Indonesian Journal of Computer Science
ISSN 2302-4364 (*print*) dan 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam, No. 1, Padang, Indonesia, Telp. (0751) 7056199, 7058325
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Perbandingan Performa Algoritma *Decision Tree* untuk Klasifikasi Penerima Beasiswa Bank Indonesia

Dimas Bayu Stiawan[1], Yusuf Sulistyo Nugroho[2]
[1]L200190150@student.ums.ac.id, [2]yusuf.nugroho@ums.ac.id
[1,2] Universitas Muhammadiyah Surakarta

Informasi Artikel

Diterima : 6 Ags 2023
Direview : 11 Ags 2023
Disetujui : 29 Ags 2023

Kata Kunci

Beasiswa, *Decision Tree,* ID3, C4.5, CART

Abstrak

Beasiswa merupakan bantuan yang bisa digunakan sebagai biaya penunjang pendidikan yang diberikan oleh institusi atau lembaga. Pelaksanaan seleksi mahasiswa di Universitas Muhammadiyah Surakarta (UMS) yang berhak menerima beasiswa didasarkan pada ketentuan pemberi beasiswa. Namun klasifikasi kandidat yang dilakukan dengan aplikasi Excel menyebabkan sasaran penerima beasiswa kurang tepat karena tidak konsisten dan adanya unsur subjektivitas. Untuk membantu mengatasi masalah tersebut, penelitian ini bertujuan untuk menerapkan 3 algoritma *Decision Tree,* yaitu ID3, C4.5, serta CART dalam melakukan klasifikasi Penerimaan Beasiswa BI dan membandingkan performa dari 3 algoritma tersebut. Data yang digunakan sebanyak 398 pendaftar Beasiswa Bank Indonesia tahun 2022 yang diperoleh dari Biro Kemahasiswaan UMS. Hasil evaluasi kinerja ketiga algoritma menunjukkan bahwa algoritma CART memiliki nilai tertinggi dalam *accuracy*, *precision*, dan *recall*, yakni masing-masing sebesar 72%, 92,59%, dan 74,62%. Selanjutnya algoritma ID3 dan C4.5 memiliki nilai *accuracy*, *precision*, dan *recall* yang sama, yakni masing-masing sebesar 67,26%, 85,71%, dan 71,74%.

Keywords

Scholarships, Decision Tree, ID3, C4.5, CART

Abstrak

Scholarships are assistance that can be used as educational support costs given by institutions or agencies. The selection of students at the Muhammadiyah University of Surakarta (UMS) who are entitled to receive scholarships is based on the provisions of the scholarship provider. However, the classification of candidates using the Excel application causes the targeting of scholarship recipients inaccurate because of inconsistency and subjectivity. To help overcome this problem, this study aims to apply 3 Decision Tree algorithms, namely ID3, C4.5, and CART, in classifying BI Scholarship Acceptance and comparing the performance of the three algorithms. The data used were 398 applicants for the 2022 Bank Indonesia Scholarship obtained from the UMS Student Affairs Bureau. The performance evaluation results of the three algorithms show that the CART algorithm has the highest scores in accuracy, precision, and recall, which are 72%, 92.59%, and 74.62%, respectively. Furthermore, the ID3 and C4.5 algorithms have the same accuracy, precision, and recall values, namely 67.26%, 85.71%, and 71.74%, respectively.

A. Pendahuluan

Beasiswa dapat diartikan sebagai bantuan yang dimaksudkan bisa digunakan sebagai biaya penunjang pendidikan yang bukan berasal dari dana pribadi ataupun orang tua, namun diberikan institusi pemerintahan, kedutaan besar, perusahaan swasta, yayasan atau universitas [1]. Bantuan biaya diberikan kepada individu yang memenuhi syarat, berdasarkan klasifikasi, kualitas, dan kompetensi mereka sebagai penerima beasiswa. Beasiswa biasanya berupa bantuan pendanaan pendidikan saja ataupun dapat berupa bantuan fasilitas pendukung pendidikan berupa buku ataupun biaya hidup. Sebagai contoh adalah beasiswa Bank Indonesia (BI) yang terdapat di Universitas Muhammadiyah Surakarta (UMS).

Program Beasiswa BI yang diberikan oleh Bank Indonesia di UMS merupakan sebuah program sosial untuk memberikan bantuan biaya kuliah kepada mahasiswa program S1 dan D3 di UMS. Tujuan dari program ini adalah untuk meningkatkan angka partisipasi pendidikan tinggi, indeks pembangunan manusia, dan daya saing bangsa. Program ini juga bertujuan untuk memotivasi generasi muda untuk menyelesaikan pendidikan tinggi, mengembangkan komunitas mahasiswa berwawasan kebanksentralan, dan menciptakan *frontlines*, *change agents*, dan *future leaders*. Selain menyediakan pembiayaan untuk biaya pendidikan, tunjangan studi, dan biaya hidup, para penerima Beasiswa BI juga memiliki kesempatan untuk bergabung dalam komunitas Generasi Baru Indonesia (GenBI). Komunitas ini menyediakan pelatihan berkala dan terencana guna meningkatkan kompetensi individu, mengembangkan karakter, dan jiwa kepemimpinan agar para penerima beasiswa dapat menjadi insan unggul dan memiliki daya saing yang tinggi [2].

Pelaksanaan seleksi mahasiswa yang berhak menerima beasiswa didasarkan pada syarat dan ketentuan lembaga pemberi beasiswa. Apabila sebuah institusi pendidikan memiliki banyak pendaftar beasiswa dan proses klasifikasi kandidat masih dilakukan dengan cara melalui lembar kerja (*Microsoft Excel/Google Spreadsheet*) dengan cara membandingkan, mengurutkan, atau menyortir sesuai dengan ketentuan, maka hal ini dapat menyebabkan sasaran penerima beasiswa kurang tepat karena tidak konsisten akibat unsur subjektivitas.

Salah satu solusi untuk menyelesaikan masalah ini adalah *educational data mining* yang merupakan metode data mining pada basis data pendidikan [3]. Metode data mining yang diaplikasikan pada penelitian ini adalah algoritma *decision tree* untuk melakukan klasifikasi. *Decision tree* adalah model berurutan yang secara logis menggabungkan urutan tes sederhana pada setiap pengujian membandingkan atribut numerik dengan nilai ambang atau atribut nominal dengan serangkaian nilai yang layak [4].

