SELEKSI ATRIBUT PADA ALGORITMA C4.5 MENGGUNAKAN GENETIK ALGORITMA DAN BAGGING UNTUK ANALISA KELAYAKAN PEMBERIAN KREDIT

Saeful Bahri
STMIK NUSA MANDIRI Jakarta
Jl. Damai No. 8, Warung Jati Barat Jakarta Selatan , 12540 Indonesia
Saeful.sel@nsamandiri.ac.id

Abstract

According to the banking ACT No. 9 of 1992 is the provision of credit or money bills which can dipersama-kan with it, based on the approval of an agreement between the bank pinjam-meminjam with other parties that require that the borrower to pay off a loan after a certain period of time with the giving of flowers. Credit analysis aims to evaluate the customer able to or not in fulfilling obligations. In analyzing the sometimes an analyst is not accurate in analyzing causing bad credit. Of the problems that existed then used a method of classification for an analysis of the feasibility of granting credit using a model algorithm Genetic Algorithm with C4.5 (AG) as a selection of attributes and bagging method to improve accuracy. After testing two models namely algorithm C4.5 and C4.5 with Genetic Algorithms (AG) and the results obtained bagging method is the algorithm C 4.5 produces a value accuracy 93,47% and AUC values 0,932 with excellent levels of Clasification diagnose but after Genetic Algorithm added (AG) and increased accuracy value bagging 2.87% to 96,34% and AUC values increased 0.044 became 0.976.

Keywords*: Credit, the algorithm C 4.5, Genetic Algorithms (GA), Bagging*

Abstrak

Menurut UU Perbankan No.9 Tahun 1992 kredit merupakan penyediaan uang atau tagihan yang dapat dipersama-kan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam-meminjam antara bank dengan pihak lain yang mewajibkan pihak peminjam untuk melunasi utangnya setelah jangka waktu tertentu dengan pemberian bunga. Analisa kredit bertujuan untuk mengevaluasi nasabah mampu atau tidak dalam memenuhi kewajiban. Dalam menganalisa terkadang seorang analis tidak akurat dalam menganalisa sehingga menyebabkan kredit macet. Dari permasalahan yang ada maka digunakan sebuah metode klasifikasi untuk analisis kelayakan pemberian kredit menggunakan model algoritma C4.5 dengan Algoritma Genetika (AG) sebagai seleksi atribut dan metode bagging untuk meningkatkan akurasi. Setelah dilakukan pengujian dua model yaitu algoritma C4.5 dan C4.5 dengan Algoritma Genetika (AG) dan metode bagging hasil yang diperoleh adalah algoritma C4.5 menghasilkan nilai akurasi 93,47 % dan nilai AUC 0,932 dengan tingkat diagnose excellent Clasification namun setelah ditambahkan Algoritma Genetika(AG) dan bagging nilai akurasi meningkat 2,87% menjadi 96,34 % dan nilai AUC meningkat 0.044 menjadi 0.976.

Kata kunci*: Kredit, Algoritma C4.5, Algoritma Genetika (AG), Bagging*

1. PENDAHULUAN

Kredit adalah penyediaan uang atau tagihan yang dapat dipersama-kan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam-meminjam antara bank dengan pihak lain yang mewajibkan pihak peminjam untuk melunasi utangnya setelah jangka waktu tertentu dengan pemberian bunga[1], analisa pemberian kredit dilakukan untuk mengevaluasi nasabah atau debitur berdasarkan data historis seperti pendapatan, usia, histori kredit sebelumnya, catatan kriminal dan sebagainya[2]. Pada umumnya bank sebagai pemberi kredit atau kreditor melakukan proses pemberian pembiayaan secara garis besar yaitu pengajuan pembiayaan, analisis usulan pembiayaan, persetujuan pihak terkait, perjanjian kredit, dan proses pencairan dana. Resiko kredit merupakan isu yang paling penting dalam dunia industri perbankan[3] karena akan merugikan terhadap kelangsungan keuangan suatu negara dan berpotensi menimbulkan kesulitan keuangan[4] untuk mengurangi resiko kredit maka analisa kredit menjadi kunci utama dalam manajemen resiko kredit [5].

