# EXPLORING COMPLEX DECISION TREES: UNVEILING DATA PATTERNS AND OPTIMAL PREDICTIVE POWER

**Ismail Setiawan[1], Renata Fina Antika Cahyani[2], Irfan Sadida[3]**
Program Studi Sistem dan Teknologi Infomasi, Universitas 'Aisyiyah Surakarta
Jl. Kapulaga No. 3 Laweyan, Surakarta
**e-mail:** *[1]ismailsetiawan@aiska-university.ac.id

***Abstract***
*This research investigates the development and analysis of decision tree models in the context of classification tasks. Decision tree models were developed without employing pruning or pre-pruning techniques and were tested on relevant datasets. The research findings demonstrate that complex models without pruning achieved the highest level of accuracy in classifying data. This study was inspired by the potential issue of students facing the risk of not completing their studies (dropout), which could lead to a decline in the college's accreditation rating. Therefore, this model was devised to assist in identifying factors that could influence this outcome as a preventative measure. Additionally, we successfully generated clear visualizations of the decision trees, enhancing the understanding of the model's decision-making process. This research provides insights into the adaptability of decision tree models within this specific case and showcases their potential for enhancing decision-making across various contexts. These findings encourage further discussions on the benefits of pruning methods within this specific context and the broader application potential of decision tree models.*

***Keyword:*** *Accuracy; Decision tree models; Dropout; Preventative measure; Pruning*

## PENDAHULUAN

Hubungan antara machine learning dan decision tree sangat erat karena decision tree merupakan salah satu algoritma yang digunakan dalam machine learning untuk melakukan tugas klasifikasi dan regresi (X. Xu et al., 2019). Decision tree adalah model prediktif yang menggambarkan serangkaian keputusan dan konsekuensi yang mungkin terjadi berdasarkan fitur-fitur input(He et al., 2022). Algoritma ini bekerja dengan memecah data menjadi subset yang lebih kecil berdasarkan atribut-atribut tertentu dan menghasilkan struktur pohon keputusan. Setiap simpul pada pohon mewakili pengujian terhadap suatu atribut, sedangkan cabang-cabangnya menggambarkan hasil dari pengujian tersebut(Aronov et al., 2023). Melalui proses pembelajaran dari data latih, decision tree dapat mengidentifikasi pola-pola dan hubungan dalam data, sehingga memungkinkan untuk melakukan prediksi atau klasifikasi pada data baru(Cao et al., 2023). Dengan keunggulan interpretabilitas yang tinggi, decision tree sangat berguna dalam pemahaman tentang bagaimana keputusan dibuat oleh model, serta dalam aplikasi di berbagai bidang seperti pengenalan pola, pengolahan citra, keuangan, dan lainnya dalam paradigma machine learning(Moshkov, 2022).

Penelitian mengenai penggunaan decision tree dalam konteks perguruan tinggi telah menjadi topik yang menarik dalam beberapa tahun terakhir (Audemard et al., 2022). Decision tree digunakan untuk menganalisis data terkait mahasiswa, termasuk performa akademik, preferensi kursus, dan faktor-faktor yang memengaruhi kelulusan (Guggari et al., 2022). Dengan memanfaatkan algoritma decision tree, perguruan tinggi dapat mengidentifikasi pola-pola yang berkontribusi terhadap kesuksesan akademik mahasiswa. Selain itu, decision tree juga digunakan untuk membangun model prediksi mengenai penerimaan mahasiswa baru berdasarkan sejumlah faktor seperti nilai ujian masuk, prestasi akademik sebelumnya, dan latar belakang sosial-ekonomi (X. Xu et al., 2019)(Mariano et al., 2022). Hasil dari penelitian ini dapat membantu perguruan tinggi dalam mengambil keputusan yang lebih baik terkait pengelolaan sumber daya, peningkatan retensi mahasiswa, dan pengembangan kurikulum yang sesuai dengan kebutuhan

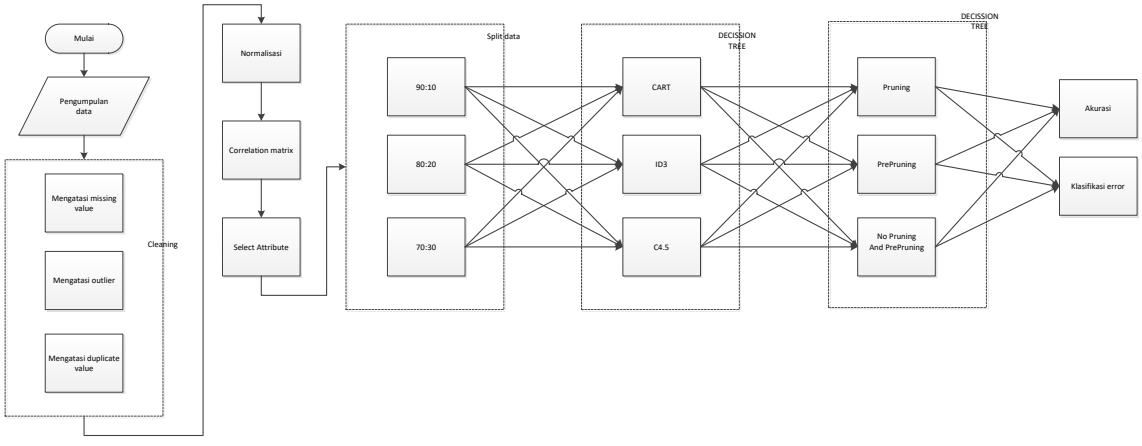
mahasiswa. Dengan memanfaatkan kekuatan interpretatif decision tree, penelitian semacam ini dapat memberikan wawasan yang berharga bagi manajemen perguruan tinggi dalam merencanakan langkah-langkah strategis untuk meningkatkan efisiensi dan efektivitas pendidikan tinggi (Viloria et al., 2019).

