Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 5 No. 1 (2021) 187 - 192 ISSN Media Elektronik: 2580-0760

Analisis Perbandingan Algoritma Optimasi pada *Random Forest* untuk Klasifikasi Data Bank Marketing

Yoga Religia[1], Agung Nugroho[2], Wahyu Hadikristanto[3]
[1,2,3]Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
[1]yoga.religia@pelitabangsa.ac.id, [2]agung@pelitabangsa.ac.id, [3]wahyu.hadikristanto@pelitabangsa.ac.id

Abstract

The world of banking requires a marketer to be able to reduce the risk of borrowing by keeping his customers from occurring non-performing loans. One way to reduce this risk is by using data mining techniques. Data mining provides a powerful technique for finding meaningful and useful information from large amounts of data by way of classification. The classification algorithm that can be used to handle imbalance problems can use the Random Forest (RF) algorithm. However, several references state that an optimization algorithm is needed to improve the classification results of the RF algorithm. Optimization of the RF algorithm can be done using Bagging and Genetic Algorithm (GA). This study aims to classify Bank Marketing data in the form of loan application receipts, which data is taken from the www.data.world site. Classification is carried out using the RF algorithm to obtain a predictive model for loan application acceptance with optimal accuracy. This study will also compare the use of optimization in the RF algorithm with Bagging and Genetic Algorithms. Based on the tests that have been done, the results show that the most optimal performance of the classification of Bank Marketing data is by using the RF algorithm with an accuracy of 88.30%, AUC (+) of 0.500 and AUC (-) of 0.000. The optimization of Bagging and Genetic Algorithm has not been able to improve the performance of the RF algorithm for classification of Bank Marketing data.

Keywords: Data Mining, Bank Marketing, Random Forest, Bagging, Genetic Algorithm

Abstrak

Dunia perbankan mengharuskan seorang *marketing* agar mampu mengurangi resiko peminjaman dengan cara menjaga nasabahnya agar tidak terjadi kredit bermasalah. Salah satu cara untuk mengurangi resiko tersebut adalah dengan menggunakan teknik data mining. Data mining menyediakan teknik yang kuat untuk menemukan informasi yang bermakna dan berguna dari sejumlah besar data dengan cara klasifikasi. Algoritma klasifikasi yang dapat digunakan untuk menangani masalah *imbalance* dapat menggunakan algoritma *Random Forest* (RF). Namun demikian beberapa referensi menyebutkan bahwa diperlukan algoritma optimasi guna meningkatkan hasil klasifikasi dari algoritma RF. Optimasi algoritma RF dapat dilakukan dengan menggunakan *Bagging* dan *Genetic Algorithm* (GA). Penelitian ini bertujuan untuk mengklasifikasikan data *Bank Marketing* berupa penerimaan pengajuan pinjaman yang mana datanya diambil dari situs www.data.world. Klasifikasi dilakukan dengan menggunakan algoritma RF untuk memperoleh model prediksi penerimaan pengajuan pinjaman dengan akurasi yang optimal. Penelitian ini juga akan membandingkan penggunaan optimasi pada algoritma RF dengan *Bagging* dan *Genetic Algorithm*. Berdasarkan pengujian yang telah dilakukan diperoleh hasil bahwa performa paling optimal dari klasifikasi data *Bank Marketing* adalah dengan menggunakan algoritma RF dengan akurasi sebesar 88,30%, AUC (+) sebesar 0,500 dan AUC(-) sebesar 0,000. Adapun penggunaan optimasi *Bagging* dan *Genetic Algorithm* ternyata belum mampu meningkatkan performa dari algoritma RF untuk klasifikasi data *Bank Marketing*.

Kata kunci: Data Mining, Bank Marketing, Random Forest, Bagging, Genetic Algorithm

1. Pendahuluan

Pembiayaan kredit adalah penyedian dana berdasarkan persetujuan pinjam meminjam antara Bank dengan nasabah atau pihak lain yang mewajibkan pihak peminjam agar melunasi pinjamannya setelah jangka waktu tertentu [1]. Bagian *marketing* perlu menyeleksi calon nasabah untuk diberikan kredit dengan beberapa pertimbangan seperti kepercayaan, tenggang waktu, tingkat risiko serta objek kredit. Hal tersebut diperlukan karena seorang *marketing* wajib menjaga nasabahnya agar tidak terjadi kredit bermasalah dimana kredit bermasalah kerap menjadi resiko utama dalam setiap pemberian kredit [2]. Salah satu cara untuk mengurangi resiko pengajuan kredit adalah dengan menambang

Diterima Redaksi: 07-01-2021 | Selesai Revisi: 27-02-2021 | Diterbitkan Online : 28-02-2021

informasi dari data yang sudah ada sebelumnya, dimana teknik yang dapat digunakan adalah data mining [3].

