Jurnal KomtekInfo

https://jkomtekinfo.org/ojs

2024 Vol. 11 No. 3 Hal: 80-88 e-ISSN: 2502-8758

Implementasi Algoritma SVM dan C4.5 dalam Klasifikasi Calon Penerimaan Beasiswa

Rahmaddeni[1]✉, Syarfi Aziz[2], Zairi Saputra[1], Hafid Azis Supahri[1], Ryan Ismanizan[1]

[1] Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia, Riau, Indonesia
[2] Program Studi Manajemen Komputer, Institut Az Zuhra, Riau, Indonesia

rahmaddeni@sar.ac.id

Abstract

In the scholarship selection process, various criteria are considered thoroughly. Grades, academic and economic status, and others are important factors. However, with the many candidates applying, the manual selection process becomes inefficient and prone to errors. The purpose of this study is to evaluate both algorithms in predicting scholarship candidates with available historical data. Grade point average, parental income, and extracurricular activities are some of the relevant features used in the dataset. Utilizing the metrics of accuracy, precision, recall, and F1 score, the performance of both algorithms was evaluated. The experimental results showed that both algorithms could provide fairly accurate predictions; the C4.5 algorithm showed superiority in interpreting the classification results, while the SVM showed superiority in prediction accuracy. The SVM and C4.5 models were tested for the classification of scholarship candidates. The C4.5 model showed better performance in all evaluation metrics compared to the SVM model. Therefore, the C4.5 model is recommended for use in the implementation of a scholarship candidate classification system. However, the SVM model remains relevant as an additional tool in prediction validation. Further development with parameter tuning and exploration of the algorithm algorithm can further improve the performance of this classification system.

Keywords: *SVM, C4.5, Selection, Slassification, Scholarship*

Abstrak

Proses seleksi penerima beasiswa, berbagai kriteria dipertimbangkan secara menyeluruh. Nilai akademik, keadaan ekonomi, dan lainnya adalah beberapa faktor penting yang harus dipertimbangkan. Namun, dengan banyaknya calon yang mendaftar, proses seleksi manual menjadi tidak efisien dan rentan terhadap kesalahan. Tujuan dari penelitian ini untuk mengevaluasi kinerja kedua algoritma dalam memprediksi kandidat penerima beasiswa dengan data historis yang tersedia. Nilai rata-rata, penghasilan orang tua, dan aktivitas ekstrakurikuler adalah beberapa fitur yang relevan yang digunakan dalam dataset. Memanfaatkan metrik akurasi, presisi, recall, dan skor F1, kinerja kedua algoritma dievaluasi. Hasil eksperimen menunjukkan bahwa kedua algoritma mampu memberikan prediksi yang cukup akurat; algoritma C4.5 menunjukkan keunggulan dalam interpretabilitas hasil klasifikasi, sementara SVM menunjukkan keunggulan dalam akurasi prediksi. Model SVM dan C4.5 telah diuji untuk klasifikasi calon penerima beasiswa. Model C4.5 menunjukkan kinerja yang lebih baik dalam semua metrik evaluasi dibandingkan model SVM. Oleh karena itu, model C4.5 disarankan untuk digunakan dalam implementasi sistem klasifikasi calon penerima beasiswa. Namun, model SVM tetap relevan sebagai alat tambahan dalam validasi prediksi. Pengembangan lebih lanjut dengan tuning parameter dan eksplorasi algoritma ensemble dapat lebih meningkatkan kinerja sistem klasifikasi.

Kata kunci: *SVM, C4.5,* Seleksi, Klasifikasi, Beasiswa

1. Pendahuluan

Banyak anak-anak di Indonesia yang memiliki kemampuan belajar yang baik dan ingin mencapai cita-cita. Namun, masalah finansial dan tekanan, terutama yang dialami oleh siswa dan anggota keluarga, menghalangi hal ini. Pemerintah saat ini melakukan banyak hal untuk menyelesaikan masalah keuangan siswa. Untuk membantu siswa, pemerintah membuat program beasiswa. [1]

Beasiswa biasanya membantu mahasiswa yang masih kuliah. Agus Lahinta (2009) menyatakan bahwa Beasiswa adalah jenis dana yang diberikan kepada seseorang dengan tujuan untuk mendukung pendidikan mereka. Pemerintah, perusahaan, atau yayasan dapat memberikan beasiswa. [2]

Menurut Grindle (dalam Mulyadi, 2015) implementasi adalah proses administratif yang umum dan dapat dipelajari pada tingkat program yang spesifik. Tachjan mengatakan "Implementasi kebijakan publik, selain dapat dipahami sebagai salah satu aktivitas administrasi publik sebagai institusi (birokrasi) dalam proses kebijakan publik, juga dapat dipahami sebagai salah satu lapangan studi administrasi publik sebagai ilmu" [3]. Serangkaian langkah-langkah dilakukan oleh banyak actor pembuat kebijakan dikenal sebagai

Diterima: 23-07-2024 | Revisi: 27-08-2024 | Diterbitkan: 30-09-2024 | doi: 10.35134/komtekinfo.v11i3.524

implementasi untuk mencapai tujuan dengan alat yang mendukung dan didasarkan pada aturan [4].

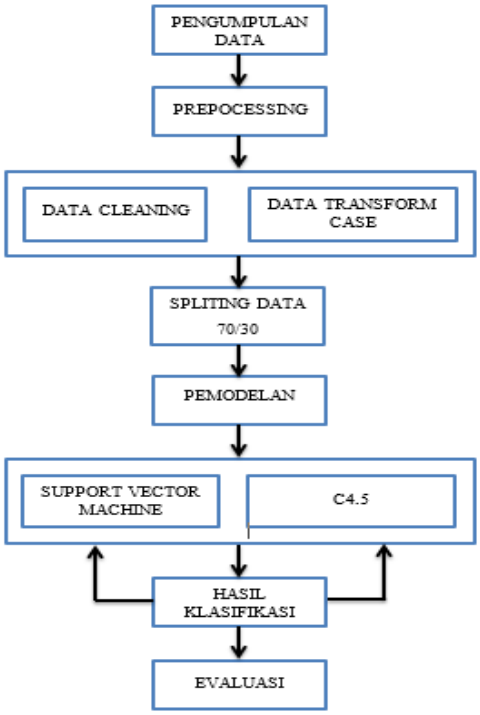
Klasifikasi adalah proses yang menunjukkan suatu objek data sebagai kategori atau kelas. Tujuan dari Proses klasifikasi mencakup pencarian model atau fitur yang dapat memberikan penjelasan atau membedakan ide atau kelas data, dan tujuan akhir dari klasifikasi adalah untuk mengetahui label objek yang kelasnya tidak diketahui. Neural network, pilihan tanaman, formula matematis, atau if-then-rules adalah beberapa contoh model. [5]. Menggabungkan objek atau entitas yang sama dan memisahkan objek atau entitas yang berbeda dikenal sebagai klasifikasi. [6]

