

**IMPLEMENTASI ALGORITMA SYNTHETIC MINORITY OVER-
SAMPLING TECHNIQUE (SMOTE) UNTUK MENANGANI
KETIDAKSEIMBANGAN KELAS PADA
DATASET KLASIFIKASI**

SKRIPSI



disusun oleh :

Gagah Gumelar

16.11.0799

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

**IMPLEMENTASI ALGORITMA SYNTHETIC MINORITY OVER-
SAMPLING TECHNIQUE (SMOTE) UNTUK MENANGANI
KETIDAKSEIMBANGAN KELAS PADA
DATASET KLASIFIKASI**

SKRIPSI

**untuk memenuhi sebagai persyaratan
mencapai gelar Sarjana
pada Program Studi Informatika**



disusun oleh

Gagah Gumelar

16.11.0799

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

PERSETUJUAN

SKRIPSI

**IMPLEMENTASI ALGORITMA SYNTHETIC MINORITY OVER-
SAMPLING TECHNIQUE (SMOTE) UNTUK MENANGANI
KETIDAKSEIMBANGAN KELAS PADA
DATASET KLASIFIKASI**

yang dipersiapkan dan disusun oleh

Gagah Gumelar

16.11.0799

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 20 Maret 2020

Dosen Pembimbing,

Mulia Sulistiyono, M.Kom

NIK. 190302248

PENGESAHAN
SKRIPSI
IMPLEMENTASI ALGORITMA SYNTHETIC MINORITY OVER-
SAMPLING TECHNIQUE (SMOTE) UNTUK MENANGANI
KETIDAKSEIMBANGAN KELAS PADA
DATASET KLASIFIKASI

yang dipersiapkan dan disusun oleh

Gagah Gumelar

16.11.0799

telah dipertahankan di depan Dewan Penguji
pada tanggal 20 Maret 2020

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Agung Nugroho, M.Kom
NIK. 190302242

