IMPLEMENTASI INFORMATION GAIN UNTUK FEATURE SELECTION DENGAN THRESHOLD NILAI MEDIAN

SKRIPSI

Steven Liong
00000042707

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2024

IMPLEMENTASI INFORMATION GAIN UNTUK FEATURE SELECTION DENGAN THRESHOLD NILAI MEDIAN

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer (S.Kom.)

Steven Liong
00000042707

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2024

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama	:	Steven Liong
Nomor Induk Mahasiswa	:	00000042707
Program Studi	:	Informatika

Skripsi dengan judul:

Implementasi Information Gain untuk Feature Selection dengan Threshold nilai Median

merupakan hasil karya saya sendiri bukan plagiat dari karya ilmiah yang ditulis oleh orang lain, dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/ penyimpangan, baik dalam pelaksanaan Skripsi maupun dalam penulisan laporan Skripsi, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk Tugas akhir yang telah saya tempuh.

Tangerang, 13 Mei 2024

(Steven Liong)

HALAMAN PENGESAHAN

Skripsi dengan judul

IMPLEMENTASI INFORMATION GAIN UNTUK FEATURE SELECTION DENGAN THRESHOLD NILAI MEDIAN

oleh

Nama	:	Steven Liong
NIM	:	00000042707
Program Studi	:	Informatika
Fakultas	:	Fakultas Teknik dan Informatika

Telah diujikan pada hari Rabu, 29 Mei 2024
Pukul 13.00 s/s 15.00 dan dinyatakan
LULUS
Dengan susunan penguji sebagai berikut

Ketua Sidang

(Adhi Kusnadi, S.T, M.Si.)
NIDN: 303037304

Penguji

5 Juni 2024

(Arya Wicaksana, S.Kom., M.Eng.Sc.
(OCA, CEH, CEI))
NIDN: 315109103

Pembimbing

(Dr. Maria Irmina Prasetiyowati, S.Kom., M.T.)
NIDN: 725057201

Ps. Ketua Program Studi Informatika,

(Dr. Eng. Niki Prastomo, S.T., M.Sc.)
NIDN: 0419128203

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan di bawah ini:

Nama	:	Steven Liong
NIM	:	00000042707
Program Studi	:	Informatika
Jenjang	:	S1
Jenis Karya	:	Skripsi

Menyatakan dengan sesungguhnya bahwa:

☐ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya di repositori Knowledge Center, sehingga dapat diakses oleh Civitas Akademika/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial dan saya juga tidak akan mencabut kembali izin yang telah saya berikan dengan alasan apapun.

☑ Saya tidak bersedia karena dalam proses pengajuan untuk diterbitkan ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*)**.

Tangerang, 13 Mei 2024

Yang menyatakan

Steven Liong

** Jika tidak bisa membuktikan LoA jurnal/HKI selama enam bulan ke depan, saya bersedia mengizinkan penuh karya ilmiah saya untuk diunggah ke KC UMN dan menjadi hak institusi UMN.

Halaman Persembahan / Motto

”A good name is to be more desired than great wealth, Favor is better than silver and gold.”

Proverbs 22:1 (NASB)

KATA PENGANTAR

Puji Syukur atas berkat dan rahmat kepada Tuhan Yang Maha Esa, atas selesainya penulisan laporan Skripsi ini dengan judul: Implementasi Information Gain untuk Feature Selection dengan Threshold nilai Median dilakukan untuk memenuhi salah satu syarat untuk mencapai gelar Sarjana Komputer Jurusan Informatika Pada Fakultas Teknik dan Informatika Universitas Multimedia Nusantara. Saya menyadari bahwa, tanpa bantuan dan bimbingan dari berbagai pihak, dari masa perkuliahan sampai pada penyusunan skripsi ini, sangatlah sulit bagi saya untuk menyelesaikan skripsi ini. Oleh karena itu, saya mengucapkan terima kasih kepada:

1. Bapak Dr. Ninok Leksono, selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Ketua Program Studi Informatika Universitas Multimedia Nusantara.
4. Ibu Dr. Maria Irmina Prasetiyowati, S.Kom., M.T., sebagai Pembimbing pertama yang telah banyak meluangkan waktu untuk memberikan bimbingan, arahan dan motivasi atas terselesainya skripsi ini.
5. Orang Tua yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan skripsi ini.
6. Untuk yang terkasih, Gracella Irwana, yang senantiasa memotivasi, memberikan dukungan dengan tulus, serta selalu menemani peneliti sehingga skripsi ini dapat terselesaikan dengan baik.

