



Department of Digital Business

Journal of Artificial Intelligence and Digital Business (RIGGS)

Homepage: <https://journal.ilmudata.co.id/index.php/RIGGS>

Vol. 5 No. 2 (2026) pp: 23754-23763

P-ISSN: 2963-9298, e-ISSN: 2963-914X

Analisis Prediksi Kelulusan Siswa Menggunakan Algoritma Gradient Boosting Berbasis Orange Data Mining

Deli Ananda Delin

Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas HKBP Nommensen

Pematangsiantar

deliananda7@gmail.com

Abstrak

Prediksi kelulusan siswa merupakan salah satu penerapan Educational Data Mining yang dapat membantu institusi pendidikan dalam mengidentifikasi siswa yang berpotensi mengalami kesulitan akademik sejak dini. Informasi tersebut dapat digunakan sebagai dasar dalam pengambilan keputusan untuk memberikan pendampingan atau intervensi akademik yang lebih tepat sasaran. Penelitian ini bertujuan untuk membangun model prediksi kelulusan siswa menggunakan algoritma Gradient Boosting dengan memanfaatkan Student Performance Dataset. Proses penelitian dilakukan menggunakan aplikasi Orange Data Mining yang meliputi tahap prapemrosesan data, pembentukan model, dan evaluasi performa. Pada tahap prapemrosesan, atribut G3 ditransformasikan menjadi atribut kategorikal Target dengan dua kelas, yaitu Lulus dan Tidak Lulus, kemudian atribut G3 dipindahkan ke bagian Ignored untuk menghindari terjadinya data leakage. Model diuji menggunakan metode 5-Fold Cross Validation dan dievaluasi berdasarkan nilai Area Under Curve (AUC), Accuracy, Precision, Recall, F1-Score, Matthews Correlation Coefficient (MCC), serta Confusion Matrix. Hasil penelitian menunjukkan bahwa algoritma Gradient Boosting menghasilkan nilai AUC sebesar 0,969, Accuracy sebesar 89,6%, Precision sebesar 0,896, Recall sebesar 0,896, F1-Score sebesar 0,896, dan MCC sebesar 0,765. Hasil tersebut menunjukkan bahwa model memiliki kemampuan klasifikasi yang sangat baik dalam membedakan siswa yang termasuk kategori Lulus dan Tidak Lulus. Dengan demikian, algoritma Gradient Boosting dapat menjadi salah satu alternatif yang efektif untuk mendukung sistem prediksi kelulusan siswa dan membantu proses pengambilan keputusan di bidang pendidikan.

Kata kunci: Gradient Boosting, Student Performance, Machine Learning, Orange Data Mining, Prediksi Kelulusan Siswa.

1. Pendahuluan

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence) telah memberikan pengaruh besar terhadap berbagai bidang, salah satunya bidang pendidikan melalui penerapan Machine Learning. Pemanfaatan Machine Learning dalam pendidikan memungkinkan proses analisis data akademik dilakukan secara otomatis untuk menemukan pola tertentu yang dapat membantu pengambilan keputusan. Salah satu penerapan yang banyak dikembangkan adalah prediksi performa akademik siswa untuk mengidentifikasi peserta didik yang memiliki risiko mengalami kegagalan akademik sejak dini. Prediksi tersebut dapat membantu institusi pendidikan dalam memberikan strategi intervensi yang lebih tepat berdasarkan karakteristik data siswa. Penelitian mengenai student performance prediction menjadi salah satu bagian penting dalam bidang Educational Data Mining karena mampu memanfaatkan data akademik, sosial, dan demografi siswa sebagai dasar analisis prediktif [1], [2].

Berbagai algoritma Machine Learning telah digunakan dalam penelitian prediksi performa siswa, seperti Decision Tree, Random Forest, Support Vector Machine, dan Neural Network. Penelitian sebelumnya menunjukkan bahwa metode berbasis pembelajaran mesin mampu menghasilkan model prediksi dengan tingkat akurasi yang baik dalam mengidentifikasi pola keberhasilan akademik siswa. Li dan Liu melakukan penelitian mengenai prediksi performa mahasiswa menggunakan pendekatan Deep Learning dan menunjukkan bahwa teknologi prediksi dapat membantu pihak akademik dalam melakukan pemantauan perkembangan mahasiswa. Selain itu, penelitian lain juga menunjukkan bahwa pendekatan Machine Learning dapat digunakan untuk membangun sistem peringatan dini (early warning system) bagi siswa yang memiliki kemungkinan mengalami kegagalan akademik.

