Kombinasi *Logistic Regression* dan *Gradient Boost Tree* untuk Mendeteksi Email Spam

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurursan Teknik Informatika

Oleh:

Arfah Anggraina

09021281520140

JURUSAN TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2019

LEMBAR PENGESAHAN TUGAS AKHIR

KOMBINASI *LOGISTIC REGRESSION* DAN *GRADIENT BOOST TREE* UNTUK MENDETEKSI EMAIL SPAM

Oleh :

ARFAH ANGGRAINA
NIM : 09021281520140

Indralaya, 26 November 2019

Pembimbing,

Rifkie Primartha, M.T.
NIP. 197706012009121004

Mengetahui,
Ketua Jurusan Teknik Informatika,

Rifkie Primartha, M.T.
NIP. 197706012009121004

ii

TANDA LULUS UJIAN SIDANG TUGAS AKHIR

Pada hari **Jumat** tanggal **22 November 2019** telah dilaksanakan ujian sidang tugas akhir oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama	:	Arfah Anggraina
N I M	:	09021281520140
Judul	:	Kombinasi *Logistic Regression* dan *Gradient Boost Tree* untuk Mendeteksi Email Spam

1. Pembimbing

   Rifkie Primartha, M.T. .....
   NIP. 197706012009121004

2. Penguji I

   Dian Palupi Rini, M.Kom., Ph.D .....
   NIP 197802232006042002

3. Penguji II

   Rizki Kurniati, M.T .....
   NIP. 199107122019032016

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha S.T. M.T.
NIP 197706012009121004

iii

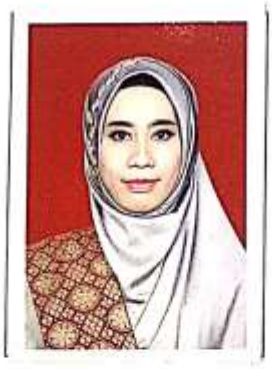
HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama	: Arfah Anggraina
NIM	: 09021281520140
Program Studi	: Teknik Informatika
Judul Skripsi	: Kombinasi *Logistic Regression* dan *Gradient Boost Tree* untuk Mendeteksi Email Spam
Hasil Pengecekan Software *iThenticate/Turnitin*	: 18%

Menyatakan bahwa Laporan Projek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan projek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian, pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.



Palembang, 26 November 2019

Arfah Anggraina
NIM. 09021281520140

iv

Motto:

- *Awali hari-harimu dengan Bismillah.*
- الـ ذ صـ ير وذ عم الـ مولـ ى ذ عم الـ وكـ يل وذ عم الله حـ سـ بن
- *Love Yourself!*

Kupersembahkan karya tulis ini kepada :

- *Orang tuaku tersayang*
- *Keluarga besarku*
- *Sahabat dan teman seperjuanganku*
- *Fakultas Ilmu Komputer Universitas Sriwijaya*

THE COMBINATION OF LOGISTIC REGRESSION AND GRADIENT BOOST TREE FOR DETECTING EMAIL SPAMS

By:
Arfah Anggraina
09021281520140

ABSTRACT

Email spam is a serious problem that is experienced by users all over the world. In 2016, it was noted that 61.66% of spam hampered the flow of world traffic. The average email received by the recipient in the form of spam containing an ad, so the need for filtering spam in order to minimize receipt of spam to the recipient. Spam filtering can be done by spam classification process which is a problem in email. In this study, the classification process is carried out using the Gradient Boost Tree method. However, this method can experience overfitting if the data used is noise. So for the data classification process, using the Gradient Boost Tree algorithm that is optimized using Logistic Regression. This study compares the results of the classification using the spambase dataset on the Gradient Boost Tree algorithm and with the addition of Logistic Regression to the feature selection process. From this study, obtained the highest accuracy results in the merging of the Gradient Boost Tree algorithm that is optimized with Logistic Regression that is equal to 95.13%

