Prosiding Seminar Nasional Indonesia
Vol. 3 No. 3 2026, hal., 46-63
e-ISSN: 3026-5169

KLASIFIKASI TINGKAT NILAI MATEMATIKA SISWA MENGGUNAKAN METODE MACHINE LEARNING BERBASIS FAKTOR SOSIAL DAN PERILAKU

Razpa Arya Wardana
Universitas Pembangunan Nasional "Veteran" Jawa Timur
220810109@student.upnjatim.ac.id

Abstract

The use of data in education is increasingly important to support data-driven decision making, particularly in monitoring students' learning progress in mathematics. This study aims to classify students' mathematics achievement levels based on social, behavioral, and academic factors using machine learning methods. The dataset used is the Student Performance Data Set from the UCI Machine Learning Repository, specifically the student-mat.csv file, which contains 395 student records with 33 attributes. The final mathematics grade (G3) is grouped into three categories: low, medium, and high. The research methodology follows the CRISP-DM approach, which includes the stages of Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. Model development is carried out using Logistic Regression, Random Forest, and XGBoost algorithms. The evaluation results show that the Random Forest algorithm achieves the best performance, with an accuracy of 0.5570 and an F1-score of 0.4975, outperforming the other algorithms. Feature analysis indicates that prior academic failures, school support, social activities, and students' absenteeism levels have a significant influence on mathematics achievement. This study is expected to help schools identify at-risk students earlier and support the planning of more targeted learning interventions.

Keywords: Student Performance, Mathematics Achievement, Machine Learning, Random Forest, Educational Data Mining

Abstrak

Pemanfaatan data dalam bidang pendidikan semakin penting untuk membantu pengambilan keputusan berdasarkan data, terutama dalam memantau kemajuan belajar siswa pada mata pelajaran matematika. Penelitian ini bertujuan untuk mengklasifikasikan tingkat nilai matematika siswa berdasarkan faktor sosial, perilaku, dan akademik menggunakan metode machine learning. Dataset yang digunakan adalah Student Performance Data Set dari UCI Machine Learning Repository, khususnya berkas student-mat.csv, yang berisi 395 data siswa dengan 33 atribut. Nilai akhir matematika (G3) dikelompokkan menjadi tiga kategori, yaitu rendah, sedang, dan tinggi. Metodologi penelitian menggunakan pendekatan CRISP-DM yang mencakup tahapan Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment. Pemrosesan model dilakukan dengan algoritma Logistic Regression, Random Forest, dan XGBoost. Hasil evaluasi menunjukkan bahwa algoritma Random Forest

memiliki kinerja terbaik dengan akurasi sebesar 0,5570 dan F1-score sebesar 0,4975, yang lebih baik dibandingkan algoritma lainnya. Analisis fitur menunjukkan bahwa kegagalan akademik sebelumnya, dukungan dari sekolah, aktivitas sosial, serta tingkat ketidakhadiran siswa berpengaruh signifikan terhadap pencapaian nilai matematika. Hasil penelitian ini diharapkan bisa membantu sekolah dalam mengidentifikasi siswa yang berisiko lebih dini serta mendukung perencanaan intervensi pembelajaran yang lebih tepat sasaran.
Kata Kunci : Student Performance, Mathematics Achievement, Machine Learning, Random Forest, Educational Data Mining

PENDAHULUAN

Matematika merupakan salah satu mata pelajaran inti yang berperan penting dalam pembentukan kemampuan berpikir logis, analitis, dan pemecahan masalah pada siswa. Capaian nilai matematika yang baik tidak hanya menjadi syarat kelulusan, tetapi juga menjadi indikator kesiapan siswa untuk melanjutkan ke jenjang pendidikan yang lebih tinggi maupun memasuki dunia kerja. Dalam konteks pendidikan modern, kemampuan numerasi dan literasi matematika juga berkaitan erat dengan kecakapan abad ke-21, seperti penalaran berbasis data, pengambilan keputusan rasional, serta kemampuan memahami informasi kuantitatif dalam kehidupan sehari-hari. Oleh karena itu, pemantauan dan peningkatan performa siswa pada mata pelajaran matematika menjadi perhatian utama bagi guru, wali kelas, maupun pihak manajemen sekolah. Kondisi tersebut menjadikan kemampuan untuk mengidentifikasi siswa dengan risiko capaian nilai matematika rendah sebagai kebutuhan strategis dalam konteks manajemen pendidikan, terutama untuk mendorong intervensi yang tepat waktu dan berkelanjutan.

Dalam praktiknya, proses identifikasi siswa berperforma rendah masih banyak bergantung pada penilaian subjektif guru serta evaluasi berkala melalui ujian harian, ujian tengah semester, dan ujian akhir semester. Pendekatan yang bersifat reaktif dan manual ini memiliki sejumlah keterbatasan. Siswa yang mengalami penurunan performa secara perlahan sering kali tidak terdeteksi sejak dini, sehingga intervensi baru diberikan setelah hasil belajar terlanjur rendah. Dampaknya, peluang perbaikan menjadi lebih sempit karena siswa telah kehilangan motivasi belajar, tertinggal materi, atau bahkan mengembangkan persepsi negatif terhadap matematika. Selain itu, guru umumnya kesulitan menelusuri secara sistematis faktor-faktor sosial dan perilaku yang berkontribusi terhadap capaian nilai, seperti latar belakang keluarga, dukungan belajar di rumah, kebiasaan belajar, jam luang, maupun gaya hidup. Faktor-faktor tersebut sering kali muncul sebagai "sinyal lemah" yang tidak terlihat pada nilai ujian semata, padahal dapat menjadi pemicu penting dari penurunan hasil belajar. Di sisi lain, berbagai data akademik dan non-akademik siswa sebenarnya telah tercatat cukup lengkap dalam sistem administrasi sekolah, namun belum dimanfaatkan secara optimal sebagai dasar

pengambilan keputusan berbasis data. Akibatnya, sekolah cenderung mengandalkan kebijakan umum yang seragam, padahal kebutuhan siswa sangat bervariasi dan memerlukan pendekatan yang lebih personal serta berbasis bukti.

