REPOSITOR, Vol. 7, No. 3, Agustus 2025, Pp. 381-392
ISSN : 2714-7975
E-ISSN : 2716-1382

Deteksi Bias dalam Model Machine Learning untuk Prediksi Kelulusan Mahasiswa Berdasarkan Aktivitas Virtual Learning Environment

Deva Putra Setya Pratama*[1], Setio Basuki[1]
Universitas Muhammadiyah Malang
devanebanderas32@webmail.umm.ac.id*

Abstrak

Revolusi digital yang cepat dalam pendidikan telah menempatkan Virtual Learning Environment (VLE) sebagai elemen penting dalam mengembangkan paradigma pembelajaran, yang sangat terlihat selama pandemi COVID-19. Penelitian ini menyelidiki dampak aktivitas siswa berbasis VLE dalam memprediksi keberhasilan akademik dan mengatasi bias dalam model machine learning yang digunakan untuk prediksi ini. Menggunakan Open University Learning Analytics Dataset (OULAD), penelitian ini mengintegrasikan teknik prapemrosesan data, pemilihan fitur, dan transformasi data untuk mengembangkan dataset yang komprehensif. Model Random Forest digunakan untuk memprediksi hasil kelulusan siswa, yang dikategorikan menjadi kelas "pass", "fail", dan "distinction". Kinerja model dievaluasi menggunakan metrik klasifikasi seperti akurasi, presisi, recall, dan F1-score, serta matriks kebingungan. Deteksi bias dilakukan menggunakan alat DALEX, dengan fokus pada atribut terlindungi seperti usia, jenis kelamin, dan disabilitas untuk memastikan keadilan. Hasilnya mengungkapkan akurasi model yang tinggi tetapi menyoroti adanya bias yang signifikan dalam beberapa kelompok demografis. Penelitian ini berkontribusi pada diskursus berkelanjutan tentang memastikan penerapan machine learning yang etis dan adil dalam pengaturan pendidikan dengan mengusulkan metode untuk meningkatkan kesetaraan dan transparansi model prediktif.

Kata Kunci: *Lingkungan Pembelajaran Virtual, Pembelajaran Mesin, Deteksi Bias, DALEX, Open University Learning Analytics Dataset, Random Forest, Penggalian Data Pendidikan*

Abstract

The rapid digital revolution in education has positioned Virtual Learning Environments (VLEs) as critical to evolving learning paradigms, particularly highlighted during the COVID-19 pandemic. This research investigates the impact of VLE-based student activity on predicting academic success and addresses the biases in machine learning models used for these predictions. Using the Open University Learning Analytics Dataset (OULAD), this study integrates data preprocessing techniques, feature selection, and data transformation to develop a comprehensive dataset. A Random Forest model is employed to predict student graduation outcomes, categorized into "pass", "fail", and "distinction" classes. The model's performance is evaluated using classification metrics such as accuracy, precision, recall, and F1-score, alongside confusion matrices. Bias detection is conducted using the DALEX tool, focusing on protected attributes like age, gender, and disability to ensure fairness. The results reveal high model accuracy but highlight significant bias in some demographic groups. This study contributes to the ongoing discourse on ensuring ethical and fair machine learning applications in educational settings by proposing methods to enhance the equity and transparency of predictive models.

Keywords: *Virtual Learning Environment, Machine Learning, Bias Detection, DALEX, Open University Learning Analytics Dataset, Random Forest, Educational Data Mining*

1. Pendahuluan

Dalam era revolusi digital di dunia pendidikan, *Virtual Learning Environment* (VLE) menjadi tiang utama evolusi pembelajaran. VLE menyajikan sebuah lingkungan pembelajaran daring yang terintegrasi, tidak sekadar sebagai platform penyedia materi, tetapi sebuah ekosistem yang memperluas paradigma tradisional pembelajaran [1]. Didefinisikan sebagai suatu konteks belajar online yang dinamis, VLE memberikan mahasiswa akses ke sumber daya pembelajaran, memfasilitasi interaksi antar sesama, dan mengubah cara pandang proses pembelajaran [2].

Penggunaan VLE semakin meningkat seiring berjalannya waktu, terutama dengan munculnya pandemi COVID-19 yang memaksa perguruan tinggi untuk mengadaptasi metode pembelajaran online [3]. Transformasi yang cepat ini menjadi respon atas kebutuhan mendesak untuk memastikan aksesibilitas dan kesinambungan pembelajaran di tengah keterbatasan fisik dan pembatasan sosial yang diakibatkan oleh pandemi [4]. Penggunaan VLE memiliki dampak yang signifikan terhadap kualitas dan hasil pembelajaran. Dampak VLE dapat bervariasi tergantung pada faktor-faktor seperti desain, motivasi, keterampilan pengguna, dan dukungan fasilitas [5]. Pada dasarnya VLE memberikan kemudahan pembelajaran, peningkatan akses sumber belajar, pengembangan pemikiran kritis, umpan balik yang cepat, dan evaluasi yang akurat [6].

