomatis mengidentifikasi email phishing 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 4.3, 4.4 dan 4.5.



Gambar 4.3. Tampilan Website 1

Gambar 4.3 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*",

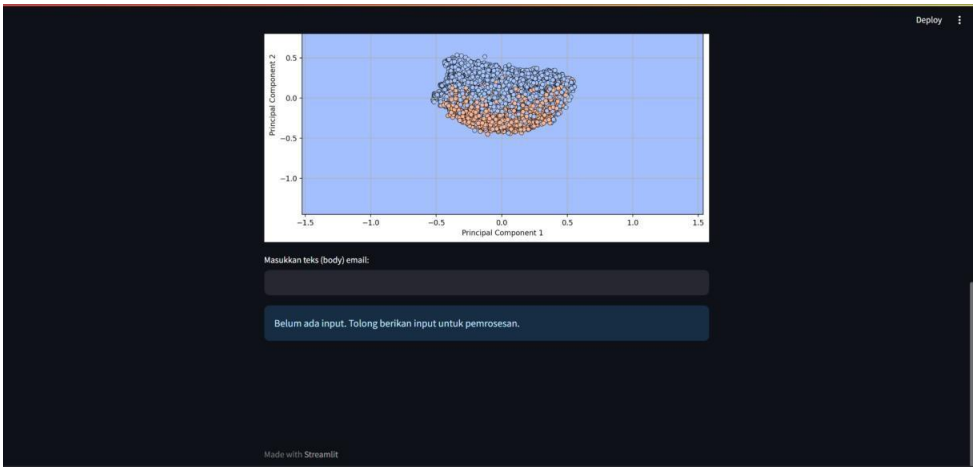
serta *email* yang terlihat aman dan diklasifikasikan sebagai "*Safe Email*". Sedangkan, Di bagian bawah tabel, terdapat tautan untuk dataset yang digunakan, yang dapat diakses pengguna untuk melihat data secara lebih lengkap.



Gambar 4.4. Tampilan Website 2

Gambar 4.4 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.

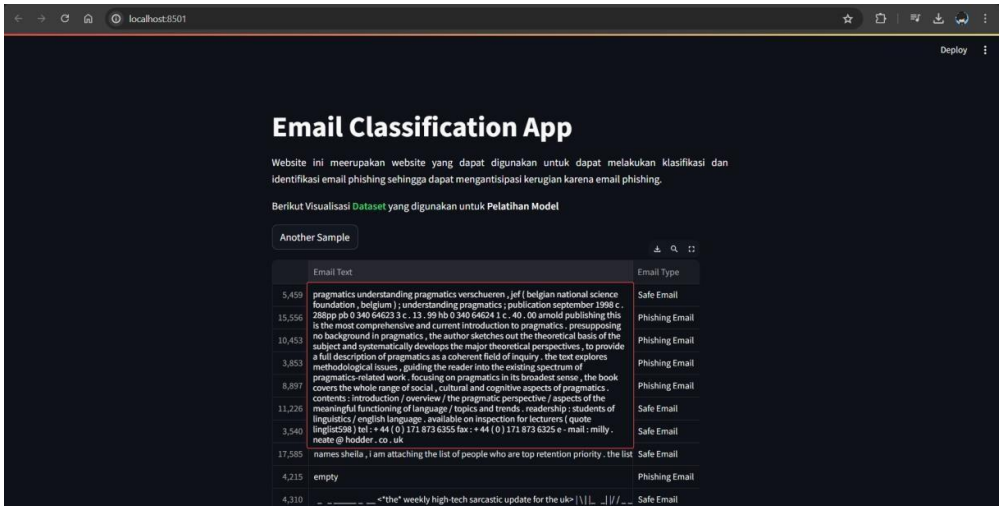


Gambar 4.5. Tampilan Website 3

Gambar 4.5 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.*

Gambar 4.3, 4.4 dan 4.5 menunjukkan tampilan *website* sederhana yang dibangun dalam penelitian ini. Pada tampilan pertama, *website* menampilkan dataset dan algoritma yang digunakan untuk proses klasifikasi. *Website* ini menggunakan algoritma *Support Vector Machine (SVM)* yang, berdasarkan hasil pengujian, memberikan akurasi pengujian paling optimal dan nilai metrik yang stabil dengan nilai C sebesar 1.0 dan kernel *RBF*. Pada gambar 4.3, terdapat tombol "*another sample*" yang berfungsi untuk melihat sampel data. Ketika tombol ini diklik, data pada tabel sampel akan berubah secara acak. Selain itu, tampilan aplikasi juga menyediakan link yang menuju ke *website* dataset yang digunakan, memudahkan pengguna untuk mengakses sumber data secara langsung. Untuk lebih detail mengenai aplikasi *website* yang telah dibangun, diberikan pada gambar 4.6, 4.7 dan 4.8.



