PERBANDINGAN METODE *OVERSAMPLING* SMOTE DAN G-SMOTE DENGAN KLASIFIKASI ADABOOST PADA PENYAKIT *LIVER*

SKRIPSI

Untuk Memenuhi Persyaratan Melakukan Penelitian

Dalam Rangka Penyusunan Skripsi Strata-1 Ilmu Komputer

Oleh

MUNAWWARAH

NIM. 1811016320004

PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU

SEPTEMBER

2023

PERBANDINGAN METODE *OVERSAMPLING* SMOTE DAN G-SMOTE DENGAN KLASIFIKASI ADABOOST PADA PENYAKIT *LIVER*

SKRIPSI

Untuk Memenuhi Persyaratan Melakukan Penelitian

Dalam Rangka Penyusunan Skripsi Strata-1 Ilmu Komputer

Oleh

MUNAWWARAH

NIM. 1811016320004

PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU

SEPTEMBER 2023

SKRIPSI

SKRIPSI

PERBANDINGAN METODE OVERSAMPLING SMOTE DAN G-SMOTE DENGAN KLASIFIKASI ADABOOST PADA PENYAKIT LIVER

Oleh:

MUNAWWARAH

NIM. 1811016320004

telah dipertahankan di depan Dosen Penguji pada tanggal 20 September 2023

Susunan Dosen Penguji:

Pembimbing Utama,	Ketua Penguji,
Triando H. Saragih, S.Kom., M.Kom.	Rudy Herteno, S.Kom., M.Kom.
NIP. 199308242019031012	NIP. 198809252022031003
Pembimbing Pendamping,	Anggota Penguji,
Dwi Kartini, S.Kom., M.Kom.	Muliadi, S.Kom., M.Cs.
NIP. 198704212012122003	NIP. 197804222010121002

Banjarbaru, 10 Oktober 2023

Ketua Program Studi Ilmu Komputer

Irwan Budiman, S.T., M.Kom.

NIP. 197703252008121001

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang sepengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Pustaka.

Banjarbaru, 10 Oktober 2023

Munawwarah

NIM. 1811016320004

ABSTRAK

PERBANDINGAN METODE OVERSAMPLING *SMOTE* DAN *G-SMOTE* DENGAN KLASIFIKASI *ADABOOST* PADA PENYAKIT LIVER (Oleh: Munawwarah; Pembimbing: Triando Hamonangan Saragih S.Kom., M.Kom dan Dwi Kartini, S.Kom., M.kom.; 2023; 58 halaman)

Liver merupakan salah satu organ tubuh manusia yang paling penting. Jika organ *liver* mengalami kerusakan maka kesehatan pun akan terganggu. Saat ini penyakit *liver* menjadi salah satu dari penyebab utama kematian didunia yang masih meningkat dari tahun ke tahun. Berdasarkan data *UK Health Security Agency* pada tahun 2014 jumlah orang yang meninggal akibat penyakit liver di Inggris meningkat 40% menjadi 10.127. Metode klasifikasi pada penelitian ini menggunakan metode *Adaptive Boosting (AdaBoost)* dan Teknik *oversampling SMOTE* dan *G-SMOTE*. Dataset akan dilakukan preprocessing *label encoder* dan normalisasi min-max, selanjutnya membagi data set menggunakan metode *Splitting Data* dengan rasio 80:20 terbagi menjadi 466 *data traning* dan 117 *data testing*, kemudian dilakukan klasifikasi dari penelitian ini diperolah hasil terbaik dari masin-masing metode baik itu yang menggunakan *oversampling* maupun tanpa *oversampling*, untuk pengklasifikasian *AdaBoost* didapatkan akurasi 78,63%, presisi 73,33%, *recall* 61,52% pada nilai learning rate 0.2 dan AUC terbaik sebesar 79,95% pada nilai learning rate 0,1, pada *AdaBoost+SMOTE* dengan parameter n estimator 350 didapatkan akurasi 76,07%, presisi 79,87%, recall 73,69% dengan learning rate 0.5 dan hasil terbaik AUC dengan learning rate 0,1 yaitu 82,42%, dan terakhir pada *AdaBoost+G-SMOTE* didapatkan hasil terbaik akurasi 79,49%, presisi 72,62%, recall 67,87%, dan AUC 78,95% pada parameter n estimator 100 dan learning rate 0,4. Dari ketiga metode yang telah dilakukan bahwa dengan penambahan Teknik penyeimbang G-SMOTE mampu meningkatkan nilai akurasi, dibanding tanpa menggunakan Teknik penyeimbang serta pada Teknik penyeimbang SMOTE diperoleh nilai presisi, recall, dan AUC yang lebih baik di banding tanpa Teknik penyeimbang. Walaupun demikian metode AdaBoost mampu mengklasifikasi dengan baik walupun tanpa penambahan *oversampling* hal ini bisa dilihat dengan membandingkan akurasi dari *AdaBoost* dan *AdaBoost+SMOTE* tanpa adanya penambahan *oversampling* bahkan *AdaBoos*t memiliki akurasi yang lebih baik hal ini membuktikan bahwa metode *AdaBoost* mampu bekerja dengan maksimal pada dataset yang attributenya sedikit.