Meskipun VLE memberikan banyak dampak positif, namun masih banyak siswa yang mengalami kesulitan dan bahkan mengundurkan diri atau tidak mencapai kelulusan [7]. Permasalah tersebut sangat penting untuk dikaji guna menemukan faktor yang mempengaruhi, salah satunya adalah aktivitas belajar virtual siswa. Oleh karena itu, saat ini diperlukan fokus pada peningkatan VLE serta pengembangan metode dan indikator efektif untuk mengukur kinerja dalam memprediksi kelulusan mahasiswa. Dalam mengatasi tantangan kelulusan mahasiswa berdasarkan aktivitas VLE, penggunaan teknik *data mining* dan model *machine learning* membuka potensi yang efektif [8]. Pentingnya model machine learning terlihat dalam kemampuannya untuk mengidentifikasi mahasiswa yang berisiko gagal atau putus kuliah [9]. Dalam hal ini, model machine learning bukan hanya sebuah alat prediktif, tetapi juga menjadi sarana efektif untuk meningkatkan efisiensi dalam mendukung kemajuan akademis mahasiswa.

Sejumlah penelitian terdahulu dalam dunia pendidikan telah membuktikan bahwa model machine learning telah berhasil dalam melakukan prediksi kelulusan mahasiswa. Tomasevic et al. [10] berhasil memprediksi kelulusan mahasiswa di *Open University* menggunakan enam model berbeda dengan ANN sebagai model terbaik. Waheed et al. [11], menggunakan data dan model yang sama, tetapi dengan teknik pengolahan data dan skema prediksi yang berbeda, berhasil melakukan prediksi kelulusan dengan baik berdasarkan kuartal *click stream*. Rizvi et al. [12], dengan menggunakan dataset yang sama, membagi prediksi kelulusan berdasarkan jenis *asessment* dan kode modul, dan berhasil memberikan akurasi yang baik dengan model *decision tree*. Selain itu Al-Zawqari et al [13] berhasil memprediksi hasil akhir mahasiswa melalui pendekatan seleksi fitur dengan sukses menggunakan model *random forest* pada dataset yang sama. Berdasarkan penelitian terdahulu, dapat disimpulkan bahwa penggunaan *machine learning* dalam prediksi kelulusan mahasiswa telah menghasilkan hasil yang signifikan. Meskipun model *machine learning* menawarkan berbagai keuntungan intervensi dan solusi yang tepat waktu. Namun, dalam model *machine learning* terdapat bias yang menyebabkan ketidaksesuaian atau ketimpangan antara hasil prediksi dengan kenyataan yang dapat mendiskriminasi kelompok tertentu [14]. Kehadiran bias dalam model *machine learning* menjadi isu sentral yang perlu mendapat perhatian serius. Bias tersebut terkait erat dengan prinsip-prinsip *fairness* (keadilan) dan transparansi, yang merupakan aspek etis dalam pengembangan model *machine learning* [15]. Fairness mengacu pada kemampuan model memberikan hasil yang adil dan merata di antara kelompok yang berbeda.

Dalam konteks ini, deteksi bias memiliki peran khusus dalam melindungi atribut yang dilindungi, seperti jenis kelamin, ras, umur, dan lain sebagainya [16]. Deteksi bias yang efektif tidak hanya menjamin bahwa model memberikan hasil prediksi yang akurat secara keseluruhan, tetapi juga melibatkan perlindungan terhadap kelompok-kelompok yang rentan terhadap diskriminasi atau ketidaksetaraan [17]. Dalam prediksi kelulusan mahasiswa, deteksi bias dapat memastikan bahwa model tidak memberikan preferensi atau diskriminasi berdasarkan jenis kelamin, ras, atau faktor lain yang dapat menjadi atribut yang dilindungi. Oleh karena itu, deteksi bias tidak hanya berfungsi sebagai alat untuk meningkatkan kualitas prediksi, tetapi juga sebagai garda terdepan dalam memastikan bahwa model *machine learning* beroperasi secara adil dan tidak memberikan perlakuan yang tidak setara terhadap atribut yang dilindungi.