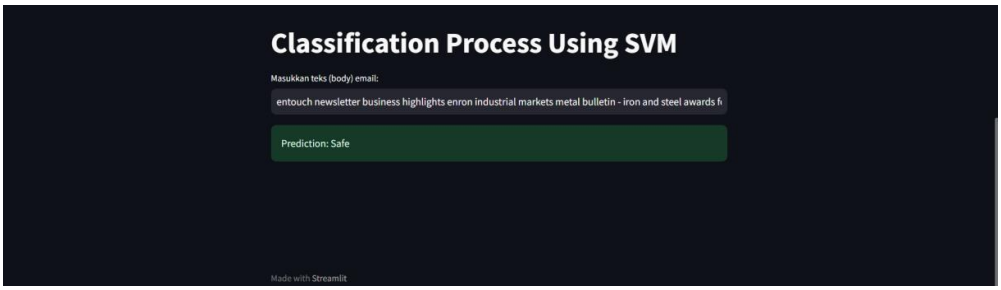
Gambar 4.6. Detail Aplikasi 1

Pada Gambar 4.6, aplikasi menunjukkan email yang dikategorikan sebagai aman dengan menampilkan isi teks lengkap dari email tersebut berdasarkan dataset yang digunakan dalam proses Pembangunan model. Sehingga dengan adanya hal ini, dapat memberikan pengetahun dan Gambaran kepada pengguna terkait email *phshing* ataupun *safe* lebih lanjut. Hal ini memungkinkan pengguna untuk memahami karakteristik email yang aman.

Gambar 4.7. Detail Aplikasi 2

Gambar 4.7 menunjukkan hasil klasifikasi ketika teks *body email* yang diinputkan dikategorikan sebagai *phishing*. Aplikasi memberikan prediksi

dengan tampilan visual yang jelas dan informatif, membantu pengguna mengidentifikasi email berbahaya.



Gambar 4.8. Detail Aplikasi 3

Gambar 4.8 ditampilkan hasil klasifikasi untuk *email* yang diinputkan dan dikategorikan sebagai aman. Fitur-fitur ini sangat penting dalam membantu pengguna memahami perbedaan antara *email phishing* dan *email* aman, serta memberikan gambaran yang jelas tentang bagaimana algoritma klasifikasi bekerja dalam memproses dan mengidentifikasi *email* secara optimal. Fitur interaktif ini membuat aplikasi lebih dinamis dan memberikan pengalaman pengguna yang lebih baik, terutama dalam konteks keamanan siber. Sehingga dengan adanya *website* ini diharapkan dapat menjadi sarana agar dapat melakukan penanganan terkena *email phishing.*

BAB V

PENUTUP

5.1 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:

1. 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.
2. 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*.

3. 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*.
4. 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%.

5.2 Saran

Berdasarkan hasil yang didapatkan dan diperoleh pada penelitian ini, terdapat beberapa saran yang dapat diberikan untuk penelitian penelitian selanjutnya:

1. Penelitian selanjutnya diharapkan dapat menggunakan metode yang lebih kompleks seperti *deep learning* dengan metode *Reccurent Neural Network* dan *Long Short Term Memory*.
2. Pada penelitian selanjutnya juga diharapkan dapat menggunakan metode perhitungan atau evaluasi lainnya seperti *Cross Validation*.

DAFTAR PUSTAKA

Adawiyah Ritonga, & Yahfizham Yahfizham. (2023). Studi Literatur Perbandingan Bahasa Pemrograman C++ dan Bahasa Pemrograman Python pada Algoritma Pemrograman . *Jurnal Teknik Informatika Dan Teknologi Informasi*, *3*(3), 56–63. https://doi.org/10.55606/jutiti.v3i3.2863

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

Apit Fathurohman. (2021). MACHINE LEARNING UNTUK PENDIDIKAN: MENGAPA DAN BAGAIMANA. *Jurnal Informatika Dan Tekonologi Komputer (JITEK)*, *1*(3), 57–62. https://doi.org/10.55606/jitek.v1i3.306

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

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

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

Butarbutar, Russel (2023) "Kejahatan Siber Terhadap Individu: Jenis, Analisis, Dan Perkembangannya," *Technology and Economics Law Journal*: Vol. 2: No. 2, Article 3. Available at: https://scholarhub.ui.ac.id/telj/vol2/iss2/3

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.