Kata kunci: Penyakit Liver, AdaBoost, Oversampling, SMOTE, G-SMOTE

ABSTRACT

COMPARISON OF SMOTE AND G-SMOTE OVERSAMPLING METHODS WITH ADABOOST CLASSIFICATION ON LIVER DISEASE (By: Munawwarah; Advisor: Triando Hamonangan Saragih S.Kom., M.Kom and Dwi Kartini, S.Kom., M.kom.; 2023; 58 pages)

The liver is one of the most important organs of the human body. If the liver is damaged, health will be disrupted. Currently liver disease is one of the leading causes of death in the world which is still increasing from year to year. Based on data from the UK Health Security Agency in 2014 the number of people who died from liver disease in the UK increased by 40% to 10,127. The classification method in this study uses the Adaptive Boosting (AdaBoost) method and the SMOTE and G-SMOTE oversampling techniques. The dataset will be preprocessed label encoder and min-max normalization, then divide the data set using the Splitting Data method with a ratio of 80:20 divided into 466 traning data and 117 testing data, then the classification of this study obtained the best results from each method both those using oversampling and without oversampling, for the AdaBoost classification obtained an accuracy of 78.63%, precision 73.33%, recall 61.52% at a learning rate value of 0. 2 and the best AUC of 79.95% at a learning rate value of 0.1, in AdaBoost + SMOTE with parameter n estimator 350 obtained accuracy of 76.07%, precision 69.87%, recall 73.69% with a learning rate of 0.5 and the best AUC results with a learning rate of 0.1 which is 82.42%, and finally in AdaBoost + G-SMOTE obtained the best results of accuracy 79.49%, precision 72.62%, recall 67.87%, and AUC 78.95% on parameter n estimator 100 and learning rate 0.4. From the three methods that have been carried out, the addition of the G-SMOTE balancing technique is able to increase the accuracy value, compared to without using the balancing technique and the SMOTE balancing technique obtained better precision, recall, and AUC values than without the balancing technique. However, the AdaBoost method is able to classify well even without the addition of oversampling, this can be seen by comparing the accuracy of AdaBoost and AdaBoost + SMOTE without the addition of oversampling, even AdaBoost has better accuracy, this proves that the AdaBoost method is able to work optimally on datasets with few attributes.

Keywords: Liver Disease, AdaBoost, Oversampling, SMOTE, G-SMOTE

PRAKATA

Assalamualaikum Warahmatullahi Wabarakatuh

Puji syukur penulis panjatkan ke hadirat Allah SWT karena atas berkat rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul "Perbandingan Metode *Oversampling* SMOTE dan G-SMOTE dengan Klasifikasi AdaBoost pada Penyakit *Liver*" untuk memenuhi syarat dalam menyelesaikan Pendidikan S1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat. Tak lupa penulis panjatkan sholawat dan salam ke hadirat Rasulullah Muhammad SAW beserta para sahabat, keluarga, dan pengikut beliau hingga *yaumul qiama.*

Pada lembar ini penulis ingin menyampaikan ucapan terimakasih kepada pihak-pihak yang sangat mendukung penulis dalam pembuatan dan penyusunan skripsi ini, adapun yang dimaksud adalah sebagai berikut:

1. Keluarga terutama kedua orang tua yaitu ibu Sri Rahmawati dan ayah H. Sa'dudin yang selalu memberikan bantuan, semangat, doa dan dukungan dalam proses menyelesaikan skripsi ini.
2. Bapak Triando Hamonangan Saragih S.Kom., M.Kom selaku dosen pembimbing utama yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
3. Ibu Dwi Kartini S.Kom., M.Kom selaku dosen pembimbing pendamping yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
4. Bapak Rudy Herteno S.Kom., M.Kom., dan Bapak Muliadi S.Kom., M. Cs. selaku dosen penguji saya yang bersedia meluangkan waktu serta saran agar terselesaikannya skripsi ini
5. Bapak Irwan Budiman, S.T., M.Kom selaku Koordinator Program Studi Komputer FMIPA ULM, atas bantuan dan izin beliau skripsi ini dapat diselesaikan.