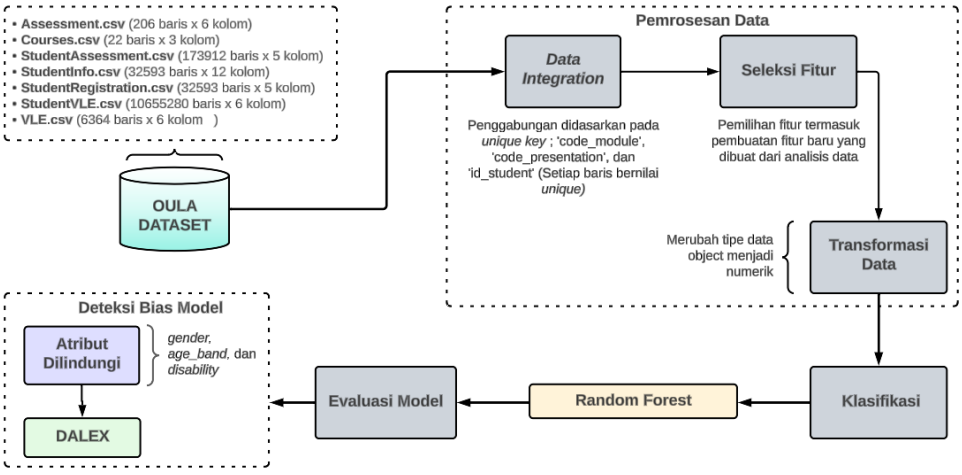
Sejumlah penelitian terdahulu, seperti yang dilakukan oleh Gardner et al. [18] dengan menggunakan teknik *Absolute Between ROC Area* (ABROCA), telah memberikan wawasan berharga dalam deteksi bias pada model *machine learning*. Penelitian ini tidak hanya fokus pada kinerja keseluruhan model, tetapi juga melakukan analisis *slicing* untuk menyoroti perbedaan kinerja berdasarkan kelompok demografis, identitas, atau kelompok lainnya. Riazy et al. [19], dengan pendekatan serupa berhasil mengeksplorasi kemungkinan diskriminasi dalam analitika pembelajaran pada data OULAD. Penelitain menemukan adanya bias terhadap kelompok-kelompok tertentu seperti gender dan disabilitas yang dideklarasikan di dalam model. Sementara

itu, pada dataset yang sama Verge et al. [20] mengembangkan metrik keadilan *Model Absolute Density Distance* (MADD) untuk menganalisis perilaku diskriminatif dari model pembelajaran mesin. Penelitian menemukan bahwa kinerja prediktif yang adil tidak menjamin perilaku model yang adil. Selain itu, masih pada dataset yang sama Riazy et al. [21] menggunakan metode *Fairness Metrics Analytics* untuk mengevaluasi keadilan model prediksi. Mereka menemukan perbedaan signifikan dalam hasil prediksi antara subkelompok mahasiswa, menyoroti ketidaksetaraan prediksi untuk subkelompok tertentu. Meskipun penelitian-penelitian sebelumnya telah berhasil mendeteksi bias dengan baik menggunakan teknik deteksi yang berbeda-beda. Analisis terperinci antar kelompok dalam atribut yang dilindungi belum sepenuhnya dilakukan dalam penelitian sebelumnya. Selain itu, penggunaan kombinasi atribut yang dilindungi dalam analisis diskriminasi antar kelompok belum diimplementasikan sepenuhnya. Perlu dicatat bahwa, kombinasi atribut yang dilindungi dapat mempengaruhi sensitifitas diskriminasi pada model.

Berdasarkan pemaparan latar belakang dan penelitian sebelumnya, seperti masih adanya siswa yang kesulitan dalam pembelajaran berbasis VLE dan mengakibatkan kegagalan akademik. Penelitaan ini bertujuan untuk mendeteksi bias dalam model *machine learning* yang digunakan dalam memprediksi kelulusan mahasiswa berdasarkan aktivitas VLE. Penelitian ini menerapkan teknik deteksi bias menggunakan moDel Agnostic Language for Exploration and eXplanation (DALEX) [22] dengan dataset dari Open University Learning Analytics Dataset (OULAD) [23], yang telah digunakan dalam penelitian sebelumnya. Pendekatan DALEX dipilih karena menyediakan antarmuka yang agnostik terhadap model, memungkinkan penjelasan interaktif, dan mendukung keadilan antar kelompok [24]. Melalui pemanfaatan DALEX dan dataset OULAD, harapannya adalah memberikan kontribusi yang signifikan dalam peningkatan kualitas dan keadilan model machine learning di konteks pendidikan. Analisis holistik dan visualisasi mendalam yang dilibatkan dalam penelitian ini diharapkan dapat membuka wawasan baru dan menjadi landasan bagi pengembangan model yang lebih adil dan transparan, khususnya dalam konteks pembelajaran virtual.