Keywords: *email, gradient boost tree, logistic regression, spam.*

Inderalaya, 26 November 2019

Supervisor,

Rifkie Primartha, M.T.
NIP. 197706012009121004

Approved,
Chairman of Informatic Engineering Department

Rifkie Primartha, M.T.
NIP. 197706012009121004

vi

KOMBINASI *LOGISTIC REGRESSION* DAN *GRADIENT BOOST TREE*
UNTUK MENDETEKSI EMAIL SPAM

Oleh:
Arfah Anggraina
09021281520140

ABSTRAK

Email spam menjadi permasalahan cukup serius yang dialami pengguna email di seluruh dunia. Pada tahun 2016, tercatat bahwa sebesar 61.66% spam menghambat jalannya lalu lintas dunia. Rata-rata email yang diterima penerima berupa spam yang berisikan sebuah iklan, sehingga perlunya *filtering* spam guna meminimalisir diterimanya spam kepada penerima. Filtering spam dapat dilakukan dengan proses klasifikasi spam yang menjadi permasalahan pada Email. Pada penelitian ini, proses klasifikasi dilakukan dengan menggunakan metode *Gradient Boost Tree*. Namun metode ini dapat mengalami *overfitting* jika data yang digunakan terdapat *noise*. Sehingga untuk proses klasifikasi data, menggunakan algoritma *Gradient Boost Tree* yang dioptimasi menggunakan *Logistic Regression*. Penelitian ini melakukan perbandingan hasil klasifikasi menggunakan dataset *spambase* pada algoritma *Gradient Boost Tree* dan dengan penambahan *Logistic Regression* untuk proses seleksi fitur. Dari penelitian ini, diperoleh hasil akurasi tertinggi pada penggabungan algoritma *Gradient Boost Tree* yang dioptimasi dengan *Logistic Regression* yaitu sebesar 95.13%.

Keywords: *email*, *gradient boost tree*, *logistic regression*, *spam*.

Inderalaya, 26 November 2019

Pembimbing,

Rifkie Primartha, M.T.
NIP. 197706012009121004

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 197706012009121004

vii

KATA PENGANTAR

Puji syukur kepada Allah SWT atas berkat dan rahmat-Nya yang telah diberikan kepada Penulis sehingga dapat menyelesaikan Tugas Akhir ini dengan baik. Tugas Akhir ini disusun untuk memenuhi salah satu syarat guna menyelesaikan pendidikan program Strata-1 pada Fakultas Ilmu Komputer Program Studi Teknik Informatika di Universitas Sriwijaya.

Dalam menyelesaikan Tugas Akhir ini banyak pihak yang telah memberikan bantuan dan dukungan baik secara langsung maupun secara tidak langsung. Pada kesempatan ini, penulis ingin menyampaikan ucapan terima kasih kepada pihak-pihak yang telah membantu penulis dalam menyelesaikan Tugas Akhir ini, yaitu kepada:

1. Orang tuaku, Dimyati Marzuki Azmi dan Almarhumah Etty Sumarni, serta seluruh saudari-saudariku Yuk Tya, Yuk Rara, Yuk Rosi, dan kakak iparku kak Adi, Kak Taufik dan kak Uki yang memberikan cinta dan kasih sayangnya untuk selalu mendoakan serta memberikan dukungan baik moril maupun materil.
2. Bapak Jaidan Jauhari, S.Pd., M.T. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
3. Bapak Rifkie Primartha, S.T., M.T. selaku Ketua Jurusan Teknik Informatika sekaligus sebagai pembimbing Tugas Akhir yang telah membimbing, mengarahkan dan memberikan motivasi penulis dalam proses perkuliahan dan pengerjaan Tugas Akhir.
4. Bapak Danny Matthew Saputra, M.Sc selaku dosen pembimbing akademik, yang telah membimbing, mengarahkan, dan memberikan motivasi penulis dalam proses perkuliahan.
5. Ibu Dian Palupi Rini, M.Kom., Ph.D selaku dosen penguji I, yang telah memberikan masukkan dan dorongan dalam proses pengerjaan Tugas Akhir.
6. Ibu Rizki Kurniati, M.T selaku dosen penguji II, yang telah memberikan masukkan dan dorongan dalam proses pengerjaan Tugas Akhir.
7. Seluruh dosen Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
8. Mbak Winda, Kak Ricy dan Kak Hafez serta seluruh staf tata usaha yang telah membantu dalam kelancaran proses administrasi dan akademik selama masa perkuliahan.
9. Sahabatku Taca Rosa, Mega Rizki, Mikyal Marshalita, Indah Ramadhona yang telah membantu dalam melancarkan proses Tugas Akhir serta yang selalu mendengarkan keluh kesah penulis dan memberikan saran terbaik kalian agar penulis terus bangkit.
10. Teman dekatku Vira Melinda, Faiz Muhammad, Ahmad Halim, Arief Rachmatullah, Ahmad Ikrom, Ade Lismita, Zakia Amalia, Devi Permata Hati yang telah memberikan semangat dan motivasi kepada penulis.
11. Rizki Pratama Putra telah membantu penulis dalam kelancaran proses program Tugas Akhir.