S.E. Goodman dan S.T. Hedetniemi menyatakan bahwa algoritma adalah suatu urutan tak terbatas yang dihasilkan oleh operasi yang sudah terdefinisi dengan baik, yang masing-masing hanya memiliki waktu dan penyimpanan yang terbatas [7]. Penelitian yang berkaitan dengan Support Vector Machine dan C4.5 seperti Analisis Sentimen Terhadap Opini Publik yang memiliki hasil pengujian analisis SVM berdasarkan data sampel yang diperoleh hasil analisis klasifikasi dengan tingkat akurasi sebesar 69,69%, recall sebesar 45.60%, precision sebesar 51.56%, dan F1-Score 46% [8]. Selanjutnya Analisis Data Mining Menggunakan Algoritma C4.5 yang memiliki hasil pengujian kinerja algoritma C4.5 menyajikan keluaran analisis prediksi dengan tingkat akurasi sebesar 99.00% [9]. Selanjutnya Klasifikasi Penyakit Diabetes Mellitus menggunakan algoritma C4.5 yang mana hasil dari penelitian ini didapati nilai akurasi sebesar 76% [10].

Permasalahan utama yang akan dibahas adalah bagaimana mengatasi kendala finansial yang dihadapi oleh siswa di Indonesia, yang berdamppak pada pencapaian cita-cita mereka. Untuk mengatasi masalah ini, penelitian ini akan focus pada implementasi program beasiswa yang dilakukan oleh pemerintah, serta bagaimana kebijakan ini dapat dioptimalkan melalui pendekatan kllasifikasi data menggunakan algoritma seperti Support Vector Machine (SVM) dan C4.5. Metode klasifikasi iniakan digunakan untuk menganalisis data siswa yang layak mendapatkan beasiswa, dengan tujuan meningkatkan akurasi dalam penyaluran beasiswa tersebut. Hasil yang diharapkan dari penelitian ini adalah peningkatan efisiennsi dan efektivitas dalam penyaluran beasiswa, sehingga lebih banyak siswa yang dapat terbantu dalam pencapaian cita-cita mereka. Manfaat dari penelitian ini adalah memberikan kontribusi dalam perbaikan kebiijakan beasiswa, yang pada akhirnya dapat membantu meningkatkan akses pendidikan bagi siswa di Indonesia

2. Metodologi Penelitian

Metodologi penelitian dijelaskan pada alur penelitian, yang menggambarkan urutan langkah-langkah sistematis yang harus diikuti selama proses penelitian. Kerangka ini berfungsi sebagai panduan yang memastikan setiap tahap penelitian berjalan sesuai dengan tujuan yang ditetapkan. Kerangka penelitian ini membantu dalam menjaga konsistensi dan integritas proses penelitian, serta memudahkan peneliti untuk mengevaluasi dan merevisi langkah-langkah yang telah diambil apabisa diperlukan. Alur penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

Alur penelitian diatas menunjuukan proses sistematis dalam pengolahan data untuk klasifikasi menggunakan algoritma SVM dan C4.5 dimulai dari pengumpulan data, proses ini melibatkan beberapa tahap penting seperti prepoceessing, pembersihan data (data cleaning), transformasi data, dan pembagian data menjadi set pelatihan dan pengujian (70 : 30).. setelah data diproses, tahap pemodelan dilakukan dengan menerapkan algoritma Support Vector Machine (SVM) dan C4.5 untuk menghasilkan hasil klasifikasi yang kemudian dievaluasi untuk menentukan akurasi model yang digunakan.

2.1. Pengumpulan Data

Pengumpulan data merupakan suatu proses sistematis untuk mengumpulkan informasi atau fakta dari berbagai sumber yang relevan guna menjawab pertanyaan penelitian, menguji hipotesis, atau membuat keputusan berdasarkan bukti yang ada. model pengumpulan data dapat bervariasi, termasuk survey, wawancara, observasi, eksperimen, dan pengambilan data, tergantung pada data yang

dibutuhkan dan tujuan dari penelitian atau analisis. Sumber data adalah bagian penting dari analisisdata. Proses mengumpulkan data kemudian dan memilah-milah data berdasarkan tema, konsep, dan kategori tertentu dikenal sebagai mengurangi data [11].

2.2. Prepocessing

Preprocessing adalah tindakan penting dalam TextMining, Natural Language Processing (NLP), dan Information Retrieval (IR). Dalam TextMining, data preprocessing digunakan supaya memperoleh informasi yang menarik dan signifikan tentang data text yang tidak terorganisir. Retrieval Informasi (IR) memilih dokumen mana yang harus dikeluarkan dari koleksi untuk memenuhi persyaratan informasi pengguna.[12]. Preprocessing adalah proses mengolah data mentah sebelum memulai proses lainnya. Dalam proses ini, sistem mengubah atau menghilangkan data yang tidak sesuai [13].

2.2.1. Data Cleaning (Pembersihan Data)

Pembersihan data adalah salah satu teknik preprocessing yang digunakan untuk mengidentifikasi dan memperbaiki kesalahan dan inkonsistensi agar data menjadi akurat. Sering terjadi kesalahan eja, kesalahan format, data yang hilang atau tidak valid. Dalam kebanyakan kasus, ini disebabkan oleh input data yang salah atau berlebihan dari sumbernya; ini biasanya disebut sebagai database atau Big Data [14]. Mendeteksi, memperbaiki, atau bahkan menghapus catatan, tabel, dan database yang salah atau tidak akurat adalah proses yang dikenal sebagai pembersihan data [15].

2.2.2. Data Transform Case

Transformasi data adalah langkah penting dalam pra-pemrosesan yang memastikan data berada dalam format yang sesuai untuk analisis lebih lanjut. Transformasi yang tepat membantu model pembelajaran mesin untuk memahami pola dalam data dengan lebih efektif dan meningkatkan akurasi prediksi. Salah satu tahapan preprocessing yang paling umum adalah transformasi case. Tahapan ini sering digunakan, terutama untuk parameter TF-IDF yang ada di Rapidminer. Pada titik ini, kalimat dalam dokumen diubah menjadi huruf kecil atau huruf bawah [16].

