AdeS2-4Des2024

by Kristiawan Nugroho

Submission date: 04-Dec-2024 03:22PM (UTC+0700)
Submission ID: 2540292654
File name: JPIT_Template_2024_AdeErmillian_Rev01.docx (810.96K)
Word count: 5387
Character count: 35055

Jurnal Informatika: Jurnal pengembangan IT
Vol. xx, No. x, Bulan 20xx
ISSN : 2477-5126, DOI: 10.30591/jpit.v9ix.xxx

Perancangan Model Deteksi Potensi Siswa Putus Sekolah Menggunakan Metode Logistic Regression dan Decision Tree

Ade Ermillian, Kristiawan Nugroho
Magister Teknologi Informasi, Fakultas Teknologi Informasi dan Industri, Unisbank, Semarang
Jln. Trilomba Juang No 1, Kota Semarang, 50272, Indonesia

Commented [R1]: Diganti Program Studi
Commented [R2]: Universitas Stikubank

Info Artikel

Riwayat Artikel:
Received month dd, yyyy
Revised month dd, yyyy
Accepted month dd, yyyy

Abstract – This study aims to design a model for detecting the potential of school dropout students using Logistic Regression and Decision Tree methods based on student data from SMA N 4 Tegal. The variables used in the analysis include demographic, academic, and social information such as absenteeism, average semester grades, parental income, and transportation type. The dataset is processed using one-hot encoding and label encoding techniques to convert categorical data into numeric values. The results indicate that both methods have their respective advantages. The Decision Tree model achieves high precision, especially in predicting students who continue their education, with a precision of 0.99 for the "Continue School" class. However, the recall for the "Dropout" class remains low (0.60), indicating the need for improvements in detecting students at risk of dropping out. On the other hand, the Logistic Regression model shows better balance in detecting both classes, with more balanced accuracy and recall. This study concludes that both models can be used to monitor the potential of school dropouts and provide data-driven recommendations for more accurate educational decision-making.

Keywords: Logistic Regression, Decision Tree, student dropout, risk detection, data analysis.

Corresponding Author:
Ade Ermillian
Email:
adeermillian0009@mhs.unisbank.ac.id

Abstrak – Penelitian ini bertujuan untuk merancang model deteksi potensi siswa putus sekolah dengan menggunakan metode Logistic Regression dan Decision Tree pada data siswa SMA N 4 Tegal. Variabel yang digunakan dalam analisis mencakup informasi demografis, akademik, dan sosial siswa, seperti ketidakhadiran, nilai rata-rata semester, penghasilan orang tua, dan jenis transportasi. Dataset yang digunakan diolah dengan teknik one-hot encoding dan label encoding untuk mengubah data kategorikal menjadi numerik. Hasil penelitian menunjukkan bahwa kedua metode memiliki keunggulan masing-masing. Model Decision Tree menghasilkan presisi tinggi, terutama dalam memprediksi siswa yang tetap melanjutkan sekolah, dengan presisi mencapai 0,99 untuk kelas "Tetap Selesai." Namun, recall untuk kelas "Putus Sekolah" masih rendah (0,60), yang menunjukkan perlunya perbaikan dalam deteksi siswa yang berisiko putus sekolah. Di sisi lain, model Logistic Regression menunjukkan keseimbangan yang lebih baik dalam mendeteksi kedua kelas, dengan akurasi dan recall yang lebih seimbang. Penelitian ini menyimpulkan bahwa kedua model dapat digunakan untuk memonitor potensi siswa putus sekolah dan memberikan rekomendasi berbasis data untuk pengambilan keputusan pendidikan yang lebih tepat.

Kata Kunci: Logistic Regression, Decision Tree, siswa putus sekolah, deteksi risiko, analisis data

I. PENDAHULUAN

Siswa merupakan elemen penting dalam sistem pendidikan yang harus mendapatkan perhatian khusus. Dalam konteks pendidikan saat ini, keberhasilan siswa tidak hanya diukur dari pencapaian akademis, tetapi juga dari kemampuan mereka menyelesaikan pendidikan tanpa terhambat oleh berbagai kendala. Namun, fenomena siswa yang putus sekolah masih menjadi salah satu tantangan besar yang dihadapi banyak institusi pendidikan. Hal ini mencerminkan adanya masalah yang mendasar dan membutuhkan solusi yang tepat agar siswa dapat menyelesaikan pendidikan dengan baik.

Jumlah siswa yang mengalami putus sekolah atau drop out cukup signifikan dan biasanya disebabkan oleh berbagai faktor. Beberapa di antaranya adalah masalah kehadiran yang rendah, tekanan ekonomi keluarga, serta masalah sosial atau psikologis. Selain itu, kurangnya motivasi belajar dan lingkungan belajar yang kurang mendukung sehingga nilai yang di dapat rendah juga menjadi penyebab lain yang perlu diperhatikan. Kondisi ini mendorong sekolah untuk mengambil kebijakan yang efektif dan proaktif dalam mengidentifikasi serta mengatasi masalah yang dapat menyebabkan siswa meninggalkan sekolah.

Dalam upaya menekan angka putus sekolah, sekolah memerlukan sistem yang mampu mendeteksi secara dini siswa yang memiliki potensi untuk tidak melanjutkan pendidikannya. Deteksi ini memungkinkan sekolah untuk memberikan perhatian dan intervensi yang tepat kepada siswa yang membutuhkan. Sayangnya, hingga saat ini belum banyak sekolah yang memiliki program atau sistem otomatis untuk mendeteksi siswa yang berisiko putus sekolah. Ketiadaan sistem ini membuat proses identifikasi seringkali lambat dan kurang efisien, sehingga intervensi yang

diberikan tidak optimal. Oleh karena itu, penting untuk mengembangkan model berbasis teknologi yang dapat memba[22]sekolah dalam menangani permasalahan ini secara lebih terarah.

Penelitian ini dilaksanakan di Sekolah Menengah Atas[29](SMA) Negeri 4 Tegal, salah satu institusi pendidikan tingkat atas yang cukup dikenal di Kota Tegal. Sekolah ini mulai menerima peserta didik baru pada tahun ajaran 1989/1990. Dengan pengalaman dan usia yang cukup matang, SMA N[27] 4 Tegal terus berupaya meningkatkan kualitas pendidikan serta pelayanan terhadap siswa. Namun, meskipun memiliki sejarah yang panjang dan reputasi yang baik, tantangan seperti siswa yang mengalami putus sekolah masih menjadi perhatian utama bagi pihak sekolah.

Berdasarkan data terbaru, jumlah siswa SMA Negeri 4 Tegal yang putus sekolah dari angka[24]2022 hingga angkatan 2024 mencapai 23 siswa. Jumlah ini cukup memprihatinkan, mengingat dampak negatifnya tidak hanya bagi siswa yang bersangkutan tetapi juga bagi citra institusi pendidikan. Data siswa yang putus sekolah telah tercatat secara rapi dalam database sekolah dan menjadi sumber informasi penting yang berpotensi untuk dimanfaatkan lebih lanjut. Salah satu pemanfaatannya adalah melalui proses penambangan data (*data mining*), yang dapat membantu menganalisis pola dan faktor-faktor penyebab siswa putus sekolah serta mendukung pembuatan kebijakan yang lebih efektif untuk mengatasinya.

Implementasi *data mining* dalam bidang pendidikan, terutama untuk mendeteksi siswa atau mahasiswa yang berpotensi mengalami putus sekolah (*drop out*), telah menjadi perhatian penting bagi berbagai institusi pendidikan. Proses ini memungkinkan lembaga untuk menganalisis data akademik secara mendalam menggunakan teknik seperti *decision tree*, *classification*, dan *clustering* untuk menemukan pola yang dapat membantu mencegah mahasiswa meninggalkan studi mereka sebelum lulus.

