

**KLASIFIKASI SPAM EMAIL MENGGUNAKAN ALGORITMA
DECISION TREE DENGAN METODE
BOOTSTRAP AGGREGATING (BAGGING) DAN
ADAPTIVE BOOSTING (ADABOOST)**

SKRIPSI

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



disusun oleh

ANITA KUSUMA WARDANI

21.11.4011

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

**KLASIFIKASI SPAM EMAIL MENGGUNAKAN ALGORITMA
DECISION TREE DENGAN METODE
BOOTSTRAP AGGREGATING (BAGGING) DAN
ADAPTIVE BOOSTING (ADABOOST)**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

ANITA KUSUMA WARDANI

21.11.4011

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

HALAMAN PERSETUJUAN

SKRIPSI

**KLASIFIKASI SPAM EMAIL MENGGUNAKAN ALGORITMA
DECISION TREE DENGAN METODE
BOOTSTRAP AGGREGATING (BAGGING) DAN
ADAPTIVE BOOSTING (ADABOOST)**

yang disusun dan diajukan oleh

Anita Kusuma Wardani

21.11.4011

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 24 Februari 2026

Dosen Pembimbing,



Windha Mega Pradnya Dhuhiita, S.Kom., M.Kom.

NIK. 190302185

HALAMAN PENGESAHAN
SKRIPSI
KLASIFIKASI SPAM EMAIL MENGGUNAKAN ALGORITMA
DECISION TREE DENGAN METODE
BOOTSTRAP AGGREGATING (BAGGING) DAN
ADAPTIVE BOOSTING (ADABOOST)

yang disusun dan diajukan oleh

Anita Kusuma Wardani

21.11.4011

Telah dipertahankan di depan Dewan Penguji
pada tanggal 24 Februari 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Muhammad Tofa Nurcholis, S.Kom., M.Kom.
NIK. 190302281

Ali Mustopa, S.Kom., M.Kom.
NIK. 190302192

Windha Mega Pradnya Dhuhiita, S.Kom., M.Kom.
NIK. 190302185



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 24 Februari 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusriani, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Anita Kusuma Wardani

NIM : 21.11.4011

Menyatakan bahwa Skripsi dengan judul berikut:

KLASIFIKASI SPAM EMAIL MENGGUNAKAN ALGORITMA DECISION TREE DENGAN METODE BOOTSTRAP AGGREGATING (BAGGING) DAN ADAPTIVE BOOSTING (ADABOOST)

Dosen Pembimbing : Windha Mega Pradnya Dhuhiita, 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 Februari 2026

Yang Menyatakan,



Anita Kusuma Wardani

HALAMAN PERSEMBAHAN

Puji syukur kepada Allah SWT atas segala rahmat dan hidayah yang telah memberikan kelancaran, kemudahan, kesehatan, dan kesabaran dalam menyusun skripsi ini. Skripsi ini saya persembahkan kepada orang-orang yang saya sayangi dan cintai :

1. Allah SWT karena sudah memberikan kemudahan kepada penulis untuk menyelesaikan skripsi ini.
2. Bapak, Ibu, kakak, dan keluarga besar karena menjadi motivasi dan selalu memberikan dukungan, semangat dan doa untuk penulis.
3. Ibu Windha Mega Pradnya Dhuhiya, S.Kom., M.Kom, selaku dosen pembimbing yang selalu membantu dan memberi masukan untuk penulis agar laporan skripsi ini selesai dengan baik.
4. Teman seperjuangan S1 Informatika yang selalu memberikan dukungan, motivasi dan petunjuk untuk menyelesaikan tugas akhir skripsi ini.
5. Terakhir, saya ucapkan terimakasih untuk Anita Kusuma Wardani, diri saya sendiri.

KATA PENGANTAR

Puji syukur kehadiran Allah SWT atas segala rahmat dan hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul "Klasifikasi Spam Email Menggunakan Algoritma Decision Tree Dengan Metode Bootstrap Aggregating (Bagging) Dan Adaptive Boosting (AdaBoost)" ini sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer di Universitas Amikom Yogyakarta.