12. Teman-teman dari kelas IF Reg B 2015, kakak tingkat, adik tingkat, serta teman-teman lainnya yang telah mendengarkan keluh kesah penulis serta memberikan berbagai masukkan selama menempuh Pendidikan di Fakultas Ilmu Komputer Universitas Sriwijaya.
13. Darmawan Abinugroho yang telah siap sedia untuk membantu penulis.
14. BPH HMIF Fasilkom Unsri, yang telah memberikan kesempatan penulis dalam berkarya serta turut andil dalam menjalankan berbagai tugas yang diberikan sehingga penulis dapat menerapkan tugas tersebut ke lingkungan yang lebih luas.
15. The Ambigus yang telah memberikan doa dan semangat untuk penulis sehingga penulis dapat menyelesaikan Tugas Akhir ini.

Penulis menyadari dalam penyusunan Tugas Akhir ini masih terdapat banyak kekurangan disebabkan keterbatasan pengetahuan dan pengalaman, oleh karena itu kritik dan saran yang membangun sangat diharapkan untuk kemajuan penelitian selanjutnya.

Akhir kata semoga Tugas Akhir ini dapat berguna dan bermanfaat bagi kita semua.

Indralaya, November 2019

Arfah Anggraina

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR LAMPIRAN

Halaman

BAB I

PENDAHULUAN

1.1 Pendahuluan

Pada bab ini membahas latar belakang masalah, rumusan masalah, tujuan, manfaat penelitian, dan batasan masalah. Bab ini akan memberikan penjelasan umum mengenai keseluruhan penelitian.

Pendahuluan dimulai dengan penjelasan mengenai proses klasifikasi email spam serta penelitian yang berkaitan dengan penerapan metode *Gradient Boost Tree* dan *Logistic Regression* yang menjadi latar belakang penelitian ini.

1.2 Latar Belakang

Email spam menjadi permasalahan yang cukup serius yang sering terjadi. Spam pada email dapat berupa iklan yang dapat mengganggu aktivitas penggunanya dalam berkomunikasi (Wijaya & Bisri, 2016). Pada tahun 2016, tercatat bahwa sebesar 61.66% spam yang menghambat jalannya lalu lintas dunia (Gomes et al., 2017). Mengingat spam dapat mengganggu aktivitas penggunanya, berbagai teknik dilakukan untuk menyaring email spam. Penyaringan email spam menjadi pusat perhatian yang terus meningkat. Berbagai macam *software* terkait *filtering* spam telah banyak bermunculan dan terus dikembangkan. Salah satu pendekatan yang popular yaitu menjadikan *filtering* spam sebagai masalah klasifikasi (Bing Zhou, Yao, & Luo, 2014).

Berbagai macam algoritma klasifikasi yang telah digunakan untuk penyaringan email spam. Salah satu penelitian mengenai klasifikasi email spam dilakukan oleh (Bing Zhou et al., 2014). Zhou dkk mengusulkan metode yang digunakan yaitu *cost-sensitive three-way* (*bayensian, thresholds, probability*) dengan hasil akurasi yang diperoleh yaitu sebesar 89.88%. Penelitian lain terkait klasifikasi email spam yaitu yang pernah diteliti oleh (Wijaya & Bisri, 2016). Pada tahun 2016, Adi Wijaya dan Achmad Bisri melakukan penelitian terkait klasifikasi email spam. Mereka mengusulkan metode untuk digunakan pada penelitiannya yaitu *Logistic Regression* dengan *FN Threshold* (LRFNT) dan *Decision Tree* (DT). Hasil yang diperoleh menggunakan algoritma tersebut yaitu akurasi sebesar 91.67%. Peningkatan terus dilakukan dengan berbagai macam metode klasifikasi yang digunakan. Salah satu metode data mining untuk mengklasifikasikan email spam yaitu *Gradient Boost Tree*.

