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

SKRIPSI

disusun oleh

Lucky Adhikrisna 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,

Muhlis Sulistiyono, 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 Kurniawan, M.Kom. **NIK. 190302244**	
Andriyan Dwi Putra, M.Kom. **NIK. 190302270**	
Mulia Sulistiyono, M.Kom. **NIK. 190302248**	

Skripsi 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 Adhikrisna 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 kehadirat 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 satu-persatu.

KATA PENGANTAR

Segala puji dan syukur kehadirat 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 Univeristas 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

DAFTAR GAMBAR

DAFTAR TABEL

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 rasio* (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 pemprosesan, 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.*