

**IMPLEMENTASI METODE ENSEMBLE PADA ALGORITMA
KLASIFIKASI TERHADAP KASUS IMBALANCED DATASET**

SKRIPSI



disusun oleh :

Adltya Ahmad Zeln

17.11.1401

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2021**

**IMPLEMENTASI METODE ENSEMBLE PADA ALGORITMA
KLASIFIKASI TERHADAP KASUS IMBALANCED DATASET**

SKRIPSI

untuk memenuhi sebagian persyaratan
mencapai gelar Sarjana
pada Program Studi Informatika



disusun oleh :

Aditya Ahmad Zeln

17.11.1401

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2021**

PERSETUJUAN

SKRIPSI

IMPLEMENTASI METODE ENSEMBLE PADA ALGORITMA KLASIFIKASI TERHADAP KASUS IMBALANCED DATASET

yang dipersiapkan dan disusun oleh

Aditya Ahmad Zeln

17.11.1401

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 28 Februari 2021

Dosen Pembimbing,

Yoga Pristyanto, S.Kom., M.Eng.

NIK: 190302412

PENGESAHAN

SKRIPSI

IMPLEMENTASI METODE ENSEMBLE PADA ALGORITMA KLASIFIKASI TERHADAP KASUS IMBALANCED DATASET

yang dipersiapkan dan disusun oleh
Aditya Ahmad Zeln

17.11.1401

telah dipertahankan di depan Dewan Penguji
pada tanggal

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Windha Mega Pradnya D, M.Kom
NIK. 190302185

Anna Balta, M.Kom
NIK. 190302290

Yoga Pristyanto, S.Kom, M.Eng
NIK. 190302412

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

DEKAN FAKULTAS ILMU KOMPUTER

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

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, 26 Februari 2021

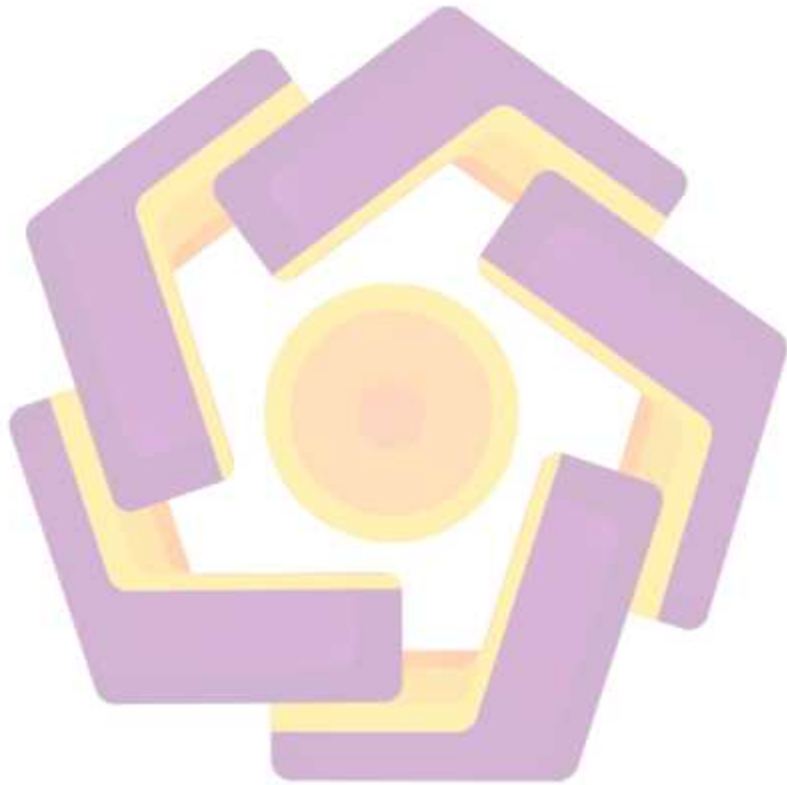


Aditya Ahmad Zein

NIM. 17.11.1401

MOTTO

“Sesungguhnya Allah tidak akan mengubah nasib suatu kaum hingga mereka mengubah diri mereka sendiri,” (QS. Ar-Ra'd:11).