Seleksi atribut pada penelitian tentang analisa pemberian kredit yang telah dilakukan diantaranya penelitian tentang memodelkan resiko kredit dengan menggunakan Bayesian Additive Classification Tree[6] perbandingan beberapa algoritma klasifikasi salah satunya adalah Decision tree atau C4.5 untuk melakukan klasifikasi dalam manajemen resiko hasilnya Decision tree memiliki tingkat akurasi paling tinggi dibanding algoritmaalgoritma yang lain[7], keuntungan pengklasifikasian menggunakan pohon keputusan memiliki kelebihan dalam memecahkan struktur kompleks menjadi struktur yang lebih sederhana sehingga lebih mudah untuk diimplementasikan[8][9]. pohon keputusan memiliki kelemahan dalam menangani data yang besar dan ketidakseimbangan data yang disebabkan oleh banyaknya atribut pada sebuah dataset[10], muncul noise data ketika salah pelabelan[9]. Untuk menangani beberapa kelemahan yang masih ada maka akan diterapkan algoritma pohon keputusan berbasis Algoritma Genetika (AG) yang akan diterapkan untuk pemilihan atribut dan bagging akan diterapkan untuk menanggulangi data noise yang dihasikan dari proses pengklasifikasian menggunakan decision tree untuk meningkatkan akurasi hasil analisa kelayakan pemberian kredit

2. METODOLOGI PENELITIAN

Dalam penelitian ada empat metode umum yang digunakan diantaranya *Action Research Experiment*, *Case Study* dan *Survey*[11], metode penelitian dalam penelitian ini menggunakan metode penelitian *experiment,*penelitian jenis ini terdiri dari :

a. Mendefinisikan hipotesis teoritis
b. Memilih sampel dari populasi yang diketahui
c. Mengalokasikan sampel untuk kondisi percobaan yang berbeda
d. Memperkenalkan perubahan yang direncanakan untuk satu atau lebih *variable.*
e. Mengukur sejumlah kecil variable
f. Mengontrol semua Variabel.

Dalam metode penelitian eksperimen, digunakan model proses CRISP-DM (*Cross-Standard Industry Process for Data Mining*) yang terdiri dari 6 tahapan[12]

a. Tahap *Business Understanding*

Data set yang digunakan pada penelitian ini ialah data sekunder dengan jumlah data sebanyak 766 record, terdiri dari 16 variabel atau atribut dan 1 class yang benilai MACET atau LANCAR. Atribut yang digunakan sebagai prediktor oleh peneliti terdahulu ada 15 atribut termasuk class diantaranya nama nasabah, jenis kelamin, umur, jumlah pinjaman, jangka waktu, jumlah angsuran perbulan, tipe pinjaman, jenis pinjaman, bi sector ekonomi, col bi golongan debitur, bi golongan penjamin,saldo nominatif, plafon teoritis, tunggakan pokok, dan tunggakan bunga dan untuk class atau tujuan adalah lancar dan macet, sedangkan pada penelitian ini atribut yang digunakan yaitu nama nasabah, jenis kelamin, Rate, plafon pinjaman, jangka waktu, jml ansuran per bulan, ln_type, main branch, no rek, region, branch, cif no, sisa angsuran, tunggakan pokok, tunggakan bunga.

b. Tahap Data Understanding

Data merupakan data sekunder yang didapat dari hasil riset, atribut atau variabel yang ada sebanyak 15 Variabel-variabel tersebut ada yang tergolong variabel prediktor atau pemrediksi (predictor variabel) yaitu variabel yang dijadikan dasar sebagai penentu lancar atau macet status dari nasabah yang bersangkutan, dan variabel tujuan yaitu variabel yang dijadikan sebagai MACET atau LANCAR. Variabel prediktor yaitu digunakan yaitu nama nasabah, jenis kelamin, Rate, plafon pinjaman, jangka waktu, jml ansuran per bulan, ln_type, main branch, no rek, region, branch, cif no, sisa angsuran, tunggakan pokok, tunggakan bunga.