Data mining (DM) menyediakan teknik yang kuat untuk menemukan informasi yang bermakna dan berguna dari sejumlah besar data, sehingga sangat berguna untuk diaplikasikan pada dunia nyata [4]. Teknik data mining secara luas dibagi menjadi dua kategori yaitu prediktif dan deskriptif. Pada metode prekdiktif dapat dilakukan dengan model klasifikasi. Klasifikasi adalah proses mengubah catatan data menjadi sekumpulan kelas yang sama [5]. Situs www.data.world telah menyediakan set data *Bank Marketing* berupa data sampel yang terdiri dari 45.211 *record* pengajuan kredit dengan 16 atribut dan tidak ada *missing value*, sehingga dapat digunakan untuk membangun model klasifikasi [6]. Penelitian ini menggunakan seluruh *record* dari set data *Bank Marketing* yang diambil dari situs www.data.world. Berdasarkan 16 atribut yang ada, jenis data dari set data *Bank Marketing* termasuk kategori data *imbalance*, sehingga dibutuhkan algoritma yang tepat untuk mengklasifikasikan data tersebut.

Menurut beberapa penelitian algortma *Random Forest* (RF) dapat digunakan untuk klasifikasi data *imbalance* dalam jumlah besar dengan memberikan hasil performa yang baik dan waktu eksekusi yang cepat [7] [8]. Penelitian yang dilakukan oleh Wei Chen (2017) mengungkapkan bahwa dalam hal klasifikasi, algoritma RF mampu memberikan akurasi yang lebih besar dibandingkan dengan algoritma *tree* yang lain seperti *logistic model tree* (LMT) dan *classification and regression tree* (CART) [9]. Namun demikian, beberapa penelitian menghimbau agar dalam model kasifikasi dilakukan optimasi untuk membuat performa yang dihasilkan dapat lebih baik [10]. Dengan melakukan optimasi dapat membantu proses evaluasi kecocokan setiap *instance* data dengan hasil yang diinginkan [11].

Eyad Elyan dan M. Medhat Gaber menyebutkan bahwa pada algoritma RF, pengaturan parameter untuk penentuan kelas data memiliki ruang pencarian yang besar dan overfitting sehingga menjadi masalah optimasi [12]. Salah satu cara optimasi yang dapat digunakan untuk mencegah overfitting dan mengurangi varians data adalah dengan menggunakan metode *Bagging* [13]. Selain menggunakan *Bagging*, optimasi algoritma RF juga dapat dilakukan menggunakan *Genetic Algorithm* (GA), karena strategi pengoptimalan pada GA dapat digambarkan sebagai prosedur pengoptimalan global dengan keuntungan tidak bergantung pada nilai awal parameter untuk mendapatkan konvergensi data [14] .

Algoritma *Bootstrap Aggregation* atau yang lebih dikenal dengan nama *Bagging* adalah metode ensemble yang digunakan untuk mengklasifikasikan data dengan akurasi yang baik [15]. *Bagging* mampu mengurangi tingkat kesalahan klasifikasi kasusnya klasifikasi dengan 50 pengulangan [16]. Algortma *Bagging* dapat digunakan untuk menghindari penurunan keragaman, meningkatkan kecepatan klasifikasi dan menurunkan kebutuhan memori [17]. Sedangkan pada GA, dapat memberikan *hyper* parameter dengan hasil yang hampir sama dengan pencarian grid dengan waktu komputasi yang lebih cepat [18]. Beberapa penelitian menyebutkan bahwa optimasi GA dapat meningkatkan performa dari algoritma RF dengan cukup signifikan [19] [20].

Berdasarkan pembahasan pada paragraph sebelumnya penelitian ini akan membahas tentang penggunaan algoritma RF untuk klasifikasi data *Bank Marketing*. Selain itu akan digunakan juga metode optimasi *Bagging* dan GA untuk dilihat apakah dengan menggunakan kedua algoritma optimasi tersebut dapat meningkatkan performa dari algoritma RF untuk klasifikasi data *Bank Marketing*.

