Jurnal Informatika: Jurnal pengembangan IT

Vol. 9, No. 3, 2024
ISSN : 2477-5126, DOI:10.30591/jpit.v9i3.8007

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

Ade Ermillian, Kristiawan Nugroho

Program Studi Magister Teknologi Informasi, Fakultas Teknologi Informasi dan Industri, Universitas Stikubank, Semarang
Jln. Trilomba Juang No 1, Kota Semarang, 50272, Indonesia

Info Artikel

Riwayat Artikel:

Received 2024-12-06
Revised 2024-12-14
Accepted 2024-12-15

Abstract – The phenomenon of student dropouts is one of the main challenges in education, influenced by various factors such as absenteeism, economic pressures on families, low academic performance, and lack of motivation. This issue not only affects the personal development of students but also tarnishes the reputation of educational institutions. Therefore, an innovative technology-based approach, such as data mining, is needed to detect students at risk of dropping out early. 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, 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 – Fenomena siswa putus sekolah menjadi salah satu tantangan utama dalam dunia pendidikan, yang dipengaruhi oleh berbagai faktor seperti ketidakhadiran, tekanan ekonomi keluarga, rendahnya nilai akademik, dan kurangnya motivasi belajar. Masalah ini tidak hanya berdampak pada perkembangan siswa tetapi juga pada citra institusi pendidikan. Oleh karena itu, diperlukan pendekatan inovatif berbasis teknologi, seperti data mining, untuk mendeteksi siswa yang berpotensi putus sekolah secara dini. 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 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. Meskipun pendidikan telah berkembang pesat, masalah siswa yang putus sekolah tetap menjadi salah satu tantangan utama yang dihadapi oleh banyak lembaga pendidikan. Hal ini mencerminkan adanya masalah yang mendasar dan membutuhkan solusi yang tepat agar siswa dapat menyelesaikan pendidikan dengan baik.

Jumlah siswa yang putus sekolah atau *drop out* cukup signifikan, yang biasanya dipengaruhi oleh berbagai faktor seperti rendahnya tingkat kehadiran, tekanan ekonomi keluarga, dan permasalahan 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 membantu sekolah dalam menangani permasalahan ini secara lebih terarah.

Penelitian ini dilakukan di Sekolah Menengah Atas (SMA) Negeri 4 Tegal, sebuah institusi pendidikan tingkat atas yang memiliki reputasi baik di Kota Tegal. Sekolah ini mulai menerima siswa baru sejak tahun ajaran 1989/1990. Dengan pengalaman yang panjang, SMA Negeri 4 Tegal terus berupaya meningkatkan mutu pendidikan dan pelayanan bagi para siswanya. Namun demikian, kasus siswa putus sekolah masih menjadi perhatian serius bagi pihak sekolah.

Berdasarkan data terbaru, jumlah siswa SMA Negeri 4 Tegal yang putus sekolah dari angkatan 2022 hingga angkatan 2024 mencapai 23 siswa. Jumlah siswa putus sekolah ini cukup mengkhawatirkan karena tidak hanya berdampak negatif pada siswa secara individu, tetapi juga memengaruhi citra institusi pendidikan secara keseluruhan. 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 termasuk dalam metode *decision tree*, digunakan untuk membangun model berdasarkan atribut seperti IPK, jumlah SKS yang diambil, dan durasi 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.

Penerapan *data mining* tidak hanya membantu institusi pendidikan dalam mendeteksi mahasiswa yang memiliki risiko tinggi untuk putus sekolah, tetapi juga memberikan wawasan mendalam terkait faktor-faktor yang memengaruhi kelulusan. Informasi seperti IPK, tingkat kehadiran, dan latar belakang demografis siswa dapat dimanfaatkan untuk mendukung pengambilan keputusan yang berbasis data ([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 pemberian beasiswa atau program bimbingan akademik.

Berdasarkan latar belakang tersebut, penelitian ini memanfaatkan metode klasifikasi *logistic regression* dan *decision tree* untuk memprediksi siswa yang berpotensi putus sekolah dan memberikan rekomendasi yang dapat digunakan oleh pihak sekolah. Proses 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.