Proses penyusunan hingga selesainya laporan skripsi ini tidak lepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Maka dari itu, saya tidak lupa mengucapkan terima kasih dan penghargaan kepada:

1. Bapak Prof. Dr. M. Suyanto, M.M, selaku Rektor Universitas AMIKOM Yogyakarta.
2. Ibu Windha Mega Pradnya Dhuhita, M.Kom., selaku dosen pembimbing yang telah memberikan arahan, bimbingan, dan motivasi dalam penyusunan skripsi ini.
3. Orang tua dan keluarga tercinta, yang selalu memberikan doa, dukungan moral, serta semangat dalam menyelesaikan skripsi ini.
4. Teman-teman seperjuangan, yang selalu memberikan bantuan, dukungan, serta kebersamaan selama masa studi.
5. Semua pihak yang tidak dapat disebutkan satu per satu, yang telah membantu dalam proses penyelesaian skripsi ini.

Penulis menyadari bahwa skripsi ini masih jauh dari kesempurnaan. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi perbaikan di masa mendatang. Semoga skripsi ini dapat memberikan manfaat bagi pembaca dan menjadi kontribusi yang berharga.

Yogyakarta, 24 Februari 2026

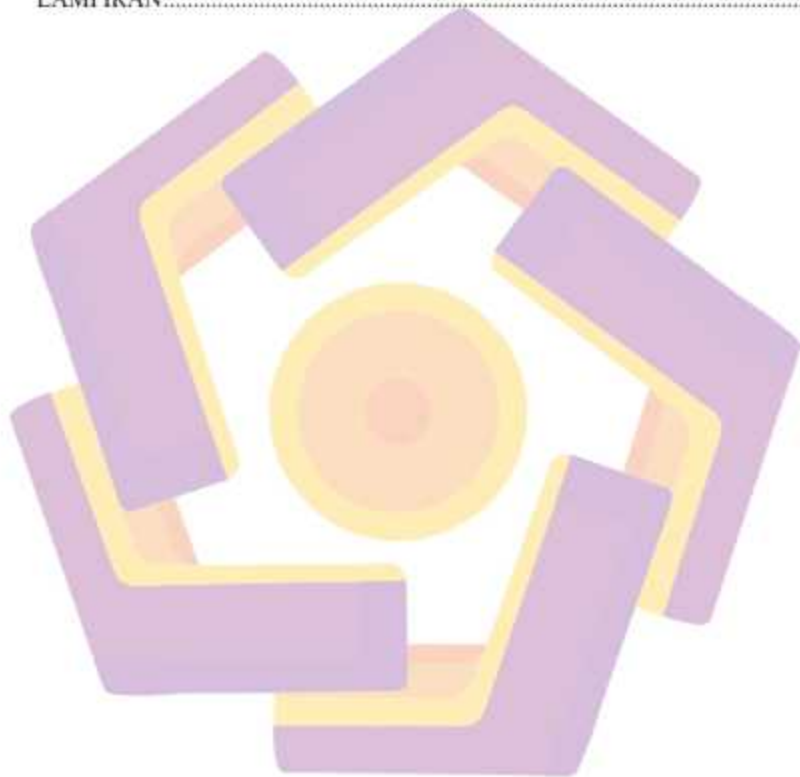
Penulis

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
DAFTAR LAMPIRAN	xii
DAFTAR LAMBANG DAN SINGKATAN	xiii
DAFTAR ISTILAH	xiv
INTISARI	xvii
ABSTRACT	xviii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	3
1.6 Sistematika Penulisan	4
BAB II TINJAUAN PUSTAKA	5
2.1 Studi Literatur	5
2.2 Dasar Teori	13
2.2.1 Email Spam	13
2.2.2 Algoritma	13
2.2.3 Klasifikasi	14
2.2.4 Preprocessing Text	15
2.2.5 Ekstraksi Fitur	16