Sebagai contoh, penelitian oleh Naruhiko Shiratori [1] di universitas-universitas Jepang menggunakan model *logistic regression* untuk mengklasifikasikan mahasiswa ke dalam dua kategori utama: *preliminary dropout state* (kemungkinan besar akan putus sekolah) dan *normal state* (kemungkinan rendah untuk putus sekolah). Model ini memanfaatkan data harian, seperti riwayat pelajaran dan nilai rata-rata, serta data sebelum penerimaan mahasiswa. Penelitian ini menemukan bahwa mahasiswa yang masuk ke dalam *preliminary dropout state* beberapa kali memiliki probabilitas lebih tinggi untuk tidak lulus. Dari 719 mahasiswa yang diteliti, 95,1% mahasiswa yang tidak pernah masuk kategori tersebut berhasil lulus. Sebaliknya, hanya 24,4% dari mereka yang masuk kategori dua kali yang berhasil lulus, sementara tidak ada satu pun mahasiswa yang masuk kategori empat kali atau lebih yang berhasil lulus.

Penelitian di Universitas Budi Luhur menganalisis data akademik mahasiswa dari tahun 2018 hingga 2022 untuk mengidentifikasi potensi *drop out*. Data ini diproses melalui tahapan *data cleaning* untuk menghilangkan nilai kosong dan *data preparation* untuk memastikan konsistensi data. Algoritma C4.5, yang merupakan salah satu metode *decision tree*, digunakan untuk memodelkan pola berdasarkan atribut seperti IPK, jumlah SKS yang diambil, dan masa studi. Hasil pengujian menunjukkan bahwa validasi silang (*cross-validation*) dengan 10 lipatan (*folds*) memberikan akurasi terbaik, sehingga menunjukkan potensi besar dalam membantu institusi mendeteksi mahasiswa berisiko secara dini [2].

Penelitian lain oleh Sanjaya dan rekan-rekan [3] di STMIK Primakara menggunakan metode serupa dengan mengintegrasikan analisis berbasis *Knowledge Discovery in Databases* (KDD). Dalam studi tersebut, data mahasiswa dari tahun 2017 hingga 2020 dianalisis untuk mendeteksi risiko *drop out*. Dengan menerapkan algoritma C4.5, penelitian ini menghasilkan aturan atau pola yang dapat digunakan institusi untuk merancang kebijakan akademik, seperti program mentoring atau pemberian beasiswa. Selain itu, regresi logistik digunakan untuk memprediksi potensi mahasiswa *drop out* berdasarkan data akademik dan latar belakang ekonomi. Studi ini menunjukkan bahwa kehadiran kurang dari 75%, kondisi ekonomi keluarga, dan kurangnya keterlibatan dalam kegiatan kampus adalah faktor dominan. Model ini menghasilkan akurasi prediksi sebesar 78% dan digunakan sebagai dasar untuk program intervensi seperti beasiswa dan pendampingan akademik.

Implementasi *data* [23]*ng* ini tidak hanya membantu institusi pendidikan dalam mendeteksi mahasiswa yang berisiko tinggi, tetapi juga memberikan wawasan yang mendalam tentang faktor-fa[37]yang memengaruhi kelulusan mahasiswa. Data seperti IPK, frekuensi kehadiran, dan latar belakang demografis dapat digunakan untuk membuat keputusan berbasis data yang lebih baik ([3], [2]). Pendekatan proaktif berbasis teknologi ini terbukti secara signifikan meningkatkan tingkat kelulusan mahasiswa. Selain itu, teknologi ini memungkinkan terciptanya pendidikan berbasis data, di mana lembaga dapat secara efektif mengalokasikan sumber daya untuk intervensi yang tepat sasaran. Dengan memanfaatkan algoritma prediksi, institusi dapat memberikan dukungan yang lebih baik kepada mahasiswa, baik dalam bentuk bimbingan akademik yang lebih intensif maupun intervensi finansial, sehingga membantu mereka menyelesaikan studi dengan sukses [3].

Penelitian yang dilakukan oleh Budiyantara dan rekan-rekan [4] di sebuah universitas negeri di Indonesia menggunakan algoritma *decision tree* untuk menganalisis risiko *drop out* mahasiswa. Penelitian ini menunjukkan bahwa faktor kehadiran di bawah 75%, IPK di bawah 2,5, dan status ekonomi keluarga rendah merupakan prediktor utama risiko *drop out*. Dengan akurasi prediksi 82%, model ini menjadi acuan untuk memberikan intervensi, seperti pembe[31]beasiswa atau program bimbingan akademik.

Berdasarkan latar belakang tersebut, penelitian ini menggunakan metode klasifikasi *logistic regression* dan *decision tree* untuk memprediksi siswa yang berpotensi putus sekolah sebagai rekomendasi bagi pihak sekolah. Proses

Commented [R4]: Cek dan pastikan lagi utk gaya sitasinya apakah pakai IEEE?

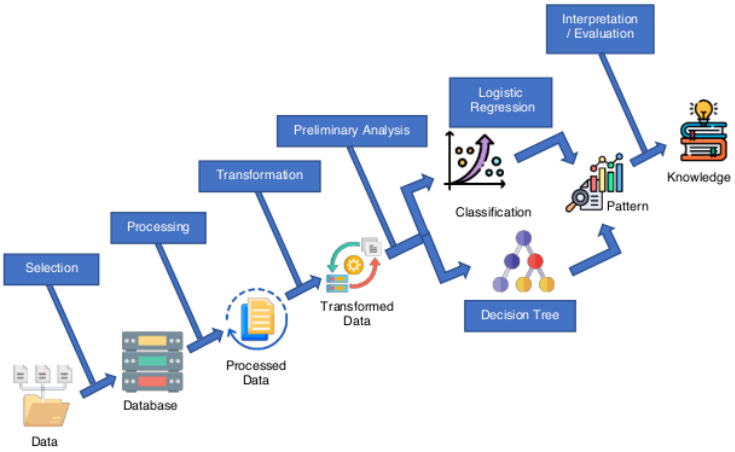
penelitian meliputi pengolahan data siswa, yang dimulai dari tahap persiapan data hingga analisis lebih lanjut. Selanjutnya, penelitian ini akan menguji performa kedua model untuk mengevaluasi keakuratan prediksi yang dihasilkan. Hasilnya diharapkan dapat memberikan dasar yang kuat bagi sekolah dalam mengambil keputusan preventif terhadap risiko putus sekolah.

II. METODE

Pada penelitian ini teknik yang digunakan adalah metode *data mining*. *Data mining* atau penambangan data adalah proses pengumpulan informasi atau data penting yang diharapkan oleh suatu pihak melalui sebuah data yang lebih besar. Metode matematika, statistika, hingga teknologi *artificial intelligence* (AI) sering dimanfaatkan dalam proses pengumpulan informasi tersebut. Teknik data mining digunakan untuk membangun model yang berfungsi mengenali informasi baru berdasarkan data yang belum diketahui [5]. semua teknik data mining memiliki satu kesamaan, yaitu penemuan otomatis hubungan baru dan ketergantungan antar atribut dalam data yang diamati. Jika tujuan analisis adalah pengelompokan data berdasarkan kelas, maka informasi baru yang dihasilkan adalah tentang kelas tempat data tersebut berada. Pola yang sebelumnya tidak tampak atau tidak begitu jelas diekstraksi menggunakan *data mining*. Teknik data mining seperti klasterisasi, hubungan antar variabel, hingga model prediksi telah diterapkan untuk menganalisis data ini, terutama untuk memprediksi siswa yang berpotensi gagal atau putus sekolah [6]

Data mining memiliki banyak sekali fungsi. Untuk fungsi utama sendiri ada dua yaitu fungsi *descriptive* dan fungsi *predictive* [7]. Fungsi *descriptive* atau deskripsi adalah sebuah fungsi untuk memahami lebih jauh tentang data yang diamati. Sedangkan fungsi *predictive* atau prediksi adalah fungsi bagaimana sebuah proses nantinya akan menemukan pola tertentu dari suatu data. Pola-pola tersebut dapat diketahui dari berbagai variabel yang ada pada data. Pola yang sudah ditemukan tersebut dapat dipakai untuk memprediksi variabel lain yang belum diketahui nilai ataupun jenisnya. Oleh karena alasan tersebut, fungsi ini memudahkan dan menguntungkan bagi siapapun yang memerlukan prediksi yang akurat.

