

**PENGARUH IMPLEMENTASI TEKNIK RESAMPLING DAN FITUR
SELEKSI PADA DATASET IMBALANCE TERHADAP KINERJA
ALGORITMA KLASIFIKASI**

SKRIPSI



disusun oleh

Lucky Adhkrisna Wirasakti

17.11.1444

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2020**

**PENGARUH IMPLEMENTASI TEKNIK RESAMPLING DAN FITUR
SELEKSI PADA DATASET IMBALANCE TERHADAP KINERJA
ALGORITMA KLASIFIKASI**

SKRIPSI

untuk memenuhi sebagian persyaratan
mencapai gelar Sarjana
pada Program Studi Informatika



disusun oleh

Lucky Adhikrisna Wirasakti

17.11.1444

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2020**

PERSETUJUAN**SKRIPSI****PENGARUH IMPLEMENTASI TEKNIK RESAMPLING DAN FITUR
SELEKSI PADA DATASET IMBALANCE TERHADAP KINERJA
ALGORITMA KLASIFIKASI**

yang dipersiapkan dan disusun oleh

Lucky Adhikrisna Wirasakti

17.11.1444

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 14 Agustus 2020

Dosen Pembimbing,

Mulla Sulistyano, M.Kom.
NIK. 190302248

PENGESAHAN**SKRIPSI****PENGARUH IMPLEMENTASI TEKNIK RESAMPLING DAN FITUR
SELEKSI PADA DATASET IMBALANCE TERHADAP KINERJA
ALGORITMA KLASIFIKASI**

yang dipersiapkan dan disusun oleh

Lucky Adhikrisna Wirasakti

17.11.1444

telah dipertahankan di depan Dewan Penguji
pada tanggal 25 Agustus 2020

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Hendra Kurnawan, M.Kom.

NIK. 190302244

Andriyan Dwi Putra, M.Kom.

NIK. 190302270

Mulla Sulistyono, M.Kom.

NIK. 190302248

Skrripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal

DEKAN FAKULTAS ILMU KOMPUTER

Krisnawati, S.Si, M.T.

NIK. 190302038

PERNYATAAN

Saya yang bertandatangan dibawah ini menyatakan bahwa, skripsi ini merupakan karya saya sendiri (ASLI), dan isi dalam skripsi ini tidak terdapat karya yang pernah diajukan oleh orang lain untuk memperoleh gelar akademis di suatu institusi pendidikan tinggi manapun, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis dan/atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Segala sesuatu yang terkait dengan naskah dan karya yang telah dibuat adalah menjadi tanggungjawab saya pribadi.

Yogyakarta, 04 September 2020



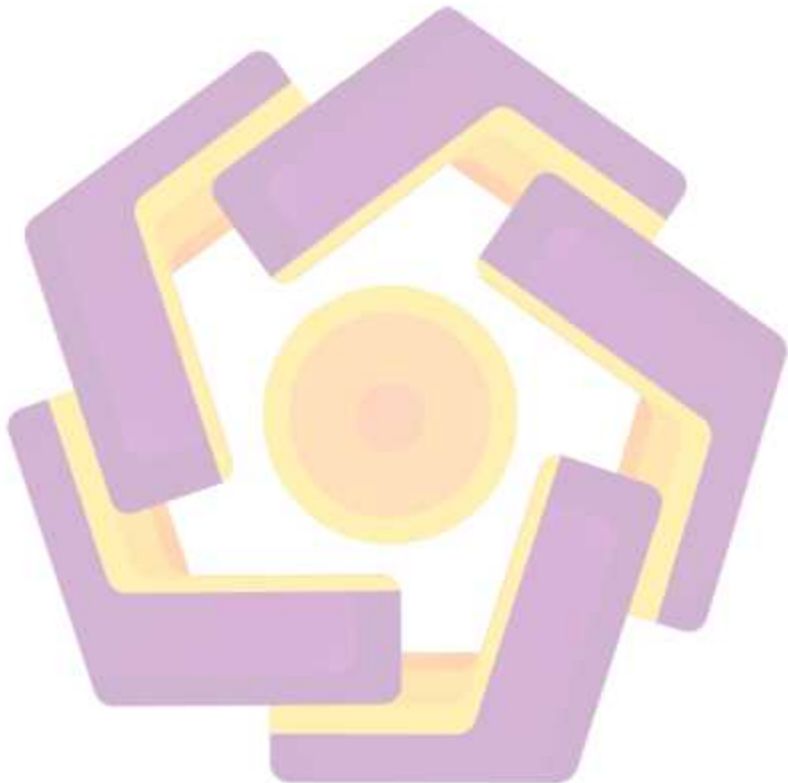
Lucky Adhukrisna Wirasakti

NIM.17.11.1444

MOTTO

“Kita tidak akan pernah tau, sebelum melakukannya. Teruslah berusaha.