2.2.1	<i>Decision Tree</i>	17
2.2.2	<i>Bootstrap Aggregating (Bagging)</i>	18
2.2.3	<i>Adaptive Boosting (AdaBoost)</i>	19
2.2.4	Evaluasi Model	21
BAB III METODE PENELITIAN		24
3.1	Objek Penelitian	24
3.2	Alur Penelitian	24
3.2.1	Pengumpulan Data	25
3.2.2	<i>Preprocessing Data</i>	25
3.2.3	TF-IDF (<i>Term Frequency-Inverse Document Frequency</i>)	26
3.2.4	Pelatihan Model	26
3.2.5	Evaluasi Model	28
3.3	Alat dan Bahan	29
3.3.1	Alat Penelitian	29
3.3.2	Bahan Penelitian	29
BAB IV HASIL DAN PEMBAHASAN		31
4.1	Pengumpulan Data	31
4.2	<i>Preprocessing Data</i>	32
4.2.1	<i>Case folding</i>	32
4.2.2	<i>Cleaning Data</i>	34
4.2.3	<i>Tokenization</i>	35
4.2.4	<i>Stopword Removal</i>	37
4.2.5	<i>Stemming</i>	38
4.3	Transformasi Label Data	39
4.4	Distribusi Data	39
4.5	TF-IDF (<i>Term Frequency-Inverse Document Frequency</i>)	40
4.6	<i>Splitting Data</i>	41
4.7	Klasifikasi	41
4.8.1	Analisa Performa <i>Decision Tree</i>	42
4.8.2	Analisa Performa <i>Decision Tree</i> + AdaBoost	44
4.8.3	Analisa Performa <i>Decision Tree</i> + Bagging	46

4.8 Perbandingan Ketiga Model	48
BAB V PENUTUP	51
5.1 Kesimpulan	51
5.2 Saran	52
REFERENSI	53
LAMPIRAN.....	56



DAFTAR TABEL

Tabel 2. 1 Keaslian Penelitian	8
Tabel 2. 2 <i>Confusion matrix</i>	21
Tabel 4. 1 Dataset	31
Tabel 4. 2 <i>Case Folding</i>	32
Tabel 4. 3 <i>Cleaning Text</i>	34
Tabel 4. 4 <i>Tokenization</i>	35
Tabel 4. 5 <i>Stopword Removal</i>	37
Tabel 4. 6 <i>Stemming</i>	38
Tabel 4. 7 Transformasi label data	39
Tabel 4. 8 Skor TF-IDF	41
Tabel 4. 9 <i>Splitting Data</i>	41
Tabel 4. 10 <i>Classification report Decision Tree</i>	43
Tabel 4. 11 <i>Classification report Decision Tree + AdaBoost</i>	45
Tabel 4. 12 <i>Classification report Decision Tree + Bagging</i>	47
Tabel 4. 13 Perbandingan Ketiga Model	48