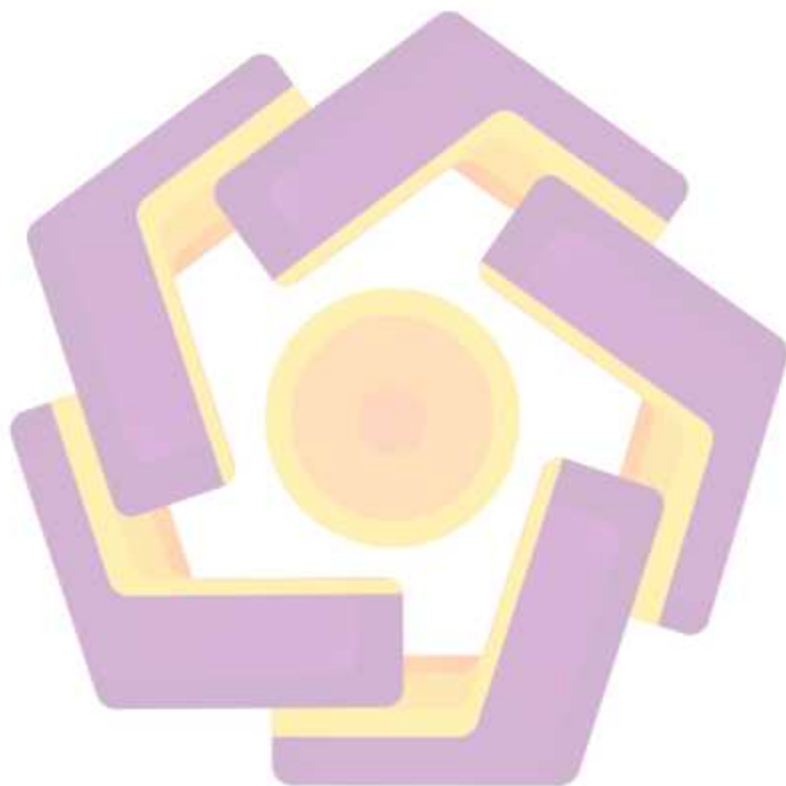
PERSEMBAHAN

Alhamdulillah penulis panjatkan puji-syukur kepada Allah SWT atas segala rahmat, taufiq, serta hidayah-Nya, sehingga diberi kesempatan untuk dapat menyelesaikan skripsi ini dengan sebaik-baiknya dengan segala kekurangan penulis. Segala syukur penulis ucapkan kepada-Mu karena telah menghadirkan mereka yang memberikan semangat dan doa disaat menjalani proses pembuatan skripsi ini. Dengan segala kerendahan hati saya persembahkan skripsi ini kepada :

1. Kedua Orang Tua, Bapak Muh. Muslimin, M. Ag Dan Ibu Hindun Sukma Dj., S. Pd yang selalu mendoakan, memberi semangat serta motivasi supaya dapat menyelesaikan skripsi ini dengan lancar serta bermanfaat bagi semua.
2. Adikku Avillian Aufa yang tidak berhenti untuk tetap menghibur hingga memberi semangat pada saya dalam mengerjakan skripsi.
3. Bpk. Yoga Pristyanto, S.Kom, M.Eng selaku dosen pembimbing dalam skripsi ini yang tidak lelah untuk tetap membimbing dan mengingatkan penulis dari awal hingga akhir proses pembuatan skripsi.
4. Sdr. Lucky Adhikrisna Wirasakti, S.Kom. selaku teman sekaligus mentor disaat pembuatan program sedang berlangsung.
5. Dosen-dosen Universitas Amikom Yogyakarta yang telah memberikan banyak ilmu baik ilmu akademik maupun ilmu non-akademik selama kuliah.
6. Keluarga besar kelas 17-S11F-08 yang telah bersama-sama menemani selama kuliah. Semoga silaturahmi kita tetap terjaga.
7. Keluarga besar UKM UKI Jashtis yang selalu bersama-sama berjalan di dalam dakwah selama kuliah di Amikom. Semoga kita tetap bersama sampai di surga nanti.
8. Serta orang-orang yang selalu membantu peneliti dalam mengerjakan skripsi yang tidak bisa disebut namanya satu-persatu.