6. Bapak Mohammad Reza Faisal, S.Si., S.T., M.T., P.hD selaku dosen pembimbing akademik yang banyak memberikan masukkan dan bimbingan selama saya berkuliah, serta seluruh dosen dan staf Program Studi Ilmu Komputer FMIPA ULM atas ilmu dan bantuan yang diberikan selama ini yang sangat bermanfaat.
7. Adik, sahabat dan teman-teman yang sudah membantu dan memberi semangat kepada saya dalam mengerjakan penelitian ini yaitu Kayla, Nelisa, Selvi, Laila, Fitri, Helma, Leha, Sora dan lain-lain yang tidak bisa saya sebutkan satu persatu.
8. Teman-teman Ilmu Komputer Angkatan 2018 terimakasih atas canda, tawa, dan perjuangan yang sudah dilewai bersama.
9. Semua pihak yang tidak dapat disebutkan satu persatu yang telah turut membantu dalam penyelesaian skripsi ini.
10. Dan terimakasih untuk diri saya sendiri karna sudah berhasil sampai ketitik ini, walau ditemani dengan air mata dan juga tawa tapi tetap tidak berhenti untuk terus menyelesaikan naskah ini.

Semoga tulisan ini dapat bermanfaat bagi ilmu pengetahuan dan pembaca khususnya serta mendapat keridhaan Allah SWT.

Banjarbaru, September 2023

Munawwarah

DAFTAR ISI

DAFTAR TABEL

Tabel **Halaman**

DAFTAR GAMBAR

Gambar **Halaman**

LAMPIRAN

Lampiran **Halaman**

BAB I
PENDAHULUAN

1.1 Latar Belakang

. Saat ini, kasus observasi seperti data penyakit seringkali memiliki ketidakseimbangan data antara kelas. Salah satu contoh kasusnya pada penyakit liver yang datanya diperoleh dari website UCI Repository. Liver merupakan organ penting dalam tubuh manusia, dan penyakit liver menjadi salah satu penyakit dengan kematian yang terus meningkat dari tahun ke tahun. Berdasarkan data *UK Health Security Agency* pada tahun 2014 jumlah orang yang meninggal akibat penyakit liver di Inggris meningkat 40%. Ketidakseimbangan ini dapat memengaruhi performa pengklasifikasian. Salah satu cara untuk mengatasi masalah ini adalah dengan menggunakan teknik oversampling, teknik yang populer saat ini yaitu SMOTE. SMOTE meningkatkan jumlah sampel kelas minoritas agar sejajar dengan kelas mayoritas, tanpa menghilangkan informasi data lain. Namun, SMOTE memiliki kelemahan, seperti overgeneralization yang dapat menyebabkan overlaping data sintetis yang serupa. Oleh karena itu, peneliti mengembangkan teknik lain seperti G-SMOTE, Untuk menghasilkan data *sintetis* pada G-SMOTE data dipilih secara acak beberapa kali, dan nilai tetangga terdekat dari data mayoritas dan minoritas dibandingkan agar data sintetis tidak tumpang tindih, serta menghindari overlapping data (Douzas et al., 2019).

Sebelum melakukan klasifikasi, dataset akan seimbangkan dengan teknik *Oversampling* SMOTE. SMOTE telah berhasil digunakan untuk mengatasi masalah ketidakseimbangan kelas dalam berbagai macam klasifikasi (Wijayanti et al., 2021). Dengan penambahan teknik ini dapat menghindari terjadinya overfitting pada data. Penelitian yang dilakukan oleh (Astuti dan Febri, 2021) membandingkan hasil klasifikasi yang hanya menggunakan *KNN* dan menggunakan *KNN+SMOTE* dataset yang digunakan yaitu *car evolution*. Hasil penelitian menunjukkan hasil akurasi tertinggi

dengan hanya menggunakan *KNN* sebesar 83.56% dan *KNN+SMOTE* sebesar 93.11%, dari hasil tersebut menunjukkan bahwa penggunaan *SMOTE* mampu menaikkan nilai akurasi sebesar 9.97%.