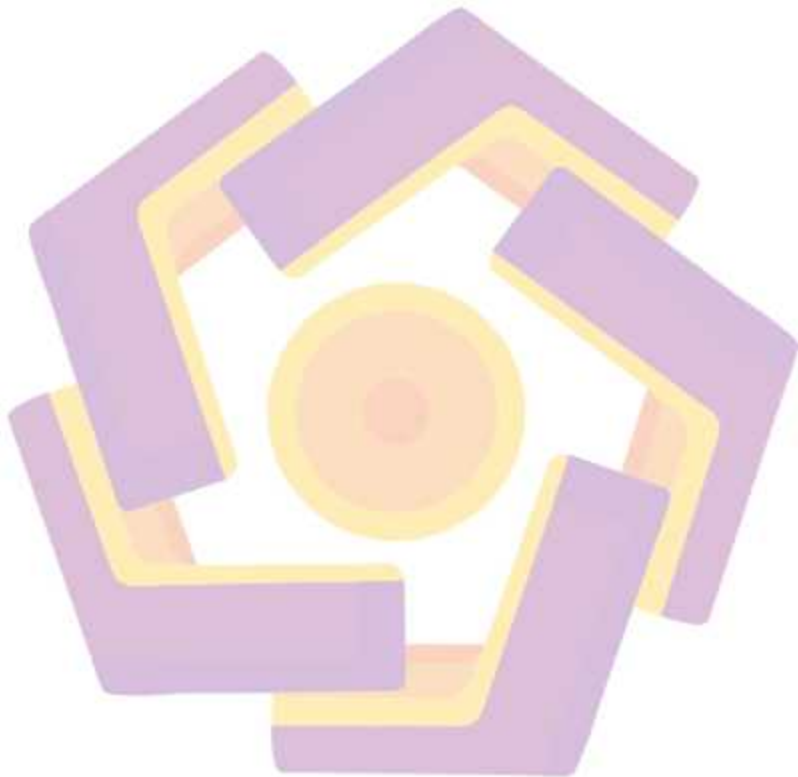
DAFTAR GAMBAR

Gambar 2. 1 Alur Diagram Metode Bagging.....	19
Gambar 3. 1 Alur Penelitian	24
Gambar 4. 1 Visualisasi distribusi label dataset	40
Gambar 4. 2 <i>Confusion matrix Decision Tree</i>	42
Gambar 4. 3 <i>Learning curve Decision Tree</i>	43
Gambar 4. 4 <i>Confusion matrix AdaBoost</i>	44
Gambar 4. 5 <i>Learning curve AdaBoost</i>	46
Gambar 4. 6 <i>Confusion matrix Bagging</i>	47
Gambar 4. 7 <i>Learning curve Bagging</i>	48
Gambar 4. 8 Perbandingan Ketiga Model.....	49
Gambar 4. 9 ROC-AUC dari ketiga model.....	50



DAFTAR LAMPIRAN

Lampiran 1 Link project	56
Lampiran 2 Deploy	56



DAFTAR LAMBANG DAN SINGKATAN



$h_t(x)$	Model lemah (Weak Learner) ke-t
α_t	Bobot model ke-t pada AdaBoost
A	Atribut
D	Jumlah seluruh dokumen
FN	False Negative
FP	False Positive
IDF	Inverse Document Frequency
N	Jumlah data
S	Himpunan data atau kasus
TF	Term Frequency
TF – IDF	Bobot term hasil perkalian TF dan IDF
TN	True Negative
TP	True Positive
W	Bobot term dalam dokumen
d	Dokumen
df(t)	Jumlah dokumen yang mengandung term (t)
f(t, d)	Frekuensi kemunculan term (t) pada dokumen (d)
n	Jumlah partisi data
t	Term atau kata
π	Proporsi data pada suatu kelas
AdaBoost	Adaptive Boosting
AUC	Area Under Curve
Bagging	Bootstrap Aggregating
NLP	Natural Language Processing
ROC	Receiver Operating Characteristic
ROC-AUC	Area Under Curve – Receiver Operating Characteristic
TF-IDF	Term Frequency – Inverse Document Frequency

DAFTAR ISTILAH

AdaBoost	Metode ensemble learning yang menggabungkan beberapa model lemah secara iteratif dengan memberi bobot lebih besar pada data yang salah diklasifikasikan.
Akurasi	Ukuran kinerja model yang menunjukkan proporsi prediksi yang benar terhadap keseluruhan data.
Bagging	Metode ensemble learning yang membangun beberapa model dari data hasil sampling bootstrap untuk mengurangi varians dan meningkatkan stabilitas model.
Case folding	Tahap preprocessing teks yang mengubah seluruh huruf menjadi huruf kecil (lowercase).
Classification report	Ringkasan hasil evaluasi model klasifikasi yang memuat nilai precision, recall, F1-Score, dan support.
Confusion matrix	Tabel evaluasi yang membandingkan hasil prediksi model dengan label aktual (TP, TN, FP, FN).
Decision tree	Algoritma supervised learning berbentuk pohon keputusan untuk klasifikasi atau prediksi data.
Email Ham (Not-spam)	Email sah yang diinginkan oleh penerima dan bukan termasuk spam.
Email Spam	Email yang dikirim tanpa izin penerima dan biasanya berisi iklan, penipuan, atau konten berbahaya.
Ensemble learning	Pendekatan pembelajaran mesin yang menggabungkan beberapa model untuk meningkatkan performa prediksi.
Entropy	Ukuran ketidakpastian data pada node dalam algoritma Decision tree.
F1-Score	Metrik evaluasi berupa rata-rata harmonik dari precision dan recall.

False Negative (FN)	Kondisi ketika data aktual positif tetapi diprediksi negatif oleh model.
False Positive (FP)	Kondisi ketika data aktual negatif tetapi diprediksi positif oleh model.
Gain (Information Gain)	Nilai yang digunakan untuk memilih atribut terbaik dalam pembentukan Decision tree berdasarkan penurunan entropy.
Klasifikasi	Proses pengelompokan data ke dalam kelas tertentu berdasarkan karakteristiknya.
Learning curve	Grafik yang menunjukkan hubungan antara ukuran data latih dan kinerja model.
Machine learning	Cabang kecerdasan buatan yang memungkinkan sistem belajar dari data.
Overfitting	Kondisi ketika model terlalu menyesuaikan data latih sehingga performanya menurun pada data uji.
Precision	Metrik evaluasi yang menunjukkan ketepatan prediksi kelas positif.
Preprocessing Text	Tahap awal pengolahan teks untuk membersihkan dan menyiapkan data sebelum analisis.
Recall	Metrik evaluasi yang menunjukkan kemampuan model mendeteksi seluruh data kelas positif.
ROC-AUC	Metode evaluasi yang mengukur kemampuan model membedakan kelas positif dan negatif berdasarkan kurva ROC.
Stemming	Proses mengubah kata menjadi bentuk dasarnya dengan menghilangkan imbuhan.
Stopword Removal	Tahap preprocessing untuk menghapus kata umum yang tidak memiliki makna penting.