Saya ucapkan terima kasih yang sebesar-besarnya untuk kalian semua. Mohon maaf jika ada salah kata atau perbuatan baik yang disengaja maupun tidak disengaja

selama ini. Sukses untuk kalian semua, semoga Allah SWT memberikan rahmat dan hidayah-Nya kepada kita semua. Dan semoga skripsi ini dapat bermanfaat dan berguna untuk kemajuan ilmu pengetahuan kedepannya.



KATA PENGANTAR

Alhamdulillah penulis panjatkan puji syukur kepada Allah SWT atas segala rahmat, taufiq, serta hidayah-Nya kepada penulis sehingga dapat menyelesaikan skripsi yang berjudul "Implementasi Metode Ensemble Pada Algoritma Klasifikasi Terhadap Kasus Imbalanced Dataset".

Selama proses pengerjaan skripsi ini penulis menyadari bahwa dalam proses penulisan skripsi ini banyak mengalami kendala, namun berkat bantuan, bimbingan, kerjasama dari berbagai pihak dan berkah dari Allah SWT sehingga kendala-kendala tersebut bisa diatasi. Selanjutnya ucapan terima kasih penulis sampaikan kepada :

1. Bapak Prof. Dr. M. Suyanto, M.M selaku Rektor Universitas Amikom Yogyakarta.
2. Bpk. Yoga Pristyanto, S.Kom, M.Eng selaku dosen pembimbing yang telah memberikan banyak masukan yang membantu membimbing dalam menyelesaikan skripsi ini.
3. Ibu Krisnawati, S.Si, M.T selaku Dekan Fakultas Ilmu Komputer Universitas Amikom Yogyakarta.
4. Bapak Sudarmawan, M.T selaku Kepala Prodi Informatika Universitas Amikom Yogyakarta.
5. Dosen Penguji (Ibu Windha Mega Pradnya D, M.Kom dan Ibu Anna Baita, M.Kom) yang telah memberikan masukan terhadap penelitian ini.
6. Kedua orang tua dan keluarga yang selalu memberikan doa, dukungan dan semangat.
7. Keluarga Besar UKM UKI Jashtis dan Keluarga Besar 17-S11F-08.
8. Serta semua pihak yang tidak bisa penulis sebutkan satu-persatu yang telah membantu dalam penyusunan skripsi ini.

Penulis Menyadari bahwa masih banyak terdapat kekurangan-kekurangan dalam mengerjakan skripsi ini, sehingga penulis mengharapkan adanya saran dan kritik yang membangun demi kesempurnaan skripsi ini.

