PERBANDINGAN ALGORITMA RANDOM FOREST DAN LOGISTIC REGRESSION DENGAN SELEKSI FITUR INFORMATION GAIN UNTUK KLASIFIKASI EMAIL SPAM

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Informatika

disusun oleh

RANGGA TEDDY PRATAMA

17.11.1702

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

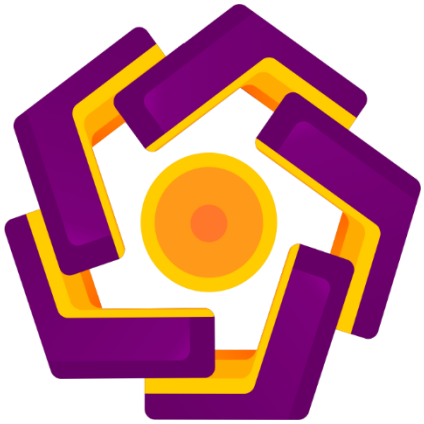
2023

PERBANDINGAN ALGORITMA RANDOM FOREST DAN LOGISTIC REGRESSION DENGAN SELEKSI FITUR INFORMATION GAIN UNTUK KLASIFIKASI EMAIL SPAM

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Informatika



disusun oleh

Rangga Teddy Pratama

17.11.1702

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2023

HALAMAN PERSETUJUAN

SKRIPSI

PERBANDINGAN ALGORITMA RANDOM FOREST DAN LOGISTIC REGRESSION DENGAN SELEKSI FITUR INFORMATION GAIN UNTUK KLASIFIKASI EMAIL SPAM

yang disusun dan diajukan oleh

Rangga Teddy Pratama

17.11.1702

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 25 Juli 2023

Dosen Pembimbing,

Lukman, M.Kom
NIK. 190302151

HALAMAN PENGESAHAN

SKRIPSI

PERBANDINGAN ALGORITMA RANDOM FOREST DAN LOGISTIC REGRESSION DENGAN SELEKSI FITUR INFORMATION GAIN UNTUK KLASIFIKASI EMAIL SPAM

yang disusun dan diajukan oleh

Rangga Teddy Pratama

17.11.1702

Telah dipertahankan di depan Dewan Penguji
pada tanggal 25 Juli 2023

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Yuli Astuti, M.Kom **NIK. 190302146**	
Banu Santoso, S.T., M.Eng **NIK. 190302327**	
Lukman, M.Kom **NIK. 190302151**	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 3 Agustus 2023

DEKAN FAKULTAS ILMU KOMPUTER

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

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Rangga Teddy Pratama
NIM : 17.11.1702

Menyatakan bahwa Skripsi dengan judul berikut:

PERBANDINGAN ALGORITMA RANDOM FOREST DAN LOGISTIC REGRESSION DENGAN SELEKSI FITUR INFORMATION GAIN UNTUK KLASIFIKASI EMAIL SPAM

Dosen Pembimbing : Lukman, 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, 25 Juli 2023

Yang Menyatakan,

Rangga Teddy Pratama

HALAMAN PERSEMBAHAN

Puji Syukur kepada Allah SWT atas rahmat dan ridho-Nya yang telah diberikan kepada saya ,karena tanpa rahmat dan ridho-Nya saya tidak mungkin bisa sampai sejauh ini. Dan terimakasih kepada orang-orang yang telah membantu serta mendukung saya dalam melakukan penelitian ini. Oleh karena itu, saya persembahkan kepada semua pihak yang terlibat secara langsung dan tidak langsung dalam proses penelitian ini.

1. Kedua orang tua (Anggraeni dan Suripto) serta seluruh keluarga yang selalu mensupport dan selalu mendoakan saya agar dapat segera menyelesaikan skripsi ini.
2. Bapak Lukman, M.Kom selaku dosen pembimbing saya yang sudah membimbing saya dalam penelitian ini dan memberikan saran-saran yang membantu dalam penyusunan penelitian ini.
3. Terima kasih banyak untuk bantuan dan dukungannya selama ini kepada Farida, Tito, dan Rizqi yang menjadi teman terdekat saya selama perkuliahan.
4. Rekan-rekan seperjuangan yang selalu membantu, mensupport yang tidak bisa saya sebutkan satu persatu yang telah menjadi tempat saya belajar dan berbagi selama saya kuliah di Universitas AMIKOM Yogyakarta.
5. Terakhir skripsi ini saya persembahkan untuk orang-orang yang selalu menanyakkan “skripsi sampe mana ngga?”, pertanyaan paling menyebalkan tetapi itu juga yang memicu saya untuk segera menyelesaikan skripsi ini, Terimakasih.