Haryoko, M.Kom
NIK. 190302268

Mulia Sulistiyono, M.Kom
NIK. 190302248

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 20 Maret 2020

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, 20 Maret 2020

Gagah Gumelar
NIM. 16.11.0799

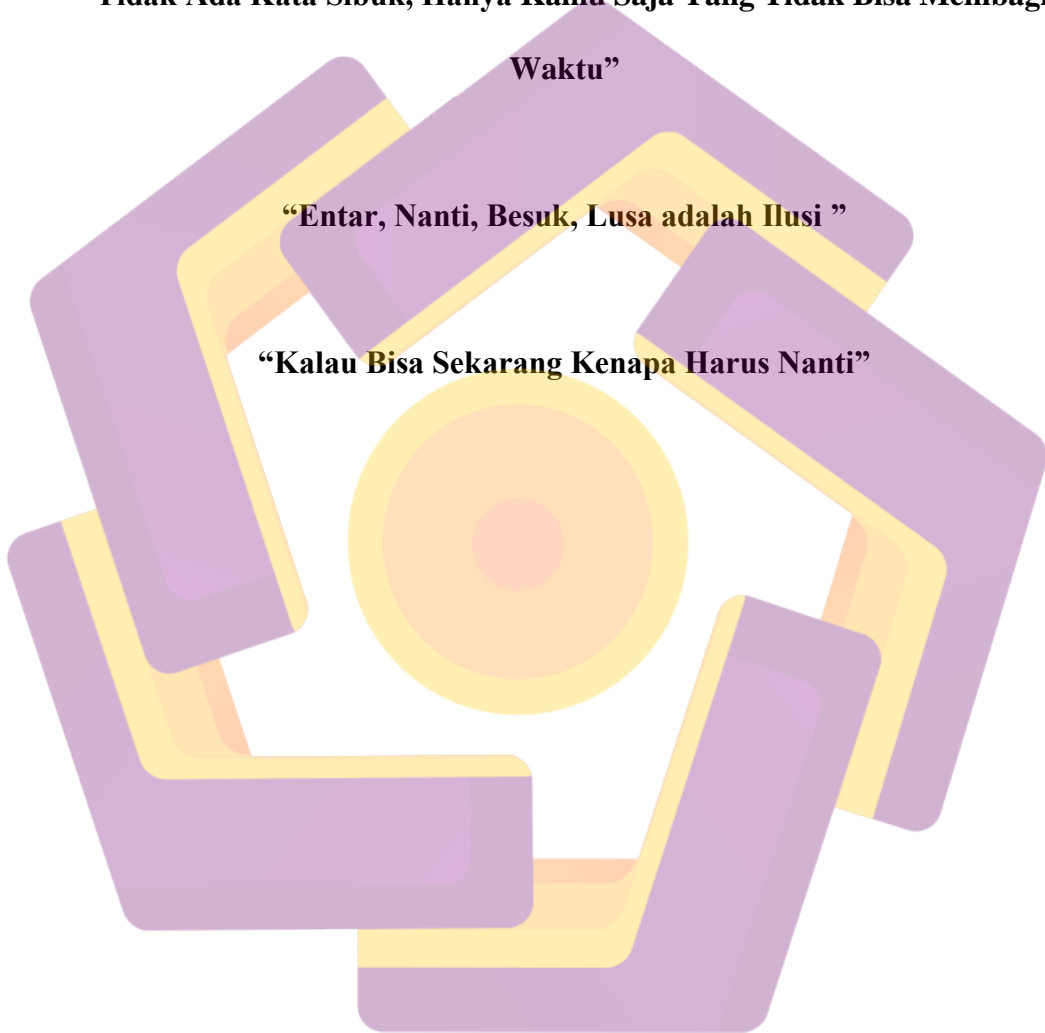
MOTTO

“Slow Progress Is Better Than No Progress”

“Tidak Ada Kata Sibuk, Hanya Kamu Saja Yang Tidak Bisa Membagi Waktu”

“Entar, Nanti, Besuk, Lusa adalah Ilusi ”

“Kalau Bisa Sekarang Kenapa Harus Nanti”



PERSEMBAHAN

Alhamdulillahirabil'alamin, puji syukur kepada Allah SWT, atas rahmat dan karunia-Nya sehingga saya bisa menjadi pribadi yang beriman, berfikir, bersabar dan memberikan kelancaran serta kemudahan dalam menyelesaikan tugas akhir ini, tentunya juga tidak lepas dari dukungan dari orang-orang yang ada disekeliling saya yang selalu memberi semangat dan doa. Untuk itu saya mengucapkan terimakasih kepada:

1. Orang tuaku Bapak Alm. Musiman dan Ibu Yatimah, Terima kasih yang tidak henti-hentinya Penulis ucapkan atas kasih sayangnya. Terima kasih sudah mendidik dan merawatku hingga detik ini. Kesuksesan anakmu adalah berkat doa yg selalu kalian panjatkan. Semoga Allah SWT selalu memberikan keberkahan, kesehatan, umur yang panjang dan keberkahan dunia maupun akhirat.
2. Seluruh keluarga dan semua pihak yang mungkin tidak dapat disebutkan satu persatu, terimakasih atas dukungan dan doanya.
3. Sela Oktavia Sari yang selalu menyemangati, dalam setiap keadaan baik suka maupun duka. Terimakasih atas dukungan selama pengerjaan skripsi ini. Semoga kebaikan kalian dibalas oleh Allah SWT dan menjadi orang sukses.
4. Teman-teman Kelas S1 IF-13 angkatan 2016, untuk semuanya yang tidak bisa disebutkan satu-satu disini tanpa terkecuali, Terimakasih atas bantuannya, dukungan dan doanya. Semoga kalian sukses dan apa yang kalian cita-citakan tercapai.

KATA PENGANTAR

Bismillahirrahmanirahim, Alhamdulillahirabil'alamin puji syukur penulis panjatkan kepada Allah SWT atas Rahmat dan karunia-Nya , sehingga penulis dapat menyelesaikan skripsi dengan judul “Implementasi Algoritma Synthetic Minority Over-Sampling Technique (SMOTE) untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi”. Untuk memenuhi syarat akademis dalam menyelesaikan Program Studi Strata Satu (S1) pada Fakultas Ilmu Komputer Universitas Amikom Yogyakarta. Mengingat keterbatasan pengetahuan dan pengalaman dalam penulisan skripsi ini, Penulis banyak mendapatkan bimbingan, petunjuk, saran dan arahan dari berbagai pihak, oleh karena itu dengan kerendahan hati dan rasa hormat Penulis mengucapkan terima kasih sebesar-besarnya kepada :

1. Bapak Prof. Dr. M. Suyanto, MM. Selaku Rektor Universitas AMIKOM Yogyakarta..
2. Bapak Mulia Sulistiyono, M.Kom. Selaku Dosen pembimbing yang telah membantu dalam penulisan skripsi ini.
3. Ibu/bapak Selaku Dosen Penguji yang telah memberikan petunjuk, serta nasehat dalam ujian skripsi ini.

