



Perancangan Model Deteksi Potensi Siswa Putus Sekolah Menggunakan Metode Decision Tree di SMK Informatika Ciputat

Andreas K.M.

Prodi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang

andreasschool322@gmail.com

Abstract

Student dropout is a multidimensional problem that impacts the quality of human resources. This research aims to design an early detection model for potential student dropouts using the Decision Tree CART (Classification and Regression Tree) algorithm at SMK Informatika Ciputat. The Decision Tree CART algorithm was selected because it produces a transparent and easily interpretable model, making it suitable for educational contexts that require a clear understanding of risk factors. This research used five predictor variables: report card scores, attendance percentage, commuting distance, parental occupation, and parental status. Data were collected from school archives covering three academic years. The research stages included data cleaning, derivation of parental occupation features based on KBJI 2014, data standardization using StandardScaler, label encoding, handling class imbalance with SMOTE, parameter optimization, and cross-validation. The dataset was split into training and testing sets with an 80:20 stratified ratio and evaluated using Stratified K-Fold Cross Validation with 5 folds. The trained model was integrated into a Laravel web application using Scikit-Learn for prediction. Evaluation using the confusion matrix shows that the model achieved 89.42% accuracy, demonstrating good performance in classifying at-risk students. Based on feature importance analysis, attendance percentage was identified as the most dominant attribute in determining dropout risk. The resulting model provides an objective, data-driven early detection alternative that supports schools in prioritizing interventions for students at potential risk of dropping out.

Keywords: Dropout, Decision Tree CART, Early Detection, SMOTE, Vocational High School.

Abstrak

Angka putus sekolah merupakan permasalahan multidimensional yang berdampak pada kualitas sumber daya manusia. Penelitian ini bertujuan merancang model deteksi dini potensi siswa putus sekolah menggunakan algoritma Decision Tree CART (Classification and Regression Tree) di SMK Informatika Ciputat. Algoritma Decision Tree CART dipilih karena menghasilkan model yang transparan dan mudah diinterpretasikan sehingga cocok digunakan dalam konteks pendidikan yang memerlukan pemahaman terhadap faktor-faktor risiko secara jelas. Penelitian ini menggunakan lima variabel prediktor, yaitu nilai rapor, persentase absensi, jarak domisili, pekerjaan orang tua, dan status orang tua. Data diperoleh dari arsip sekolah mencakup tiga tahun ajaran. Tahapan penelitian meliputi pembersihan data, derivasi fitur pekerjaan orang tua berdasarkan KBJI 2014, standarisasi data menggunakan StandardScaler, label encoding, penanganan ketidakseimbangan kelas menggunakan SMOTE, serta optimasi parameter dan validasi silang. Dataset dibagi menjadi data latih dan data uji dengan perbandingan 80:20 secara stratified, kemudian dievaluasi menggunakan Stratified K-Fold Cross Validation dengan 5 fold untuk memastikan kemampuan generalisasi model. Model yang telah dilatih kemudian diintegrasikan ke dalam aplikasi web berbasis Laravel dengan memanfaatkan pustaka Scikit-Learn untuk proses prediksi. Hasil evaluasi menggunakan confusion matrix menunjukkan bahwa model memperoleh akurasi sebesar 89,42%, sehingga mampu mengklasifikasikan potensi siswa putus sekolah dengan kinerja yang baik. Berdasarkan analisis tingkat kepentingan atribut yang dihasilkan oleh pohon keputusan, persentase absensi merupakan atribut yang paling dominan dalam menentukan risiko putus sekolah. Model yang dihasilkan mampu memberikan alternatif deteksi dini yang objektif dan berbasis data sehingga dapat mendukung pihak sekolah dalam menentukan prioritas pendampingan terhadap siswa yang berpotensi mengalami putus sekolah.

Kata Kunci: Putus Sekolah, Decision Tree CART, Deteksi Dini, SMOTE, SMK.

1. Pendahuluan

Permasalahan putus sekolah merupakan isu multidimensional yang berdampak langsung terhadap kualitas sumber daya manusia serta perkembangan sosial dan ekonomi suatu negara. Penelitian menunjukkan bahwa permasalahan ini masih menjadi tantangan global yang memerlukan pendekatan analitis dan teknologi untuk mengidentifikasi faktor penyebabnya secara lebih akurat (Roda-Segarra et al., 2024). Faktor ekonomi dan keterbatasan fasilitas pendidikan menjadi penyebab utama terjadinya putus sekolah, di mana sekolah yang berada di wilayah dengan tingkat ekonomi rendah memiliki kecenderungan angka putus sekolah yang lebih tinggi (Silalahi et al., 2025). Selain faktor ekonomi, faktor psikologis dan lingkungan sosial juga berperan penting, di mana siswa dengan dukungan sosial yang positif cenderung memiliki keterlibatan belajar yang lebih tinggi (Magister et al., 2021). Pendekatan sistem peringatan dini atau early warning system telah diterapkan untuk mengidentifikasi siswa yang