2.3. Spliting Data

Dalam proses pengolahan data, penting untuk memastikan bahwa data yang digunakan untuk pelatihan dan pengujian diacak secara adil untuk menghindari bias. Pembagian data yang tepat membantu dalam menghasilkan model yang lebih generalis dan mampu bekerja baik pada data baru. Pada tahap splitting data kita menggabungkan data menjadi dua set yang berbeda: satu untuk melatih model (70% dari data) dan satu lagi untuk menguji model (30% dari data).

2.4. Pemodelan

Pemodelan adalah proses representasi atau penyederhanaan dari suatu sistem nyata ke dalam bentuk model yang lebih mudah dipahami, dianalisis, atau dimanipulasi. Model ini bisa berupa matematis, fisik, atau berbasis komputer dan dugunakan untuk memprediksi, memanipulasikan, atau menguji berbagai skenario dalam sistem yang diwakili, sehingga membantu dalam mengambil keputusan lebih lanjut. Menurut Adipraja & Sulistyo (2018), proses membuat atau membangun model dari suatu sistim nyata dalam bahasa formal tertentu dikenal sebagai permodelan [17].

2.4.1. Support Vector Machine

Vapnik pertama kali membuat Support Vector Machine pada tahun 1992, menggabungkan ide-ide terhebat dalam pengidentifikasian pola [18]. Teori optimalisasi menyatakan, Support Vector Machines (SVM) dapat menangani masalah yang tidak linier karena mereka memiliki keterampilan dalam fungsi linier menggunakan algoritma pembelajaran untuk fitur multidimensi [13]. Selain itu, SVM dapat memasukkan konsep trik kernel ke dalam ruang kerja dimensi [14].

SVM sangat penting dalam teori pembelajaran mesin. Mereka juga sangat efektif dalam banyak bidang sains dan teknik, terutama berkaitan dengan problem dengan klasifikasi (identifikasi pola) [17]. Dalam contoh ini, m sampel pengamatan (mengatur pelatihan), (xi, yi), dengan i = 1, 2,..., dan m sampel pengamatan, maka:

$$x_i^T = (xi1, \dots xid) \in R^d \quad (2)$$

Dalam kasus dimana sampel x_i^T diberikan kelas positif, yi mendapat +1, sedangkan ketika sampel xi diberikan kelas negatif, yi mendapat -1. Menggunakan hyperplane wTxi + b = 0, di mana vektor bobot w dan predisposition b, set instruksi ini dapat dipecahkan. Pada (3) dan (4), persamaan hyperplane minimal H1 dan H2 ditunjukkan.

$$H_1: (w^T xi + b) = 1 \quad (3)$$

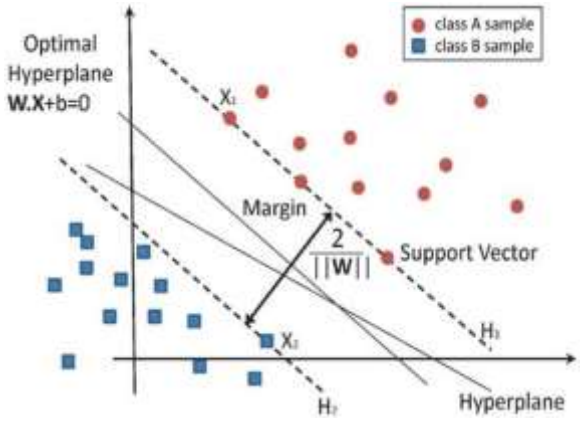
$$H_2: (w^T xi + b) = -1 \quad (4)$$

Persamaan 3-4 merupakan persamaan yang digunakan untuk melihat ketidaksamaan dipenuhi oleh titik yang diklasifikasikan secara akurat. Dalam hal ini, y_i adalah label kelas dari titik data tersebut, w adalah vektor bobot, xi adalah vector fitur, dan b adalah bias yang

menentukan margin dalam model Support Vector Machine (SVM). Ketidaksamaan ini memastikan bahwa titik data berada di sisi yang benar dari hyperlane dengan jarak minimal 1 dari margin, menjamin klasifikasi yang akurat.

$$y_i: (w^T xi + b) \geq 1 \quad (5)$$

Persamaan 5 menyatakan xi, i=1,2,...,m, jarak yang membuat perbedaan hyperplanes negligible sebanding dengan.$\frac{2}{\|w\|}$. Gambar 3 menunjukkan sisi tes vektor pendukung yang menentukan batas instruksi yang ada di hyperplanes H1 atau H2. Setiap vektor pendukunng memiliki peran krusial dalam menentukan margin maksimum antara dua kelas, memastikan bahwa data dipisahkan dengan optimal oleh hyperplane yang dihasilkan.



Gambar 2. Klasifikasi Support Vector Machine

Penggunaan bagian Straight, Spiral Premise Work (RBF), dan Polynomial dalam analisis Support Vector Machine (SVM) sangat berdampak pada parameter dan fungsi bagian yang digunakan [19]. Peggunaan fungsi-fungsi ini memungkinkan SVM untuk menangani data yang tidak dapat dipisahkan secara linier. Fungsi kernel Radial Basis Function (RBF) mampu memetakan data ke dimensi yang lebih tinggu sehingga meningkatkan akurasi klasifikasi. Selain itu fungsi Polynominal dapat menyesuaikan kompleksitas model berdasarkan derajat polynominal yang digunakan, memungkinkan SVM untuk mengatasi pola data yang lebih rumit.

2.4.2. C4.5

Algoritma C4.5 ialah evolusi algoritma dari ID3 sebelumnya. Perkembangannya termasuk kemampuan untuk menangani pruning, kesalahan data, dan data kontinu [20]. Mengatasi kerugian yang hilang, mengatasi data bertipe terus menerus, dan dibuat pemangkasan pohon adalah kelebihan algoritma C4.5 berbeda dengan algoritma ID3 [21]. Peneliti sering mengenakan algoritma berkode C4.5 untuk klasifikasi. Perhitungan algoritma berkode C4.5 menghasilkan pohon keputusan [22].