Data mining juga dikenal dengan istilah *knowledge discovery in databases* atau KDD. Banyak sekali teknik dan konsep yang dapat diaplikasikan pada aktivitas data mining. Agar dapat memperoleh data sesuai dengan yang diharapkan, proses tersebut memerlukan sejumlah langkah. Proses KDD terdiri dari *selection, processing, transformation, data mining, dan interpretation/evaluation* [8] Tahapan proses KDD dalam *Data Mining* dapat dilihat pada gambar 1.



Gambar 1. Tahapan Proses KDD dalam Data Mining

A. *Selection*

Selection data adalah proses seleksi data yang sesuai untuk dianalisis, yang diambil dari sumber data atau basis data (*database*). Seleksi data dilakukan karena tidak semua data yang ada pada basis data relevan atau diperlukan dalam analisis. Proses pemilihan data yang sesuai ini dilakukan sebelum memasuki tahap penggalian informasi dalam *Knowledge Discovery in Databases* (KDD). Tahap ini bertujuan untuk memilih data yang relevan dan representatif, yang akan digunakan dalam proses *data mining* [9]. Pada tahapan ini, data melalui proses pemilahan, di mana data yang digunakan adalah data peserta didik dari angkatan 2022 hingga 2024 dengan total 963 siswa. Dengan adanya proses seleksi ini, pengolahan data dapat berlangsung lebih baik, efektif, dan sesuai dengan tujuan penelitian.

B. *Processing*

Processing data merupakan proses pembersihan data yang mencakup penanganan data kosong, kurang, maupun yang mengandung kesalahan. Sebelum proses *data mining* dapat dilakukan, terdapat serangkaian tahapan yang perlu dilalui untuk membersihkan dan mempersiapkan informasi yang menjadi fokus dalam *Knowledge Discovery in Databases* (KDD). Tahapan ini meliputi proses pembersihan data, verifikasi terhadap data yang kontradiktif, serta perbaikan informasi yang mungkin mengandung kesalahan tipografi [9]. Pada tahap ini, sebanyak 963 data siswa diproses, dan setelah pembersihan data, diperoleh 959 data siswa yang dapat digunakan. Dari 14 variabel yang tersedia, sebanyak 11 variabel dipilih untuk digunakan dalam penelitian ini.

C. *Transformation*

Transformation data dilakukan setelah tahap *Processing* dan *Cleaning Data*, yaitu tahap untuk menghasilkan data yang siap ditambang dengan mengubah data dari bentuk asalnya sesuai dengan kebutuhan. Data peserta didik ditransformasi menjadi format yang sesuai dengan kriteria yang digunakan dalam perhitungan, sehingga memudahkan proses penggalian data untuk menemukan pengetahuan baru. Proses *coding* dalam *Knowledge Discovery in Databases* (KDD) merupakan tahap penting dalam transformasi data. Pada tahap ini, data yang telah tersedia diubah atau dimodifikasi agar memenuhi persyaratan dan kebutuhan proses *data mining*. Tujuan dari proses *coding* adalah mengonversi data mentah ke dalam bentuk yang dapat diproses dan dianalisis menggunakan algoritma atau teknik *data mining* tertentu [9].

D. *Data Mining*

Data mining adalah proses menemukan pola atau informasi menarik pada data terpilih dengan menggunakan alat atau metode tertentu [10]. Dalam era transformasi digital, kebutuhan untuk memahami pola tersembunyi di dalam data semakin meningkat, seiring dengan pertumbuhan eksponensial volume data. Teknologi *data mining* telah menjadi alat yang penggung analitik di berbagai bidang, termasuk bisnis, pendidikan, kesehatan, dan ilmu sosial, dengan memberikan wawasan mendalam untuk mendukung pengambilan keputusan berbasis data [4].

Proses *data mining* terdiri dari berbagai tahapan, seperti pengumpulan data, prapemrosesan data, analisis, dan interpretasi hasil. Setiap tahapan ini membutuhkan kombinasi teknik yang melibatkan statistik, pembelajaran mesin (*machine learning*), dan kecerdasan buatan (*artificial intelligence*) [11]. Salah satu keunggulan utama dari teknologi ini adalah kemampuannya mengubah data yang tidak terstruktur menjadi informasi yang terorganisasi dan bermakna, sehingga dapat digunakan untuk berbagai keperluan analitik dan pengambilan keputusan.

E. *Logistic regression*

Pengklasifikasian adalah salah satu metode yang digunakan untuk mengelompokkan data yang telah tersusun secara sistematis. Salah satu metode pengelompokan yang banyak diterapkan dalam penelitian adalah *logistic regression*. *Logistic regression* merupakan teknik statistika yang digunakan untuk menganalisis data dengan tujuan mengetahui hubungan antara beberapa variabel. Dalam teknik ini, variabel respon bersifat kategorik, baik nominal maupun ordinal, sedangkan variabel penjelas dapat bersifat kategorik maupun kontinu [12]. Jika variabel respon terdiri atas dua kategori (*dikotomis*), analisis yang digunakan disebut *binary logistic regression*. Namun, jika variabel respon memiliki lebih dari dua kategori, maka digunakan *multinomial logistic regression*.

Dalam statistika, *logistic regression* digunakan untuk memahami hubungan antara variabel respon yang bersifat kategorik dengan variabel penjelas yang bersifat kontinu atau kategorik [12]. Penelitian ini secara khusus menggunakan *binary logistic regression*, yang bertujuan menganalisis hubungan antara variabel independen (X) dan variabel dependen (Y) yang memiliki nilai dikotomis atau biner. Menurut Harlan [13], *binary logistic regression* digunakan untuk mengukur pengaruh variabel independen terhadap variabel dependen, baik ketika variabel independen bersifat kontinu maupun kategorik.

F. *Decision tree*

Selain *logistic regression*, metode pengelompokan atau klasifikasi lain yang sering diterapkan adalah metode *decision tree* atau pohon keputusan. Klasifikasi *decision tree* menyediakan metode yang cepat dan efektif untuk mengelompokkan dataset [14]. Metode ini juga telah mengalami pengembangan untuk menghasilkan pendekatan yang dapat mengklasifikasikan data yang bersifat sensitif [15]. Hasil klasifikasi dari pohon keputusan tersebut kemudian digunakan untuk memprediksi siswa yang berpotensi putus sekolah.

Decision tree adalah salah satu metode yang dapat diterapkan untuk mengklasifikasikan data atau objek untuk menghasilkan sebuah keputusan. Pendekatan ini terdiri dari serangkaian node pilihan yang dihubungkan melalui

cabang, bergerak menurun dari simpul akar hingga mencapai simpul daun. Pengembangan *decision tree* dimulai dari simpul akar, yang terutama didasarkan pada konvensi yang diposisikan di bagian atas diagram pohon keputusan. Semua atribut dievaluasi pada simpul seleksi, dengan setiap outcome yang mungkin menghasilkan cabang. Setiap cabang dapat mengarah ke *decision node* lain atau ke *leaf node* [16].

Dalam konteks pendidikan, *decision tree* banyak digunakan untuk memprediksi berbagai fenomena, seperti tingkat kelulusan siswa, prediksi *dropout*, atau efektivitas metode pengajaran. Misalnya, model *decision tree* telah diterapkan untuk mengidentifikasi faktor-faktor yang memengaruhi mahasiswa untuk lulus tepat waktu atau putus sekolah, berdasarkan atribut seperti nilai akademik, latar belakang keluarga, dan tingkat kehadiran [4].