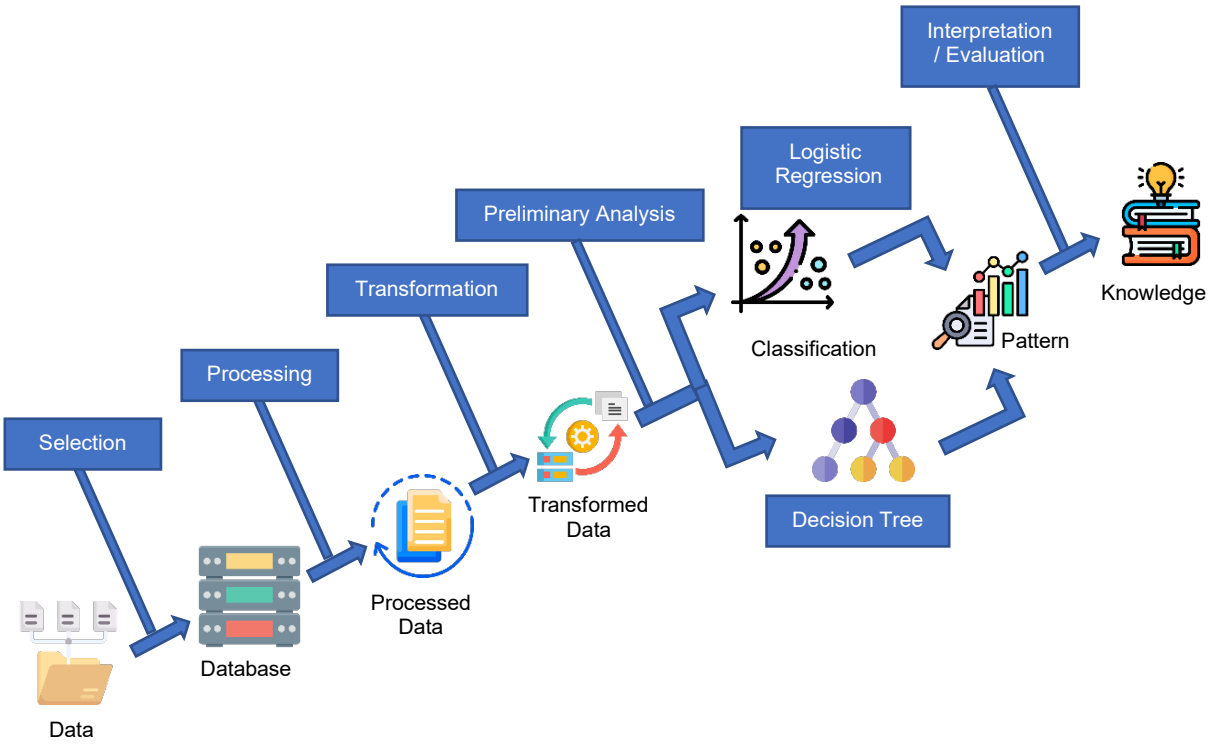
Penelitian ini bertujuan untuk mengembangkan model prediksi yang mampu mendeteksi potensi siswa putus sekolah secara akurat di SMA N 4 Tegal. Model yang dirancang menggunakan dua pendekatan utama, yaitu *Logistic Regression* dan *Decision Tree*, dengan memanfaatkan berbagai variabel seperti kehadiran siswa, rata-rata nilai semester, penghasilan orang tua, dan jenis transportasi. Melalui penelitian ini, diharapkan dapat diidentifikasi faktor-faktor signifikan yang memengaruhi risiko siswa putus sekolah serta membandingkan performa kedua metode dalam hal akurasi, presisi, dan sensitivitas. Hasil dari penelitian ini diharapkan dapat menjadi dasar pengambilan keputusan berbasis data bagi pihak sekolah untuk merancang strategi intervensi yang lebih efektif guna menekan angka putus sekolah.

II. METODE

Dalam penelitian ini teknik yang digunakan adalah metode *data mining*. *Data mining*, atau penambangan data, merupakan proses untuk memperoleh informasi atau data penting dari kumpulan data berukuran besar. Dalam proses ini, berbagai metode seperti matematika, statistika, hingga teknologi kecerdasan buatan (*artificial intelligence* atau AI) sering digunakan untuk menganalisis dan mendapatkan informasi yang dibutuhkan. 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-pola dalam data yang sebelumnya tidak terlihat atau sulit dikenali dapat diekstraksi dengan menggunakan teknik *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 berbagai fungsi, yang secara umum terbagi menjadi dua kategori utama, yaitu fungsi deskriptif (*descriptive*) dan fungsi prediktif (*predictive*) [7]. Fungsi deskriptif bertujuan untuk memahami secara mendalam data yang sedang diamati, sedangkan fungsi prediktif digunakan untuk menemukan pola tertentu dalam data yang dapat digunakan untuk memprediksi variabel lain yang belum diketahui nilainya. Dengan pola-pola ini, *data mining* menjadi alat yang sangat berguna untuk menghasilkan prediksi yang akurat, memberikan keuntungan bagi pengguna dalam pengambilan keputusan yang berbasis data.

Selain dikenal sebagai *data mining*, proses ini juga sering disebut sebagai *knowledge discovery in databases* (KDD). KDD melibatkan berbagai teknik dan konsep yang dapat diterapkan untuk menghasilkan informasi yang relevan. Proses ini terdiri dari beberapa langkah penting, yaitu seleksi (*selection*), pengolahan (*processing*), transformasi (*transformation*), penambangan data (*data mining*), dan evaluasi/interpretasi (*interpretation/evaluation*) [8]. Tahapan-tahapan ini memastikan bahwa data yang diproses dapat memberikan hasil yang sesuai dengan kebutuhan analisis, sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Proses *knowledge discovery in databases*