Penerapan decision tree dalam penelitian pada konteks perguruan tinggi dapat memberikan solusi yang beragam terhadap sejumlah masalah yang dihadapi (Viloria et al., 2019). Pertama-tama, penggunaan decision tree dalam analisis performa akademik mahasiswa memungkinkan identifikasi faktor-faktor yang berpengaruh terhadap keberhasilan atau kesulitan belajar mahasiswa(Viloria et al., 2019). Dengan menemukan pola-pola ini, perguruan tinggi dapat mengintervensi dengan program mentoring atau bimbingan akademik yang lebih tepat sasaran.

Upaya identifikasi faktor-faktor yang memengaruhi keberhasilan atau kesulitan belajar mahasiswa melibatkan sejumlah masalah yang kompleks (Aning & Przybyła-Kasperek, 2022). Pertama, adanya multi-faktor dan interkoneksi antara berbagai aspek, baik akademis, sosial, ekonomi, maupun psikologis, dapat membingungkan analisis (Moshkov, 2022). Kedua, kualitas data menjadi hal utama, karena keakuratan dan kelengkapan data sangat memengaruhi hasil kesimpulan (X. Xu et al., 2019). Ketiga, dimensi data yang tinggi menghadirkan tantangan visualisasi dan pemodelan yang kompleks (N & K, 2022). Keempat, memahami hubungan sebab-akibat dalam analisis observasional bisa menjadi sulit dan memerlukan perhatian ekstra terhadap faktor-faktor yang tidak terdeteksi (Han et al., 2023). Kelima, perubahan konteks dari waktu ke waktu dapat mempengaruhi dinamika faktor-faktor tersebut (Loyola-González et al., 2023). Keenam, hasil analisis yang berhasil di satu institusi belum tentu dapat diterapkan begitu saja pada institusi lain karena perbedaan karakteristik (Y. Li et al., 2023). Ketujuh, adanya bias pemilihan variabel dan metode analisis serta interpretasi yang subyektif juga harus diperhatikan (Cembranel et al., 2023). Kedelapan, faktor perubahan sosial dan budaya turut memengaruhi relevansi faktor-faktor yang ditemukan (Camerlingo et al., 2022). Kesembilan, walaupun faktor-faktor dapat teridentifikasi, merumuskan intervensi atau tindakan yang tepat memerlukan penelitian lanjutan (M. Li et al., 2022). Dalam menghadapi kompleksitas ini, diperlukan pendekatan hati-hati yang memadukan metodologi analisis yang tepat, pengumpulan data yang akurat, dan pemahaman mendalam tentang konteks institusi dan mahasiswa yang bersangkutan(Gao et al., 2022).

## METODE PENELITIAN

Penelitian ini menggunakan framework CRISP-DM, namun memotong 2 langkah pertama bisnis understanding dan data understanding.




Gambar 1. Metode Penelitian

Proses pengumpulan data adalah tahapan krusial dalam penelitian di mana informasi relevan diperoleh untuk menjawab pertanyaan penelitian. Langkah pertama melibatkan perencanaan tujuan dan metode pengumpulan yang tepat, seperti survei, wawancara, atau observasi. Instrumen pengumpulan data dirancang dan diuji sebelum data dikumpulkan secara substansial. Data kemudian diambil melalui interaksi langsung, pengamatan, atau pengisian kuesioner, diikuti oleh validasi dan pengolahan data. Untuk kasus ini data didapatkan dari Open

University Learning Analytics.

Proses pembersihan data (data cleaning) adalah langkah penting dalam analisis data di mana data yang dikumpulkan dari berbagai sumber dievaluasi, dikoreksi, dan disusun untuk memastikan kualitas, akurasi, dan konsistensi. Langkah-langkah dalam proses pembersihan data melibatkan identifikasi dan penanganan data yang hilang, duplikat, tidak relevan, atau anomali. Data yang rusak atau tidak lengkap diperbaiki atau diisi ulang, duplikat dihapus, dan kesalahan manusia atau perangkat diperbaiki. Tujuannya adalah untuk memastikan bahwa data yang digunakan dalam analisis akurat, andal, dan dapat diandalkan, sehingga hasil analisis yang dihasilkan memberikan pandangan yang jelas dan benar tentang fenomena yang diteliti.

Proses normalisasi data melibatkan pengubahan skala nilai dalam dataset agar seragam, mempermudah analisis, dan meminimalkan efek outlier (Mariano et al., 2022). Metode normalisasi umum termasuk Min-Max Scaling yang mengubah nilai ke rentang 0-1, Z-Score Normalization yang mengubah nilai menjadi distribusi normal dengan rata-rata 0 dan simpangan baku 1, serta Robust Scaling yang menangani outlier dengan menggunakan kuartil. Selain itu, terdapat juga normalisasi L2 untuk vektor, transformasi logaritma untuk mendekati distribusi normal, dan pemisahan data kategorikal dengan One-Hot Encoding. Pemilihan metode normalisasi harus disesuaikan dengan jenis data yang hendak diolah.