Algoritma *Gradient Boost Tree* dapat digunakan pada proses pengklasifikasian dan *Regression* (Bingyin Zhou, Lu, & Wang, 2016). *Gradient Boost Tree* memiliki kelebihan yaitu mampu memperbaiki kesalahan yang terjadi pada pohon sebelumnya. Namun *Gradient Boost Tree* dapat mengalami *overfitting* jika data yang digunakan terdapat *noise*. *Overfitting* yaitu perbedaan antara tingkat kesalahan klasifikasi model sebenarnya dan tingkat yang diakui berdasarkan pergantian dari dataset pelatihan (Khoshgoftaar, Allen, & Deng, 2001). Untuk mengatasi kekurangan pada *Gradient Boost Tree,* maka algoritma yang dapat digunakan yaitu *Logistic Regression*.

Algoritma *Logistic Regression* yaitu gabungan dari sejumlah teknik untuk menampilkan dan mengevaluasi beberapa faktor penekanannya pada hubungan antara variabel bawaan serta merupakan strategi penilaian untuk memecahkan kumpulan data dimana setidaknya ada satu faktor otonom yang menentukan hasil (Bhargava & Katarya, 2017). Algoritma ini bersifat linier untuk memprediksi probabilitas (Wei, Wang, & Wang, 2012). Penelitian yang telah dilakukan oleh (Wijaya & Bisri, 2016) menjelaskan bahwa *Logistic Regression* juga mampu mengatasi *noise* sebelum data diumpan ke *Gradient Boost Tree* yang dapat mengakibatkan *overfitting*. Sehingga kelebihan *Logistic Regression* dapat dimanfaatkan untuk menutupi kekurangan *Gradient Boost Tree*. Untuk itu, penelitian ini akan mengusulkan algoritma *Gradient Boost Tree* dan *Logistic Regression* untuk mengklasifikasi email spam.

1.3 Rumusan Masalah

Rumusan masalah dari penelitian ini yaitu bagaimana peranan *Logistic Regression* dalam meningkatkan kinerja terhadap algoritma *Gradient Boost Tree*. Untuk menjawab rumusan masalah tersebut, maka diuraikan beberapa *research question* sebagai berikut:

1. Bagaimana penerapan *Logistic Regression* pada klasifikasi email spam dengan *Gradient Boost Tree*?
2. Bagaimana pengaruh *Logistic Regression* terhadap *Gradient Boost Tree* dalam peningkatan akurasi klasifikasi email spam?

1.4 Tujuan Penelitian

Tujuan dilakukannya penelitian sebagai berikut:

1. Mengetahui cara penerapan *Logistic Regression* pada klasifikasi email spam dengan algoritma *Gradient Boost Tree*;
2. Mengetahui pengaruh *Logistic Regression* terhadap *Gradient Boost Tree* dalam peningkatan akurasi klasifikasi email spam.

1.5 Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut:

1. Memahami algoritma *Gradient Boost Tree* untuk klasifikasi email spam;
2. Memahami peranan *Logistic Regression* yang bekerja dengan metode *Gradient Boost Tree* dalam meningkatkan akurasi;
3. Mampu menerapkan metode *Logistic Regression* dengan *Gradient Boost Tree* untuk klasifikasi email spam.

1.6 Batasan Masalah

Batasan masalah dari penelitian ini adalah sebagai berikut:

1. Karena kasus yang digunakan yaitu email spam, maka data yang digunakan merupakan data Spambase yang terdapat di UCI Machine Learning Repository.
2. Format berkas yang dapat dimasukkan hanya *.xlsx*

1.7 Sistematika Penulisan

Sistematika penulisan tugas akhir ini mengikuti standar penulisan tugas akhir Fakultas Ilmu Komputer Universitas Sriwijaya yaitu sebagai berikut.

BAB I PENDAHULUAN

Pada bab ini akan diuraikan mengenai latar belakang, perumusan masalah, tujuan dan manfaat penelitian, batasan masalah/ruang lingkup, metodologi penelitian dan sistematika penulisan.

BAB II KAJIAN LITERATUR

Pada bab ini akan membahas seluruh dasar-dasar teori yang digunakan mulai dari definisi sistem, informasi mengenai domain, dan semua yang digunakan pada tahapan analisis, perancangan, dan implementasi.

BAB III METODOLOGI PENELITIAN

Pada bab ini akan membahas mengenai tahap-tahap yang akan diterapkan pada penelitian. Setiap rencana dari tahapan penelitian dideskripsikan secara rinci berdasarkan kerangka kerja. Dilanjutkan dengan perancangan manajemen proyek dalam pelaksanaan penelitian.