Salah satu algoritma yang memiliki kemampuan tinggi dalam melakukan proses klasifikasi adalah Gradient Boosting. Algoritma ini termasuk metode ensemble learning yang menggabungkan beberapa model sederhana (weak learner) secara bertahap untuk meningkatkan kemampuan prediksi. Gradient Boosting memiliki keunggulan dalam menangani hubungan kompleks antar atribut serta mampu menghasilkan performa yang baik pada data

dengan berbagai karakteristik. Dalam penelitian prediksi akademik, metode berbasis ensemble learning menunjukkan kemampuan yang kompetitif dibandingkan beberapa metode pembelajaran mesin lainnya.

Meskipun penelitian mengenai prediksi performa siswa menggunakan Machine Learning telah banyak dilakukan, sebagian penelitian masih berfokus pada perbandingan banyak algoritma atau menggunakan pendekatan pemrograman secara langsung. Penggunaan platform berbasis visual seperti Orange Data Mining masih memiliki peluang untuk dikembangkan karena menyediakan lingkungan analisis data yang lebih sederhana melalui konsep workflow tanpa membutuhkan proses pemrograman yang kompleks. Selain itu, sebagian penelitian sebelumnya menggunakan nilai akademik sebagai variabel utama, sehingga berpotensi menghasilkan prediksi yang kurang menggambarkan faktor pendukung keberhasilan siswa apabila terjadi ketergantungan terhadap nilai sebelumnya.

Berdasarkan permasalahan tersebut, penelitian ini melakukan implementasi algoritma Gradient Boosting menggunakan Orange Data Mining untuk melakukan prediksi kelulusan siswa berdasarkan faktor akademik, sosial, dan demografi. Dataset yang digunakan adalah Student Performance Dataset yang diperoleh dari Kaggle. Pada penelitian ini, nilai akhir siswa (G3) dikonversi menjadi kelas kategorikal yaitu Lulus dan Tidak Lulus untuk kebutuhan klasifikasi. Berbeda dengan penelitian sebelumnya yang banyak melakukan perbandingan beberapa algoritma, penelitian ini berfokus pada analisis penerapan satu algoritma yaitu Gradient Boosting dengan pendekatan visualisasi data mining menggunakan Orange Data Mining.

Tujuan penelitian ini adalah menerapkan algoritma Gradient Boosting untuk melakukan prediksi kelulusan siswa serta mengevaluasi performa model berdasarkan nilai Accuracy, Precision, Recall, F1-Score, dan Confusion Matrix. Hasil penelitian ini diharapkan dapat memberikan gambaran mengenai kemampuan Gradient Boosting dalam mendukung proses pengambilan keputusan pada bidang pendidikan berbasis data.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode klasifikasi menggunakan algoritma Gradient Boosting yang diimplementasikan melalui perangkat lunak Orange Data Mining. Penelitian bertujuan untuk memprediksi kelulusan siswa berdasarkan karakteristik akademik, sosial, dan demografi yang terdapat pada dataset Student Performance Dataset. Tahapan penelitian meliputi pengumpulan data, prapemrosesan data, pembentukan model klasifikasi, serta evaluasi performa model menggunakan metode Cross Validation.

2.1 Dataset Penelitian

Dataset yang digunakan dalam penelitian ini adalah Student Performance Dataset yang diperoleh dari platform Kaggle. Dataset terdiri dari 395 data siswa dengan 33 atribut yang mencakup informasi mengenai karakteristik siswa, kondisi keluarga, aktivitas belajar, serta nilai akademik. Variabel target penelitian dibentuk berdasarkan nilai akhir siswa (G3) yang kemudian dikonversi menjadi dua kelas, yaitu Lulus dan Tidak Lulus. Konversi tersebut dilakukan agar target memiliki tipe data kategorikal sehingga dapat digunakan pada proses klasifikasi menggunakan Orange Data Mining.

Tabel 1. Informasi Dataset

Keterangan	Deskripsi
Nama Dataset	Student Performance Dataset
Sumber Dataset	Kaggle
Jumlah Data	395
Jumlah Atribut	33
Metode	Gradient Boosting

Software	Orange Data Mining
Target	Lulus dan Tidak Lulus

2.2 Tahapan Penelitian

Tahapan penelitian dimulai dengan mengumpulkan dataset dari Kaggle yang kemudian diimpor ke dalam Orange Data Mining. Selanjutnya dilakukan proses prapemrosesan data untuk menentukan atribut yang digunakan sebagai fitur dan target klasifikasi. Setelah data siap digunakan, model dibangun menggunakan algoritma Gradient Boosting. Tahap terakhir adalah melakukan evaluasi performa model menggunakan metode 5-Fold Cross Validation serta menghitung nilai Area Under Curve (AUC), Accuracy (CA), Precision, Recall, F1-Score, dan Matthews Correlation Coefficient (MCC).