Proses pembuatan matriks korelasi melibatkan perhitungan hubungan statistik antara setiap pasangan variabel dalam suatu dataset. Langkah pertama adalah mengumpulkan data numerik yang ingin dianalisis. Kemudian, untuk setiap pasangan variabel, dihitung koefisien korelasi Pearson atau metode lainnya seperti Spearman jika data tidak terdistribusi normal. Koefisien korelasi mengukur tingkat hubungan linier antara variabel, dengan nilai berkisar antara -1 (hubungan terbalik sempurna) hingga 1 (hubungan positif sempurna), serta 0 untuk ketidakhubungan. Hasil perhitungan ini disusun dalam bentuk matriks korelasi, di mana setiap elemen mewakili korelasi antara dua variabel. Matriks korelasi membantu mengidentifikasi pola keterkaitan antara variabel dalam dataset, mendukung pemilihan fitur, dan memberikan wawasan tentang bagaimana variabel berkorelasi satu sama lain.

Proses seleksi atribut dengan metode filter melibatkan penggunaan metrik statistik atau perhitungan heuristik untuk mengevaluasi setiap atribut secara terpisah dari model yang akan digunakan. Atribut dinilai berdasarkan hubungan mereka dengan target atau kelas yang ingin diprediksi, serta hubungan antara atribut-atribut tersebut. Atribut-atribut dinilai dengan metrik seperti nilai p-value, koefisien korelasi, atau skor informasi. Atribut yang memenuhi ambang batas atau skor tertentu akan dipilih untuk membentuk subset atribut yang lebih kecil dan lebih relevan untuk analisis atau pemodelan selanjutnya. Metode ini memungkinkan pemilihan atribut tanpa melibatkan model pembelajaran yang kompleks, tetapi memerlukan pemahaman yang baik tentang data dan konteksnya.

Pembagian data dengan komposisi 70:30, 80:20, dan 90:10 adalah pendekatan umum dalam membagi dataset menjadi subset pelatihan dan pengujian. Dalam pembagian 70:30, 70% data digunakan untuk melatih model, sementara 30% digunakan untuk menguji performa model. Pembagian 80:20 memiliki proporsi serupa, di mana 80% data digunakan untuk pelatihan dan 20% untuk pengujian. Sedangkan, pada pembagian 90:10, dataset pelatihan mencakup 90% data, dan 10% sisanya digunakan untuk pengujian. Pemisahan ini penting untuk menghindari overfitting, memastikan bahwa model dapat generalisasi dengan baik pada data baru, dan memberikan gambaran yang akurat tentang performa model pada data yang tidak dikenal sebelumnya. Pemilihan rasio tergantung pada ukuran dataset, kompleksitas model, dan kebutuhan spesifik analisis atau pemodelan.

Pada tahap pengujian dengan metode pengambilan keputusan CART, ID3, dan C4.5 (algoritma pohon keputusan yang berbeda), dataset akan dipecah menjadi subset pelatihan dan pengujian sesuai dengan rasio yang telah ditentukan. Setiap algoritma akan digunakan untuk membangun model pohon keputusan berdasarkan subset pelatihan. Selanjutnya, model-model tersebut akan diuji menggunakan subset pengujian untuk mengukur performa dan akurasi prediksi mereka. Hasil pengujian akan memberikan wawasan tentang bagaimana masing-masing algoritma berkinerja dalam konteks dataset yang diberikan, serta membantu dalam pemilihan algoritma yang paling sesuai untuk tugas klasifikasi atau pengambilan keputusan tersebut.

Dalam tahap pengujian dengan penerapan teknik pruning, prepruning, dan kombinasi

keduanya pada model pohon keputusan, dataset akan dibagi menjadi subset pelatihan dan pengujian sesuai dengan rasio yang telah ditetapkan. Kemudian, model pohon keputusan akan dibangun dengan menerapkan metode pruning, di mana cabang-cabang yang dianggap kurang signifikan atau kompleks akan dihapus untuk mencegah overfitting. Selanjutnya, metode prepruning akan digunakan saat membangun model pohon, di mana pertumbuhan pohon akan dihentikan jika kriteria pra-ditetapkan (misalnya, batasan kedalaman atau jumlah sampel per daun) tercapai. Pada eksperimen kombinasi keduanya, model akan dibangun dengan menerapkan kedua teknik ini secara bersamaan. Hasil pengujian akan memberikan gambaran tentang bagaimana setiap metode mempengaruhi kinerja model pohon keputusan, serta membantu dalam pemilihan pendekatan terbaik untuk mencegah overfitting dan mengoptimalkan akurasi prediksi.

Pada tahap akhir, kinerja model pohon keputusan yang telah dibangun dengan berbagai metode pruning, prepruning, dan kombinasi keduanya akan dievaluasi berdasarkan dua metrik utama: akurasi dan klasifikasi error. Akurasi mengukur seberapa akurat model dalam melakukan prediksi dengan membandingkan jumlah prediksi benar dengan total prediksi. Sedangkan klasifikasi error mencakup berbagai jenis error seperti false positive, false negative, true positive, dan true negative, memberikan wawasan lebih mendalam tentang jenis kesalahan yang dilakukan oleh model dalam konteks klasifikasi. Dengan memperhatikan kedua metrik ini, kita dapat memahami sejauh mana model mampu membuat prediksi yang benar dan jenis kesalahan yang mungkin terjadi. Ini akan membantu dalam pemilihan strategi optimal untuk membangun model pohon keputusan yang memberikan hasil terbaik sesuai dengan tujuan dan konteks analisis atau pemodelan yang dilakukan.

