Available online at: https://journal.ikippgriptk.ac.id/index.php/saintek
Jurnal Pendidikan Informatika dan Sains
13(2), 2024, 141-153
E-ISSN: 2407-1536
P-ISSN: 2089-2802

MODEL DECISION TREE UNTUK PREDIKSI PRESTASI AKADEMIK MATEMATIKA SISWA KELAS VIII SMP FRATER DON BOSCO MANADO

Monica Tiara Gunawan, Jeane Yosefa Tine, Chatarina Enny Murwaningtyas*
Pendidikan Matematika Program Magister, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Sanata Dharma, Yogyakarta, Indonesia
*email: enny@usd.ac.id

Received: 2024-07-14 Accepted: 2024-09-07 Published: 2024-12-25

Abstrak

Penelitian ini bertujuan untuk mengembangkan model *Decision Tree* yang dapat memprediksi prestasi akademik matematika siswa kelas VIII di SMP Frater Don Bosco Manado, serta untuk mengidentifikasi dan menganalisis faktor-faktor penting yang perlu diperhatikan oleh orang tua dalam upaya meningkatkan prestasi akademik anak mereka. Data dikumpulkan melalui dokumentasi nilai akademik siswa, catatan kehadiran, dan kuesioner yang diisi oleh siswa untuk memperoleh informasi tentang dukungan keluarga, banyaknya kegiatan ekstrakurikuler yang diikuti, lama belajar, dan tingkat pendidikan orang tua. Data tersebut dianalisis menggunakan pendekatan *data mining* dengan model *Decision Tree*. Dua model dikembangkan dan dibandingkan: model pertama tanpa seleksi fitur dan model kedua dengan seleksi fitur menggunakan metode *SelectKBest*. Model tanpa seleksi fitur mencapai akurasi 93,33%, sementara model dengan seleksi fitur mencapai akurasi 95,56%. Evaluasi terhadap pentingnya fitur menunjukkan bahwa tanpa seleksi fitur, nilai rapor matematika semester sebelumnya menjadi fitur yang paling dominan, diikuti oleh nilai ulangan harian dan banyaknya kegiatan ekstrakurikuler yang diikuti. Sebaliknya, dalam model dengan *SelectKBest*, durasi belajar menjadi fitur yang paling signifikan, diikuti oleh tingkat pendidikan ayah, dukungan keluarga, dan nilai ulangan harian. Temuan ini menunjukkan bahwa penggunaan seleksi fitur tidak hanya meningkatkan akurasi prediksi tetapi juga membantu mengidentifikasi faktor-faktor kunci yang perlu difokuskan oleh orang tua, seperti durasi belajar, pendidikan orang tua, dukungan keluarga, partisipasi dalam kegiatan ekstrakurikuler, dan nilai akademik sebelumnya, untuk meningkatkan prestasi akademik siswa.
Kata kunci: Prediksi prestasi akademik matematika, *Decision Tree*, Tanpa seleksi fitur dan seleksi fitur, *Feature importances*

Abstract

This study aims to develop a Decision Tree model to predict the academic performance in mathematics of 8th-grade students at SMP Frater Don Bosco Manado and to identify and analyze key factors that parents should focus on to improve their children's academic outcomes. Data were collected through documentation of students' academic records, attendance logs, and questionnaires completed by students to gather information about family support, the number of extracurricular activities they participated in, study time, and parents' education levels. The target variable in this study is the mathematics exam score, while the independent variables include previous report card grades, attendance, family support, the number of extracurricular activities, study time, and parents' education levels. The data were analyzed using a data mining approach with a Decision Tree model. Two models were developed and compared: one without feature selection and another with feature selection using the SelectKBest method. The model without feature selection achieved an accuracy of 93.33%, while the model with feature selection achieved an accuracy of 95.56%. An evaluation of feature importance revealed that, without feature selection, the previous semester's mathematics report card grade was the most dominant feature, followed by daily test scores and the number of extracurricular activities.

Conversely, in the model with SelectKBest, study time emerged as the most significant feature, followed by the father's education level, family support, and daily test scores. These findings suggest that feature selection not only improves prediction accuracy but also helps identify key factors that parents should focus on, such as study time, parents' education, family support, participation in extracurricular activities, and previous academic performance, to enhance students' academic achievement.
Keywords: *Academic achievement mathematics prediction, Decision Tree, Without feature selection and feature selection, Feature importances*

How to cite (in APA style): Gunawan, M. T., Tine, J. Y., & Murwaningtyas, C. E. (2024). Model decision tree untuk prediksi prestasi akademik matematika siswa kelas VIII SMP Frater Don Bosco Manado. *Jurnal Pendidikan Informatika Dan Sains*, *13*(2), 141–153. https://doi.org/10.31571/saintek.v13i2.7696

DOI: 10.31571/saintek.v13i2.7696

PENDAHULUAN

Siswa yang menonjol dengan prestasi akademis biasanya ditandai dengan skor tinggi dalam ujian dan tugas, serta memiliki pemahaman yang sangat baik terhadap materi yang diajarkan, terutama dalam subjek yang menuntut logika dan analisis mendalam seperti matematika. Kemampuan ini tidak hanya mencerminkan kecerdasan analitis, tetapi juga menunjukkan kemampuan siswa untuk mengintegrasikan dan menerapkan konsep dalam berbagai situasi pemecahan masalah (Saragih et al., 2023). Faktor yang memengaruhi prestasi ini sangat beragam, mulai dari penguasaan materi yang efektif, kehadiran yang konsisten, hingga latar belakang sosial dan ekonomi yang mendukung. Studi menunjukkan bahwa kombinasi sumber daya internal dan eksternal ini sangat krusial dalam menunjang kesuksesan akademik siswa (Yulianto & Firmansyah, 2022).

Kegiatan ekstrakurikuler memainkan peran vital dalam pengembangan komprehensif siswa, menawarkan lebih dari sekadar kegiatan sampingan tetapi sebagai bagian integral dari pendidikan karakter. Melalui partisipasi dalam klub debat, olahraga, musik, atau seni, siswa mengembangkan keterampilan sosial, kepemimpinan, dan resiliensi yang membantu mereka tidak hanya di lingkungan sekolah tetapi juga dalam kehidupan pribadi dan profesional di masa depan. Kegiatan ini juga menumbuhkan nilai-nilai seperti kerja sama tim dan integritas, serta menawarkan peluang bagi siswa untuk mengeksplorasi minat dan bakat di luar lingkungan kelas tradisional (Sanjaya, 2020).

Selanjutnya, penelitian oleh Susanto dan Sudiyatno (2014) menyatakan bahwa disiplin dan latar belakang sosial ekonomi keluarga memegang peranan penting dalam membangun dasar bagi keberhasilan akademik siswa. Disiplin yang konsisten, yang diterapkan di rumah dan di sekolah, menyiapkan siswa dengan kebiasaan belajar yang baik yang esensial untuk pencapaian akademis. Kehadiran yang teratur di kelas memungkinkan siswa untuk terus terlibat dalam pembelajaran interaktif, yang sangat penting untuk pemahaman mendalam dan retensi jangka panjang (Retnowati & Khotimah, 2020). Di sisi lain, kondisi ekonomi keluarga sering kali memengaruhi jenis sumber daya yang dapat disediakan untuk mendukung pendidikan anak, termasuk akses ke materi pendidikan berkualitas, teknologi pendidikan, dan kegiatan pengayaan yang semuanya berkontribusi langsung pada prestasi siswa (Tulus, 2004).