Pada metode oversampling juga ada teknik *G-SMOTE*. G-SMOTE membangkitkan sample data minoritas agar kedua kelas seimbang dengan melakukan pengacakan data yang lebih luas berdasarkan nilai dari tetangganya. Berdasarkan penelitian yang dilakukan oleh Douzas et al., (2019). Penelitian ini menggunakan dataset data lahan dari barat laut Portugal, dengan membandingkan metode *Oversampling* terhadap beberapa metode klasifikasi, maka hasil yang didapatkan bahwa metode G-SMOTE memiliki hasil yang lebih baik dibanding metode penyeimbang lain, yaitu F-Score *g-smote+lr* 31%, g-smote+*knn* 28%, *g-smote+gbc* 32%, *g-smote+rf* 34% dan G-Mean g-smote+lr 56%, g-smote+knn 50%, g-smote+dt 51%, g-smote+gbc 55%, g-smote+rf 57%.

Banyak penelitian mengembangkan algoritma klasifikasi untuk mendeteksi penyakit dengan sistem pendukung keputusan medis cerdas, seperti AdaBoost, sebuah algoritma pohon keputusan yang memadukan hasil dari weak classifier. AdaBoost memiliki keunggulan dalam memperbaiki klasifikasi lemah dengan mengoreksi kesalahan pada data sebelumnya (Saragih et al., 2020) dan mampu menangani dataset yang tidak terlalu kompleks dan tidak seimbang dengan baik. Penelitian yang dilakukan Prianti et al. (2020) yaitu membandingkan KNN dan AdaBoost dalam model, dari hasil ini menunjukkan AdaBoost memiliki kinerja akurasi yang lebih bagus yaitu 84%, dibandingkan dengan KNN yang hanya mencapai 82% dalam memprediksi kinerja perusahaan di Indonesia

Berdasarkan latar belakang yang telah disebutkan, maka akan dilakukan perbandingan klasifikasi AdaBoost dengan menggunakan Teknik *oversampling* dan tanpa Teknik *oversampling* pada dataset *Indian Liver Patient* untuk melihat performa kinerja mana yang terbaik dari akurasi, presisi, *recall*, dan *AUC*.

1.2 Rumusan Masalah

Berdasarkan uraian dari latar belakang, maka dapat dirumuskan permasalahan yang akan dibahas pada penelitian ini yaitu:

1. Berapa kinerja pada algoritma AdaBoost untuk klasifikasi penyakit *liver*?
2. Berapa kinerja algoritma AdaBoost pada klasifikasi penyakit *liver* dengan menggunakan SMOTE ?
3. Berapa kinerja algoritma AdaBoost pada klasifikasi penyakit *liver* dengan menggunakan G-SMOTE?

1.3 Batasan Masalah

Agar penelitian ini dapat berjalan sesuai dengan tujuan yang ingin dicapai dan tetap berjalan pada ruang lingkupnya, maka berikut ini Batasan masalah pada penelitian ini:

1. Dataset yang digunakan yaitu *Indian Liver Patient Dataset* (ILDP) yang didapat dari *website* UCI Repository.
2. Metode yang digunakan yaitu SMOTE dan G-SMOTE untuk penyeimbangan data serta pengklasifikasian menggunakan metode AdaBoost
3. Fokus pada penelitian yang akan dilakukan yaitu mengetahui perbandingan nilai kinerja performa menggunakan metode SMOTE dan G-SMOTE menggunakan klasifikasi AdaBoost dan tanpa menggunakan penyeimbang data pada pengklasifikasian AdaBoost.

1.4 Tujuan Penelitian

Adapun tujuan yang akan dicapai pada penelitian ini yaitu:

1. Mengetahui kinerja performa algoritma AdaBoost pada klasifikasi penyakit *liver*.
2. Mengetahui kinerja performa algoritma AdaBoost pada klasifikasi penyakit *liver* menggunakan SMOTE.
3. Mengetahui kinerja performa algoritma AdaBoost pada klasifikasi penyakit *liver* menggunakan G-SMOTE.

1.5 Manfaat Penelitian

Adapun manfaat dari penelitian ini yaitu:

1. Membantu penangan deteksi penyakit *liver* sejak dini sehingga penderita bisa

mendapat penangan lebih cepat.

2. Dapat memberi pengetahuan mengenai metode AdaBoost pada klasifikasi penyakit *liver* dengan menggunakan metode *oversampling* SMOTE, G-SMOTE dan tanpa penyeimbang data.