Penulis mendoakan untuk semua pihak yang telah membantu dalam penulisan skripsi ini semoga diberikan balasan dan berkah dari Allah SWT. Penulis menyadari masih banyak kekurangan dalam penulisan skripsi ini untuk itu saran, kritik dan perbaikan yang bersifat membangun sangat diharapkan. Akhir kata Penulis berharap semoga skripsi ini bermanfaat untuk semua pihak yang membutuhkan.

Yogyakarta, 20 Maret 2020
Penulis,

Gagah Gumelar

DAFTAR ISI

PERSETUJUAN	ii
PERNYATAAN	iv
MOTTO.....	v
PERSEMBAHAN	vi
KATA PENGANTAR.....	vii
DAFTAR ISI.....	viii
DAFTAR TABEL	xi
DAFTAR GAMBAR.....	xviii
DAFTAR LAMPIRAN	xix
DAFTAR ISTILAH	xx
INTISARI	xxi
ABSTRACT	xxii
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Batasan Masalah	3
1.4 Maksud dan Tujuan Penelitian	4
1.5 Manfaat Penelitian.....	5
1.6 Metode Penelitian.....	5
1.6.1 Metode mengatasi ketidakseimbangan kelas	5
1.6.2 Metode Klasifikasi	6
1.6.3 Metode Evaluasi	6
1.7 Sistematika Penulisan.....	6
BAB 2 LANDASAN TEORI	8
2.1 Kajian Pustaka	8
2.2 Landasan Teori	14
2.2.1 Imbalance Class	14
2.2.2 Resampling.....	14
2.2.3 Synthetic Minority Over-sampling Technique (SMOTE)	15
2.2.4 Data Splitting	16

2.2.5	C 4.5 atau Decision Tree	17
2.2.6	Naïve Bayes	20
2.2.7	KNN	20
2.2.8	SVM	21
2.2.9	Confusion Matrix	22
2.2.10	Python	24
BAB 3	METODE PENELITIAN.....	27
3.1	Gambaran Umum	27
3.2	Instrumen Penelitian	27
3.2.1	Kebutuhan Perangkat Keras	27
3.2.2	Kebutuhan Perangkat Lunak.....	28
3.3	Alur Penelitian	29
3.4	Analisa dan Perancangan Sistem	30
3.4.1	Analisa Sistem	30
3.4.2	Analisa Data.....	30
3.4.3	Proses SMOTE	33
3.5	Pengujian Akurasi	37
BAB 4	HASIL DAN PEMBAHASAN	40
4.1	Implementasi Algoritma Oversampling terhadap Dataset	40
4.1.1	Implementasi Algoritma Oversampling terhadap Dataset ecoli IR 3,3	41
4.1.2	Implementasi Algoritma Oversampling terhadap Dataset ecoli IR 5,4	43
4.1.3	Implementasi Algoritma Oversampling terhadap Dataset ecoli IR 8,6	45
4.1.4	Implementasi Algoritma Oversampling terhadap Dataset ecoli IR 15,8	47
4.2	Implementasi Algoritma Klasifikasi untuk Pengujian	49
4.2.1	Implementasi Algoritma Klasifikasi pada Dataset Ecoli dengan IR 3,3	49
4.2.2	Implemenntasi Algoritma Klasifikasi Pada Dataset Ecoli Dengan IR 5,4	74
4.2.3	Implemenntasi Algoritma Klasifikasi Pada Dataset Ecoli Dengan IR 8,6	90
4.2.4	Implemenntasi Algoritma Klasifikasi Pada Dataset Ecoli Dengan IR 15,8106	

BAB 5	KESIMPULAN DAN SARAN	124
5.1	Kesimpulan.....	124
5.2	Saran	127
DAFTAR PUSTAKA		128
LAMPIRAN.....		130