Yogyakarta, 28 Februari 2021

Aditya Ahmad Zein
NIM. 17.11.1401

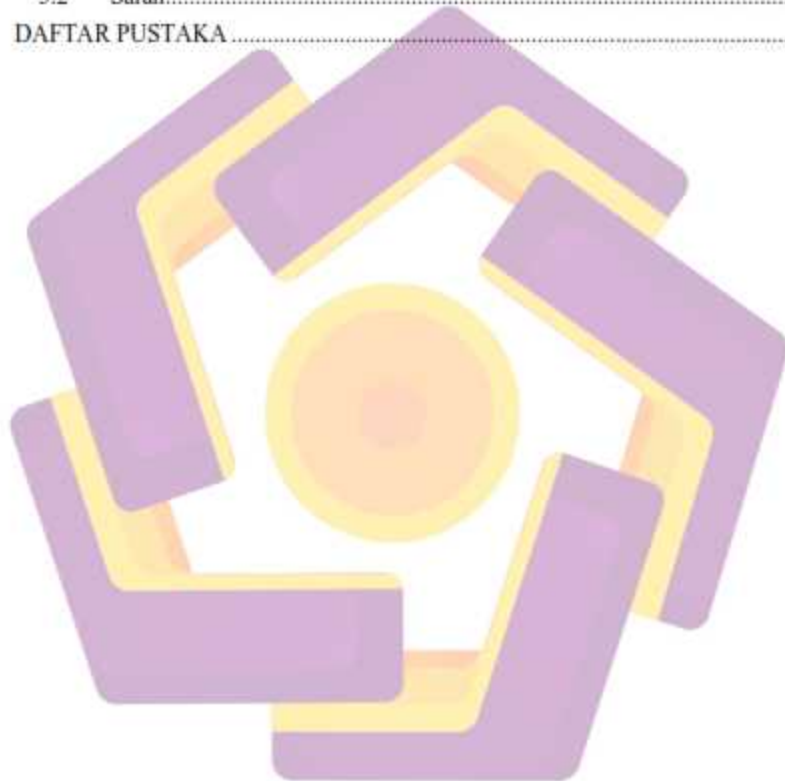


DAFTAR ISI

HALAMAN SAMPUL	i
HALAMAN JUDUL	ii
PERSETUJUAN	iii
PENGESAHAN	iv
PERNYATAAN	v
MOTTO	vi
PERSEMBAHAN	vii
KATA PENGANTAR	ix
DAFTAR ISI	xi
DAFTAR GAMBAR	xiv
DAFTAR TABEL	xv
INTISARI	xvi
ABSTRACT	xvii
BAB I	1
PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	5
1.3 Batasan Penelitian	5
1.4 Tujuan Penelitian	6
1.5 Manfaat Penelitian	7
1.6 Metode Penelitian	7
1.6.1 Metode Pengumpulan Data	7
1.6.2 Metode Klasifikasi dan Penanganan Ketidakseimbangan Kelas	8
1.6.3 Metode Evaluasi	8
1.7 Sistematika Penulisan	8
BAB II	10
TINJUAN PUSTAKA DAN LANDASAN TEORI	10
2.1 Tinjauan Pustaka	10
2.2 Landasan Teori	23

2.2.1	Data Mining	23
2.2.2	Bagging (Bootstrap Aggregating)	23
2.2.3	Balanced-Bagging (Undersampling-Bagging)	25
2.2.4	Decision Tree	27
2.2.5	Logistic Regression	33
2.2.6	Data Preprocessing	35
2.2.7	Stratified K-Fold Cross Validation	36
2.2.8	Evaluasi Model	37
BAB III		40
METODOLOGI PENELITIAN		40
3.1	Gambaran Umum	40
3.2	Alat dan Bahan	40
3.2.1	Alat	40
3.2.2	Bahan	41
3.3	Alur Penelitian	45
3.4	Pra Pemrosesan Data	49
3.5	Klasifikasi dan Penanganan Ketidakseimbangan Kelas	49
3.5.1	Klasifikasi dan Penanganan Ketidakseimbangan Kelas Menggunakan Bagging	49
3.5.2	Klasifikasi dan Penanganan Ketidakseimbangan Kelas Menggunakan Balanced-Bagging	53
3.6	Evaluasi	57
3.6.1	Evaluasi Bagging	57
3.6.2	Evaluasi Balanced-Bagging	58
BAB IV		60
HASIL DAN PEMBAHASAN		60
4.1	Dataset	60
4.2	Pra Pemrosesan Data	61
4.3	Implementasi Algoritma Klasifikasi	62
4.3.1	Implementasi Algoritma Klasifikasi (tanpa Bagging)	63
4.3.2	Implementasi Algoritma Klasifikasi dengan Bagging	65
4.3.3	Implementasi Algoritma Klasifikasi dengan Balanced-Bagging	67