3. Hasil dan Pembahasan

3.1. Pengumpulan Data

Studi dataset ini berjumlah total 1042 data mengenai calon penerima beasiswa yang dikumpulkan melalui dataset terbuka Kaggle (kaggle.com). Dataset ini mengandung berbagai variable yang mencakup informasi akademis dan non-akademis yang relevan dengan proses seleksi beasiswa. Salah satu karakteristik dataset ini adalah Nama Lengkap, Prodi, Jenis Kelamin, Jarak Tempat Tinggal kekampus (Km), Asal Sekolah, Tahun Lulus, SKS, Ikut Organisasi, Ikut UKM, IPK, Pekerjaan Orang Tua, Penghasilan, Tanggungan, Status Beasiswa. Penelitian ini untuk menemukan pola-pola signifikan dalam data yang dapat digunakan sebagai dasar dalam mengambil keputusan untuk pemberian beasiswa. Berikut merupakan sampel dataset calon penerima beasiswa.

Tabel 1. Sampel Dataset Calon Penerima Beasiswa

Nama Lengkap	Prodi	Jenis Kelamin	Jarak	Asal Sekolah	Tahun Lulus	SKS	Ikut Organisasi	Ikut UKM	IPK	Pekerjaan Orang Tua	Penghasilan	Tanggungan	Status Beasiswa
GALAN PRASETIO	Bimbingan dan	L	Dekat	SMAN 1 GEDONG TATAAN	2020	21	Ikut	Ikut	3.57	Wiraswasta	Sedang	4	Terima
FINGKY RANDIANSYAH	Bimbingan dan	L	Dekat	SMK HAMPAR BAIDURI	2020	21	Tidak	Ikut	2.95	Buruh	Sedang	2	Tidak
ADELIA PANE	Bimbingan dan	P	Dekat	SMK HAMPAR BAIDURI	2020	21	Ikut	Ikut	3.67	Petani	Sedang	4	Terima
DWI HANDOKO	Bimbingan dan	L	Dekat	SMA MA'ARIF NU	2020	21	Tidak	Ikut	3.19	Wiraswasta	Tinggi	2	Tidak
DESTRI FERAWANTI GUSTINI	Bimbingan dan	P	Jauh	SMA 2 NEGERI AGUNG	2020	21	Tidak	Ikut	3.19	Wiraswasta	Sedang	2	Tidak

Gambar 3 merupakan Sampel Dataset Calon Penerima Beasiswa yang akan digunakan dalam penelitian ini. Data sampel yang digunakan sebanyak 5 sampel. Data tersebut digunakan untuk mengklasifikasi data calon penerima beasiswa menggunakan algoritma SVM dan C4.5.

3.2. Prepocessing

3.2.1. Cleaning Data (Pembersihan Data)

Pada tahap Cleaning Data tabel sampel akan di lakukan pembersihan data. Menghapus kolom yang tidak relevan, mengonversi kategori menjadi numerik, dan memastikan tidak ada nilai NaN atau inf. Dari kumpulan data yang diperoleh tidak ada data yang perlu dibersihkan karna informasi data sebanyak 1042 data tidak ada yang mengandung data nan atau inf terdapat pada Gambar 3.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1042 entries, 0 to 1041
Data columns (total 15 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   No                                  1042 non-null   int64
 1   Nama Lengkap                        1042 non-null   object
 2   Prodi                               1042 non-null   object
 3   Jenis Kelamin                       1042 non-null   int64
 4   Jarak Tempat Tinggal kekampus (Km)  1042 non-null   int64
 5   Asal Sekolah                        1042 non-null   object
 6   Tahun Lulus                         1042 non-null   int64
 7   SKS                                 1042 non-null   int64
 8   Ikut Organisasi                     1042 non-null   int64
 9   Ikut UKM                            1042 non-null   int64
 10  IPK                                 1042 non-null   float64
 11  Pekerjaan Orang Tua                 1042 non-null   int16
 12  Penghasilan                         1042 non-null   int64
 13  Tanggungan                          1042 non-null   int64
 14  Status Beasiswa                     1042 non-null   int64
dtypes: float64(1), int16(1), int64(10), object(3)
memory usage: 116.1+ KB
None
```

Gambar 3. Struktur kerangka data yang sudah dibersihkan

Gambar 3 menunjukkan informasi sebuah dataframe dalam pustaka ‘pandas’ pada Phyton. Dataframe diatas memiliki 15 kolom dan mencakup 1042 baris, dengan berbagai tipe data seperti ‘int64’, ‘float64’, dan ‘object’. Kolom-kolom tersebut mencakup informasi seperti nomor, nama lengkap, prodi, jenis kelamin, jarak tempat tinggal kekampus (Km), asal sekolah, tahun lulus, SKS, ikut organisasi, ikut UKM, IPK, pekerjaan orang tua, penghasilan, tanggungan dan status beasiswa.

3.2.2. Data Transform Case

Transformasi data adalah metode perubahan atau mengubah data untuk mengevaluasi atau mensimulasikan kasus atau masalah tertentu. Ini dapat mencakup berbagai proses penghitungan statistik baru, normalisasi, penggantian nilai, atau penggabungan.

Tabel 2 tersebut berisi data dengan beberapa variabel penting yang dapat digunakan untuk analisis lebih lanjut. Misalnya, klasifikasi untuk penerimaan beasiswa adalah salah satu contohnya. Untuk variabel "Jenis Kelamin", nilai 1 diberikan kepada laki-laki dan nilai 0 diberikan kepada perempuan. "Jarak Rumah ke Kampus (Km)" menunjukkan jarak dalam kilometer antara rumah mahasiswa dan kampus. Tabel ini menunjukkan tahun 2020 sebagai "Tahun Lulus".

Setiap siswa menerima 21 SKS, menurut kolon "SKS". 1 menunjukkan partisipasi dalam organisasi dan UKM dan 0 menunjukkan partisipasi tidak. "IPK" adalah Indeks Prestasi Kumulatif yang menunjukkan prestasi akademik siswa, dengan nilai 2,95–3,67. Kode tertentu digunakan untuk menunjukkan pekerjaan orang tua, dan "Penghasilan Orang Tua" menunjukkan pendapatan. Terakhir, jumlah ditunjukkan dalam "Tanggungan".

Tabel 2. Sampel Data setelah dilakukan transform case

Jenis Kelamin	Jarak Tempat Tinggal ke Kampus (Km)	Tahun Lulus	SKS	Ikut Organisasi	Ikut UKM	IPK	Pekerjaan Orang Tua	Penghasilan	Tanggungan
1	1	2020	21	1	1	3.57	169	1	4
1	1	2020	21	0	1	2.95	30	1	2
0	1	2020	21	0	1	3.67	104	1	4
1	1	2020	21	0	1	3.19	169	2	2
0	0	2020	21	0	1	3.19	16	1	2

3.3.Spliting Data

Pembagian data berfungsi memisahkan fitur dan label, serta membagi data menjadi training dan testing set dengan rasio 70:30. Pembagian ini penting untuk memastikan model dilatih dengan data yang memadai dan dapat diuji engan data yang tidak pernah dulihat sebelumnya. Dengan demikian, model dapat diukur performnya secara akurat dan menghindari overfitting disajikan pada Gambar 4

