

**OPTIMASI ALGORITMA *SUPPORT VECTOR MACHINE* DAN
RANDOM FOREST TERHADAP DETEKSI SPAM EMAIL**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

MUHAMMAD ALI IMRAN KAHFI

19.11.3274

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2024

**OPTIMASI ALGORITMA *SUPPORT VECTOR MACHINE* DAN
RANDOM FOREST TERHADAP DETEKSI SPAM EMAIL**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

MUHAMMAD ALI IMRAN KAHFI

19.11.3274

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA**

YOGYAKARTA

2024

SKRIPSI

OPTIMASI ALGORITMA *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* TERHADAP DETEKSI SPAM EMAIL

yang disusun dan diajukan oleh

MUHAMMAD ALI IMRAN KAHFI

19.11.3274

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 24 September 2024

Dosen Pembimbing,



Anna Baita, M.Kom
NIK. 190302290

HALAMAN PENGESAHAN
SKRIPSI
OPTIMASI ALGORITMA *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* TERHADAP DETEKSI SPAM EMAIL

yang disusun dan diajukan oleh

MUHAMMAD ALI IMRAN KAHFI

19.11.3274

Telah dipertahankan di depan Dewan Penguji
pada tanggal 24 September 2024

Susunan Dewan Penguji

Nama Penguji

Arifiyanto Hadinegoro, S.Kom, MT
NIK. 190302289

Yoga Pristyanto, S.Kom., M.Eng.
NIK. 190302412

Anna Baita, M.Kom
NIK. 190302290

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 24 September 2024

DEKAN FAKULTAS ILMU KOMPUTER



Hanif Al Fatta, S.Kom., M.Kom., Ph.D.
NIK. 190302096

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : MUHAMMAD ALI IMRAN KAHFI
NIM : 19.11.3274

Menyatakan bahwa Skripsi dengan judul berikut:

OPTIMASI ALGORITMA SUPPORT VECTOR MACHINE DAN RANDOM FOREST TERHADAP DETEKSI SPAM EMAIL

Dosen Pembimbing : Anna Baita, M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 24 September 2024

Yang Menyatakan,



Muhammad Ali Imran Kahfi

HALAMAN PERSEMBAHAN

Puji dan syukur penulis ucapkan kepada Allah SWT atas anugerah dan nikmat yang tak terkira sehingga penulis dapat menyelesaikan karya tulis ini. Pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih yang sebesar- besarnya kepada :

1. Kedua orang tua saya Ibu dan Bapak yang selalu memberikan doa, dukungan dan motivasi dalam pengerjaan skripsi ini.
2. Seluruh sahabat saya Terima kasih atas segala doa dan dukungannya untuk saya.
3. Ibu Anna Baita M, Kom selaku pembimbing, yang banyak memberikan arahan, masukan dan motivasi kepada saya
4. Bapak Prof. Dr. M. Suyanto, M.M. selaku Rektor UNIVERSITAS AMIKOM Yogyakarta.

KATA PENGANTAR

Bagian ini berisi pernyataan resmi yang ingin disampaikan oleh penulis kepada pihak lain, misalnya ucapan terima kasih kepada Dosen Pembimbing, Tim Dosen Penguji, dan semua pihak yang terkait dalam penyelesaian skripsi termasuk orang tua dan penyandang dana.

Nama harus ditulis secara lengkap termasuk gelar akademik dan harus dihindari ucapan terima kasih kepada pihak yang tidak terkait. Bahasa yang digunakan harus mengikuti kaidah bahasa Indonesia yang baku.

