

Peningkatan Akurasi Metode C4.5 Untuk Memprediksi Kelayakan Kredit Berbasis Stratified Sampling Dan Optimize Selection

Ardiyansyah¹, Rabiatus Sa'adah², Lisnawanty³, Deasy Purwaningtias⁴

Universitas Bina Sarana Informatika Kampus Pontianak

Jl. Abdul Rahman Saleh No.18, Bangka Belitung Laut, Kec. Pontianak Tenggara, Kota Pontianak, Kalimantan Barat 78124

1ardiyansyah.arq@bsi.ac.id, 2rabiatus.rbh@bsi.ac.id, 3lisnawanty.lsy@bsi.ac.id,

4deasy.dwg@bsi.ac.id

Abstract

Credit financing is a provider of funds that provides loans based on an agreement between the customer and the bank with the condition that the borrower must pay off the loan at a certain time. The purpose of this research is to improve the accuracy of the C4.5 algorithm in predicting credit worthiness. The stratified sampling algorithm is used to deal with large amounts of data, while the optimize selection algorithm is used to determine the best attribute. The application of stratified sampling and optimize selection algorithms in C4.5 has an accuracy rate of 80%, higher than other algorithms. So it can be concluded that in this study the stratified sampling and optimize selection algorithms are the best algorithms in increasing accuracy and overcoming the shortcomings of the C4.5 algorithm.

Keywords: Data Mining, Decision Tree, Stratified Sampling, Optimize Selection

Abstrak

Pembiayaan kredit merupakan penyedia dana yang memberikan pinjaman yang berdasarkan perjanjian antara nasabah dan bank dengan syarat peminjam harus melunasi pinjamannya pada waktu tertentu. Tujuan dari penelitian ini dilakukan adalah untuk meningkatkan akurasi algoritma C4.5 dalam memprediksi kelayakan kredit. Algoritma stratified sampling digunakan untuk mengatasi jumlah data yang besar, sedangkan algoritma optimize selection digunakan untuk menentukan atribut terbaik. Penerapan algoritma stratified sampling dan optimize selection pada C4.5 memiliki tingkat akurasi sebesar 80%, lebih tinggi dari algoritma yang lainnya. Sehingga dapat disimpulkan bahwa pada penelitian ini algoritma stratified sampling dan optimize selection merupakan algoritma terbaik dalam meningkatkan akurasi serta mengatasi kekurangan algoritma C4.5.

Kata kunci: Data Mining, Decision Tree, Stratified Sampling, Optimize Selection

1. PENDAHULUAN

Pembiayaan kredit merupakan penyedia dana yang memberikan pinjaman yang berdasarkan perjanjian antara nasabah dan bank dengan syarat peminjam harus melunasi pinjamannya pada waktu tertentu [1]. Dalam memilih calon nasabah, setiap marketing melakukan beberapa analisa terhadap calon nasabah seperti kepercayaan, tingkat risiko, objek kredit dan jangka waktu [1]. Hal tersebut dilakukan agar mengurangi resiko terjadinya keterlambatan atau penundaan dalam pembayaran pinjaman yang sudah disepakati. Namun, cara yang dilakukan masih menjadi masalah seperti banyak kredit yang bermasalah atau kredit macet [2].

Data mining ialah proses pengumpulan dan pengolahan data agar dapat diidentifikasi menjadi sebuah informasi dan keputusan [3]. Dengan adanya data mining, masalah yang terjadi di berbagai bidang khususnya di bidang finansial dapat diatasi dengan penerapan data mining.

Berbagai macam algoritma data mining yang pernah digunakan oleh para peneliti dalam melakukan penelitian adalah algoritma C4.5 [2], support vector machine [4], artificial neural network [5], random forest [6], cart dan naïve bayes [7]. Salah satu algoritma yang populer di kalangan peneliti yaitu algoritma C4.5.

Algoritma C4.5 merupakan algoritma yang membentuk pohon keputusan berdasarkan kriteria pembentuk keputusan [8]. Algoritma ini banyak digunakan peneliti untuk mengklasifikasi kelayakan kredit. Sehingga algoritma yang di pilih dalam penelitian ini adalah algoritma C4.5.

Kelebihan dari algoritma C4.5 ini adalah dapat mengolah data numeric dan diskret serta mampu menangani nilai atribut yang hilang [9], selain itu dapat mengatasi kerentanan atribut non-informatif dan pengukuran akurasi yang mengandalkan satu probabilitas saja dengan membuat model pohon keputusan [2]. Namun algoritma C4.5 juga memiliki kekurangan seperti dapat menimbulkan noisy-data, atribut yang kurang relevan dan mengakibatkan *over fitting* [10], serta lemah terhadap atribut dan jumlah data yang besar [2].