c. Tahap Data Preparation

Data pada penelitian ini berjumlah 766 yang kemudian dibagi kedalam 10 set menjadi masing-masing set 76 tupel, dengan rincian 9 set untuk *data training* dan 1 *set* untuk *data testing1*, proses berulang hingga 10 kali *iterasi* sehingga dari sebagai langkah persiapan penelitian untuk mendapatkan *dataset* yang berkualitas tinggi, terdapat beberapa teknik yang dapat dilakukan digunakan dalam analisis data mining diantaranya adalah:

1) *Data Cleaning* untuk membersihkan nilai yang kosong atau tupel yang kosong.
2) *Data Integration* yang berfungsi menyatukan tempat peyimpanan yang berbeda kedalam satu data. Dalam kasus ini data yang diambil dari dari sistem informasi debitur, di satukan dalam sebuah file dengan format excel.
3) *Data reduction* jumlah atribut yang ada pada data nasabah sebanyak 31 atribut kemudian direduksi menjadi sekitar 15 atribut yang berpengaruh langsung terhadap pengambilan keputusan dalam analisa pemberian kredit Berikut atribut hasil *data reduction* digunakan yaitu nama nasabah, jenis kelamin*, Rate,* plafon pinjaman, jangka waktu, jml ansuran per bulan, *ln_type*, *main branch*, no rek, region, branch, *cif no*, sisa angsuran, tunggakan pokok, tunggakan bunga untuk kemudian ditentukan pembuatan kandidat pohon, penentuan kandidat pohon dilakukan dengan cara memasukan seluruh atribut yang untuk kemudian dilakukan penilaian pada atribut-atribut sehingga menghasilkan atribut yang mempengaruhi dalam klasifikasi kemudian di tentukan pohon.

Tabel 1. Candidat splite dan rule atribut algoritma C4.5

Candidat split	*Child Node*	
1	Tunggakan Pokok	Tunggakan Pokok
	≤ 84199.865	> 84199.865
	≤ 166833.330	> 166833.330
	≤ 313750.005	> 313750.005
	≤ 236166.645	> 236166.645
	≤ 22670.550	> 22670.550
	≤ 6722	> 6722
2	Plafon Pinjaman	Plafon Pinjamann
	≤ 1831667	> 1831667
	≤ 780000	> 780000
	≤ 960000	> 960000
	≤ 441875	> 441875
3	LN *Type* =H5	
	LN *type* = HA	
	LN *Type* =H5	
	LN *type* = HA	
	LN *type* = HI	
	LN *type* = HU	
	LN *type* = HY	
	LN *type* = KB	
	LN *type* = KJ	
	LN *type* = LI	
4.	No Rekening	No Rekening
	≤ 441301005318601.500	>441301005318601.500
5.	Jml angsuran per bulan	Jml angsuran per bulan
	≤ 221250.115	> 221250.115
	≤ 44267.690	> 44267.690
	≤ 40231.260	> 40231.260
	≤ 15238.030	> 15238.030
6.	Tunggakan Bunga	Tunggakan Bunga
	≤ 1756	> 1756
	≤ 2500	> 2500

Tabel 2. Candidat splite dan rule atribut algoritma C4.5 dengan GA dan Bagging

Candidat split	*Child Node*	
1	Tunggakan Pokok	Tunggakan Pokok
	≤ 7055.330	> 7055.330
	≤ 166833.330	> 166833.330
	≤ 72916.665	> 72916.665
2.	Plafon Pinjaman	Plafon Pinjaman
	≤ 925208.380	> 925208.380