2. Metode Penelitian

Penelitian ini menggunakan algoritma random forest untuk memprediksi kelulusan mahasiswa dan menggunakan model DALEX untuk mendeteksi bias pada model prediksi. Rancangan skema eksperimen pada Gambar 1 yang akan diterapkan untuk mendapatkan hasil berupa model evaluasi dan deteksi bias pada model prediksi kelulusan mahasiswa.

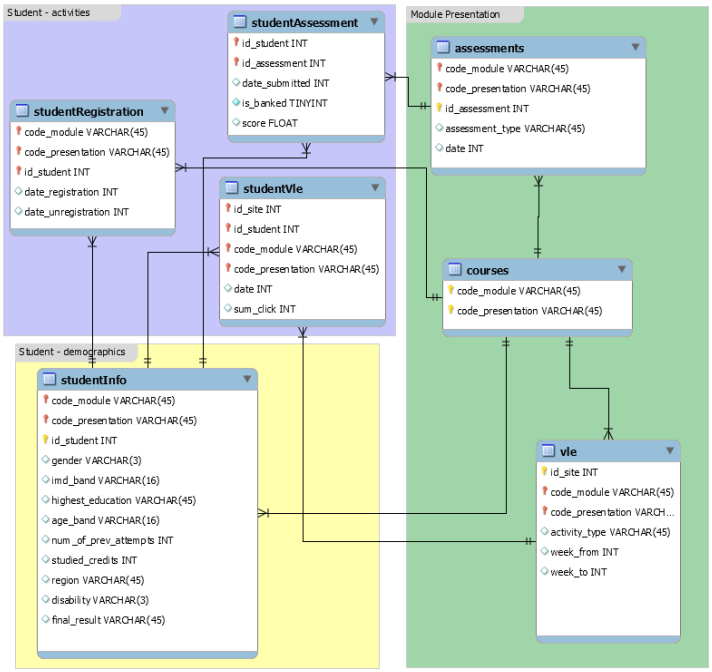


Gambar 1. Alur Penelitian

2.1 *Dataset*

Dalam penelitian ini, data yang digunakan berasal dari *Open University Learning Analytics Dataset* (OULAD) yang merupakan dataset dari hasil pengumpulan lingkungan pembelajaran daring *Open University* [23]. OULAD terdiri dari 7 file yang terpisah sesuai dengan skema database yang digunakan oleh Open University. OULAD berisi informasi lengkap aktivitas belajar, penilaian, dan atribut demografis dari 32.593 siswa dengan label hasil akhir sebagai "*fail*", "*withdrawn*", "*pass*", atau "*distinction*". Dataset ini dikembangkan untuk mendukung analisis dan

penelitian di bidang pembelajaran online, dan menjadi sumber data kritis dalam konteks penelitian prediksi kelulusan mahasiswa berbasis aktivitas pada *Virtual Learning Environment* (VLE). Skema database OULAD ditunjukkan pada Gambar 2.



Gambar 2. Skema Database OULAD

2.2 *Data Preprocessing*

Dalam setiap penelitian, proses preprocessing dataset menjadi tahapan yang sangat krusial untuk mencegah hasil yang kurang optimal dari model yang dikembangkan. Ada berbagai metode yang dapat diterapkan untuk memastikan bahwa dataset yang digunakan memiliki kualitas yang memadai, termasuk pengecekan keberadaan nilai yang hilang (*missing value*), deteksi duplikasi data, analisis setiap fitur, dan sejumlah pertimbangan lainnya. Oleh karena itu, pada penelitian ini terdapat beberapa langkah yang harus dilakukan, meliputi integrasi data, seleksi fitur, dan transformasi data.

Tahap pertama adalah integrasi data, di mana data OULAD digabungkan berdasarkan kriteria tertentu untuk membentuk satu set data yang lengkap dan terpadu. Integrasi data ini esensial untuk menciptakan suatu entitas data yang mencakup informasi komprehensif dan relevan dari berbagai sumber, memungkinkan analisis holistik terhadap pola aktivitas mahasiswa di VLE. Setelah integrasi data, langkah selanjutnya adalah melakukan seleksi fitur untuk memahami kontribusi setiap atribut terhadap prediksi kelulusan mahasiswa. Seleksi fitur bertujuan mengidentifikasi atribut yang signifikan, mengurangi dimensi data, dan meningkatkan akurasi model tanpa *overfitting*. Analisis atribut juga membuka peluang penambahan fitur baru berdasarkan korelasi, pola, atau informasi tambahan yang dapat memperkaya model *machine learning*, dengan tujuan meningkatkan kemampuan model dalam memprediksi kelulusan mahasiswa dengan lebih akurat.