Berdasarkan penelitian [11] [2] metode *stratified sampling* dapat mengatasi jumlah data yang besar karena penarikan sampel acak terstruktur dilakukan dengan membagi anggota populasi dalam beberapa sub kelompok [12]. Sedangkan metode *optimize selection* digunakan untuk mencari dan memilih nilai atribut yang memiliki nilai terbaik dari seluruh subset [13][14]. Pada penelitian ini, metode *stratified sampling* dan *optimize selection* akan di terapkan pada algoritma C4.5 untuk meningkatkan akurasi dalam memprediksi kelayakan kredit.

2. METODOLOGI PENELITIAN

Pada penelitian ini berfokus pada memprediksi kelayakan kredit dengan menggunakan algoritma C4.5, kemudian akan dilakukan peningkatan akurasi

dengan menerapkan *stratified sampling* dan *optimize selection* pada C4.5. Metode penelitian yang digunakan adalah eksperimen. Penelitian eksperimen yaitu penelitian yang dilakukan dengan uji coba yang di control sendiri dengan melibatkan sebab akibat [15]. Tahapan penelitian yang dilakukan dalam penelitian ini, terdiri dari [15]:

A. Data Gathering

Pengumpulan data dalam penelitian ini menggunakan dataset *public* yang diperoleh dari situs UCI *Machine Learning Repository* <https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data>.

Dataset ini memiliki 21 atribut dengan 1 atribut sebagai class. Dataset ini juga memiliki 1000 record yang terdiri dari 700 record data sebagai status kredit lancar dan 300 record data sebagai status kredit macet.

B. Data Pre-Processing

Pada *preprocessing* data ini, metode *stratified sampling* dan *optimize selection* digunakan dalam tahap ini untuk mengatasi jumlah data yang besar dan menentukan atribut yang terbaik.

C. Proposed Method

Metode yang digunakan pada tahap ini adalah algoritma C4.5, kemudian algoritma C4.5 yang di kombinasikan dengan metode *stratified sampling* dan *optimize selection*.

D. Result Evaluation and Validation

Pada tahap terakhir ini, akan di lakukan pengujian model dengan membandingkan antara algoritma C4.5, C4.5+*stratified sampling+optimize selection*, C4.5+*stratified sampling+feature selection*, C4.5+ *stratified sampling+backward elimination*. Data akan divalidasi dengan *k-fold cross validation* yang dibagi menjadi 90% data training dan 10% data testing. Sedangkan pengukuran kinerja algoritma dilakukan menggunakan *confusion matrix* dan AUC.

2.1. Algoritma C4.5

Algoritma C4.5 adalah algoritma yang digunakan untuk membentuk pohon keputusan berdasarkan kriteria pembentuk keputusan[8]. Algoritma C4.5 merupakan pohon keputusan yang paling populer[16]. Ada beberapa tahapan dalam algoritma C4.5, diantaranya:

- a. menyiapkan data training
- b. menentukan akar yang diambil dari atribut yang terpilih. Pemilihan akar berdasarkan nilai gain yang paling tinggi diantara atribut yang lainnya. Sebelum mendapatkan nilai gain, nilai entropy harus di dapatkan terlebih dahulu dengan rumus sebagai berikut:

$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2 p_i$$

Keterangan:

S : himpunan kasus

A : atribut

n : jumlah partisi S
pi : proporsi dari Si terhadap S

- c. Kemudian cari nilai gain dengan rumus sebagai berikut:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i)$$

Keterangan:

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

- d. Mencari nilai *Split Info* dengan rumus berikut:

$$SplitInfo_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right)$$

Keterangan:

Dj = jumlah kasus pada partisi ke-i

D = jumlah kasus dalam D

- e. Mencari nilai *Gain Ratio* dengan rumus berikut:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

- f. Setelah mendapatkan nilai akar, maka buat cabang berdasarkan nilai atribut akar tersebut.
g. Ulangi semua proses hingga membentuk pohon keputusan.