Sejalan dengan penelitian terdahulu [5], metode *decision tree* merupakan metode yang mencoba untuk menentukan fungsi pendekatan yang memiliki nilai diskrit dan tahan terhadap data dengan kesalahan (*noisy* data) serta bisa mempelajari ekspresi *disjunctive* (ekspresi *OR*). Terdapat tiga algoritma *decision tree* yang paling umum dipakai dan digunakan dalam penelitian ini dalam mengetahui perbandingan performa algoritma *decision tree* untuk klasifikasi penerima Beasiswa Bank Indonesia. Algoritma yang digunakan diantaranya ID3 (*Iterative Dichotomizer* 3) dikembangkan oleh J.R Quinlan pada tahun 1986 [6][7]. C4.5 merupakan evolusi dari ID3, yang disajikan oleh pencipta yang sama oleh Quinlan pada tahun 1993 [8]. Selanjutnya Algoritma CART adalah singkatan dari

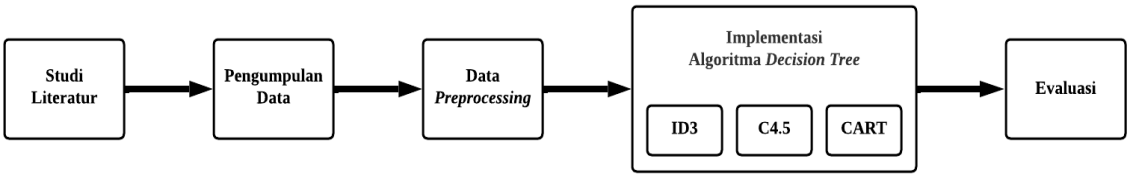
Classification and Regression Trees yang dikembangkan oleh Breiman et al. pada tahun 1984 [9].

Penelitian sebelumnya dalam perbandingan kinerja algoritme C4.5 dan *Naïve Bayes* pada penerimaan beasiswa menunjukkan bahwa algoritma C4.5 memiliki tingkat akurasi 96.40% dan algoritma *Naïve Bayes* sebesar 95.11% [10]. Dalam penelitian yang lain [11], perbandingan algoritma *Naïve Bayes*, *Decision Tree*, dan SVM dalam klasifikasi persetujuan pembiayaan nasabah koperasi syariah memiliki hasil akurasi masing-masing 77,29%; 89,02%; dan 89,86%. Penelitian lain yang juga dilakukan menghasilkan bahwa algoritma C4.5 memiliki akurasi sebesar 95,76% yang lebih baik daripada CART dengan akurasi hanya 95,11% [12]. Algoritma C4.5 menghasilkan akurasi sebesar 85,55% pada proses klasifikasi beasiswa keluarga kurang mampu [13]. Dari analisis yang dilakukan pada ID3, C4.5, dan CART, dapat disimpulkan bahwa algoritma pembelajaran pohon keputusan memberikan tingkat akurasi yang cukup tinggi. Namun, setiap algoritma harus diimplementasikan berdasarkan kondisi dataset. Untuk dataset umum, ID3 akan memberikan hasil yang memuaskan, tetapi jika pemangkasan pohon diperlukan, C4.5 akan memberikan hasil yang diharapkan. Jika dataset mengandung ketidakmurnian, algoritma CART akan menggunakan Indeks Gini untuk memisahkan atribut secara biner [14].

Berdasarkan permasalahan yang telah dijabarkan dan penelitian sebelumnya, penelitian ini membandingkan 3 algoritma dalam *decision tree,* yaitu ID3, C4.5, dan CART untuk melakukan klasifikasi Penerima Beasiswa Bank Indonesia pada seleksi tingkat universitas di UMS. Performa masing-masing algoritma dibandingkan berdasarkan nilai *accuracy, precision*, dan *recall.* Hasil penelitian menunjukkan bahwa algoritma CART menghasilkan nilai *accuracy, precision*, dan *recall* tertinggi dibandingkan dengan algoritma ID3 dan C4.5. Sehingga dapat disimpulkan bahwa algoritma CART lebih efektif dan andal dalam mengklasifikasi Penerima Beasiswa Bank Indonesia di UMS. Keunggulan performa CART menunjukkan bahwa algoritma ini mungkin lebih mampu menangani karakteristik khusus dari dataset penelitian, memberikan keputusan yang lebih tepat dan mengurangi kesalahan.

B. Metode Penelitian

Tahap penelitian yang digunakan untuk membandingkan performa 3 algoritma *decision tree* bagi penerima Beasiswa Bank Indonesia dapat ditunjukkan pada Gambar 1.



Gambar 1. Tahap penelitian yang terdiri dari 5 langkah utama, yaitu studi literatur, pengumpulan data, *data preprocessing*, implementasi *decision tree*, dan evaluasi

1. Studi Literatur

   Tahap studi literatur ini merupakan langkah penelitian untuk mengumpulkan dasar teori-teori yang diperlukan dalam penelitian [12]. Pengumpulan informasi dengan topik yang sesuai merupakan langkah awal dari

tahapan studi literatur. Informasi yang digunakan sebagai acuan penelitian diperoleh dari laporan penelitian sebelumnya, karya ilmiah, dan buku.

2. Pengumpulan Data

Data yang digunakan dalam proses penelitian ini merupakan data Pendaftar Beasiswa Bank Indonesia di Universitas Muhammadiyah Surakarta tahun 2022 yang diperoleh dari Biro Kemahasiswaan UMS. Dataset yang diperoleh memiliki data sebanyak 398 baris data.

3. Data *Preprocessing*

Data *preprocessing* adalah proses transformasi, menggabungkan, atau mengubah data menjadi bentuk yang sesuai, agar dapat diproses dengan perhitungan algoritma *decision tree* [15]. Data penelitian yang diperoleh merupakan data mentah (data asli atau primer) yang perlu dilakukan *preprocessing.* Tahap ini termasuk menentukan atribut yang akan digunakan dalam proses klasifikasi dan melakukan modifikasi pada data dengan menangani data yang hilang, data ganda, dan mengubah data atribut menjadi tipe kategori. Terakhir, untuk melindungi serta menjaga privasi data pribadi dengan membuat anonim dan menghapus atribut data (tidak diperlukan dalam proses klasifikasi) seperti nama, NIM, nomor telepon, rekening, dan lain-lain.

4. Implementasi Algoritma *Decision Tree*

Tahap ini dilakukan klasifikasi pada dataset pendaftar beasiswa menggunakan bahasa pemrograman *Python*. Proses klasifikasi menggunakan algoritma *decision tree* dengan metode ID3, C4.5, dan CART untuk menghasilkan sebuah pohon keputusan. Hasil analisis digunakan untuk mengetahui performa dari masing-masing metode algoritma *decision tree* yang paling baik dalam proses klasifikasi pendaftar Beasiswa Bank Indonesia di UMS.