Pekerjaan yang kita lakukan tidak akan sia – sia.”



PERSEMBAHAN

Segala puji dan syukur kehadiran Allah SWT atas berkat, rahmat dan hidayah-Nya sehingga penulis mampu menyelesaikan skripsi dengan baik. Penulis juga mengucapkan terima kasih kepada pihak – pihak yang telah berkontribusi baik secara langsung maupun tidak langsung baik dalam penelitian maupun dalam penyusunan naskah. Skripsi dipersembahkan kepada:

1. Orang tua dan saudara penulis yang senantiasa memberikan dukungan dan doa sehingga dapat menempuh pendidikan dan menyelesaikan skripsi ini dengan baik.
2. Bapak Mulia Sulistiyono, M. Kom selaku pembimbing yang telah dengan sabar dan tulus membimbing dan memberikan ilmu kepada penulis.
3. Bapak Yoga Pristyanto, S. Kom, M. Eng yang telah berkenan berbagi pengetahuan sebagai bekal dan memberikan dukungan sehingga penelitian ini dapat terlaksana.
4. Teman teman mahasiswa S1 – Informatika – 08 yang telah banyak bekerja sama dalam masa menempuh pendidikan.
5. Semua pihak keluarga besar Universitas Amikom Yogyakarta yang tidak bisa kami sebutkan safu-persatu.

KATA PENGANTAR

Segala puji dan syukur kehadiran Allah SWT atas berkat, rahmat dan hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi dengan judul “Pengaruh Implementasi Teknik *Resampling* Dan Fitur Seleksi Pada *Dataset Imbalance* Terhadap Kinerja Algoritma Klasifikasi” sebagai syarat untuk menyelesaikan Program Sarjana (S1) pada Fakultas Ilmu Komputer Universitas Amikom Yogyakarta.

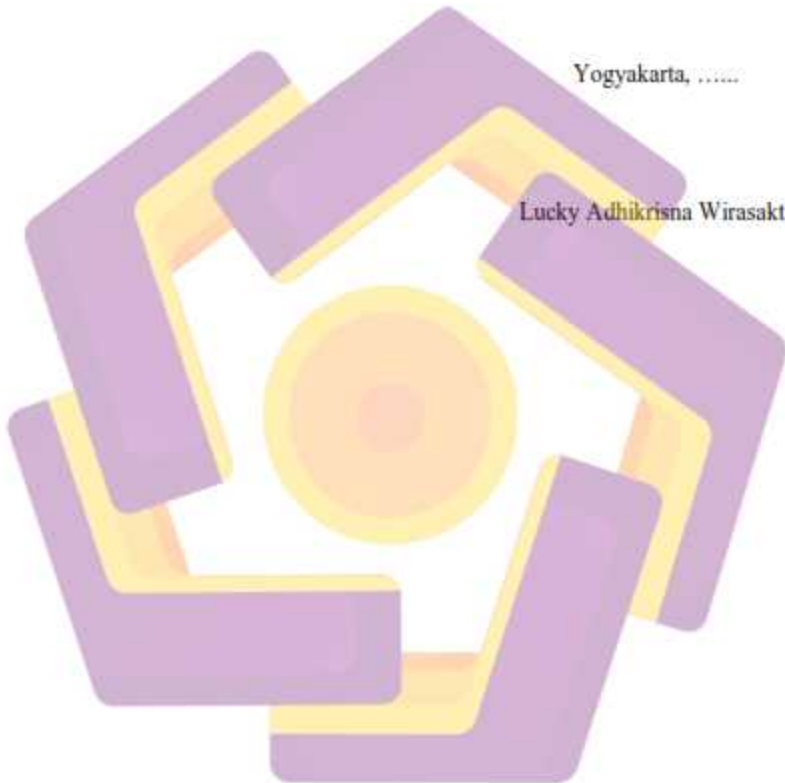
Dalam penyusunan skripsi banyak hambatan dan rintangan yang dihadapi namun pada akhirnya dapat dilalui karena bantuan dan bimbingan dari berbagai pihak baik secara moral dan spiritual. Pada kesempatan ini penulis mengucapkan terima kasih kepada:

1. Bapak M. Suyanto, Prof., Dr., M.M. selaku Rektor Universitas AMIKOM Yogyakarta.
2. Bapak Mulia Sulistiyono, M. Kom selaku pembimbing yang telah dengan sabar dan tulus membimbing dan memberikan ilmu kepada penulis.
3. Bapak Yoga Pristyanto, S. Kom, M. Eng yang telah berkenan berbagi pengetahuan sebagai bekal dan memberikan dukungan sehingga penelitian ini dapat terlaksana.
4. Seluruh jajaran dosen Program Studi S1 Informatika Fakultas Ilmu Komputer yang telah memberikan bekal ilmu pengetahuan selama perkuliahan.
5. Semua pihak tidak bisa disebutkan satu persatu yang telah membantu sehingga penelitian ini bisa dilaksanakan.

Penulis menyadari bahwa laporan skripsi ini masih jauh dari kata sempurna, untuk itu kritik dan saran yang membangun sangat diharapkan. Akhir kata semoga tulisan ini dapat memberikan manfaat bagi pembaca dan khususnya penulis sendiri.

Yogyakarta,

Lucky Adhikrisna Wirasakti



DAFTAR ISI

HALAMAN SAMPUL.....	i
HALAMAN JUDUL.....	ii
PERSETUJUAN.....	iii
PENGESAHAN.....	iv
PERNYATAAN.....	v
MOTTO.....	vi
PERSEMBAHAN.....	vii
KATA PENGANTAR.....	viii
DAFTAR ISI.....	x
DAFTAR GAMBAR.....	xiv
DAFTAR TABEL.....	xvi
ARTI LAMBANG DAN SINGKATAN.....	xviii
INTISARI.....	xix
<i>ABSTRACT</i>	xx
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang Masalah.....	1
1.2 Rumusan Masalah.....	4
1.3 Batasan Penelitian.....	4

1.4	Tujuan Penelitian	5
1.5	Manfaat Penelitian	6
1.6	Metode Penelitian	7
1.7	Sistematika Penulisan	8
BAB II	TINJAUAN PUSTAKA DAN LANDASAN TEORI	10
2.1	Tinjauan Pustaka	10
2.2	Landasan Teori	20
2.2.1	Data Mining	20
2.2.2	Adaptive Synthetic k – Nearest Neighbor	20
2.2.3	Fitur Seleksi	24
2.2.3.1	Attribute Selection Measure	24
2.2.3.2	Information Gain	25
2.2.4	Algoritma Classifier	26
2.2.4.1	Naive Bayes	26
2.2.4.2	Decision Tree (C.45)	29
2.2.5	Data Preprocessing	33
2.2.6	Dataset Splitting	35
2.2.7	Evaluasi Model	35
BAB III	METODE PENELITIAN	38
3.1	Gambaran Umum	38

3.2	Alat dan Bahan.....	38
3.1.1	Alat.....	38
3.1.2	Bahan.....	39
3.2	Jalannya Penelitian.....	42
3.3	Pra Pemrosesan Data.....	43
3.3.1	Penanganan Ketidakseimbangan Kelas.....	44
3.3.2	Proses Transformasi Data (Normalisasi).....	50
3.4	Fitur Seleksi Atribut.....	52
3.5	Klasifikasi.....	55
3.6	Evaluasi.....	56
BAB IV	HASIL DAN PEMBAHASAN.....	57
4.1	Implementasi Resampling terhadap Dataset.....	57
4.1.1	Dataset.....	57
4.2	Pra Pemrosesan Data.....	58
4.2.1	Implementasi Algoritma ADASYN k-Nearest Neighbor.....	59
4.2.2	Implementasi Algoritma Min Max Normalization.....	68
4.3	Fitur Seleksi.....	70
4.3.1	Implementasi Algoritma Information Gain.....	70
4.4	Implementasi Algoritma Klasifikasi.....	73
4.4.1	Implementasi Algoritma Naïve Bayes.....	73

4.4.1.1	Performa Akurasi Algoritma Naïve Bayes	73
4.4.1.2	Performa Geometric Mean Algoritma Naïve Bayes	74
4.4.2	Implementasi Algoritma Decision Tree	76
4.4.2.1	Performa Akurasi Algoritma Decision Tree	76
4.4.2.2	Performa Geometric Mean Decision Tree	77
BAB V	KESIMPULAN DAN SARAN	81
5.1	Kesimpulan	81
5.2	Saran.....	83
DAFTAR PUSTAKA	84