2.3 Prapemrosesan Data

Tahap prapemrosesan dilakukan untuk mempersiapkan dataset sebelum proses pembentukan model. Pada penelitian ini, nilai akhir siswa (G3) digunakan sebagai dasar pembentukan target klasifikasi. Nilai tersebut dikonversi menjadi dua kategori, yaitu Lulus apabila nilai $G3 \geq 10$ dan Tidak Lulus apabila nilai $G3 < 10$. Setelah proses konversi selesai, atribut Target ditetapkan sebagai Class Variable, sedangkan atribut G3 diatur sebagai Ignored pada Orange Data Mining agar tidak terjadi *data leakage* selama proses pelatihan model.

Tabel 2. Transformasi Target Klasifikasi

Nilai G3	Target
$G3 < 10$	Tidak Lulus
$G3 \geq 10$	Lulus

Selain itu, atribut lain seperti school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, guardian, studytime, failures, schoolsup, famsup, paid, activities, internet, romantic, famrel, freetime, goout, Dalc, Walc, health, absences, G1, dan G2 digunakan sebagai atribut prediktor dalam proses pembentukan model.

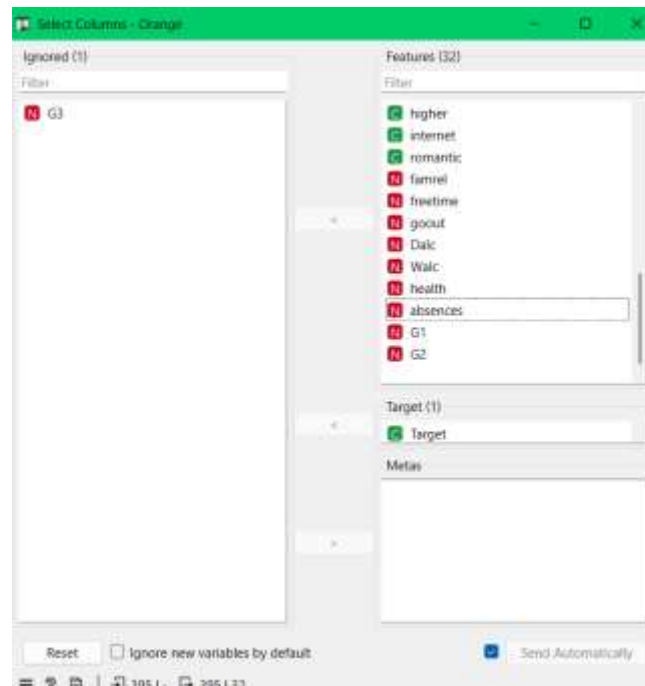
2.4 Implementasi Algoritma Gradient Boosting

Model klasifikasi dibangun menggunakan algoritma Gradient Boosting yang tersedia pada Orange Data Mining. Gradient Boosting merupakan metode *ensemble learning* yang membentuk model secara bertahap dengan menggabungkan beberapa *weak learner* sehingga mampu meningkatkan kemampuan klasifikasi. Algoritma ini bekerja dengan memperbaiki kesalahan model sebelumnya pada setiap iterasi sehingga menghasilkan model yang memiliki tingkat akurasi yang lebih baik.

Implementasi model dilakukan melalui workflow pada Orange Data Mining yang terdiri dari widget File, Select Columns, Gradient Boosting, Test & Score, dan Confusion Matrix. Widget Select Columns digunakan untuk menentukan atribut prediktor, atribut target, serta atribut yang diabaikan (*Ignored*). Selanjutnya model dilatih menggunakan widget Gradient Boosting dan dievaluasi menggunakan widget Test & Score.

transformasi selesai, atribut Target ditetapkan sebagai Class Variable, sedangkan atribut G3 dipindahkan ke bagian Ignored sehingga tidak digunakan sebagai atribut prediktor dalam proses pembentukan model.

Gambar 3. Hasil Pengaturan Atribut pada Orange Data Mining



Pengaturan tersebut dilakukan untuk menghindari terjadinya data leakage, yaitu kondisi ketika model memperoleh informasi target secara langsung selama proses pelatihan. Apabila atribut G3 tetap digunakan sebagai fitur prediktor, model akan menghasilkan performa yang sangat tinggi karena atribut tersebut merupakan sumber pembentukan variabel Target, sehingga hasil evaluasi tidak lagi mencerminkan kemampuan model dalam melakukan prediksi terhadap data baru. Menurut Zhang et al. (2021), proses pemilihan atribut (*feature selection*) merupakan salah satu faktor utama yang memengaruhi kualitas model klasifikasi pada penelitian Student Performance Prediction, karena atribut yang tidak tepat dapat menyebabkan model mengalami *overfitting* dan menghasilkan evaluasi yang bias.