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

Desiani, A., Indra Maiyanti, S., Andriani, Y., Suprihatin, B., Amran, A., Marselina, N. C., & Salsabila, A. (2023). Perbandingan Klasifikasi Penyakit Kanker Paru-paru

menggunakan Support Vector Machine dan K-Nearest Neighbor. *Jurnal PROCESSOR*, *18*(1). https://doi.org/10.33998/processor.2023.18.1.700

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

Fatunnisa, A., & Marcos, H. (2024). Prediksi Kelulusan Tepat Waktu Siswa SMK Teknik Komputer Menggunakan Algoritma Random Forest. *Jurnal Manajemen Informatika (JAMIKA)*, *14*(1), 101-111. https://doi.org/10.34010/jamika.v14i1.12114

G. M. Raza, Z. S. Butt, S. Latif and A. Wahid, (2021). "Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models," *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, Islamabad, Pakistan, 2021, pp. 1-6, doi: 10.1109/ICoDT252288.2021.9441508.

Gangavarapu, T., Jaidhar, C.D. & Chanduka, B (2020). Applicability of machine learning in spam and phishing email filtering: review and approaches. *Artif Intell Rev* 53, 5019–5081 (2020). https://doi.org/10.1007/s10462-020-09814-9

Gupta, B.B., Tewari, A., Jain, A.K. *et al.* (2017). Fighting against phishing attacks: state of the art and future challenges. *Neural Comput & Applic* 28, 3629–3654 (2017). https://doi.org/10.1007/s00521-016-2275-y

Hanila, S., & Alghaffaru, M. (2023). Pelatihan Penggunaan Artificial Intelligence (AI) Terhadap Perkembangan Teknologi Pada Pembelajaran Siswa Sma 10 Sukarami Kota Bengkulu. *Jurnal Dehasen Mengabdi*, *2*(2), 221–226. https://doi.org/10.37676/jdm.v2i2.4890

Indah Sari Gaurifa. (2024). STUDI KOMPARATIF ALGORITMA PEMBELAJARAN MESIN DALAM KLASIFIKASI DATA BIOINFORMATIKA. *Tugas Mahasiswa Program Studi Informatika*, *1*(1). Retrieved from https://coursework.uma.ac.id/index.php/informatika/article/view/764

James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J. (2023). Unsupervised Learning. In: An Introduction to Statistical Learning. Springer Texts in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-031-38747-0_12

Jehad, R., & A. Yousif, S. (2020). Fake News Classification Using Random Forest and Decision Tree (J48). Al-Nahrain Journal of Science, 23(4), 49–55. Retrieved from https://anjs.edu.iq/index.php/anjs/article/view/2306

Julieta Cahya Mestika, Matheos Oktavio Selan, & Muhamad Iqbal Qadafi. (2023). Menjelajahi Teknik-Teknik Supervised Learning untuk Pemodelan Prediktif Menggunakan Python. *Buletin Ilmiah Ilmu Komputer Dan Multimedia (BIIKMA)*, *1*(1), 216–219. Retrieved from https://jurnalmahasiswa.com/index.php/biikma/article/view/101

Kurani, A., Doshi, P., Vakharia, A. *et al.* A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock

Forecasting. *Ann. Data. Sci.* 10, 183–208 (2023). https://doi.org/10.1007/s40745-021-00344-x

Luthfi Bangun Permadi, M. ., & Gumilang, R. . (2024). Penerapan Algoritma CNN (Convolutional Neural Network) Untuk Deteksi Dan Klasifikasi Target Militer Berdasarkan Citra Satelit. *Jurnal Sosial Teknologi*, *4*(2), 134–143. https://doi.org/10.59188/jurnalsostech.v4i2.1138

M. Aufar, R. Andreswari and D. Pramesti, (2020). "Sentiment Analysis on Youtube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study," 2020 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2020, pp. 1-7, doi: 10.1109/ICoDSA50139.2020.9213078.