A. Selection

Proses *selection* data merupakan tahap pemilihan data yang relevan untuk dianalisis, yang diambil dari sumber data atau basis data (database). Langkah ini dilakukan karena tidak semua data dalam basis data bersifat penting atau diperlukan dalam analisis. Pemilihan data yang sesuai dilakukan sebelum tahap penggalian informasi dalam proses *Knowledge Discovery in Databases* (KDD). Tujuan utama dari tahap ini adalah memilih data yang relevan dan representatif untuk digunakan dalam proses *data mining* [9]. Pada penelitian ini, data yang diseleksi mencakup data siswa dari angkatan 2022 hingga 2024 dengan total 963 siswa. Proses seleksi dilakukan untuk memastikan hanya data yang relevan dan representatif yang diikutsertakan dalam analisis, sehingga hasil penelitian dapat menggambarkan kondisi nyata di SMA N 4 Tegal. Dengan adanya seleksi ini, pengolahan data dapat dilakukan secara lebih efisien dan tepat sasaran, 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 dilaksanakan, diperlukan beberapa tahapan untuk membersihkan dan mempersiapkan informasi yang akan dianalisis dalam KDD. Tahapan ini mencakup proses pembersihan data, verifikasi terhadap data yang bertentangan, dan perbaikan informasi yang mungkin mengandung kesalahan seperti kesalahan tipografi. Langkah-langkah ini memastikan bahwa data yang digunakan dalam analisis memiliki kualitas yang baik dan bebas dari inkonsistensi. [9]. Tahap ini mencakup identifikasi dan penanganan data yang tidak lengkap, duplikat, atau inkonsisten. Data yang memiliki nilai kosong atau kontradiktif diperbaiki atau dihapus sesuai kebutuhan agar dataset tetap berkualitas tinggi dan bebas dari kesalahan. 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

Transformasi data dilakukan setelah tahap *processing* dan *cleaning*. Tahap ini bertujuan untuk mengubah data dari format aslinya ke format yang sesuai dengan kebutuhan analisis, sehingga data siap untuk dieksplorasi lebih lanjut. Dalam konteks penelitian ini, data siswa ditransformasi menjadi format yang sesuai dengan kriteria tertentu, seperti yang dibutuhkan dalam algoritma *data mining*. Transformasi ini juga mencakup proses pengkodean, di mana data mentah diubah atau dimodifikasi agar memenuhi persyaratan teknik analisis. Proses pengkodean merupakan langkah penting dalam KDD, karena memastikan data dapat dianalisis dengan alat atau algoritma yang sesuai [9]. Data kategori, seperti jenis transportasi dan penghasilan orang tua, diubah menjadi

format numerik menggunakan teknik *one-hot encoding* atau *label encoding*. Proses ini memastikan bahwa data dapat diolah secara akurat oleh algoritma *data mining* tanpa kehilangan makna aslinya.

D. Data Mining

Data mining merupakan proses untuk mengidentifikasi pola atau informasi penting dari kumpulan data tertentu dengan memanfaatkan berbagai alat atau metode analisis [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 komponen utama dalam analitik di berbagai bidang, seperti bisnis, pendidikan, kesehatan, dan ilmu sosial. Teknologi ini memberikan wawasan berharga yang membantu pengambilan keputusan berdasarkan data secara lebih akurat dan efisien [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, *machine learning*, dan *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.

Dataset dibagi menjadi *training set* (80%) dan *test set* (20%) menggunakan metode *train_test_split()* secara acak. Langkah ini memastikan bahwa model yang dihasilkan diuji pada data yang tidak pernah dilatih, sehingga performanya lebih dapat diandalkan. Untuk menjaga reliabilitas, dataset diuji konsistensinya dengan memeriksa distribusi data untuk setiap variabel melalui histogram dan Matriks Korelasi. Hal ini dilakukan untuk mendeteksi anomali atau pola yang tidak wajar yang dapat memengaruhi hasil analisis.

E. Logistic regression

Klasifikasi merupakan salah satu metode yang digunakan untuk mengelompokkan data yang telah tersusun secara sistematis. Salah satu teknik klasifikasi yang banyak diterapkan dalam penelitian adalah *logistic regression*. *Logistic regression* adalah metode statistika yang digunakan untuk menganalisis data guna memahami hubungan antara beberapa variabel. Dalam metode ini, variabel respon bersifat kategorik, baik nominal maupun ordinal, sedangkan variabel independen dapat berupa data kategorik maupun kontinu [12]. Jika variabel respon memiliki dua kategori (dikotomis), teknik yang digunakan disebut *binary logistic regression*. Sementara itu, jika variabel respon memiliki lebih dari dua kategori, metode yang digunakan adalah *multinomial logistic regression*.

Dalam bidang statistika, *logistic regression* digunakan untuk menjelaskan hubungan antara variabel respon yang bersifat kategorik dengan variabel independen yang bisa berupa data kontinu atau kategorik [12]. Penelitian ini secara khusus menggunakan *binary logistic regression*, yang bertujuan untuk menganalisis hubungan antara variabel independen (X) dengan variabel dependen (Y) yang memiliki nilai biner. Menurut Harlan [13], *binary logistic regression* berfungsi untuk mengukur pengaruh variabel independen terhadap variabel dependen, baik dalam bentuk kontinu maupun kategorik.