4.4	Lampiran	69
BAB V		74
KESIMPULAN DAN SARAN		74
5.1	Kesimpulan	74
5.2	Saran	75
DAFTAR PUSTAKA		76



DAFTAR GAMBAR

Gambar 2.2.1 Tahapan pada metode bagging	24
Gambar 2.2.2 Ilustrasi Plot Klasifikasi pada Pohon Keputusan	27
Gambar 2.2.3 Model Pohon Keputusan	29
Gambar 2.2.4 Ilustrasi Skenario Bergantung Karakteristik Data [2].	32
Gambar 2.2.5 Ilustrasi Alur Kerja Dari Stratified K-Fold Cross Validation.	37
Gambar 3.3.1 Skenario Pertama Pada Tiap-tiap Proses Klasifikasi	46
Gambar 3.3.2 Skenario Kedua Alur Penelitian Pada Tiap-tiap Proses Klasifikasi Menggunakan bagging	47
Gambar 3.3.3 Skenario Ketiga Alur Penelitian Pada Tiap-tiap Proses Klasifikasi Menggunakan balanced-bagging.	48
Gambar 3.5.1 Diagram Alur Proses Class Balancing (Bootstrap Aggregating)	50
Gambar 3.5.2 Diagram Alur Proses Balanced-Bagging	54
Gambar 3.5.3 Diagram Alur Proses Penyeimbangan Data Oleh Undersampling.....	56
Gambar 4.1.1 Grafik Distribusi Kelas Dataset.....	61
Gambar 4.3.1 Diagram Hasil Akurasi Decision Tree	63
Gambar 4.3.2 Diagram Hasil Akurasi Logistic Regression.....	64
Gambar 4.3.3 Diagram Hasil Geometric Mean Decision Tree.....	64
Gambar 4.3.4 Diagram Hasil Geometric Mean Logistic Regression.....	65
Gambar 4.3.5 Diagram Hasil Akurasi Bagging Pada Decision Tree dan Logistic Regression	66
Gambar 4.3.6 Diagram Hasil Geometric Mean Bagging Pada Decision Tree dan Logistic Regression.....	66
Gambar 4.3.7 Diagram Perbandingan Hasil Akurasi Bagging dan Balanced-Bagging Pada Decision Tree	67
Gambar 4.3.8 Diagram Perbandingan Hasil Akurasi Bagging dan Balanced-Bagging Pada Logistic Regression	68
Gambar 4.3.9 Diagram Perbandingan Hasil Geometric Mean Bagging dan Balanced-Bagging pada Decision Tree	68
Gambar 4.3.9 Diagram Perbandingan Hasil Geometric Mean Bagging dan Balanced-Bagging Pada Logistic Regression	69

DAFTAR TABEL

Tabel 2.1.1 Tabel Perbandingan Metode Bagging.....	18
Tabel 2.2.1 Confusion Matrix.....	38
Tabel 3.2.1 Karakteristik Dataset.....	41
Tabel 3.2.2 pen_digits.....	42
Tabel 3.2.3 us_crime.....	42
Tabel 3.2.4 yeast_me2.....	43
Tabel 3.2.5 protein_homo.....	44
Tabel 3.2.6 mammography.....	45
Tabel 3.5.1 Dataset Asli Sebelum Melalui Proses Bootstrap.....	51
Tabel 3.5.2 Bootstrap Sample yang Dibuat Dari Dataset Asli.....	51
Tabel 3.5.3 Agregasi dari Bootstrap Classifier Untuk Mencari Perkiraan Probabilitas.....	52
Tabel 3.5.4 Subset Bootstrap yang Dibuat Dari Dataset Asli.....	54
Tabel 3.5.5 Agregasi dari Bootstrap Classifier Setelah Proses Undersampling.....	56
Tabel 3.6.1 Bagging Confusion Matrix.....	57
Tabel 3.6.2 Bagging Confusion Matrix.....	58
Tabel 4.2.1 Hasil Dataset yang Menunjukkan Nilai Missing Value.....	61
Tabel 4.2.2 Hasil dari Proses Splitting Dataset Menggunakan 5-Fold Cross Validation.....	62
Tabel 4.4.1 Lampiran Nilai Akurasi Decision Tree.....	69
Tabel 4.4.2 Lampiran Nilai Geometric Mean Decision Tree.....	70
Tabel 4.4.3 Lampiran Nilai Akurasi Logistic Regression.....	70
Tabel 4.4.4 Lampiran Nilai Geometric Mean Logistic Regression.....	70
Tabel 4.4.5 Lampiran Perbandingan Nilai Akurasi Bagging Berdasarkan Base Classifier.....	71
Tabel 4.4.6 Lampiran Perbandingan Nilai G-Mean Bagging Berdasarkan Base Classifier.....	71
Tabel 4.4.7 Lampiran Perbandingan Nilai Akurasi Bagging Balanced-Bagging Pada Decision Tree.....	71
Tabel 4.4.8 Lampiran Perbandingan Nilai Geometric Mean Bagging Balanced-Bagging Pada Decision Tree.....	72
Tabel 4.4.9 Lampiran Perbandingan Nilai Accuracy Bagging Balanced-Bagging Pada Logistic Regression.....	72
Tabel 4.4.10 Lampiran Perbandingan Nilai Geometric Mean Bagging Balanced-Bagging Pada Logistic Regression.....	73