Perkembangan teknik data mining dan machine learning membuka peluang untuk mengolah data pendidikan secara lebih komprehensif guna menghasilkan pola dan pengetahuan yang mendukung pengambilan keputusan. Kajian dalam bidang Educational Data Mining (EDM) menunjukkan bahwa model klasifikasi dapat digunakan untuk memprediksi performa akademik siswa melalui kombinasi atribut demografis, sosial, perilaku, dan rekam jejak akademik (Romero & Ventura, 2020). Secara konseptual, EDM berfokus pada eksplorasi data pendidikan untuk menemukan pola pembelajaran, perilaku siswa, serta faktor-faktor yang memengaruhi hasil belajar, sehingga institusi pendidikan dapat mengembangkan strategi pembelajaran yang lebih adaptif dan efektif. Salah satu dataset yang banyak digunakan dalam penelitian EDM adalah Student Performance Data Set, khususnya berkas student-mat.csv untuk mata pelajaran matematika, yang memuat atribut demografis, sosial, dan perilaku siswa, serta tiga nilai rapor (G1, G2, G3) (Cortez & Silva, 2008), (Dua & Graff, 2019). Atribut-atribut pada dataset ini mencerminkan kondisi yang relevan dengan konteks sekolah, misalnya dukungan keluarga, waktu belajar, keterlibatan sosial, kebiasaan di luar sekolah, hingga tingkat ketidakhadiran. Keberadaan variabel yang beragam membuat dataset ini cocok untuk menguji model prediksi akademik, sekaligus menilai bagaimana faktor non-akademik dapat berhubungan dengan capaian matematika siswa.

Namun demikian, sejumlah penelitian terdahulu masih memiliki keterbatasan, misalnya pemanfaatan subset variabel yang relatif terbatas, fokus pada skema klasifikasi biner (misalnya lulus/tidak lulus), atau penggunaan satu algoritma saja tanpa perbandingan sistematis dengan metode klasifikasi lain. Pada klasifikasi biner, informasi tentang tingkat capaian menjadi kurang rinci karena siswa dengan nilai "cukup" dan "sangat baik" dapat masuk dalam kelompok yang sama, sehingga rekomendasi intervensi tidak dapat dibedakan secara jelas. Padahal dalam implementasi di sekolah, kebutuhan siswa dengan nilai rendah biasanya berbeda dengan siswa kategori sedang, dan berbeda pula dengan siswa kategori tinggi yang membutuhkan pengayaan. Selain itu, penggunaan satu algoritma tanpa pembandingan berpotensi menghasilkan kesimpulan yang kurang kuat, karena setiap algoritma memiliki asumsi dan karakteristik berbeda terhadap struktur data. Oleh sebab itu, penelitian yang membandingkan beberapa algoritma dan menggunakan skema multikelas menjadi penting agar hasil lebih komprehensif, sekaligus meningkatkan relevansi implementasi dalam manajemen pendidikan.

Berdasarkan konteks tersebut, masih terdapat ruang pengembangan berupa pemodelan klasifikasi tingkat nilai matematika dalam beberapa kategori (rendah, sedang, tinggi) dengan memanfaatkan secara lebih luas atribut sosial dan perilaku yang tersedia pada student-mat.csv, sekaligus membandingkan kinerja beberapa algoritma

klasifikasi dengan karakteristik berbeda. Skema multikelas dipilih karena mampu menggambarkan variasi performa siswa secara lebih realistis, sehingga dapat mendukung pengambilan keputusan yang lebih detail, misalnya penentuan prioritas program remedial untuk kategori rendah, program pendampingan untuk kategori sedang, serta program pengayaan atau kompetisi untuk kategori tinggi. Di samping itu, penelitian ini juga menempatkan hasil prediksi sebagai bahan dukungan keputusan (decision support), bukan sebagai alat labeling permanen. Artinya, kategori yang dihasilkan model seharusnya digunakan sebagai indikator awal untuk memandu perhatian guru, bukan untuk menstigma siswa.

Penelitian ini menggunakan Logistic Regression sebagai pendekatan linear yang relatif mudah diinterpretasi (Cox, 1958), Random Forest sebagai metode ansambel berbasis pohon keputusan yang mampu menangani hubungan nonlinier secara efektif (Breiman, 2001), serta XGBoost sebagai algoritma gradient boosting yang dikenal unggul pada data tabular dengan interaksi fitur yang kompleks (Chen & Guestrin, 2016). Logistic Regression dipilih sebagai baseline karena sederhana, efisien, dan dapat memberikan gambaran awal mengenai pengaruh variabel terhadap peluang suatu kelas, terutama jika dilakukan analisis koefisien atau odds ratio. Meskipun demikian, model linear memiliki keterbatasan ketika hubungan antar variabel bersifat nonlinier atau melibatkan interaksi kompleks. Random Forest dipilih karena mampu menggabungkan banyak pohon keputusan dan mengurangi risiko overfitting melalui teknik bagging, sehingga sering memberikan performa stabil pada data dengan banyak atribut yang heterogen. XGBoost dipilih karena merupakan metode boosting yang mengoptimalkan model secara bertahap (sequential) dengan fokus pada perbaikan kesalahan prediksi sebelumnya, sehingga dikenal memiliki performa tinggi pada dataset tabular yang kompleks. Dengan membandingkan ketiga algoritma ini, penelitian dapat mengevaluasi trade-off antara interpretabilitas, kemampuan menangkap nonlinieritas, serta performa prediksi.

Selain menghasilkan model prediktif, penelitian ini juga menekankan identifikasi atribut paling berpengaruh agar temuan dapat ditindaklanjuti oleh pihak sekolah. Analisis fitur penting (feature importance) atau kontribusi variabel menjadi aspek krusial karena sekolah membutuhkan "alasan" yang dapat diterjemahkan ke kebijakan. Jika model hanya memberikan hasil klasifikasi tanpa menjelaskan faktor pendukungnya, maka sulit bagi sekolah untuk merancang intervensi. Sebaliknya, jika diketahui bahwa ketidakhadiran tinggi atau riwayat kegagalan akademik sebelumnya merupakan indikator kuat, sekolah dapat merancang kebijakan monitoring kehadiran, konseling akademik, atau pendampingan belajar. Jika dukungan dari sekolah atau aspek sosial tertentu berpengaruh, sekolah dapat memperkuat program dukungan belajar, pembinaan, dan komunikasi dengan orang tua. Dengan demikian, luaran penelitian tidak berhenti pada metrik akurasi, tetapi juga menghasilkan wawasan praktis yang relevan.