G. Interpretation/Evaluation

Interpretation/evaluation adalah tahap untuk melakukan evaluasi hasil setelah melalui semua proses *data mining* [17]. Tahap ini bertujuan untuk membahas hasil yang diperoleh dari *data mining* menggunakan *logistic regression* dan *decision tree*. Hasil tersebut kemudian dievaluasi atau diinterpretasikan agar dapat dipahami dengan mudah. Proses ini memastikan bahwa informasi yang dihasilkan dapat memberikan wawasan yang jelas dan bermanfaat.

III. HASIL DAN PEMBAHASAN

Hasil penelitian yang diperoleh melalui penerapan metode *Logistic Regression* dan *Decision Tree* dalam mendeteksi potensi siswa putus sekolah di SMA N 4 Tegal. Analisis dilakukan berdasarkan data siswa yang mencakup variabel seperti NIS, nama, tahun masuk, jenis kelamin, ketidakhadiran dalam satu semester, rata-rata nilai semester 1, jarak dari rumah (km), penghasilan orang tua, jenis tinggal, alat transportasi, status penerimaan KIP, anak keberapa, jumlah saudara kandung, dan status kelulusan (tetap selesai atau putus sekolah). Tidak semua variabel tersebut digunakan sebagai fitur untuk klasifikasi. Kolom NIS, nama, dan tahun masuk dikeluarkan sejak awal karena prediksi yang dilakukan tidak bersifat subjektif. Prediksi ini hanya berfokus pada hasil akhir siswa, yakni apakah tetap selesai sekolah atau putus sekolah.

Hasil dari masing-masing metode disajikan untuk menunjukkan performa model, akurasi, sensitivitas, dan presisi dalam mengidentifikasi siswa dengan risiko tinggi untuk putus sekolah. Selanjutnya, perbandingan antara kedua metode dilakukan untuk mengevaluasi tingkat akurasi masing-masing pendekatan, serta validitas model dalam konteks pendidikan.

A. Analisis Data

Pada tahap analisis data awal ini dilakukan eksplorasi terhadap dataset siswa untuk memahami pola dan hubungan antar-variabel menggunakan Matriks Korelasi. Dataset ini mencakup variabel-variabel jenis kelamin, ketidakhadiran dalam 1 semester, rata-rata nilai semester 1, jarak dari rumah (km), penghasilan orang tua, jenis tinggal, alat transportasi, penerima KIP atau bukan, anak keberapa, jumlah saudara kandung, dan tetap selesai atau putus sebagai goalnya seperti terlihat pada gambar 2.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 963 entries, 0 to 962
Data columns (total 11 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   JK                      963 non-null    object
 1   Ketidakhadiran1semester 963 non-null    int64
 2   Rata2Nilai1Semester1    963 non-null    float64
 3   JarakDariRumahKM        963 non-null    int64
 4   PenghasilanOrangTua     963 non-null    object
 5   JenisTinggal            963 non-null    object
 6   AlatTransportasi        963 non-null    object
 7   PenerimaKIP             963 non-null    object
 8   AnakKeberapa            959 non-null    float64
 9   JmlSaudaraKandung       963 non-null    int64
 10  TetapSelesaiAtauPutus   963 non-null    object
dtypes: float64(2), int64(3), object(6)
memory usage: 82.9+ KB
```

Gambar 2. Variabel pada dataset siswa

Langkah berikutnya pada tahap analisis data awal adalah mendeteksi nilai unik untuk kolom tertentu. Contohnya, pada kolom Penghasilan Orang Tua dengan kategori: "Kurang dari Rp500.000," "Rp500.000 - Rp999.999," "Rp1.000.000 - Rp1.999.999," "Rp2.000.000 - Rp4.999.999," "Rp5.000.000 - Rp20.000.000," dan "Lebih dari Rp20.000.000." Pada kolom Jenis Tinggal terdapat kategori seperti: "Bersama orang tua" dan "Wali." Untuk kolom Alat Transportasi mencakup kategori: "Jalan kaki," "Sepeda," "Sepeda motor," "Mobil pribadi," dan "Lainnya." Sementara itu, kolom Tetap Selesai atau Putus memiliki dua kategori, yaitu: "Tetap Selesai Sekolah" dan "Putus Sekolah." Data lima belas baris pertama ditampilkan sebagaimana terlihat pada Gambar 3.

	JK	KetidakhadiranSemester	Rata2NilaiSemester1	JarakDariRumahKM	PenghasilanOrangTua	JenisTinggal	AlatTransportasi	PenerimaKIP	AnakKeberapa	JmlSaudaraKandung	TetapSelesaiAtauPutus
0	L	0	82.43	3	Rp. 1,000,000 - Rp. 1,999,999	Bersama orang tua	Sepeda motor	Tidak	3.0	4	Tetap Selesai Sekolah
1	P	4	81.43	1	Rp. 1,000,000 - Rp. 1,999,999	Bersama orang tua	Sepeda motor	Tidak	2.0	3	Tetap Selesai Sekolah
2	L	0	82.43	1	Kurang dari Rp. 500,000	Bersama orang tua	Sepeda	Tidak	1.0	3	Tetap Selesai Sekolah
3	L	11	75.02	1	Rp. 2,000,000 - Rp. 4,999,999	Bersama orang tua	Lainnya	Tidak	1.0	0	Tetap Selesai Sekolah
4	L	3	75.19	1	Rp. 500,000 - Rp. 999,999	Bersama orang tua	Lainnya	Ya	1.0	0	Tetap Selesai Sekolah
5	P	3	80.79	1	Rp. 1,000,000 - Rp. 1,999,999	Bersama orang tua	Sepeda	Tidak	2.0	5	Tetap Selesai Sekolah
6	P	3	87.00	1	Rp. 1,000,000 - Rp. 1,999,999	Bersama orang tua	Jalan kaki	Tidak	3.0	2	Tetap Selesai Sekolah
7	P	5	81.07	0	Rp. 1,000,000 - Rp. 1,999,999	Bersama orang tua	Jalan kaki	Tidak	1.0	2	Tetap Selesai Sekolah
8	P	0	83.68	1	Kurang dari Rp. 500,000	Bersama orang tua	Sepeda motor	Tidak	1.0	2	Tetap Selesai Sekolah
9	P	0	87.14	1	Rp. 1,000,000 - Rp. 1,999,999	Bersama orang tua	Jalan kaki	Tidak	2.0	2	Tetap Selesai Sekolah
10	P	24	75.25	2	Rp. 1,000,000 - Rp. 1,999,999	Bersama orang tua	Mobil pribadi	Tidak	1.0	1	Putus Sekolah
11	L	2	82.86	1	Rp. 1,000,000 - Rp. 1,999,999	Bersama orang tua	Sepeda motor	Tidak	1.0	2	Tetap Selesai Sekolah
12	L	0	87.93	1	Rp. 1,000,000 - Rp. 1,999,999	Bersama orang tua	Sepeda motor	Tidak	5.0	2	Tetap Selesai Sekolah
13	P	7	76.14	1	Rp. 500,000 - Rp. 999,999	Bersama orang tua	Sepeda	Tidak	1.0	2	Tetap Selesai Sekolah
14	L	0	83.64	1	Rp. 2,000,000 - Rp. 4,999,999	Bersama orang tua	Mobil pribadi	Tidak	1.0	4	Tetap Selesai Sekolah

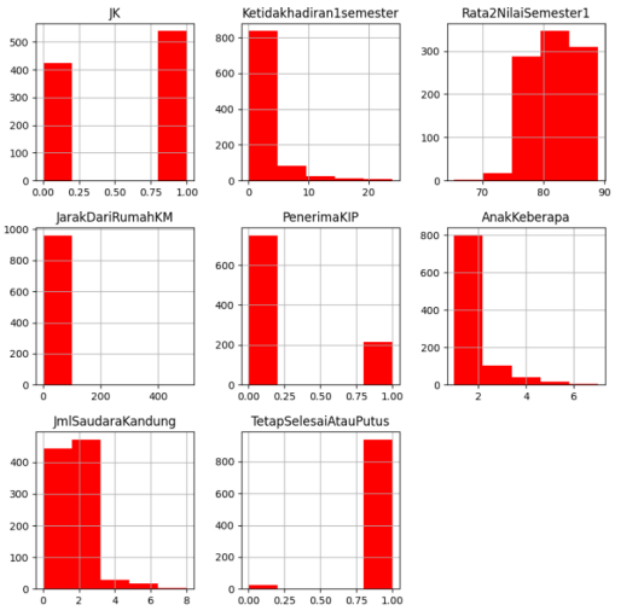
Gambar 3. Tampilan data 15 baris pertama

Variabel yang bernilai string diubah menjadi data integer menggunakan teknik *one-hot encoding* atau *label encoding*. Teknik *one-hot encoding* atau *label encoding* mengonversi sebuah daftar string menjadi angka berdasarkan urutan abjad. Pada beberapa kolom, nilai string diubah menjadi integer secara manual tanpa menggunakan *one-hot encoding* atau *label encoding*, karena semakin besar nilainya seharusnya merepresentasikan angka yang semakin besar. Hal ini diterapkan pada kolom Penghasilan Orang Tua, Jenis Tinggal, dan Alat Transportasi.

	JK	KetidakhadiranSemester	Rata2NilaiSemester1	JarakDariRumahKM	PenghasilanOrangTua	JenisTinggal	AlatTransportasi	PenerimaKIP	AnakKeberapa	JmlSaudaraKandung	TetapSelesaiAtauPutus
0	0	0	82.43	3	3	1	2	0	3.0	4	1
1	1	4	81.43	1	3	1	2	0	2.0	3	1
2	0	0	82.43	1	1	1	1	0	1.0	3	1
3	0	11	75.02	1	4	1	4	0	1.0	0	1
4	0	3	75.19	1	2	1	4	1	1.0	0	1
5	1	3	80.79	1	3	1	1	0	2.0	5	1
6	1	3	87.00	1	3	1	0	0	3.0	2	1
7	1	5	81.07	0	3	1	0	0	1.0	2	1
8	1	0	83.68	1	1	1	2	0	1.0	2	1
9	1	0	87.14	1	3	1	0	0	2.0	2	1
10	1	24	75.25	2	3	1	3	0	1.0	1	0
11	0	2	82.86	1	3	1	2	0	1.0	2	1
12	0	0	87.93	1	3	1	2	0	5.0	2	1
13	1	7	76.14	1	2	1	1	0	1.0	2	1
14	0	0	83.64	1	4	1	3	0	1.0	4	1

Gambar 4. Tampilan data 15 baris pertama yang sudah berubah menjadi nilai integer

Gambar 4 merupakan dataset siswa yang sudah berubah menjadi nilai integer, ditampilkan data lima belas baris pertama. Setelah semua data sudah berubah menjadi nilai integer selanjutnya melakukan visualisasi diagram untuk masing-masing kolom menggunakan histogram seperti terlihat pada gambar 5.



Gambar 5. Tampilan histogram awal

Dengan melakukan visualisasi histogram untuk setiap kolom, kita dapat memahami karakteristik awal dataset, mengidentifikasi distribusi data, serta mendeteksi pola atau anomali yang mungkin memengaruhi analisis lebih lanjut. Pada Gambar 5, histogram hanya menampilkan 8 dari 11 kolom, yang menunjukkan bahwa masih terdapat 3 kolom yang belum divisualisasikan. Oleh karena itu, penting untuk mengetahui tipe data dari setiap kolom agar analisis dapat dilakukan secara menyeluruh dan akurat.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 963 entries, 0 to 962
Data columns (total 11 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   JK                      963 non-null    int64
 1   Ketidakhadiran1semester  963 non-null    int64
 2   Rata2NilaiSemester1     963 non-null    float64
 3   JarakDariRumahKM        963 non-null    int64
 4   PenghasilanOrangTua     963 non-null    object
 5   JenisTinggal            963 non-null    object
 6   AlatTransportasi        963 non-null    object
 7   PenerimaKIP             963 non-null    int64
 8   AnakKeberapa            959 non-null    float64
 9   JmlSaudaraKandung       963 non-null    int64
 10  TetapSelesaiAtauPutus   963 non-null    int64
dtypes: float64(2), int64(6), object(3)
memory usage: 82.9+ KB
```