BAB II
TINJAUAN PUSTAKA

2.1 Kajian Terdahulu

Penelitian terdahulu yang memiliki keterkaitan dengan penelitian yang akan dilakukan penulis dan menjadi acuan penulisan serta bertujuan untuk menemukan perbedaan penelitian yang akan dilakukan dengan penelitian terdahulu. Berikut beberapa penelitian terdahulu yang berkaitan dengan penelitian ini

1. Penelitian yang dilakukan oleh Hidayat et al., (2021). Penelitian ini membandingkan antara Algoritma *ADASYN* dan *SMOTE* pada dataset airnbd dengan menggunakan metode *SVM*. Tujuan pada penelitian ini yaitu mengetahui performa teknik *oversampling ADASYN* dan *SMOTE* pada algoritma *SVM*. Teknik yang digunakan untuk menganalisis data adalah oversampling yang bertujuan menangani dataset yang tidak seimbang, dan *confusion matrix* digunakan untuk pengujian Presisi*, Recall, dan F1-SCORE,* serta Akurasi. Hasil temuan menunjukkan bahwa *SMOTE SVM* meningkatkan tingkat akurasi sebesar 1 persen dari 80% menjadi 81% yang dipengaruhi dari kenaikan hasil pengujian label True (minoritas) dan terjadi penurunan pada hasil pengujian label False (mayoritas), metode *SMOTE SVM* memiliki tingkat akurasi paling unggul dibandingkan *ADASYN SVM* dan *SVM* tanpa oversampling.
2. Penelitian oleh Prianti et al., (2020). Penelitian ini tentang menemukan metode yang paling tepat untuk mengetahui tingkat kesehatan keuangan suatu perusahaan agar perusahaan tersebut dapat meningkatkan efisiensi usaha dalam memperoleh laba. Analisis menggunakan metode *KNN* dan *AdaBoost* dengan sepuluh kali perulangan pada masing-masing model menghasilkan model terbaik *KNN* adalah model dengan proporsi pembagian 80%:20% dan $k = 5$ dengan akurasi sebesar 0,82087 sedangkan model terbaik *AdaBoost* adalah model dengan proporsi pembagian 80%:20% dan $M = 30$ dengan akurasi sebesar 0,84522.

3. Penelitian yang dilakukan oleh Qadrini et al., (2021). Penelitian ini menggunakan data sekunder penduduk Desa yang memperoleh Bantuan Langsung Tunai di Kelurahan Banggae, Data yang digunakan dalam penelitian ini adalah data Tahun 2020. Variabel yang digunakan adalah numerik dan kategorik, penelitian ini bertujuan untuk menemukan hasil akurasi yang lebih tinggi maka dilakukan 2 metode kombinasi yaitu metode *Decision Tree* dan *AdaBoost*, pada Seleksi Fitur digunakan untuk menyeleksi data yang rusak/tidak lengkap menggunakan fitur "Input Missing Value" dan "Rename Unused Value" dengan menggunakan metode Decision Tree sehingga didapatkan data set murni. Berdasarkan hasil penelitian yang telah dilakukan menggunakan metode klasifikasi *Decision tree* dan *Adaboost* didapatkan hasil akurasi sebesar 94% dan 95%.
4. Penelitian yang dilakukan oleh Arifiyanti & Wahyuni (2020). Penelitian ini menggunakan data mengenai penipuan kartu kredit dengan Jumlah instances dalam dataset berjumlah 1000 dengan 19 attribute dan 1 attribute klasifikasi yang memiliki 2 kelas yaitu kelas 1 (transaksi penipuan) dan kelas 0 (bukan transaksi penipuan). Penelitian ini bertujuan untuk membuktikan apakah diperlukan teknik resampling jika dihadapkan pada situasi ketidakseimbangan data dengan cara menguji performa (presisi, recall, F1 score, kurva ROC dan AUC) dari berbagai macam algoritma classifier yaitu Logistic Regression, KNN, Decision Tree, dan Naive Bayes. Hasil dari penelitian ini berupa model klasifikasi yang dihasilkan oleh logistic linear, KNN, dan Naive Bayes menunjukkan bahwa metode SMOTE meningkatkan performa model klasifikasi pada perhitungan recall pada klasifikasi Logistic Regression dari 18% menjadi 70%, KNN dari 42% menjadi 64%, dan Naive Bayes dari 32% menjadi 52%, sedangkan decision tree tidak menunjukkan hasil yang berbeda baik sebelum oversampling maupun setelah oversampling.
5. Penelitian yang dilakukan oleh Mualfah et al., (2022)Penelitian ini memakai dataset penyakit stroke dengan total 43.000 pasien, dan 783 mengalami stroke. Tujuan penelitian ini yaitu membuktikan apakah deteksi penyakit stroke menggunakan algoritma random forest menggunakan SMOTE sebagai