Selain menghilangkan potensi *data leakage*, proses prapemrosesan juga meningkatkan objektivitas model karena prediksi dilakukan berdasarkan atribut lain, seperti jenis kelamin (*sex*), usia (*age*), tingkat pendidikan orang tua (*Medu* dan *Fedu*), waktu belajar (*studytime*), jumlah kegagalan (*failures*), ketidakhadiran (*absences*), serta nilai ujian sebelumnya (*G1* dan *G2*). Dengan demikian, model mempelajari hubungan antar atribut yang tersedia sebelum nilai akhir diperoleh, sehingga lebih mencerminkan kondisi nyata pada saat proses prediksi dilakukan. Hasil penelitian sebelumnya juga menunjukkan bahwa penggunaan atribut akademik, sosial, dan demografi secara bersamaan mampu meningkatkan kemampuan model dalam mengenali pola performa belajar siswa dibandingkan hanya menggunakan satu kelompok atribut saja.

Tahap prapemrosesan yang dilakukan pada penelitian ini juga sejalan dengan konsep Knowledge Discovery in Databases (KDD), yaitu memastikan bahwa data yang digunakan telah melalui proses seleksi, transformasi, dan pembersihan sebelum memasuki tahap pemodelan. Proses tersebut bertujuan menghasilkan model yang memiliki kemampuan generalisasi yang baik ketika diterapkan pada data yang belum pernah dipelajari sebelumnya. Penelitian mengenai prediksi performa siswa juga menegaskan bahwa kualitas data masukan memiliki pengaruh yang sangat besar terhadap performa model *machine learning*, bahkan sering kali lebih berpengaruh dibandingkan pemilihan algoritma itu sendiri.

Berdasarkan hasil prapemrosesan, seluruh atribut telah memiliki tipe data yang sesuai dan siap digunakan pada proses pembentukan model Gradient Boosting. Penambahan atribut Target sebagai variabel kelas serta pengaturan

atribut G3 ke bagian Ignored menunjukkan bahwa dataset telah memenuhi kebutuhan proses klasifikasi pada Orange Data Mining. Tahapan ini menjadi dasar penting sebelum dilakukan proses pelatihan dan evaluasi model, karena kualitas data yang baik akan menghasilkan model yang lebih akurat, stabil, dan mampu melakukan prediksi secara objektif terhadap data baru. Hasil tersebut mendukung temuan Zhang et al. (2021) bahwa keberhasilan penelitian *Educational Data Mining* sangat bergantung pada kualitas tahapan prapemrosesan yang dilakukan sebelum proses pembelajaran model.

3.2 Hasil Pengujian Model Gradient Boosting

Setelah tahap prapemrosesan selesai, proses berikutnya adalah membangun model klasifikasi menggunakan algoritma Gradient Boosting pada Orange Data Mining. Model dievaluasi menggunakan metode 5-Fold Cross Validation, yaitu teknik validasi yang membagi dataset menjadi lima bagian. Pada setiap iterasi, empat bagian digunakan sebagai data pelatihan (*training set*) dan satu bagian digunakan sebagai data pengujian (*testing set*). Proses ini dilakukan secara bergantian hingga seluruh data memperoleh kesempatan menjadi data uji. Teknik *cross validation* dipilih karena mampu memberikan estimasi performa model yang lebih stabil dan mengurangi risiko *overfitting* dibandingkan metode pembagian data tunggal.

Model dibangun menggunakan atribut-atribut prediktor yang telah melalui tahap prapemrosesan. Pada penelitian ini, atribut **Target** digunakan sebagai variabel kelas (*class variable*), sedangkan atribut **G3** ditempatkan pada bagian **Ignored** sehingga tidak ikut digunakan sebagai fitur prediktor. Pengaturan tersebut memastikan bahwa proses pembelajaran model dilakukan berdasarkan karakteristik siswa, seperti kondisi keluarga, kebiasaan belajar, nilai sebelumnya (**G1** dan **G2**), serta faktor sosial lainnya tanpa memanfaatkan nilai akhir secara langsung. Dengan demikian, hasil evaluasi yang diperoleh mencerminkan kemampuan model dalam melakukan prediksi terhadap data baru, bukan akibat adanya *data leakage*.