Bagian ini tidak perlu dituliskan hal-hal yang bersifat ilmiah. Kata Pengantar diakhiri dengan mencantumkan kota dan tanggal penulisan diikuti di bawahnya dengan **kata “Penulis” tanpa perlu menyebutkan nama dan tanda tangan.**

Yogyakarta, 24 September 2024

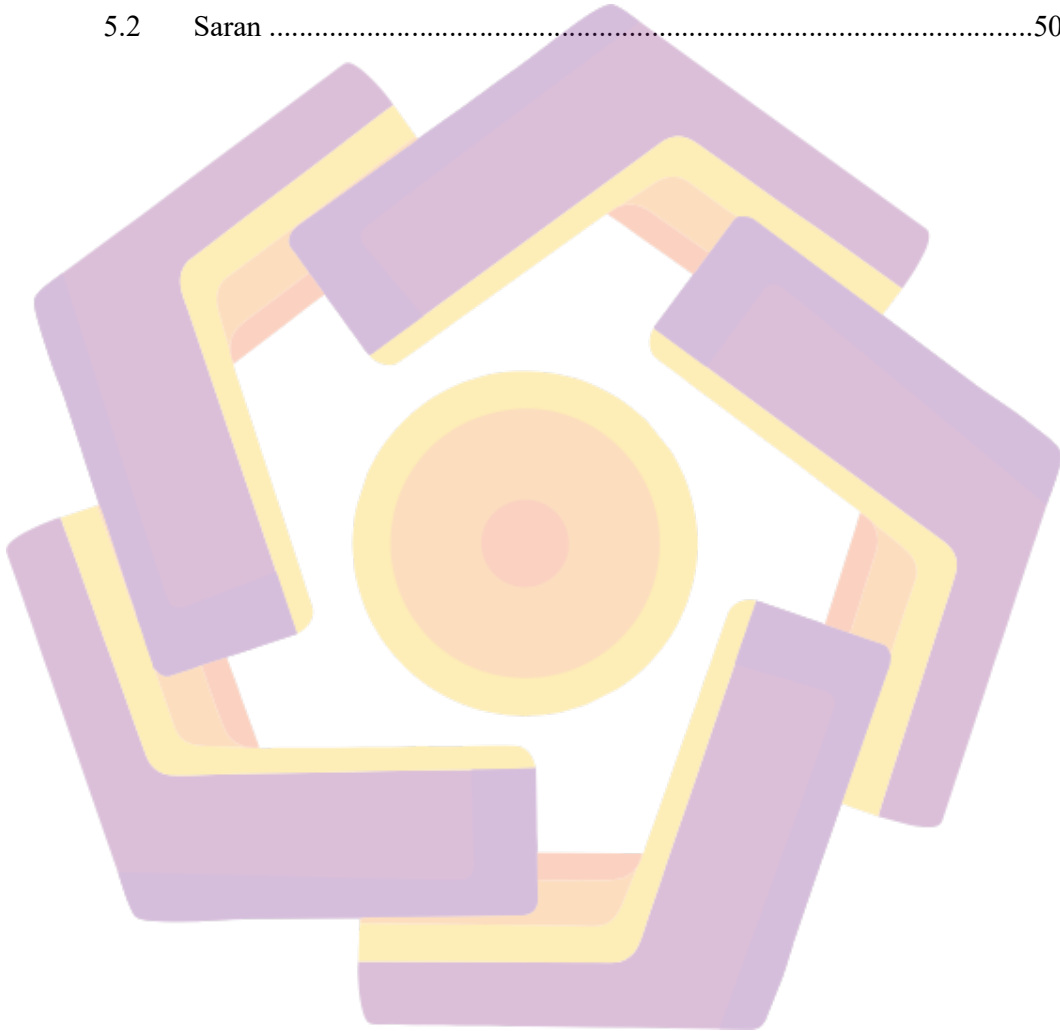
Penulis

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI.....	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vi
DAFTAR ISI.....	vii
DAFTAR TABEL.....	x
DAFTAR GAMBAR.....	xi
INTISARI	xiii
<i>ABSTRACT</i>	xiv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian.....	3
1.6 Sistematika Penulisan	4
BAB II TINJAUAN PUSTAKA	6
2.1 Studi Literatur	6
2.2 Dasar Teori.....	11
2.2.1 Spam Email.....	11
2.2.2 <i>Exploratory Data Analysis</i>	12