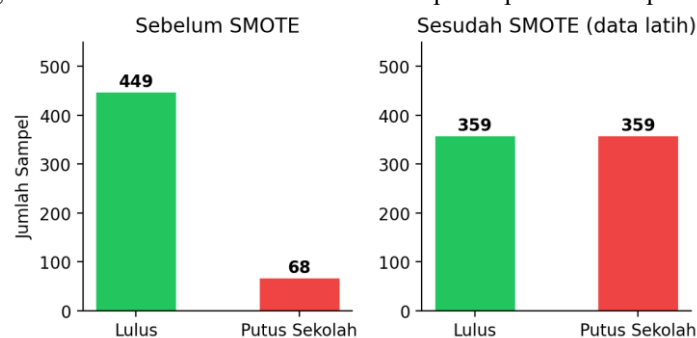
berisiko putus sekolah melalui pemantauan indikator tertentu seperti motivasi belajar, kehadiran, dan kondisi sosial ekonomi, sehingga sekolah dapat melakukan intervensi secara lebih cepat dan tepat waktu (Hidayat et al., 2024). Dalam konteks analisis data pendidikan, pendekatan berbasis machine learning telah banyak digunakan untuk mengidentifikasi faktor-faktor yang memengaruhi prestasi dan keberlanjutan pendidikan siswa. Metode machine learning mampu mengolah data dalam jumlah besar dan menghasilkan pola prediksi yang dapat digunakan untuk pengambilan keputusan strategis di bidang pendidikan (Sarker, 2021). Beberapa penelitian telah menerapkan algoritma Decision Tree untuk deteksi potensi putus sekolah dan menunjukkan hasil yang menjanjikan (Ermillian & Nugroho, 2024). Algoritma Decision Tree seperti ID3, CART, dan C4.5 memiliki kemampuan klasifikasi yang baik serta mampu menghasilkan aturan keputusan yang mudah dipahami oleh pengguna (Bishnoi & Hooda, 2022). Keunggulan utama Decision Tree terletak pada kemampuannya menghasilkan model yang transparan dan mudah diinterpretasikan, sehingga sangat sesuai digunakan dalam sistem pengambilan keputusan berbasis data di bidang pendidikan yang memerlukan pemahaman terhadap faktor-faktor risiko secara jelas (Blockeel et al., 2023). Meskipun berbagai penelitian telah menunjukkan efektivitas machine learning dalam deteksi risiko putus sekolah, implementasinya di lingkungan sekolah menengah kejuruan (SMK) masih terbatas. Proses identifikasi siswa berisiko di sekolah saat ini masih dilakukan secara manual oleh guru bimbingan konseling atau wali kelas sehingga bersifat subjektif dan kurang efisien. SMK Informatika Ciputat sebagai salah satu sekolah menengah kejuruan belum memiliki sistem berbasis data yang mampu mendeteksi secara dini siswa yang berpotensi putus sekolah. Oleh karena itu, diperlukan suatu model deteksi yang objektif dan berbasis data untuk membantu pihak sekolah dalam mengidentifikasi siswa berisiko secara lebih akurat. Algoritma Decision Tree CART (Classification and Regression Tree) dipilih dalam penelitian ini karena menghasilkan model yang transparan, efisien dalam memproses data tabular, serta memberikan informasi mengenai tingkat kepentingan setiap atribut (feature importance) yang dapat digunakan untuk menganalisis faktor dominan penyebab putus sekolah. Penelitian ini bertujuan merancang model deteksi dini potensi siswa putus sekolah menggunakan algoritma Decision Tree CART dengan lima variabel prediktor, yaitu nilai rapor, persentase absensi, jarak domisili, pekerjaan orang tua, dan status orang tua. Model yang dihasilkan diintegrasikan ke dalam aplikasi web berbasis Laravel dengan memanfaatkan pustaka Scikit-Learn untuk proses prediksi sehingga pihak sekolah dapat dengan mudah menggunakannya dalam kegiatan operasional sehari-hari. Sistem yang dikembangkan diharapkan mampu memberikan alternatif deteksi yang objektif dan berbasis data sehingga intervensi dapat dilakukan secara lebih dini dan terarah terhadap siswa yang berpotensi mengalami putus sekolah.

2. Metode Penelitian

2.1. Sumber Data dan Persiapan Awal

Penelitian ini menggunakan data historis siswa SMK Informatika Ciputat yang dikumpulkan dari arsip sekolah mencakup tiga tahun ajaran, yaitu 2023/2024, 2024/2025, dan 2025/2026. Data awal diperoleh dalam format Excel yang terdiri dari data peserta didik dan rekap nilai, kemudian digabungkan menjadi satu dataset. Dataset awal mengandung kolom administratif seperti kode siswa, nama, kelas, jurusan, serta data pekerjaan ayah dan ibu yang masih terpisah. Proses data cleaning dilakukan melalui lima tahapan berurutan: resolusi nama kolom agar sesuai dengan format standar model, pemetaan nilai target dari tiga kategori ("AKTIF", "PUTUS SEKOLAH", "PINDAH SEKOLAH") menjadi dua kelas biner ("Lulus" dan "Putus Sekolah"), penghapusan baris duplikat, penghapusan baris dengan status "PINDAH SEKOLAH" atau target kosong, serta imputasi missing values pada fitur numerik menggunakan median.

Setelah data cleaning, diperoleh 517 sampel dengan lima fitur prediktor dan satu target biner. Gambar 1 menyajikan perbandingan distribusi kelas sebelum dan sesudah penerapan SMOTE pada data latih.