Dukungan keluarga memegang peranan yang sangat krusial dalam membantu anak mencapai prestasi belajar yang optimal. Menurut Khafid (2007), keluarga bukan hanya menyediakan dukungan emosional dan motivasi, tetapi juga berperan sebagai lingkungan pembelajaran utama yang menentukan fondasi keberhasilan pendidikan anak. Dalam konteks ini, rumah dianggap sebagai ruang pertama dan paling efektif di mana anak mulai mempelajari berbagai keterampilan dan pengetahuan dasar. Puspita (2023) menambahkan bahwa proses pembelajaran di rumah terjadi secara langsung

dan berkelanjutan, memungkinkan anak untuk menyerap nilai, norma, dan keterampilan penting yang diperlukan untuk keberhasilan di lingkungan akademis dan sosial.

Lebih lanjut, tingkat pendidikan orang tua memiliki dampak signifikan terhadap cara mereka mendidik dan membimbing anak-anak mereka. Orang tua yang memiliki pendidikan lebih tinggi biasanya lebih mampu memberikan bimbingan akademis yang efektif, memanfaatkan metode pembelajaran yang lebih terstruktur, dan menyediakan sumber belajar yang lebih kaya. Tingkat pendidikan ini juga sering kali dikaitkan dengan sikap yang lebih disiplin terhadap pendidikan, yang secara tidak langsung ditularkan kepada anak-anak. Dengan demikian, anak-anak dari orang tua yang berpendidikan tinggi cenderung mengembangkan disiplin belajar yang lebih baik, keterampilan manajemen waktu, dan pendekatan yang lebih serius terhadap tugas-tugas sekolah.

Setiawan (2015) menjelaskan lebih lanjut bahwa perbedaan tingkat pendidikan di antara orang tua bisa sangat memengaruhi strategi dan kualitas bimbingan yang diberikan kepada anak dalam proses belajar di rumah. Misalnya, orang tua yang memiliki latar belakang akademis yang kuat mungkin lebih mampu membantu dengan tugas-tugas sekolah yang kompleks, memberikan penjelasan yang mendalam tentang materi, dan menggunakan sumber daya pendidikan tambahan seperti buku atau teknologi pembelajaran. Hal ini secara langsung memengaruhi seberapa efektif anak dapat memahami materi pelajaran dan, pada akhirnya, berdampak positif pada prestasi belajar mereka di sekolah.

Dengan demikian, keluarga berperan sebagai pilar utama dalam mendukung pendidikan anak, mulai dari membentuk kebiasaan belajar yang baik, menyediakan dukungan emosional, hingga mengimplementasikan strategi pembelajaran yang efektif berdasarkan pengetahuan dan pengalaman orang tua. Dukungan ini adalah komponen penting yang tidak hanya mempengaruhi prestasi akademik, tetapi juga pengembangan sosial dan emosional anak, membekali mereka dengan keterampilan yang diperlukan untuk berhasil dalam kehidupan mereka yang lebih luas (Eccles & Roeser, 2011), selain itu melalui dukungan emosional dan sosial yang diberikan kepada siswa, membantu siswa merasa dihargai dan didukung dalam proses belajar mereka, yang pada gilirannya dapat meningkatkan keterlibatan siswa dalam pembelajaran matematika.

Seiring dengan berkembangnya era digital pada abad ke-21, munculnya *data mining* dilatarbelakangi oleh ledakan informasi yang terjadi belakangan ini. Banyak instansi telah menggumpulkan informasi dalam jumlah besar selama ini. *Data mining* bertujuan untuk mengolah data mentah menjadi informasi yang bermanfaat, sehingga dapat digunakan untuk membuat keputusan yang tepat dan mendorong terciptanya inovasi baru.

Data mining ini dapat menyelesaikan masalah kompleks terkait dengan jumlah data besar di berbagai bidang seperti pendidikan. Salah satu metode klasifikasi yang terkenal dan digunakan sebagai prediksi adalah *Decision Tree* (Hanif & Setiaji, 2022). Cara kerja metode klasifikasi pada *Decision Tree* berusaha memprediksi kelas target dengan presisi tertinggi serta menemukan hubungan antara atribut *input* dan atribut *output* untuk membangun sebuah model yang merupakan proses pelatihan atau *training* (Jijo & Abdulazeez, 2021).

Model *Decision Tree* memiliki keunggulan dalam memodelkan hubungan antara variabel dan menghasilkan keputusan yang mudah dipahami (Fauzan et al., 2024). Struktur dari *Decision Tree* memiliki kemiripan dengan proses pengambilan keputusan manusia, sehingga konsepnya mudah untuk dipahami (Patel & Prajapati, 2018). Selain itu, model *Decision Tree* dapat digunakan untuk menyelesaikan permasalahan, baik ketika tipe data kategorikal maupun numerikal.

Penggunaan model *Decision Tree* untuk memprediksi prestasi siswa di SD N 3 Bayalangu Kidul menunjukkan tingkat keberhasilan prediksi kelulusan mencapai 98,04%, sementara prediksi ketidaklulusan mencapai 83,33% (Amani & Hayati, 2024). Hal ini menunjukkan bahwa model ini efektif dalam memprediksi prestasi siswa dengan tingkat akurasi yang tinggi. Keberhasilan ini menegaskan bahwa *Decision Tree* mampu mengidentifikasi pola-pola penting dalam data akademik siswa yang dapat digunakan oleh sekolah untuk meningkatkan strategi pembelajaran dan intervensi

pendidikan. Rajagukguk (2021) juga menunjukkan bahwa pembelajaran mesin termasuk model *Decision Tree* relevan untuk digunakan dalam memprediksi prestasi belajar siswa.

Pada penelitian yang dilakukan oleh Qisthiano et al. (2023), hasil penelitian menunjukkan bahwa dengan menggunakan *Decision Tree* yang diterapkan pada 1739 data, dengan pembagian 90% data *training* dan 10% data *testing*, diperoleh akurasi prediksi tertinggi sebesar 87,93%. Temuan ini memperkuat bukti bahwa *Decision Tree* tidak hanya akurat tetapi juga konsisten dalam memprediksi hasil akademik di berbagai konteks pendidikan. Pembagian data yang digunakan dalam penelitian ini juga menekankan pentingnya memiliki *set* data *training* yang besar untuk melatih model secara efektif.