BAB IV METODOLOGI PENELITIAN

Pada bab ini akan membahas perancangan dan lingkungan klasifikasi, implementasi program hasil klasifikasi dengan *Gradient*

Boost Tree dan *Logistic Regression*, hasil eksekusi dan hasil pengujian.

BAB V HASIL DAN ANALISIS PENELITIAN

Pada bab ini, hasil penelitian yang telah dilakukan akan disajikan. Analisis yang disajikan sebagai basis dari kesimpulan yang diambil dari penelitian ini.

BAB VI KESIMPULAN DAN SARAN

Pada bab ini berisikan kesimpulan dari semua uraian yang telah dibahas pada bab sebelumnya dan saran yang diharapkan dapat berguna untuk pengembangan pada penelitian selanjutnya.

1.8 Kesimpulan

Kesimpulan pada bab ini adalah sebagai berikut:

1. Penerapan algoritma *Logistic Regression* pada *Gradient Boost Tree* untuk klasifikasi email spam;
2. Tingkat akurasi dipengaruhi oleh kinerja seleksi fitur dengan menggunakan algoritma *Logistic Regression.*

DAFTAR PUSTAKA

Bhargava, K., & Katarya, R. (2017, 12-14 Oct. 2017). *An improved lexicon using logistic regression for sentiment analysis.* Paper presented at the 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN).

Düntsch, I., & Gediga, G. (2019). *Confusion matrices and rough set data analysis.* Paper presented at the Journal of Physics: Conference Series.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Gomes, S. R., Saroar, S. G., Mosfaiul, M., Telot, A., Khan, B. N., Chakrabarty, A., & Mostakim, M. (2017). *A comparative approach to email classification using Naive Bayes classifier and hidden Markov model.* Paper presented at the Advances in Electrical Engineering (ICAEE), 2017 4th International Conference on.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Hastuti, K. (2012). Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif. *Semantik, 2*(1).

Khoshgoftaar, T. M., Allen, E. B., & Deng, J. (2001). *Controlling overfitting in software quality models: Experiments with regression trees and classification.* Paper presented at the metrics.

Kim, T., Lee, D., Choi, J., Spurlock, A., Sim, A., Todd, A., & Wu, K. (2015). *Extracting Baseline Electricity Usage with Gradient Tree Boosting.* Paper presented at the Proceedings of 2015 International Conference on Big Data Intelligence and Computing (DataCom 2015).

Kruchten, P. (2003). *Introdução ao RUP: rational unified process*: Ciência Moderna.

Laksana, E. A., & Sulianta, F. (2017). ANALISIS DAN STUDI KOMPARATIF ALGORITMA KLASIFIKASI GENRE MUSIK. *SEMNASTEKNOMEDIA ONLINE, 5*(1), 2-1-67.

Rindskopf, D., & Shrout, P. E. (2019). Logistic regression with floor and ceiling effects. *Advances in Latent Class Analysis: A Festschrift in Honor of C. Mitchell Dayton*, 147.

Rushin, G., Stancil, C., Sun, M., Adams, S., & Beling, P. (2017). *Horse race analysis in credit card fraud—deep learning, logistic regression, and Gradient Boosted Tree.* Paper presented at the Systems and Information Engineering Design Symposium (SIEDS), 2017.

Salehi, S., Selamat, A., Kuca, K., Krejcar, O., & Sabbah, T. (2017). Fuzzy granular classifier approach for spam detection. *Journal of Intelligent & Fuzzy Systems, 32*(2), 1355-1363.

Singh, L., Kaur, N., & Chetty, G. (2018). *Customer Life Time Value Model Framework Using Gradient Boost Trees with RANSAC Response Regularization.* Paper presented at the 2018 International Joint Conference on Neural Networks (IJCNN).

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, *10*(5), 293-302.

Wei, D., Wang, T., & Wang, J. (2012). A logistic regression model for Semantic Web service matchmaking. *Science China Information Sciences*, *55*(7), 1715-1720.

Wijaya, A., & Bisri, A. (2016). *Hybrid decision tree and logistic regression classifier for email spam detection.* Paper presented at the 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE).

Zhou, B., Lu, M., & Wang, Y. (2016). *Counting people using gradient boosted trees.* Paper presented at the Information Technology, Networking, Electronic and Automation Control Conference, IEEE.

Zhou, B., Yao, Y., & Luo, J. (2014). Cost-sensitive three-way email spam filtering. *Journal of Intelligent Information Systems, 42*(1), 19-45.