Berdasarkan hasil pengujian menggunakan widget Test & Score pada Orange Data Mining, algoritma Gradient Boosting menunjukkan performa klasifikasi yang sangat baik. Nilai Area Under Curve (AUC) sebesar **0,969** menunjukkan bahwa model memiliki kemampuan yang sangat tinggi dalam membedakan kelas Lulus dan Tidak Lulus. Menurut Zhang et al. (2021), nilai AUC yang mendekati 1 menunjukkan bahwa model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan dua kelas sehingga peluang terjadinya kesalahan klasifikasi menjadi lebih kecil. Selain itu, penelitian Khan dan Ghosh (2021) juga menyatakan bahwa nilai AUC merupakan salah satu indikator yang paling penting dalam mengevaluasi model prediksi performa siswa karena mampu menggambarkan kemampuan model secara keseluruhan tanpa dipengaruhi oleh distribusi kelas.

Selain AUC, model menghasilkan nilai Accuracy (CA) sebesar 0,896, yang berarti sekitar **89,6%** data berhasil diklasifikasikan sesuai dengan kelas sebenarnya. Nilai akurasi tersebut menunjukkan bahwa algoritma Gradient Boosting mampu mengenali pola hubungan antar atribut dengan baik sehingga menghasilkan prediksi yang konsisten. Hasil ini mendukung penelitian Roslan dan Chen (2022) yang menyatakan bahwa algoritma berbasis *ensemble learning* cenderung memberikan tingkat akurasi yang lebih tinggi dibandingkan algoritma klasifikasi tunggal karena mampu memperbaiki kesalahan prediksi secara bertahap melalui proses *boosting*.

Nilai Precision, Recall, dan **F1-Score** pada penelitian ini masing-masing sebesar **0,896**. Kesamaan nilai ketiga metrik tersebut menunjukkan bahwa model memiliki keseimbangan yang baik dalam mengidentifikasi kedua kelas tanpa menunjukkan kecenderungan menghasilkan prediksi yang bias terhadap salah satu kelas. Precision yang tinggi menunjukkan bahwa sebagian besar data yang diprediksi sebagai **Lulus** memang benar termasuk dalam kelas tersebut. Di sisi lain, Recall yang tinggi menunjukkan bahwa model mampu menemukan sebagian besar data yang benar-benar termasuk ke dalam masing-masing kelas. Nilai F1-Score yang sama dengan Precision dan Recall mengindikasikan bahwa model memiliki keseimbangan antara ketepatan dan kelengkapan prediksi. Kondisi ini sesuai dengan hasil kajian sistematis mengenai prediksi performa siswa yang menyatakan bahwa evaluasi model sebaiknya tidak hanya bergantung pada nilai akurasi, tetapi juga mempertimbangkan Precision, Recall, dan F1-Score agar kualitas klasifikasi dapat dinilai secara lebih komprehensif.

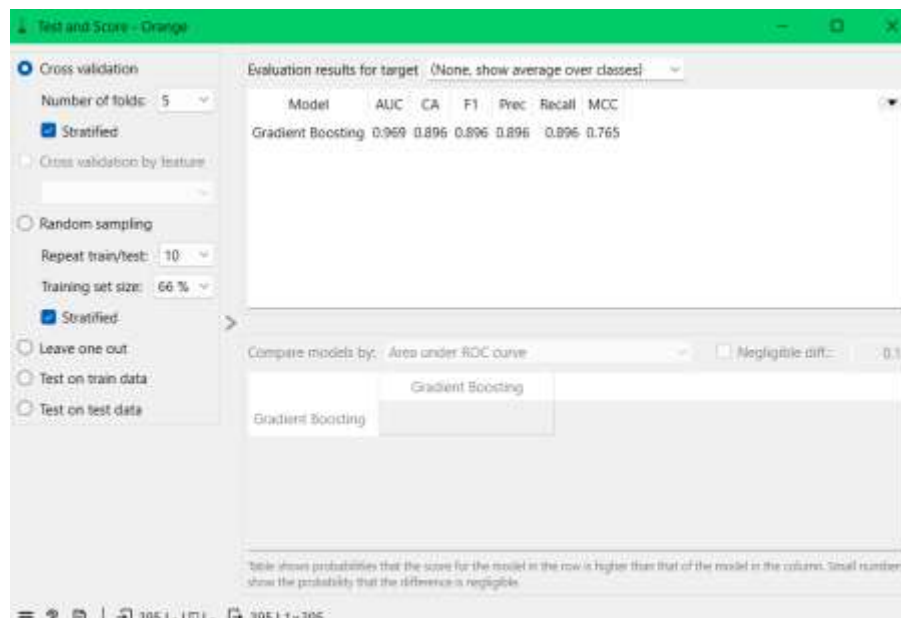
Model juga menghasilkan nilai Matthews Correlation Coefficient (MCC) sebesar 0,765. Nilai tersebut menunjukkan adanya hubungan yang kuat antara hasil prediksi model dengan kondisi sebenarnya. MCC merupakan metrik evaluasi yang cukup sensitif terhadap distribusi kelas sehingga mampu memberikan gambaran performa model secara lebih objektif, terutama ketika jumlah data antar kelas tidak sepenuhnya seimbang. Oleh

karena itu, kombinasi nilai AUC, Accuracy, Precision, Recall, F1-Score, dan MCC memberikan keyakinan bahwa model yang dibangun memiliki performa yang baik serta mampu melakukan klasifikasi secara konsisten.