INTISARI

Algoritma klasifikasi merupakan algoritma yang sangat sering digunakan beriringan dengan kebutuhan manusia, namun peneliti sering menjumpai kendala saat menggunakan algoritma klasifikasi ini. Salah satu permasalahan yang sering sekali dijumpai oleh peneliti ialah kasus *imbalanced dataset*. Sehingga dalam penelitian ini peneliti menyarankan *ensemble method* untuk mengatasinya salah satu algoritma *ensemble method* yang terkenal ialah *bagging*. Dan disini peneliti juga akan menyarankan algoritma *balanced-bagging* untuk meningkatkan kemampuan dari *bagging*.

Dalam penelitian ini melibatkan tiga proses klasifikasi berbeda dengan lima dataset yang memiliki *imbalanced ratio* (IR) yang berbeda – beda yang dievaluasi berdasarkan hasil akurasi (*balanced accuracy*) dan *geometric mean*. Proses Pertama merupakan proses klasifikasi pada *Base Classifier* (tanpa *Bagging*) dimana mencapai performa terbaiknya pada dataset *pen_digits* dengan rata - rata hasil akurasi (92,66%) dan *g-mean* (91,85%). Kemudian pada proses kedua merupakan proses klasifikasi menggunakan *Base Classifier* (dengan *Bagging*) dengan dataset yang sama terjadi peningkatan hasil evaluasi yaitu dengan rata - rata hasil akurasi (92,68%) dan *g-mean* (92,31%). Namun pada beberapa dataset yang kurang berisik metode *bagging* mengalami penurunan kinerja sehingga dilakukan proses ketiga. Pada proses ketiga ini dilakukan klasifikasi menggunakan *Base Classifier* (dengan *Balanced-Bagging*) dimana terjadi peningkatan yang hasil evaluasi dengan dataset yang sama yaitu rata - rata hasil akurasi dan *g-mean* meningkat menjadi (95,90%), peningkatan hasil evaluasi ini terjadi pada setiap dataset.

Dengan peningkatan hasil evaluasi baik nilai akurasi maupun *g-mean*, menunjukkan bahwa kemampuan klasifikasi juga meningkat dan lebih peka terhadap dataset minoritas sehingga hasil penelitian ini dapat dijadikan referensi terhadap penanganan kasus *imbalanced dataset*.

Kata Kunci: *Bagging, Balanced-Bagging, Ketidakseimbangan Kelas, Klasifikasi.*

ABSTRACT

The classification algorithm is an algorithm that is very often used in conjunction with human needs, but researchers often encounter problems when using this classification algorithm. One of the problems that researchers often encounter is the case of imbalanced dataset. So that in this study the researcher suggests the ensemble method to overcome it. One of the well-known ensemble method algorithms is bagging. And here the researcher will also suggest a balanced-bagging algorithm to improve the ability of bagging.

In this research, it involves three different classification processes with five datasets that have different imbalanced ratio (IR) which are evaluated based on the results of balanced accuracy and geometric mean. The first process is a classification process on the Base Classifier (without Bagging) which achieves its best performance on the pen_digits dataset with an average accuracy of results (92.66%) and g-mean (91.85%). Then in the second is the classification process using the Base Classifier (with Bagging) with a dataset that has an increase in evaluation results, namely the average process result (92.68%) and the g-mean (92.31%). However, in some datasets that were less noisy, the bagging method experienced a decrease in performance so a third process was carried out. In this third process, classification is carried out using a Base Classifier (with Balanced-Bagging) where there is an increase in the evaluation results with the same dataset, namely the average value of the value and the g-mean increases to (95.90%), an increase in the results of this evaluation occurs in each data set.

With an increase in the results of the evaluation both the accuracy and g-mean values, it shows that the classification ability also increases and is more sensitive to minority datasets so that the results of this study can be used as a reference for handling cases of imbalanced dataset.

Keywords: *Bagging, Balanced-Bagging, Class Imbalance, Decision Tree, Classification.*