2. Metode Penelitian

2.1. Data yang digunakan

Penelitian ini menggunakan data sekunder yang diambil dari situs www.data.world pada tanggal 02 November 2020 berupa set data *Bank Marketing* [6]. Set data *Bank Marketing* terdiri dari tiga variabel yaitu data klien bank, kontak terakhir klien, dan label. Setiap variabel memiliki atribut masing-masing dimana atribut dari setiap variabel dapat dilihat pada Tabel 1. Adapun jumlah *record* data yang terdapat pada data *Bank Marketing* adalah sebanyak 45.211 record yang terdiri dari 16 atribut dan tidak terdapat *missing value*, sehingga tidak memerlukan *pre-processing* data.

Tabel 1. Atribut Data *Bank Marketing*

Variabel	Atribut	Type Data
Data Klien Bank	Age	Integer
	Job	String
	Marital	String
	Education	String
	Default	Boolean
	Balance	Integer
	Housing	Boolean
	Loan	Boolean
kontak terakhir klien	Day	Integer
	Month	String
	Duration	Integer
	Campaign	Integer
	Pdays	Integer
	Previous	Integer
	Poutcome	String
Label	Accapted (y)	Boolean

2.2. Model Penelitian

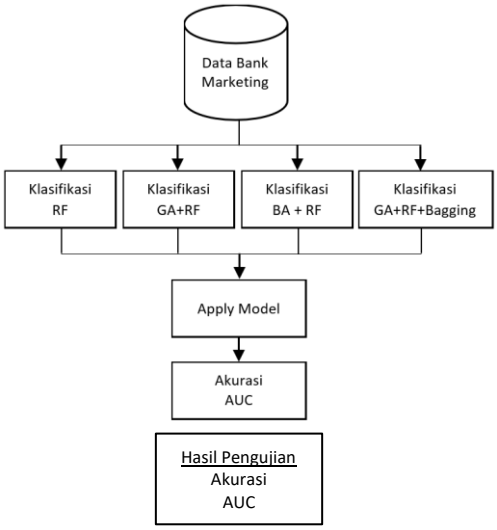
Penelitian ini menggunakan model perbandingan antara penggunaan *Random Forest* (RF) dan RF dengan optimasi *Bagging* dan *Genetic Algoritm* (GA) untuk klasifikasi data *Bank Marketing*. Adapun perbandingan dilakukan sebanyak empat kali yaitu:

1) Klasifikasi data *Bank Marketing* menggunakan RF
2) Klasifikasi data *Bank Marketing* menggunakan RF dengan optimasi *Bagging*

DOI: https://doi.org/10.29207/resti.v5i1.2813

3) Klasifikasi data *Bank Marketing* menggunakan RF dengan optimasi GA
4) Klasifikasi data *Bank Marketing* menggunakan RF dengan optimasi *Bagging* dan GA.

Proses validasi pada penelitian ini menggunakan *cross* validation. Adapun secara lebih jelas dari model penelitian yang digunakan pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Model Penelitian yang dibangun

Pada tahapan *apply model* dilakukan penerapan model yang telah dilatih sebelumnya menggunakan data *training* pada data *testing*. Berdasarkan Gambar 1 dapat diketahui bahwa hasil pengujian model adalah berupa nilai akurasi dan nilai *area under the curve* (AUC). Beberapa penelitian mungkin hanya menguji performa algoritma klasifikasi dengan melihat nilai akurasinya saja, akan tetapi untuk beberapa data tertentu juga membutuhkan pengukuran performa yang lain. Penelitian ini juga menggunakan pengujian AUC untuk melihat performa dari model penelitian, karena pengujian AUC dapat digunakan untuk melihat hasil yang lebih detail dari pengujian akurasi [21].

2.3. Klasifikasi dengan *Random Forest*

Random Forest (RF) adalah algoritma yang menggunakan metode pemisahan biner rekursif untuk mencapai node akhir dalam struktur pohon berdasarkan pada pohon klasifikasi dan regresi [22]. Breiman tahun 2001 memperkenalkan algoritma RF dengan menunjukkan beberapa kelebihan diantaranya mampu menghasilkan error yang relatif rendah, performa yang baik dalam klasifikasi, dapat mengatasi data pelatihan dalam jumlah besar secara efisien, serta metode yang efektif untuk mengestimasi missing data. RF menghasilkan banyak pohon independen dengan subset yang dipilih secara acak melalui *bootstrap* dari sampel pelatihan dan dari variabel input disetiap node. RF melakukan klasifikasi dengan cara mengadopsi pendekatan ansambel dari berbagai pohon melalui kemunculan mayoritas untuk mencapai keputusan akhir [23].