Selanjutnya, penelitian yang dilakukan oleh Suriani (2023) menunjukkan hasil bahwa pengujian model prediksi tingkat kelulusan mahasiswa menggunakan metode klasifikasi *Decision Tree* mencapai 99,64%. Ini menandakan model yang digunakan memiliki kemampuan klasifikasi yang hampir sempurna. Dengan tingkat akurasi yang sangat tinggi, *Decision Tree* dapat diandalkan untuk memprediksi kelulusan mahasiswa dengan sangat akurat, yang sangat penting bagi institusi pendidikan tinggi untuk merancang program yang mendukung keberhasilan mahasiswa mereka. Penelitian ini juga menyoroti potensi *Decision Tree* sebagai alat yang dapat memberikan nilai tambah signifikan dalam pengambilan keputusan berbasis data di sektor pendidikan.

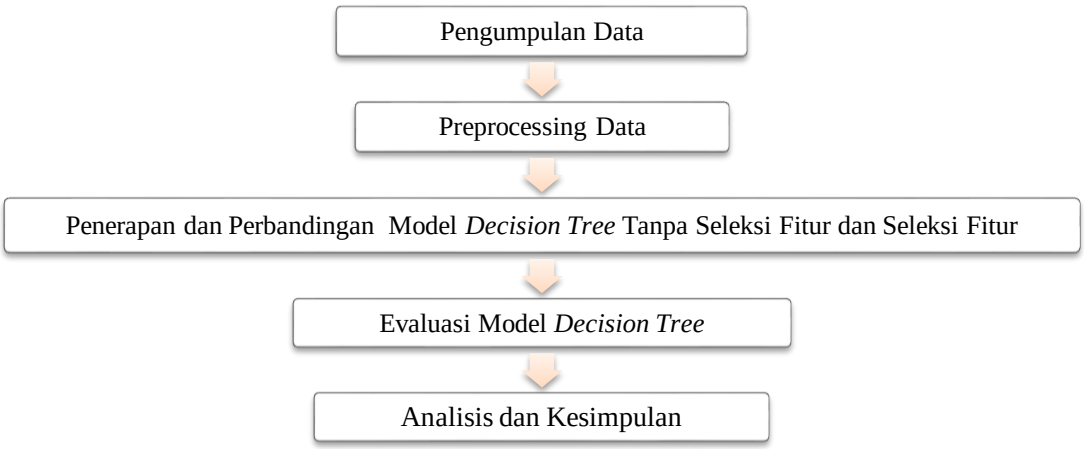
Penelitian ini bertujuan untuk menentukan model prediksi terbaik serta mengidentifikasi faktor yang memiliki pengaruh kuat dan faktor dominan terhadap prestasi akademik matematika siswa kelas VIII SMP Frater Don Bosco Manado. Penelitian ini menggunakan dua model *Decision Tree*, yakni tanpa seleksi fitur dan dengan seleksi fitur. Dalam penelitian ini, variabel target untuk model *Decision Tree* adalah ketuntasan siswa, yang dikategorikan berdasarkan nilai ujian tengah semester. Nilai 70 ke atas dinyatakan tuntas, sementara nilai di bawah 70 dinyatakan tidak tuntas. Analisis ini membantu dalam memahami bagaimana model klasifikasi dapat diterapkan untuk mendukung pencapaian ketuntasan akademik.

Hasil dari penelitian ini berkontribusi dalam penerapan model klasifikasi, khususnya *Decision Tree*, untuk memprediksi prestasi akademik matematika siswa. Perbandingan model dengan dan tanpa seleksi fitur membantu mengidentifikasi faktor-faktor yang berpengaruh kuat dan dominan untuk memprediksi prestasi akademik siswa kelas VIII SMP Frater Don Bosco Manado serta memberikan wawasan dalam pengembangan model klasifikasi di bidang pendidikan.

METODE

Penelitian ini menggunakan metode deskriptif kuantitatif dengan model *Decision Tree*, yang dipilih karena kemampuannya dalam mengidentifikasi faktor-faktor dominan yang mempengaruhi prestasi akademik serta menentukan model prediksi yang optimal. Metode ini dianggap relevan karena mampu menangani variabel-variabel kompleks dan interaksi antarvariabel, seperti yang telah dibuktikan dalam penelitian oleh Matzavela dan Alepis (2021), di mana *Decision Tree* digunakan untuk memprediksi kinerja akademik dalam konteks pembelajaran *mobile* (*M-Learning*).

Dalam penelitian ini, populasi terdiri dari 150 siswa kelas VIII di SMP Frater Don Bosco Manado. Data dikumpulkan melalui dokumentasi nilai akademik siswa, catatan kehadiran, serta kuesioner yang diisi oleh siswa untuk mendapatkan informasi mengenai dukungan keluarga, jumlah kegiatan ekstrakurikuler yang diikuti, lama belajar, dan tingkat pendidikan orang tua. Instrumen yang digunakan meliputi rekapitulasi nilai untuk variabel kuantitatif dan kuesioner untuk mengukur faktor-faktor kualitatif. Analisis data dilakukan dalam lima tahap: (1) pengumpulan data, (2) *preprocessing* data untuk model *Decision Tree*, (3) pengujian model tanpa seleksi fitur dan pengujian model dengan seleksi fitur menggunakan metode *SelectKBest*, (4) evaluasi hasil model untuk menentukan faktor-faktor yang paling signifikan dalam memprediksi prestasi akademik, dan (5) analisis dan kesimpulan. Tahapan penelitian ini disajikan secara rinci pada Gambar 1, yang menggambarkan langkah-langkah analisis dari awal pengumpulan data hingga penyusunan kesimpulan.



Gambar 1. Tahapan Penelitian

1. Pengumpulan Data

Pada tahap ini, peneliti menyajikan data yang telah dikumpulkan melalui rekapitulasi nilai tes, tugas, dan hasil kuesioner siswa kelas VIII SMP Frater Don Bosco Manado. Kuesioner tersebut mencakup informasi mengenai nama siswa, status ekonomi keluarga, dukungan keluarga, partisipasi dalam kegiatan ekstrakurikuler, lama belajar per hari, dan tingkat pendidikan orang tua. Jumlah data yang digunakan dalam penelitian ini adalah 150 sampel. Proses pengumpulan data dimulai dengan peneliti memberikan kuesioner yang wajib diisi oleh seluruh siswa kelas VIII. Selain itu, peneliti juga mengumpulkan rekapitulasi nilai yang dibutuhkan, seperti nilai rapor kelas VIII Semester 1, rata-rata nilai ulangan harian, dan nilai UTS kelas VIII. Setelah semua data terkumpul, peneliti menentukan variabel target dan variabel bebas yang akan digunakan dalam analisis. Variabel target dan variabel bebas dalam penelitian ini disajikan pada Tabel 1.