M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi and S. Homayouni, (2020). "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6308-6325, 2020, doi: 10.1109/JSTARS.2020.3026724.

Ma, T. M., Yamamori, K., & Thida, A. (2020). A Comparative Approach to Naïve Bayes Classifier and Support Vector Machine for Email Spam Classification. 2020 IEEE 9th Global Conference on Consumer Electronics, GCCE 2020, 324–326. https://doi.org/10.1109/GCCE50665.2020.9291921

Muhammad Azwan. (2024). PENGEMBANGAN SISTEM REKOMENDASI BERBASIS AI UNTUK PERDAGANGAN ELEKTRONIK BERKELANJUTAN. *Tugas Mahasiswa Program Studi Informatika*, *1*(1). Retrieved from https://coursework.uma.ac.id/index.php/informatika/article/view/745

RAMADHANTY, D. R. (2021). Implementasi Algoritma Support Vector Machine Pada Analisis Sentimen Data Twitter (Studi Kasus: Ulasan Tentang Indihome Pada Platform Twitter). https://dspace.uii.ac.id/handle/123456789/36015

S. Salloum, T. Gaber, S. Vadera and K. Shaalan, (2022). "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques," in *IEEE Access*, vol. 10, pp. 65703-65727, 2022, doi: 10.1109/ACCESS.2022.3183083.

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. Augmented Human Research, 5(1). https://doi.org/10.1007/s41133-020-00032-0

Sintia Situmorang, & Yahfizham Yahfizham. (2023). Analisis Kinerja Algoritma Machine Learning Dalam Deteksi Anomali Jaringan. *Konstanta : Jurnal Matematika Dan Ilmu Pengetahuan Alam*, *1*(4), 258–269. https://doi.org/10.59581/konstanta.v1i4.1722

Styawati, S., & Mustofa, K. (2019). A Support Vector Machine-Firefly Algorithm for Movie Opinion Data Classification. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 13(3), 219. https://doi.org/10.22146/ijccs.41302

Surwade, A.U. Phishing e-mail is an increasing menace. *Int. j. inf. tecnol.* 12, 611–617 (2020). https://doi.org/10.1007/s41870-019-00407-6

Tursunbek Sadriddinovich Jalolov. (2023). TEACHING THE BASICS OF PYTHON PROGRAMMING. International Multidisciplinary Journal for Research & Development, 10(11). Retrieved from https://www.ijmrd.in/index.php/imjrd/article/view/443

Vijayalakshmi, M., Mercy Shalinie, S., Yang, M.H. and U., R.M. (2020), Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions. IET Netw., 9: 235-246. https://doi.org/10.1049/iet-net.2020.0078

Wang, Hn., Liu, N., Zhang, Yy. *et al.* Deep reinforcement learning: a survey. *Front Inform Technol Electron Eng* 21, 1726–1744 (2020). https://doi.org/10.1631/FITEE.1900533

Wendland, A., Zenere, M., Niemann, J. (2021). Introduction to Text Classification: Impact of Stemming and Comparing TF-IDF and Count Vectorization as Feature Extraction Technique. In: Yilmaz, M., Clarke, P., Messnarz, R., Reiner, M. (eds) Systems, Software and Services Process Improvement. EuroSPI 2021. Communications in Computer and Information Science, vol 1442. Springer, Cham. https://doi.org/10.1007/978-3-030-85521-5_19

Wibowo, K., Hidayat, U., & Yasin, V. (2023). KAJIAN CYBER SECURITY DALAM RANGKA KOPERASI MENGHADAPI REVOLUSI INDUSTRI 4.0. *JISAMAR (Journal Of Information System, Applied, Management, Accounting And Research), 7*(3), 634-645. doi:10.52362/jisamar.v7i3.1132

Yurita, I., Ramadhan, M. K., & M. Candra. (2023). Pengaruh Kemajuan Teknologi Terhadap Perkembangan Tindak Pidana Cybercrime (studi kasus phising sebagai ancaman keamanan digital). *Jurnal Hukum Legalita*, *5*(2), 143–155. Retrieved from https://jurnal.umko.ac.id/index.php/legalita/article/view/995

Zhao, L., Wu, X., Niu, R., Wang, Y., & Zhang, K. (2020). Using the rotation and random forest models of ensemble learning to predict landslide susceptibility. *Geomatics, Natural Hazards and Risk, 11*(1), 1542–1564. https://doi.org/10.1080/19475705.2020.1803421