Set data pelatihan pada algoritma RF diformulasikan sebagai $S = \{(x_i, y_j), i = 1, 2, \ldots, N; j = 1, 2, \ldots, M\}$, dimana x adalah sampel dan y adalah variabel fitur S. N adalah jumlah sampel pelatihan, dan ada variabel fitur M di setiap sampel [24]. Adapun dalam pembangunan algoritma RF terdiri dari 3 langkah yaitu: (1) *Sampling* himpunan bagian pelatihan k, (2) Pembuatan setiap model pohon keputusan, dan (3) Pengumpulan k pohon ke dalam model RF. Penggunaan algoritma RF untuk klasifikasi dapat diterapkan pada data *imbalance* dalam jumlah besar dengan memberikan hasil performa yang baik dan waktu eksekusi yang cepat [7] [8].

2.4. Optimasi dengan *Bagging*

Bootstrap aggregating atau biasa disebut *Bagging* adalah metode ensembel yang biasa diterapkan pada kasus klasifikasi yang bertujuan untuk meningkatkan akurasi klasifikasi dengan menggabungkan klasifikasi tunggal, dan hasilnya lebih baik daripada random *sampling* [25]. *Bagging* melakukan *bootstrap* atau teknik *sampling* pada data asli sebanyak n kali dengan penggantian untuk membuat set pelatihan [26]. Setiap data latih dibuat pohon klasifikasi dan proses agregat atau suara terbanyak untuk kasus klasifikasi dan rata-rata untuk kasus regresi [27].

Penerapan *Bagging* pada algoritma *Tree* dilakukan dengan cara pohon keputusan diturunkan dengan membangun pengklasifikasi dasar C_1, C_2,…, C_n pada sampel *bootstrap* D_1, D_2, .., D_n dengan penggantian dari kumpulan data D. Kemudian model akhir atau pohon keputusan diturunkan sebagai a kombinasi dari semua pengklasifikasi dasar C_1, C_2,…, C_n dengan suara terbanyak. Cara ini dapat diterapkan pada semua pengklasifikasi dengan pohon keputusan seperti REP *Tree*, *random forest*, C4.5 J48 dll [28].

2.5. Optimasi dengan *Genetic Algorithm*

Genetic Algorithm (GA) adalah algoritma dengan pendekatan pencarian heuristik yang dapat diterapkan pada berbagai masalah optimasi [29]. GA melakukan proses optimasi berdasarkan pada populasi sampel. GA mengembangkan populasi solusi kandidat menuju solusi yang lebih baik [30]. Selama pelaksanaan GA, populasi individu mengalami operator genetik seperti: seleksi, persilangan dan mutasi. Pada tahap seleksi terdapat fungsi yang digunakan untuk mengevaluasi individu-individu pada populasi yang bertujuan untuk mengukur kesesuaian solusi tertentu terhadap masalah yang diberikan. Individu yang paling cocok selama langkah ini dipilih untuk mereproduksi dan menghasilkan individu baru yang mungkin merupakan solusi yang baik untuk masalah tersebut [31].

Menurut teori Darwin, individu yang paling bugar cenderung memiliki kemungkinan lebih baik untuk menyebarkan gen mereka ke generasi mendatang dan

DOI: https://doi.org/10.29207/resti.v5i1.2813

dengan demikian menghasilkan keturunan yang lebih beradaptasi dengan lingkungan tempat mereka tinggal [32]. GA dikembangkan berdasarkan mekanisme seleksi alam dan genetika meniru makhluk hidup untuk memecahkan masalah sulit dengan kompleksitas tinggi dan struktur yang tidak diinginkan. GA dapat dikembangkan sebagai solusi untuk masalah dunia nyata jika dikodekan dengan tepat.