Tabel 1. Variabel Target dan Variabel Bebas

No	Atribut	Keterangan
1	Nilai UTS Matematika Siswa Kelas VIII (Tuntas, Tidak Tuntas)	Variabel Tak Bebas (Variabel Target)
2	Nilai Rapor Matematika Kelas VIII Semester 1	Variabel Bebas
3	Nilai Ulangan Harian 1 Materi Teorema Phytagoras Kelas VIII Semester 2	Variabel Bebas
4	Nilai Ulangan Harian 2 Materi Lingkaran Kelas VIII Semester 2	Variabel Bebas
5	Nilai Ulangan Harian 3 Materi Statistika Kelas VIII Semester 2	Variabel Bebas
6	Jumlah Kehadiran	Variabel Bebas
7	Dukungan Keluarga	Variabel Bebas
8	Kegiatan Ekstrakurikuler	Variabel Bebas
9	Lama Belajar per Hari	Variabel Bebas
10	Tingkat Pendidikan Orang Tua	Variabel Bebas

2. *Preprocessing* Data

Pada tahap ini, peneliti melakukan pembersihan data (*data cleaning*) dimana pada langkah tersebut peneliti membuang data yang tidak konsisten, nilai yang hilang, dan data berisik atau *noise* yang tidak diperlukan dalam proses *pre-processing*. Setelah data dibersihkan, selanjutnya data tersebut dikelompokan menjadi dua bagian yakni *data training* dan *data testing*. Selanjutnya, peneliti melakukan tahap *encoding* pada data kategorikal untuk memudahkan proses pemrosesan data oleh model klasifikasi.

3. Penerapan dan Perbandingan Model *Decision Tree* Tanpa Seleksi Fitur dan Seleksi Fitur

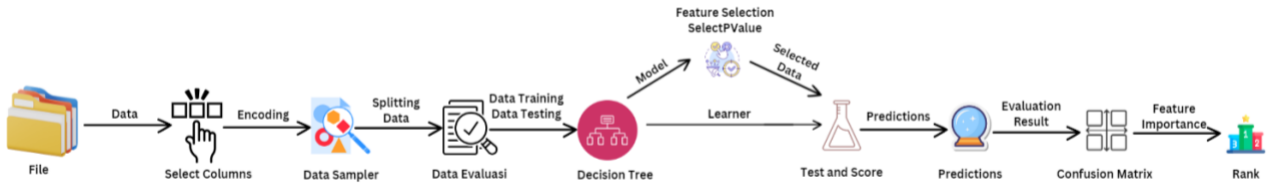
   Pada tahap ini, peneliti melakukan uji coba dengan menerapkan model *Decision Tree*. Selain itu, peneliti juga membandingkan penggunaan model *Decision Tree* antara tanpa seleksi fitur dan dengan seleksi fitur. Tujuan dari perbandingan ini adalah untuk menentukan model prediksi terbaik serta melihat fitur-fitur mana yang memiliki pengaruh kuat dan dominan untuk memprediksi prestasi akademik siswa kelas VIII. Apabila dalam penerapan kinerja model tidak sesuai yang diharapkan maka model ini perlu dievaluasi lebih lanjut agar mendapatkan kinerja yang lebih baik lagi.
4. Evaluasi Model *Decision Tree*

   Pada tahap ini, peneliti melakukan evaluasi terkait dengan pemrosesan data pada model *Decision Tree*. Tujuannya adalah untuk mendapatkan hasil kinerja yang lebih baik lagi, dengan membandingkan antara model tanpa seleksi fitur dan model dengan seleksi fitur.
5. Analisis dan Kesimpulan

   Pada tahap ini, peneliti melakukan analisis untuk menyimpulkan hasil perbandingan kedua model yakni tanpa dan dengan seleksi fitur untuk menentukan model prediksi yang terbaik serta fitur-fitur yang berpengaruh dan dominan dengan mempertimbangkan *feature importances* pada kedua model prediksi tersebut.

HASIL DAN PEMBAHASAN

Secara umum berikut disajikan gambaran tahapan sesuai dengan Gambar 2 dalam penelitian ini untuk membantu visualisasi pemrosesan data dari awal sampai akhir. Berikut disajikan visualisasi pemrosesan data.



Gambar 2. Visualisasi Tahapan Klasifikasi

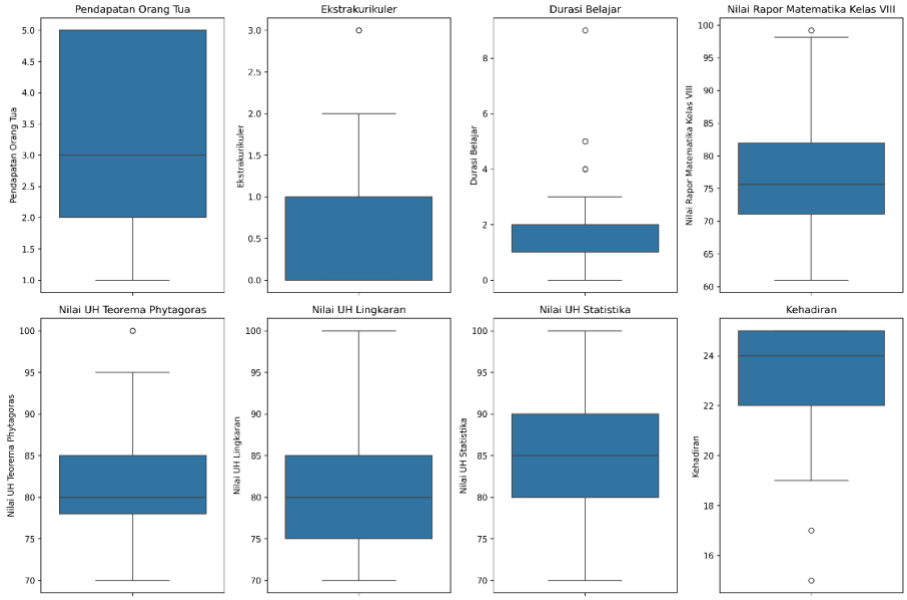
Penelitian ini bertujuan untuk menjawab rumusan masalah, yaitu menentukan model prediksi terbaik antara *Decision Tree* tanpa seleksi fitur dan dengan seleksi fitur, serta mengidentifikasi faktor-faktor dominan yang memengaruhi prestasi akademik matematika siswa kelas VIII SMP Frater Don Bosco Manado. Proses pembangunan model dimulai dengan meng*input* data dari file Excel. Pada tahap *pre-processing*, peneliti melakukan pembersihan data, terutama untuk menangani nilai yang hilang (*missing value*) sesuai metode yang dijelaskan oleh Huang et al. (2015). Selanjutnya, *dataset* dipisahkan menjadi data numerik dan kategorik untuk memudahkan pemrosesan menggunakan bahasa pemrograman Python pada aplikasi Jupyter Notebook. Langkah ini penting agar model dapat membentuk pola yang akurat pada data *input* maupun *output*, serta memastikan kebutuhan data yang akan dikumpulkan (Njeri, 2022).