2.2. Stratified Sampling

Pengambilan sampel dapat digunakan untuk mereduksi data. Populasi data yang besar dapat diwakili oleh sampel acak yang ukurannya jauh lebih kecil. Ada beberapa teknik pengambilan sampel, salah satunya Stratified Sampling. Stratified Sampling merupakan teknik pengambilan sampel secara acak dengan memperhatikan tingkatan (strata) pada populasi. Dataset dibagi menjadi beberapa bagian yang terpisah (strata), kemudian sampel diambil secara acak berdasarkan strata yang sudah dibuat [2]. Adapun tahapan Stratified Sampling sebagai berikut: 1. Pertama, populasi N dibagi menjadi beberapa sub-sub populasi yang masing-masing sub-sub populasi tersebut terdiri dari elemen $N_1, N_2, N_3, \dots, N_L$. 2. Kemudian, diantara sub-sub populasi, tidak boleh ada tumpang tindih, sehingga $N_1 + N_2 + N_3 + \dots + N_L = N$. 3. Terakhir, ambil sampel secara acak dari masing-masing sub populasi dengan alokasi sampel yang proporsional. Sebelum pengambilan sampel, menentukan ukuran sampel merupakan hal yang penting. Sampel yang diambil harus mencerminkan populasi. Ada beberapa cara dalam

menentukan ukuran sampel. Salah satunya yang paling banyak dan umum menggunakan teori slovin dijelaskan dengan rumus sebagai berikut:

$$n = \frac{N}{1 + Ne^2} \quad [2]$$

Keterangan:

n = besar sampel

N = ukuran populasi atau jumlah elemen dalam populasi

e = nilai presisi atau tingkat signifikansi yang telah ditentukan.

3. HASIL DAN PEMBAHASAN

Pada bagian ini, pembahasan hasil pengujian menggunakan Rapid Miner Studio Versi 10. Kemudian dataset yang digunakan merupakan dataset public yang diperoleh dari situs *UCI Machine Learning Repository* yaitu *German Credit Data*. *Preprocessing* data menggunakan *cross validation* dengan pembagian 9:1.

3.1. Hasil Pengujian Algoritma C4.5

Pada tahap pertama, dataset hanya menggunakan *cross validation*, tidak dilakukan penerapan metode *stratified sampling* dan *optimize selection* serta langsung melakukan pengujian menggunakan metode C4.5. Hasil dapat dilihat pada tabel 1 dibawah ini:

Tabel 1. Hasil Pengujian Algoritma C4.5 Tanpa *stratified sampling* + *optimize selection*

Algoritma	Precision	Recall	Accuracy
C4.5	46.77	31.33	68.7

Pada tabel 1 di atas menunjukkan bahwa nilai yang didapatkan dengan hanya menggunakan algoritma C4.5 tanpa menerapkan metode *stratified sampling* dan *optimize selection* yaitu nilai precision sebesar 46.77%, nilai Recall sebesar 31.33% dan nilai akurasi sebesar 68.70%.

3.2. Hasil Pengujian C4.5+*Stratified Sampling*+*Optimize Selection*

Pada tahap ini, pengujian dilakukan dengan algoritma C4.5 dengan menerapkan metode *stratified sampling*+*optimize selection*. Hasil pengukuran yang dilakukan menggunakan rapid miner studio, hasil nilai *Area Under ROC Curve* (AUC) yang dihasilkan sebesar 0,686. Nilai AUC dapat dilihat berdasarkan grafik kurva ROC di bawah ini:



Gambar 1. Kurva ROC C4.5+ *Stratified Sampling* + *Optimize Selection*

Kemudian dibuat pula tabel perbandingan antara algoritma C4.5 dengan algoritma C4.5 yang sudah menerapkan metode *stratified sampling* + *optimize selection*. Tabel ini untuk menampilkan seberapa besar peningkatan nilai yang dihasilkan. Nilai perbandingan tersebut dapat dilihat pada tabel 2 dibawah ini:

Tabel 2. Hasil Pengujian Algoritma C4.5 + *stratified sampling* + *optimize selection*

Algoritma	Precision	Recall	Accuracy
C4.5	46.77	31.33	68.7
C4.5+SS+OS	70.83	56.67	80

Pada tabel 2 diatas, menunjukkan bahwa precision memiliki nilai sebesar 70.83%, recall memiliki nilai sebesar 56.67% dan akurasi memiliki nilai sebesar 80%. Sehingga dapat di artikan bahwa dengan menerapkan metode *stratified sampling* serta menggunakan metode *optimize selection* dapat meningkatkan akurasi algoritma C4.5 sebesar 11.3%. Selain itu, nilai precision dan recall juga terjadi peningkatan dari algoritma C4.5 tanpa optimasi.