Sejalan dengan rumusan masalah tersebut, penelitian ini bertujuan untuk: (1) mengembangkan model klasifikasi tingkat nilai matematika siswa berdasarkan nilai akhir G3 dengan memanfaatkan atribut sosial dan perilaku pada student-mat.csv; (2) membandingkan kinerja algoritma Logistic Regression, Random Forest, dan XGBoost dalam mengklasifikasikan tingkat performa matematika siswa; dan (3) mengidentifikasi atribut-atribut yang berpengaruh besar terhadap klasifikasi kategori nilai matematika. Untuk mencapai tujuan tersebut, penelitian mengadopsi kerangka CRISP-DM sebagai metodologi kerja yang sistematis, dimulai dari pemahaman kebutuhan sekolah sebagai "business problem" hingga rencana penerapan (deployment) sebagai sistem pendukung keputusan. Tahapan Data Preparation menjadi krusial karena dataset mengandung kombinasi data numerik dan kategorikal, sehingga diperlukan transformasi yang tepat agar algoritma dapat belajar secara optimal. Tahap Evaluation tidak hanya menilai akurasi, tetapi juga metrik yang lebih representatif untuk data multikelas seperti F1-score, mengingat distribusi kelas dapat tidak seimbang dan setiap kelas memiliki konsekuensi intervensi yang berbeda.

Hasil penelitian diharapkan berkontribusi praktis bagi sekolah dalam deteksi dini siswa berisiko serta perencanaan intervensi pembelajaran yang lebih tepat sasaran. Dengan adanya model klasifikasi, sekolah dapat melakukan pemetaan siswa secara lebih cepat dan konsisten berdasarkan data, sehingga guru dapat memprioritaskan pendampingan pada kelompok yang memerlukan. Intervensi yang lebih dini dapat berupa program remedial terstruktur, bimbingan belajar tambahan, pelatihan strategi belajar matematika, serta kolaborasi dengan wali kelas dan orang tua untuk memperbaiki kebiasaan belajar di rumah. Selain itu, model juga dapat mendukung monitoring berkala tanpa menunggu nilai ujian akhir, sehingga sekolah dapat mengurangi potensi keterlambatan penanganan. Dari sisi akademik, penelitian ini diharapkan memberikan kontribusi berupa studi perbandingan kinerja beberapa algoritma klasifikasi pada Student Performance Data Set untuk tugas klasifikasi multikelas tingkat nilai matematika. Perbandingan ini dapat memperkaya literatur EDM, terutama terkait pemilihan algoritma yang sesuai untuk data pendidikan dengan karakter heterogen, sekaligus menjadi referensi bagi penelitian lanjutan dalam pengembangan model yang lebih akurat, lebih adil, serta lebih mudah diimplementasikan di lingkungan sekolah.

Pada akhirnya, pemanfaatan data pendidikan melalui machine learning bukan semata untuk meningkatkan nilai metrik model, melainkan untuk mendukung proses pendidikan yang lebih responsif, terukur, dan berorientasi pada kebutuhan siswa. Dengan memadukan pendekatan prediktif dan analisis faktor, penelitian ini diharapkan mampu menjembatani kebutuhan praktis sekolah dan pengembangan ilmiah di bidang EDM, sehingga data yang selama ini tersimpan dalam administrasi sekolah dapat diubah menjadi pengetahuan yang bermanfaat bagi peningkatan kualitas pembelajaran matematika.

METODE PENELITIAN

Metodologi penelitian ini dimulai dari tahap identifikasi masalah, yaitu proses awal untuk memformulasikan isu utama yang akan dikaji secara terarah dan terukur. Pada tahap ini peneliti melakukan penelaahan kondisi nyata di lingkungan pendidikan, khususnya terkait tantangan sekolah dalam mendeteksi secara dini siswa yang berpotensi memperoleh nilai matematika rendah. Meskipun sekolah umumnya memiliki data akademik serta data sosial dan perilaku siswa yang relatif lengkap (misalnya data ketidakhadiran, waktu belajar, dukungan keluarga, hingga aktivitas di luar sekolah), data tersebut sering belum dimanfaatkan secara optimal untuk mendukung keputusan berbasis data. Oleh sebab itu, peneliti merumuskan fokus penelitian, menyusun pertanyaan penelitian, serta menetapkan tujuan penelitian yang ingin dicapai. Tujuan utama penelitian mencakup pengembangan model klasifikasi tingkat capaian nilai matematika siswa (misalnya kategori rendah, sedang, dan tinggi) serta identifikasi faktor-faktor sosial dan perilaku yang paling berpengaruh terhadap hasil klasifikasi. Tahap ini juga mencakup penentuan ruang lingkup penelitian (misalnya hanya menggunakan dataset *student-mat.csv*), batasan penelitian, serta luaran yang diharapkan, yaitu model yang tidak hanya akurat tetapi juga mampu memberikan wawasan untuk intervensi pembelajaran.

Tahap berikutnya adalah data understanding, yang berfokus pada pemahaman mendalam terhadap dataset dan konteks datanya. Pada tahap ini peneliti mempelajari struktur data secara menyeluruh, termasuk jumlah sampel, banyaknya atribut, definisi setiap variabel, serta tipe data yang digunakan (numerik, kategorikal, maupun ordinal). Peneliti juga memeriksa distribusi nilai target, yaitu nilai akhir matematika (G3), untuk memahami sebaran performa siswa serta mengidentifikasi kemungkinan terjadinya ketidakseimbangan kelas (*class imbalance*) ketika G3 dikonversi menjadi beberapa kategori. Selain itu, dilakukan eksplorasi terhadap fitur-fitur yang secara teoritis dan empiris relevan dengan prestasi matematika, seperti tingkat ketidakhadiran, intensitas waktu belajar, status dukungan keluarga, akses terhadap sumber belajar, dan aktivitas sosial. Analisis pada tahap ini umumnya dilengkapi statistik deskriptif (rata-rata, median, standar deviasi), visualisasi awal (histogram, bar chart, boxplot), serta pemeriksaan korelasi atau keterkaitan antarvariabel untuk mengenali pola dasar dan indikasi hubungan potensial. Tahap data understanding juga mencakup deteksi awal terhadap masalah kualitas data, misalnya nilai hilang (*missing value*), inkonsistensi penulisan kategori, atau nilai ekstrem yang tidak wajar.