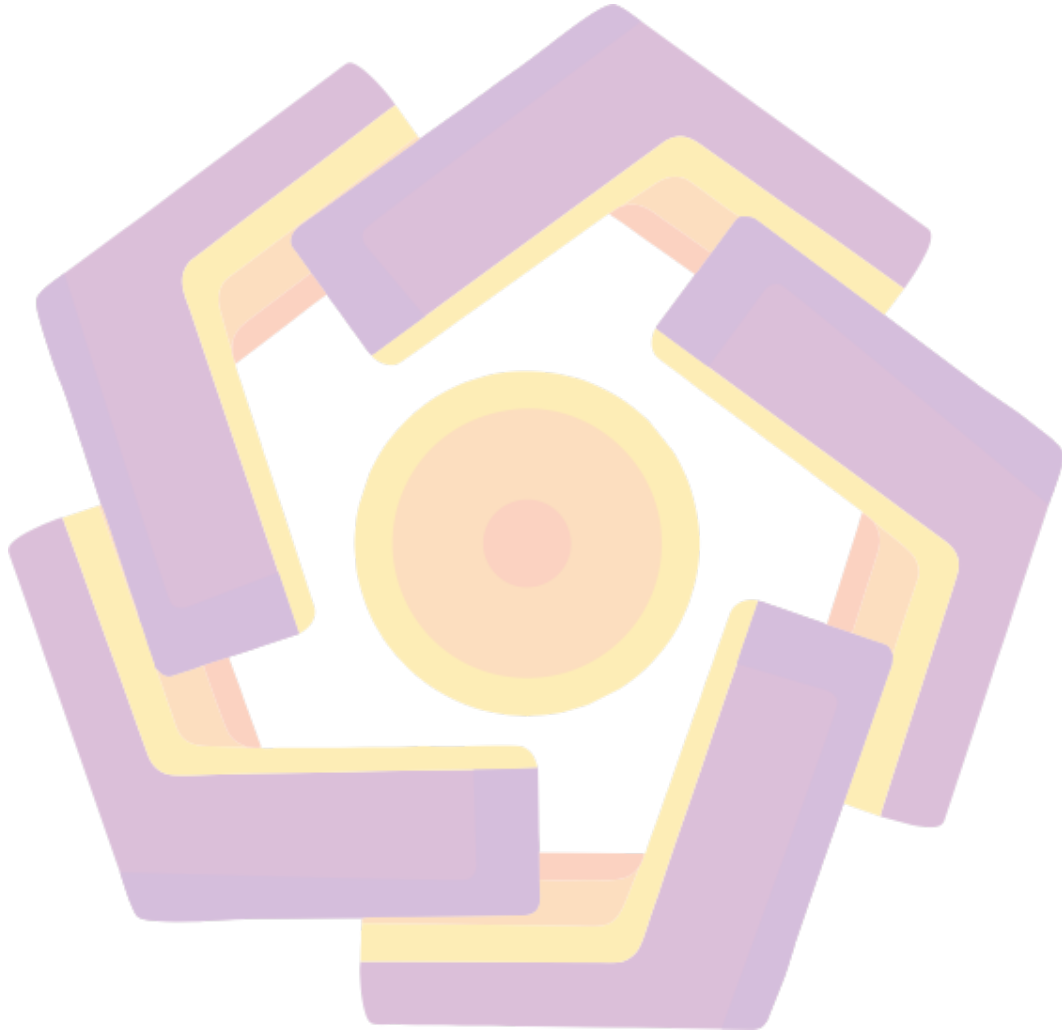
2.2.3	Tahapan <i>Text Mining</i>	13
2.2.4	<i>Hyperparameter Optimization</i>	14
2.2.3	<i>Machine Learning</i>	17
2.2.4	<i>Support Vector Machine (SVM)</i>	19
2.2.5	<i>Random Forest</i>	20
2.2.6	Matriks Pengukuran	24
2.2.7	AUC ROC	25
BAB III METODE PENELITIAN		27
3.1	Objek Penelitian	27
3.2	Alur Penelitian	27
3.3	Alat dan Bahan	30
BAB IV HASIL DAN PEMBAHASAN		31
4.1	Pengumpulan Data	31
4.2	<i>Data Preprocessing</i>	31
4.3	<i>Exploratory Data Analysis(EDA)</i>	33
4.4	<i>Text Preprocessing</i>	34
4.5	<i>Feature Extraxtion</i>	37
4.6	<i>Hyperparameter Optimization</i>	38
4.6.1	Optimasi <i>Support Vector Machine</i>	38
4.6.2	Optimasi <i>Random Forest</i>	39
4.7	Model Training	41
4.7.1	<i>Support Vector Machine</i>	41
4.7.2	<i>Random Forest</i>	41
4.8	Evaluasi dan Validasi	42
4.8.1	Evaluasi dan Validasi <i>Support Vector Machine</i>	42