Gambar 1. Distribusi Kelas Sebelum dan Sesudah SMOTE

Kelima fitur prediktor terdiri dari tiga fitur numerik (Nilai Rapor semester 1, Persentase Absensi semester 1, dan Jarak domisili dalam kilometer) serta dua fitur kategorikal (Pekerjaan Orang Tua dan Status Orang Tua). Fitur Pekerjaan Orang Tua tidak tersedia secara langsung dalam dataset asli, melainkan diturunkan dari tiga kolom terpisah (Pekerjaan Ayah, Pekerjaan Ibu, dan Status Orang Tua) melalui pipeline tiga tahap. Tahap pertama

menentukan pekerjaan mana yang digunakan berdasarkan status orang tua. Tahap kedua menormalisasi variasi penulisan seperti "driver gojek", "driver online", dan "ojek online" menjadi "Driver Online" melalui kamus sinonim. Tahap ketiga memetakan jabatan yang telah ternormalisasi ke dalam Klasifikasi Baku Jabatan Indonesia (KBJI) 2014 yang terdiri dari sepuluh major group (KBJI 0 hingga KBJI 9). Jabatan yang tidak memiliki padanan dalam KBJI dikategorikan sebagai "Belum Terklasifikasi", sementara "Ibu Rumah Tangga", "Data Tidak Tersedia", dan "Tidak Berlaku" diperlakukan sebagai kategori final.

2.2. Preprocessing Data

Fitur kategorikal dikonversi ke representasi numerik menggunakan Label Encoding. Algoritma Decision Tree pada Scikit-Learn hanya memproses data numerik, sehingga nilai kategorikal harus diubah. Label Encoding dipilih karena Decision Tree tidak sensitif terhadap urutan numerik dan mampu menangani multi-level secara langsung, berbeda dengan One-Hot Encoding yang akan memperluas dimensi secara tidak perlu. Fitur Status Orang Tua memiliki lima kategori (LENGKAP, CERAI, YATIM, PIATU, YATIM PIATU) yang dipetakan ke nilai 0 hingga 4. Fitur Pekerjaan Orang Tua memiliki empat belas kategori (KBJI 0 hingga KBJI 9, Ibu Rumah Tangga, Belum Terklasifikasi, Data Tidak Tersedia, Tidak Berlaku) yang dipetakan ke nilai 0 hingga 13. Apabila ditemukan nilai yang tidak terdapat dalam kamus encoding, sistem memberikan kode -1 sebagai indikator nilai tidak dikenali.

Fitur numerik distandarisasi menggunakan StandardScaler yang mengubah data melalui perhitungan z-score sehingga rata-rata mendekati nol dan simpangan baku mendekati satu. Persamaan StandardScaler dinyatakan sebagai berikut.

$$z = \frac{x - \mu}{\sigma}$$

StandardScaler dipilih karena mempertahankan distribusi asli data dan lebih robust terhadap pencilan dibandingkan metode normalisasi lainnya (Sarker, 2021). Parameter mean dan standard deviation dihitung dari seluruh data latih, kemudian diterapkan secara konsisten pada data uji dan data baru tanpa menghitung ulang. Dataset yang digunakan memiliki ketidakseimbangan kelas yang signifikan, yaitu 449 sampel Lulus (86,8%) dan 68 sampel Putus Sekolah (13,2%). Ketidakseimbangan ini ditangani menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) yang membangkitkan data sintesis pada kelas minoritas melalui interpolasi antar tetangga terdekat (Putri & Rachmatika, 2025). Proses interpolasi SMOTE dinyatakan dalam persamaan berikut.

$$x_{baru} = x_i + \lambda \cdot (x_{zi} - x_i)$$

dengan x_i adalah sampel minoritas, x_{zi} tetangga terdekat, dan λ bilangan acak dalam $[0,1]$. SMOTE diterapkan hanya pada data latih setelah pembagian dataset untuk mencegah data leakage, dengan parameter $k\text{-neighbors} = 5$ dan $\text{random_state} = 42$. Setelah SMOTE, jumlah sampel kelas Putus Sekolah pada data latih meningkat dari 54 menjadi 359, menyamai jumlah kelas Lulus sehingga total data latih menjadi 718 sampel. Distribusi kelas sebelum dan sesudah SMOTE disajikan pada Tabel 1.

Tabel 1. Distribusi Kelas Sebelum dan Sesudah SMOTE

Dataset	Lulus	Putus Sekolah	Total	Rasio Minoritas
Data Latih (sebelum SMOTE)	359	54	413	15,04%
Data Uji	90	14	104	15,56%
Data Latih (sesudah SMOTE)	359	359	718	100%

2.3. Pemodelan Decision Tree CART

Dataset dibagi menjadi data latih dan data uji dengan perbandingan 80:20 menggunakan stratified sampling untuk menjaga proporsi kelas pada kedua bagian. Pembagian menghasilkan 413 sampel data latih dan 104 sampel data uji. Stratified sampling memastikan bahwa rasio kelas minoritas pada data latih (13,1%) dan data uji (13,5%) tetap konsisten dengan dataset asli (13,2%). Model dibangun menggunakan algoritma Decision Tree CART (Classification and Regression Trees) dari Scikit-Learn yang membangun pohon keputusan biner secara rekursif dengan memilih pemisahan terbaik pada setiap simpul berdasarkan kriteria entropi (information gain) (Sarker, 2021). Entropy dihitung dengan persamaan berikut.

$$H = - \sum_{j=1}^c p_j \log_2(p_j)$$

dengan p_j adalah proporsi kelas j dalam suatu simpul dan c adalah jumlah kelas. Pemilihan pemisahan terbaik secara rekursif ini menghasilkan model Decision Tree yang transparan dan mudah diinterpretasikan (Blockeel et al., 2023). Algoritma CART dipilih karena menghasilkan model yang transparan dan mudah diinterpretasikan, sehingga cocok digunakan dalam konteks pendidikan yang memerlukan pemahaman terhadap faktor-faktor risiko secara jelas (Bishnoi & Hooda, 2022).