3.3. Perbandingan Metode *Feature Selection*

Pada bagian ini, perbandingan tidak hanya dilakukan antara algoritma C4.5 tanpa optimasi dengan C4.5+SS+OS. Tetapi akan dilakukan juga

perbandingan algoritma C4.5 yang diterapkan menggunakan metode optimasi lainnya seperti *stratified sampling+forward selection* dan *stratified sampling+backward elimination*. Pada perbandingan ini, akan ditampilkan hasil dari pengujian antara *stratified sampling + optimize selection*, *stratified sampling+forward selection* dan *stratified sampling+backward elimination*. Perbandingan ini untuk menampilkan mana model yang terbaik berdasarkan nilai akurasi yang didapatkan.

Berikut adalah hasil pengujian C4.5+*stratified sampling+forward selection* yang dapat di lihat pada tabel 3 di bawah ini:

Tabel 3. hasil pengujian C4.5+*stratified sampling+forward selection*

Algoritma	Precision	Recall	Accuracy
C4.5+SS+FS	70.83	56.67	79.11

Berdasarkan tabel 3 di atas, nilai precision yang didapatkan sebesar 70.83%, nilai recall sebesar 56.67% dan nilai accuracy sebesar 79.11%.

Berikut adalah hasil pengujian C4.5+*stratified sampling+ backward elimination* yang dapat di lihat pada tabel 4 di bawah ini:

Tabel 4. C4.5+*stratified sampling+ backward elimination*

Algoritma	Precision	Recall	Accuracy
DT+SS+BE	57.14	53.33	74

Berdasarkan tabel 4 di atas, menyatakan bahwa nilai precision yang didapatkan sebesar 57.14%, nilai recall sebesar 53.33% dan nilai accuracy sebesar 74%.

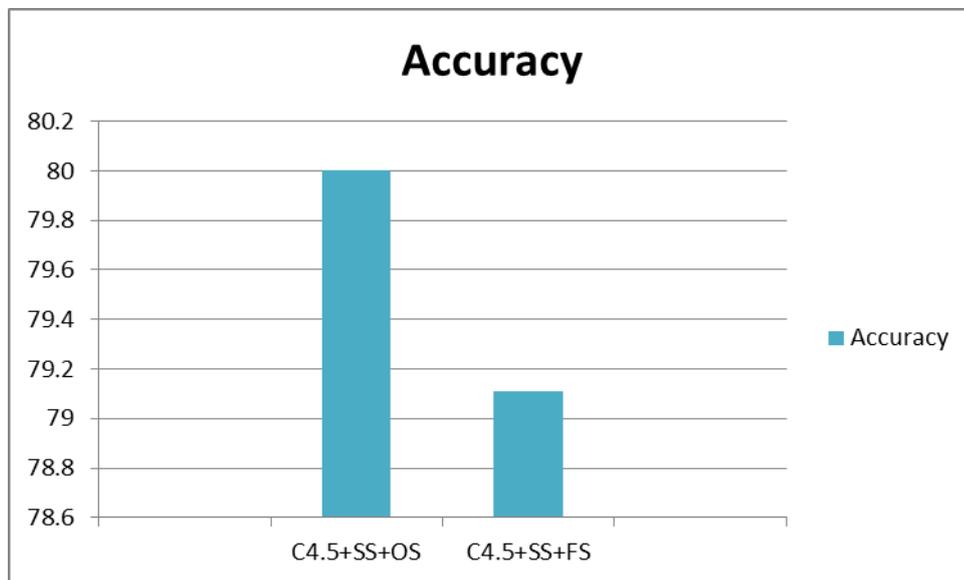
Dari hasil pengukuran yang sudah dilakukan, perbandingan nilai akurasi algoritma C4.5 setelah dilakukan optimasi menggunakan metode *feature selection* yaitu *optimize selection*, *forward selection* dan *backward elimination* akan ditampilkan dalam bentuk tabel dan grafik perbandingan berdasarkan nilai akurasi. Berikut tabel perbandingan nilai yang dapat di lihat pada tabel dibawah ini:

Tabel 5. Perbandingan Algoritma Setelah Optimasi

Algoritma	Precision	Recall
C4.5	46.77	31.33
C4.5+SS+OS	70.83	56.67
C4.5+SS+FS	70.83	56.67
C4.5+SS+BE	57.14	53.33

Berdasarkan tabel diatas, menunjukkan bahwa algoritma C4.5 memiliki nilai precision dan recall paling rendah dibandingkan dengan algoritma C4.5 yang sudah di optimasi. Pada algoritma C4.5+SS+BE memiliki nilai terendah kedua dari algoritma yang lainnya. Tetapi tetap mengalami peningkatan dari algoritma C4.5 tanpa optimasi yaitu nilai precision memiliki peningkatan sebesar 10.37% dan nilai recall memiliki peningkatan sebesar 22%. Sedangkan pada algoritma C4.5+SS+OS dan C4.5+SS+FS memiliki nilai precision dan recall sama besar.