Tahap berikutnya adalah perhitungan kardinalitas untuk setiap atribut kualitatif. Dalam penelitian ini, empat atribut kualitatif dianalisis, yaitu "Dukungan Keluarga," "Tingkat Pendidikan Ibu," "Tingkat Pendidikan Ayah," dan "Nilai UTS" yang merupakan variabel target dan dinyatakan dalam data kategori. Oleh karena itu, proses *encoding* data dilakukan dengan mempertimbangkan jumlah kardinalitas dan jenis data, karena hal ini dapat memengaruhi cara data diolah dan bagaimana model *machine learning* memproses informasi tersebut secara efektif (Maharana et al., 2022). Peneliti menggunakan label *encoding* untuk data dengan dua kategori, yaitu "Tuntas" dan "Tidak Tuntas," serta ordinal *encoding* untuk data dengan urutan kategori yang jelas, seperti sangat rendah, rendah, cukup, tinggi, dan sangat tinggi (Potdar et al., 2017). Gambar 3 menyajikan hasil transformasi data yang dilakukan untuk mengubah tipe atribut pada data.

	Pendapatan Orang Tua	Dukungan Keluarga	Ekstrakurikuler	Durasi Belajar	Tingkat Pendidikan Ibu	Tingkat Pendidikan Ayah	Nilai Rapor Matematika Kelas VIII	Nilai UH Teorema Phytagoras	Nilai UH Lingkaran	Nilai UH Statistika	Kehadiran	Nilai UTS
0	1	3	0	2.0	4	4	99.2	100	100	100	22	1
1	1	3	0	1.0	3	3	84.2	85	80	90	24	1
2	5	2	0	2.0	3	4	76.4	80	85	80	25	0
3	4	2	0	2.0	3	3	84.0	90	90	90	23	0
4	2	1	0	1.0	4	3	98.2	100	95	100	24	1

Gambar 3. Transformasi Data

Proses pemisahan (*splitting data*) dilakukan menjadi dua bagian, yaitu data *training* dan data *testing*. Data *training* digunakan untuk melatih model, sedangkan data *testing* digunakan untuk menguji kinerja model yang telah terbentuk (Birba, 2020). Setelah melakukan *splitting* data, tahap berikutnya adalah penerapan model *Decision Tree*.



Gambar 4. Boxplot Data Numerik

Berdasarkan visualisasi boxplot yang ditunjukkan pada Gambar 4, terlihat adanya *outliers* pada beberapa variabel seperti Ekstrakurikuler, Durasi Belajar, Nilai Rapor Matematika Kelas VIII, Nilai UH Teorema Phytagoras dan Kehadiran. Pengecekan *outliers* (nilai ekstrim) sangat penting dilakukan karena *outliers* dalam suatu *dataset* dapat menyebabkan distribusi data menjadi tidak berdistribusi normal, sehingga ketika dilakukan analisis statistik uji normalitas data tentu menjadi kurang akurat (Sihombing et al., 2023). Namun, penanganan *outliers* ini juga perlu mempertimbangkan beberapa faktor seperti tujuan analisis, karakteristik data, dan model yang digunakan.

Selain itu, hasil penerapan model ini menunjukkan adanya ketidakseimbangan kelas (*class imbalance*) pada model yang dijalankan. Penanganan *outliers* dan ketidakseimbangan kelas (*class imbalance*) dalam *dataset* dapat diatasi menggunakan teknik *pipeline*. Teknik ini digunakan untuk menyusun serangkaian langkah pemrosesan data dan pelatihan model secara berurutan. Fungsi *pipeline* ini digunakan untuk memastikan bahwa setiap langkah dalam proses pemodelan dilakukan dengan cara yang konsisten dan dapat diulang, sehingga meminimalkan risiko kesalahan dan meningkatkan efisiensi.

Tahap *pipeline* dalam penelitian ini meliputi beberapa langkah penting. Pertama, *Variance Threshold* diterapkan dengan tujuan untuk menghapus fitur-fitur yang konstan atau hampir konstan

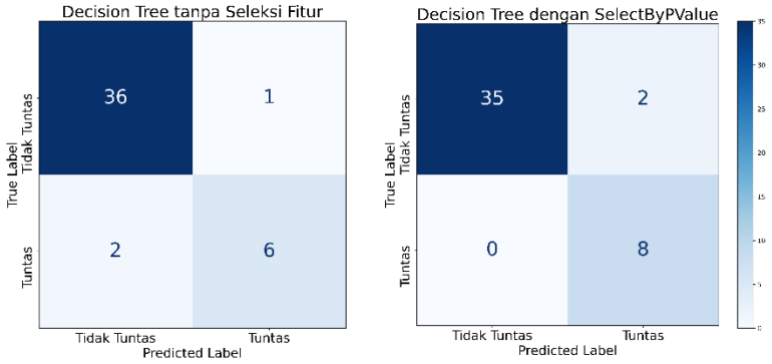
(fitur dengan variabilitas rendah) dari data. Kedua, *Robust Scaler* digunakan untuk menyeimbangkan skala fitur-fitur dengan mengurangi pengaruh *outliers*. Selanjutnya, untuk menangani ketidakseimbangan kelas (*class imbalance*) digunakan teknik SMOTE (*Synthetic Minority Over-sampling Technique*).

SMOTE bekerja dengan menduplikasi data dari kelas minoritas secara acak untuk meningkatkan jumlah sampel dari kelas tersebut dalam *dataset* yang tidak seimbang. Penggunaan SMOTE perlu dilakukan dengan hati-hati karena dapat berkontribusi pada *overfitting*, yaitu ketika model pembelajaran mesin terlalu menyesuaikan diri dengan data pelatihan, bahkan menangkap variasi kecil atau tidak relevan dalam data tersebut. Namun, SMOTE dapat meningkatkan akurasi prediksi baik untuk kelas minoritas maupun kelas mayoritas dengan menginterpolasi data dari kelas minoritas yang ada.

Selanjutnya, dalam *pipeline* disajikan model *Decision Tree* baik tanpa seleksi fitur maupun dengan seleksi fitur. Tujuannya adalah untuk menentukan model prediksi yang terbaik serta mengidentifikasi fitur-fitur yang mempengaruhi prediksi prestasi akademik siswa kelas VIII di SMP Frater Don Bosco Manado. Adanya perbandingan tanpa seleksi fitur dan dengan seleksi fitur tentu akan berpengaruh pada performa prediksi dan penentuan fitur-fitur penting yang berpengaruh kuat dan dominan terhadap prediksi prestasi akademik siswa. Melalui perbandingan tersebut, dapat dipahami apakah model tanpa seleksi fitur atau dengan seleksi fitur dapat meningkatkan akurasi model dan mengidentifikasi fitur-fitur utama yang mempengaruhi prestasi akademik siswa kelas VIII. Hal ini membantu memberikan pemahaman yang lebih jelas tentang faktor-faktor penting yang perlu dipertimbangkan dalam upaya meningkatkan prestasi akademik siswa.