```
# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
print("Training and testing data split successfully.")
```

```
Training and testing data split successfully.
```

Gambar 4. Teknik Spliting Data

Potongan kode Phyton yang digunakan pada Gambar 4 digunakan untuk membagi data menjadi set pelatihan dan set pengujian. Kode tersebut menggunakan fungsi ‘train_test_split’ dari pustaka ‘sklearn.model_selection’ untuk memisahkan fitur (X) dan label (Y) menjadi dua bagian :data pelatihan dan data pengujian. Dalam hal ini, data dibagi dengan rasio 70 : 30, yang berarti 70% dari data digunakan untuk melatih model, sementara 30% sisanya digunakan untuk menguji kinerja model. Argumen ‘random_state=42’ digunakan untuk memastikan bahwa pemisahan data ini dapat direproduksi, artinya hasil pemisahan yang sama akan diperoleh setiap kali kode dijalankan. Pesan “Training and testing data split successfully.” Di tampilkan setelah pemisah selesai, menandakan bahwa proses berjalan dengan sukses.

3.4. Pemodelan

Pemodelan adalah proses membuat model, representasi, atau ide sistem untuk pemahaman, penjelasan, prediksi, atau desain. Dalam pemodelan ini, model SVM dan C4.5 dilatih menggunakan set pelatihan. Setelah dilatih, model SVM dan C4.5 digunakan untuk mengklasifikasi data baru berdasarkan pola yang telah dipelajari. Hasil dari klasifikasi tersebut kemudian dievaluasi untuk mengukur kinerja model dalam menkategorikan data.

3.4.1. Support Vector Machine

Model SVC berhasil dilatih menggunakan data pelatihan. Hasil dari proses pelatihan ini adalah model yang siap untuk digunakan dapal memprediksi label pada data uji atau data baru. Efektivitas model ini kemudian dapat dievaluasi dengan mengukur kinerja prediksinya, seperti akurasi atau numerik lainnya, berdasarkan data yang tidak termasuk dalam pelatihan terdapat pada Gambar 5.

```
svm_model = SVC(kernel='rbf', C=1, gamma='auto')
svm_model.fit(X_train_data, y_train)
```

```
SVC
SVC(C=1, gamma='auto')
```

Gambar 5. pemodelan algoritma Support Vector Machine(SVM)

Gambar 5 merupakan sajian algoritma Support Vector Classifier (SVC) dari pustaka *scikit-learn.* Kode tersebut mendefinisikan model SVC dengan kernel RBF, dimana parameter ‘c’ diatur ke nilai 1 untuk menentukan tingkat regulasi, dan ‘gamma’ diatur ke ‘auto’ untuk menentukan seberapa jauh pengaruh dari satu titik data. Model ini kemudian dilatih pada data pelatihan ‘X_train_data’ dan label ‘y_train’, yang berarti model sdang mempelajari pola dari data tersebut untuk membuat prediksi pada data baru di masa depan.

3.4.2. C4.5 (Decission Tree)

Decision Tree berhasil di buat dengan kriteria pemisah menggunakan entropi. Model ini kemudian dilatih menggunakan dataset pelatihan ‘X_train’ dan ‘y_train’. Hasil dari proses ini merupakan model yang sudah siap digunakan dalam memprediksi label pada data baru berdasarkan pola yang dipelajari dari data pelatihan tersaji pada Gambar 6.

```
c45_model = DecisionTreeClassifier(criterion='entropy')
c45_model.fit(X_train, y_train)
```