F. Decision tree

Selain *logistic regression*, metode klasifikasi lain yang sering digunakan adalah *decision tree* atau pohon keputusan. Metode *decision tree* menawarkan pendekatan yang cepat dan efektif dalam mengelompokkan data [14]. Selain itu, metode ini telah dikembangkan untuk menangani data yang bersifat sensitif [15]. Hasil dari klasifikasi *decision tree* dapat dimanfaatkan untuk memprediksi siswa yang berisiko putus sekolah.

Decision tree adalah salah satu metode klasifikasi yang digunakan untuk mengorganisasi data atau objek ke dalam kelompok yang menghasilkan keputusan. Proses ini terdiri dari simpul-simpul pilihan yang terhubung oleh cabang, bergerak dari simpul akar hingga ke simpul daun. Pembangunan pohon keputusan dimulai dari simpul akar, yang ditempatkan di bagian atas diagram pohon, di mana atribut-atribut dievaluasi pada simpul seleksi. Setiap hasil yang mungkin menghasilkan cabang baru, yang dapat mengarah ke simpul keputusan lainnya atau langsung ke simpul daun [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

Interpretasi dan evaluasi merupakan tahap akhir dalam proses *data mining* yang bertujuan untuk mengevaluasi hasil yang telah diperoleh [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.

Setelah model *Logistic Regression* dan *Decision Tree* diterapkan, dilakukan evaluasi hasil prediksi dengan membandingkan data aktual dan data prediksi. *Confusion matrix* digunakan untuk mengidentifikasi jumlah *True Positives*, *True Negatives*, *False Positives*, dan *False Negatives*. Metode ini membantu mengevaluasi apakah model telah bekerja sesuai dengan data dan menghasilkan klasifikasi yang akurat.

Model yang telah dilatih pada *training set* diuji pada *test set* untuk mengevaluasi kemampuan generalisasi model. Hasil evaluasi dari data uji memastikan bahwa model tidak mengalami *overfitting* atau *underfitting* terhadap dataset yang digunakan. Hasil dari interpretasi model dibandingkan kembali dengan data asli untuk memastikan bahwa prediksi dan pola yang ditemukan selaras dengan kondisi sebenarnya. Hal ini bertujuan untuk menjaga relevansi hasil analisis dengan realitas di SMA N 4 Tegal.

Pemilihan metode *Logistic Regression* dan *Decision Tree* dalam penelitian ini didasarkan pada karakteristik permasalahan dan jenis data yang digunakan. Kedua metode ini memiliki keunggulan masing-masing yang membuatnya sesuai untuk mendeteksi potensi siswa putus sekolah berdasarkan variabel yang bersifat numerik maupun kategorikal. Kedua metode ini dipilih karena mampu menangani permasalahan klasifikasi biner dengan baik. *Logistic Regression* memberikan model yang linier dan mudah diinterpretasikan, sementara *Decision Tree* menawarkan fleksibilitas dan kemampuan untuk menangkap interaksi antar variabel. Kombinasi kedua metode ini memungkinkan evaluasi komparatif, sehingga dapat menentukan pendekatan mana yang lebih efektif dalam konteks penelitian ini.

Metode *Logistic Regression* dipilih karena merupakan teknik statistika yang sederhana namun sangat efektif dalam memodelkan hubungan antara variabel independen dan variabel dependen biner. Dalam konteks penelitian ini, *Logistic Regression* digunakan untuk memprediksi apakah seorang siswa akan tetap melanjutkan sekolah atau putus sekolah (variabel biner). Keunggulan utama *Logistic Regression* adalah kemampuannya menghasilkan model yang mudah diinterpretasikan, sehingga pihak sekolah dapat memahami pengaruh masing-masing variabel prediktor, seperti kehadiran siswa dan rata-rata nilai semester. Selain itu, metode ini memungkinkan pengujian signifikan terhadap setiap variabel prediktor, yang membantu mengidentifikasi faktor-faktor paling berpengaruh.

Metode *Decision Tree* dipilih karena kemampuannya untuk menangkap pola yang kompleks dalam data dan menghasilkan aturan klasifikasi yang dapat dengan mudah dipahami. Dalam penelitian ini, *Decision Tree* digunakan untuk memetakan hubungan antara variabel seperti penghasilan orang tua, jenis transportasi, dan status kelulusan. Keunggulan utama *Decision Tree* adalah struktur hierarkisnya, yang memberikan visualisasi intuitif tentang bagaimana keputusan dibuat berdasarkan atribut-atribut tertentu. Selain itu, metode ini sangat fleksibel dalam menangani data kategorikal dan numerik, menjadikannya alat yang ideal untuk analisis data siswa yang memiliki berbagai tipe variabel.