DAFTAR TABEL

Tabel 2.1 Matrik Literatur Review dan Posisi Penelitian	10
Tabel 3.1 Contoh dataset sebelum dilakukan proses SMOTE	34
Tabel 3.2 Contoh komposisi dataset sebelum dan setelah dilakukan proses SMOTE	36
Tabel 3.3 Contoh dataset setelah dilakukan proses SMOTE	37
Tabel 4.1 Confusion matrix klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3	49
Tabel 4.2 Hasil klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3	49
Tabel 4.3 Confusion matrix klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3 diulang sebanyak 50 kali	51
Tabel 4.4 Hasil klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3	51
Tabel 4.5 Confusion matrix klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3	52
Tabel 4.6 Hasil klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3	52
Tabel 4.7 Confusion matrix klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3 diulang sebanyak 50 kali	54
Tabel 4.8 Hasil klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3	54
Tabel 4.9 Confusion matrix klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3	55
Tabel 4.10 Hasil klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3	55
Tabel 4.11 Confusion matrix klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3 diulang sebanyak 50 kali	57
Tabel 4.12 Hasil klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3	57
Tabel 4.13 Confusion matrix klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3	58
Tabel 4.14 Hasil klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3	58
Tabel 4.15 Confusion matrix klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3 diulang sebanyak 50 kali	60
Tabel 4.16 Hasil klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3	60
Tabel 4.17 Confusion matrix klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3 k = 5	61
Tabel 4.18 Hasil klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3	61
Tabel 4.19 Confusion matrix klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3 k = 5	63

Tabel 4.20 Hasil klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3 diulang sampai dengan 50 kali.....	63
Tabel 4.21 Confusion matrix klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3 k = 5	64
Tabel 4.22 Hasil klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3	64
Tabel 4.23 Confusion matrix klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3 k = 5	66
Tabel 4.24 Hasil klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3	66
Tabel 4.25 Confusion matrix klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3.....	67
Tabel 4.26 Hasil klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3.....	67
Tabel 4.27 Confusion matrix klasifikasi Decision Tree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3 diulang sebanyak 50 kali	69
Tabel 4.28 Hasil klasifikasi Decision Tree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3.....	69
Tabel 4.29 Confusion matrix klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3.....	70
Tabel 4.30 Hasil klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3.....	71
Tabel 4.31 Confusion matrix klasifikasi DecisionTree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3	72
Tabel 4.32 Hasil klasifikasi DecisionTree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 3,3 diulang sebanyak 50 kali	72
Tabel 4.33 Confusion matrix klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	74
Tabel 4.34 Hasil klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4.....	74
Tabel 4.35 Confusion matrix klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	75
Tabel 4.36 Hasil klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4 diulang sebanyak 50 kali	75
Tabel 4.37 Confusion matrix klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	76
Tabel 4.38 Hasil klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4.....	76
Tabel 4.39 Confusion matrix klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	77
Tabel 4.40 Hasil klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4.....	77
Tabel 4.41 Confusion matrix klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	78

Tabel 4.42 Hasil klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	78
Tabel 4.43 Confusion matrix klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	79
Tabel 4.44 Hasil klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	79
Tabel 4.45 Confusion matrix klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	80
Tabel 4.46 Hasil klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	80
Tabel 4.47 Confusion matrix klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4 diulang sebanyak 50 kali	81
Tabel 4.48 Hasil klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	81
Tabel 4.49 Confusion matrix klasifikasi K-NN k = 7 dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	82
Tabel 4.50 Hasil klasifikasi K-NN k = 7 dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	82
Tabel 4.51 Confusion matrix klasifikasi K-NN k = 7 dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	83
Tabel 4.52 Hasil klasifikasi K-NN k = 7 dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	83
Tabel 4.53 Confusion matrix klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4 k = 3	84
Tabel 4.54 Hasil klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4 k = 3	84
Tabel 4.55 Confusion matrix klasifikasi K-NN k = 3 dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	85
Tabel 4.56 Hasil klasifikasi K-NN k = 3 dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	85
Tabel 4.57 Confusion matrix klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	86
Tabel 4.58 Hasil klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	86
Tabel 4.59 Confusion matrix klasifikasi Decision Tree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	87
Tabel 4.60 Hasil klasifikasi Decision Tree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	87
Tabel 4.61 Confusion matrix klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	88
Tabel 4.62 Hasil klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 5,4	88
Tabel 4.63 Confusion matrix klasifikasi DecisionTree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4	89

Tabel 4.64 Hasil klasifikasi DecisionTree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 5,4.....	89
Tabel 4.65 Confusion matrix klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6	90
Tabel 4.66 Hasil klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6.....	90
Tabel 4.67 Confusion matrix klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6	91
Tabel 4.68 Hasil klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6.....	91
Tabel 4.69 Confusion matrix klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6	92
Tabel 4.70 Hasil klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6.....	92
Tabel 4.71 Confusion matrix klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6	93
Tabel 4.72 Hasil klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6.....	93
Tabel 4.73 Confusion matrix klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6	94
Tabel 4.74 Hasil klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6	94
Tabel 4.75 Confusion matrix klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6	95
Tabel 4.76 Hasil klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6	95
Tabel 4.77 Confusion matrix klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6	96
Tabel 4.78 Hasil klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6	96
Tabel 4.79 Confusion matrix klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6	97
Tabel 4.80 Hasil klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6	97
Tabel 4.81 Confusion matrix klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6 k = 5	98
Tabel 4.82 Hasil klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6	98
Tabel 4.83 Confusion matrix klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6 k = 5	99
Tabel 4.84 Hasil klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6	99
Tabel 4.85 Confusion matrix klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6 k = 5	100