Tahap akhir dari pemrosesan data penelitian ini adalah transformasi data. Transformasi dilakukan untuk menyesuaikan format dan distribusi data, memastikan keseragaman, dan mengoptimalkan performa model. Proses transformasi ini mencakup pengkodean variabel kategorikal dan langkah-langkah lain yang diperlukan untuk menghasilkan data yang siap digunakan oleh model machine learning. Selain ketiga tahapan tersebut, pengecekan nilai hilang dan duplikasi data juga harus dilakukan untuk memaksimalkan pemrosesan data. Berikut ini adalah tabel dari setiap fitur OULAD beserta detail pemrosesan data yang akan dilakukan.

Berikut ini merupakan fitur-fitur dari pemrosesan data yang telah dilakukan sesuai dengan analisis dan kebutuhan model pada Tabel 1.

Tabel 1. Fitur-Fitur dari Tahap Pemrosesan Data

Fitur	Nama Fitur	Deskripsi
1	code_module	Kode unik setiap modul
2	code_presentation	Kode unik setiap semester
3	gender	Jenis kelamin siswa
4	region	Daerah asal siswa
5	highest_education	Pendidikan tertinggi terakhir siswa
6	imd_band	Indeks kemiskinan wilayah siswa
7	age_band	Rentang umur siswa
8	num_of_prev_attempts	Kehadiran siswa pada setiap modul
9	studied_credits	Jumlah sks siswa yang ditempuh
10	disability	Status disabilitas siswa
11	final_result	Hasil akhir dari setiap siswa
12	all_activity_sumclick	Jumlah klik keseluruhan setiap siswa pada vle
13	after_clicks_sum	Jumlah klik keseluruhan setiap siswa pada vle setelah kursus dimulai
14	before_clicks_sum	Jumlah klik keseluruhan setiap siswa pada vle sebelum kursus dimulai
15-34	{activity_type}_sum	Jumlah klik keseluruhan setiap siswa pada masing-masing tipe aktivitas vle
35	date_registration	Tanggal pendaftaran kursus siswa
36	module_presentation_lenght	Panjang modul pembelajaran
37	CMA_sumScore	Jumlah total skor CMA setiap siswa
38	TMA_sumScore	Jumlah total skor TMA setiap siswa
39	Exam_sumScore	Jumlah total skor EXAM setiap siswa

2.3 *Random Forest*

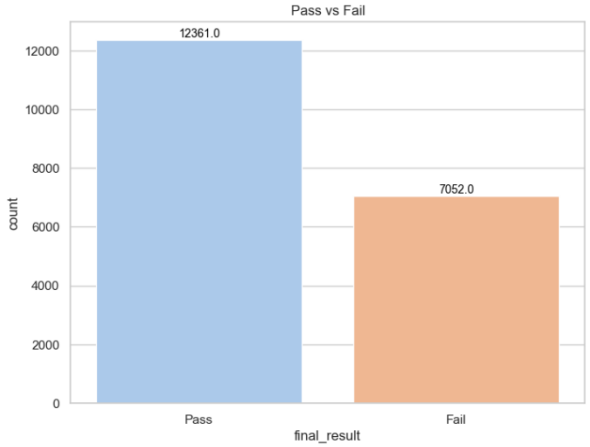
Penelitian ini menggunakan model *random forest*, sebuah metode ensemble yang menggabungkan sejumlah *decision tree* untuk meningkatkan akurasi prediksi [13]. Dalam *random forest*, teknik *bootstrapping* digunakan untuk membangun model dengan setiap pohon yang dibangun dari subset data yang dipilih secara acak. Untuk pengambilan keputusan, hasil prediksi dari berbagai pohon digabungkan untuk menghasilkan prediksi akhir yang lebih stabil dan akurat. Pendekatan ini membantu *random forest* mengatasi *overfitting* dan meningkatkan generalisasi pada dataset yang belum pernah dilihat sebelumnya. Karena fleksibilitasnya dan kemampuannya untuk menangani berbagai jenis data, Random Forest sering digunakan dalam berbagai aplikasi *machine learning* termasuk klasifikasi, regresi, dan pengelompokan data. Dalam penelitian ini, prediksi dilakukan dengan parameter *default* dari model *random forest* karena fokus utama adalah mendeteksi bias dalam model Machine Learning yang digunakan. Model ini dilatih dengan data yang dibagi menjadi 80% data pelatihan dan 20% data pengujian. Parameter model yang digunakan dalam penelitian ini ditunjukkan pada Tabel 2.