Hyperparameter tuning dilakukan menggunakan GridSearchCV dengan 5-fold Stratified Cross-Validation untuk menemukan kombinasi parameter terbaik (Ekka Pujo Ariesanto Akhmad, 2025). Parameter yang dioptimasi dan nilai yang diuji disajikan pada Tabel 2.

Tabel 2. Parameter yang Diuji dalam GridSearchCV

Parameter	Nilai yang Diuji
criterion	gini, entropy
max_depth	3, 5, 7, 10, 15, None
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4

Total kombinasi yang diuji adalah 108 kombinasi. Scoring menggunakan F1-macro untuk menyeimbangkan performa pada kedua kelas. Parameter terbaik yang diperoleh adalah criterion = entropy, max_depth = 15, min_samples_split = 5, dan min_samples_leaf = 1 dengan skor CV rata-rata sebesar 0,9554. Kombinasi ini memberikan keseimbangan antara kompleksitas dan generalisasi model. Hasil 5-fold Stratified Cross-Validation pada model dengan parameter terbaik disajikan pada Tabel 3.

Tabel 3. Hasil K-Fold Cross Validation

Fold	Akurasi	F1-Score
1	0,9444	0,9444
2	0,9444	0,9444
3	0,9861	0,9861
4	0,9371	0,9371
5	0,9650	0,9650
Mean	0,9554	0,9554
Std	0,0180	0,0180

Proses rekursif menghasilkan pohon keputusan dengan 61 simpul (30 internal, 31 daun) dan kedalaman maksimum 10 level, dengan Absensi sebagai fitur pemisah akar.

2.4. Evaluasi Model dan Integrasi Sistem

Performa model dievaluasi menggunakan confusion matrix, akurasi, precision, recall, dan F1-score pada data uji yang tidak digunakan selama pelatihan. Akurasi dihitung sebagai proporsi prediksi benar terhadap total sampel menggunakan persamaan berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

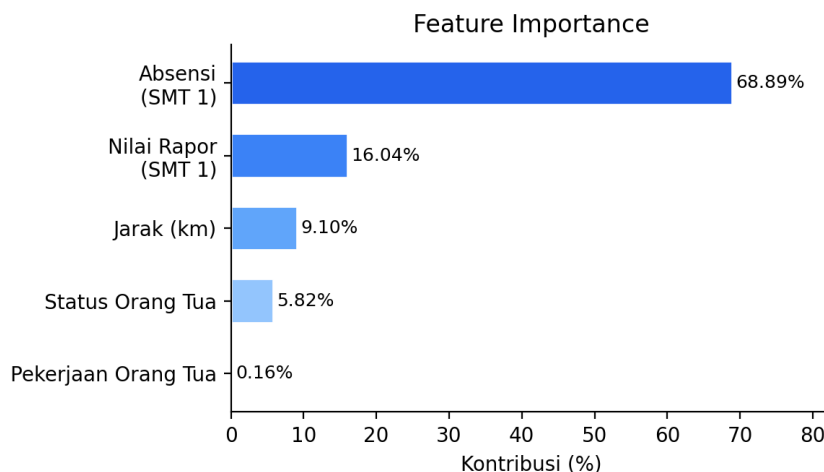
Confusion matrix menyajikan perbandingan antara prediksi model dengan nilai aktual, sedangkan precision dan recall memberikan gambaran performa pada masing-masing kelas. Analisis feature importance juga dilakukan untuk mengidentifikasi kontribusi setiap fitur terhadap keputusan model, dihitung berdasarkan seberapa sering suatu fitur digunakan sebagai pemisah dan seberapa besar pengaruhnya terhadap pengurangan entropi (Blockeel et al., 2023).

Model yang telah dilatih diintegrasikan ke dalam aplikasi web berbasis Laravel yang berkomunikasi dengan Python Scikit-Learn melalui mekanisme subprocess. Laravel mengirimkan data siswa dalam format JSON ke script Python melalui shell_exec, kemudian Python memuat model dan artefak preprocessing yang telah disimpan dalam format .pkl menggunakan pustaka Joblib. Empat artefak yang disimpan meliputi model.pkl (objek DecisionTreeClassifier terlatih), encoders.pkl (kamus LabelEncoder), scaler.pkl (objek StandardScaler dengan parameter μ dan σ), dan target_encoder.pkl (LabelEncoder untuk target). Python melakukan preprocessing dengan encoding dan standarisasi menggunakan parameter yang identik dengan saat pelatihan, menjalankan inferensi, menghitung probabilitas dan kontribusi fitur melalui decision path, kemudian mengembalikan hasil dalam format JSON. Pendekatan subprocess ini memisahkan tanggung jawab aplikasi web (otorisasi, form, ORM) dari komputasi numerik dan model machine learning tanpa memerlukan API server terpisah (Ermillian & Nugroho, 2024).