Tabel 4.86 Hasil klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6	100
Tabel 4.87 Confusion matrix klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6 k = 5	101
Tabel 4.88 Hasil klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6	101
Tabel 4.89 Confusion matrix klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6.....	102
Tabel 4.90 Hasil klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6.....	102
Tabel 4.91 Confusion matrix klasifikasi Decision Tree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6	103
Tabel 4.92 Hasil klasifikasi Decision Tree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6.....	103
Tabel 4.93 Confusion matrix klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6.....	104
Tabel 4.94 Hasil klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 8,6.....	104
Tabel 4.95 Confusion matrix klasifikasi DecisionTree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6	105
Tabel 4.96 Hasil klasifikasi DecisionTree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 8,6.....	105
Tabel 4.97 Confusion matrix klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	106
Tabel 4.98 Hasil klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	106
Tabel 4.99 Confusion matrix klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	107
Tabel 4.100 Hasil klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	107
Tabel 4.101 Confusion matrix klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	108
Tabel 4.102 Hasil klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	108
Tabel 4.103 Confusion matrix klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	109
Tabel 4.104 Hasil klasifikasi Naïve Bayes dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	109
Tabel 4.105 Confusion matrix klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	110
Tabel 4.106 Hasil klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8	110
Tabel 4.107 Confusion matrix klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8	111

Tabel 4.108 Hasil klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	111
Tabel 4.109 Confusion matrix klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	112
Tabel 4.110 Hasil klasifikasi SVM dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	112
Tabel 4.111 Confusion matrix klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	113
Tabel 4.112 Hasil klasifikasi SVM dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	113
Tabel 4.113 Confusion matrix klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8 k = 7.....	114
Tabel 4.114 Hasil klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	114
Tabel 4.115 Confusion matrix klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	115
Tabel 4.116 Hasil klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	115
Tabel 4.117 Confusion matrix klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8 k = 3.....	116
Tabel 4.118 Hasil klasifikasi K-NN dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	116
Tabel 4.119 Confusion matrix klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8 k = 3.....	117
Tabel 4.120 Hasil klasifikasi K-NN dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	117
Tabel 4.121 Confusion matrix klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	118
Tabel 4.122 Hasil klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	118
Tabel 4.123 Confusion matrix klasifikasi Decision Tree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	119
Tabel 4.124 Hasil klasifikasi Decision Tree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	119
Tabel 4.125 Confusion matrix klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	120
Tabel 4.126 Hasil klasifikasi Decision Tree dengan Cross-validation fold 10 pada Dataset ecoli IR 15,8.....	120
Tabel 4.127 Confusion matrix klasifikasi DecisionTree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	121
Tabel 4.128 Hasil klasifikasi DecisionTree dengan 80% data training dan 20% data testing pada Dataset ecoli IR 15,8.....	121
Tabel 4.129 Rangkuman Hasil Akurasi Algoritma Klasifikasi.....	122
Tabel 4.130 Rangkuman Hasil G-Mean Algoritma Klasifikasi.....	123