Tabel 1. Parameter Model Random Forest

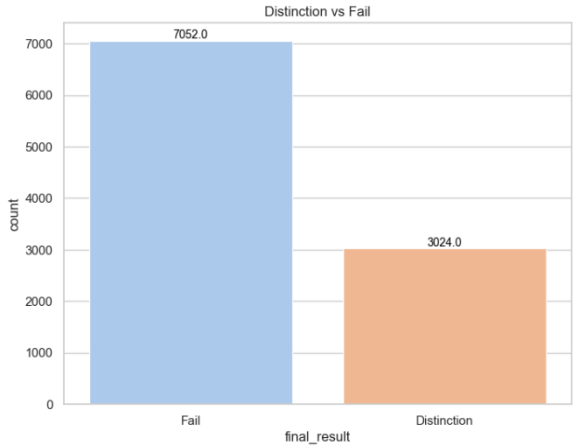
Parameter	Nilai
n_estimators	100
n_jobs	None
oob_score	False
bootstrap	True
verbose	0
warm_start	False
monotonic_cst	None
criterion	gini
class_weight	None
min_weight_fraction_leaf	0
min_samples_split	2
min_samples_leaf	1
min_impurity_decrease	0

max_samples	None
max_leaf_nodes	None
max_features	sqrt
max_depth	None
random_state	280

Skema klasifikasi yang diterapkan dalam penelitian ini adalah *binary classification*. Penelitian ini mengkategorikan hasil prediksi menjadi "*pass*", "*fail*", dan "*distinction*" untuk memberikan identifikasi yang lebih terperinci terhadap performa siswa. Kategori "withdrawn" tidak digunakan karena status "*withdrawn*" menunjukkan bahwa mahasiswa secara sukarela menarik diri dari kursus atau program, kemungkinan disebabkan oleh faktor non-akademis seperti masalah pribadi atau kehidupan, bukan karena kinerja akademis yang buruk. Penelitian ini menggabungkan kelas "*pass*" dengan "*fail*" dan juga kelas "*distinction*" dengan "*fail*". Strategi ini tidak hanya meningkatkan ketajaman penilaian hasil, tetapi juga memberikan wawasan baru dalam evaluasi kinerja siswa. Gambar 3 dan Gambar 4 menunjukkan distribusi data dari setiap skema klasifikasi yang digunakan dalam penelitian ini.



Gambar 3. Distribusi Data Skema Pass vs Fail



Gambar 4. Distribusi Data Skema Distinction vs Fail

2.4 Evaluasi

Penelitian ini menggunakan *classification report* dengan metrik *accuracy, precision, recall*, dan *F1-score* untuk mengevaluasi kinerja model [25]. *Classification report* memberikan gambaran komprehensif tentang ketepatan dan cakupan model dalam mengklasifikasikan data, memungkinkan peneliti menilai performa secara menyeluruh dan mengidentifikasi trade-off antara precision dan recall. Berikut adalah rumus untuk metrik *accuracy*, *precision, recall*, dan *F1-score* pada Persamaan 1, Persamaan 2, Persamaan 3, dan Persamaan 4.

$$accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (1)$$

$$precision = \frac{TP}{(TP + FP)} \quad (2)$$

$$recall = \frac{TP}{(TP + FN)} \quad (3)$$

$$F1 - score = 2 \times \frac{precision \times recall}{(precision + recall)} \quad (4)$$

Selain itu, penelitian ini juga menggunakan *confusion matrix* untuk menganalisis secara lebih rinci distribusi prediksi yang benar dan salah oleh model [26]. Dengan menggunakan confusion matrix, peneliti dapat memperoleh wawasan tambahan tentang seberapa baik model dalam mengklasifikasikan setiap kelas.

2.5 Deteksi Bias

Dalam penelitian ini, deteksi bias dilakukan menggunakan fungsi fairness pada model explainer DALEX. DALEX, singkatan dari *Model Agnostic Language for Exploration and Explanation* memungkinkan peneliti menjelajahi dan menjelaskan kinerja model secara mendalam, menyediakan visualisasi yang kaya, dan mengidentifikasi potensi bias dalam prediksi [25]. DALEX unggul dalam mengintegrasikan berbagai jenis model machine learning dan menyajikan interpretasi yang terstruktur, yang bermanfaat untuk mendeteksi dan mengatasi bias dalam prediksi kelulusan mahasiswa berdasarkan aktivitas VLE. Namun, library DALEX hanya mendukung *binary classification* dalam model deteksi biasnya.