Selanjutnya, penelitian memasuki tahap data preparation, yaitu tahap yang bertujuan menyiapkan data agar layak dan optimal digunakan dalam pemodelan. Kegiatan utama pada tahap ini meliputi pembersihan data, penanganan *missing value* (misalnya dengan imputasi atau penghapusan jika proporsinya kecil), penyelarasan kategori (misalnya menyamakan format kategori yang tidak konsisten), serta penanganan outlier yang tidak wajar agar tidak mengganggu proses pelatihan model.

Karena dataset memuat variabel kategorikal, peneliti melakukan proses *encoding* seperti *one-hot encoding* atau *label encoding* sesuai kebutuhan algoritma yang digunakan. Untuk fitur numerik, penskalaan (*scaling*) dapat dilakukan bila diperlukan, terutama untuk algoritma yang sensitif terhadap skala, seperti Logistic Regression. Tahap ini juga mencakup pembentukan variabel target dengan mengonversi nilai G3 ke dalam kelas-kelas tertentu (rendah, sedang, tinggi) berdasarkan aturan kategorisasi yang ditetapkan. Setelah data bersih dan tertransformasi, peneliti melakukan pemisahan data menjadi himpunan latih, validasi, dan uji. Pemisahan dilakukan secara *stratified* agar proporsi kelas pada masing-masing subset tetap seimbang, sehingga evaluasi model menjadi lebih representatif dan tidak bias.

Tahap modelling merupakan proses pengembangan model klasifikasi menggunakan data yang telah dipersiapkan. Pada tahap ini peneliti membangun model baseline sebagai pembanding awal, kemudian melatih model utama menggunakan beberapa algoritma, yakni Logistic Regression, Random Forest, dan XGBoost. Logistic Regression digunakan sebagai model yang relatif sederhana dan interpretatif, Random Forest digunakan karena mampu menangani hubungan nonlinier dan interaksi fitur melalui pendekatan ansambel, sedangkan XGBoost digunakan karena dikenal kuat pada data tabular dengan kompleksitas interaksi fitur yang tinggi. Setiap algoritma dikonfigurasi melalui penentuan parameter awal, lalu dapat dilakukan *hyperparameter tuning* untuk meningkatkan performa, misalnya dengan *grid search* atau *random search* yang dikombinasikan dengan validasi silang (*cross-validation*). Validasi silang membantu memastikan model tidak hanya bagus pada satu pembagian data, tetapi juga stabil pada beberapa skenario pembagian. Model yang telah dilatih kemudian disimpan (misalnya dalam bentuk file model) agar dapat diuji pada data uji dan digunakan pada tahap implementasi berikutnya.

Tahap evaluasi bertujuan untuk menilai kinerja model yang telah dibangun menggunakan himpunan data uji yang tidak pernah terlibat dalam proses pelatihan. Pada tahap ini peneliti menghitung metrik evaluasi seperti akurasi, precision, recall, dan F1-score, baik secara keseluruhan maupun per kelas. Analisis confusion matrix dilakukan untuk melihat pola kesalahan prediksi, terutama pada kelas nilai rendah karena kelas ini biasanya menjadi prioritas intervensi sekolah. Evaluasi juga dapat mencakup perbandingan antar model untuk menentukan algoritma terbaik berdasarkan metrik yang paling relevan. Selain itu, peneliti melakukan analisis interpretabilitas melalui *feature importance* (pada Random Forest dan XGBoost) atau koefisien model (pada Logistic Regression) untuk mengidentifikasi variabel sosial dan perilaku yang paling berpengaruh. Hasil analisis ini menjadi dasar rekomendasi praktis bagi sekolah, misalnya memfokuskan intervensi pada peningkatan kehadiran, pendampingan belajar, atau penguatan dukungan sekolah dan keluarga.

Tahap terakhir adalah penyusunan laporan, yaitu proses mengintegrasikan seluruh tahapan dan temuan penelitian dalam bentuk dokumen ilmiah yang sistematis,

baik berupa skripsi maupun artikel jurnal. Pada tahap ini peneliti menyusun bagian pendahuluan yang memuat latar belakang dan urgensi penelitian, tinjauan pustaka yang mengkaji teori terkait EDM dan algoritma klasifikasi, metodologi yang menjelaskan tahapan penelitian secara rinci, hasil dan pembahasan yang menampilkan performa model serta interpretasi faktor berpengaruh, hingga kesimpulan dan saran. Penyusunan laporan juga melibatkan penyajian tabel, grafik, dan visualisasi yang mendukung argumentasi ilmiah, serta pencantuman daftar pustaka sesuai gaya sitasi yang ditentukan. Dengan demikian, metodologi ini tidak hanya menghasilkan model klasifikasi, tetapi juga menyediakan kerangka ilmiah yang kuat dan dapat direplikasi untuk penelitian atau implementasi lanjutan di lingkungan sekolah.

HASIL DAN PEMBAHASAN

1. Business Understanding

Tahap Business Understanding bertujuan untuk memahami masalah utama yang mendasari penelitian ini serta tujuan yang ingin dicapai dengan menggunakan data dan teknik machine learning. Dalam bidang pendidikan, khususnya pada pelajaran matematika, nilai siswa sering dianggap sebagai indikator utama kesuksesan proses belajar mengajar. Akan tetapi, cara evaluasi yang dilakukan saat ini masih bersifat reaktif, yaitu dilakukan setelah nilai siswa menurun secara signifikan. Selain itu, guru dan pihak sekolah masih kesulitan mengenali siswa yang berpotensi memiliki nilai matematika rendah secara dini, terutama jika penurunan performa terjadi perlahan. Faktor-faktor seperti kebiasaan belajar, dukungan keluarga, waktu luang, dan gaya hidup sebenarnya juga mempengaruhi capaian nilai matematika siswa. Data-data terkait sudah ada di dalam sistem administrasi sekolah, tetapi belum dimanfaatkan secara optimal dalam pengambilan keputusan berbasis data. Berdasarkan situasi tersebut, tujuan bisnis dari penelitian ini adalah membantu sekolah dalam mengidentifikasi tingkat nilai matematika siswa secara lebih terstruktur dan objektif.