2.6. Penilaian Klasifikasi

Indikator penilaian sangat penting untuk mengevaluasi kinerja setiap algoritma pembelajaran mesin. Terdapat banyak indikator penilaian dibidang klasifikasi diantaranya akurasi dan *area under the curve* (AUC). Akurasi adalah persentase sampel target dan non-target yang diprediksi dengan benar dan mencerminkan kemampuan pengklasifikasi sebagian data sampel (data *testing*) untuk menentukan seluruh sampel (data *training*) [33]. Pengukuran akurasi tidak dipengaruhi oleh banyak atau sedikitnya data saja melainkan juga dari imbalancing dari data yang digunakan. Adapun akurasi dapat diukur dengan persamaan berikut:

$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN} \times 100$$

True Positive (TP) adalah jumlah sampel positif yang diprediksi benar; *False Positive* (FP) adalah banyaknya sampel positif yang prediksi salah; *True Negative* (TN) adalah jumlah sampel negatif yang diprediksi dengan benar; *False Negative* (FN) adalah jumlah sampel negatif yang diprediksi salah.

Area under the curve (AUC) adalah parameter yang secara intuitif mengukur probabilitas skor sampel positif lebih besar dari sampel negatif saat pengambilan sampel positif dan sampel negatif secara acak. AUC adalah penilaian kinerja model yang sudah biasa digunakan. Nilai AUC berkisar dari 0 hingga 1, di mana 1 mewakili kinerja optimal dan 0 mewakili kinerja terburuk [21]. pengujian AUC juga dapat digunakan untuk melihat hasil yang lebih detail dari pengujian akurasi.

3. Hasil dan Pembahasan

Penelitian ini menggunakan dataset sekunder berupa data *bank marketing* yang diambil dari situs www.data.world yang kemudian diklasifikasikan menggunakan algoritma RF. Algoritma optimasi juga digunakan pada penelitian ini untuk dibandingkan apakah dengan menggunakan algoritma optimasi dapat memberikan peningkatan akurasi dari algoritma RF. Adapun algoritma optimasi yang digunakan pada penelitian ini ada 2, yaitu GA dan *Bagging*. Setelah dilakukan pengujian akan diperoleh nilai akurasi dan nilai AUC untuk kemudian dilakukan analisa hasil pengujian.

3.1. Hasil Penelitian

Pengujian yang dilakukan menggunakan *cross validation*. Penggunaan *cross validation* dipilih karena dapat membagi *dataset* menjadi k bagian *dataset* dengan ukuran yang sama. Setiap kali berjalan, satu pecahan berperan sebagai *dataset testing* sedangkan pecahan lainnya menjadi *dataset training*. Hasil dari pengujian klasifikasi data *bank marketing* dengan optimasi GA dan *Bagging* dapat dilihat pada Tabel 2.

Tabel 2. Hasil Akurasi Pengujian Model Penelitian

Algoritma	Akurasi
RF	88,30%
RF + Bagging	88,30%
GA + RF	88,30%
GA + RF + Bagging	88,30%

Berdasarkan Tabel 2 dapat diketahui bahwa nilai akurasi dari algoritma *RF* untuk klasifikasi data *Bank Marketing* adalah sebesar 88,30%. O ptimasi algoritma RF dengan algoritma *Bagging* untuk klasifikasi data *Bank Marketing* juga memperoleh akurasi sebesar 88,30%, sedangkan optimasi RF dengan GA untuk klasifikasi data *Bank Marketing* memperoleh akurasi 88,30%, bahkan kombinasi optimasi menggunakan GA dan *Bagging* pada algoritma RF untuk klasifikasi data *Bank Marketing* juga memperoleh akurasi yang sama, yaitu 88,30%. Hasil pengujian yang diperoleh menunjukkan bahwa optimasi penggunaan algoritma *Bagging*, GA ataupun kombinasi dari kedua algoritma optimasi tersebut ternyata belum dapat meningkatkan akurasi dari algoritma RF untuk klasifikasi data *Bank Marketing*.

Hasil pengujian dari penelitian ini selain digunakan untuk mengetahui nilai akurasi, juga digunakan untuk mengetahui nilai *area under the curve* (AUC), dimana nilai AUC diperoleh dari penghitungan hubungan antara *false positive* dan *true positive*. Semakin tinggi nilai AUC, maka semakin baik baik model klasifikasi yang digunakan. Memeriksa "kecuraman" kurva penting bagi penelitian ini karena menggambarkan maksimalisasi rasio *true positive* sambil meminimalkan rasio *false positive*. Adapun hasil AUC dari pengujian model penelitian ini dapat dilihat pada Tabel 3.