4.8.2	Evaluasi dan Validasi <i>Random Forest</i>	45
4.9	Analisis Hasil	48
BAB V PENUTUP		50
5.1	Kesimpulan	50
5.2	Saran	50



DAFTAR TABEL

Tabel 2.1 Keaslian Penelitian	7
Tabel 2.2 <i>Matriks Confusion</i>	24
Tabel 4.1 Perbandingan Optimasi Dan Tanpa Optimasi	49



DAFTAR GAMBAR

Gambar 2.1 Tahapan EDA	13
Gambar 2.2. Ilustrasi Pengaplikasian Hyperparameter Optimization	15
Gambar 2.3 Prinsip Kerja SVM	20
Gambar 2.4 Prinsip Kerja Random Forest	23
Gambar 3.1 Alur Penelitian	27
Gambar 4.1 Dataset	31
Gambar 4.2 Pengecekan Data Null	31
Gambar 4.3 Pengecekan Data Duplikasi	32
Gambar 4.4 Penghapusan Data Duplikasi	32
Gambar 4.5 Pelabelan Variabel <i>Category</i>	32
Gambar 4.6 Persen Ham dan Spam	33
Gambar 4.7 Pairplot Data Target	33
Gambar 4.8 Hasil <i>Correlation</i>	34
Gambar 4.9 <i>Correlation Heatmap</i>	34
Gambar 4.10 Proses <i>Text Preprocessing</i>	35
Gambar 4.11 Penambahan Kolom Baru	36
Gambar 4.14 Proses CountVectorizer,dan TfidfVectorizer	37
Gambar 4.15 Split Data	38
Gambar 4.16 Standarisasi Data Training	38
Gambar 4.17 Optimasi <i>Support Vector Machine</i>	39
Gambar 4.18 Optimasi <i>Random Forest</i>	40
Gambar 4.19 Proses Model Training SVM Tanpa Optimasi	41
Gambar 4.20 Proses Model <i>Training</i> SVM Dengan Optimasi	41
Gambar 4.21 Proses Model Training <i>Random Forest</i> Tanpa Optimasi	42
Gambar 4.22 Proses Model Training <i>Random Forest</i> Dengan Optimasi	42
Gambar 4.23 Proses <i>Confusion Matriks</i> SVM Tanpa Optimasi	43
Gambar 4.24 Hasil <i>Confusion Matriks</i> SVM Tanpa Optimasi	43
Gambar 4.25 Proses <i>Confusion Matriks</i> SVM Dengan Optimasi	43
Gambar 4.26 Hasil <i>Confusion Matriks</i> SVM Dengan Optimasi	43

Gambar 4.27 Proses AUC ROC SVM Tanpa Optimasi	44
Gambar 4.28 Hasil AUC ROC SVM Tanpa Optimasi	44
Gambar 4.30 Hasil ROC dan AUC SVM Dengan Optimasi	45
Gambar 4.31 Proses <i>Confusion Matriks Random Forest</i> Tanpa Optimasi	45
Gambar 4.32 Hasil <i>Confusion Matriks Random Forest</i> Tanpa Optimasi	45
Gambar 4.33 Proses <i>Confusion Matriks SVM</i> Dengan Optimasi	46
Gambar 4.34 Hasil <i>Confusion Matriks Random Forest</i> Dengan Optimasi	46
Gambar 4.35 Proses AUC ROC <i>Random Forest</i> Tanpa Optimasi	46
Gambar 4.36 Hasil AUC ROC <i>Random Forest</i> Tanpa Optimasi	47
Gambar 4.38 Hasil AUC ROC <i>Random Forest</i> Dengan Optimasi	48