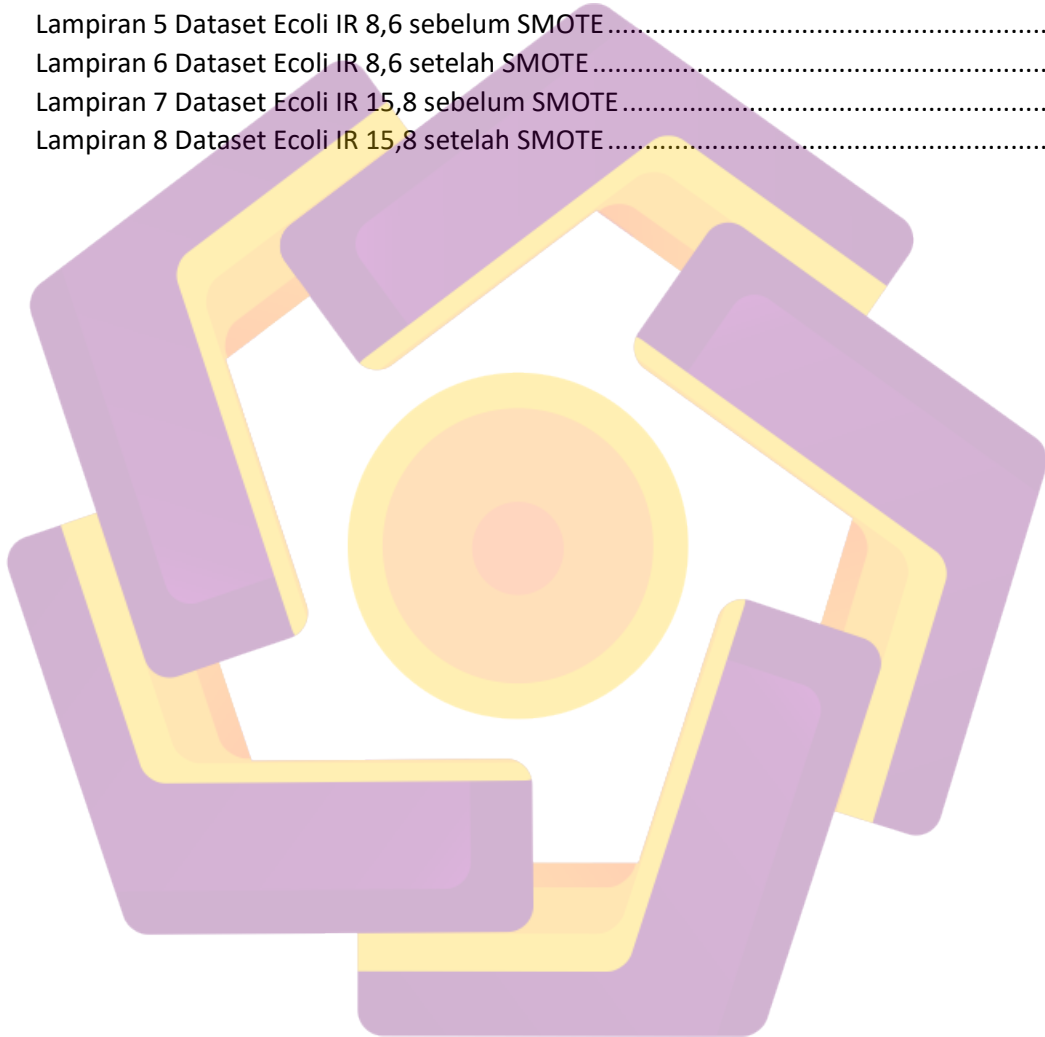


DAFTAR GAMBAR

Gambar 2.1 Pembagian data.....	16
Gambar 2.2 Pembagian data dengan 10 fold cross validation	17
Gambar 2.3 Confusion matrix	23
Gambar 2.4 python programming language	24
Gambar 2.5 programming languages for data science.....	25
Gambar 3.1 Alur penelitian	29
Gambar 3.2 Distribusi kelas pada Dataset ecoli IR 3,3 sebelum dilakukan SMOTE	31
Gambar 3.3 Distribusi kelas pada Dataset ecoli IR 5,4 sebelum dilakukan SMOTE	31
Gambar 3.4 Distribusi kelas pada Dataset ecoli IR 8,6 sebelum dilakukan SMOTE	32
Gambar 3.5 Distribusi kelas pada Dataset ecoli IR 15,8 sebelum dilakukan SMOTE	32
Gambar 3.6 proses uji akurasi data dengan klasifikasi algoritma C45.....	38
Gambar 4.1 Distribusi kelas pada dataset ecoli IR 3,3 sebelum dilakukan SMOTE	41
Gambar 4.2 Ilustrasi proses SMOTE pada dataset ecoli IR 3,3	41
Gambar 4.3 Distribusi kelas pada dataset ecoli IR 3,3 setelah dilakukan SMOTE	42
Gambar 4.4 Distribusi kelas pada dataset ecoli IR 5,4 sebelum dilakukan SMOTE	43
Gambar 4.5 Ilustrasi proses SMOTE pada dataset ecoli IR 5,4	43
Gambar 4.6 Distribusi kelas pada dataset ecoli IR 5,4 setelah dilakukan SMOTE	44
Gambar 4.7 Distribusi kelas pada dataset ecoli IR 8,6 sebelum dilakukan SMOTE	45
Gambar 4.8 Ilustrasi proses SMOTE pada dataset ecoli dengan IR 8,6	45
Gambar 4.9 Distribusi kelas dataset ecoli IR 8,6 setelah dilakukan SMOTE	46
Gambar 4.10 Distribusi kelas dataset ecoli IR 15,8 sebelum dilakukan SMOTE	47
Gambar 4.11 Ilustrasi proses SMOTE pada dataset ecoli IR 15,8	47
Gambar 4.12 Distribusi kelas dataset ecoli IR 25,4 setelah dilakukan proses SMOTE.....	48