3. Hasil dan Pembahasan

3.1. Analisis Faktor Risiko

Analisis feature importance dari model Decision Tree yang telah dilatih memberikan gambaran mengenai kontribusi setiap fitur dalam menentukan prediksi putus sekolah. Fitur dengan kontribusi lebih besar menunjukkan pengaruh yang lebih signifikan terhadap keputusan model. Nilai feature importance dihitung berdasarkan seberapa sering suatu fitur digunakan sebagai pemisah dalam pohon keputusan dan seberapa besar kontribusinya terhadap penurunan entropi secara keseluruhan. Semakin tinggi nilai feature importance suatu fitur, semakin besar perannya dalam membentuk struktur pohon keputusan dan menentukan hasil klasifikasi. Gambar 2 menyajikan kontribusi seluruh fitur secara visual.



Gambar 2. Feature Importance

Hasil analisis menunjukkan bahwa Persentase Absensi merupakan faktor yang paling dominan dengan kontribusi sebesar 68,89% terhadap keputusan model. Nilai Rapor menempati posisi kedua dengan kontribusi 16,04%, diikuti oleh Jarak domisili sebesar 9,10%, Status Orang Tua sebesar 5,82%, dan Pekerjaan Orang Tua sebesar 0,16%. Ketiga fitur teratas (Absensi, Nilai Rapor, dan Jarak) secara kumulatif menyumbang 94,02% dari total pengaruh, menunjukkan bahwa faktor akademik dan geografis jauh lebih dominan dibandingkan faktor demografis orang tua. Rincian kontribusi setiap fitur disajikan pada Tabel 4.

Tabel 4. Feature Importance

Peringkat	Fitur	Kontribusi
1	Persentase Absensi	68,89%
2	Nilai Rapor	16,04%
3	Jarak Domisili	9,10%
4	Status Orang Tua	5,82%
5	Pekerjaan Orang Tua	0,16%

Dominasi Absensi sebagai faktor utama mengonfirmasi bahwa tingkat kehadiran siswa merupakan indikator peringatan dini yang paling sensitif dalam mengidentifikasi risiko putus sekolah. Temuan ini sejalan dengan konsep early warning system di bidang pendidikan yang menempatkan kehadiran sebagai salah satu indikator utama (Hidayat et al., 2024). Siswa dengan persentase absensi rendah cenderung mengalami ketertinggalan materi dan menurunnya motivasi belajar, yang pada akhirnya meningkatkan risiko putus sekolah. Nilai Rapor sebagai faktor kedua menunjukkan bahwa performa akademik juga berperan penting, di mana siswa dengan nilai rendah memiliki risiko lebih tinggi. Jarak domisili yang cukup berkontribusi mengindikasikan bahwa aksesibilitas geografis turut memengaruhi konsistensi kehadiran siswa. Sementara itu, rendahnya kontribusi Status Orang Tua dan Pekerjaan Orang Tua menunjukkan bahwa faktor demografis orang tua memiliki pengaruh yang lebih kecil dalam konteks deteksi dini di SMK Informatika Ciputat.

3.2. Evaluasi Model

Pengujian model dilakukan pada data uji sebanyak 104 sampel yang tidak digunakan selama proses pelatihan, sehingga memberikan gambaran objektif tentang kemampuan generalisasi model terhadap data baru. Seluruh sampel data uji merupakan data historis siswa yang telah diketahui status kelulusannya, sehingga hasil prediksi dapat dibandingkan secara langsung dengan nilai aktual. Data uji dipilih secara stratified untuk memastikan proporsi kelas Lulus dan Putus Sekolah tetap representatif sesuai distribusi dataset asli. Model mencapai akurasi sebesar 89,42% dengan confusion matrix yang ditampilkan pada Gambar 3.

Confusion Matrix

Aktual	Lulus	84	6
	Putus Sekolah	5	9
		Lulus	Putus Sekolah
		Prediksi	

Gambar 3. Confusion Matrix

Confusion matrix pada Tabel 5 menunjukkan 84 siswa Lulus diprediksi dengan benar (true negative), 9 siswa Putus Sekolah diprediksi dengan benar (true positive), 6 siswa Lulus salah diprediksi sebagai Putus Sekolah (false positive), dan 5 siswa Putus Sekolah salah diprediksi sebagai Lulus (false negative).

Tabel 5. Confusion Matrix

	Prediksi Lulus	Prediksi Putus Sekolah
Aktual Lulus	84	6
Aktual Putus Sekolah	5	9

Metrik evaluasi per kelas pada Tabel 6 menunjukkan bahwa performa model pada kelas Lulus sangat baik dengan precision 0,9438, recall 0,9333, dan F1-score 0,9385. Pada kelas Putus Sekolah, model memperoleh precision 0,6000, recall 0,6429, dan F1-score 0,6207. Rata-rata makro (macro average) untuk seluruh metrik berada di atas 0,77 yang menunjukkan kinerja yang cukup baik secara keseluruhan.