## HASIL DAN PEMBAHASAN

Dataset "Open University Learning Analytics" yang disumbangkan pada 20 Desember 2015 dan tersedia di UCI Machine Learning Repository memiliki beberapa manfaat penting. Dataset ini menyediakan wawasan tentang perilaku belajar siswa dalam lingkungan pembelajaran online, yang dapat digunakan untuk analisis dan pemodelan dalam bidang pendidikan dan pembelajaran. Informasi tentang interaksi siswa dengan platform pembelajaran, waktu belajar, partisipasi, dan hasil evaluasi memberikan pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi kinerja akademik. Dataset ini memungkinkan peneliti dan praktisi untuk menjalankan analisis statistik dan eksplorasi data untuk mengidentifikasi pola, tren, dan pengaruh yang berpotensi memengaruhi hasil belajar siswa. Selain itu, dataset ini juga memberikan peluang untuk mengembangkan dan menguji model pembelajaran mesin yang dapat meramalkan kinerja siswa, mengidentifikasi faktor risiko, atau memberikan wawasan untuk meningkatkan efektivitas pengajaran dan pembelajaran dalam konteks pendidikan online.

Operator replace missing value dalam RapidMiner adalah alat yang efektif dalam menangani masalah missing value atau nilai yang hilang dalam dataset. Dengan menggunakan operator ini, permasalahan yang timbul akibat nilai-nilai yang tidak lengkap atau kosong dapat diatasi dengan lebih mudah. Operator missing value memungkinkan pengguna untuk memilih metode yang sesuai untuk mengisi atau mengganti nilai yang hilang. Ini termasuk metode statistik seperti rata-rata, median, atau modus, serta metode berdasarkan nilai terdekat atau pola yang ditemukan dalam data. Dengan mengintegrasikan metode ini, operator missing value dapat memproses dan memperbaiki data yang memiliki nilai yang hilang, sehingga meningkatkan kualitas dataset dan mengurangi potensi bias atau ketidakakuratan dalam analisis atau pemodelan selanjutnya. Dengan demikian, operator missing value di RapidMiner membantu pengguna dalam menjaga integritas data dan memastikan bahwa nilai yang hilang tidak menghalangi analisis yang akurat dan hasil yang dapat diandalkan.

Dalam analisis data, data pada baris 369, 611, 892, 1025, 1269, 2512, 2623, 2698, 2975, dan 4163 telah terdeteksi sebagai outlier. Outlier merujuk pada nilai yang signifikan secara statistik berbeda dari nilai-nilai lain dalam dataset. Deteksi outlier penting karena dapat mempengaruhi hasil analisis dan pemodelan secara keseluruhan. Pada data tersebut, nilai-nilai tersebut memiliki karakteristik yang secara signifikan berbeda dari data lainnya, dan mungkin perlu diperiksa lebih lanjut untuk memahami penyebabnya atau menentukan apakah mereka merupakan data yang tidak valid atau hanya representasi langka dari distribusi data yang lebih luas. Lihat gambar 1 untuk lebih jelasnya.

*P-ISSN: 2656-1719*
*E-ISSN: 2656-2774*

| Row No. | Target | outlier ↓ | Marital status | Application ... | Application o... | Course | Daytime/eve... | Previous qu... | Previous qu... |
|---|---|---|---|---|---|---|---|---|---|
| 369 | Graduate | true | -0.295 | -1.011 | -0.554 | 0.483 | 0.350 | -0.350 | -0.805 |
| 611 | Graduate | true | -0.295 | -1.011 | 0.207 | 0.483 | 0.350 | -0.350 | -1.563 |
| 892 | Dropout | true | -0.295 | 1.163 | -0.554 | -4.209 | 0.350 | -0.350 | 1.091 |
| 1025 | Graduate | true | -0.295 | -0.038 | 0.207 | 0.483 | 0.350 | -0.350 | -1.639 |
| 1269 | Enrolled | true | -0.295 | 1.963 | -0.554 | 0.550 | -2.856 | -0.350 | 0.560 |
| 2512 | Enrolled | true | -0.295 | -0.210 | -0.554 | 0.312 | 0.350 | -0.350 | 4.351 |
| 2623 | Graduate | true | -0.295 | 1.849 | -0.554 | -0.408 | -2.856 | 3.663 | 0.560 |
| 2698 | Enrolled | true | -0.295 | -0.095 | -0.554 | 0.394 | 0.350 | -0.350 | -1.336 |
| 2975 | Dropout | true | -0.295 | -1.011 | -0.554 | -4.209 | 0.350 | -0.350 | -0.653 |
| 4163 | Dropout | true | 4.658 | 1.392 | -0.554 | 0.483 | 0.350 | -0.350 | 0.037 |
| 1 | Dropout | false | -0.295 | -0.095 | 2.491 | -4.209 | 0.350 | -0.350 | -0.805 |
| 2 | Graduate | false | -0.295 | -0.210 | -0.554 | 0.193 | 0.350 | -0.350 | 2.077 |
| 3 | Dropout | false | -0.295 | -1.011 | 2.491 | 0.103 | 0.350 | -0.350 | -0.805 |
| 4 | Graduate | false | -0.295 | -0.095 | 0.207 | 0.444 | 0.350 | -0.350 | -0.805 |
| 5 | Graduate | false | 1.356 | 1.163 | -0.554 | -0.408 | -2.856 | -0.350 | -2.473 |
| 6 | Graduate | false | 1.356 | 1.163 | -0.554 | 0.550 | -2.856 | 1.412 | 0.037 |
| 7 | Graduate | false | -0.295 | -1.011 | -0.554 | 0.312 | 0.350 | -0.350 | 0.712 |
| 8 | Dropout | false | -0.295 | -0.038 | 1.729 | 0.193 | 0.350 | -0.350 | -1.032 |