```
DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy')
```

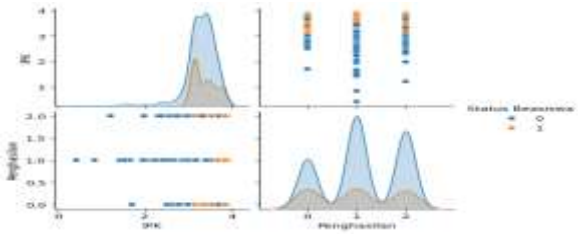
Gambar 6. Pemodelan C,45 (Decision Tree)

Gambar 6 tersebut menunjukkan implementasi algoritma Decission Tree menggunakan library scikit-learn di Python. Kode tersebut membuat model Decision Tree dengan kriteria pemisah berdasarkan entropi. Model tersebut kemudian dilatih menggunakan data pelatihan yang diberikan (‘X_train’ dan ‘y_train’)

3.5. Hasil Klasifikasi

3.5.1. Hasil Klasifikasi Support Vector Machine

Hasil Klasifikasi menampilkan visualisasi hubungan antara IPK dan Penghasilan dengan status beasiswa. Pada scatter plot, terlihat bahwa distribusi IPK dan Penghasilan dipisahkan berdasarkan status beasiswa (0 untuk tidak menerima beasiswa dan 1 untuk menerima beasiswa). Diagonal plot menunjukkan distribusi atau sebaran nilai IPK dan Penghasilan, dimana terdapat beberapa puncak yang menandakan adanya variasi dalam data terdapat pada Gambar 7.



Gambar 7. Hasil Klasifikasi Support Vector Machine

Histogram dan KDE Plot untuk "IPK" dan "Penghasilan" IPK (Grafik di diagonal kiri atas): Histogram dan plot Kernel Density Estimation (KDE) menunjukkan distribusi nilai IPK. Sebagian besar data IPK berada di kisaran 3.5 hingga 4.0, dengan distribusi yang mirip untuk kedua kategori status beasiswa (0 dan 1).

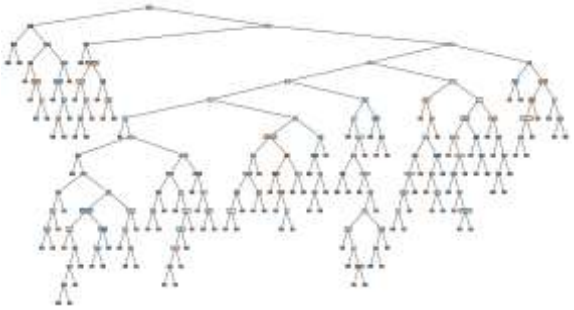
Penghasilan (Grafik di diagonal kanan bawah): Histogram dan plot KDE menunjukkan distribusi nilai penghasilan. Ada beberapa puncak dalam distribusi penghasilan, menunjukkan adanya beberapa kelompok atau klasifikasi penghasilan.

Scatter Plot antara "IPK" dan "Penghasilan" IPK vs. Penghasilan (Grafik di luar diagonal): Scatter plot menunjukkan hubungan antara IPK dan Penghasilan untuk kedua kategori status beasiswa (0 dan 1). Dari plot ini, terlihat bahwa meskipun ada variasi dalam data, tidak ada hubungan yang sangat jelas antara IPK dan Penghasilan dalam menentukan status beasiswa.

3.5.2. Hasil Klasifikasi C4.5

Hasil Klasifikasi C4.5 menunjukkan visualisasi dari sebuah pohon keputusan (Decision Tree) yang digunakan dalam proses klasifikasi. Setiap node pada pohon ini mempresentasikan suatu kondisi atau fitur yang digunakan untuk memisahkan data ke dalam dua

atau lebih kelompok. Cabang-vabang yang menggambarkan jalur pengambilan keputusan berdasarkan kondisi tersebut, hingga mencapai nide daun yang menunjukkan keputusan akhir atau hasil klasifikasi terdapat pada Gambar 8.



Gambar 8. Hasil Klasifikasi C4.5

Akar Pohon (Root Node) adalah simpul pertama di bagian atas grafik. Ini adalah fitur pertama yang digunakan untuk membagi data. Dalam konteks ini, fitur pertama yang dipilih mungkin adalah variabel yang paling penting atau yang memiliki kekuatan prediksi terbesar untuk klasifikasi status beasiswa. Cabang (Branches), Setiap cabang mewakili hasil dari tes pada simpul (node) tertentu, memisahkan dataset ke dalam subset yang lebih kecil. Misalnya, cabang mungkin menunjukkan apakah nilai IPK di atas atau di bawah suatu titik. Simpul Daun (Leaf Nodes) adalah titik akhir dari pohon dan menunjukkan prediksi akhir untuk subset data tersebut. Simpul daun berisi label kelas (0 atau 1 dalam hal ini) dan mungkin juga menunjukkan jumlah atau proporsi data dalam subset tersebut yang masuk dalam kategori itu.

Simpul Internal (Internal Nodes) adalah simpul di antara akar dan daun yang melakukan pengujian pada satu atau lebih fitur dan menentukan arah cabang berikutnya. Warna dan Teks, Warna pada simpul mungkin menunjukkan distribusi dari kategori yang ada di simpul tersebut, dengan warna yang berbeda untuk masing-masing status beasiswa (0 atau 1). Teks di dalam setiap simpul biasanya menunjukkan kondisi tes, jumlah sampel yang mencapai simpul tersebut, dan distribusi kelas dari sampel tersebut.

3.6. Hasi Evaluasi

Menilai sebuah kinerja model yang mana dengan metrik akurasi, presisi, recall., skor F1, dan confusion matrix. Matrix-matrix ini membantu dalam mengidentifikasi kekuatan dan kelemahan model, serta menentukan apakah model tersebut sudah cukup baik untuk di gunakan. Interpretasi yang tepat dari matrix ini sangat penting untuk mengambil keputusan yang lebih baik dalam pengembangan dan penerapan model.

3.6.1. Accurasi Pemodelan Support Vector Machine

Pada accurassi pemodelan Support Vector Machine menampilkan hasil evaluasi kinerja model Support Vector Machine (SVM) dalam melakukan prediksi pada data uji. Terdapat empat metric utama yang ditampilkan, yaitu Accuracy, Precision, Recall, dan F1 Score. Hasil menunjukkan bahwa model SVM memiliki tingkat acurasi sebesar 75.7%, namun nilai recall yang lebih rendahh, yaitu 32.5%, menunjukkan bahwa model ini tidak terlalu baik dalam mengidentifikasi semua kasus positif secara tepat tersaji pada Gambar 9.

```
y_pred_svm = svm_model.predict(X_test_data)
print("SVM Accuracy:", accuracy_score(y_test, y_pred_svm))
print("SVM Precision:", precision_score(y_test, y_pred_svm))
print("SVM Recall:", recall_score(y_test, y_pred_svm))
print("SVM F1 Score:", f1_score(y_test, y_pred_svm))
```

```
SVM Accuracy: 0.7571884984025559
SVM Precision: 0.574468085106383
SVM Recall: 0.3253012048192771
SVM F1 Score: 0.41538461538461535
```

Gambar 9. Hasil Accurasi pemodelan Support Vector Machine(SVM)

Secara keseluruhan, model SVM menunjukkan tingkat akurasi yang cukup baik, nilai precision dan recall yang relative rendah mengidentifikasi kelas tertentu. Hal ini tercermin dalam nilai F1 Score yang hanya mencapai 41.5%, menunjukkan bahwa model ini kurang optimal dalam menangani trade-off antara precision dan recall. Untuk meningkatkan performa model, ungkin diperlukan penyesuaian hyperparameter atau penerapan baik teknik penyeimbang data.

3.6.2. Accurasi Pemodelan C4.5(Decission Tree)

Hasil evaluasi model C4.5 menunjukkan kinerja cukup baik pada data uji. Dari metrik yang ditampilkan, kita dapat melihat bahwa akaurasi model C4.5 mencapai sekitar 78%, yang menunjukkan tingkat keakuratan prediksi yang tinggi. Precision model ini berada di angka 60%, yang berarti model ini cukup andal dalam mengidentifikasi positif dengan benar, meskipun ada ruang untuk peningkatan tersaji pada Gambar 10.