Decision tree atau pohon keputusan adalah salah satu metode yang termasuk dalam metode klasifikasi *unsupervised* yang berbentuk seperti pohon [16]. Algoritma ini dapat membantu dalam menentukan keputusan serta memiliki struktur yang mirip dengan pohon yang biasanya divisualisasikan melalui gambar daun dan cabang [17]. Pada seluruh simpul (*node*) pohon mewakili atribut yang sudah diuji. Setiap cabang mempresentasikan pembagian hasil uji dan *node* daun (*leaf*) merupakan kelompok kelas tertentu. *Node* pada level teratas dari sebuah pohon keputusan merupakan akar (*root*), biasanya berupa atribut dengan pengaruh terbesar pada suatu kelas tertentu [18]. Algoritma *decision tree* memiliki beberapa macam, namun dalam penelitian ini hanya membandingkan algoritma ID3, C4.5, dan CART.

a. Algoritma ID3

Algoritma ID3 (*Iterative Dichotomizer* 3) merupakan algoritma *decision tree* paling dasar yang dikembangkan pertama kali oleh J. Ross Quinlan [19]. ID3 membangun *decision tree* didasari dengan pencarian secara *top-down* (dari atas ke bawah), secara menyeluruh (*greedy*) melalui kumpulan data untuk menguji setiap atribut di setiap simpul pohon dari akar hingga ke daun [3].

Algoritma ID3 bekerja dengan cara penentuan *entropy* dan *information gain* digunakan untuk pemilihan atribut. Persamaan (1) dan (2) menunjukkan rumus mencari *entropy* dan *information gain* berturut-urut, serta langkah penentuan dengan algoritma ID3 secara singkat [19].

$$Entropy\ (S) = -\sum_{i=1}^{k} pi * log_2 pi \quad (1)$$

Keterangan:

S : Himpunan kasus
k : Jumlah partisi S
pi : Probabilitas yang diperoleh dari sum (ya) dibagi dengan total kasus

Perhitungan *entropy* kemudian dilakukan pada setiap kasus, dimana *entropy* merupakan ukuran ketidakpastian, yaitu perbedaan dalam keputusan tentang nilai atribut tertentu [19]. Perbedaan keputusan (ketidakpastian) akan semakin tinggi jika nilai *entropy* semakin tinggi. Selanjutnya adalah menghitung *information gain* dengan persamaan (2), dimana *information gain* adalah langkah menentukan tingkat pengaruh suatu atribut dalam membagi data menjadi kelompok yang lebih homogen.

$$Gain\ (S, A) = Entropy\ (S) -\sum_{i=1}^{k} \frac{|Si|}{|S|} * Entropy\ (S_i) \quad (2)$$

Keterangan:

S : Himpunan kasus
A : Atribut
$|S_i|$: Jumlah kasus dalam partisi ke i
$|S|$: Jumlah kasus pada S

Langkah selanjutnya adalah memilih atribut dengan *information gain* terbesar, kemudian membentuk simpul dengan atribut yang telah dipilih pada langkah sebelumnya. Tahap perhitungan *information gain* diulang secara terus-menerus sampai seluruh data masuk pada kelas yang sama. Pada atribut yang telah dipilih tidak dimasukkan kembali dalam perhitungan nilai *information gain*.

b. Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membangun model pohon keputusan (*decision tree*). Algoritma C4.5 dikembangkan oleh Ross Quinlan pada tahun 1993 sebagai evolusi dari algoritma ID3 [19]. Algoritma ini dapat digunakan untuk data yang berbasis kategorikal dan numerik. Ada empat hal yang membedakan algoritma C4.5 dan ID3, diantaranya adalah kemampuan terhadap *noise* data yang lebih baik, dapat mengatasi variabel dengan tipe kontinu dan diskret, mampu menangani variabel dengan *missing value,* serta kemampuan memangkas cabang dari pohon keputusan [19]. Tahapan algoritma C4.5 dimulai dengan membagi data menjadi data latih (*training*) dan data uji(*testing*)[20]. Data *training* adalah data latih guna membangun pohon keputusan yang sebelumnya telah diuji keabsahannya. Data *testing* yaitu kumpulan data yang akan menjadi parameter pada klasifikasi dataset [12].

Tahap berikutnya adalah menghitung nilai *split info,* dimana dapat ditentukan dengan rumus pada persamaan (3).

$$Split\ Info\ (S, A) = \sum_{i=1}^{k} \frac{Sj}{S} * log_2 \frac{Sj}{S} \quad (3)$$

Keterangan:

S : Ruang *sample*

A : Atribut

S_j : Jumlah sample untuk atribut ke j

Selanjutnya adalah tahap menentukan nilai *gain ratio* dengan rumus yang tertera pada persamaan (4). *Gain ratio* dengan nilai tertinggi akan digunakan untuk atribut akar. Maka, dengan begitu akan terbentuk pohon keputusan sebagai *node* 1.

$$Gain\ Ratio\ (S, A) = \frac{Gain\ (S,A)}{Split(S,A)} \quad (4)$$

Keterangan:

A : Atribut

Gain (S, A) : *Information gain* pada atribut *(S, A)*

Split (S, A) : *Split information* pada atribut *(S, A)*

Terakhir mengulangi seluruh proses hingga semua cabang mempunyai kelas yang sama. Proses percabangan akan berhenti saat semua kasus pada simpul n memperoleh kelas yang sama, variabel independen telah tidak ada pada kasus yang di partisi lagi, dan tidak terdapat kasus pada cabang yang kosong.

c. Algoritma CART

CART (*Classification and Regression Tree*) adalah salah satu metode dari *decision tree* (pohon keputusan) yang menggunakan algoritma penyekatan rekursif dengan cara biner (*binary recursive partitioning*) [21]. CART memiliki kemampuan untuk melakukan klasifikasi dan analisis regresi [12]. Tujuan CART adalah untuk mengklasifikasikan objek menjadi dua atau lebih kelompok [17]. Pada sekumpulan data yang terdiri dari p variabel independen dan satu variabel dependen, bila yang dimiliki bertipe kategori, maka CART menghasilkan pohon klasifikasi, sedangkan saat variabel dependen dengan tipe kontinu atau numerik maka CART menghasilkan pohon regresi [17].