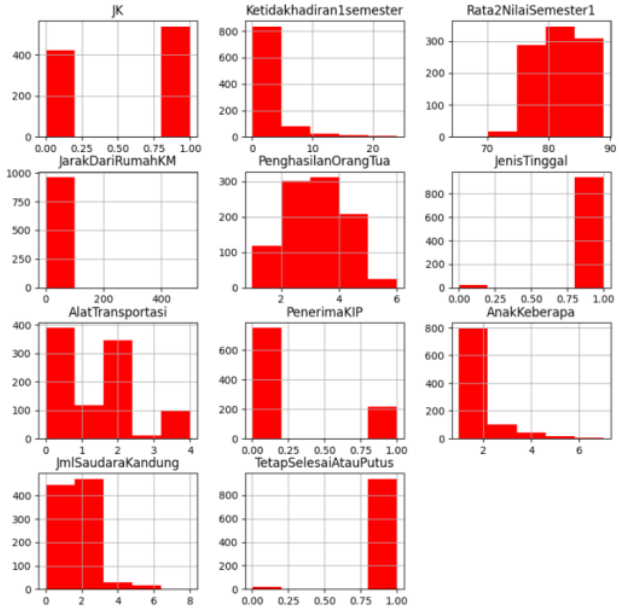
(a)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 963 entries, 0 to 962
Data columns (total 11 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   JK                      963 non-null    int64
 1   Ketidakhadiran1semester  963 non-null    int64
 2   Rata2NilaiSemester1     963 non-null    float64
 3   JarakDariRumahKM        963 non-null    int64
 4   PenghasilanOrangTua     963 non-null    int64
 5   JenisTinggal            963 non-null    int64
 6   AlatTransportasi        963 non-null    int64
 7   PenerimaKIP             963 non-null    int64
 8   AnakKeberapa            959 non-null    float64
 9   JmlSaudaraKandung       963 non-null    int64
 10  TetapSelesaiAtauPutus   963 non-null    int64
dtypes: float64(2), int64(9)
memory usage: 82.9 KB
```

(b)

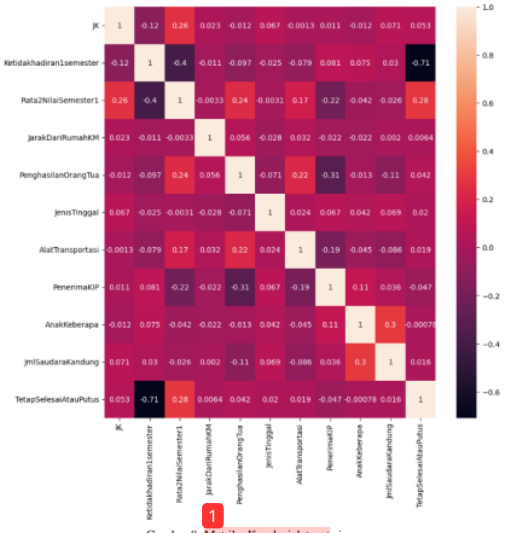
Gambar 6. Tipe kolom (a) sebelum dilakukan perubahan dan (b) setelah dilakukan perubahan

Gambar 6 menunjukkan perubahan tipe kolom sebelum dan sesudah dilakukan modifikasi. Pada bagian (a), tipe kolom masih bertipe object, yang sering digunakan untuk data teks atau karakter. Setelah dilakukan perubahan, seperti ditampilkan pada bagian (b), tipe kolom tersebut diubah menjadi integer untuk mempermudah pengolahan data numerik. Dengan demikian, ketika divisualisasikan kembali dalam bentuk histogram, semua kolom dapat terlihat dengan jelas, seperti yang ditunjukkan pada Gambar 7.



Gambar 7. Tampilan histogram semua kolom

Setelah semua kolom terlihat, langkah selanjutnya adalah visualisasi Matriks Korelasi. Matriks Korelasi digunakan untuk mengukur hubungan linier antara variabel numerik, dengan koefisien korelasi Pearson (rr) sebagai ukuran hubungan. Nilai rr berkisar antara -1 hingga 1, dengan r=1adalah Korelasi positif sempurna, r=−1 adalah Korelasi negatif sempurna, dan r=0 adalah Tidak ada korelasi. Visualisasi Matriks Korelasi seperti yang ditunjukkan pada Gambar 8 memudahkan dalam memahami pola hubungan antar-variabel yang diuji.



Gambar 8. Matriks Korelasi dataset siswa

Hasil visualisasi Matriks Korelasi pada Gambar 8 dari dataset siswa menunjukkan adanya korelasi antara setiap variabel dalam dataset dengan variabel target, yaitu "Tetap Selesai atau Putus Sekolah." Meskipun tidak semua variabel memiliki korelasi yang kuat, terdapat dua variabel yang menunjukkan korelasi signifikan dengan variabel target, yaitu variabel "Ketidakhadiran dalam 1 Semester" dan "Rata-rata Nilai Semester 1." Hal ini ditunjukkan oleh nilai korelasi yang mendekati satu atau minus satu.

Variabel "Ketidakhadiran dalam 1 Semester" memiliki korelasi negatif yang kuat (r=−0.71) dengan status siswa tetap selesai sekolah, yang berarti semakin tinggi jumlah ketidakhadiran, semakin besar risiko putus sekolah. Sementara itu, variabel "Rata-rata Nilai Semester 1" menunjukkan korelasi positif yang cukup signifikan (r=0.28) dengan status siswa tetap selesai sekolah, yang menunjukkan bahwa siswa dengan nilai akademik rendah lebih cenderung putus sekolah.

B. Analisis Logistic Regression

Pada tahap ini, dilakukan penerapan metode *Logistic Regression* untuk menganalisis faktor-faktor yang memengaruhi kemungkinan siswa putus sekolah berdasarkan dataset yang telah dianalisis sebelumnya. *Logistic Regression* dipilih karena mampu menangani masalah klasifikasi biner, yaitu menentukan apakah siswa akan "Tetap Selesai" atau "Putus Sekolah," berdasarkan variabel independen yang tersedia.

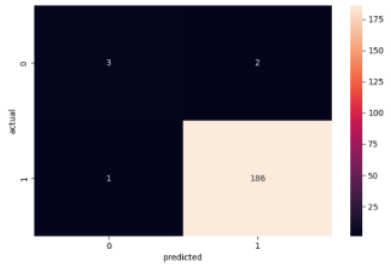
Selanjutnya, untuk memahami kinerja model, dataset dibagi menjadi *training set* dan *test set*. Pemisahan dataset dilakukan menggunakan fungsi *train_test_split()*. Proporsi *test set* yang digunakan adalah sebesar 20% dari total data. Setelah modul *Logistic Regression* diimpor, dibuat objek *classifier Logistic Regression* menggunakan fungsi *LogisticRegression()*. Model yang terbentuk kemudian digunakan untuk menghasilkan prediksi berdasarkan *test set*. Gambar 9 menunjukkan kelas *Logistic Regression* dan hasil perbandingan antara data asli dengan data prediksi, yang mengindikasikan apakah prediksi tersebut sudah sesuai atau belum.