Dengan demikian, intervensi pembelajaran dapat dilakukan lebih awal dan tepat sasaran. Sistem klasifikasi berbasis data diharapkan dapat meningkatkan kualitas pengambilan keputusan dalam merencanakan strategi pembelajaran, pendampingan siswa, serta kebijakan akademik lainnya. Sementara itu, tujuan dari sisi data mining adalah membangun model klasifikasi tingkat nilai matematika siswa menggunakan algoritma machine learning, yaitu Logistic Regression, Random Forest, dan XGBoost. Model yang dikembangkan diharapkan mampu mengelompokkan siswa ke dalam kategori nilai matematika (rendah, sedang, tinggi), serta mampu mengidentifikasi faktor-faktor sosial dan perilaku yang paling berpengaruh dalam menentukan hasil klasifikasi tersebut.

2. Data Understanding

Data Understanding dilakukan dengan menggunakan dataset Student Performance Data Set yang diambil dari UCI Machine Learning Repository, terutama file **student-mat.csv.** Dataset ini berisi informasi tentang siswa sekolah menengah dalam pelajaran matematika, terdiri dari 395 siswa dan 33 atribut. Atribut yang digunakan mencakup data demografis, latar belakang keluarga, faktor akademik, serta aspek sosial dan perilaku siswa, dengan nilai akhir G3 sebagai variabel yang ingin dicari hubungannya.

Gambar 3. 1 Data Understanding

3. Data Preparation

Data preparation adalah tahap dalam proses data mining yang berfokus pada penyiapan data mentah agar layak dan siap digunakan untuk pemodelan. Pada tahap ini, data yang semula masih bercampur antara nilai tidak lengkap, format yang beragam, serta skala yang berbeda-beda diolah terlebih dahulu sehingga menjadi himpunan data yang bersih, konsisten, dan terstruktur sesuai kebutuhan algoritma yang akan digunakan. Tujuannya adalah meminimalkan noise dan kesalahan, mengurangi bias, serta memastikan bahwa pola yang dipelajari model benar-benar berasal dari informasi yang relevan, bukan dari artefak kualitas data yang buruk.

Dalam konteks penelitian ini dengan menggunakan dataset student-mat.csv, tahap *data preparation* meliputi beberapa langkah utama, yaitu: melakukan pemeriksaan dan pembersihan awal terhadap data, membentuk variabel kelas target berdasarkan nilai akhir G3, melakukan *encoding* terhadap seluruh atribut kategorikal (demografis, keluarga, dukungan belajar, dan gaya hidup), melakukan penskalaan pada fitur numerik untuk mendukung kinerja algoritma Logistic Regression, menangani kemungkinan ketidakseimbangan jumlah siswa pada masing-masing kategori nilai, serta membagi data menjadi himpunan latih, validasi, dan uji yang selanjutnya digunakan pada tahap pemodelan dengan Logistic Regression, Random Forest, dan XGBoost.

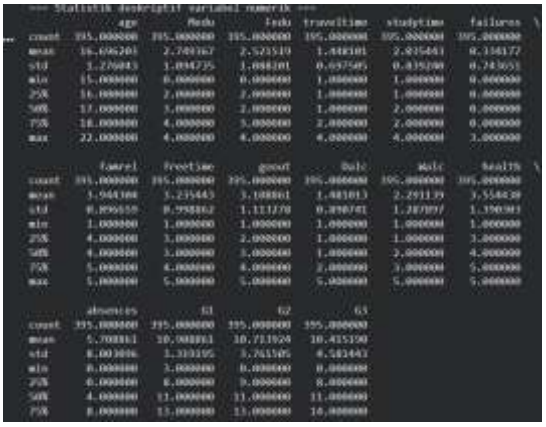
Gambar 3. 2 Data Preparation

```
=== Info dataframe ===
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      395 non-null    object
 1   sex         395 non-null    object
 2   age         395 non-null    int64
 3   address     395 non-null    object
 4   famsize     395 non-null    object
 5   Pstatus     395 non-null    object
 6   Medu        395 non-null    int64
 7   Fedu        395 non-null    int64
 8   Mjob        395 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    395 non-null    object
 12  traveltime  395 non-null    int64
 13  studytime   395 non-null    int64
 14  failures    395 non-null    int64
```

Gambar 3. 3 Jumlah Missing Value

```
=== Jumlah missing value per kolom ===
school        0
sex           0
age           0
address       0
famsize       0
Pstatus       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
nursery       0
higher        0
internet      0
romantic      0
famrel        0
freetime      0
goout         0
Dalc          0
Walc          0
```

Gambar 3. 4 Statistik Deskriptif



4. Modelling

Tahap modelling adalah bagian utama dalam penelitian ini yang bertujuan untuk membuat model klasifikasi nilai matematika siswa berdasarkan dataset yang sudah melewati proses persiapan data. Di tahap ini, algoritma machine learning digunakan untuk belajar tentang pola hubungan antara atribut sosial, perilaku, dan akademik siswa dengan tingkat nilai matematika, yang dinyatakan dalam nilai akhir (G3). Sebelum memulai proses pemodelan, data dibagi menjadi dua bagian, yaitu data latih dan data uji.

Pembagian ini dilakukan agar model yang dibuat bisa diuji secara objektif menggunakan data yang belum pernah dilihat sebelumnya. Biasanya, data latih mencakup 80% dari total data, sedangkan data uji hanya 20%. Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk mengevaluasi kemampuan model dalam memprediksi data baru.

a. Logistic Regression

Algoritma pertama yang digunakan dalam penelitian ini adalah Logistic Regression. Logistic Regression adalah metode klasifikasi yang berbasis model linier, digunakan untuk menggambarkan hubungan antara variabel independen dengan kemungkinan suatu kelas tertentu. Algoritma ini dipilih karena memiliki kelebihan dalam hal kemudahannya untuk diinterpretasikan, sehingga pengaruh setiap variabel terhadap hasil klasifikasi bisa dilihat melalui nilai koefisien model. Dalam penelitian ini, Logistic Regression digunakan sebagai model dasar untuk membandingkan hasilnya dengan algoritma lain yang lebih rumit.

Model Logistic Regression dilatih menggunakan data latih yang sudah diproses sebelumnya, kemudian melakukan prediksi terhadap data uji untuk mendapatkan hasil klasifikasi mengenai nilai matematika siswa. Hasil dari model ini berfungsi sebagai acuan awal dalam mengevaluasi efektivitas algoritma lain yang digunakan dalam penelitian.