Pembuatan pohon keputusan dengan algoritma CART dilakukan melalui perhitungan indeks gini untuk setiap kelas [12] dengan menggunakan formula dalam persamaan 5.

$$j\ (t) = 1 - \sum_{i,j=1} P^2(j|t) \quad (5)$$

Pada persamaan 5, *node* t dibagi menjadi dua subset, yaitu D1 dan D2, dengan ukuran b1 dan b2 masing-masing. Selanjutnya, indeks gini total untuk pembelahan dihitung dihitung berdasarkan subset-subset tersebut dan ditemukan dalam persamaan 6.

$$Gini_{pembelahan}\ (t) = \frac{b_1}{b} gini\ (D_1) + \frac{b_2}{b} gini\ (D_2) \quad (6)$$

Keterangan:

Ginipembelahan : Nilai indeks gini untuk setiap variabel

Gini(D1) : Nilai indeks gini dari subset D1 untuk setiap variabel

Gini(D2) : Nilai indeks gini dari subset D2 untuk setiap variabel

B : Banyaknya data dalam sebuah variabel

b1 : Banyaknya data dalam subset d1

b2 : Banyaknya data dalam subset d2

Jika jumlah sampel dalam suatu kelas sama atau kurang dari 5, maka *node* tersebut ditetapkan sebagai *node* terminal [22]. Penentuan label untuk *node* terminal ini dilakukan berdasarkan persamaan 7, dengan mempertimbangkan jumlah yang paling banyak, sebagai berikut:

$$P(j_0|t) = max_j P(\text{j}|t) + max_j \frac{m_j(t)}{m(t)} \quad (7)$$

5. Evaluasi

Hasil klasifikasi yang diperoleh dari algoritma ID3, C4.5, dan CART, selanjutnya dievaluasi menggunakan *Confusion matrix,* yang memiliki fungsi untuk menganalisis *classifier* dalam mengenali tupel dari kelas yang berbeda [23]. *Confusion matrix* juga digunakan untuk mengetahui nilai *accuracy*, *precision*, dan *recall* hasil klasifikasi [15]. Nilai *accuracy*, *precision*, dan *recall* yang diperoleh kemudian dibandingkan untuk mengetahui algoritma yang memiliki performa terbaik dalam melakukan klasifikasi penerima Beasiswa Bank Indonesia di UMS.

Nilai *confusion matrix* dapat dihitung berdasarkan contoh pada Tabel 1, serta dapat digunakan untuk menghitung nilai *accuracy, precision* dan *recall* dengan rumus masing-masing pada persamaan (10), (11), dan (12)[24].

Tabel 1. *Confusion Matrix*

Klasifikasi yang benar	Diklasifikasikan sebagai	
	+	-
+	*True Positive*	*False Negative*
-	*False Positive*	*True Negative*

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \text{ x } 100\% \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \text{ x } 100\% \quad (11)$$

$$Recall = \frac{TP}{TP+FN} \text{ x } 100\% \quad (12)$$

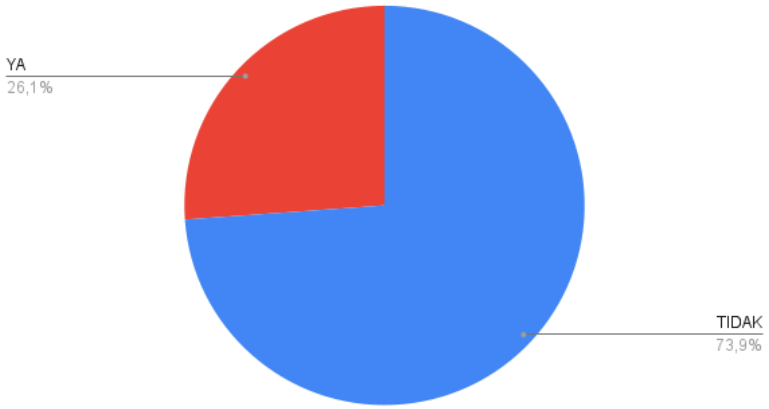
Keterangan:

TP (*True Positive*) : Jumlah data positif dengan prediksi positif
FP (*False Positive*) : Jumlah data negatif dengan prediksi positif
FN (*False Negative*) : Jumlah data positif dengan prediksi negatif
TN (*True Negative*) : Jumlah data negatif dengan prediksi negatif

C. Hasil dan Pembahasan

1. Pengumpulan Data

Jumlah data yang diolah setelah menghilangkan duplikasi adalah 375 data. Data tersebut memiliki variabel dependen (Y) yang terdistribusi dalam 2 kelas, yaitu kelas YA (26,1%) dan TIDAK (73,9%) dari total data, yang dapat dilihat pada Gambar 2.



Gambar 2. Perbandingan jumlah kelas pada variabel dependen (Y)

2. Data *Preprocessing*

Proses klasifikasi menggunakan 9 atribut yang terdiri dari dua variabel, yaitu 8 variabel independen (X) dan 1 variabel dependen (Y). Definisi masing-masing variabel yang digunakan dalam penelitian ini dijelaskan dalam Tabel 2.

Tabel 2. Daftar atribut yang digunakan dalam penelitian

Variabel	Atribut	Kelas
X1	Semester	1. Angkatan A (Semester 3 dan 4) 2. Angkatan B (Semester 5 dan 6) 3. Angkatan C (Semester 7 dan 8)
X2	SKS	1. Kelas A (0 – 100) 2. Kelas B (< 100)
X3	IPK	1. Kurang (< 3,0) 2. Baik (3,0 <= IPK < 3,5) 3. Sangat Baik (>= 3,5)
X4	Minat Bakat *Life Skill*	SENI, OLAHRAGA, AKADEMIK
X5	Potensi Diri	SENI, OLAHRAGA, AKADEMIK
X6	Aktivitas Sosial	1. Memiliki 1 kegiatan sosial 2. Memiliki 2 kegiatan sosial 3. Memiliki lebih dari 2 kegiatan sosial
X7	Alasan	1. Memiliki alasan kurang (melakukan pekerjaan sebatas tuntutan tugas dan tanggungjawab) 2. Memiliki alasan baik (mengerti dan mendukung misi dan tujuan organisasi) 3. Memiliki alasan baik sekali (menempatkan kepentingan organisasi diatas kepentingan dan keinginan pribadi)
X8	Surat Rekomendasi GenBI	YA dan TIDAK
Y	Hasil	YA (Diterima) dan TIDAK (Ditolak)

Tabel 3 menunjukkan penggalan dataset yang telah dilakukan proses *data preprocessing* dan mengelompokkan dataset dengan mengubahnya menjadi data kategori.

Tabel 3. Penggalan data setelah *preprocessing*

No.	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	Angkatan B	A	Sangat Baik	Akademik	Akademik	Aktif	Baik	Tidak	Tidak
2	Angkatan C	B	Sangat Baik	Seni	Seni	Kurang	Kurang	Tidak	Tidak
3	Angkatan B	A	Baik	Olahraga	Seni	Aktif	Baik	Tidak	Tidak
4	Angkatan B	B	Sangat Baik	Olahraga	Olahraga	Kurang	Kurang	Tidak	Tidak
5	Angkatan B	A	Sangat Baik	Seni	Akademik	Kurang	Kurang	Tidak	Tidak
6	Angkatan C	B	Sangat Baik	Akademik	Akademik	Kurang	Lebih Baik	Tidak	Tidak
7	Angkatan C	B	Sangat Baik	Seni	Akademik	Sangat Aktif	Lebih Baik	Tidak	Ya
8	Angkatan C	B	Sangat Baik	Akademik	Akademik	Kurang	Baik	Tidak	Tidak
9	Angkatan C	B	Sangat Baik	Akademik	Akademik	Aktif	Kurang	Tidak	Tidak
10	Angkatan C	B	Sangat Baik	Akademik	Akademik	Aktif	Lebih Baik	Tidak	Tidak