Gambar 2. Nomor baris data yang terdeteksi sebagai outlier

Meskipun nilai duplikat tidak terdeteksi, proses normalisasi tetap dilanjutkan. Normalisasi adalah langkah yang penting dalam analisis data untuk mengubah skala nilai menjadi seragam, yang dapat meningkatkan efektivitas analisis dan pemodelan. Namun, perlu diperhatikan bahwa data duplikat dapat memengaruhi hasil normalisasi dan interpretasi akhir. Sebelum melanjutkan normalisasi, penting untuk memastikan bahwa dataset telah dibersihkan dengan benar dari nilai duplikat dan outlier. Dengan demikian, proses normalisasi dapat berjalan lebih efektif dan akurat, memberikan hasil yang lebih bermakna dalam analisis data Anda.

Setelah proses normalisasi selesai dilakukan pada dataset, langkah selanjutnya adalah melakukan perhitungan matriks korelasi. Proses normalisasi akan mengubah skala nilai dalam dataset sehingga variabel memiliki distribusi yang lebih seragam, memudahkan dalam perbandingan dan analisis. Setelah normalisasi, Anda dapat menghitung matriks korelasi untuk mengevaluasi hubungan statistik antara setiap pasangan variabel. Ini akan memberikan wawasan tentang sejauh mana variabel berhubungan satu sama lain setelah diubah ke dalam skala yang seragam. Dengan demikian, langkah ini membantu dalam mengidentifikasi pola keterkaitan yang mungkin tidak terlihat sebelum normalisasi, memudahkan dalam pengambilan keputusan dan pemodelan berikutnya.

*P-ISSN: 2656-1719*
*E-ISSN: 2656-2774*


Gambar 3.Hasil Pendeteksian Atribut Yang Paling Berpengaruh

Hasil pengujian dilakukan terhadap 4 parameter yaitu penggunaan pruning, Prepruning, tanpa pruning dan prepurning, dan menggunakan pruning dan prepurning secara bersamaan. Dalam pengujian model Decision Tree, kami menganalisis kinerja model untuk tugas klasifikasi yang kami hadapi. Kami mengukur akurasi model untuk memahami sejauh mana model mampu mengklasifikasikan data dengan benar. Selain itu, kami menggunakan matriks confusion untuk mendapatkan wawasan lebih rinci tentang jenis-jenis kesalahan yang dilakukan oleh model, termasuk True Positive, True Negative, False Positive, dan False Negative (J. Xu et al., 2020). Dari matriks ini, kami dapat menghitung metrik seperti presisi, recall, dan F1-score yang memberikan gambaran lebih komprehensif tentang kemampuan model. Selain itu, kami memvisualisasikan pohon keputusan yang dihasilkan oleh model untuk memahami cara model mengambil keputusan berdasarkan fitur-fitur tertentu. Dalam upaya mencegah overfitting, kami juga mempertimbangkan teknik pruning untuk menyederhanakan struktur pohon. Kami juga mengaplikasikan validasi silang (cross-validation) untuk menghindari bias dalam evaluasi kinerja model. Melalui pengujian ini, kami berharap dapat mendapatkan pemahaman yang lebih baik tentang bagaimana model Decision Tree kami berperforma, serta menentukan langkah-langkah perbaikan yang mungkin diperlukan.

Table 1 Memperlihatkan Hasil Performance Algoritma Decision Tree.

| No | Metode | Performance | Classification error |
|---|---|---|---|
| 1 | Pruning | 72.40% | 27.60% |
| 2 | Prepurning | 69.00% | 31.00% |
| 3 | Pruning dan prepurning | 72.40% | 27.60% |
| 4 | Tanpa keduanya | 72.62% | 27.38% |

*P-ISSN: 2656-1719*
*E-ISSN: 2656-2774*

Table 2. Komparasi Hasil Perhitungan Matriks Confusion

| No | Metode | Matriks Confusion | | | | |
|---|---|---|---|---|---|---|
| 1 | Pruning | | true Dropout | true Graduate | true Enrolled | class precision |
| | | pred. Dropout | 107 | 9 | 22 | 77.54% |
| | | pred. Graduate | 34 | 212 | 56 | 70.20% |
| | | pred. Enrolled | 1 | 0 | 1 | 50.00% |
| | | class recall | 75.35% | 95.93% | 1.27% | |
| 2 | Prepruning | | true Dropout | true Graduate | true Enrolled | class precision |
| | | pred. Dropout | 78 | 3 | 12 | 83.87% |
| | | pred. Graduate | 34 | 197 | 37 | 73.51% |
| | | pred. Enrolled | 30 | 21 | 30 | 37.04% |
| | | class recall | 54.93% | 89.14% | 37.97% | |
| 3 | Tanpa pruning dan Prepruning | | true Dropout | true Graduate | true Enrolled | class precision |
| | | pred. Dropout | 114 | 15 | 32 | 70.81% |
| | | pred. Graduate | 21 | 192 | 32 | 78.37% |
| | | pred. Enrolled | 7 | 14 | 15 | 41.67% |
| | | class recall | 80.28% | 86.88% | 18.99% | |
| 4 | Dengan pruning dan Prepruning | | true Dropout | true Graduate | true Enrolled | class precision |
| | | pred. Dropout | 91 | 4 | 13 | 84.26% |
| | | pred. Graduate | 24 | 195 | 32 | 77.69% |
| | | pred. Enrolled | 27 | 22 | 34 | 40.96% |
| | | class recall | 64.08% | 88.24% | 43.04% | |