Tabel 6. Precision, Recall, dan F1-Score per Kelas

Kelas	Precision	Recall	F1-Score
Lulus	0,9438	0,9333	0,9385
Putus Sekolah	0,6000	0,6429	0,6207
Macro Average	0,7719	0,7881	0,7796

Akurasi data latih tercatat sebesar 99,27% dengan selisih sekitar 10% terhadap data uji, mengindikasikan adanya overfitting ringan yang umum terjadi pada model Decision Tree. Namun, hasil Stratified K-Fold Cross Validation dengan 5 lipatan menunjukkan skor rata-rata akurasi 95,54% dengan standar deviasi 1,80%, serta skor F1-macro rata-rata 0,9554 yang setara dengan skor validasi silang terbaik (best CV score) saat hyperparameter tuning. Konsistensi antara skor CV dan akurasi data uji menunjukkan bahwa model memiliki kemampuan generalisasi yang memadai untuk digunakan sebagai alat deteksi dini. Penelitian serupa oleh (Ermillian & Nugroho, 2024) yang menggunakan algoritma Decision Tree untuk deteksi putus sekolah memperoleh akurasi sebesar 90,3%, sementara (Pallathadka et al., 2023) melaporkan bahwa pendekatan machine learning di bidang pendidikan mampu memberikan akurasi di atas 85% untuk berbagai kasus klasifikasi. Hasil akurasi 89,42% pada penelitian ini berada dalam rentang yang sebanding dengan penelitian-penelitian tersebut.

3.3. Pembahasan

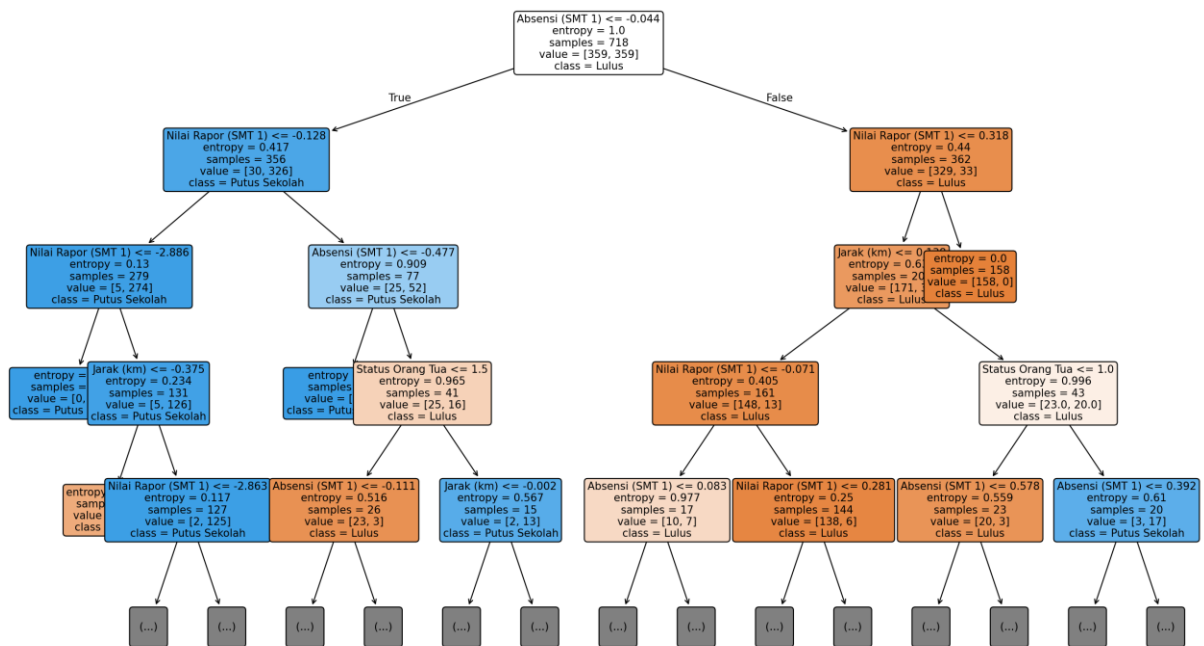
Model Decision Tree CART yang dihasilkan memiliki struktur yang transparan dengan 61 simpul dan kedalaman maksimum 10 level, sehingga keputusan yang diambil oleh model dapat ditelusuri dan dipahami dengan jelas oleh pihak sekolah. **Error! Reference source not found.** menyajikan visualisasi struktur pohon keputusan pada level awal.

Fitur Absensi menjadi pemisah pada simpul akar dengan ambang -0,04 (dalam skala standar), mengonfirmasi perannya sebagai indikator paling awal dalam mengidentifikasi risiko putus sekolah. Keunggulan interpretabilitas ini menjadikan Decision Tree lebih unggul dibandingkan metode black-box seperti neural network atau support vector machine dalam konteks pendidikan, di mana pemahaman terhadap faktor-faktor risiko sama pentingnya dengan akurasi prediksi (Bishnoi & Hooda, 2022). (Song et al., 2024) juga menegaskan bahwa Decision Tree merupakan metode yang efektif untuk menganalisis faktor-faktor psikologis dan sosial dalam konteks pendidikan karena kemampuannya mengidentifikasi pola interaksi antar variabel secara hierarkis. Temuan ini didukung oleh (Lee et al., 2024) yang menunjukkan bahwa pendekatan evaluasi stabilitas pada Decision Tree mampu menghasilkan aturan klasifikasi yang lebih konsisten dibandingkan pendekatan konvensional. Penerapan SMOTE terbukti efektif dalam meningkatkan kemampuan deteksi kelas minoritas, di mana recall kelas Putus Sekolah mencapai 64,29% meskipun precision masih tergolong rendah (0,6000). Dalam konteks deteksi dini, recall lebih diutamakan daripada precision yang sempurna, karena lebih baik melakukan intervensi pada siswa yang terindikasi berisiko meskipun terdapat kemungkinan beberapa di antaranya sebenarnya akan bertahan. Hasil ini sejalan dengan temuan (Putri & Rachmatika, 2025) yang menerapkan SMOTE pada dataset 1.690 siswa SMK dan berhasil meningkatkan recall dari 0% menjadi 99,09% dengan akurasi 99,54%, menunjukkan bahwa SMOTE merupakan teknik yang sangat efektif untuk deteksi putus sekolah di lingkungan SMK.