3. Implementasi *Decision Tree*

Proses klasifikasi dilakukan dengan menerapkan bahasa pemrograman *Python* dengan *library* Scikit-Learn untuk *machine learning.* Implementasi algoritma *decision tree* dengan membagi data menjadi 2 kelompok yaitu data *training* dan data *testing* dengan rasio 70% : 30%, dapat dilihat pada Gambar 2.

```
# Membagi data menjadi data latih dan data uji
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.3,
                                                    random_state=42)
```

Gambar 3. Pembagian data *training* dan *testing* menggunakan *Python*

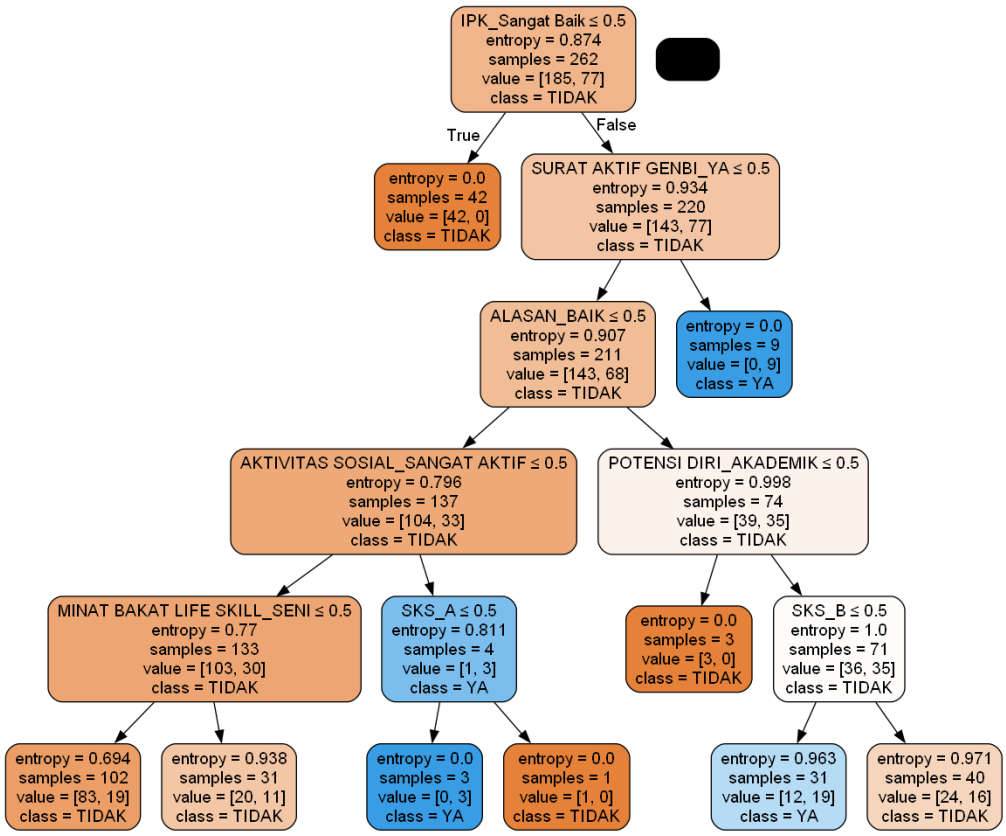
a. Algoritma ID3

Algoritma ID3 diterapkan menggunakan *library* Scikit-Learn dengan memanfaatkan *class DecisionTreeClassifier* dengan *criterion* berupa *entropy* dan *max_depth* sebesar 10, seperti yang ditunjukkan pada Gambar 4.

```
# Membangun model ID3
id3_model = DecisionTreeClassifier(criterion='entropy', max_depth=10)
id3_model.fit(X_train, y_train)
```

Gambar 4. Penerapan model algoritma ID3 menggunakan *Python*

Penerapan algoritma ID3 ini menghasilkan pohon keputusan yang dapat ditunjukkan pada Gambar 5. Sedangkan implementasi algoritma ID3 telah menghasilkan nilai *confusion matrix* yang ditunjukkan pada Tabel 4.



Gambar 5. *Decision tree* algoritma ID3

Tabel 4. *Confusion Matrix* algoritma ID3

	Pred. TIDAK (Ditolak)	*Pred.* YA (Diterima)
True TIDAK (Ditolak)	66	26
True YA (Diterima)	11	10

Berdasarkan Tabel 4, diperoleh bahwa 66 *record* diprediksi "TIDAK" pada kelompok data "TIDAK" dan sebanyak 26 *record* diprediksi "YA" pada kelompok data "TIDAK". Selanjutnya sebanyak 11 *record* diprediksi "TIDAK" pada kelompok data "YA" dan 10 record diprediksi "YA" pada kelompok data "YA".

Nilai *accuracy*, *precision*, dan *recall* yang dihasilkan sebagai berikut:

$$Accuracy = \frac{66+10}{113} \text{ x } 100\% = 67{,}26\%$$

$$Precision = \frac{66}{66+11} \text{ x } 100\% = 85{,}71\%$$

$$Recall = \frac{66}{66+26} \text{ x } 100\% = 71{,}74\%$$

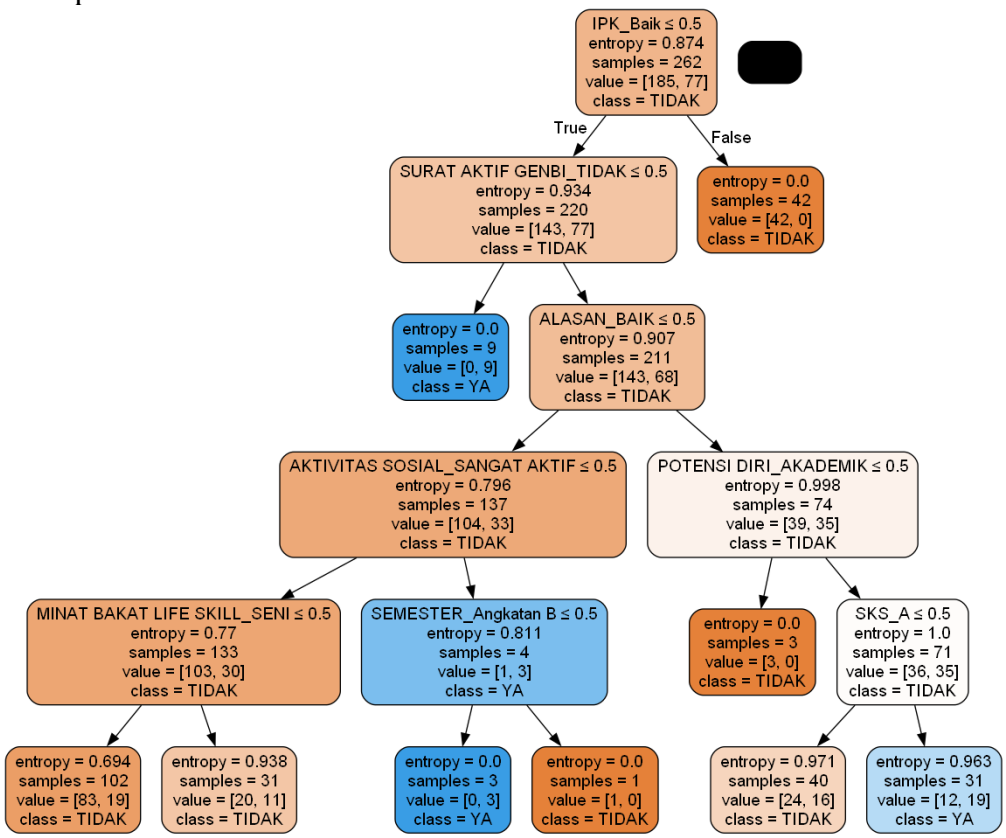
b. Algoritma C4.5

Penerapan *Python* untuk klasifikasi data menggunakan algoritma C4.5 dengan *library* Scikit-Learn memanfaatkan *class DecisionTreeClassifier.* Sedangkan *criterion* ditentukan berupa *entropy*, dengan *splitter* dipilih *best,* dan *max_depth* sebesar 10, seperti ditunjukkan pada Gambar 6.

```
# Membangun model C4.5
c45_model = DecisionTreeClassifier(criterion='entropy', splitter='best')
c45_model.fit(X_train, y_train)
```

Gambar 6. Penerapan model algoritma C4.5 dengan *Python*

Algoritma C4.5 menghasilkan sebuah pohon keputusan seperti yang dapat dilihat pada Gambar 7.



Gambar 7. *Decision tree* algoritma C4.*5*

Implementasi algoritma C4.5 tersebut menghasilkan nilai *confusion matrix* yang ditunjukkan pada Tabel 5.

Tabel 5. *Confusion Matrix* algoritma C4.5

	Pred. TIDAK (Ditolak)	*Pred.* YA (Diterima)
True TIDAK (Ditolak)	66	26
True YA (Diterima)	11	10

Tabel 5 menjelaskan bahwa 66 *record* diprediksi "TIDAK" pada kelompok data "TIDAK" dan sebanyak 26 *record* diprediksi "YA" pada kelompok data "TIDAK". Selanjutnya sebanyak 11 *record* diprediksi "TIDAK" pada kelompok data "YA" dan 10 record diprediksi "YA" pada kelompok data "YA".

Nilai *accuracy*, *precision*, dan *recall* yang dihasilkan sebagai berikut:

$$Accuracy = \frac{66+10}{113} \text{ x } 100\% = 67{,}26\%$$

$$Precision = \frac{66}{66+11} \text{ x } 100\% = 85{,}71\%$$

$$Recall = \frac{66}{66+26} \text{ x } 100\% = 71{,}74\%$$

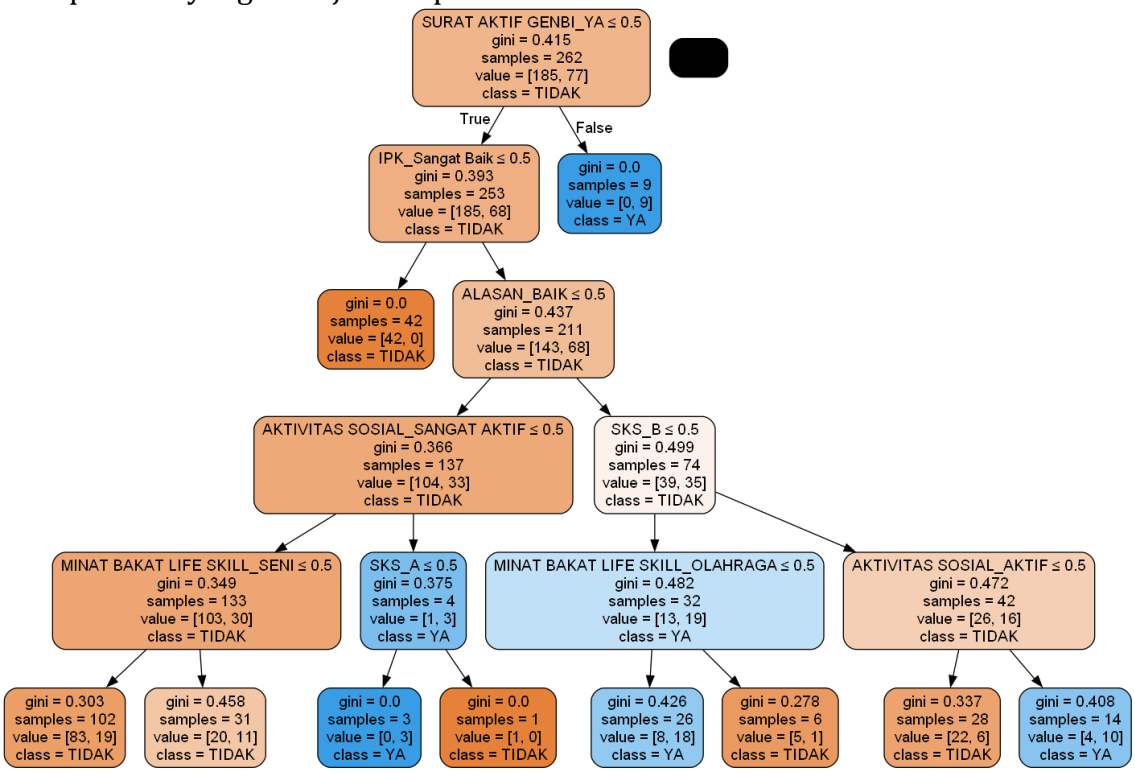
c. Algoritma CART

Adapun algoritma CART diterapkan menggunakan *library Scikit-Learn* dengan memanfaatkan *class DecisionTreeClassifier* dengan *criterion* berupa *gini* dan *max_depth* sebesar 10 seperti ditunjukkan pada Gambar 8.

```
# Membangun model CART
cart_model = DecisionTreeClassifier(criterion='gini')
cart_model.fit(X_train, y_train)
```

Gambar 8. Penerapan model algoritma CART dengan *Python*

Penerapan algoritma CART tersebut menghasilkan sebuah pohon keputusan yang ditunjukkan pada Gambar 9.