Setelah menjalani serangkaian kegiatan pengembangan model, kami berhasil mendapatkan hasil menarik. Melalui pengujian yang kami lakukan, kami menemukan bahwa model Decision Tree yang dikembangkan tanpa menggunakan metode pruning dan pre-pruning menghasilkan akurasi tertinggi dibandingkan dengan variasi lainnya (Lazebnik & Bunimovich-Mendrazitsky, 2023)(Zhou et al., 2023). Hasil ini menunjukkan bahwa dalam kasus ini, model yang lebih kompleks dan tidak mengalami penyederhanaan struktural lebih baik dalam memahami pola-pola yang ada dalam data. Meskipun penggunaan pruning dan pre-pruning sering digunakan untuk mengurangi overfitting dan meningkatkan generalisasi model, hasil ini menunjukkan adanya konteks di mana model yang kompleks lebih sesuai. Pengujian ini memberikan wawasan berharga tentang dinamika kinerja model Decision Tree dan dapat membantu panduan keputusan lebih lanjut dalam pengembangan model yang lebih baik. Maka pohon keputusan dapat dibangun dengan mengikuti rumus dibawah ini.

```
Curricular units 2nd sem (approved) > 0.175
|   Tuition fees up to date > 0.500
|   |   Curricular units 1st sem (grade) > 0.534
|   |   |   Curricular units 2nd sem (grade) > 0.564
|   |   |   |   Curricular units 1st sem (approved) > 0.058
|   |   |   |   |   Curricular units 2nd sem (approved) > 0.225
|   |   |   |   |   |   Mother's occupation > 0.997: Enrolled
{Dropout=0, Graduate=0, Enrolled=3}
|   |   |   |   |   |   Mother's occupation ≤ 0.997
|   |   |   |   |   |   |   Father's qualification > 0.895: Dropout
{Dropout=1, Graduate=1, Enrolled=0}
|   |   |   |   |   |   |   Father's qualification ≤ 0.895
|   |   |   |   |   |   |   |   Curricular units 2nd sem (grade) >
0.953: Dropout {Dropout=1, Graduate=1, Enrolled=0}
|   |   |   |   |   |   |   |   Curricular units 2nd sem (grade) ≤
0.953: Graduate {Dropout=168, Graduate=1728, Enrolled=271}
|   |   |   |   |   Curricular units 2nd sem (approved) ≤ 0.225
|   |   |   |   |   |   Curricular units 2nd sem (enrolled) > 0.435:
Dropout {Dropout=2, Graduate=0, Enrolled=0}
|   |   |   |   |   Curricular units 2nd sem (enrolled) ≤ 0.435
|   |   |   |   |   |   Mother's occupation > 0.577: Graduate
{Dropout=0, Graduate=5, Enrolled=0}
|   |   |   |   |   |   Mother's occupation ≤ 0.577
|   |   |   |   |   |   |   Age at enrollment > 0.642: Graduate
{Dropout=0, Graduate=3, Enrolled=0}
|   |   |   |   |   |   |   Age at enrollment ≤ 0.642: Enrolled
{Dropout=57, Graduate=116, Enrolled=120}
|   |   |   |   Curricular units 1st sem (approved) ≤ 0.058
|   |   |   |   |   Application mode > 0.482: Dropout {Dropout=3,
Graduate=1, Enrolled=0}
|   |   |   |   |   Application mode ≤ 0.482: Enrolled {Dropout=0,
Graduate=0, Enrolled=2}
|   |   |   Curricular units 2nd sem (grade) ≤ 0.564
|   |   |   |   Application order > 0.278
|   |   |   |   |   Course > 0.925: Dropout {Dropout=4, Graduate=0,
Enrolled=0}
|   |   |   |   |   Course ≤ 0.925: Graduate {Dropout=0, Graduate=3,
Enrolled=0}
|   |   |   |   Application order ≤ 0.278: Enrolled {Dropout=3,
Graduate=3, Enrolled=17}
|   |   Curricular units 1st sem (grade) ≤ 0.534
|   |   |   Application order > 0.222: Graduate {Dropout=0,
Graduate=1, Enrolled=1}
```

*P-ISSN: 2656-1719*
*E-ISSN: 2656-2774*

```
|   |   |   Application order ≤ 0.222
|   |   |   |   Application mode > 0.143: Enrolled {Dropout=0,
Graduate=0, Enrolled=5}
|   |   |   |   Application mode ≤ 0.143
|   |   |   |   |   Course > 0.920: Enrolled {Dropout=0, Graduate=0,
Enrolled=2}
|   |   |   |   |   Course ≤ 0.920: Dropout {Dropout=2, Graduate=0,
Enrolled=0}
|   Tuition fees up to date ≤ 0.500: Dropout {Dropout=89, Graduate=24,
Enrolled=29}
Curricular units 2nd sem (approved) ≤ 0.175
|   Father's occupation > 0.521
|   |   Curricular units 1st sem (evaluations) > 0.044
|   |   |   Application mode > 0.125: Enrolled {Dropout=0, Graduate=0,
Enrolled=16}
|   |   |   Application mode ≤ 0.125: Dropout {Dropout=1, Graduate=0,
Enrolled=1}
|   |   Curricular units 1st sem (evaluations) ≤ 0.044: Dropout
{Dropout=2, Graduate=0, Enrolled=0}
|   Father's occupation ≤ 0.521: Dropout {Dropout=946, Graduate=102,
Enrolled=248}
```

Proses ini melibatkan analisis mendalam terhadap data, pemilihan fitur yang paling relevan, dan mengatur parameter yang sesuai untuk membangun pohon keputusan yang akurat (Shanthi et al., 2022). Setelah mengikuti langkah-langkah ini, saya berhasil menghasilkan visualisasi yang jelas dan intuitif dari pohon keputusan. Gambar pohon keputusan ini memberikan pandangan visual tentang bagaimana model membuat keputusan berdasarkan kriteria yang ada dalam data. Saya sangat antusias tentang potensi pemanfaatan gambar pohon keputusan ini untuk mengkomunikasikan hasil model dengan tim dan pemangku kepentingan lainnya. Selanjutnya, saya berencana untuk terus memantau dan mengevaluasi kinerja model ini serta mempertimbangkan langkah-langkah lanjutan untuk pengembangan model yang lebih baik di masa mendatang.