Tabel 3. 1 Logistic Regression

Logistic Regression	
Evaluation	Score
Accuracy	0,4430
Precision_macro	0,3921
Recall_macro	0,3663
F1_macro	0,3655

b. Random Forest

Algoritma kedua yang digunakan adalah Random Forest, yaitu metode klasifikasi yang bekerja dengan menggabungkan banyak pohon keputusan (decision tree). Cara kerjanya adalah dengan mengambil hasil dari setiap pohon keputusan lalu menyatukannya menjadi keputusan akhir yang lebih akurat dan stabil. Kelebihan utama dari Random Forest adalah kemampuannya menghadapi data yang mempunyai hubungan yang tidak langsung, serta mencegah keadaan overfitting.

Dalam penelitian ini, Random Forest digunakan untuk menggambarkan hubungan yang rumit antara faktor sosial, perilaku, dan akademik terhadap hasil nilai matematika siswa. Selain itu, metode ini juga bisa memberikan informasi tentang pentingnya setiap variabel, yaitu seberapa besar kontribusi dari masing-masing atribut dalam mempengaruhi hasil klasifikasi. Informasi ini sangat penting untuk mengetahui faktor-faktor utama yang mempengaruhi prestasi belajar matematika siswa.

Model Random Forest dilatih menggunakan data latih, kemudian diuji dengan data uji untuk mendapatkan hasil prediksi. Hasil kinerja model ini kemudian dibandingkan dengan model Logistic Regression untuk mengetahui peningkatan akurasi dan kestabilan prediksi.

Tabel 3. 2 Random Forest

Random Forest	
Evaluation	Score
Accuracy	0,5570
Precision_macro	0,5299
Recall_macro	0,4937

F1_macro	0,4975

c. XGBoost

Algoritma ketiga yang digunakan adalah XGBoost (Extreme Gradient Boosting). XGBoost adalah metode boosting yang bekerja dengan membangun model secara bertahap, di mana setiap model baru berusaha memperbaiki kesalahan yang dibuat oleh model sebelumnya. Algoritma ini terkenal memiliki performa yang sangat bagus dalam mengolah data berbentuk tabel dan mampu memahami hubungan antar fitur yang rumit.

Dalam penelitian ini, XGBoost digunakan untuk mencapai hasil klasifikasi yang terbaik. Pelatihan model dilakukan menggunakan data latih yang sama dengan algoritma sebelumnya agar perbandingan hasilnya adil. Selain itu, parameter-parameter model (hyperparameter tuning) juga diatur dengan tepat untuk meningkatkan kinerja, seperti menentukan nilai learning rate, jumlah pohon (n_estimators), dan tingkat kedalaman pohon (max_depth).

Setelah pelatihan selesai, model XGBoost diuji menggunakan data uji untuk memprediksi tingkat nilai matematika siswa. Hasil prediksi dari model ini kemudian dibandingkan dengan Logistic Regression dan Random Forest, agar dapat mengetahui algoritma mana yang memiliki performa terbaik dalam mengklasifikasikan tingkat nilai matematika siswa.

Tabel 3. 3 XGBoost

XGBoost	
Evaluation	Score
Accuracy	0,4810
Precision_macro	0,4517
Recall_macro	0,4370
F1_macro	0,4380

5. Evaluation

Evaluasi dilakukan untuk mengetahui seberapa baik model mampu mengklasifikasikan tingkat nilai matematika siswa secara akurat dan seimbang di ketiga kelas yaitu rendah, sedang, dan tinggi. Data dibagi menjadi data latih dan data uji dengan perbandingan 80:20 secara stratified, agar proporsi setiap kelas dalam data uji tetap sesuai dengan distribusi awalnya. Kinerja model diukur menggunakan beberapa metrik yaitu Accuracy, Precision_macro, Recall_macro, dan F1_macro. Pendekatan

macro dipilih karena penelitian ini fokus pada performa rata-rata di semua kelas, bukan hanya kelas yang jumlah datanya banyak saja.

Hasil uji menunjukkan bahwa Logistic Regression sebagai model dasar menghasilkan nilai Accuracy sekitar 0,4430, Precision_macro 0,3921, Recall_macro 0,3663, dan F1_macro 0,3655. Nilai tersebut menunjukkan bahwa pendekatan linear sederhana belum mampu menggambarkan hubungan yang lebih rumit antara atribut sosial, perilaku, dan akademik dengan kategori nilai G3. Di model Random Forest, performa meningkat cukup tajam dengan nilai Accuracy sekitar 0,5570, Precision_macro 0,5299, Recall_macro 0,4937, dan F1_macro 0,4975. Ini menunjukkan bahwa metode ensemble yang menggunakan banyak pohon keputusan lebih baik dalam memahami hubungan nonlinier dan interaksi antar fitur di dataset student-mat. Sementara itu, XGBoost memberikan nilai Accuracy sekitar 0,4810, Precision_macro 0,4517, Recall_macro 0,4370, dan F1_macro 0,4380. Performa XGBoost secara umum masih di bawah Random Forest, meskipun lebih baik dari Logistic Regression.

Analisis confusion matrix menunjukkan bahwa model belum sepenuhnya seimbang dalam mengenali ketiga kelas. Namun, performa pada kelas yang jumlah datanya lebih besar cenderung lebih baik dibanding kelas dengan jumlah data yang lebih sedikit. Selain itu, analisis penting fitur dari model berbasis pohon menunjukkan bahwa variabel seperti jumlah kegagalan sebelumnya (failures), dukungan sekolah (schoolsup), intensitas keluar bersama teman (goout), dan jumlah ketidakhadiran (absences) memiliki kontribusi besar dalam menentukan kategori nilai matematika. Berdasarkan hasil evaluasi keseluruhan, model Random Forest dipilih sebagai model terbaik yang akan digunakan pada tahap implementasi karena mampu memberikan keseimbangan yang relatif baik antara akurasi dan F1-score pada tugas klasifikasi multikelas.

6. Deployment

Tahap deployment dalam penelitian ini fokus pada bagaimana hasil dari model dan temuan analisis dapat digunakan dalam situasi nyata di lingkungan pendidikan. Implementasi awal dilakukan dengan cara menulis artikel ilmiah sesuai format jurnal, yang secara rapi mencatat seluruh tahapan CRISP-DM, mulai dari memahami bisnis hingga evaluasi, serta menampilkan hasil perbandingan kinerja antara model Logistic Regression, Random Forest, dan XGBoost.