DAFTAR GAMBAR

Gambar 2.2.1 Distribusi Koefisien Beta Algoritma ADASYN [21].	21
Gambar 2.2.2 Ilustrasi Skenario Bergantung Karakteristik Data [22].	33
Gambar 2.2.3 Distribusi Min Max Normalization [22].	34
Gambar 2.2.4. Ilustrasi Skema Data Splitting	35
Gambar 3.2.1. Skenario Pertama Alur Penelitian	43
Gambar 3.2.2 Skenario Kedua Alur Penelitian	43
Gambar 3.3.1. Diagram Alur Proses <i>Class Balancing</i>	44
Gambar 3.5.1 Skenario Perbandingan Klasifikasi	55
Gambar 4.1.1 Grafik Distribusi Kelas Dataset	58
Gambar 4.2.1 Ecoli IR 8,6 Sebelum Proses ADASYN	59
Gambar 4.2.2 Ecoli IR 8,6 Setelah Proses ADASYN	60
Gambar 4.2.3 Wine Quality IR 26 Sebelum Proses ADASYN	61
Gambar 4.2.4 Wine Quality IR 26 Setelah Proses ADASYN	62
Gambar 4.2.5 Mammography IR 42 Sebelum Proses ADASYN	63
Gambar 4.2.6 Mammography IR 42 Setelah Proses ADASYN	64
Gambar 4.2.7 Mammography IR 14 Sebelum Proses ADASYN	65
Gambar 4.2.8 Mammography IR 14 Setelah Proses ADASYN	66
Gambar 4.2.9 Mammography IR 14 Sebelum Proses ADASYN	67
Gambar 4.2.10 Mammography IR 14 Setelah Proses ADASYN	68
Gambar 4.2.11 Distribusi Dataset Sebelum dan Setelah Proses Min Max Normalization	70

Gambar 4.4.1 Nilai Akurasi Pelatihan Naïve Bayes	73
Gambar 4.4.2 Nilai Akurasi Pengujian Naïve Bayes	74
Gambar 4.4.3 Nilai Geometric Mean Pelatihan Naïve Bayes	75
Berbeda dengan pada proses pelatihan pada saat pengujian model dataset dengan skenario ke 3 dengan fitur seleksi lebih mendominasi dengan hasil lebih baik, dan terlihat penurunan performa untuk nilai geometric mean pada skenario 1 (dataset asli). Perhatikan Gambar 4.4.4 dibawah:	75
Gambar 4.4.4 Nilai Geometric Mean Pelatihan Naïve Bayes	75
Gambar 4.4.5 Nilai Akurasi Pelatihan Decision Trees	76
Gambar 4.4.6 Nilai Akurasi Pengujian Decision Trees	77
Gambar 4.4.7 Nilai Geometric Mean Pelatihan Decision Trees	77
Gambar 4.4.8 Nilai Geometric Mean Pengujian Decision Trees	78