Supervised Learning	Metode pembelajaran mesin yang menggunakan data berlabel dalam proses pelatihan model.
TF-IDF	Metode ekstraksi fitur yang memberi bobot kata berdasarkan frekuensi dan keunikannya dalam dokumen.
Tokenization	Proses memecah teks menjadi unit-unit kecil berupa token atau kata.
True Negative (TN)	Kondisi ketika data aktual dan prediksi model sama-sama negatif.
True Positive (TP)	Kondisi ketika data aktual dan prediksi model sama-sama positif.



INTISARI

Pesatnya penggunaan *email* sebagai media komunikasi digital diikuti meningkatnya volume *spam email* yang mengandung iklan tidak diinginkan, penipuan, *malware*, hingga ancaman keamanan data. Hal ini mengganggu produktivitas pengguna, menurunkan kepercayaan terhadap komunikasi elektronik, serta meningkatkan risiko serangan siber. Penelitian ini bertujuan mengatasi permasalahan tersebut dengan membandingkan kinerja algoritma *Decision Tree* tunggal terhadap kombinasi *ensemble* Bagging (*Bootstrap Aggregating*) dan AdaBoost (*Adaptive Boosting*) dalam klasifikasi *spam email*. Dataset publik dari Kaggle sebanyak 3.796 *email* (1.763 *spam* dan 1.891 *not-spam*) diproses melalui tahapan *preprocessing* teks (*case folding*, *cleaning*, *tokenization*, *stopword removal*, *stemming*), diikuti ekstraksi fitur menggunakan TF-IDF, *splitting data* (80:20), serta evaluasi menggunakan metrik *accuracy*, *precision*, *recall*, *F1-Score*, *confusion matrix*, *learning curve*, dan ROC-AUC. Hasil penelitian menunjukkan bahwa *Decision Tree* + AdaBoost memberikan performa terbaik dengan akurasi 98,08%, *precision* 98,09%, *recall* 98,08%, *F1-Score* 98,08%, serta AUC-ROC 0,9969, sekaligus berhasil meminimalkan *overfitting* secara signifikan dibandingkan *Decision Tree* tunggal (akurasi 91,24%) dan *Decision Tree* + Bagging (akurasi 91,66%).

Kata kunci: *spam email*, klasifikasi, *Decision Tree*, AdaBoost, Bagging

ABSTRACT

The rapid adoption of email as a digital communication tool has been accompanied by a significant increase in spam email containing unsolicited advertisements, scams, malware, and security threats. This phenomenon disrupts user productivity, erodes trust in electronic communication, and heightens cybersecurity risks. This study aims to address the problem by comparing the performance of a standalone Decision Tree algorithm with its ensemble variants using Bagging (Bootstrap Aggregating) and AdaBoost (Adaptive Boosting) for spam email classification. A public Kaggle dataset consisting of 3,796 emails (1,763 spam and 1,891 not-spam) was preprocessed through text preprocessing stages (case folding, cleaning, tokenization, stopword removal, stemming), followed by TF-IDF feature extraction, 80:20 data splitting, and comprehensive evaluation using accuracy, precision, recall, F1-Score, confusion matrix, learning curves, and ROC-AUC metrics. The results demonstrate that Decision Tree + AdaBoost achieved the best performance with an accuracy of 98.08%, precision 98.09%, recall 98.08%, F1-Score 98.08%, and ROC-AUC of 0.9969, while significantly reducing overfitting compared to standalone Decision Tree (accuracy 91.24%) and Decision Tree + Bagging (accuracy 91.66%).

Keyword: spam email, classification, Decision Tree, AdaBoost, Bagging