Dalam penelitian pendidikan, metode *Logistic Regression* sering digunakan untuk mengidentifikasi faktor-faktor yang memengaruhi kelulusan siswa, sedangkan *Decision Tree* sering diterapkan untuk menemukan pola yang membantu memprediksi risiko siswa putus sekolah. Kedua metode ini telah digunakan dalam berbagai studi sebelumnya yang membahas prediksi dalam bidang pendidikan, sehingga membuktikan relevansi dan keandalannya dalam konteks ini. Dengan mempertimbangkan keunggulan masing-masing metode dan relevansinya terhadap jenis data serta tujuan penelitian, *Logistic Regression* dan *Decision Tree* dipilih sebagai pendekatan yang paling sesuai untuk mendeteksi potensi siswa putus sekolah di SMA N 4 Tegal.

III. HASIL DAN PEMBAHASAN

Penelitian ini menghasilkan temuan melalui penerapan metode *Logistic Regression* dan *Decision Tree* untuk 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 tidak digunakan dalam analisis sejak awal karena prediksi yang dilakukan bersifat objektif. Fokus utama prediksi adalah pada hasil akhir siswa, yaitu apakah siswa tetap menyelesaikan 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   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    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	Ketidakhadiran1semester	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

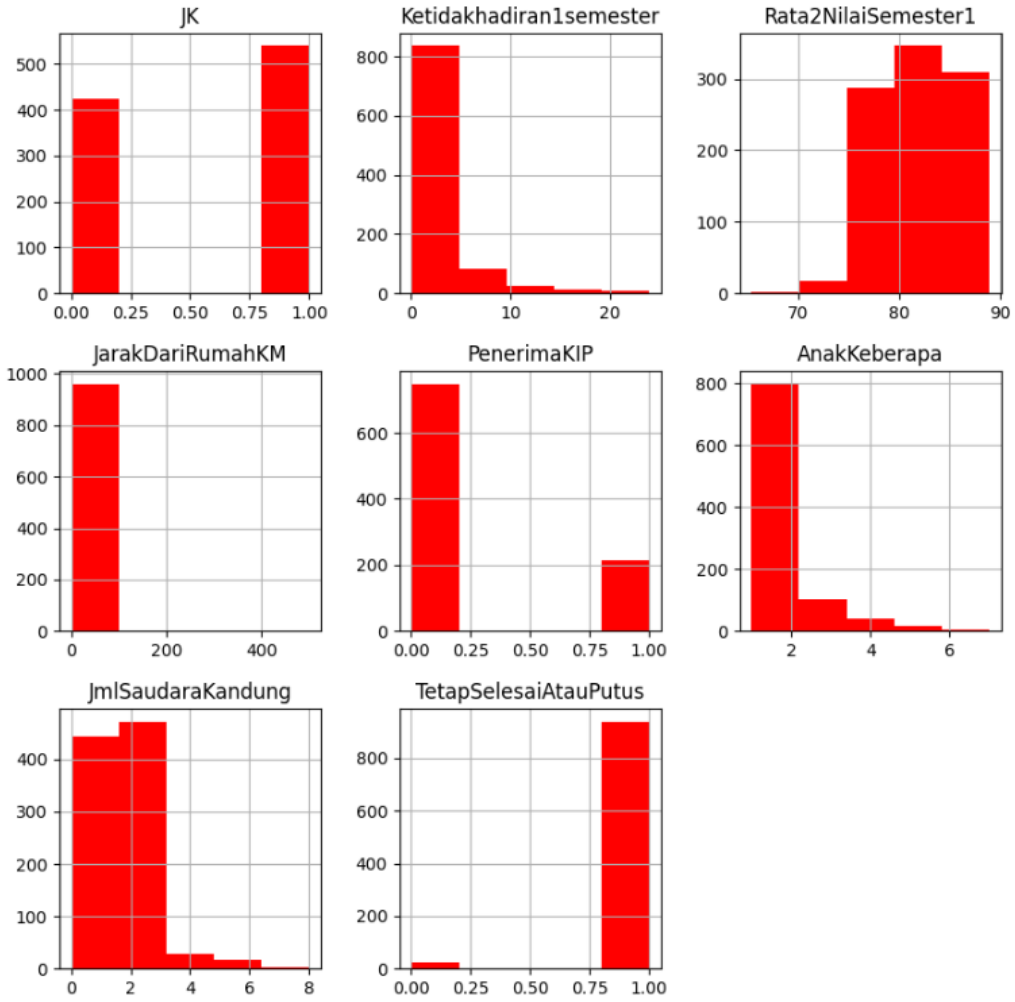
Gambar 3. Tampilan data 15 baris pertama

Variabel yang berupa string dikonversi menjadi data numerik (integer) menggunakan teknik *one-hot encoding* atau *label encoding*. Teknik ini mengubah daftar string menjadi nilai angka berdasarkan urutan abjad. Namun, pada beberapa kolom, nilai string dikonversi secara manual menjadi integer tanpa menggunakan *one-hot encoding* atau *label encoding*, karena nilai yang lebih besar secara logis merepresentasikan angka yang lebih tinggi. Hal ini diterapkan pada kolom Penghasilan Orang Tua, Jenis Tinggal, dan Alat Transportasi.