DAFTAR LAMPIRAN

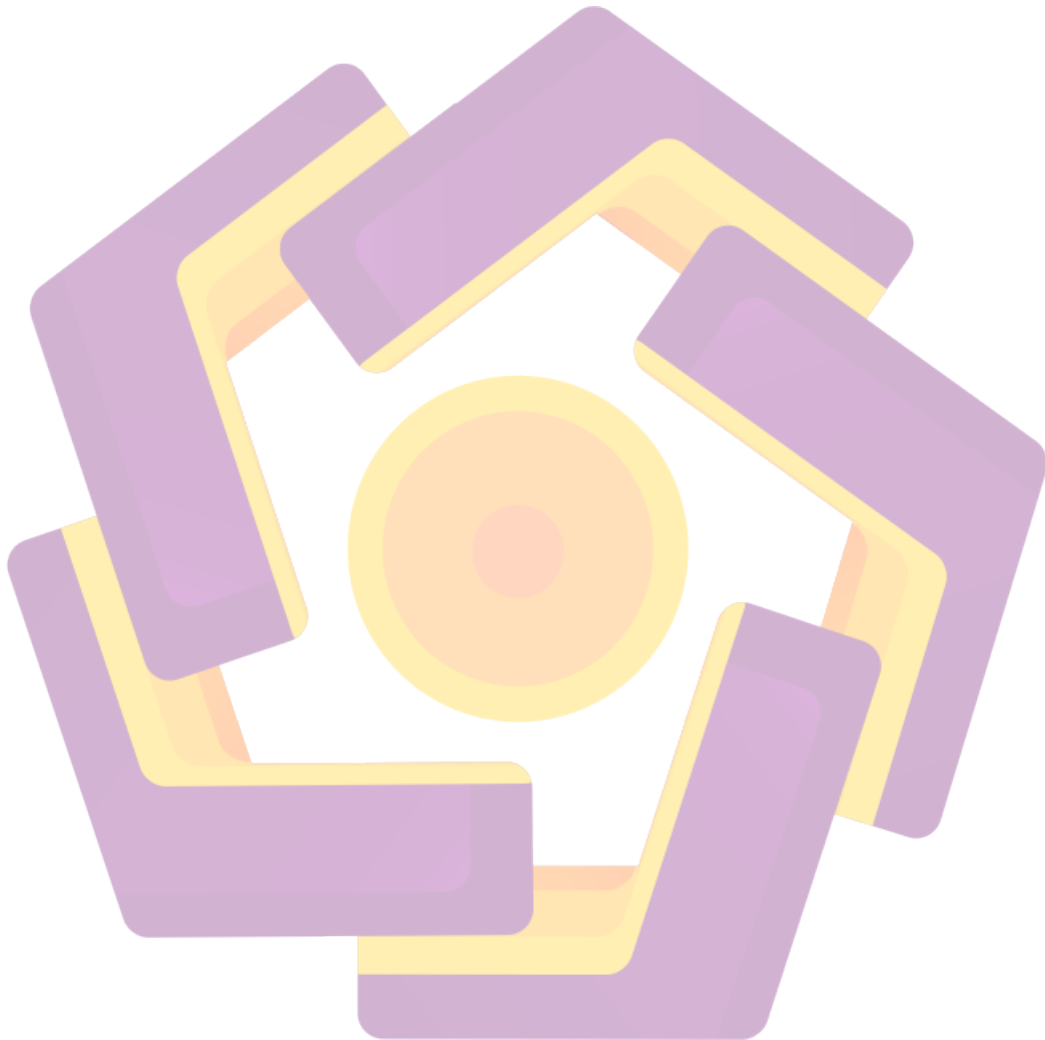
Lampiran 1 Dataset Ecoli IR 3,3 sebelum SMOTE.....	130
Lampiran 2 Dataset Ecoli IR 3,3 setelah SMOTE.....	138
Lampiran 3 Dataset Ecoli IR 5,4 sebelum SMOTE.....	149
Lampiran 4 Dataset Ecoli IR 5,4 setelah SMOTE.....	157
Lampiran 5 Dataset Ecoli IR 8,6 sebelum SMOTE.....	170
Lampiran 6 Dataset Ecoli IR 8,6 setelah SMOTE.....	177
Lampiran 7 Dataset Ecoli IR 15,8 sebelum SMOTE.....	190
Lampiran 8 Dataset Ecoli IR 15,8 setelah SMOTE.....	198