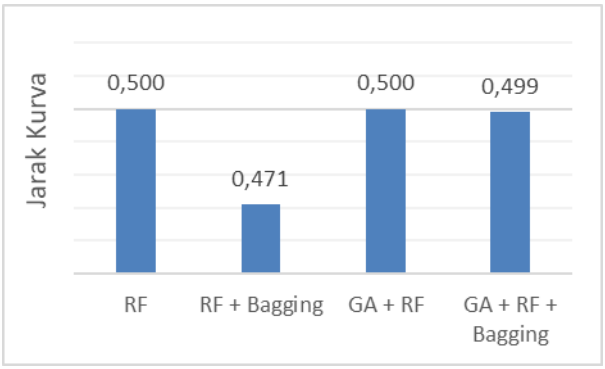
Tabel 3. Hasil AUC Pengujian Model Penelitian

Algoritma	AUC (+)	AUC (-)
RF	0,500	0,000
RF + Bagging	0,514	0,043
GA + RF	0,500	0,000
GA + RF + Bagging	0,567	0,068

Berdasarkan Tabel 2 dapat dilihat bahwa nilai AUC dari algoritma RF untuk klasifikasi data *Bank Marketing* sama dengan nilai AUC dari algoritma RF yang dioptimasi dengan *Genetic Algoritm* dimana kurva *true positive* bernilai 0,500 sedangkan kurva *false positive* bernilai 0,000, sehingga jarak dari kurva *true positive* terhadap kurva *false positive* adalah sebesar 0,500. Hasil AUC dari algoritma RF dengan optimasi *Bagging* untuk klasifikasi data *Bank Marketing* memperoleh kurva *true*

DOI: https://doi.org/10.29207/resti.v5i1.2813

positive bernilai 0,514 sedangkan kurva *false positive* bernilai 0,043, sehingga jarak dari kurva *true positive* terhadap kurva *false positive* adalah sebesar 0,471. Hasil AUC dari algoritma RF dengan optimasi *Genetic Algoritm* dan *Bagging* untuk klasifikasi data *Bank Marketing* memperoleh kurva *true positive* bernilai 0,567 sedangkan kurva *false positive* bernilai 0,068, sehingga jarak dari kurva *true positive* terhadap kurva *false positive* adalah sebesar 0,499. Adapun jarak dari kurva *true positive* terhadap kurva *false positive* pada setiap algoritma yang digunakan pada penelitian ini dapat dilihat pada Gambar 2.



Gambar 2. Jarak Kurva *True Positive* Terhadap Kurva *False Positive*

3.2. Analisa Hasil Penelitian

Berdasarkan hasil akurasi dari pengujian model klasifikasi yang digunakan pada penelitian ini dapat diketahui bahwa penggunaan optimasi *Bagging* dan GA ternyata belum mampu meningkatkan akurasi dari algoritma RF untuk klasifikasi data *Bank Marketing*. Namun demikian berdasarkan nilai AUC memperoleh hasil yang berbeda. Berdasarkan jarak dari kurva *true positive* terhadap kurva *false positive* pada setiap algoritma yang digunakan dapat diketahui bahwa:

1) Penggunan algoritma RF untuk klasifikasi data *Bank Marketing* sama baiknya dengan penggunaan algoritma RF dengan optimasi GA dengan dengan jarak antar kurva 0,500. Artinya, dengan adanya optimasi GA atau tanpa adanya optimasi GA pada algoritma RF tidak memberikan dampak pada hasil klasifikasi data *Bank Marketing*.
2) Penggunaan optimasi *Bagging* pada algotitma RF ataupun penggunaan optimasi *Genetic Algoritm* dan *Bagging* pada algoritma RF untuk klasifikasi data *Bank Marketing* tidak lebih baik dari penggunaan algoritma RF saja. Artinya optimasi *Bagging* pada algotitma RF ataupun penggunaan optimasi GA dan *Bagging* pada algoritma RF tidak diperlukan untuk klasifikasi data *Bank Marketing*.

4. Kesimpulan

Penelitian ini telah menguji algoritma RF dengan optimasi GA dan *Bagging*. Hasil pengujian menunjukkan bahwa dengan penggunaan optimasi GA ataupun *Bagging* belum mampu meningkatkan akurasi dari algoritma RF untuk klasifikasi set data *Bank Marketing*, dimana dengan menggunakan optimasi ataupun tidak akurasi yang diperoleh adalah sebesar 88,30%. Optimasi yang dilakukan belum berhasil disebabkan data yang digunakan terlalu *imbalance*, sehingga ketika dilakukan proses *bagging* ataupun GA sebaran data yang dihasilkan masih cukup banyak. *Imbalance* data dapat dilihat dari setiap atribut pada set data *Bank Marketing* yang memiliki type data yang berbeda dan sebaran data yang terlalu bervariasi. Adapun nilai AUC memperoleh hasil yang berbeda. Namun demikian jarak dari kurva *true positive* terhadap kurva *false positive* optimal berada pada penggunaan algoritma RF tanpa optimasi. Berdasarkan temuan tersebut dapat dikatakan bahwa optimasi *Bagging* pada algoritma RF ataupun penggunaan optimasi GA dan *Bagging* pada algoritma RF tidak diperlukan untuk klasifikasi data *Bank Marketing*.