Pada tahap prediksi, model yang telah dilatih sebelumnya diterapkan pada data baru untuk melakukan prediksi, guna memahami dan mengambil keputusan berdasarkan data-data baru yang masuk. Selanjutnya, dilakukan evaluasi kinerja model klasifikasi menggunakan matriks kebingungan (*confusion matrix*). Matriks kebingungan digunakan untuk menghitung berbagai metrik evaluasi kinerja model klasifikasi, seperti akurasi (*accuracy*), presisi (*precision*), *recall*, dan *F1-Score* (Hartanto, 2023). Metrik-metrik ini memberikan wawasan mengenai efektivitas model dalam memprediksi setiap kelas. Ini juga membantu menemukan area yang perlu ditingkatkan agar model prediksi dapat bekerja lebih optimal. Matriks kebingungan bertujuan untuk menunjukkan seberapa sering model membuat prediksi yang benar atau salah untuk setiap kelas serta bagaimana kesalahan prediksi terdistribusi di antara kelas-kelas tersebut. Matriks kebingungan adalah tabel yang menggambarkan hasil prediksi model dibandingkan dengan nilai sebenarnya (Vujović, 2021). Tabel ini mencakup empat komponen utama untuk setiap kelas dalam kasus biner (dua kelas), yaitu *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), dan *False Negatives* (FN). *True Positives* (TP) adalah jumlah sampel yang benar-benar berada dalam kelas positif dan diprediksi dengan tepat sebagai positif. *True Negatives* (TN) adalah jumlah sampel yang benar-benar berada dalam kelas negatif dan diprediksi dengan tepat sebagai negatif. *False Positives* (FP) adalah jumlah sampel yang sebenarnya berada dalam kelas negatif, namun salah diprediksi sebagai positif. *False Negatives* (FN) adalah jumlah sampel yang sebenarnya berada dalam kelas positif, namun salah diprediksi sebagai negatif.

Penelitian ini menggunakan model *Decision Tree* untuk memprediksi prestasi akademik siswa kelas VIII SMP Frater Don Bosco Manado, dengan fokus pada variabel apakah siswa dinyatakan tuntas dalam Ujian Tengah Semester atau tidak. Dua model *Decision Tree* dibandingkan, satu tanpa seleksi fitur dan satu dengan seleksi fitur menggunakan metode *SelectByPValue*. Metode *SelectByPValue* yang digunakan adalah *SelectKBest* dengan fungsi statistik *f_classif* dan nilai batas *p-value* 0,05. *SelectKBest* adalah teknik seleksi fitur yang memilih fitur terbaik berdasarkan skor tertinggi dari uji statistik tertentu, dalam hal ini uji ANOVA (*f_classif*). Uji ANOVA mengukur hubungan antara fitur individu dengan variabel target, dan hanya fitur dengan *p-value* kurang dari atau sama dengan 0,05 yang dipilih, menunjukkan bahwa fitur-fitur tersebut memiliki hubungan signifikan dengan variabel target.



Gambar 5. Mtriks Konfusi (*Confusion Matrix*)

Hasil penelitian ditampilkan pada Gambar 5. Model tanpa seleksi fitur memiliki akurasi 93,33%, dengan 36 siswa yang tidak tuntas diprediksi benar (*True Negative*), 6 siswa yang tuntas diprediksi benar (*True Positive*), 1 siswa yang tidak tuntas diprediksi sebagai tuntas (*False Positive*), dan 2 siswa yang tuntas diprediksi sebagai tidak tuntas (*False Negative*). Sementara itu, model dengan seleksi fitur memiliki akurasi lebih tinggi yaitu 95,56%, dengan 35 siswa yang tidak tuntas diprediksi benar (*True Negative*), 8 siswa yang tuntas diprediksi benar (*True Positive*), 2 siswa yang tidak tuntas diprediksi sebagai tuntas (*False Positive*), dan tidak ada siswa yang tuntas diprediksi sebagai tidak tuntas (*False Negative*). Dengan demikian, model dengan seleksi fitur lebih efektif dalam mengurangi kesalahan prediksi untuk siswa yang tuntas, namun sedikit meningkatkan kesalahan prediksi untuk siswa yang tidak tuntas.

Pemodelan ini menggunakan teknik *tuning hiperparameter* untuk mengoptimalkan kinerja model. *Hiperparameter* adalah parameter yang ditetapkan sebelum proses pelatihan model dimulai dan tidak dapat dipelajari dari data (Wu et al., 2019). Dalam penelitian ini, model *Decision Tree* tanpa seleksi fitur menggunakan parameter terbaik yang diperoleh melalui proses pencarian grid (*GridSearch*). Parameter terbaik yang digunakan adalah *criterion entropy*, *max depth* 3, *min samples leaf* 2, dan *min samples split* 5.

	precision	recall	f1-score	support
Tidak Tuntas	0.95	0.97	0.96	37
Tuntas	0.86	0.75	0.80	8
accuracy			0.93	45
macro avg	0.90	0.86	0.88	45
weighted avg	0.93	0.93	0.93	45

Gambar 6. Laporan Kinerja Model *Decision Tree* Tanpa Seleksi Fitur

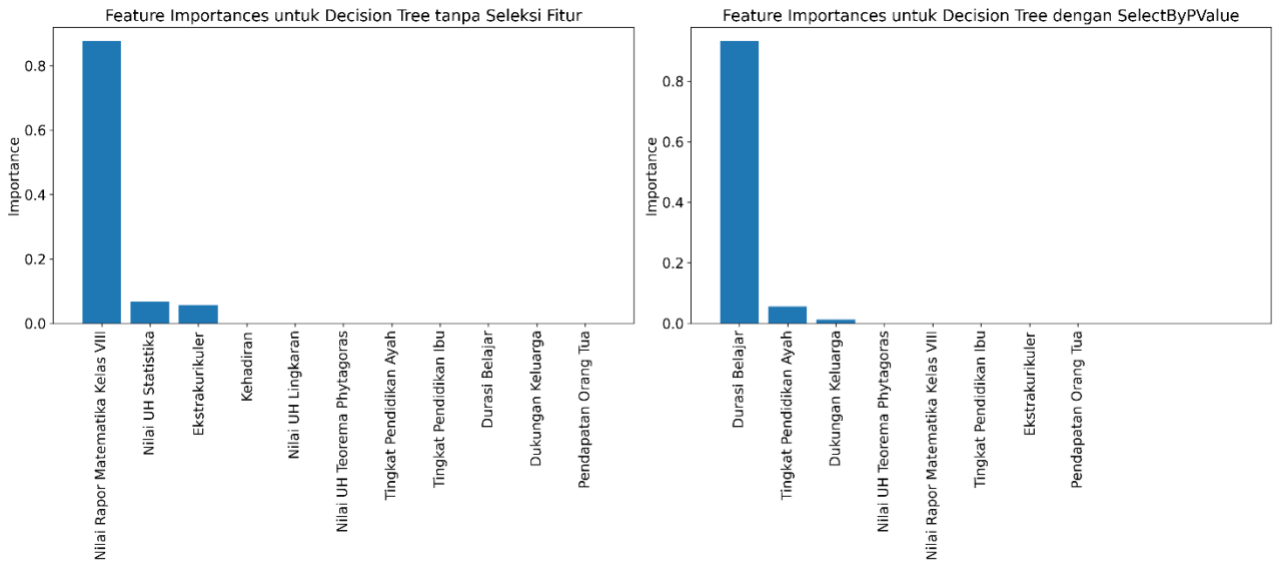
Model *Decision Tree* tanpa seleksi fitur menghasilkan skor akurasi sebesar 0,9333. Skor *precision* model ini adalah 0,9313, skor *recall* adalah 0,9333, dan skor *f1-score* adalah 0,9316. Berdasarkan laporan klasifikasi, Gambar 6, model ini menunjukkan kinerja yang baik dalam memprediksi kategori "Tidak Tuntas" dengan *precision* 0,95, *recall* 0,97, dan *f1-score* 0,96 dari 37 sampel. Untuk kategori "Tuntas", *precision* yang diperoleh adalah 0,86, *recall* 0,75, dan *f1-score* 0,80 dari 8 sampel. Secara keseluruhan, model ini memiliki akurasi 93%, dengan rata-rata makro (*macro avg*) *precision* 0,90, *recall* 0,86, dan *f1-score* 0,88, serta rata-rata tertimbang (*weighted avg*) *precision* 0,93, *recall* 0,93, dan *f1-score* 0,93. Hasil ini menunjukkan bahwa model *Decision Tree* tanpa seleksi fitur memiliki kinerja yang konsisten dan akurat dalam memprediksi apakah siswa tuntas atau tidak dalam ujian sisipan.