```
LogisticRegression()

Accuracy 98.4375 %
```

Gambar 9. *Class Logistic Regression* dan hasil perbandingan data asli dengan data prediksi

Setelah model *Logistic Regression* terbentuk, langkah selanjutnya adalah menguji model tersebut untuk memprediksi hasil dari data tes. Kumpulan hasil prediksi tersebut dinyatakan sebagai *y_pred*. Tahap berikutnya adalah mengevaluasi model tersebut. Untuk mengetahui seberapa baik model yang terbentuk, perlu dibuat *confusion matrix* yang membandingkan hasil prediksi yang dilakukan oleh model dengan hasil sebenarnya dari data tes.



Gambar 10. *confusion matrix* dari *Logistic Regression*

Berdasarkan *confusion matrix* pada Gambar 10, diketahui bahwa model *Logistic Regression* berhasil memprediksi data tes, yang merupakan 20% dari total data (959) atau sekitar 192 data, dengan rincian sebagai berikut: *True Positive (TP)* sebanyak 186 siswa, yaitu siswa yang diprediksi sebagai "Tetap Selesai" dan memang benar-benar tetap selesai sesuai data aktual; *True Negative (TN)* sebanyak 3 siswa, yaitu siswa yang diprediksi sebagai "Putus Sekolah" dan benar-benar putus sekolah; *False Positive (FP)* sebanyak 1 siswa, yaitu siswa yang diprediksi sebagai "Tetap Selesai" tetapi sebenarnya putus sekolah; dan *False Negative (FN)* sebanyak 2 siswa, yaitu siswa yang diprediksi sebagai "Putus Sekolah" tetapi sebenarnya tetap selesai. Hasil perhitungan tersebut sesuai dengan perhitungan Google Colaboratory menggunakan fungsi *accuracy_score*() dengan perolehan nilai akurasi sebesar $\frac{189}{192} = 0.984375$ untuk model *Logistic Regression* yang terbentuk.

Hasil akurasi ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengklasifikasikan siswa yang tetap selesai maupun yang putus sekolah. Namun, akurasi saja tidak cukup untuk mengevaluasi model secara menyeluruh, terutama jika data tidak seimbang, sehingga metrik lain seperti presisi, recall, dan *F1-Score* juga perlu dipertimbangkan.

	precision	recall	f1-score	support
0	0.75	0.60	0.67	5
1	0.99	0.99	0.99	187
accuracy			0.98	192
macro avg	0.87	0.80	0.83	192
weighted avg	0.98	0.98	0.98	192

Gambar 11. Hasil *matrix* evaluasi dari *Logistic Regression*

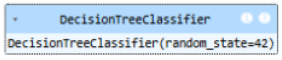
Gambar 11 menunjukan hasil metrik evaluasi dari *Logistic Regression*. Hasil evaluasi model *Logistic Regression* menggunakan metrik presisi, recall, dan *F1-score* menunjukkan perbedaan kinerja antara prediksi untuk kelas "Putus Sekolah" (0) dan "Tetap Selesai" (1).

Pada metrik presisi, model memiliki tingkat akurasi 75% dalam memprediksi siswa yang "Putus Sekolah," yang berarti 25% prediksi untuk kelas ini salah. Sebaliknya, untuk kelas "Tetap Selesai," presisi mencapai 99%, menunjukkan model hampir tidak membuat kesalahan dalam memprediksi siswa yang tetap melanjutkan sekolah. Pada metrik recall, model hanya mampu mendeteksi 60% dari siswa yang benar-benar "Putus Sekolah," yang

menunjukkan bahwa 40% siswa dalam kategori ini tidak teridentifikasi dengan benar dan justru diprediksi sebagai "Tetap Selesai." Sementara itu, untuk kelas "Tetap Selesai," recall sangat tinggi, yaitu 99%, menunjukkan hampir semua siswa yang tetap melanjutkan sekolah teridentifikasi dengan baik oleh model. Selanjutnya, metrik *F1-score*, yang menggabungkan presisi dan recall, menunjukkan bahwa model memiliki performa yang cukup baik untuk kelas "Tetap Selesai" dengan nilai 0.99, tetapi performa untuk kelas "Putus Sekolah" masih kurang optimal dengan nilai 0.67.

C. *Analisis Decision Tree*

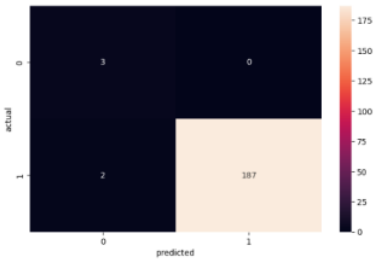
Melanjutkan proses *data mining* sebelumnya dengan menggunakan *Logistic Regression*, proses klasifikasi *Decision Tree* dilakukan pada dataset siswa yang telah melalui seleksi fitur dan dipisahkan antara *training set* dan *test set*. Pemisahan dataset dilakukan menggunakan fungsi *train_test_split()* dengan proporsi *test set* sebesar 20% dari total data. Setelah modul *Decision Tree* diimpor, dibuat objek *classifier Decision Tree* menggunakan fungsi *DecisionTreeClassifier()*. Model yang terbentuk kemudian digunakan untuk menghasilkan prediksi berdasarkan *test set*. Gambar 12 menunjukkan kelas *Decision Tree* dan hasil perbandingan antara data asli dengan data prediksi, yang mengindikasikan apakah prediksi tersebut sudah sesuai atau belum.