Penelitian ini belum memberikan nilai akurasi yang cukup baik untuk klasifikasi data *Bank Marketing*, sehingga diperlukan penelitian lebih lanjut tentang metode optimasi yang mampu meningkatkan performa algoritma RF untuk klasifikasi data *Bank Marketing* dengan lebih memperhatikan *imbalencing* data, misalkan menggunakan optimasi *feature selection*. Penggunaan optimasi *feature selection* memungkinkan untuk melakukan pemilihan subset dari fitur yang relevan untuk digunakan dalam konstruksi model sehingga diharapkan dapat mengatasi masalah *imbalencing* data.

Daftar Rujukan

[1] A. T. Rahmawati, M. Saifi and R. R. Hidayat, "Analisis Keputusan Pemberian Kredit dalam Langkah Meminimalisir Kredit Bermasalah," *Jurnal Administrasi Bisnis*, vol. 35, no. 1, pp. 179-186, 2016.

[2] S. Somadiyono and T. Tresya, "Tanggung Jawab Pidana Marketing Menurut Undang Undang Perbankan Terhadap Pembiayaan Bermasalah di Bank Muamalat Indonesia,Tbk," *Jurnal Lex Specialis*, vol. 21, pp. 22-38, 2015.

[3] S. Masripah, "Komparasi Algoritma Klasifikasi Data Mining untuk Evaluasi Pemberian Kredit," *Bina Insani ICT Journal*, vol. 3, no. 1, pp. 187-193, 2016.

[4] W. Gan, J. C.-W. C. H.-C. Lin and J. Zhan, "Data mining in Distributed Environment: A Survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* , vol. 7, no. 6, pp. 1-19, 2017.

[5] S. Umadevi and K. S. J. Marseline, "A Survey on Data Mining Classification Algorithms," in *International Conference on Signal Processing and Communication*, Coimbatore, India, 2017.

[6] "Data.World," Data.World, Inc, 2016. [Online]. Available: https://data.world/uci/bank-marketing. [Accessed 1 Desember 2020].

[7] A. S. More and D. P. Rana, "Review of Random Forest Classification Techniques to Resolve Data Imbalance," in *International Conference on Intelligent Systems and Information Management*, Aurangabad, India, 2017.

[8] A. Parmar, R. Katariya and V. Patel, "A Review on Random Forest: An Ensemble Classifier," in *International Conference*

DOI: https://doi.org/10.29207/resti.v5i1.2813

on Intelligent Data Communication Technologies and Internet of Things, Springer, Cham, 2018.

[9] W. Chen, X. Xie, B. Pradhan, H. Hong, D. T. Bui, Z. Duan and J. Ma, "A Comparative Study of Logistic Model Tree, Random Forest, and Classification and Regression Tree Models for Spatial Prediction of Landslide Susceptibility," *Catena* , vol. 151 , pp. 147-160, 2017.

[10] F. Burger and J. Pauli, "Understanding the Interplay of Simultaneous Model Selection and Representation Optimization for Classification Tasks," in *International Conference on Pattern Recognition Applications and Methods*, Lisbon, Portugal, 2016.

[11] A. P. D. Silva, "Optimization Approaches to Supervised Classification," *European Journal of Operational Research*, vol. 261, no. 2, pp. 772-788, 2017.

[12] E. Elyan and M. M. Gaber, "A Genetic Algorithm Approach to Optimising Random Forests Applied to Class Engineered Data," *Information Sciences*, vol. 384, no. 1, pp. 220-234, 2017.

[13] A., Arfiani, Z. and Rustam, "Ovarian Cancer Data Classification Using Bagging and Random Forest," in *AIP Conference Proceedings*, Depok, 2019.

[14] S. A. Naghibi, K. Ahmadi and A. Daneshi, "Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping," *Water Resour Manage*, vol. 31, no. 9, p. 2761–2775, 2017.