Tabel 3. Hasil Evaluasi Algoritma Gradient Boosting

Parameter	Nilai
AUC	0,969
Accuracy (CA)	0,896
Precision	0,896
Recall	0,896
F1-Score	0,896
MCC	0,765

Gambar 4. Hasil Evaluasi Menggunakan Widget Test & Score Orange Data Mining



Berdasarkan hasil evaluasi tersebut, algoritma **Gradient Boosting** mampu membangun model klasifikasi yang efektif dalam memprediksi kelulusan siswa. Tingginya nilai AUC dan Accuracy menunjukkan bahwa model mampu mengenali pola hubungan antar atribut secara optimal, sedangkan nilai Precision, Recall, dan F1-Score yang seimbang menunjukkan bahwa model memiliki kemampuan klasifikasi yang stabil pada kedua kelas. Hasil ini mendukung berbagai penelitian sebelumnya yang menyatakan bahwa algoritma berbasis *ensemble learning* merupakan salah satu pendekatan yang efektif dalam menyelesaikan permasalahan Student Performance Prediction karena mampu menghasilkan model dengan tingkat akurasi tinggi dan kemampuan generalisasi yang baik terhadap data baru.

3.3 Analisis Confusion Matrix dan Pembahasan

Setelah dilakukan evaluasi menggunakan **Test & Score**, penelitian ini dilanjutkan dengan analisis **Confusion Matrix** untuk mengetahui kemampuan model dalam mengklasifikasikan setiap kelas secara lebih rinci. Berbeda dengan metrik Accuracy, Precision, Recall, dan F1-Score yang hanya memberikan nilai evaluasi secara

keseluruhan, **Confusion Matrix** mampu menunjukkan jumlah data yang berhasil maupun gagal diklasifikasikan pada masing-masing kelas. Oleh karena itu, Confusion Matrix menjadi salah satu metode evaluasi yang penting dalam penelitian klasifikasi karena dapat menggambarkan distribusi kesalahan prediksi secara lebih detail. Menurut Alwarthan et al. (2022), penggunaan Confusion Matrix sangat dianjurkan pada penelitian prediksi performa siswa karena mampu menunjukkan hubungan antara prediksi model dengan kondisi sebenarnya sehingga kualitas model dapat dianalisis secara lebih komprehensif.

Berdasarkan hasil pengujian menggunakan widget Confusion Matrix pada Orange Data Mining, algoritma Gradient Boosting mampu mengklasifikasikan sebagian besar data dengan benar. Hal tersebut terlihat dari dominasi nilai pada diagonal utama matriks yang menunjukkan jumlah prediksi yang sesuai dengan kelas sebenarnya, baik pada kategori Lulus maupun Tidak Lulus. Sebaliknya, jumlah data yang berada di luar diagonal utama relatif sedikit sehingga menunjukkan bahwa tingkat kesalahan klasifikasi masih berada pada batas yang dapat diterima. Kondisi ini mengindikasikan bahwa model memiliki kemampuan yang baik dalam mengenali pola hubungan antar atribut siswa sehingga mampu membedakan kedua kelas secara efektif.

Gambar 5. Hasil Confusion Matrix Algoritma Gradient Boosting



Hasil Confusion Matrix juga memperkuat nilai evaluasi yang telah diperoleh pada tahap sebelumnya, yaitu Accuracy sebesar 89,6%, Precision sebesar 0,896, Recall sebesar 0,896, dan F1-Score sebesar 0,896. Nilai-nilai tersebut menunjukkan bahwa model tidak hanya memiliki tingkat ketepatan prediksi yang tinggi, tetapi juga mampu mengenali sebagian besar data pada masing-masing kelas. Menurut Roslan dan Chen (2022), kombinasi Accuracy, Precision, Recall, F1-Score, dan Confusion Matrix memberikan gambaran yang lebih lengkap mengenai kualitas model dibandingkan hanya menggunakan satu metrik evaluasi saja. Oleh karena itu, hasil penelitian ini menunjukkan bahwa algoritma Gradient Boosting mampu menghasilkan performa klasifikasi yang stabil dan konsisten.