Berdasarkan hasil diatas, maka akan ditampilkan grafik perbandingan berdasarkan nilai akurasi dari kedua algoritma tersebut. Grafik tersebut dapat di lihat pada gambar dibawah ini:



Gambar 2. Grafik Nilai Akurasi

Berdasarkan grafik diatas, diketahui bahwa nilai akurasi dari setiap algoritma berbeda-beda. Nilai akurasi terbaik pada penelitian ini terdapat pada algoritma C4.5+SS+OS yaitu sebesar 80%, nilai akurasi terbaik kedua terdapat pada algoritma C4.5+SS+FS yaitu sebesar 79.11%. Algoritma C4.5+SS+OS memiliki nilai akurasi sedikit lebih tinggi dibandingkan algoritma C4.5+SS+FS yaitu sebesar 0.89%.

Berikut hasil perbandingan penerapan metode sampling dan feature selection yang digunakan untuk melakukan peningkatan akurasi algoritma C4.5 yang disajikan pada tabel dibawah ini:

Tabel 6. Hasil Pengukuran Metode Sampling Dan Feature Selection Pada Algoritma C4.5

Algoritma	Precision	Recall	Accuracy
C4.5	46.77	31.33	68.7

C4.5+SS+OS	70.83	56.67	80
C4.5+SS+FS	70.83	56.67	79.11
C4.5+SS+BE	57.14	53.33	74

4. SIMPULAN

Berdasarkan pembahasan yang dijelaskan diatas, metode sampling yaitu *stratified sampling* digunakan untuk mengatasi jumlah data yang besar. Selain itu, beberapa algoritma *feature selection* juga dilakukan pengujian dan digunakan untuk mengatasi kelemahan dari algoritma C4.5 serta digunakan untuk melakukan peningkatan akurasi algoritma C4.5, algoritma tersebut seperti *optimize selection*, *forward selection* dan *backward elimination*. Berdasarkan hasil pengujian, algoritma C4.5 yang di optimasi menggunakan metode *stratified sampling* dan *optimize selection* memiliki nilai akurasi yang paling tinggi di antara metode yang diujikan, nilai akurasi yang diperoleh sebesar 80%. Sehingga dapat disimpulkan bahwa penerapan algoritma *stratified sampling* dan *optimize selection* pada algoritma C4.5 mampu meningkatkan akurasi algoritma C4.5 dalam memprediksi kelayakan kredit yang di uji pada dataset *german credit data*.

DAFTAR PUSTAKA

- [1] Y. Religia, G. T. Pranoto, and E. D. Santosa, "South German Credit Data Classification Using Random Forest Algorithm to Predict Bank Credit Receipts," *JISA (Jurnal Inform. dan Sains)*, vol. 3, no. 2, pp. 62–66, 2020, doi: 10.31326/jisa.v3i2.837.
- [2] I. Ubaedi and Y. M. Djaksana, "Optimasi Algoritma C4.5 Menggunakan Metode Forward Selection Dan Stratified Sampling Untuk Prediksi Kelayakan Kredit," *JSil (Jurnal Sist. Informasi)*, vol. 9, no. 1, pp. 17–26, 2022, doi: 10.30656/jsii.v9i1.3505.
- [3] A. Muhidin and M. Casdi, "OPTIMASI ALGORITMA NAÏVE BAYES BERBASIS PARTICLE SWARM OPTIMIZATION (PSO) DAN STRATIFIED UNTUK MENINGKATKAN AKURASI PREDIKSI PENYAKIT DIABETES," *SIGMA – J. Teknol. Pelita Bangsa*, vol. 10, no. September, pp. 151–157, 2019.
- [4] Syafi'i, O. Nurdiawan, and G. Dwilestari, "PENERAPAN MACHINE LEARNING UNTUK MENENTUKAN KELAYAKAN KREDIT MENGGUNAKAN METODE SUPPORT VEKTOR MACHINE," *J. Sist. Inf. dan Manaj.*, vol. 10, no. 2, 2022.
- [5] A. Yani, "Analisa Kelayakan Kredit Menggunakan Artificial Neural Network dan Backpropogation (Studi Kasus German Credit Data)," *J. Ilm. KOMPUTASI*, vol. 18, no. 4, pp. 385–390, 2019.