Dalam deteksi bias, DALEX mengadopsi aturan empat perlima sebagai standar untuk menilai tingkat diskriminasi [22]. Batas ini diimplementasikan dengan ε default sebesar 0,8, namun dapat disesuaikan sesuai kebutuhan. Rasio antara skor metrik yang mendekati angka 1 menunjukkan tingkat keadilan yang tinggi dalam model. Skor dari setiap metrik didefinisikan dengan rumus pada Persamaan 5.

$$\forall_{i \in \{a,b,\ldots,z\}^{\varepsilon}} < \frac{metric_i}{metric_{privileged}} < \frac{1}{\varepsilon}. \quad (5)$$

Berdasarkan penjelasan sebelumnya, inisialisasi epsilon pada model DALEX akan diisi dengan nilai *default*, yaitu 0,8, sesuai dengan aturan empat perlima yang digunakan sebagai standar untuk mengukur tingkat diskriminasi. Selain itu, atribut terlindungi (*protected attributes*) yang digunakan adalah atribut yang dikombinasikan. Penggunaan atribut terlindungi yang dikombinasikan melibatkan penggabungan atribut tertentu untuk memberikan informasi lebih rinci tentang keragaman karakteristik yang diidentifikasi. Kombinasi atribut terlindungi dilakukan pada atribut "age_band" dengan "gender" dan atribut "age_band" dengan "disability".

3. Hasil Penelitian dan Pembahasan

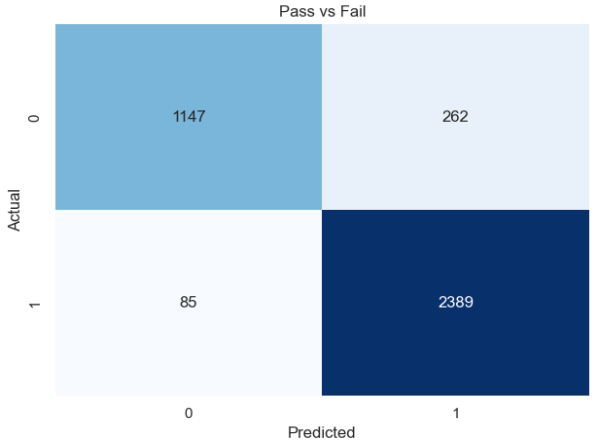
Pada bagian ini dipaparkan skenario eksperimen untuk hasil klasifikasi dan deteksi bias pada dataset. Proses analisis hasil klasifikasi berbasis *machine learning* dimulai dengan melakukan pemrosesan data melalui beberapa tahap, termasuk integrasi data, seleksi fitur, dan transformasi data. Selanjutnya, model machine learning dibangun menggunakan algoritma *random forest*. Setelah pembentukan model, evaluasi dilakukan menggunakan *classification report* dan *confusion matrix*. Tahap terakhir analisis ini mencakup deteksi bias pada model yang telah dikembangkan. Pada tahap ini, setiap skema klasifikasi dan kombinasi atribut yang dilindungi dianalisis untuk mengidentifikasi serta mengatasi potensi bias yang mungkin muncul. Dengan demikian, bagian ini memberikan gambaran menyeluruh tentang proses analisis hasil klasifikasi hingga deteksi bias pada model yang dikembangkan.

3.1 Evaluasi Hasil Pengujian

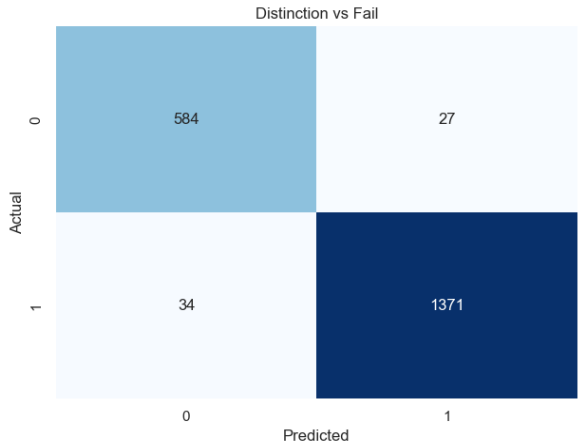
Dalam penelitian ini, evaluasi dilakukan menggunakan *classification report* dan *confusion matrix* dari setiap skema klasifikasi yang diterapkan. Berikut adalah hasil evaluasi model dari setiap skema klasifikasi yang ditunjukkan pada Tabel 3, Gambar 5, dan Gambar 6.