penyeimbang data mampu mengatasi pendeteksian penyakit stroke yang lebih baik. Hasil yang didapatkan dari penelitian ini ialah nilai akurasi, presisi, recall, dan f1-score pada algoritma random forest tanpa SMOTE sebesar 0.98, 0.69, 0.51, dan 0.51. Sedangkan algoritma random forest dengan SMOTE mendapatkan masing-masing sebesar 0.91, 0.92, 0.91, 0.91. Terjadi kenaikan signifikan pada presisi, recall, dan f1-score.

6. Penelitian yang dilakukan oleh Douzas et al., (2019). Dataset yang digunakan yaitu LUCAS 2015 yang merupakan data lahan dari barat laut Portugal yang terdiri dari lahan buatan, lahan pertanian, hutan, semak belukar, padang rumput, lahan kosong, air, dan lahan basah. Penelitian ini bertujuan untuk mengetahui metode penyeimbang mana yang lebih baik digunakan pada model klasifikasi. Hasil dari penelitian ini berdasarkan akurasi, f-score, dan G-Mean pada metode klasifikasi LR, KNN, DT, GBC, RF dengan penyeimbang Oversampling ROS, SMOTE, B-SMOTE, ADASYN, G-SMOTE didapatkan hasil tertinggi menggunakan metode penyeimbang G-SMOTE dengan nilai F-Score LR 31%, KNN 28%, GBC 32%, RF 34% dan G-Mean LR 56%, KNN 50%,DT, 51%, GBC 55%, RF 57% dari hasil ini maka didapatkan bahwa untuk perhitungan F-Score dan G-Mean metode penyeimbang G-SMOTE dapat memberikan hasil yang lebih baik dibanding metode penyeimbang oversampling lain.

Pada penelitian ini akan dilakukan penyeimbangan kelas pada dataset penyakit *liver* menggunakan metode *SMOTE* dan *G-SMOTE* yang kemudian diklasifikasikan dengan metode *AdaBoost* dan klasifikasi tanpa penyeimbang data. Tingkat kemungkinan seseorang terkena penyakit liver dibagi kedalam 2 kelas : patient liver (i) dan non patient liver (ii).

Untuk menunjukkan perbedaan antara penelitian terdahulu dengan penelitian yang akan dilakukan dapat dilihat dalam tabel 1 dan untuk rancangan penelitian yang dilakukan dilihat dalam tabel 2.

Tabel 1. Keaslian Penelitian

No	Peneliti	Sampling	Klasifikasi	Data	Hasil
1	Hidayat	ADASYN	SVM	Dataset	Hasil temuan

No	Peneliti	Sampling	Klasifikasi	Data	Hasil
	et al., (2021)	dan SMOTE		Airbnb	menunjukkan bahwa *SMOTE SVM* meningkatkan tingkat akurasi sebesar 1 persen dari 80% menjadi 81% yang dipengaruhi dari kenaikan hasil pengujian label True (minoritas) dan terjadi penurunan pada hasil pengujian label False (mayoritas), metode *SMOTE SVM* memiliki tingkat akurasi paling unggul dibandingkan *ADASYN SVM* dan *SVM* tanpa oversampling.
2	Prianti et al., (2020)	-	AdaBoost dan KNN	Data sekunder laporan keuangan periode ke-2 tahun 2019 dari 575 perusahaan yang tercatat di Bursa Efek Indonesia yang diperoleh dari website resmi milik Bursa Efek Indonesia	Analisis menggunakan metode KNN dan AdaBoost dengan sepuluh kali perulangan pada masing-masing model menghasilkan model terbaik KNN adalah model dengan proporsi pembagian 80%:20% dan k = 5 dengan akurasi sebesar 0,82087 atau sebesar 82% sedangkan model terbaik AdaBoost adalah model dengan proporsi pembagian 80%:20% dan M = 30 dengan akurasi sebesar 0,84522 atau