	JK	Ketidakhadiran1semester	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
```

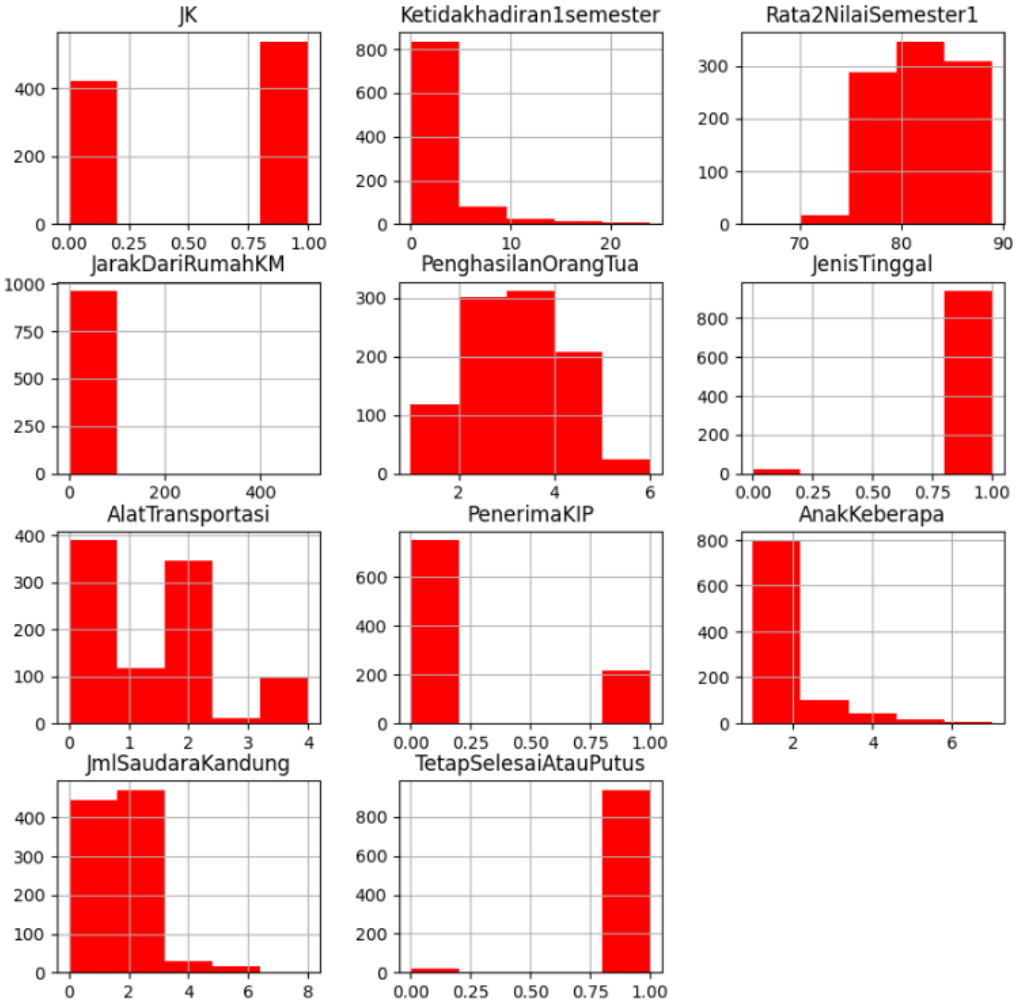
(a)

```
<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)

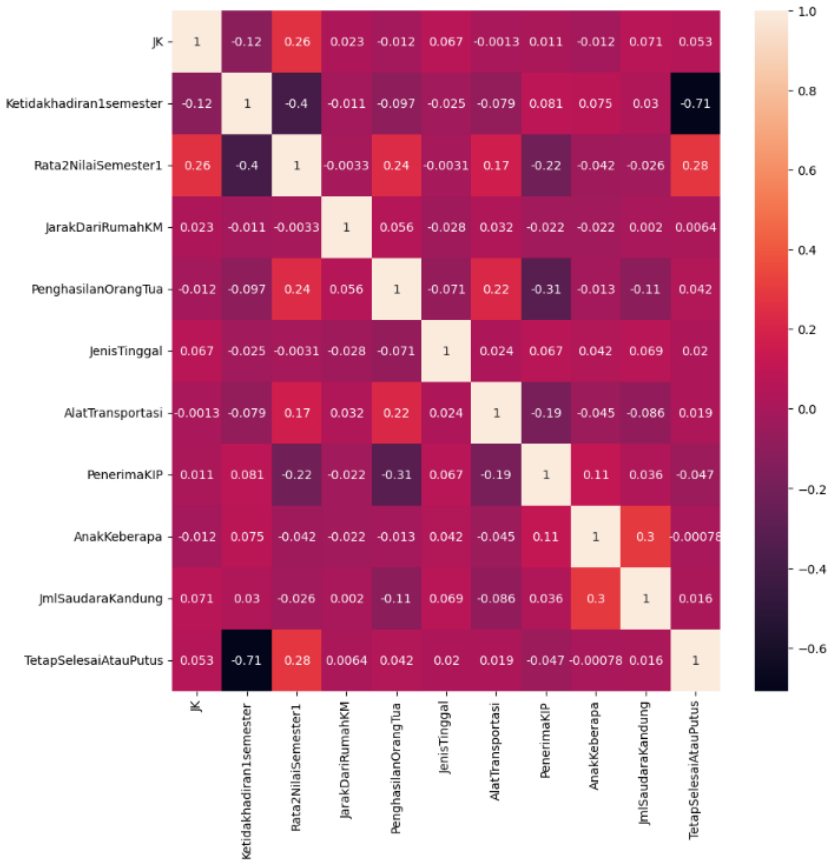
Gambar 6. Tipe kolom (a) sebelum dilakukan perubahan dan (b) setelah dilakukan perubahan