Secara langsung, model terbaik yaitu Random Forest bisa digunakan sebagai bagian dari sistem bantuan pengambilan keputusan di sekolah. Dengan memanfaatkan data sosial, perilaku, dan akademik siswa yang sudah ada, sekolah dapat menggunakan model ini untuk mengelompokkan siswa ke dalam kategori risiko nilai matematika rendah, sedang, atau tinggi sejak awal masa pembelajaran. Informasi ini bisa menjadi dasar bagi guru, wali kelas, atau konselor dalam merancang program intervensi yang lebih tepat, misalnya bimbingan belajar tambahan, pemantauan kehadiran, atau peningkatan dukungan belajar di rumah.

Di masa depan, penerapan model ini bisa dikembangkan lebih jauh dalam bentuk aplikasi atau dashboard sederhana yang memungkinkan pengguna seperti guru atau staf akademik untuk memasukkan data siswa dan langsung memperoleh prediksi kategori nilai serta fitur-fitur yang berpengaruh terhadap hasil klasifikasi. Dengan demikian, hasil penelitian ini tidak hanya berhenti di tingkat akademik, tetapi juga memberikan kontribusi nyata dalam meningkatkan kualitas pembelajaran matematika di jenjang sekolah menengah.

Analisis/Diskusi

Pada bagian analisis dan diskusi, temuan penelitian ini ditafsirkan dalam kerangka *Educational Data Mining* (EDM), yaitu pendekatan yang memanfaatkan data pendidikan untuk menghasilkan pola pengetahuan yang dapat mendukung pengambilan keputusan berbasis data. Dalam konteks ini, tujuan analisis tidak hanya menilai "seberapa tinggi akurasi model", tetapi juga menjelaskan mengapa model bekerja demikian, faktor apa yang paling berkontribusi, serta bagaimana hasilnya dapat diterjemahkan menjadi rekomendasi intervensi yang relevan bagi sekolah. Evaluasi dilakukan untuk memastikan bahwa model mampu mengklasifikasikan tingkat nilai matematika siswa secara akurat dan seimbang pada tiga kelas, yaitu rendah, sedang, dan tinggi, dengan skema pembagian data latih dan data uji 80:20 secara stratified agar proporsi setiap kelas di data uji tetap mencerminkan distribusi awal. Pendekatan stratified ini penting karena pada klasifikasi multikelas, ketidakseimbangan distribusi kelas dapat membuat model cenderung "menang" di kelas yang jumlahnya lebih besar namun lemah di kelas minoritas, padahal kelas minoritas seperti kategori nilai rendah sering menjadi prioritas utama untuk deteksi dini.

Kinerja model diukur menggunakan Accuracy, Precision_macro, Recall_macro, dan F1_macro. Pemilihan metrik *macro* menunjukkan bahwa penelitian ini menekankan performa rata-rata pada seluruh kelas, bukan hanya kelas dominan. Dengan perspektif EDM, penggunaan macro menjadi relevan karena sekolah membutuhkan sistem yang adil dan informatif untuk semua kategori nilai, terutama untuk menghindari kondisi ketika model tampak "baik" secara akurasi total, namun sebenarnya gagal mengenali kelompok siswa berisiko. Hasil uji menunjukkan bahwa Logistic Regression sebagai model dasar memperoleh Accuracy 0,4430; Precision_macro 0,3921; Recall_macro 0,3663; dan F1_macro 0,3655. Nilai ini mengindikasikan bahwa pendekatan linear sederhana belum cukup mampu menggambarkan hubungan yang lebih kompleks antara atribut sosial, perilaku, dan akademik dengan kategori nilai G3. Secara interpretatif, hal tersebut masuk akal karena variabel pendidikan sering berinteraksi secara nonlinier; misalnya, dampak absensi terhadap performa matematika dapat berbeda pada siswa dengan riwayat kegagalan akademik tinggi dibanding siswa tanpa kegagalan, atau efek dukungan sekolah dapat lebih terasa pada kelompok tertentu.

Dalam kondisi seperti ini, model linear cenderung kurang fleksibel menangkap interaksi yang tidak eksplisit.

Performa meningkat ketika menggunakan Random Forest. Model ini mencapai Accuracy 0,5570; Precision_macro 0,5299; Recall_macro 0,4937; dan F1_macro 0,4975. Peningkatan yang cukup tajam ini memperlihatkan bahwa metode ansambel berbasis banyak pohon keputusan lebih efektif dalam memodelkan hubungan nonlinier serta interaksi antarfitur di dataset *student-mat*. Temuan ini memperkuat argumen teoritis bahwa data pendidikan yang memuat kombinasi aspek akademik, sosial, dan perilaku membutuhkan pendekatan yang mampu menangkap pola yang tidak selalu linier dan tidak selalu memiliki kontribusi yang berdiri sendiri. Sementara itu, XGBoost menghasilkan Accuracy 0,4810; Precision_macro 0,4517; Recall_macro 0,4370; dan F1_macro 0,4380, yakni lebih baik daripada Logistic Regression namun masih di bawah Random Forest pada konfigurasi penelitian ini. Secara analitis, hasil tersebut dapat dipahami mengingat ukuran data yang relatif terbatas (395 siswa); pada data kecil, algoritma boosting dapat menjadi sensitif terhadap pengaturan hiperparameter dan berisiko tidak mencapai titik optimal jika tuning belum sepenuhnya sesuai. Walaupun pada dokumen disebutkan bahwa hiperparameter XGBoost diatur untuk meningkatkan kinerja (misalnya *learning rate*, jumlah pohon, dan kedalaman pohon), performanya tetap belum melampaui Random Forest dalam skenario ini. Oleh karena itu, pilihan Random Forest sebagai model terbaik menjadi beralasan karena menawarkan keseimbangan yang lebih kuat antara akurasi dan kestabilan prediksi pada tugas multikelas.