Integrasi model ke dalam aplikasi web berbasis Laravel memungkinkan pihak sekolah untuk menggunakan model secara langsung tanpa keahlian teknis di bidang machine learning. Guru bimbingan konseling dan wali kelas dapat melakukan prediksi secara mandiri, melihat hasil dalam bentuk visual yang mudah dipahami, serta mengidentifikasi faktor risiko yang dominan pada setiap siswa melalui kontribusi fitur yang ditampilkan pada halaman detail prediksi. Sistem ini melengkapi pendekatan early warning system yang telah direkomendasikan dalam penelitian sebelumnya (Hidayat et al., 2024) dengan menyediakan alat deteksi yang objektif dan terukur. Keterbatasan penelitian ini antara lain penggunaan lima variabel prediktor yang masih terbatas, serta data yang hanya berasal dari satu sekolah sehingga generalisasi ke populasi yang lebih luas memerlukan validasi tambahan. Penelitian selanjutnya dapat menambahkan variabel non-akademik seperti motivasi belajar, keterlibatan orang tua, dan kondisi psikososial, serta mengeksplorasi algoritma lain seperti Random Forest atau XGBoost untuk meningkatkan akurasi dan stabilitas model (Pallathadka et al., 2023). Evaluasi dan retraining model secara berkala setiap semester atau tahun ajaran diperlukan agar model tetap relevan dengan kondisi aktual siswa. Pengembangan fitur notifikasi otomatis dan modul pencatatan intervensi juga dapat meningkatkan efektivitas sistem sebagai alat pendukung keputusan bagi pihak sekolah.

DOI: <https://doi.org/10.69693/ijmst.v4i2.12550>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)



Gambar 4. Struktur Decision Tree

4. Kesimpulan

Penelitian ini berhasil merancang model deteksi dini potensi putus sekolah menggunakan algoritma Decision Tree CART di SMK Informatika Ciputat. Model dibangun berdasarkan lima variabel prediktor, yaitu nilai rapor, persentase absensi, jarak domisili, pekerjaan orang tua, dan status orang tua, dengan parameter optimal entropy, kedalaman maksimum 15, minimal sampel pemisah 5, dan minimal sampel daun 1. Model menghasilkan pohon keputusan dengan 61 simpul dan kedalaman 10 level yang bersifat transparan dan mudah diinterpretasikan.

Persentase absensi teridentifikasi sebagai faktor paling dominan dengan kontribusi 68,89% terhadap prediksi, diikuti oleh nilai rapor sebesar 16,04%, jarak domisili sebesar 9,10%, status orang tua sebesar 5,82%, dan pekerjaan orang tua sebesar 0,16%. Temuan ini mengonfirmasi bahwa tingkat kehadiran siswa merupakan indikator peringatan dini yang paling sensitif dalam mengidentifikasi risiko putus sekolah. Model mencapai akurasi sebesar 89,42% pada data uji dengan kemampuan deteksi kelas minoritas yang cukup baik setelah penerapan SMOTE, yaitu recall kelas Putus Sekolah sebesar 64,29%. Penerapan SMOTE berhasil mengatasi ketidakseimbangan kelas dengan meningkatkan sampel minoritas dari 54 menjadi 359 pada data latih.

Sistem berbasis web yang dikembangkan dengan Laravel dan Scikit-Learn memungkinkan pihak sekolah melakukan prediksi secara mandiri tanpa keahlian teknis di bidang machine learning. Sistem menyediakan fitur manajemen data siswa, prediksi individual dan batch, visualisasi hasil prediksi, analisis faktor risiko, serta laporan PDF. Model yang dihasilkan memberikan alternatif deteksi dini yang objektif dan berbasis data sehingga dapat mendukung pihak sekolah dalam menentukan prioritas pendampingan terhadap siswa yang berpotensi mengalami putus sekolah. Penelitian selanjutnya disarankan untuk menambahkan variabel non-akademik serta mengeksplorasi algoritma lain guna meningkatkan akurasi dan stabilitas model.