Gambar 4. Pohon keputuan yang dibuat dengan akurasi model yang paling tinggi

## KESIMPULAN

Dalam diskusi di atas, kita telah membahas tentang model pohon keputusan dalam konteks tugas klasifikasi. Model pohon keputusan ini adalah alat yang kuat dalam analisis data dan pengambilan keputusan berdasarkan fitur-fitur yang relevan. Kita telah membicarakan tentang pengujian model, termasuk pengukuran akurasi, matriks confusion, dan metrik evaluasi lainnya. Selain itu, kita juga telah menjelaskan pentingnya penggunaan pruning atau pre-pruning dalam menghindari *overfitting*. Namun, kami telah membagikan hasil menarik bahwa dalam pengembangan model, metode tanpa pruning dan pre-pruning telah menghasilkan akurasi tertinggi. Ini mengindikasikan bahwa dalam konteks spesifik ini, model kompleks lebih mampu

memahami pola yang ada dalam data. Kami juga telah mencapai pencapaian penting dalam berhasil membuat gambar pohon keputusan yang akan menjadi alat komunikasi yang efektif untuk mempresentasikan hasil model kepada pemangku kepentingan.

Kesimpulannya, diskusi ini telah menggarisbawahi pentingnya penyesuaian model dengan konteks tertentu, kemampuan model pohon keputusan dalam menganalisis data, dan pentingnya komunikasi hasil kepada pihak terkait. Semoga ini menjadi langkah awal menuju pemahaman yang lebih dalam tentang data dan pengambilan keputusan yang lebih baik di masa depan.

## SARAN

Berikut beberapa saran untuk mengembangkan penelitian kami berdasarkan hasil pengembangan model pohon keputusan yang telah di lakukan:

1. Analisis Lebih Mendalam: Luangkan waktu untuk melakukan analisis yang lebih mendalam terhadap hasil dari model pohon keputusan yang telah Anda buat. Cobalah untuk mengidentifikasi pola-pola atau kecenderungan yang mungkin tersembunyi dalam data.
2. Pembandingan Model: Pertimbangkan untuk membandingkan model pohon keputusan Anda dengan model-model lain, seperti Random Forest, Gradient Boosting, atau Neural Networks, terutama jika Anda memiliki kumpulan data yang lebih besar. Ini dapat memberikan wawasan lebih lanjut tentang mana yang merupakan model terbaik untuk tugas tertentu.
3. Feature Engineering: Lakukan eksplorasi lebih lanjut terhadap fitur-fitur dalam dataset Anda. Mungkin ada cara untuk meningkatkan performa model dengan menghasilkan fitur-fitur baru atau melakukan seleksi fitur yang lebih baik.
4. Validasi Eksternal: Validasi model Anda dengan menggunakan dataset eksternal atau data yang belum pernah dilihat sebelumnya. Ini akan membantu Anda memastikan bahwa model Anda mampu melakukan generalisasi dengan baik pada data baru.
5. Interpretabilitas: Model pohon keputusan memiliki keuntungan besar dalam hal interpretabilitas. Cobalah untuk memahami bagaimana hasil model dapat digunakan untuk mengambil keputusan yang lebih baik dalam domain spesifik Anda.
6. Publikasi dan Kolaborasi: Pertimbangkan untuk mempublikasikan hasil penelitian Anda dalam jurnal ilmiah atau konferensi yang relevan. Ini akan membantu berbagi pengetahuan Anda dengan komunitas ilmiah yang lebih luas dan mungkin membuka pintu untuk kolaborasi dengan peneliti lain.
7. Eksplorasi Kasus Penggunaan: Terapkan model Anda dalam situasi nyata atau kasus penggunaan dalam industri atau organisasi yang relevan. Ini akan membantu Anda mengukur dampak praktis dari model Anda dan mungkin memberikan peluang untuk meningkatkan efisiensi atau pengambilan keputusan di dunia nyata.
8. Perbaikan Model: Terus memantau kinerja model Anda dan jangan ragu untuk memperbaikinya jika diperlukan. Data dan situasi bisnis dapat berubah, dan model perlu disesuaikan agar tetap relevan.