Selain itu, nilai Area Under Curve (AUC) sebesar 0,969 menunjukkan bahwa model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan kelas Lulus dan Tidak Lulus. Nilai AUC yang mendekati angka 1 mengindikasikan bahwa probabilitas model dalam memberikan prediksi yang benar jauh lebih tinggi dibandingkan prediksi yang salah. Xiao et al. (2022) menjelaskan bahwa model dengan nilai AUC di atas 0,90 dapat dikategorikan memiliki performa klasifikasi yang sangat baik karena mampu memisahkan dua kelas secara efektif. Hasil penelitian ini menunjukkan bahwa Gradient Boosting berhasil memanfaatkan hubungan antar atribut akademik dan demografi siswa untuk membangun model prediksi yang memiliki kemampuan generalisasi yang tinggi.

Keberhasilan model pada penelitian ini tidak hanya dipengaruhi oleh algoritma yang digunakan, tetapi juga oleh proses prapemrosesan data. Penambahan atribut Target sebagai variabel kelas dan pemindahan atribut G3 ke bagian Ignored berhasil menghindari terjadinya data leakage, yaitu kondisi ketika model memperoleh informasi target secara langsung selama proses pelatihan. Apabila atribut G3 tetap digunakan sebagai fitur prediktor, model akan menghasilkan tingkat akurasi yang sangat tinggi namun tidak mencerminkan kemampuan prediksi terhadap

data baru. Zhang et al. (2021) menjelaskan bahwa kualitas tahapan prapemrosesan memiliki pengaruh yang sangat besar terhadap keberhasilan model prediksi pada bidang Educational Data Mining, bahkan sering kali lebih menentukan dibandingkan pemilihan algoritma klasifikasi itu sendiri.

Hasil penelitian ini juga sejalan dengan berbagai penelitian terdahulu mengenai Student Performance Prediction. Alwarthan et al. (2022) menyatakan bahwa algoritma berbasis machine learning mampu membantu institusi pendidikan dalam mengidentifikasi siswa yang berisiko mengalami kegagalan akademik sehingga dapat dilakukan tindakan pencegahan sejak dini. Sementara itu, Chaka (2022) menyimpulkan bahwa algoritma ensemble learning, termasuk Gradient Boosting, Random Forest, dan XGBoost, merupakan pendekatan yang banyak digunakan karena memiliki kemampuan yang baik dalam menangani hubungan kompleks antar atribut pendidikan. Selain itu, Xiao et al. (2022) juga menegaskan bahwa keberhasilan model prediksi dipengaruhi oleh kombinasi antara kualitas data, proses prapemrosesan, pemilihan fitur, dan metode evaluasi yang digunakan.

Secara keseluruhan, hasil penelitian menunjukkan bahwa algoritma Gradient Boosting mampu menghasilkan model klasifikasi yang efektif dalam memprediksi kelulusan siswa berdasarkan karakteristik akademik, sosial, dan demografi. Model memperoleh Accuracy sebesar 89,6%, AUC sebesar 0,969, serta nilai Precision, Recall, dan F1-Score yang seimbang. Hasil tersebut menunjukkan bahwa model memiliki kemampuan klasifikasi yang baik dan dapat digunakan sebagai salah satu alternatif dalam mendukung proses pengambilan keputusan di bidang pendidikan, khususnya untuk mengidentifikasi siswa yang memerlukan perhatian atau pendampingan akademik lebih awal. Temuan ini juga mendukung perkembangan penerapan Educational Data Mining sebagai salah satu pendekatan berbasis data dalam meningkatkan kualitas proses pembelajaran dan evaluasi pendidikan.

4. Kesimpulan

Penelitian ini berhasil menerapkan algoritma Gradient Boosting menggunakan aplikasi Orange Data Mining untuk memprediksi kelulusan siswa berdasarkan Student Performance Dataset. Sebelum proses pemodelan dilakukan, data melalui tahap prapemrosesan dengan mengubah atribut G3 menjadi atribut kategorikal Target yang terdiri atas kelas Lulus dan Tidak Lulus, kemudian atribut G3 dipindahkan ke bagian Ignored untuk menghindari *data leakage*. Berdasarkan hasil pengujian menggunakan metode 5-Fold Cross Validation, model menghasilkan nilai AUC sebesar 0,969, Accuracy sebesar 89,6%, Precision sebesar 0,896, Recall sebesar 0,896, F1-Score sebesar 0,896, dan Matthews Correlation Coefficient (MCC) sebesar 0,765. Hasil tersebut menunjukkan bahwa algoritma Gradient Boosting mampu membangun model klasifikasi dengan tingkat akurasi yang tinggi serta memiliki kemampuan yang baik dalam membedakan siswa yang termasuk kategori Lulus dan Tidak Lulus. Penerapan model ini berpotensi digunakan sebagai sistem pendukung keputusan di lingkungan pendidikan untuk membantu sekolah mengidentifikasi siswa yang berisiko tidak lulus sejak dini. Dengan adanya informasi tersebut, pihak sekolah dapat melakukan pendampingan akademik, bimbingan belajar, maupun pemberian intervensi yang lebih tepat sasaran sehingga dapat meningkatkan keberhasilan belajar siswa. Untuk penelitian selanjutnya, disarankan melakukan perbandingan performa Gradient Boosting dengan algoritma *machine learning* lainnya, seperti Random Forest, XGBoost, atau Support Vector Machine, serta menerapkan teknik feature selection dan hyperparameter tuning agar diperoleh model dengan performa yang lebih optimal.