```
Accuracy 98.95833333333334 %
```

Gambar 12. *Class Decision Tree* dan hasil perbandingan data asli dengan data prediksi

Setelah model *Decision Tree* terbentuk, langkah selanjutnya adalah menguji model tersebut untuk memprediksi hasil dari data tes. Kumpulan hasil prediksi tersebut dinyatakan sebagai *y_pred _dec_tree*. Tahap berikutnya adalah mengevaluasi model tersebut. Untuk mengetahui seberapa baik model yang terbentuk, perlu dibuat *confusion matrix* yang membandingkan hasil prediksi yang dilakukan oleh model dengan hasil sebenarnya dari data tes.



Gambar 13. *confusion matrix* dari *Decision Tree*

Berdasarkan *confusion matrix* pada Gambar 13, diketahui bahwa model *Decision Tree* berhasil memprediksi data tes, yang merupakan 20% dari total data (959) atau sekitar 192 data, dengan rincian sebagai berikut: *True Positive (TP)* sebanyak 187 siswa, yaitu siswa yang diprediksi sebagai "Tetap Selesai" dan memang benar-benar tetap selesai sesuai data aktual; *True Negative (TN)* sebanyak 3 siswa, yaitu siswa yang diprediksi sebagai "Putus Sekolah" dan benar-benar putus sekolah; *False Positive (FP)* sebanyak 2 siswa, yaitu siswa yang diprediksi sebagai "Tetap Selesai" tetapi sebenarnya putus sekolah; dan *False Negative (FN)* sebanyak 0 siswa, yaitu siswa yang diprediksi sebagai "Putus Sekolah" tetapi sebenarnya tetap selesai. Hasil perhitungan tersebut sesuai dengan perhitungan Google Colaboratory menggunakan fungsi *accuracy_score*() dengan perolehan nilai akurasi sebesar $\frac{190}{192} \cong 0.9895$ untuk model *Decision Tree* yang terbentuk.

	precision	recall	f1-score	support
0	1.00	0.60	0.75	5
1	0.99	1.00	0.99	187
accuracy			0.99	192
macro avg	0.99	0.80	0.87	192
weighted avg	0.99	0.99	0.99	192

Gambar 14. Hasil *matrix* evaluasi dari *Decision Tree*

Gambar 14 menunjukan hasil metrik evaluasi dari *Decision Tree*. Hasil evaluasi model *Decision Tree* menunjukkan performa yang sangat baik pada kelas mayoritas "Tetap Selesai," tetapi kinerjanya untuk kelas minoritas "Putus Sekolah" masih dapat ditingkatkan.

Dari segi presisi, model memiliki nilai sempurna (1.00) untuk kelas "Putus Sekolah," yang berarti semua siswa yang diprediksi sebagai "Putus Sekolah" benar-benar sesuai dengan data aktual. Sementara itu, presisi untuk kelas "Tetap Selesai" juga sangat tinggi, yaitu 0.99, yang menunjukkan tingkat kesalahan prediksi yang sangat kecil. Namun, pada metrik recall, model hanya berhasil mendeteksi 60% dari siswa yang benar-benar "Putus Sekolah," sehingga masih ada 40% siswa dalam kategori ini yang tidak teridentifikasi dengan benar dan diprediksi sebagai "Tetap Selesai." Sebaliknya, recall untuk kelas "Tetap Selesai" mencapai nilai sempurna (1.00), menunjukkan bahwa semua siswa yang benar-benar tetap selesai berhasil terdeteksi dengan tepat.

Pada metrik *F1-score*, yang merupakan rata-rata harmonis dari presisi dan recall, nilai untuk kelas "Putus Sekolah" adalah 0.75. Hal ini mencerminkan bahwa meskipun model memiliki presisi yang sangat tinggi, performa keseluruhan untuk kelas ini terbatas oleh recall yang rendah. Sebaliknya, *F1-score* untuk kelas "Tetap Selesai" mencapai 0.99, menegaskan bahwa model bekerja hampir sempurna pada kelas mayoritas.

IV. SIMPULAN

Berdasarkan penelitian yang telah dilakukan, model *Logistic Regression* dan *Decision Tree* menunjukkan kemampuan yang signifikan dalam mendeteksi potensi siswa putus sekolah di SMA N 4 Tegal. Hasil analisis menunjukkan bahwa *Decision Tree* memiliki kinerja yang lebih baik dalam hal presisi, terutama pada prediksi siswa yang tetap melanjutkan sekolah, dengan nilai presisi mencapai 0.99 untuk kelas "Tetap Selesai" dan nilai sempurna (1.00) untuk kelas "Putus Sekolah." Namun, pada metrik recall, *Logistic Regression* menunjukkan performa yang lebih seimbang dalam mendeteksi siswa yang tetap selesai maupun putus sekolah.

Pada analisis data awal menggunakan matriks korelasi, ditemukan bahwa variabel-variabel seperti ketidakhadiran siswa, rata-rata nilai semester, penghasilan orang tua, dan jenis transportasi memiliki hubungan signifikan terhadap keputusan siswa untuk tetap bersekolah atau putus sekolah. Model *Decision Tree* menunjukkan sensitivitas yang lebih tinggi terhadap pola dalam data, tetapi kinerjanya terhadap kelas minoritas seperti "Putus Sekolah" masih terbatas, terutama dalam recall yang hanya mencapai 0.60. Hal ini menunjukkan perlunya penyesuaian lebih lanjut untuk meningkatkan akurasi model, khususnya untuk kelas minoritas yang penting dalam konteks pendidikan.

Secara keseluruhan, penelitian ini menunjukkan bahwa kedua metode dapat menjadi alat yang efektif dalam mendukung pihak sekolah untuk memonitor potensi siswa yang berisiko putus sekolah. *Logistic Regression* memberikan interpretasi yang lebih mudah terhadap hubungan antar variabel, sedangkan *Decision Tree* menawarkan keunggulan dalam memetakan pola data secara visual dan prediktif. Implementasi model ini dapat memberikan kontribusi signifikan dalam pengambilan keputusan berbasis data untuk mengurangi tingkat putus sekolah, dengan tetap memperhatikan perlunya pengolahan data yang seimbang dan pemilihan model yang sesuai dengan kebutuhan analisis.

DAFTAR PUSTAKA

[1] N. Shiratori, "Derivation of Student Patterns in a Preliminary Dropout State and Identification of Measures for Reducing Student Dropouts," in *Proceedings - 2018 7th International Congress on Advanced Applied Informatics, IIAI-AAI 2018*, Institute of Electrical and Electronics Engineers Inc., Jul. 2018, pp. 497–500. doi: 10.1109/IIAI-AAI.2018.00108.

[2] Debora; R. M. et al., "PENERAPAN ALGORITME C.45 UNTUK KLASIFIKASI MAHASISWA BERPOTENSI DROP OUT PADA UNIVERSITAS BUDI LUHUR," *SENAFTI*, vol. 2, no. 1, pp. 316–325, Apr. 2023.

[3] Sanjaya; D. et al., "Penerapan Data Mining untuk Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Algoritma C4.5: Studi Kasus STMIK Primakara," *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, vol. 6, no. 1, pp. 84–97, Jun. 2022.