## DAFTAR PUSTAKA

Aning, S., & Przybyła-Kasperek, M. (2022). Comparative Study of Twoing and Entropy Criterion for Decision Tree Classification of Dispersed Data. *Procedia Computer Science*, *207*, 2434–2443. https://doi.org/https://doi.org/10.1016/j.procs.2022.09.301

Aronov, B., de Berg, M., Cardinal, J., Ezra, E., Iacono, J., & Sharir, M. (2023). Subquadratic algorithms for some 3Sum-hard geometric problems in the algebraic decision-tree model. *Computational Geometry*, *109*, 101945.

https://doi.org/https://doi.org/10.1016/j.comgeo.2022.101945

Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J.-M., & Marquis, P. (2022). On the explanatory power of Boolean decision trees. *Data & Knowledge Engineering*, *142*, 102088. https://doi.org/https://doi.org/10.1016/j.datak.2022.102088

Camerlingo, N., Vettoretti, M., Del Favero, S., Facchinetti, A., Choudhary, P., & Sparacino, G. (2022). Generation of post-meal insulin correction boluses in type 1 diabetes simulation models for in-silico clinical trials: More realistic scenarios obtained using a decision tree approach. *Computer Methods and Programs in Biomedicine*, *221*, 106862. https://doi.org/https://doi.org/10.1016/j.cmpb.2022.106862

Cao, Y., Zhao, H., Liang, G., Zhao, J., Liao, H., & Yang, C. (2023). Fast and explainable warm-start point learning for AC Optimal Power Flow using decision tree. *International Journal of Electrical Power & Energy Systems*, *153*, 109369. https://doi.org/https://doi.org/10.1016/j.ijepes.2023.109369

Cembranel, P., Teixeira Dias, F., Silva, C. G. da, Finatto, C. P., & Guerra, J. B. S. O. de A. (2023). Sustainable universities: The LGBTQIAP+ inclusive model. *Evaluation and Program Planning*, *100*, 102351. https://doi.org/https://doi.org/10.1016/j.evalprogplan.2023.102351

Gao, W., Wang, J., Zhou, L., Luo, Q., Lao, Y., Lyu, H., & Guo, S. (2022). Prediction of acute kidney injury in ICU with gradient boosting decision tree algorithms. *Computers in Biology and Medicine*, *140*, 105097. https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.105097

Guggari, S., Kadappa, V., Umadevi, V., & Abraham, A. (2022). Music rhythm tree based partitioning approach to decision tree classifier. *Journal of King Saud University - Computer and Information Sciences*, *34*(6, Part A), 3040–3054. https://doi.org/https://doi.org/10.1016/j.jksuci.2020.03.015

Han, X., Zhu, X., Pedrycz, W., & Li, Z. (2023). A three-way classification with fuzzy decision trees. *Applied Soft Computing*, *132*, 109788. https://doi.org/https://doi.org/10.1016/j.asoc.2022.109788

He, W., Wang, Y., Zhou, M., & Wang, B. (2022). A novel parameters correction and multivariable decision tree method for edge computing enabled HGR system. *Neurocomputing*, *487*, 203–213. https://doi.org/https://doi.org/10.1016/j.neucom.2021.08.147

Lazebnik, T., & Bunimovich-Mendrazitsky, S. (2023). Decision tree post-pruning without loss of accuracy using the SAT-PP algorithm with an empirical evaluation on clinical data. *Data & Knowledge Engineering*, *145*, 102173. https://doi.org/https://doi.org/10.1016/j.datak.2023.102173

Li, M., Vanberkel, P., & Zhong, X. (2022). Predicting ambulance offload delay using a hybrid decision tree model. *Socio-Economic Planning Sciences*, *80*, 101146. https://doi.org/https://doi.org/10.1016/j.seps.2021.101146

Li, Y., Feng, Y., & Qian, Q. (2023). FDPBoost: Federated differential privacy gradient boosting decision trees. *Journal of Information Security and Applications*, *74*, 103468. https://doi.org/https://doi.org/10.1016/j.jisa.2023.103468

Loyola-González, O., Ramírez-Sáyago, E., & Medina-Pérez, M. A. (2023). Towards improving decision tree induction by combining split evaluation measures. *Knowledge-Based Systems*, *277*, 110832. https://doi.org/https://doi.org/10.1016/j.knosys.2023.110832

Mariano, A. M., Ferreira, A. B. de M. L., Santos, M. R., Castilho, M. L., & Bastos, A. C. F. L. C. (2022). Decision trees for predicting dropout in Engineering Course students in Brazil. *Procedia Computer Science*, *214*, 1113–1120. https://doi.org/https://doi.org/10.1016/j.procs.2022.11.285

Moshkov, M. (2022). Decision trees for regular factorial languages. *Array*, *15*, 100203. https://doi.org/https://doi.org/10.1016/j.array.2022.100203

N, S., & K, V. (2022). Detection of Intrusion behavior in cloud applications using Pearson's chi-squared distribution and decision tree classifiers. *Pattern Recognition Letters*, *162*, 15–21. https://doi.org/https://doi.org/10.1016/j.patrec.2022.08.008

Shanthi, J., Rani, D. G. N., & Rajaram, S. (2022). A C4.5 decision tree classifier based floorplanning algorithm for System-on-Chip design. *Microelectronics Journal*, *121*, 105361. https://doi.org/https://doi.org/10.1016/j.mejo.2022.105361

Viloria, A., Padilla, J. G., Vargas-Mercado, C., Hernández-Palma, H., Llinas, N. O., & David, M. A. (2019). Integration of Data Technology for Analyzing University Dropout. *Procedia Computer Science*, *155*, 569–574. https://doi.org/https://doi.org/10.1016/j.procs.2019.08.079

Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, *507*, 772–794. https://doi.org/https://doi.org/10.1016/j.ins.2019.06.064

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*, 166–173. https://doi.org/https://doi.org/10.1016/j.chb.2019.04.015

Zhou, X., Chen, S., Peng, N., Zhou, X., & Wang, X. (2023). Uncertainty guided pruning of classification model tree. *Knowledge-Based Systems*, *259*, 110067. https://doi.org/https://doi.org/10.1016/j.knosys.2022.110067