Reference

- [1] FAlwarthan, S., Alshahrani, A., & Alghamdi, A. (2022). *Educational data mining and machine learning for student performance prediction: A review*. **Computational Intelligence and Neuroscience**. <https://doi.org/10.1155/2022/8924028>
- [2] Chaka, C. (2022). *Machine learning and educational data mining in student performance prediction: A systematic review*. **Journal of e-Learning and Knowledge Society**, 18(3), 1–15.
- [3] Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). *Educational data mining for predicting students' academic performance using machine learning algorithms*. **Materials Today: Proceedings**, 47, 5260–5267. <https://doi.org/10.1016/j.matpr.2021.05.646>
- [4] Gerlache, H. A. M., Moreno Ger, P., & de la Fuente Valentín, L. (2021). *Towards the grade's prediction: A study of different machine learning approaches to predict grades from student interaction data*. **International Journal of Interactive Multimedia and Artificial Intelligence**. <https://doi.org/10.9781/ijimai.2021.11.007>
- [5] Handayani, L., & Priyadi. (2025). *A machine learning-based early warning system for student performance prediction: System development and empirical evaluation in higher education*. **Journal of Artificial Intelligence in Information Technology**. <https://doi.org/10.51903/92j5wj58>
- [6] Li, S., & Liu, T. (2021). *Performance prediction for higher education students using deep learning*. **Complexity**, 2021, Article ID 9958203. <https://doi.org/10.1155/2021/9958203>
- [7] Orji, F. A., & Vassileva, J. (2022). *Machine learning approach for predicting students' academic performance and study strategies based on their motivation*. **arXiv**. <https://doi.org/10.48550/arXiv.2210.08186>
- [8] P. S., A. T., & D. D. (2021). *Student performance prediction using machine learning*. **Advances in Parallel Computing Technologies**

DOI: <https://doi.org/10.31004/riggs.v5i2.12588>

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

- and Applications.** <https://doi.org/10.3233/APC210137>
- [9] Roslan, M. H., & Chen, C. J. (2022). *Educational data mining for student performance prediction: A systematic literature review (2015–2021)*. **International Journal of Emerging Technologies in Learning (iJET)**, 17(5), 147–179. <https://doi.org/10.3991/ijet.v17i05.27685>
- [10] Trakunphutthirak, R., & Lee, V. C. S. (2022). *Application of educational data mining approach for student academic performance prediction using progressive temporal data*. **Journal of Educational Computing Research**, 60(3), 742–776. <https://doi.org/10.1177/07356331211048777>
- [11] Xiao, Y., Ji, M., & Hu, Y. (2022). *Machine learning-based prediction models for student academic performance: A systematic review*. **Engineering Reports**. <https://doi.org/10.1002/eng2.12482>
- [12] Xiong, Z., Li, H., Liu, Z., Chen, Z., Zhou, H., Rong, W., & Ouyang, Y. (2024). *A review of data mining in personalized education: Current trends and future prospects*. **arXiv**. <https://doi.org/10.48550/arXiv.2401.08774>
- [13] Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). *Educational data mining techniques for student performance prediction: Method review and comparison analysis*. **Frontiers in Psychology**, 12, 698490. <https://doi.org/10.3389/fpsyg.2021.698490>
- [14] Pan, J., Zhao, Z., & Han, D. (2025). *Academic performance prediction using machine learning approaches: A survey*. **IEEE Transactions on Learning Technologies**, 18, 351–368. <https://doi.org/10.1109/TLT.2025.3554174>
- [15] Jain, A., Sharma, P., & Singh, R. (2022). *Academic performance prediction using machine learning: A comprehensive and systematic review*. **Proceedings of the 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)**. <https://doi.org/10.1109/ICESIC53714.2022.9783512>