Candidat split	*Child Node*	
	≤ 14692499.500	> 14692499.500
3.	Jml angsuran per bulan ≤ 12750 ≤ 280000 ≤ 1311416.475	Jml angsuran per bulan > 12750 > 280000 > 1311416.475
4.	Tunggakan Bunga ≤ 1756 ≤ 2500	Tunggakan Bunga > 1756 > 2500

4) Tahap Meodeling

Tahap *modelling* dilakukan untuk menerapkan teknik yang tepat guna mendapatkan hasil yang optimal dalam analisis kelayakan pemberian kredit. Pada penelitian ini model yang digunakan yaitu algoritma terpilih pohon keputusan C4.5 dan algoritma pohon keputusan C4.5 dengan Algoritma Genetika (AG) berbasis *bagging* sebagai penyeleksi atribut. Tahap *modelling* dilakukan dengan dua cara yaitu cara manual dan dengan menggunakan *software rapid miner*. Perhitungan manual untuk pembuatan model dengan menggunakan algoritma pohon keputusan C4.5 dilakukan dengan cara mencari nilai *gain* tertinggi dari setiap atribut, sedangkan perhitungan manual untuk model algoritma pohon keputusan C4.5.

2.1. Model Klasifikasi C4.5

Untuk dapat membuat pohon keputusan, langkah pertama adalah menghitung jumlah *class* yang terpilih dan tidak dari masing-masing *class* berdasarkan atribut yang telah ditentukan dengan menggunakan data *training*.

a. Kemudian menghitung *Entropy* (Total) menggunakan persamaan

$$Entropy(total) = (-\frac{210}{766} * \log_2(\frac{210}{766})) + (-\frac{556}{766} * \log_2(\frac{556}{766}))$$

$$Entropy(Total) = 0.00291$$

b. Kemudian hitung masing-masing Gain berdasarkan atribut pada tabel 3.4
Sebagai contoh plafon pinjman :

≤ 925208.380 = $\frac{200}{766}$

> 925208.380= $\frac{356}{766}$

Untuk record plafond pinjaman $\leq$ 925208.380 terdiri dari 10 LANCAR dan 200 MACET, untuk plafond pinjaman > 925208.380 terdiri dari 200 LANCAR dan 356 MACET. Dapat dihitung *Entropy*nya sebagai berikut:

$$Entropy < 925208.380 = (-\frac{10}{788} * \log_2(\frac{10}{766})) + (-\frac{200}{766} * \log_2(\frac{200}{766}))$$

$$Entropy < 925208.380 = 0.27620$$

$$Entropy > 925208.380 = (-\frac{200}{766} * \log_2(\frac{200}{766})) + (-\frac{356}{766} * \log_2(\frac{356}{766}))$$

$$Entropy > 925208.380 = 0.94244$$

Selengkapnya akan disajikan dalam tabel contoh perhitungan *entropy* pada tabel dibawah ini :

Tabel. 3. Perhitungan Entropy dan Gain Algoritma C4.5

Node	Atribut		Jml Kasus (S)	Macet (Si)	Lancar (Si)	Entropy	Gain
1	Total		766	556	210	0.84736	
	Jenis Kelamin						0.00291
		Laki-Laki	304	210	94	0.89226	
		Perempuan	462	346	116	0.81299	
	Rate						0.43296
		10	4	3	1	0.81128	
		11	9	7	2	0.76420	
		12	49	24	25	0.99970	
		14	67	43	24	0.94119	
		19	228	164	64	0.85641	
		20	1	0	1	0.00000	
		24	1	0	1	0.00000	
	Plafon Pinjaman						0.08757
		≤ 925208.380	210	200	10	0.27620	
		>925208.380	556	356	200	0.94244	
	Jangka Waktu						0.07560
		1	10	6	4	0.97095	
		2	8	5	3	0.95443	
		3	24	14	10	0.97987	
		4	24	19	5	0.73828	
		5	47	29	18	0.96012	
		6	47	36	11	0.78499	
		7	22	21	1	0.26676	
		8	50	39	11	0.76017	
		10	84	70	14	0.65002	
		11	8	7	1	0.54356	
		12	127	84	43	0.92346	
		14	4	3	1	0.81128	
		15	7	7	0	0.00000	
		16	76	45	31	0.97538	
		20	73	50	23	0.89894	
		24	58	37	21	0.94439	
		25	2	1	1	1.00000	
		30	3	2	1	0.91830	
		32	3	2	1	0.91830	
		36	1	0	1	0.00000	
		38	1	1	0	0.00000	