Gambar 9. *Decision tree* algoritma CART

Implementasi algoritma CART menghasilkan nilai *confusion matrix* yang ditunjukkan pada Tabel 6.

Tabel 6. *Confusion Matrix* algoritma CART

	Pred. TIDAK (Ditolak)	*Pred.* YA (Diterima)
True TIDAK (Ditolak)	67	25
True YA (Diterima)	11	10

Tabel 6 menjelaskan bahwa 50 *record* diprediksi "TIDAK" pada kelompok data "Tidak" dan sebanyak 17 *record* diprediksi "YA" pada kelompok data "TIDAK". Selanjutnya sebanyak 4 *record* diprediksi "TIDAK" pada kelompok data "YA" dan 4 *record* diprediksi "YA" pada kelompok data "YA".

Nilai *accuracy*, *precision*, dan *recall* yang dihasilkan sebagai berikut:

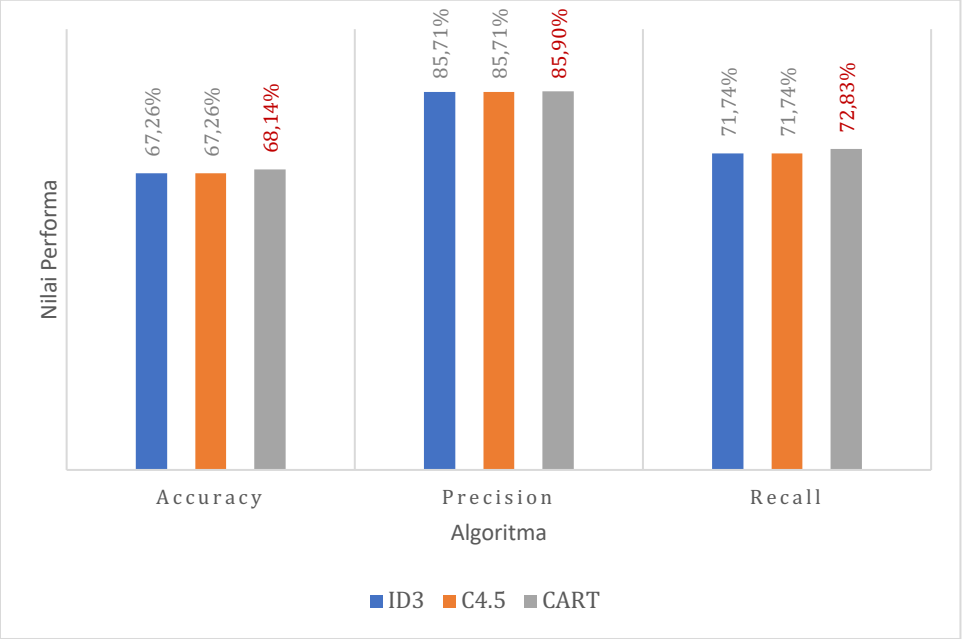
$$Accuracy = \frac{67+10}{113} \text{ x } 100\% = 68{,}14\%$$

$$Precision = \frac{67}{67+11} \text{ x } 100\% = 85{,}9\%$$

$$Recall = \frac{67}{67+25} \text{ x } 100\% = 72{,}83\%$$

4. Evaluasi

Perbandingan performa dari masing-masing algoritma yang digunakan dalam penelitian ini dilakukan dengan menggunakan nilai *accuracy*, *precision*, dan *recall.* Hasil perbandingan performa tersebut dapat dilihat pada Gambar 9.



Gambar 10. Perbandingan nilai *accuracy*, *precision*, dan *recall* dari 3 algoritma *Decision Tree*

Gambar 9 menjelaskan perbandingan nilai performa dari masing-masing algoritma ID3, C4.5, dan CART terhadap hasil klasifikasi data Pendaftar Beasiswa Bank Indonesia di Universitas Muhammadiyah Surakarta tahun 2022. Hasil perbandingan menunjukkan bahwa algoritma CART memiliki nilai tertinggi dalam hal *accuracy*, *precision*, dan *recall*, yaitu masing-masing sebesar 68,14%, 85,9%, dan 72,83%. Selanjutnya algoritma ID3 dan C4.5 memiliki nilai *accuracy*, *precision*, dan *recall* yang sama, yaitu masing-masing sebesar 67,26%, 85,71%, dan 71,74%. Hal ini mengindikasikan bahwa algoritma CART lebih efektif dan andal dalam mengklasifikasi data mahasiswa penerima beasiswa Bank Indonesia di UMS.

D. Simpulan

Pada penelitian ini, penggunaan atribut pada klasifikasi pohon keputusan untuk dataset Pendaftar Beasiswa Bank Indonesia di Universitas Muhammadiyah Surakarta tahun 2022 diimplementasikan menggunakan tiga metode *decision tree*, yaitu ID3, C4.5 dan CART. Analisis terhadap model pohon keputusan yang dihasilkan oleh ketiga algoritma tersebut mengungkap bahwa faktor Surat Aktif GenBI merupakan atribut paling diperhitungkan terhadap kemungkinan seseorang diterima beasiswa BI berdasarkan algoritma CART dengan *criterion* berupa *gini*. Sedangkan atribut IPK merupakan atribut paling diperhitungkan terhadap kemungkinan seseorang diterima beasiswa BI berdasarkan algoritma ID3 dan C4.5 dengan *criterion* berupa *entropy*.

Hasil evaluasi kinerja ketiga algoritma menunjukkan bahwa algoritma CART mencapai hasil tertinggi dalam *accuracy*, *precision*, dan *recall*, yakni masing-masing sebesar 72%, 92,59%, dan 74,62%. Selanjutnya algoritma ID3 dan C4.5 memiliki nilai *accuracy*, *precision*, dan *recall* yang sama, yakni masing-masing sebesar 67,26%, 85,71%, dan 71,74%.

Penelitian ini telah berupaya untuk menyediakan wawasan tentang perbandingan performa algoritma *decision tree* untuk klasifikasi penerima beasiswa Bank Indonesia di UMS, terdapat beberapa keterbatasan yang perlu diperhatikan. Pertama, ukuran sampel data yang digunakan dalam penelitian ini mungkin belum mewakili keragaman penuh dari implementasi *machine learning*. Serta penggunaan *library* scikit-learn yang menggunakan versi optimal dari algoritma CART.

Berdasarkan temuan dan kekurangan yang diidentifikasi, terdapat beberapa saran yang dapat menjadi arah bagi penelitian selanjutnya. Pertama, akan sangat bermanfaat untuk bisa menggunakan dataset yang lebih banyak dan beragam agar prediksi yang dihasilkan lebih akurat. Selanjutnya dapat menggunakan *library* yang berbeda pada Python ataupun menggunakan *software* pendukung lainnya.