Semoga skripsi ini bermanfaat, baik sebagai sumber informasi maupun sumber inspirasi, bagi para pembaca.

Tangerang, 13 Mei 2024

Steven Liong

IMPLEMENTASI INFORMATION GAIN UNTUK FEATURE SELECTION DENGAN THRESHOLD NILAI MEDIAN

Steven Liong

ABSTRAK

Perkembangan ilmu data dan penemuan pengetahuan dalam aplikasi digital menuntut proses *feature selection* yang bertujuan untuk mengurangi dimensi data dan mengatasi biaya komputasi yang tinggi. Salah satu metode *feature selection* adalah *information gain*. Selain itu, nilai *threshold* yang ditentukan dari nilai *information gain* dapat menggunakan perhitungan statistika. Oleh karena itu, penelitian ini mengusulkan penentuan nilai *threshold* menggunakan nilai median dari *information gain* yang dihasilkan oleh setiap *feature* dalam dataset. Penentuan nilai *threshold* diuji pada 8 dataset yang diklasifikasikan menggunakan algoritma *logistic regression*. Dataset yang digunakan memiliki lebih dari 50 *feature* dengan kategori 6 dataset *multivariate*, 1 dataset *sequential*, dan 1 dataset *univariate*. Proses pengujian dilakukan dengan menghitung nilai *information gain* untuk setiap fitur di setiap dataset, kemudian menentukan nilai *threshold* berdasarkan nilai median. Uji model dilakukan menggunakan *k-fold cross validation* dengan nilai k=10. Hasil dataset yang sudah melalui proses *feature selection* diuji menggunakan model klasifikasi *logistic regression, decision tree, random forest*, dan *naive bayes*, dengan tujuan membandingkan kinerja *logistic regression* dengan algoritma lain. Pengujian algoritma *logistic regression* pada 8 dataset menunjukkan bahwa nilai *accuracy* yang diperoleh untuk semua dataset adalah lebih dari 76%. Berdasarkan hasil pengujian, algoritma *logistic regression* menunjukkan performa unggul dalam klasifikasi menggunakan dataset hasil *feature selection*, dengan *accuracy* yang lebih tinggi dibandingkan algoritma lainnya. *Logistic regression* unggul pada 5 dari 8 uji dataset. Sementara itu, *decision tree* dan *naive bayes* tidak berhasil unggul pada satu pun dataset, sedangkan *random forest* berhasil unggul pada 3 dataset.

Kata kunci: *Feature selection*, *Information gain*, *Logistic regression*, Median, *Threshold*

Implementation of Information Gain for Feature Selection with Threshold Median Value

Steven Liong

ABSTRACT

The development of data science and knowledge discovery in digital applications demands a feature selection process aimed at reducing data dimensions and addressing high computational costs. One of the feature selection methods is information gain. Additionally, the threshold value determined from the information gain can use statistical calculations. Therefore, this research proposes determining the threshold value using the median value of the information gain generated by each feature in the dataset. The determination of the threshold value is tested on 8 datasets classified using logistic regression algorithm. The datasets used have more than 50 features with 6 multivariate dataset categories, 1 sequential dataset, and 1 univariate dataset. The testing process is done by calculating the information gain value for each feature in each dataset, then determining the threshold value based on the median value. Model testing is done using k-fold cross-validation with k=10. The results of datasets that have undergone feature selection are tested using logistic regression, decision tree, random forest, and naive Bayes classification models, with the aim of comparing the performance of logistic regression with other algorithms. Testing logistic regression algorithm on 8 datasets shows that the accuracy obtained for all datasets is more than 76%. Based on the test results, the logistic regression algorithm shows superior performance in classification using feature-selected datasets, with higher accuracy compared to other algorithms. Logistic regression excels in 5 out of 8 dataset tests. Meanwhile, decision tree and naive Bayes did not excel in any dataset, while random forest excelled in 3 datasets.

Keywords: *Feature selection*, *Information gain*, *Logistic regression*, Median, *Threshold*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR KODE

DAFTAR LAMPIRAN