2.2. Model Klasifikasi C4.5 dengan Genetik Algorima dan bagging

Proses yang dilakukan sama dengan proses pembuatan model klasifikasi C4.5 namum perbedaannya dari atribut yang digunakan, atribut yang digunakan merupakan atribut terpilih hasil seleksi atribut terbaik menggunakan Algortima Genetika, untuk meningkatkan stabilitas dalam klasifikasi maka bagging diterpakan selain berguna untuk meningkatkan stabilitas klasifikasi, *bagging* juga berguna untuk mengurangi variansi data dan untuk menghindari *overfitting data*, untuk penjelasanya sebagai berikut

Menghitung *Entropy* (Total) menggunakan persamaan.

$$Entropy(total) = (-\frac{210}{766} * \log_2(\frac{210}{766})) + (-\frac{556}{766} * \log_2(\frac{556}{766}))$$

$$Entropy(Total) = 0.00291$$

Untuk mendapatkan *gain* tiap atribut, maka harus dihitung *entropy* seperti pada algoritma C4.5 sebelumnya diatas namun perbedaannya atribut yang di hitung hanya tunggakan pokok, plafon pinjaman, jumlah angsuran, tunggakan pokok, tunggakan bunga.

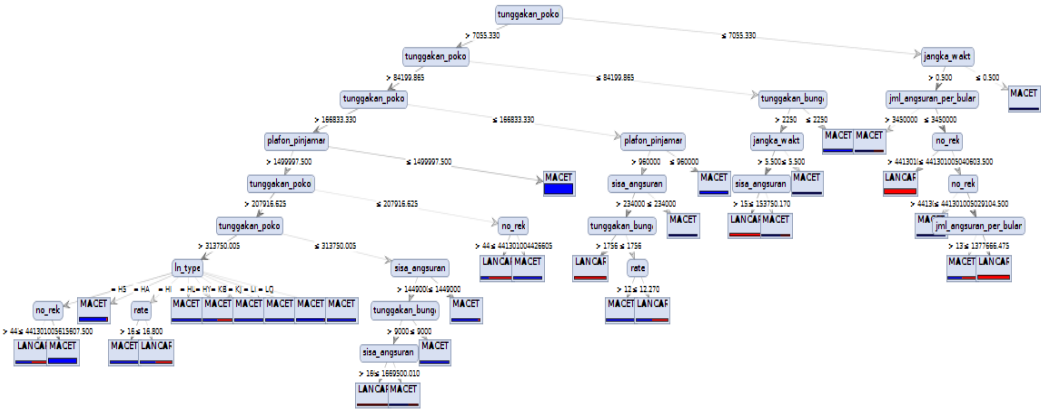
3. HASIL DAN PEMBAHASAN

Selanjutnya melakukan pengukuran dari model yang diajukan, pengukuran dilakukan dengan menggunakan tools, salah satu tools yang digunakan adalah rapid miner.

3.1. Hasil Penelitian

Membandingkan model terpilih C4.5 dengan algortima C4.5 yang dioptimalkan dalam pemilihan atribut dengan Algoritma genetika dan *bagging*. Data akan dianalisa kemudian model terbaik yang akan digunakan dalam perancangan GUI.