	precision	recall	f1-score	support
Tidak Tuntas	1.00	0.95	0.97	37
Tuntas	0.80	1.00	0.89	8
accuracy			0.96	45
macro avg	0.90	0.97	0.93	45
weighted avg	0.96	0.96	0.96	45

Gambar 7. Laporan Kinerja Model *Decision Tree* Menggunakan Seleksi Fitur

Model *Decision Tree* dengan *SelectByPValue* menggunakan parameter terbaik yang diperoleh melalui proses pencarian grid. Parameter terbaik yang digunakan adalah *criterion gini*, *max depth* 3, *min samples leaf* 5, dan *min samples split* 5. Model ini menghasilkan skor akurasi sebesar 0,9556. Skor *precision* model ini adalah 0,9644, skor *recall* adalah 0,9556, dan skor *f1-Score* adalah 0,9574. Berdasarkan laporan klasifikasi, Gambar 7, model ini menunjukkan kinerja yang sangat baik dalam memprediksi kategori "Tidak Tuntas" dengan *precision* 1,00, *recall* 0,95, dan *f1-score* 0,97 dari 37 sampel. Untuk kategori "Tuntas", *precision* yang diperoleh adalah 0,80, *recall* 1,00, dan *f1-score* 0,89 dari 8 sampel. Secara keseluruhan, model ini memiliki akurasi 96%, dengan rata-rata makro (*macro avg*) *precision* 0,90, *recall* 0,97, dan *f1-score* 0,93, serta rata-rata tertimbang (*weighted avg*) *precision* 0,96, *recall* 0,96, dan *f1-score* 0,96.

Hasil ini menunjukkan bahwa model *Decision Tree* dengan seleksi fitur *SelectByPValue* memberikan kinerja yang lebih baik dibandingkan dengan model tanpa seleksi fitur, terutama dalam hal akurasi dan pengurangan kesalahan prediksi untuk siswa yang tuntas dalam ujian sisipan. Model dengan seleksi fitur lebih efektif dalam mengidentifikasi siswa yang tuntas, sementara tetap mempertahankan kinerja yang baik dalam mengidentifikasi siswa yang tidak tuntas.



Gambar 8. *Feature Importances* Untuk Decision Tree Dengan Dan Tanpa Seleksi Fitur

Gambar 8 menampilkan pentingnya fitur (*feature importances*) untuk model *Decision Tree* tanpa seleksi fitur dan dengan *SelectByPValue*. Pada model tanpa seleksi fitur, fitur yang paling penting adalah "Nilai Rapor Matematika Kelas VIII", diikuti oleh "Nilai UH Statistika" dan "Ekstrakurikuler". Fitur "Nilai Rapor Matematika Kelas VIII" memiliki pentingnya fitur yang sangat dominan, menunjukkan bahwa kemampuan matematika siswa di kelas VIII adalah indikator utama dalam menentukan apakah siswa akan tuntas dalam ujian sisipan. Selain itu, "Nilai UH Statistika"

dan partisipasi dalam kegiatan "Ekstrakurikuler" juga memiliki pengaruh, meskipun tidak sebesar "Nilai Rapor Matematika".

Untuk model dengan *SelectByPValue*, fitur yang paling penting adalah "Durasi Belajar", diikuti oleh "Tingkat Pendidikan Ayah" dan "Dukungan Keluarga". Fitur "Durasi Belajar" memiliki pengaruh yang sangat besar dalam model ini, menunjukkan bahwa waktu yang dihabiskan siswa untuk belajar adalah faktor kritis dalam keberhasilan akademis mereka. "Tingkat Pendidikan Ayah" sebagai fitur penting menunjukkan bahwa latar belakang pendidikan orang tua juga berperan dalam prestasi siswa. "Dukungan Keluarga" yang tinggi memberikan lingkungan yang mendukung bagi siswa untuk belajar lebih efektif dan mencapai hasil yang lebih baik.

Arti dari fitur-fitur ini menunjukkan bahwa untuk meningkatkan prestasi akademik siswa, perhatian khusus perlu diberikan pada faktor-faktor seperti "Durasi Belajar", "Tingkat Pendidikan Ayah", dan "Dukungan Keluarga". Intervensi yang meningkatkan durasi belajar siswa dapat berupa program tambahan belajar, bimbingan belajar, atau mengurangi gangguan selama waktu belajar. Meningkatkan keterlibatan dan pendidikan orang tua dapat dilakukan melalui program pendidikan orang tua yang menekankan pentingnya pendidikan dan bagaimana mereka dapat mendukung anak-anak mereka. Dukungan keluarga secara keseluruhan dapat ditingkatkan dengan menciptakan lingkungan rumah yang kondusif untuk belajar, menyediakan sumber daya yang diperlukan, dan memberikan motivasi serta dukungan emosional kepada siswa. Fokus pada pengembangan dan peningkatan fitur-fitur ini dapat memberikan dampak positif yang signifikan pada prestasi akademik siswa, membantu mereka mencapai potensi penuh mereka.

SIMPULAN

Kesimpulan dari hasil penelitian ini menunjukkan bahwa model Decision Tree dengan seleksi fitur *SelectByPValue* memberikan kinerja yang lebih baik dibandingkan dengan model tanpa seleksi fitur dalam memprediksi prestasi akademik siswa kelas VIII SMP Frater Don Bosco Manado. Model dengan seleksi fitur mencapai akurasi 95,56%, sementara model tanpa seleksi fitur mencapai akurasi 93,33%. Selain itu, model dengan seleksi fitur menunjukkan peningkatan pada *precision*, *recall*, dan *f1-score*.

Analisis *feature importances* menunjukkan bahwa untuk model tanpa seleksi fitur, "Nilai Rapor Matematika Kelas VIII" adalah fitur yang paling dominan. Namun, setelah menggunakan seleksi fitur *SelectByPValue*, fitur "Durasi Belajar", "Tingkat Pendidikan Ayah", dan "Dukungan Keluarga" menjadi lebih signifikan. Hal ini mengindikasikan bahwa faktor-faktor seperti durasi belajar, tingkat pendidikan orang tua, dan dukungan keluarga memiliki peran penting dalam menentukan prestasi akademik siswa.