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 menganalisis hubungan linier antar variabel numerik dengan koefisien korelasi Pearson (*r*) sebagai indikator kekuatan hubungan. Nilai r 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

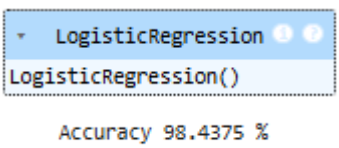
Visualisasi Matriks Korelasi pada Gambar 8 menunjukkan adanya hubungan antara variabel-variabel dalam dataset siswa dengan variabel target, yaitu "Tetap Selesai atau Putus Sekolah." Meskipun tidak semua variabel memiliki korelasi yang kuat, terdapat dua variabel yang menunjukkan hubungan signifikan dengan variabel target, yaitu "Ketidakhadiran dalam 1 Semester" dan "Rata-rata Nilai Semester 1." Nilai korelasi yang mendekati satu atau minus satu menunjukkan kekuatan hubungan tersebut.

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.

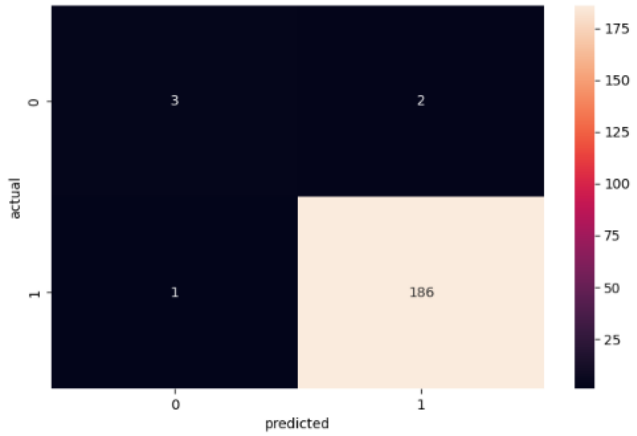
Untuk mengevaluasi kinerja model, dataset dibagi menjadi *training set* dan *test set* menggunakan fungsi *train_test_split()*. Proporsi *test set* ditetapkan sebesar 20% dari total data. Setelah modul Logistic Regression diimpor, dibuat objek klasifikasi Logistic Regression menggunakan fungsi *LogisticRegression()*. Model yang dihasilkan kemudian digunakan untuk melakukan prediksi pada *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.




```
Accuracy 98.4375 %
```

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

Setelah model *Logistic Regression* dibangun, langkah berikutnya adalah mengujinya untuk memprediksi hasil dari data uji. Hasil prediksi tersebut kemudian disebut sebagai *y_pred*. Tahap selanjutnya adalah mengevaluasi kinerja model. Untuk menilai seberapa baik model tersebut, dibuatlah *confusion matrix* yang membandingkan hasil prediksi model dengan data tes yang sebenarnya.



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

Berdasarkan *confusion matrix* pada Gambar 10, dapat dilihat 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 yang dilakukan di Google Colaboratory menggunakan fungsi *accuracy_score()*, yang menghasilkan nilai akurasi sebesar $\frac{189}{192} = 0.984375$ untuk model *Logistic Regression* yang dibangun.

Hasil akurasi ini menunjukkan bahwa model memiliki kinerja 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

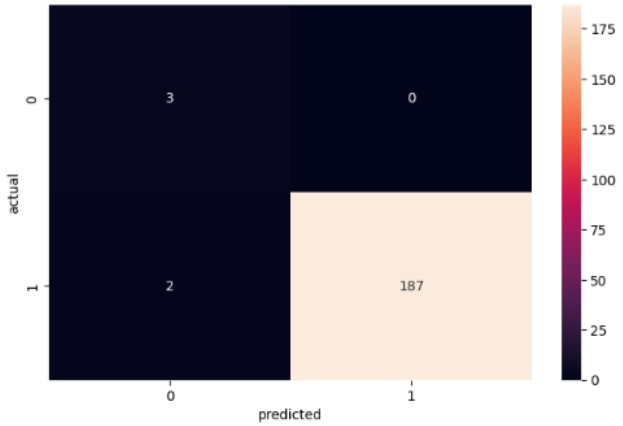
Melanjutkan proses *data mining* yang telah dilakukan dengan menggunakan *Logistic Regression*, tahap klasifikasi menggunakan *Decision Tree* dilakukan pada dataset siswa yang telah melalui seleksi fitur dan dibagi antara *training set* dan *test set*. Pembagian dataset dilakukan dengan menggunakan fungsi *train_test_split()* dengan proporsi *test set* sebesar 20% dari total data. Setelah modul *Decision Tree* diimpor, objek *classifier Decision Tree* dibuat menggunakan fungsi *DecisionTreeClassifier()*. Model yang dihasilkan kemudian digunakan untuk menghasilkan prediksi berdasarkan *test set*.

```
DecisionTreeClassifier
DecisionTreeClassifier(random_state=42)

Accuracy 98.95833333333334 %
```