Pemodelan menggunakan algoritma C4.5 akan menghasilkan model berupa pohon keputusan pada gambar 1. merupakan model pohon keputusan yang dihasilkan oleh algoritma C4.5

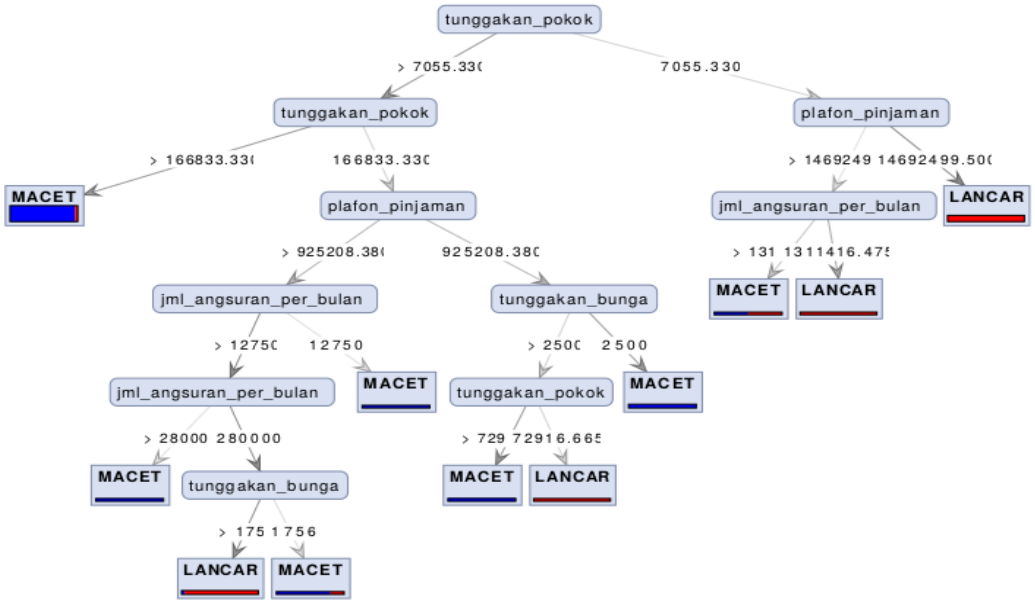


Gambar 1. Model Pohon Keputusan Algoritma C4.5

Tabel 4. Nilai Accuracy, Sensitivity, Specificity, PPV *dan* NPV Algoritma C4.5

	Nilai
Accuracy	0,934
Sensitivity	0,943
Specificity	0,908
PPV	0,967
NPV	0,847

Pemodelan menggunakan algoritma C4.5 menggunakan Algoritma Genetika (AG) dan Bagging akan menghasilkan model berupa pohon keputusan yang akan dijadikan sebagai acuan dalam pengembangan aplikasi pada gambar 2. merupakan model pohon keputusan yang dihasilkan oleh algoritma C4.5 dengan Algoritma Genetika dan Bagging



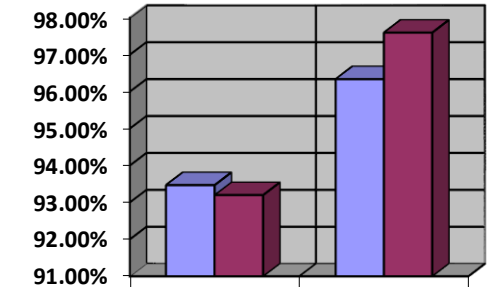
Gambar 2. Model Pohon Keputusan dengan Algoritma Genetika dan Bagging