Dengan demikian, untuk meningkatkan prestasi akademik siswa, perhatian khusus perlu diberikan pada peningkatan durasi belajar siswa, pendidikan orang tua, dan dukungan keluarga. Intervensi yang menargetkan faktor-faktor ini dapat membantu siswa mencapai hasil akademik yang lebih baik dan meningkatkan kinerja prediksi model dalam konteks pendidikan.

REFERENSI

Birba, D. E. (2020). A Comparative study of data splitting algorithms for machine learning model selection. *Degree Project in Computer Science and Engineering*, *2020*(1).

Eccles, J. S., & Roeser, R. W. (2011). Schools as developmental contexts during adolescence. *Journal of Research on Adolescence*, *21*(1). https://doi.org/10.1111/j.1532-7795.2010.00725.x

Fauzan, A. S., Irma, A., Sari, P., & Ali, I. (2024). Analisis perbandingan algoritma decision tree dan naive bayes untuk mengevaluasi prestasi belajar siswa studi kasus: SMK Al-Musyawirin. *Jurnal Mahasiswa Teknik Informatika*, *8*(1). https://doi.org/https://doi.org/10.36040/jati.v8i1.8403

Hanif, M. B., & Setiaji, G. G. (2022). Meningkatkan kinerja decision tree C4.5 dengan seleksi fitur korelasi pearson pada deteksi penyakit diabetes. *Indonesian Journal of Computer Science*. https://doi.org/https://doi.org/10.33022/ijcs.v11i2.3087

Hartanto, P. A. (2023). Penerapan algoritma decision tree untuk seleksi penerima beasiswa (studi kasus: SMPN 1 Soreang). *JS: Journal of Comprehensive Science*, *2*(7). https://doi.org/https://doi.org/10.59188/jcs.v2i7.452

Huang, J., Li, Y. F., & Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and Software Technology*, *67*. https://doi.org/10.1016/j.infsof.2015.07.004

Jijo, B. T., & Abdulazeez, A. M. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Khafid, M. (2007). Pengaruh disiplin belajar dan lingkungan keluarga terhadap hasil belajar ekonomi. *Dinamika Pendidikan*, *2*(2), 185. https://doi.org/https://doi.org/10.15294/dp.v2i2.447

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1). https://doi.org/10.1016/j.gltp.2022.04.020

Matzavela, V., & Alepis, E. (2021). Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments. *Computers and Education: Artificial Intelligence*, *2*. https://doi.org/10.1016/j.caeai.2021.100035

Nailil Amani, N., & Hayati, U. (2024). Penggunaan algoritma decision tree untuk prediksi prestasi siswa di Sekolah Dasar Negeri 3 Bayalangu Kidul. *Jurnal Mahasiswa Teknik Informatika*, *8*(1). https://doi.org/https://doi.org/10.36040/jati.v8i1.8355

Njeri, N. R. (2022). Data preparation for machine learning modelling. *International Journal of Computer Applications Technology and Research*, *11*(06). https://doi.org/10.7753/ijcatr1106.1008

Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *JCSE: International Journal of Computer Sciences and Engineering*, *6*(10), 74–78. https://doi.org/10.26438/ijcse/v6i10.7478

Potdar, K., S., T., & D., C. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, *175*(4). https://doi.org/10.5120/ijca2017915495

Puspita, E. D. (t.t.). *Pengaruh dukungan keluarga terhadap prestasi belajar siswa pada mata pelajaran ekonomi di sekolah menengah atas negeri olahraga Provinsi Riau.*

Qisthiano, M. R., Prayesy, P. A., & Ruswita, I. (2023). Penerapan algoritma decision tree dalam klasifikasi data prediksi kelulusan mahasiswa. *G-Tech: Jurnal Teknologi Terapan*, *7*(1), 21–28. https://doi.org/10.33379/gtech.v7i1.1850

Rajagukguk, S. A. (2021). Tinjauan pustaka sistematis: Prediksi prestasi belajar peserta didik dengan algoritma pembelajaran mesin. *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*, *1*(1). https://doi.org/10.20885/snati.v1i1.4

Retnowati, P., & Khotimah, T. (2020). Aplikasi forecasting kehadiran siswa di SMP 2 Jekulo menggunakan metode regresi linear. *Jurnal SIMETRIS*, *11*(2). https://doi.org/https://doi.org/10.24176/simet.v11i2.4886

Sanjaya, R. (2020). Pengaruh kegiatan ekstrakurikuler palang merah remaja terhadap prestasi siswa di SMPN 20 Kota Bengkulu. *Jurnal Pendidikan Agama Islam Indonesia (JPAII)*, *1*(1), 9–22. https://doi.org/10.37251/jpaii.v1i1.61

Saragih, R., Gunawan, I., Parlina, I., Tunas Bangsa, S., & Artikel, G. (2023). Implementasi metode regresi linier berganda untuk prediksi siswa berprestasi berdasarkan status sosial dan kedisiplinan implementation of the multiple linear regression method to predict student

achievement based on social status and discipline. *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, *2*(2), 2828–9099. https://doi.org/10.55123/jomlai.v2i2.3128

Setiawan, A. Y. (2015). *Pengaruh tingkat pendidikan orang tua dan disiplin belajar siswa terhadap prestasi belajar akuntansi siswa Kelas XI*.

Sihombing, P. R., Suryadiningrat, S., Sunarjo, D. A., & Yuda, Y. P. A. C. (2023). Identifikasi data outlier (pencilan) dan kenormalan data pada data univariat serta alternatif penyelesaiannya. *Jurnal Ekonomi Dan Statistik Indonesia*, *2*(3), 307–316. https://doi.org/10.11594/jesi.02.03.07

Suriani, U. (2023). Penerapan data mining untuk memprediksi tingkat kelulusan mahasiswa menggunakan algoritma decision tree C4.5. *Journal of Computer and Information Systems Ampera*, *3*(2). https://doi.org/10.51519/journalcisa.v4i2.393

Susanto, H., & Sudiyatno. (2014). Data mining untuk memprediksi prestasi siswa berdasarkan sosial ekonomi, motivasi, kedisiplinan dan prestasi masa lalu. *Jurnal Pendidikan Vokasi*, *4*(2). https://doi.org/http://dx.doi.org/10.21831/jpv.v4i2.2547

Tulus, T. (2004). *Peran disiplin pada perilaku dan prestasi belajar siswa.* Grasindo : Mataram., 2004.

Vujović, Ž. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6). https://doi.org/10.14569/IJACSA.2021.0120670

Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, *17*(1). https://doi.org/10.11989/JEST.1674-862X.80904120

Yulianto, A., & Firmansyah, F. (2022). Prediksi hasil belajar peserta didik menggunakan model multiple linier regression. *REMIK: Riset dan E-Jurnal Manajemen Informatika Komputer*, *6*(4), 654–663. https://doi.org/10.33395/remik.v6i4.11763