No	Peneliti	Sampling	Klasifikasi	Data	Hasil
					sebesar 84%
3	Qadrini et al., (2021)	-	AdaBoost dan DT	Data sekunder penduduk Desa yang memperoleh Bantuan Langsung Tunai di Kelurahan Banggae tahun 2020	Hasil penelitian yang telah dilakukan menggunakan metode klasifikasi *Decision tree* dan *Adaboost* didapatkan hasil akurasi sebesar 94% dan 95%.
4	Arifiyanti dan Eka, (2020)	SMOTE	LR, KNN dan NB	Data mengenai penipuan kartu kredit dengan Jumlah instances dalam dataset berjumlah 1000 dengan 19 attribute dan 1 attribute klasifikasi yang memiliki 2 kelas yaitu kelas 1 (transaksi penipuan) dan kelas 0 (bukan transaksi penipuan)	Hasil dari penelitian ini berupa model klasifikasi yang dihasilkan oleh Logistic Regression, KNN, dan Naive Bayes menunjukkan bahwa metode SMOTE meningkatkan performa model klasifikasi pada perhitungan recall pada klasifikasi Logistic Regression dari 18% menjadi 70%, KNN dari 42% menjadi 64%, dan Naive Bayes dari 32% menjadi 52%, sedangkan decision tree tidak menunjukkan hasil yang berbeda baik sebelum oversampling maupun setelah oversampling.
5	Mualfah et al.,(2022)	SMOTE	RF	Dataset penyakit stroke dengan total	Hasil yang didapatkan dari penelitian ini ialah nilai akurasi, presisi,

No	Peneliti	Sampling	Klasifikasi	Data	Hasil
				43.000 pasien, dan 783 mengalami stroke	recall, dan f1-score pada algoritma random forest tanpa SMOTE sebesar 0.98, 0.69, 0.51, dan 0.51. Sedangkan algoritma random forest dengan SMOTE mendapatkan masing-masing sebesar 0.91, 0.92, 0.91, 0.91. Terjadi kenaikan signifikan pada presisi, recall, dan f1-score.
6	Douzas et al.,(2019)	ROS, SMOTE, B-SMOTE, ADASYN, dan G-SMOTE	LR, KNN, DT, GBC dan RF	Dataset LUCAS 2015 yang merupakan data lahan dari barat laut Portugal yang terdiri dari lahan buatan, lahan pertanian, hutan, semak belukar, padang rumput, lahan kosong, air, dan lahan basah	Hasil dari penelitian ini berdasarkan akurasi, f-score, dan G-Mean pada metode klasifikasi LR, KNN, DT, GBC, RF dengan penyeimbang Oversampling ROS, SMOTE, B-SMOTE, ADASYN, G-SMOTE didapatkan hasil tertinggi menggunakan metode penyeimbang G-SMOTE dengan nilai F-Score LR 31%, KNN 28%, GBC 32%, RF 34% dan G-Mean LR 56%, KNN 50%,DT, 51%, GBC 55%, RF 57% dari hasil ini maka didapatkan bahwa untuk perhitungan F-Score dan G-Mean metode

No	Peneliti	Sampling	Klasifikasi	Data	Hasil
					penyeimbang G-SMOTE dapat memberikan hasil yang lebih baik dibanding metode penyeimbang oversampling lain, sedangkan untuk nilai akurasi tanpa penambahan metode penyeimbang maka hasilnya lebih baik.

Tabel 2. Penelitian yang akan dilakukan

No	Peneliti	Sampling	Klasifikasi	Data	Hasil
1	Munawwarah	SMOTE dan G-SMOTE	AdaBoost	Indian Liver Patient Dataset (ILPD)	Mengetahui nilai akurasi dan kinerja performa yang dihasilkan oleh metode AdaBoost menggunakan Teknik penyeimbang Oversampling SMOTE dan G-SMOTE dan tanpa menggunakan metode penyeimbang data dalam mengklasifikasi penyakit liver

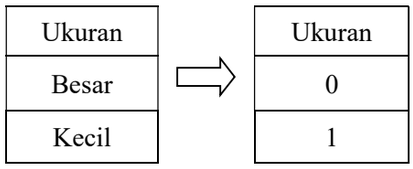
2.1 Liver

Liver adalah organ yang vital bagi manusia. Terdapat beberapa fungsi kerja *liver* antara lain sebagai penawar dan penetralisir racun, mengatur sirkulasi hormon, mengatur komposisi darah yang mengandung lemak, gula, protein, dan zat lain. *Liver* juga berfungsi membuat empedu, zat yang membantu pencernaan lemak. Penyakit *liver* merupakan suatu gangguan pada setiap fungsi *liver*. *Liver* bertanggung jawab untuk fungsi-fungsi kritis dalam tubuh, dimana hilangnya fungsi-fungsi tersebut dapat menyebabkan kerusakan yang signifikan pada tubuh. *Liver* adalah satu-satunya organ dalam tubuh yang dapat dengan mudah mengganti sel-sel yang rusak, tetapi jika sel-sel itu hilang,

maka *liver* tidak mungkin dapat memenuhi kebutuhan tubuh. Penyakit *liver* sering disebut sebagai pembunuh diam-diam karena kemungkinan tidak timbulnya gejala (Pusporani et.al, 2019).