[15] V. Chaurasia and S. Pal, "Data Mining Approach to Detect Heart Dieses," *International Journal of Advanced Computer Science and Information Technology*, vol. 2, no. 4, pp. 56-66, 2013.

[16] L. Bieman, "Bagging Predictors," *Machine Learning*, vol. 24, pp. 123-140, 1996.

[17] S. E. Roshan and S. Asadi, "Improvement of Bagging Performance for Classification of Imbalanced Datasets Using Evolutionary Multi-objective Optimization," *Engineering Applications of Artificial Intelligence*, vol. 87, pp. 1-19, 2020.

[18] A. S. Wicaksono and A. A. Supianto, "Hyper Parameter Optimization using Genetic Algorithm on Machine Learning Methods for Online News Popularity Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 12, pp. 263-267, 2018.

[19] P. Saqib, U. Qamar, A. Aslam and A. Ahmad, "Hybrid of Filters and Genetic Algorithm - Random Forests Based Wrapper Approach for Feature Selection and Prediction," in *Advances in Intelligent Systems and Computing*, Springer, Cham, 2019.

[20] Y. Grichi, Y. Beauregard and T. M. Dao, "Optimization of Obsolescence Forcasting Using New Hybrid Approach Based on The RF Method and The Meta-heuristic Genetic Algorithm," *American Journal of Management*, vol. 18, no. 2, pp. 27-38, 2018.

[21] J. Huang and C. X. Ling, "Using AUC and Accuracy in Evaluating Learning Algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299-310, 2013.

[22] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[23] C. Yoo, D. Han, J. Ima and B. Bechtel, "Comparison Between Convolutional Neural Networks and Random Forest for Local Climate Zone Classification in Mega Urban Areas Using Landsat Images," *Journal of Photogrammetry and Remote Sensing*, vol. 157, pp. 155-170, 2019.

[24] J. Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng and K. Li, "A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 4, pp. 919-933, 2017.

[25] E. Alfaro, M. Gamez and N. García, "An R Package for Classification with Boosting and Bagging," *Journal of Statistical Software*, vol. 54, no. 32, pp. 11-35, 2013.

[26] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap, New York: Chapman & Hall., 1993.

[27] L. Hakim, B. Sartono and A. Saefuddin, "Bagging Based Ensemble Classification Method on Imbalance Datasets," *International Journal of Computer Science and Network*, vol. 6, no. 6, pp. 670-676, 2017.

[28] A. J. Olalekan, F. Ogwueleka and P. O. Odion, "Effective and Accurate Bootstrap Aggregating (Bagging) Ensemble Algorithm Model for Prediction and Classification of Hypothyroid Disease," *International Journal of Computer Applications*, vol. 176, no. 39, pp. 40-48, 2020.

[29] K. Oliver, "Genetic Algorithms," in *Genetic Algorithm Essentials* , Springer, Cham, 2017, pp. 11-19.

[30] E. Habibi, M. Salehi, G. Yadegarfar and A. Taheri, "Optimization of ANFIS Using A Genetic Algorithm for Physical Work Rate Classification," *International Journal of Occupational Safety and Ergonomics*, vol. 26, no. 3, pp. 436-443, 2020.

[31] A. A. M. Lima, F. K. H. Barros, V. H. Yoshizumi, D. H. Spatti and M. E. Dajer, "Optimized Artifcial Neural Network for Biosignals Classifcation Using Genetic Algorithm," *Journal of Control, Automation and Electrical Systems*, vol. 30, p. 371–379, 2019.

[32] A. Malik, "A Study of Genetic Algorithm and Crossover Techniques," *International Journal of Computer Science and Mobile Computing*, vol. 8, no. 3, pp. 335-344, 2019.

[33] J. Lin, H. Chen, S. Li, Y. Liu, X. Li and B. Yu, "Accurate Prediction of Potential Druggable Proteins Based on Genetic Algorithm and Bagging-SVM Ensemble Classifier," *Artificial Intelligence In Medicine*, vol. 98, pp. 35-47, 2019.

[34] T. Shi, G. He and Y. Mu, "Random Forest Algorithm Based on Genetic Algorithm Optimization for Property-Related Crime Prediction," in *International Conference on Computer, Network, Communication and Information Systems*, Atlantis Press, 2019.

DOI: https://doi.org/10.29207/resti.v5i1.2813