```
y_pred_c45 = c45_model.predict(X_test)
print("C4.5 Accuracy:", accuracy_score(y_test, y_pred_c45))
print("C4.5 Precision:", precision_score(y_test, y_pred_c45))
print("C4.5 Recall:", recall_score(y_test, y_pred_c45))
print("C4.5 F1 Score:", f1_score(y_test, y_pred_c45))
```

```
C4.5 Accuracy: 0.7827476038338658
C4.5 Precision: 0.6027397260273972
C4.5 Recall: 0.5301204819277109
C4.5 F1 Score: 0.564102564102564
```

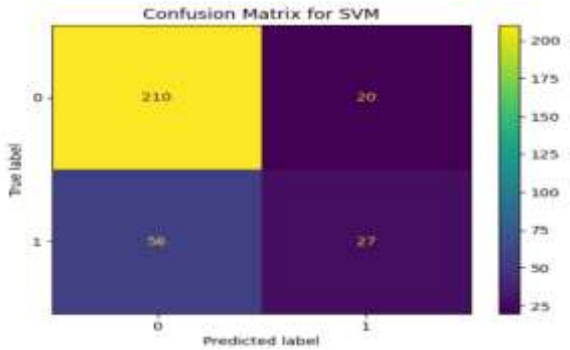
Gambar 10. Hasil Accurasi pemodelan C.45 (Decision Tree)

Berdasarkan gambar diatas, meskipu model C4.5 memiliki aurasi yag cukup tinggi sebesar 78%, nilai recall yang hanya mencapai sekitar 53% menunjukkan bahwa model ini cenderung melewatkan beberapa

kasus positif. Nilai F1 Score sebesar 56% mengindikasi adanya ketidak seimbangan antara precision dan recall, yang berarti bahwa model ini masih memiliki ruang untuk perbaikan dalam hal keseimbangan antara sensitivitas dan spesifisitas. Secara keseluruhan, konerja model C4.5 cukup baik.

3.6.3. Confusion Matrix Support Vector Machine

Confusion matrix berikut menunjukkan performa model Support Vector Machine (SVM) dala mengklasifikasi data uji, dari matriks ini, kita melihat bahwa model SVM berhasil mengklasifikasi 210 sampel benar-benar negative dan benar, sementara 20 sampel negative salah diklasifikasikan sebagai positif. Sebaliknya, terdapat 56 sampel positif yang salah diklasifikasikan sebagai negative, dan hanya 27 sampel positif yang diklasifikasikan dengan benar oleh model SVM tersaji pada Gambar 11.

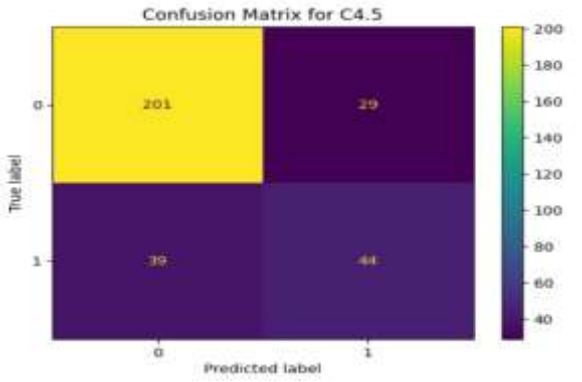


Gambar 11. Hasil Confusion Matrix Support Vector Machine

Berdasarkan gambar Confusion Matrix di atas, meskipun model SVM memiliki kemampuan yang baik dalam mengidentifikasi kelas negatif dengan benar, hal ini tidak diimbangi dengan kemampuannya untuk mengidentifikasi kelas positif, yang terlihat dari 56 sampel positif yang salah diklasifikasikan sebagai negative. Rasio kesalahan klasifikasi yang lebih tinggi pada kelas positif menunjukkan bahwa model SVM ini mungkin kurang sensitif dalam mendeteksi sampel yang sebenarnya positif. Untuk meningkatkan performa model SVM, perlu dilakukan lebih lanjut atau mungkin diperlukan penggunaan teknik lain yang lebih sesuai untuk mendeteksi kelas positif secara lebih akurat.

3.6.4. Confusion Matrix C4.5 (Decision Tree)

Confusion matrix berikut menggambarkan performa model C4.5 dalam klasiifikasi data uji. Dari matriks ini, terlihat bahwa model C4.5 berhasil mengklasifikasi 201 sampel negatif dengan benar, namun terdapat 29 sampel negatif yang salah diklasifikasikan sebagai positif. Selain iyu, model C4.5juga mampu mengidentifikasi 44 sampel positif dengan benar, meskipun masih terdapat 39 sampel positif yang salah diklasifikasikan sebagai negatif tersaji pada Gambar 12.



Gambar 12. Hasil Confusion Matrix C.45 (Decision Tree)

Berdasarkan gambar diatas model C4.5 menunjukkan kemampuan yang cukup baik dalam meng-identifikasikan sampel negatif dengan benar, meskipun masih ada beberapa kesalahan dalam mengidentifikasi sampel positif. Dengan 39 sampel positif yang salah diklasifikasikan sebagai negatif, model ini menunjukkan adanya ruang untuk peningkatan dalam sensitivitas terhadap kelas positif. Meskipun demikian, performa keseluruhan model C4.5 cukup seimbang antara deteksi kelas positif dan negatif, tetapi perbaikan lebih lanjut dapat dilakukan untuk meningkatkan akurasi dalam mendeteksi sampel positif.

4. Kesimpulan

Berdasarkan analisis klasifikasi calon penerima beasiswa menggunakan algoritma SVM dan C4.5, diperoleh performa seperti akurasi model C4.5 mencapai akurasi sebesar 78.27%, yang lebih tinggi dibandingkan model SVM yaitu sebesar 75.72%. Ini menunjukkan bahwa model C4.5 lebih sering membuat prediksi yang benar dibandingkan model SVM. Presisi model C4.5 adalah 60.27%, menunjukkan bahwa model ini cukup efektif dalam meminimalkan prediksi positif palsu sedangkan model svm hanya 57.45% . Recall model C4.5 adalah 53.01%, jauh lebih tinggi dibandingkan model SVM yaitu 32.53%. Ini berarti model C4.5 lebih baik dalam menangkap semua kasus positif yang sebenarnya, yang sangat penting dalam konteks penerimaan beasiswa. F1 Score dari model C4.5 adalah 56.41%, menggambar keseimbangan yang lebih di antara presisi dan recall dibandingkan dengan model SVM 41.54%. Model SVM dan C4.5 telah diuji untuk klasifikasi calon penerima beasiswa. Model C4.5 menunjukkan kinerja yang lebih baik dalam semua metrik evaluasi dibandingkan model SVM. Oleh karena itu, model C4.5 disarankan untuk digunakan dalam implementasi sistem klasifikasi calon penerima beasiswa. Namun, model SVM tetap relevan sebagai alat tambahan dalam validasi prediksi. Pengembangan lebih lanjut dengan tuning parameter dan eksplorasi algoritma ensemble dapat lebih meningkatkan kinerja sistem klasifikasi ini.