2.2 Label Encoder

Dalam pembelajaran supervised pada *machine learning*, jika data dalam bentuk angka maka algoritma dapat membaca atau mengoperasikan secara langsung. Namun, sering kali label data dapat berupa bentuk kategorik ataupun bentuk teks, dimana harus diubah kedalam bentuk angka sehingga algoritma dapat menggunakannya langsung untuk memulai proses training data (Joshi, Python Machine Learning Cookbook, 2016). Contoh penggunaan *label encoder* dapat dilihat pada gambar 1.



Gambar 1. Contoh label encoder

2.4 Min-Max Normalization

Min-Max Normalization berfungsi untuk mengubah ukuran data pada rentang asli, yaitu semua nilai berada dalam kisaran 0 dan 1. Persamaan *min-max normalization* ditunjukkan pada persamaan (Li & Liu, 2011). Tujuan dari normalisasi data adalah untuk memberikan semua atribut data memiliki bobot nilai yang sama. *min-max normalization* ini dapat mempermudah perbandingan nilai dalam sebuah data yang memiliki ukuran angka yang berbeda. Fungsi *min-max normalization* dapat dilihat pada persamaan (Virmani, 2015):

$$Z = \frac{x - \min(x)}{\max(x) - \min(x)} \quad \ldots(1)$$

Keterangan :

z = hasil normalisasi

x = nilai awal

min(x) = nilai minimal dari attribut yang dihitung

max(x) = nilai maksimal dari attribut yang dihitung

2.5 *Splitting Data*

Splitting Data adalah pembagian data menjadi 2 atau lebih sub data biasanya kita kenal dengan data *training* dan data *testing*. *splitting Data* ini menjadi penting dalam data sains terutama pada bagian pembuatan model data. Pada pembagian 2 data umumnya akan terbentuk data *training* yaitu data yang digunakan untuk membangun model. Data *testing* adalah data yang digunakan setelah proses *training* model selesai. Rasio pembagian data itu dibagi berdasarkan ukuran dari dataset yang dimiliki hal ini karena tidak adanya panduan atau metric dalam *splitting data*. Dalam pembagian data biasanya digunakan pembagian 80:20 atau dikenal dengan *Pareto Principle* yang sering dipakai dalam matematika, ekonomi dan komputer.

2.6 AdaBoost

AdaBoost merupakan salah satu dari beberapa varian pada algoritma *boosting* (Liu, 2015). Adaboost dan variannya telah sukses diterapkan pada beberapa bidang (domain) karena dasar teorinya yang kuat, prediksi yang akurat, dan kesederhanaan yang besar (Li et al., 2008). AdaBoost bersifat adiktif yang dalam artian pembelajaran klasifikasi yang lemah akan di-*tweak* sebagai contoh dari pengklasifikasian yang salah klasifikasi pada data sebelumnya (Saragih et al., 2021). Dalam menentukan sebuah keputasan dalam pengklasifikasian sederhananya melalui proses iterasi, yang mana perhitungan akan diputuskan berdasarkan hasil *weak-classifier* yang telah melalui proses perhitungan, *weak-classifier* yang terbentuk akan digabung dalam menentukan keputusan pada klasifikasi dengan syarat *weak-classifier* yang digunakan memiliki tingkat *error* harus kurang dari 0,5 (Prasvita, 2016). Langkah-langkah perhitungan pada algoritma Adaboost adalah sebagai berikut dan bisa dilihat juga pada gambar 2:

a. Masukkan suatu kumpulan sampel penelitian dengan label $\{(x_i , y_i), \dots , (x_N, y_N)\}$, suatu *component learn* algoritma, jumlah perputaran T.
b. Hitung bobot suatu sampel pelatihan $w_t^1 = {}^1/_m$ …(2)
   untuk semua i = 1, 2, 3… , m
c. Untuk t= 1, ...,T