

**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**



**Steven Liong**  
**00000042707**

**UMMN**

**UNIVERSITAS  
MULTIMEDIA  
NUSANTARA**

**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)

UMM  
UNIVERSITAS  
MULTIMEDIA  
NUSANTARA

## 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 Kushadi, S.T, M.Si.)

NIDN: 303037304

Penguji

  
5 Juni 2024

(Arya Wicaksana, S.Kom., M.Eng.Sc.)

(OCA, CEH, CEI)

NIDN: 315109103

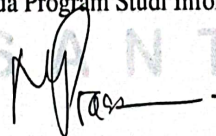
Pembimbing



(Dr. Maria Irmina Prasetyowati, S.Kom., M.T.)

NIDN: 725057201

Pjs. 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

U M M N  
U N I V E R S I T A S  
M U L T I M E D I A  
N U S A N T A R A

\*\* 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)



**UMMN**  
UNIVERSITAS  
MULTIMEDIA  
NUSANTARA

## 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 Prasetyowati, 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*

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA



***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*

U N I V E R S I T A S  
M U L T I M E D I A  
N U S A N T A R A

## DAFTAR ISI

HALAMAN JUDUL . . . . .	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT . . . . .	ii
HALAMAN PENGESAHAN . . . . .	iii
HALAMAN PERSETUJUAN PUBLIKASI ILMIAH . . . . .	iv
HALAMAN PERSEMBAHAN/MOTO . . . . .	v
KATA PENGANTAR . . . . .	vi
ABSTRAK . . . . .	vii
ABSTRACT . . . . .	viii
DAFTAR ISI . . . . .	ix
DAFTAR GAMBAR . . . . .	xi
DAFTAR TABEL . . . . .	xii
DAFTAR KODE . . . . .	xiii
DAFTAR KODE . . . . .	xiii
DAFTAR LAMPIRAN . . . . .	xiv
BAB 1 PENDAHULUAN . . . . .	1
1.1 Latar Belakang Masalah . . . . .	1
1.2 Rumusan Masalah . . . . .	4
1.3 Batasan Permasalahan . . . . .	5
1.4 Tujuan Penelitian . . . . .	5
1.5 Manfaat Penelitian . . . . .	5
1.6 Sistematika Penulisan . . . . .	6
BAB 2 LANDASAN TEORI . . . . .	7
2.1 Tinjauan Teori . . . . .	7
2.1.1 Feature Selection . . . . .	7
2.1.2 Entropy . . . . .	8
2.1.3 Logistic Regression . . . . .	8
2.1.4 Information Gain . . . . .	9
2.1.5 Threshold . . . . .	9
2.1.6 K-Fold Cross Validation . . . . .	10
2.1.7 Confusion Matrix . . . . .	10
BAB 3 METODOLOGI PENELITIAN . . . . .	12
3.1 Metodologi Penelitian . . . . .	12
3.1.1 Studi Literatur . . . . .	12
3.1.2 Pengumpulan dan Pengolahan Data . . . . .	12
3.1.3 Perancangan Model . . . . .	12
3.1.4 Pengujian dan Evaluasi Model . . . . .	13
3.1.5 Penulisan Laporan . . . . .	13
3.2 Perancangan Sistem . . . . .	13
3.2.1 Import Libraries . . . . .	14
3.2.2 Preprocessing . . . . .	14
3.2.3 Calculate Entropy & Information Gain . . . . .	15
3.2.4 Threshold Median . . . . .	18
3.2.5 Feature Selection . . . . .	20
3.2.6 Evaluate Model . . . . .	21
BAB 4 HASIL DAN DISKUSI . . . . .	23
4.1 Spesifikasi Sistem . . . . .	23
4.2 Pengumpulan Dataset . . . . .	23
4.3 Implementasi Sistem . . . . .	24

4.3.1	Import Libraries . . . . .	24
4.3.2	Import Dataset . . . . .	25
4.3.3	Preprocessing . . . . .	26
4.3.4	Calculate Entropy & Information Gain . . . . .	27
4.3.5	Calculate Threshold Median . . . . .	32
4.3.6	Feature Selection . . . . .	33
4.3.7	Evaluate Model . . . . .	34
4.4	Hasil Uji Coba . . . . .	35
4.4.1	AP_Breast_Omentum . . . . .	36
4.4.2	Musk Version 2 . . . . .	36
4.4.3	Arcene . . . . .	37
4.4.4	Internet Advertisement . . . . .	37
4.4.5	Bioresponse . . . . .	38
4.4.6	AP_Colon_Kidney . . . . .	38
4.4.7	Hill Valley . . . . .	39
4.4.8	Nomao . . . . .	40
4.5	Evaluasi Hasil Uji Coba . . . . .	40
BAB 5	SIMPULAN DAN SARAN . . . . .	42
5.1	Simpulan . . . . .	42
5.2	Saran . . . . .	43
	DAFTAR PUSTAKA . . . . .	44