Tabel 5. Nilai Accuracy, Sensitivity, Specificity, PPV *dan* NPV

	Nilai
Accuracy	0,963
Sensitivity	0,964
Specificity	0,959
PPV	0,964
NPV	0,904

3.2. Analisis dan Validasi Model

Berdasarkan hasil pengujian menggunakan *confusion matrix* maupun ROC *curve* diatas terbukti bahwa, algoritma C4.5 dengan Algorima Genetika (AG) dan *Bagging* mampu meningkatkan akurasi hasil klasifikasi C4.5. sedangkan nilai akurasi untuk C4.5 itu sendiri adalah sebesar 93,47% dan nilai akurasi C4.5 dengan Algoritma Genetika dan *Bagging* adalah sebesar 96,34 % dengan selisih akurasi 2,87% dapat dilihat pada gambar dibawah ini:



	C4.5	C4.5+GA+ BAGGING
Accuracy	93.47%	96.34%
AUC	0.932	0.976

Gambar 3. Hasil Perbandingan

Untuk evaluasi menggunakan ROC *curve* sehingga menghasilkan nilai AUC(*Area Under Curve*) untuk model algoritma klasifikasi C4.5 mengasilkan nilai 0.932 dengan nilai diagnosa *Excelent Clasificatian*, sedangkan untuk algoritma klasifikasi C4.5 dengan Algoritma Genetika dan Bagging menghasilkan nilai 0.976 dengan nilai diagnosa *Excelent Classification*, dan selisih nilai keduanya sebesar 0.044

4. SIMPULAN

Dari hasil penelitian untuk akurasi algoritma klasifikasi C4.5 sebesar 93,47%, sedangkan untuk akurasi algoritma klasifikasi dengan GA dan *Bagging* sebesar 96,36%, sehingga didapat selisih peningkatan akurasi sebesar 2,87%. Hasil evaluasi keduanya menggunakan Curva ROC yaitu, algoritma klasifikasi C4.5 bernilai 0,932 dengan tingkat diagnosa *excellent clasification*, sedangkan untuk algoritma klasifikasi C4.5 dengan Algoritma Genetika dan *bagging* senilai 0,976 dengan tingkat diagnosa *excellent clasification* maka didapatlah selisih nilai sekitar nilai AUC 0,044.

Dapat disimpukan bahwa penggunaan algoritma genetika dan *bagging* pada algoritma klasifikasi C4.5 dapat meningkatkan akurasi pada algortima klasifikasi dengan C4.5.

DAFTAR PUSTAKA

[1] R. Indonesia, *UU Perbankan No. 9 1992*. 1995, pp. 1–20.

[2] O. Akbilgic and H. Bozdogan, "A new supervised classification of credit approval data via the hybridized RBF neural network model using information complexity," in *Studies in Classification, Data Analysis, and Knowledge Organization*, vol. 48, 2015, pp. 13–27.

[3] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert Syst. Appl.*, vol. 41, no. 4 PART 2, pp. 2052–2064, 2014.

[4] J. Zurada, "Could decision trees improve the classification accuracy and interpretability of loan granting decisions?," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2010.

[5] A. S. U. Refailzadeh Payam, Lei Tang, Huan Liu, "Cross-Validation."

[6] J. L. Zhang and W. K. Härdle, "The Bayesian Additive Classification Tree applied to credit risk modelling," *Comput. Stat. Data Anal.*, vol. 54, no. 5, pp.

1197–1205, 2010.
[7] L. Yu, G. Chen, A. Koronios, S. Zhu, and X. Guo, "Application and Comparison of Classification Techniques in Controlling Credit Risk," *World*, pp. 2007–2007.
[8] C. Jun, Y. Cho, and H. Lee, "Improving Tree-based Classification Rules Using a Particle Swarm Optimization," 2017.
[9] J. Abellán and A. R. Masegosa, "Bagging schemes on the presence of class noise in classification," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 6827–6837, 2012.
[10] B. K. Sarkar, S. S. Sana, and K. Chaudhuri, "Selecting informative rules with parallel genetic algorithm in classification problem," *Appl. Math. Comput.*, vol. 218, no. 7, 2011.
[11] C. W. Dawson, "Project in computing and information system," 2009.
[12] T. D. Larose, "Discovering Knowledge in Data an Introduction to Data Mining," 2005.