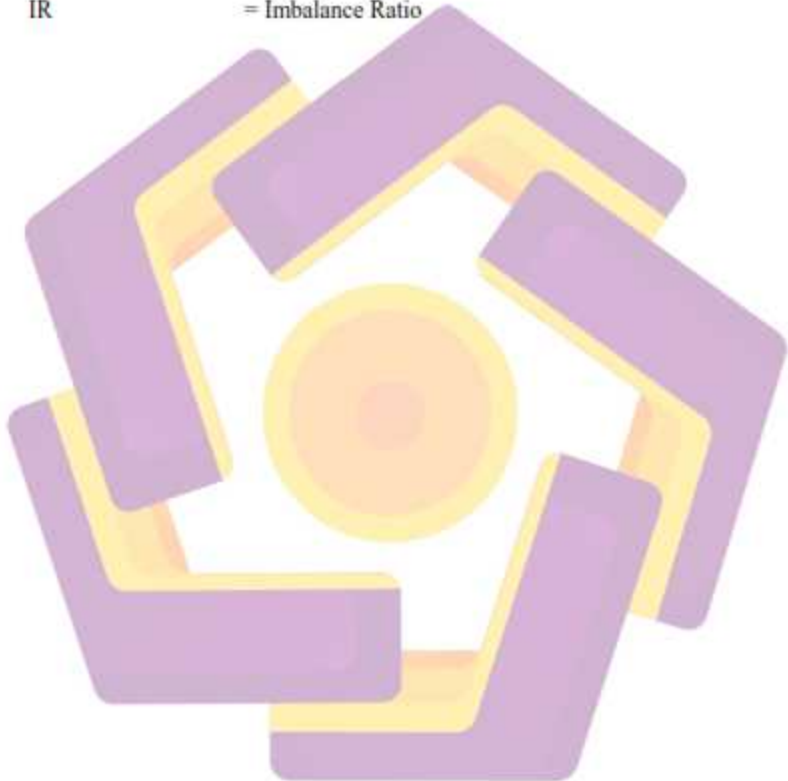
DAFTAR TABEL

Tabel 2.1.1 Tabel Perbandingan Resampling ADASYN.....	15
Tabel 2.1.2 Tabel Perbandingan Seleksi Fitur Information Gain	19
Tabel 2.2.1 Confusion Matrix	36
Tabel 3.2.1. Karakteristik Dataset.....	39
Tabel 3.2.2. Ecoli	40
Tabel 3.2.3. Yeast M18.....	40
Tabel 3.2.4. Libras Move.....	41
Tabel 3.2.5. Wine Quality.....	41
Tabel 3.2.6. Mammography.....	42
Tabel 3.3.1 Dataset Asli Sebelum Proses ADASYN.....	45
Tabel 3.3.2 Nearest Neighbor untuk Data 1 (D1).....	46
Tabel 3.3.3 Nearest Neighbor Untuk Setiap Data Minoritas	47
Tabel 3.3.4 Density Distribution Untuk Setiap Data Minoritas.....	47
Tabel 3.3.5 Jumlah Duplikasi Data Sintesis	48
Tabel 3.3.6 Majority Voting Untuk Data 1.....	48
Tabel 3.3.7 Majority Voting Untuk Data 2.....	48
Tabel 3.3.8 Majority Voting Untuk Data 3.....	49
Tabel 3.3.9 Dataset Setelah Proses Sampling ADASYN	49
Tabel 3.3.10 Dataset Sebelum Proses Normalisasi.....	50
Tabel 3.3.11 Dataset Setelah Proses Normalisasi	51
Tabel 3.4.1 Dataset Sebelum Proses Seleksi Fitur.....	52

Tabel 3.4.2 Nilai Entropi Pada Masing Masing Fitur	53
Tabel 3.4.3 Nilai Gain Pada Setiap Fitur	54
Tabel 3.4.4 Dataset Setelah Proses Fitur Seleksi	54
Tabel 4.3.1 Peringkat Fitur Ecoli Berdasar Nilai Gain	70
Tabel 4.3.2 Peringkat Fitur Wine Quality Berdasar Nilai Gain	71
Tabel 4.3.3 Peringkat Fitur Mammography Berdasar Nilai Gain	71
Tabel 4.3.4 Hasil Fitur Seleksi	72
Tabel 4.4.1 Lampiran Hasil Akurasi Pelatihan Algoritma Decision Tree	79
Tabel 4.4.2 Lampiran Hasil Geometric Mean Pelatihan Algoritma Decision Tree	79
Tabel 4.4.3 Lampiran Hasil Akurasi Pengujian Algoritma Naïve Bayes	80
Tabel 4.4.4 Lampiran Hasil Geometric Mean Pengujian Algoritma Naïve Bayes	80

ARTI LAMBANG DAN SINGKATAN

ADASYN k-NN	= Adaptive Synthetic k-Nearest Neighbor
IG	= Information Gain
G-mean	= Geometric Mean
IR	= Imbalance Ratio



INTISARI

Permasalahan klasifikasi yang sering dijumpai adalah ketidakseimbangan kelas yang berpotensi menimbulkan resiko terjadinya kesalahan klasifikasi akibat dominasi kelas mayoritas dan menganggap kelas minoritas sebagai *noise sample*. Penelitian ini ditujukan sebagai tindakan penanganan terhadap ketidakseimbangan kelas pada *dataset* digunakan teknik *resampling Adaptive Synthetic k - Nearest Neighbor* dan reduksi dimensi fitur berdasarkan nilai rangking dari relevansi fitur menggunakan fitur seleksi *Information Gain*.