## DAFTAR GAMBAR

Gambar 2.1	Confusion Matrix . . . . .	11
Gambar 3.1	Flowchart Utama . . . . .	13
Gambar 3.2	<i>Preprocessing</i> . . . . .	14
Gambar 3.3	<i>Calculate Total Entropy</i> setiap <i>Feature</i> . . . . .	15
Gambar 3.4	<i>Calculate Entropy Unique Value &amp; Information Gain</i> . . . . .	17
Gambar 3.5	<i>Threshold Median</i> . . . . .	19
Gambar 3.6	<i>Feature Selection</i> . . . . .	20
Gambar 3.7	<i>Evaluate Model</i> . . . . .	21



## DAFTAR TABEL

Tabel 4.1	Dataset . . . . .	24
Tabel 4.2	Classification Report Dataset AP_Breast_Omentum . . . . .	36
Tabel 4.3	Classification Report Dataset Musk (Version 2) . . . . .	37
Tabel 4.4	Classification Report Dataset Arcene . . . . .	37
Tabel 4.5	Classification Report Dataset Internet Advertisements . . . . .	38
Tabel 4.6	Classification Report Dataset Bioresponse . . . . .	38
Tabel 4.7	Classification Report Dataset AP_Colon_Kidney . . . . .	39
Tabel 4.8	Classification Report Dataset Hill Valley . . . . .	39
Tabel 4.9	Classification Report Dataset Nomao . . . . .	40
Tabel 4.10	Dataset Unggul Setiap Model Algoritma . . . . .	40



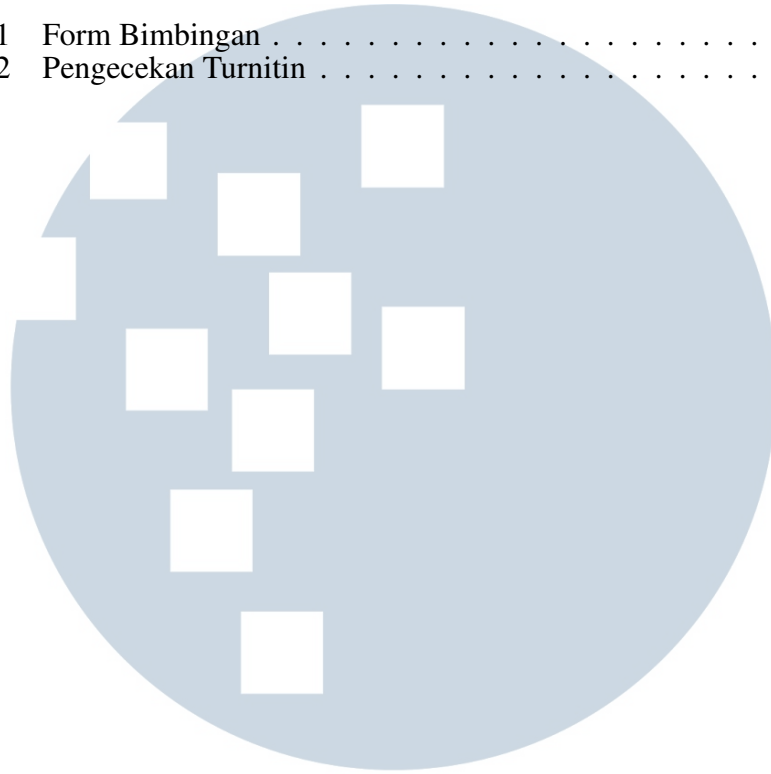
## DAFTAR KODE

Kode 4.1	Potongan kode import libraries . . . . .	24
Kode 4.2	Potongan kode import dataset . . . . .	25
Kode 4.3	Potongan kode label encoding . . . . .	26
Kode 4.4	Potongan kode check null . . . . .	26
Kode 4.5	Potongan kode drop data null . . . . .	26
Kode 4.6	Potongan kode function hitung_ig . . . . .	27
Kode 4.7	Potongan kode entropy total feature . . . . .	28
Kode 4.8	Potongan kode entropy setiap feature . . . . .	30
Kode 4.9	Potongan kode perhitungan information gain . . . . .	31
Kode 4.10	Potongan kode perhitungan threshold nilai median . . . . .	32
Kode 4.11	Potongan kode feature selection . . . . .	33
Kode 4.12	Potongan kode untuk menentukan variabel X_selected dan y . . . . .	34
Kode 4.13	Potongan kode untuk uji dan evaluasi model . . . . .	34



## DAFTAR LAMPIRAN

Lampiran 1	Form Bimbingan . . . . .	47
Lampiran 2	Pengecekan Turnitin . . . . .	48



**UMN**  
UNIVERSITAS  
MULTIMEDIA  
NUSANTARA