DAFTAR ISTILAH

SMOTE = Synthetic Minority Over-sampling Technique

Imbalance class = kelas tidak seimbang



INTISARI

Sebuah kelas dikatakan tidak seimbang apabila ada suatu kelas yang memiliki data yang lebih banyak dibandingkan dengan kelas lainnya. Kelompok kelas dengan jumlah data yang banyak disebut dengan kelas mayoritas, sedangkan kelompok kelas dengan jumlah yang sedikit disebut dengan kelas minoritas. Perbandingan antara kelas minoritas dengan kelas mayoritas disebut dengan *Imbalance Ratio (IR)*. Semakin besar perbedaan antara kelas minoritas dengan kelas mayoritas maka nilai dari *Imbalance Ratio (IR)* semakin besar.

Ketidakseimbangan dataset pada data mining adalah masalah yang serius. Dataset yang tidak seimbang menyebabkan misleading atau kesesatan dalam Hasil klasifikasi dimana data kelas minoritas sering diklasifikasikan sebagai kelas mayoritas. Penerapan algoritma klasifikasi tanpa memperhatikan keseimbangan kelas mengakibatkan prediksi yang baik bagi kelas mayoritas dan kelas minoritas diabaikan. Oleh karena itu pada penelitian ini diimplementasikan algoritma *Syntethic Minority Over-Sampling Technique (SMOTE)* untuk menyeimbangkan dataset. Penelitian ini menggunakan 4 dataset dengan *Imbalance Ratio* yang berbeda dan menggunakan algoritma klasifikasi C45, Naïve Bayes, K-NN, dan SVM. Kemudian dibandingkan antara sebelum dilakukan SMOTE dan setelah dilakukan SMOTE.

Dari hasil penelitian yang sudah dilakukan nilai akurasi dan nilai G-mean algoritma Naïve Bayes konsisten dengan performanya pada setiap level imbalance ratio, sebelum implementasi SMOTE memiliki performa yang tidak baik, sedangkan setelah diimplementasikan SMOTE algoritma Naïve Bayes memiliki peningkatan akurasi yang konsisten. Sehingga dapat ditarik kesimpulan bahwa kombinasi SMOTE + Naïve Bayes paling efektif digunakan pada dataset imbalance dengan level yang berbeda-beda pada skema 10 fold cross validation maupun 80% data testing yang diujikan sebanyak 50 kali.

Kata kunci : klasifikasi dengan dataset tidak seimbang, *Syntethic Minority Over-Sampling Technique (SMOTE)*

ABSTRACT

A class to be imbalance when there is a class that has more data than other classes. The group of classes with a lot of data is called the majority class, while the class groups with fewer numbers are called minority classes. A comparison between minority classes and the majority class is called Imbalance Ratio (IR). The greater the difference between the minority class and the majority class the value of the Imbalance Ratio (IR) is getting larger.

Dataset imbalance in data mining is a serious problem. The imbalanced Dataset causes misleading or error in the classification results where minority class data is often classified as a majority class. The application of the classification algorithm regardless of class balance resulted in a good prediction for the majority class and a neglected minority class. Therefore in this research implemented Syntethic Minority Over-Sampling Technique (SMOTE) algorithm to balance the dataset. The study used 4 datasets with different Imbalance Ratio and used classification algorithms, C45, Naïve Bayes, K-NN, and SVM. Then compared between before SMOTE and after doing SMOTE.

From the research results that have been done accuracy value and value G-mean Naïve Bayes algorithm is consistent with its performance at each level of imbalance ratio, before the implementation of SMOTE has no good performance, whereas after the implemented SMOTE algorithm Naïve Bayes has a consistent increase in accuracy. So it can be concluded that the combination SMOTE + Naïve Bayes most effectively used in the imbalance dataset with different levels in the scheme of 10 fold cross validation and 80% data testing tested as much as 50 times.

Keywords: classification with unbalanced datasets, Syntethic Minority Over-Sampling Technique (SMOTE)