- [6] N. L. Hanun, A. U. Zailani, P. Studi, T. Informatika, and U. Pamulang, "PENERAPAN ALGORITMA KLASIFIKASI RANDOM FOREST UNTUK PENENTUAN KELAYAKAN PEMBERIAN KREDIT DI KOPERASI MITRA SEJAHTERA," *J. Technol. Inf.*, vol. 6, no. 1, pp. 7–14, 2020.
- [7] E. A. Riyanto *et al.*, "ANALISIS KINERJA ALGORITMA CART DAN NAIVE BAYES BERBASIS PARTICLE SWARM OPTIMIZATION (PSO) UNTUK KLASIFIKASI KELAYAKAN PERFORMANCE ANALYSIS CART AND NAIVE BAYES ALGORITHM BASED ON PARTICLE SWARM OPTIMIZATION (PSO) TO CLASSIFY THE CREDIT PROPERNESS OF ," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, pp. 55–60, 2021, doi: 10.25126/jtiik.202182988.
- [8] A. Saputra and T. A. Yoga, "OPTIMASI CHI SQUARE DAN PERBAIKAN TEKNIK PRUNNING UNTUK PENINGKATAN AKURASI ALGORITMA C4.5 DALAM MODEL KASUS PREDIKSI KETERLAMBATAN BIAYA KULIAH," *JIKO-Jurnal Inform. dan Komput.*, vol. 6, no. 2, pp. 231–241, 2022.
- [9] M. Mahpuz, A. Muliawan Nur, and L. M. Samsu, "Penerapan Algoritma C4.5 Dalam Mengklasifikasi Status Gizi Balita Pada Posyandu Desa Dames Damai Kabupaten Lombok Timur," *Infotek J. Inform. dan Teknol.*, vol. 5, no. 1, pp. 72–81, 2022, doi: 10.29408/jit.v5i1.4414.
- [10] A. K. B. Ginting, M. S. Lydia, and E. M. Zamzami, "Reduksi Atribut Menggunakan Chi Square untuk Optimasi Kinerja Metode Decision Tree C4.5," *J. Edukasi dan Penelit. Inform.*, vol. 9, no. 1, pp. 44–49, 2023.
- [11] T. Saegusa, "Nonparametric inference for distribution functions with stratified samples," *J. Stat. Plan. Inference*, vol. 215, pp. 356–367, 2021, doi: 10.1016/j.jspi.2021.05.001.
- [12] M. Subzar, S. Ahmad Lone, M. Aslam, A. Hussein AL-Marshadi, and S. Maqbool, "Exponential ratio estimator of the median: An alternative to the regression estimator of the median under stratified sampling," *J. King Saud Univ. - Sci.*, vol. 35, no. 3, p. 102536, 2023, doi: 10.1016/j.jksus.2022.102536.
- [13] R. T. Prasetyo and E. Ripandi, "Optimasi Klasifikasi Jenis Hutan Menggunakan Deep Learning Berbasis Optimize Selection," *J. Inform.*, vol. 6, no. 1, pp. 100–106, 2019, doi: 10.31311/ji.v6i1.5176.
- [14] K. Anam, B. Nurhakim, and C. Juliane, "Komparasi Algoritma Klasifikasi Data Mining Menggunakan Optimize Selection untuk Peminatan Program Studi," *Build. Informatics, Technol. Sci.*, vol. 4, no. 2, pp. 606–613, 2022, doi: 10.47065/bits.v4i2.2160.
- [15] Ardiyansyah and P. A. Rahayuningsih, "PENERAPAN TEKNIK SAMPLING UNTUK MENGATASI IMBALANCE CLASS PADA KLASIFIKASI ONLINE SHOPPERS INTENTION," *J. Tek. Inf. dan Komput.*, vol. 4, no. 1, pp. 7–15, 2020.
- [16] A. Rifai, R. Aulianita, and D. Mining, "Komparasi Algoritma Klasifikasi C4.5 dan Naïve Bayes Berbasis Particle Swarm Optimization Untuk

Penentuan Resiko Kredit," *J. Sentra Penelit. Eng. dan Edukasi*, vol. 10,
no. 2, 2018.