Analisis juga diperkuat melalui *confusion matrix*. Hasilnya menunjukkan bahwa model "belum sepenuhnya seimbang" dalam mengenali ketiga kelas, dan performa pada kelas dengan jumlah data lebih besar cenderung lebih baik dibanding kelas dengan jumlah data lebih sedikit. Temuan ini penting dalam diskusi karena menyiratkan adanya tantangan khas pada klasifikasi multikelas di domain pendidikan, yaitu potensi bias kelas. Dalam implementasi sekolah, konsekuensi dari bias ini bisa cukup signifikan: jika model kurang sensitif pada kelas "nilai rendah", maka siswa yang seharusnya menjadi target intervensi dini justru berpotensi tidak terdeteksi. Sebagai ide peneliti untuk memperkaya analisis, hasil confusion matrix dapat dihubungkan langsung dengan strategi evaluasi yang lebih berorientasi risiko, misalnya memprioritaskan interpretasi pada kesalahan yang berkaitan dengan kelas rendah, karena kesalahan pada kelas ini memiliki dampak praktis yang lebih besar bagi sekolah.

Selanjutnya, penelitian tidak berhenti pada pembandingan performa, tetapi juga memaknai hasil melalui analisis kontribusi fitur. Pada model berbasis pohon, disebutkan bahwa variabel failures, schoolsup, goout, dan absences memiliki kontribusi besar dalam menentukan kategori nilai matematika. Dari sudut pandang teori dan logika pendidikan, *failures* (kegagalan akademik sebelumnya) dapat dipahami sebagai indikator adanya kesenjangan konsep yang bersifat kumulatif; matematika memiliki

struktur materi yang bertingkat, sehingga kesulitan pada tahap awal cenderung menimbulkan efek lanjutan pada materi berikutnya. Variabel *absences* (ketidakhadiran) dapat ditafsirkan sebagai representasi berkurangnya waktu pembelajaran efektif dan hilangnya kesempatan memperoleh umpan balik dari guru, yang pada akhirnya berdampak pada kesiapan siswa menghadapi evaluasi. Sementara itu, *schoolsup* (dukungan sekolah) perlu ditafsirkan secara hati-hati karena bisa menunjukkan dua kemungkinan yang sama-sama masuk akal: siswa yang menerima dukungan sekolah mungkin memang berada pada kondisi berisiko (sehingga dukungan diberikan), atau dukungan sekolah berkontribusi memperbaiki performa; di sinilah ide peneliti dapat dimasukkan bahwa interpretasi variabel dukungan sebaiknya dibaca bersama variabel akademik lain (misalnya failures) agar maknanya tidak disalahpahami. Adapun *goout* (intensitas keluar bersama teman) dapat merepresentasikan aspek pengelolaan waktu dan kebiasaan sosial; ketika intensitasnya tinggi, waktu belajar dan kualitas istirahat bisa terganggu, sehingga berdampak pada performa matematika. Dengan demikian, temuan feature importance menghasilkan wawasan yang dapat ditindaklanjuti, bukan hanya angka statistik.

Berdasarkan keseluruhan evaluasi, dokumen menyatakan bahwa Random Forest dipilih sebagai model terbaik untuk tahap implementasi karena mampu memberikan keseimbangan yang relatif baik antara akurasi dan F1-score pada klasifikasi multikelas. Pemilihan ini selaras dengan tujuan EDM yang tidak sekadar mengejar akurasi total, melainkan menekankan kegunaan model untuk pengambilan keputusan praktis. Dalam konteks penerapan, hasil penelitian dapat diarahkan sebagai sistem bantu pengambilan keputusan di sekolah: dengan memanfaatkan data sosial, perilaku, dan akademik yang sudah tersedia, sekolah dapat mengelompokkan siswa ke dalam kategori risiko nilai matematika sejak awal masa pembelajaran dan menjadikannya dasar perancangan intervensi yang lebih tepat, seperti bimbingan belajar tambahan, pemantauan kehadiran, atau penguatan dukungan belajar di rumah. Sebagai gagasan pengembangan dari peneliti, implementasi ini dapat dibuat lebih operasional melalui aplikasi atau dashboard sederhana yang memungkinkan guru/staf akademik memasukkan data siswa dan memperoleh prediksi kategori nilai beserta variabel yang paling berpengaruh terhadap hasil klasifikasi. Dengan cara ini, keluaran penelitian tidak berhenti sebagai laporan akademik, tetapi menjadi instrumen praktis yang dapat memperkuat strategi deteksi dini dan meningkatkan kualitas pembelajaran matematika secara lebih terarah.

KESIMPULAN

Penelitian ini berhasil mengelompokkan tingkat nilai matematika siswa berdasarkan faktor sosial, perilaku, dan akademik menggunakan metode machine learning. Data yang digunakan berasal dari Student Performance Data Set di UCI Machine Learning Repository, dengan pendekatan CRISP-DM yang terdiri dari beberapa

tahapan, yaitu Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment. Dalam penelitian ini, tiga algoritma klasifikasi yaitu Logistic Regression, Random Forest, dan XGBoost digunakan untuk membandingkan kemampuan masing-masing model dalam mengelompokkan siswa ke dalam kategori nilai rendah, sedang, dan tinggi. Hasil evaluasi menunjukkan bahwa algoritma Random Forest memiliki performa terbaik dibandingkan algoritma lainnya, dengan nilai akurasi dan F1-score tertinggi. Hal ini menunjukkan bahwa metode berbasis pohon keputusan lebih efektif dalam menangkap hubungan yang kompleks antara faktor sosial, perilaku, dan akademik dengan hasil nilai matematika siswa. Selain itu, analisis fitur menunjukkan bahwa faktor seperti catatan akademik yang buruk, support dari sekolah, intensitas aktivitas sosial, dan tingkat ketidakhadiran siswa memiliki dampak signifikan terhadap prestasi matematika. Secara keseluruhan, penelitian ini membuktikan bahwa machine learning dapat membantu mendeteksi lebih awal siswa yang berisiko mengalami penurunan nilai matematika, serta menjadi dasar pengambilan keputusan berbasis data dalam merancang strategi intervensi pembelajaran yang lebih tepat sasaran di lingkungan pendidikan.

DAFTAR PUSTAKA

[1]. C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, e1355, 2020, doi: 10.1002/widm.1355.

[2]. P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," in *Proc. 5th Annual Future Business Technology Conf. (FUBUTEC 2008)*, Porto, Portugal, Apr. 9–11, 2008, pp. 5–12.

[3]. D. Dua and C. Graff, "UCI Machine Learning Repository," Univ. of California, Irvine, School of Information and Computer Sciences, 2019.

[4]. D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc. Ser. B*, vol. 20, no. 2, pp. 215–242, 1958.

[5]. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[6]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.