Reference

- Bañeres, D., Rodríguez-González, M. E., Guerrero-Roldán, A. E., & Cortadas, P. (2023). An early warning system to identify and intervene online dropout learners. *International Journal of Educational Technology in Higher Education*, 20, 3. <https://doi.org/10.1186/s41239-022-00371-5>
- Bishnoi, S., & Hooda, B. K. (2022). Decision Tree Algorithms and their Applicability in Agriculture for Classification. *Journal of Experimental Agriculture International*, 20–27. <https://doi.org/10.9734/jeai/2022/v44i730833>
- Blockeel, H., Devos, L., Frénay, B., Nanfack, G., & Nijssen, S. (2023). Decision trees: from efficient prediction to responsible AI. In *Frontiers in Artificial Intelligence* (Vol. 6). Frontiers Media SA. <https://doi.org/10.3389/frai.2023.1124553>
- Ekka Pujo Ariesanto Akhmad. (2025). Enhancing the Accuracy of Airline Review Classification Using SMOTE and Grid Search with Cross Validation for Hyperparameter Tuning. *Journal of Information Systems Engineering and Management*, 10(25s), 850–861. <https://doi.org/10.52783/jisem.v10i25s.4174>
- Ermilian, A., & Nugroho, K. (2024). Perancangan Model Deteksi Potensi Siswa Putus Sekolah Menggunakan Metode Logistic Regression Dan Decision Tree. *Jurnal Informatika: Jurnal Pengembangan IT*, 9(3), 281–295. <https://doi.org/10.30591/jpit.v9i3.8007>
- Hidayat, T., An Nur, A. M., Hasri, H., Genisa, M. U., & Rizal, H. P. (2024). Implementasi Program Early Warning Systems Sebagai Upaya Mitigasi Risiko Anak Putus Sekolah. *Jurnal Pengabdian Kepada Masyarakat (JPKM) TABIKPUN*, 5(1), 37–46. <https://doi.org/10.23960/jpkmt.v5i1.148>

DOI: <https://doi.org/10.69693/ijmst.v4i2.12550>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

- Lee, J., Sim, M. K., & Hong, J. S. (2024). Assessing Decision Tree Stability: A Comprehensive Method for Generating a Stable Decision Tree. *IEEE Access*, 12, 90061–90072. <https://doi.org/10.1109/ACCESS.2024.3419228>
- Magister, M. J., Sains, P., Psikologi, F., Dahlan, U. A., Magister, K. B., Hidayah, N., & Sains, M. P. (2021). Gambaran Student Engagement pada Siswa SMA (Studi Kasus pada Siswa MAN 1 Magelang). 17(1). <https://doi.org/10.32528/ins.v%vi%oi.3615>
- 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/10.1016/j.procs.2022.11.285>
- Pallathadka, H., Sajja, G. S., Mustafa, M., Kaliyaperumal, K., Patra, I., Ray, S., & Raghuvanshi, A. (2023). Machine learning in education, finance and management: Applications and future trends. 090003. <https://doi.org/10.1063/5.0150535>
- Putri, S. A., & Rachmatika, R. (2025). Penerapan Algoritma Random Forest dan SMOTE untuk Prediksi Risiko Putus Sekolah Siswa Sekolah Menengah Kejuruan. *Decode: Jurnal Pendidikan Teknologi Informasi*, 5(3), 903–910. <https://doi.org/10.51454/decode.v5i3.1360>
- Roda-Segarra, J., De-La-peña, C., & Mengual-Andrés, S. (2024). Effectiveness of Artificial Intelligence Models for Predicting School Dropout: A Meta-Analysis. *Multidisciplinary Journal of Educational Research*, 14(3), 317–340. <https://doi.org/10.17583/remie.13342>
- Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3), 160. <https://doi.org/10.1007/s42979-021-00592-x>
- Scheidig, F., & Holmeier, M. (2023). Predicting student teachers' dropout with machine learning: Potentials of using student and study progress data from the campus management system. *Unterrichtswissenschaft*, 51, 489–509. <https://doi.org/10.1007/s42010-023-00182-1>
- Selim, K. S., & Rezk, S. S. (2023). On predicting school dropouts in Egypt: A machine learning approach. *Education and Information Technologies*, 28, 9235–9266. <https://doi.org/10.1007/s10639-022-11571-x>
- Shiao, Y.-T., Chen, C.-H., Wu, K.-F., Chen, B.-L., Chou, Y.-H., & Wu, T.-N. (2023). Reducing dropout rate through a deep learning model for sustainable education: Long-term tracking of learning outcomes of an undergraduate cohort from 2018 to 2021. *Smart Learning Environments*, 10, 55. <https://doi.org/10.1186/s40561-023-00274-6>
- Silalahi, Y. N., Panjaitan, R., & Harahap, L. S. (2025). ANALISIS PERBANDINGAN ANGKA PUTUS SEKOLAH DASAR NEGERI DAN SWASTA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS (K-NN). *Jurnal Informatika Dan Teknik Elektro Terapan*, 13(1). <https://doi.org/10.23960/jitet.v13i1.6045>
- Song, P., Cai, X., Qin, D., Wang, Q., Liu, X., Zhong, M., Li, L., & Yang, Y. (2024). Analyzing psychological resilience in college students: A decision tree model. *Heliyon*, 10(11). <https://doi.org/10.1016/j.heliyon.2024.e32583>
- Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence*, 4, 2. <https://doi.org/10.1007/s44163-023-00079-z>
- Xia, X., & Qi, W. (2023). Dropout prediction and decision feedback supported by multi temporal sequences of learning behavior in MOOCs. *International Journal of Educational Technology in Higher Education*, 20, 32. <https://doi.org/10.1186/s41239-023-00400-x>