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

Setelah model *Decision Tree* terbentuk, langkah selanjutnya adalah mengujinya untuk memprediksi hasil dari data uji. Hasil prediksi tersebut disimpan dalam *y_pred_dec_tree*. Tahap berikutnya adalah mengevaluasi kinerja model tersebut. Untuk menilai seberapa baik model yang terbentuk, dibuatlah *confusion matrix* yang membandingkan hasil prediksi yang dihasilkan oleh model dengan hasil sebenarnya dari data uji.



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

Berdasarkan *confusion matrix* pada Gambar 13, dapat dilihat 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 yang dilakukan di Google Colaboratory menggunakan fungsi *accuracy_score()*, yang menghasilkan nilai akurasi sebesar $\frac{190}{192} \cong 0.9895$ untuk model *Decision Tree* yang dibangun.

	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 antar 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.

Meskipun penelitian ini memberikan hasil yang signifikan dalam mendeteksi potensi siswa putus sekolah menggunakan metode *Logistic Regression* dan *Decision Tree*, terdapat beberapa keterbatasan yang perlu diperhatikan karena dapat memengaruhi hasil yang diperoleh. Dataset yang digunakan dalam penelitian ini terbatas pada siswa dari SMA N 4 Tegal angkatan 2022 hingga 2024, dengan total 963 siswa. Data ini hanya mencakup populasi tertentu dan tidak mencerminkan variasi siswa di sekolah lain, terutama yang memiliki latar belakang sosial, ekonomi, atau budaya yang berbeda. Hal ini dapat memengaruhi generalisasi hasil penelitian ke konteks yang lebih luas.

Ketidakseimbangan data kelas juga menjadi salah satu keterbatasan, di mana jumlah siswa yang tetap melanjutkan sekolah jauh lebih banyak dibandingkan siswa yang putus sekolah. Ketidakseimbangan ini memengaruhi performa model, terutama dalam mendeteksi kelas minoritas seperti siswa yang berisiko putus sekolah. Hasil recall yang rendah untuk kelas "Putus Sekolah" pada metode *Decision Tree* menjadi salah satu indikator dari masalah ini.

Keterbatasan tersebut dapat memengaruhi keakuratan, sensitivitas, dan kemampuan generalisasi model. Ketidakseimbangan kelas dapat menyebabkan hasil prediksi untuk kelas minoritas menjadi kurang optimal, sementara keterbatasan data variabel dapat mengurangi kemampuan model untuk menangkap faktor-faktor penting lainnya yang memengaruhi risiko putus sekolah. Selain itu, generalisasi hasil ke populasi siswa yang lebih luas atau ke sekolah dengan karakteristik berbeda perlu dilakukan dengan hati-hati.

Untuk mengatasi keterbatasan ini, penelitian lanjutan dapat dilakukan dengan mengumpulkan dataset yang lebih besar dan mencakup populasi siswa dari berbagai sekolah. Selain itu, teknik penyeimbangan data, seperti *oversampling* atau *undersampling*, dapat diterapkan untuk meningkatkan performa model pada kelas minoritas. Penggunaan variabel tambahan, seperti data psikologis atau lingkungan keluarga, juga dapat membantu menciptakan model yang lebih komprehensif. Pembaruan model secara berkala juga penting untuk memastikan relevansi model terhadap kondisi siswa di masa depan.

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 performa yang lebih unggul 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.

Dalam 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.

Penelitian ini juga menyoroti pentingnya pemanfaatan data berbasis teknologi untuk mendukung pengambilan keputusan yang lebih efektif di bidang pendidikan. Dengan menggunakan pendekatan *data mining*, sekolah dapat lebih proaktif dalam mengidentifikasi siswa yang berisiko putus sekolah dan memberikan intervensi yang tepat waktu.

Selain itu, hasil penelitian ini dapat dijadikan dasar untuk pengembangan model serupa di institusi pendidikan lain, dengan menyesuaikan karakteristik dan kebutuhan data masing-masing.

Untuk penelitian selanjutnya, disarankan untuk menggunakan dataset yang lebih besar dan lebih beragam, mencakup berbagai latar belakang sosial dan budaya. Selain itu, memasukkan variabel tambahan, seperti motivasi siswa, dukungan keluarga, dan lingkungan belajar, dapat membantu menciptakan model prediksi yang lebih komprehensif dan akurat. Pembaruan model secara berkala juga penting agar dapat terus relevan dengan perubahan pola dan dinamika siswa di masa depan.

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.