E. Ucapan Terima Kasih

Terima kasih yang tulus kami sampaikan kepada Universitas Muhammadiyah Surakarta, khususnya kepada Biro Kemahasiswaan UMS yang telah menyediakan data untuk penelitian ini, atas dukungan yang diberikan.

F. Referensi

[1] P. Putriani and A. Mardiana, "Sistem Pendukung Keputusan Penerimaan Beasiswa Bank Indonesia Menggunakan Metode Logika Fuzzy Dan Saw (Studi Kasus Universitas Majalengka)," *Infotech journal*, vol. 8, no. 1, pp. 13–21, 2022.

[2] "Generasi Baru Indonesia (GenBI)." https://www.generasibaruindonesia.com/ (accessed Mar. 07, 2023).

[3] I. M. K. Karo, M. Y. Fajari, N. U. Fadhilah, and W. Y. Wardani, "Benchmarking Naïve Bayes and ID3 Algorithm for Prediction Student Scholarship," *IOP Conference Series: Materials Science and Engineering*, vol. 1232, no. 1, p. 012002, 2022, doi: 10.1088/1757-899x/1232/1/012002.

[4] S. B. Kotsiantis, "Decision trees: A recent overview," *Artificial Intelligence Review*, vol. 39, no. 4. pp. 261–283, 2013, doi: 10.1007/s10462-011-9272-4.

[5] W. A. Damanik and Prihandoko, "Analisis Penentuan Pemberian Beasiswa Berprestasi Menggunakan Metode Decision Tree dan SVM (Support Vector Machine)," *Jurnal Teknik Dan Informatika*, vol. 6, pp. 2018–2020, 2019, [Online]. Available: http://jurnal.pancabudi.ac.id/index.php/Juti/article/view/480.

[6] J. R. Quinlan, "Improved Use of Continuous Attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, no. 1996, pp. 77–90, 1996, doi: 10.1613/jair.279.

[7] I. H. Witten, E. Frank, and M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)," Third Edit., Boston: Morgan Kaufmann, 2011.

[8] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Machine Learning*, vol. 16, no. 3, pp. 235–240, 1994, doi: 10.1007/BF00993309.

[9] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Wadsworth International Group, 1984.

[10] C. Anam and H. B. Santoso, "Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa," *Energy - Jurnal Ilmiah Ilmu-Ilmu Teknik*, vol. 8, no. 1, pp. 13–19, 2018, [Online]. Available: https://ejournal.upm.ac.id/index.php/energy/article/view/111.

[11] N. Nurajijah and D. Riana, "Algoritma Naïve Bayes, Decision Tree, dan SVM untuk Klasifikasi Persetujuan Pembiayaan Nasabah Koperasi Syariah," *Jurnal Teknologi dan Sistem Komputer*, vol. 7, no. 2, pp. 77–82, 2019, doi: 10.14710/jtsiskom.7.2.2019.77-82.

[12] Suryani, D. Rahmadani, A. A. Muzafar, A. Hamid, R. Annisa, and Mustakim, "Analisis Perbandingan Algoritma C4.5 dan CART untuk Klasifikasi Penyakit Stroke," in *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat*, 2022, pp. 197–206, [Online]. Available: https://journal.irpi.or.id/index.php/sentimas.

[13] Y. Kustiyahningsih, B. K. Khotimah, D. R. Anamisa, M. Yusuf, T. Rahayu, and J. Purnama, "Decision Tree C 4.5 Algorithm for Classification of Poor Family Scholarship Recipients," *IOP Conference Series: Materials Science and Engineering*, vol. 1125, no. 1, p. 012048, 2021, doi: 10.1088/1757-899x/1125/1/012048.

[14] F. M. Javed Mehedi Shamrat, R. Ranjan, K. M. Hasib, A. Yadav, and A. H. Siddique, "Performance Evaluation Among ID3, C4.5, and CART Decision Tree Algorithm," *Lecture Notes in Networks and Systems*, vol. 317, no. March 2021, pp. 127–142, 2022, doi: 10.1007/978-981-16-5640-8_11.

[15] A. A. Aldino and H. Sulistiani, "Decision Tree C4.5 Algorithm for Tuition Aid Grant Program Classification (Case Study: Department of Information System, Universitas Teknokrat Indonesia)," *Edutic - Scientific Journal of Informatics Education*, vol. 7, no. 1, pp. 40–50, 2020, doi: 10.21107/edutic.v7i1.8849.

[16] H. Imaduddin, F. Y. A'la, A. Fatmawati, and B. A. Hermansyah, "Comparison of transfer learning method for COVID-19 detection using convolution neural network," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, pp. 1091–1099, 2022, doi: 10.11591/eei.v11i2.3525.

[17] Irawadi and S. Sunendiari, "Penerapan dan Perbandingan Tiga Metode

Analisis Pohon Keputusan pada Klasifikasi Penderita Kanker Payudara," *Jurnal Riset Statistika*, vol. 1, no. 1, pp. 19–27, 2021, doi: 10.29313/jrs.v1i1.22.

[18] A. H. Nasrullah, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *Jurnal Pilar Nusa Mandiri*, vol. 7, no. 2, p. 217, 2021.

[19] P. B. N. Setio, D. R. S. Saputro, and Bowo Winarno, "Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5," *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 3, pp. 64–71, 2020.

[20] R. P. S. Putri and I. Waspada, "Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika," *Khazanah Informatika : Jurnal Ilmu Komputer dan Informatika*, vol. 4, no. 1, pp. 1–7, 2018, doi: 10.23917/khif.v4i1.5975.

[21] S. H. Sumartini, "Penggunaan Metode Classification and Regression Trees (CART) untuk Klasifikasi Rekurensi Pasien Kanker Serviks di RSUD Dr. Soetomo Surabaya," *Jurnal Sains dan Seni ITS*, vol. 4, no. 2, pp. 211–216, 2015.

[22] F. E. Pratiwi and I. Zain, "Klasifikasi Pengangguran Terbuka Menggunakan CART (Classification and Regression Tree) di Provinsi Sulawesi Utara," *Jurnal Sains dan Seni ITS*, vol. 3, no. 1, pp. D54–D59, 2014, [Online]. Available: http://www.ejurnal.its.ac.id/index.php/sains_seni/article/view/6129.

[23] L. D. Yulianto, A. Triayudi, & I. D. Sholihati, "Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4.5," *Jurnal Mantik*, vol. 4, no. 1, pp. 441–451, 2020, [Online]. Available: https://iocscience.org/ejournal/index.php/mantik/index.

[24] D. Indrajaya, A. Setiawan, D. Hartanto, and H. Hariyanto, "Object Detection to Identify Shapes of Swallow Nests Using a Deep Learning Algorithm," *Khazanah Informatika : Jurnal Ilmu Komputer dan Informatika*, vol. 8, no. 2, pp. 139–148, 2022, doi: 10.23917/khif.v8i2.16489.