Daftar Rujukan

[1] Q. A'yuni, A. Nazir, L. Handayani, and I. Afrianty, "Penerapan Algoritma K-Means Clustering untuk Mengetahui Pola Penerima Beasiswa Bank Indonesia (BI)," *J. Comput. Syst. Informatics*, vol. 4, no. 3, pp. 530–539, 2023.

[2] A. H. Wijaya, P. B. R. Putri, F. S. Bufra, and ..., "Sistem Pendukung Keputusan Penerima Beasiswa Pendidikan Badan Amil Zakat (Baznas) Kabupaten Pesisir Selatan Menggunakan …," *J. Rev. Pendidik. dan Pengajaran*, vol. 6, no. 3, pp. 1–9, 2023.

[3] A. Ningsih, S. Nurhaliza, and E. Priyanti, "Implementasi Sistem Keuangan Desa Dalam Transparansi Pengelolaan Alokasi Dana Desa Di Desa Bulak Kabupaten Indramayu," *J. Gov. Sci. J. Ilmu Pemerintah.*, vol. 3, no. 1, pp. 1–21, 2022.

[4] A. Siregar, E. Yunita, I. Sofia, R. E. Maulina, and T. Y. Hidayatullah, "Implementasi Manajemen Strategik dalam Meningkatkan Manajemen Pendidikan Islam," *J. Pendidik. dan Konseling*, vol. 4, no. 5, p. 5518, 2022.

[5] Y. Partogi and A. Pasaribu, "Perancangan Metode Decision Tree Terhadap Sistem Perpustakaan STMIK Kuwera," *J. Sist. Inf. dan Teknol.*, vol. 1, no. 2, pp. 20–25, 2022.

[6] M. Mayasari, D. Iskandar Mulyana, M. Betty Yel, and S. Tinggi Ilmu Komputer Cipta Karya Informatika Jl Raden, "Komparasi Klasifikasi Jenis Tanaman Rimpang Menggunakan Principal Component Analiysis, Support Vector Machine, K-Nearest Neighbor Dan Decision Tree," *J. Tek. Inform. Kaputama*, vol. 6, no. 2, pp. 644–655, 2022.

[7] M. Rajagukguk, "Implementasi Association Rule Mining Untuk Menentukan Pola Kombinasi Makanan Dengan Algoritma Apriori," *J. Fasilkom*, vol. 10, no. 3, pp. 248–254, 2020.

[8] Ade Dwi Dayani, Yuhandri, and G. Widi Nurcahyo, "Analisis Sentimen Terhadap Opini Publik pada Sosial Media Twitter Menggunakan Metode Support Vector Machine," *J. KomtekInfo*, vol. 11, pp. 1–10, 2024.

[9] S. Mulyanda, S. Defit, and Sumijan, "Analisis Data Mining Menggunakan Algoritma C4.5 Untuk Prediksi Harga Pasar Mobil Bekas," *J. KomtekInfo*, vol. 10, pp. 116–121, 2023.

[10] R. P. Fadhillah, R. Rahma, A. Sepharni, R. Mufidah, B. N. Sari, and A. Pangestu, "Klasifikasi Penyakit Diabetes Mellitus Berdasarkan Faktor-Faktor Penyebab Diabetes menggunakan Algoritma C4.5," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 7, no. 4, pp. 1265–1270, 2022.

[11] A. S. Millah, D. Arobiah, E. S. Febriani, and E. Ramdhani, "Analisis Data dalam Penelitian Tindakan Kelas," vol. 1, no. 2, pp. 140–153, 2023.

[12] L. Hermawan and M. Bellaniar Ismiati, "Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval," *J. Transform.*, vol. 17, no. 2, p. 188, 2020.

[13] S. Parsaoran Tamba, A. Laia, Y. Kristian Butar Butar, and F. Sains dan Teknologi, "Penerapan Data Mining Untuk Klasifikasi Berita Hoax Menggunakan Algoritma Naive Bayes," *J. TEKINKOM*, vol. 6, no. 2, p. 2023, 2023.

[14] T. Z. Dessiaming, S. Anraeni, and S. Pomalingo, "College Academic Data Analysis Using Data Visualization," *J. Tek. Inform.*, vol. 3, no. 5, pp. 1203–1212, 2022.

[15] Baiq Nurul Azmi, Arief Hermawan, and Donny Avianto, "Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 4, no. 4, pp. 281–290, 2023.

[16] N. Q. Rizkina and F. N. Hasan, "Analisis Sentimen Komentar Netizen Terhadap Pembubaran Konser NCT 127 Menggunakan Metode Naive Bayes," *J. Inf. Syst. Res.*, vol. 4, no. 4, pp. 1136–1144, 2023.

[17] H. A. Putra and A. Arista, "Jurnal Comasie PERMODELAN STRUKTUR MATERIAL," vol. 3, pp. 94–101, 2020.

[18] M. Rizki, D. Devrika, I. H. Umam, and F. S. Lubis, "Aplikasi Data Mining dalam Penentuan Layout Swalayan dengan Menggunakan Metode MBA," *J. Tek. Ind. J. Has. Penelit. dan Karya Ilm. dalam Bid. Tek. Ind.*, vol. 5, no. 2, p. 130, 2020.

[19] N. G. Ramadhan and A. Khoirunnisa, "Klasifikasi Data Malaria Menggunakan Metode Support Vector Machine," *J. Media Inform. Budidarma*, vol. 5, no. 4, p. 1580, 2021.

[20] M. A. Wiratama and W. M. Pradnya, "Optimasi Algoritma Data Mining Menggunakan Backward Elimination untuk Klasifikasi Penyakit Diabetes," *J. Nas. Pendidik. Tek. Inform.*, vol. 11, no. 1, p. 1, 2022.

[21] L. Y. L. Gaol, M. Safii, and D. Suhendro, "Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi," vol. 2, no. 2, pp. 97–106, 2021.

[22] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 32–39, 2020.