[4] Agus Budiyantara et al., "KOMPARASI ALGORITMA DECISION TREE, NAIVE BAYES DAN K-NEAREST NEIGHBOR UNTUK MEMPREDIKSI MAHASISWA LULUS TEPAT WAKTU," *JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER*, vol. 5, pp. 265–270, Feb. 2020.

[5] E. ; Osmanbegovic and M. Suljic, "Data Mining Approach for Predicting Student Performance," 2012. [Online]. Available: https://hdl.handle.net/10419/193806

[6] K. O. T. U. Otgontsetseg Sukhbaatar, "Mining Educational Data to Predict Academic Dropouts: a Case Study in Blended Learning Course," *Proceedings of TENCON*, vol. 10, pp. 2205–2208, Oct. 2018.

[7] R. Amalia, "Penerapan Data Mining untuk Memprediksi Hasil Kelulusan Siswa Menggunakan Metode Naïve Bayes," *JUISI*, vol. 06, no. 01, 2020.
[8] P. B. Mayank Pareek, "A REVIEW REPORT ON KNOWLEDGE DISCOVERY IN DATABASES AND VARIOUS TECHNIQUES OF DATA MINING," *OAIJSE*, vol. 5, no. 12, pp. 79–82, Dec. 2020.
[9] M. Nurizki, W. Apriandari, and A. Asriyanik, "Algoritma Naïve Bayes untuk Rekomendasi Seleksi Peserta Paskibraka Berbasis Website," *Journal of Information System Research (JOSH)*, vol. 4, no. 4, pp. 1486–1493, Jul. 2023, doi: 10.47065/josh.v4i4.3574.
[10] M. Atalya, A. Leza, W. Utami, P. Anugrah, and C. Dewi, "PREDIKSI PRESTASI SISWA SMAS KATOLIK SANTO YOSEPH DENPASAR BERDASARKAN KEDISIPLINAN DAN TINGKAT EKONOMI ORANG TUA MENGGUNAKAN METODE KNOWLEDGE DISCOVERY IN DATABASE DAN ALGORITMA REGRESI LINIER BERGANDA," 2024.
[11] A. N. Putri, N. Wakhidah, and V. G. Utomo, "Pemanfaatan Data Mining untuk Media Pembelajaran di SMK Hidayah Semarang," *Jurnal Pengabdian kepada Masyarakat*, vol. 13, no. 3, pp. 487–491, [Online]. Available: http://journal.upgris.ac.id/index.php/e-dimas
[12] D. Yuniarti and dan Rito Goejantoro, "Perbandingan Metode Klasifikasi Regresi Logistik Dengan Jaringan Saraf Tiruan (Studi Kasus: Pemilihan Jurusan Bahasa dan IPS pada SMAN 2 Samarinda Tahun Ajaran 2011/2012) Comparison of Classification Methods Between Logistic Regression and Artificial Neural Network (Case Study: Selection of Language and Social Studies Depertement at SMAN 2 Samarinda academic year 2011/2012)," *Jurnal EKSPONENSIAL*, vol. 4, no. 1, 2013.
[13] Y.- Brahmantyo, R. Riaman, and F. Sukono, "Willingness to Pay of Fishermen Insurance Using Logistic Regression with Parameter Estimated by Maximum Likelihood Estimation Based on Newton Raphson Iteration," *Jurnal Matematika Integratif*, vol. 17, no. 1, p. 15, Aug. 2021, doi: 10.24198/jmi.v17.n1.32037.15-21.
[14] M. J. Aitkenhead, "A co-evolving decision tree classification method," *Expert Syst Appl*, vol. 34, no. 1, pp. 18–25, Jan. 2008, doi: 10.1016/j.eswa.2006.08.008.
[15] B. Aviad and G. Roy, "Classification by clustering decision tree-like classifier based on adjusted clusters," *Expert Syst Appl*, vol. 38, no. 7, pp. 8220–8228, Jul. 2011, doi: 10.1016/j.eswa.2011.01.001.
[16] Sugianto C. A., "PENERAPAN TEKNIK DATA MINING UNTUK MENENTUKAN HASIL SELEKSI MASUK SMAN 1 GIBEBER UNTUK SISWA BARU MENGGUNAKAN DECISION TREE," *TEDC*, vol. 9, no. 1, pp. 39–43, Jan. 2015.
[17] W. Yustanti, "Studi Komparasi Local Outlier Factor (LOF) dan Isolation Forest (IF) pada Analisis Anomali Kinerja Dosen," *Journal of Informatics and Computer Science*, vol. 06, 2024.

AdeS2-4Des2024

ORIGINALITY REPORT

23%	22%	7%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	journal.maranatha.edu Internet Source	13%
2	ejurnal.seminar-id.com Internet Source	2%
3	online-journal.unja.ac.id Internet Source	1%
4	Submitted to Submitted on 1688618575702 Student Paper	1%
5	www.researchgate.net Internet Source	1%
6	Zelvi Gustiana. "PERFORMANCE EVALUATION ALGORITMA C 4.5 PADA KLASIFIKASI DATA", Djtechno: Jurnal Teknologi Informasi, 2024 Publication	<1%
7	ejournal.itn.ac.id Internet Source	<1%
8	repository.its.ac.id Internet Source	<1%

9	Submitted to Universitas Katolik Indonesia Atma Jaya Student Paper	<1 %
10	Submitted to UIN Sultan Syarif Kasim Riau Student Paper	<1 %
11	Submitted to Higher Education Commission Pakistan Student Paper	<1 %
12	Submitted to Universiti Sains Malaysia Student Paper	<1 %
13	repository.undar.ac.id Internet Source	<1 %
14	123dok.com Internet Source	<1 %
15	ksp0fc.tatestreetart.com Internet Source	<1 %
16	repository.umy.ac.id Internet Source	<1 %
17	core.ac.uk Internet Source	<1 %
18	www.neliti.com Internet Source	<1 %
19	Abdul Rahim, Pareza Alam Jusia. "Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Algoritma Decision Tree	<1 %

Classifier", Indonesian Journal of Computer Science, 2024
Publication

20	Muhamad Faiz Harby, Eka Dyar Wahyuni, Nur Cahyo Wibowo. "REKOMENDASI STRATEGI PENJUALAN BUNDLING DI CAFE SZ POINT MENGGUNAKAN ALGORITMA FREQUENT PATTERN GROWTH", Jurnal Informatika dan Teknik Elektro Terapan, 2024 Publication	<1 %
21	apbsrilanka.org Internet Source	<1 %
22	docplayer.info Internet Source	<1 %
23	journal.upy.ac.id Internet Source	<1 %
24	www.scribd.com Internet Source	<1 %
25	apps.dtic.mil Internet Source	<1 %
26	bisnisindonesia.id Internet Source	<1 %
27	candranaya.blogspot.com Internet Source	<1 %
28	ichi.pro Internet Source	<1 %

29	issuu.com Internet Source	<1 %
30	journal.uny.ac.id Internet Source	<1 %
31	jurnal.mdp.ac.id Internet Source	<1 %
32	lib.unnes.ac.id Internet Source	<1 %
33	manhijismd.wordpress.com Internet Source	<1 %
34	repositori.usu.ac.id Internet Source	<1 %
35	repository.ub.ac.id Internet Source	<1 %
36	www.kebijakanaidsindonesia.net Internet Source	<1 %
37	Akram Kemal Dewantara. "Optimasi Pengambilan Keputusan dengan Neural Network: Menuju Era Keputusan Pintar", The Indonesian Journal of Computer Science, 2024 Publication	<1 %
38	Daniel Zapata-Medina, Albeiro Espinosa-Bedoya, Jovani Alberto Jiménez-Builes. "Improving the Automatic Detection of Dropout Risk in Middle and High School	<1 %

Students: A Comparative Study of Feature Selection Techniques", Mathematics, 2024

Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		