INTISARI

Email merupakan alat komunikasi yang sangat penting di era digital, digunakan oleh individu dan organisasi untuk berbagai keperluan. Namun, dengan meningkatnya penggunaan terdapat kendala terkait dengan spam email. Spam email tidak hanya mengganggu penerima tetapi juga membebani infrastruktur penyedia layanan email. Ancaman seperti tautan berbahaya, serangan phishing, dan malware dapat merusak perangkat pengguna, mengurangi efisiensi komunikasi, dan mengganggu produktivitas. Oleh karena itu, pengelolaan spam email menjadi perhatian penting dalam lingkungan digital yang semakin kompleks.

Untuk mengatasi masalah tersebut pada penelitian ini dilakukan pendekatan berbasis machine learning yaitu *Support Vector Machine* (SVM) dan *Random Forest* untuk menunjukkan kinerja yang signifikan dalam optimasi terhadap spam email dan termasuk deteksi spam email. Penelitian ini bertujuan untuk mengoptimalkan algoritma SVM dan *Random Forest* dalam deteksi spam email menggunakan *Hyperparameter Optimization* dengan *RandomizedSearchCV*. Fokus penelitian meliputi penentuan parameter optimal menggunakan *Support Vector Machine* (SVM), seperti *kernel*, *C*, dan *gamma*, serta penyesuaian jumlah pohon dan fitur dalam *Random Forest*.

Selain itu, metode pemilihan fitur yang relevan juga akan diperhatikan. Evaluasi kinerja akan dilakukan dengan menggunakan metrik seperti akurasi, presisi, recall, dan F1-score dan AUC ROC. Untuk penggunaan optimasi berpengaruh pada model training *Support Vector Machine* dan *Random Forest* yaitu dengan *Support Vector Machine* tanpa optimasi menghasilkan akurasi sebesar 95 % dan dengan optimasi menghasilkan akurasi sebesar 96% dan *Random Forest* tanpa optimasi menghasilkan akurasi sebesar 97% dan dengan optimasi menghasilkan akurasi sebesar 98%.

Kata kunci: *Support Vector Machine* (SVM), *Random Forest*, Email Spam, *Hyperparameter Optimization*

ABSTRACT

Email is a very important communication tool in the digital era, used by individuals and organizations for various purposes. However, with increasing use there are problems related to email spam. Email spam not only annoys recipients but also burdens the infrastructure of email service providers. Threats such as malicious links, phishing attacks, and malware can damage users' devices, reduce communication efficiency, and disrupt productivity. Therefore, managing email spam is an important concern in an increasingly complex digital environment.

To overcome this problem, this research used a machine learning-based approach, namely Support Vector Machine (SVM) and Random Forest, to show significant performance in optimizing email spam and including email spam detection. This research aims to optimize the SVM and Random Forest algorithms in email spam detection using Hyperparameter Optimization with RandomizedSearchCV. The research focus includes determining optimal parameters using Support Vector Machine (SVM), such as kernel, C, and gamma, as well as adjusting the number of trees and features in Random Forest.

In addition, the method of selecting relevant features will also be considered. Performance evaluation will be carried out using metrics such as accuracy, precision, recall, and F1-score and AUC ROC. The use of optimization has an effect on the Support Vector Machine and Random Forest training models, namely Support Vector Machine without optimization produces an accuracy of 95% and with optimization produces an accuracy of 96% and Random Forest without optimization produces an accuracy of 97% and with optimization produces an accuracy of 98%.

Keyword: *Support Vector Machine (SVM), Random Forest, Email Spam, Hyperparameter Optimization*