Dalam penelitian ini melibatkan tiga skenario dengan lima *dataset* yang memiliki *imbalance ratio* (IR) yang berbeda – beda yang dievaluasi berdasarkan akurasi dan *geometric – mean*. Skenario 1 mencapai performa terbaik pada *dataset mammography* dengan nilai *imbalance ratio* 42 dilakukan klasifikasi menggunakan algoritma *Decision Tree* mendapat hasil *geometric mean* (*train*: 98,59%; *test*: 72,72%) dan akurasi (*train*: 99,93%; *test*: 98,25%) dan algoritma *Naïve Bayes* mendapat hasil *geometric mean* (*train*: 82,97%; *test*: 86,03%) dan akurasi (*train*: 95,53%; *test*: 94,81%), Skenario 2 mencapai performa terbaik pada *dataset libras move* dengan nilai *imbalance ratio* 14 diklasifikasi menggunakan algoritma *Decision Tree* mendapat hasil *geometric mean* (*train*: 4,16%; *test*: 99,31%) dan akurasi (*train*: 4,16%; *test*: 99,25%) dan algoritma *Naïve Bayes* mendapat hasil *geometric mean* (*train*: 87,11%; *test*: 84,03%) dan akurasi (*train*: 87,15%; *test*: 85,18%), dan Skenario 3 mencapai performa terbaik pada *dataset libras move* dengan *imbalance ratio* 14 diklasifikasi menggunakan algoritma *Decision Tree* mendapat hasil *geometric mean* (*train*: 4,16%; *test*: 98,62%) dan akurasi (*train*: 4,16%; *test*: 98,51%) sedangkan pada algoritma *Naïve Bayes* mencapai performa paling baik pada *dataset ecoli* dengan *imbalance ratio* 8,6 dengan hasil *geometric mean* (*train*: 90,29%; *test*: 89,97%) dan akurasi (*train*: 90,62%; *test*: 90,08%).

Dengan peningkatan hasil performa *geometric mean*, maka kemampuan klasifikasi juga meningkat dan lebih peka terhadap data minoritas sehingga hasil penelitian ini diharapkan mampu dijadikan referensi terhadap penanganan kasus ketidakseimbangan kelas.

Kata Kunci: Adaptive Synthetic k - Nearest Neighbor, Information Gain, Ketidakseimbangan Kelas, Pra pemrosesan, Decision Tree, Naïve Bayes, Fitur Seleksi.

ABSTRACT

Classification problems that are often encountered are class imbalances that have the potential to cause the risk of misclassification due to the domination of the majority class and to regard the minority class as a noise sample. This research is intended as an action to treat class imbalance in the dataset using the Adaptive Synthetic k - Nearest Neighbor resampling technique and feature dimension reduction based on the ranking value of feature relevance using the Information Gain selection feature.

This research involved three scenarios with five datasets that have different imbalance ratio (IR) evaluated based on their accuracy and geometric - mean. Scenario 1 achieves the best performance on the mammography dataset with an imbalance ratio of 42 and is classified using the Decision Tree algorithm to get the geometric mean (train: 98.59%; test: 72.72%) and accuracy (train: 99.93%; test: 98.25%) and the Naïve Bayes algorithm gets geometric mean results (train: 82.97%; test: 86.03%) and accuracy (train: 95.53%; test: 94.81%), Scenario 2 achieves performance The best on the Libras move dataset with an imbalance ratio value of 14 is classified using the Decision Tree algorithm, the results are geometric mean (train: 4.16%; test: 99.31%) and accuracy (train: 4.16%; test: 99.25%)) and the Naïve Bayes algorithm gets geometric mean results (train: 87.11%; test: 84.03%) and accuracy (train: 87.15%; test: 85.18%), and Scenario 3 achieves the best performance on the dataset libras moves with an imbalance ratio of 14 are classified using the Decision Tree algorithm to obtain geometric mean results (train: 4.16%; test: 98.62%) and accuracy (train: 4.16%; test: 98.51%) s while the Naïve Bayes algorithm achieves the best performance on the ecoli dataset with an imbalance ratio of 8.6 with the geometric mean results (train: 90.29%; test: 89.97%) and accuracy (train: 90.62%; test: 90.08%).

With the increase in the results of geometric mean performance, the classification ability also increases and is more sensitive to minority data so that the results of this study are expected to be used as a reference for handling cases of class imbalance.

Keywords: Adaptive Synthetic k - Nearest Neighbor, Information Gain, Class Imbalance, Pre-processing, Decision Tree, Naïve Bayes, Selection Features.