KATA PENGANTAR

Puji syukur kepada Allah SWT atas limpahan rahmat, nikmat, serta hidayahnya sehingga saya dapat menyelesaikan skripsi ini dengan judul "**Perbandingan Algoritma Random Forest dan Logistic Regression dengan Seleksi Fitur Information Gain Untuk Klasifikasi Email SPAM "** yang disusun sebagai salah satu syarat kelulusan dan menjadi bukti bahwa telah menyelesaikan jenjang program studi Strata-1 dan memperoleh gelar Sarjana Komputer.

Penyelesaian skripsi ini juga tidak terlepas dari bantuan berbagai pihak yang tidak bisa penulis sebutkan satu per satu, karena itu pada kesempatan ini penulis ingin menyampaikan rasa hormat dan terima kasih kepada orangtua saya,keluarga, dosen, teman-teman, dan seluruh pihak yang membantu sehingga skripsi ini dapat terselesaikan.semoga Allah SWT memberikan balasan kebaikan kepada semua pihak yang telah membantu penulis dalam menyelesaikan skripsi ini.

Dalam penyusunan skripsi ini, penulis memahami bahwa masih terdapat banyak aspek yang perlu ditingkatkan dan di perbaiki. Oleh karena itu, demi perbaikan selanjutnya segala saran dan kritik yang membangun akan diterima dengan senang hati dan rasa terima kasih. Semoga skripsi ini bisa memberikan manfaat bagi penulis, serta bagi semua pihak yang membaca. Amin

Klaten, 17 Agustus 2023

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Email merupakan surat elektronik yang memungkinkan seseorang mengirimkan dan menerima pesan lebih dari satu orang. *Email* memiliki kelebihan diantara lain cepat, hemat, dan dapat mengirimkan pesan dalam bentuk dolumen, excel, foto, audio, video, dll sehingga banyak yang menggunakannya diseluruh dunia. Banyaknya lalulintas *email* setiap harinya ini sering disalahgunakan oleh berbagai pihak untuk mengirim informasi iklan produk jasa dan berbagai informasinya yang tidak berguna atau tidak diinginkan oleh user. Itulah yang sering disebut dengan *email SPAM*.

Untuk penyaringan *email SPAM* digunakan banyak algoritma klasifikasi, algoritma tersebut diantarannya adalah *random forest* dan *logistic regression*. Pada penelitian ini akan dilakukan perbandingan algoritma *random forest* dan *logistic regression*. Untuk meningkatkan kinerja algoritma klasifikasi dilakukan seleksi fitur menggunakan metode *information gain* yang bertujuan untuk menyeleksi fitur/atribut yang paling berpengaruh.

Berdasarkan hasil pengujian yang menggunakan data sebanyak 4601 baris data dan 58 fitur, menggunakan algoritma *random forest* dan *logistic regression* diperoleh hasil akurasi sebesar 95,3% dan 82,9%, setelah dilakukan seleksi fitur dengan *information gain* akurasi yang dihasilkan sebesar 95,6% dan 82,3%. Dari uji coba terjadi peningkata akurasi pada *random forest* dan terjadi penurunan pada *logistic regression*. Hal ini membuktikan bahwa penerapan seleksi fitur *information gain* dapat menghilangkan atribut redundan.

Kata Kunci: *email SPAM*, klasifikasi, *random forest, logistic regression, information gain.*

ABSTRACT

Email is electronic mail that allows someone to send and receive messages from more than one person. Email has advantages including fast, economical, and can send messages in the form of documents, excel, photos, audio, video, etc. so many people use them all over the world. The amount of e-mail traffic every day is often misused by various parties to send advertising information for service products and various information that is not useful or unwanted by users. That is what is often referred to as SPAM e-mail.

For SPAM email filtering, many classification algorithms are used, the algorithms of which are random forest and logistic regression. In this study, a comparison of random forest and logistic regression algorithms will be carried out. To improve the performance of the classification algorithm, feature selection is carried out using the information gain method which aims to select the most influential features/attributes.

Based on the test results using data as many as 4601 data lines and 58 features, using random forest and logistic regression algorithms obtained accuracy results of 95.3% and 82.9%, after feature selection with information gain the resulting accuracy is 95.6% and 82.3%. From the trial there was an increase in accuracy in the random forest and a decrease in logistic regression. This proves that the application of information gain feature selection can eliminate redundant attributes.

Keyword: *email SPAM, classification, random forest, logistic regression, information gaim.*