Tabel 3. Evaluasi Model Setiap Skema Klasifikasi

Skema	Precision	Recall	F1-Score	Accuracy
Pass vs Fail	0.96	0.90	0.93	0.91
Distinction vs Fail	0.97	0.97	0.97	0.96



Gambar 5. Confusion Matrix Skema Pass vs Fail



Gambar 6. Confusion Matrix Skema Distinction vs Fail

Berdasarkan hasil tersebut, evaluasi model dari kedua skema klasifikasi menunjukkan kinerja yang sangat baik, terutama pada skema "*distinction vs fail*". Pada skema klasifikasi ini, model mencapai presisi, recall, dan F1-Score yang tinggi, mencapai 97% untuk presisi dan F1-Score, serta 96% untuk recall, dengan tingkat akurasi sebesar 96%. Sementara itu, pada skema klasifikasi "*pass vs fail*", model menunjukkan presisi sebesar 96%, recall sebesar 90%, dan F1-Score sebesar 93%, dengan tingkat akurasi mencapai 91%. Analisis dari hasil evaluasi ini mengindikasikan bahwa model mampu secara efektif membedakan antara kelas "*pass*" dan "*fail*" serta antara "*distinction*" dan "*fail*". Tingkat presisi dan recall yang tinggi menunjukkan bahwa model memberikan prediksi yang akurat untuk kedua skema klasifikasi tersebut. Akurasi yang tinggi juga menunjukkan bahwa model secara keseluruhan berhasil dalam memprediksi kelulusan mahasiswa dengan sangat baik. Namun demikian, untuk penelitian lebih lanjut, penting untuk menguji kinerja model pada data dengan jumlah kelas yang seimbang. Ketidakseimbangan kelas pada data dapat memengaruhi keakuratan akurasi model, di mana model cenderung

memprediksi kelas mayoritas lebih sering. Oleh karena itu, penting untuk menguji model pada data uji yang seimbang untuk memahami kemampuan model dalam membedakan kelas minoritas secara menyeluruh.

3.2 Hasil Deteksi Bias

Tahapan terakhir dalam penelitian ini adalah mendeteksi kemungkinan adanya bias dalam model prediksi kelulusan mahasiswa. Deteksi bias dilakukan menggunakan model DALEX dengan keluaran berupa metriks sebagai indikator nilai bias. Keluaran metriks tersebut dikatakan bias jika rentang nilainya diluar nilai epsilon yang telah ditetapkan sebelumnya. Jika nilai epsilon yang ditetapkan adalah 8, maka setiap skema klasifikasi harus memiliki nilai metriks dalam rentang 0,8-1,25 agar model dapat dikatakan adil. Dari kedua skema klasifikasi yang telah dilakukan sebelumnya, masing-masing skema memiliki 2 tabel deteksi bias dengan *protected attribute* yang berbeda. Kedua tabel tersebut adalah kombinasi *protected attribute* antara fitur umur dengan jenis kelamin dan fitur disabilitas dengan jenis kelamin. Masing-masing tabel memiliki *priviledge* yang diacukan pada jenis kelamin, yaitu "**male_old**" dan "**male_normal**". Berikut ini adalah hasil deteksi bias dari setiap skema klasifikasi dan kombinasi protected attribute yang ditunjukkan pada Tabel 4:

Tabel 4. Deteksi Bias Setiap Skema Klasifikasi

Skema Klasifikasi	Kombinasi Atribut	Atribut Dilindungi	TPR	ACC	PPV	FPR	STP
Pass vs Fail	gender + age_band	female_old	1.00	0.99	0.99	1.48	1.14
		female_young	0.97	0.97	0.98	1.24	1.03
		male_old	1.01	1.01	1.03	1.06	1.12
	gender + disability	female_disability	1.00	0.96	0.93	1.40	0.98
		female_normal	0.98	0.97	0.98	1.25	1.02
		male_disability	0.99	0.99	0.98	0.82	0.87
Distinction vs Fail	gender + age_band	female_old	0.94	0.96	0.98	1.02	0.76
		female_young	0.97	0.97	0.98	1.90	0.94
		male_old	1.01	1.01	1.01	0.20	0.85
	gender + disability	female_disability	0.96	0.98	1.01	NaN	0.92
		female_normal	0.96	0.96	0.97	2.50	0.93
		male_disability	1.01	1.00	0.99	1.86	1.12

Skema Pass vs Fail (gender dan age_band) tidak bisa dikatakan adil karena memiliki bias pada salah satu metriksnya. Hasil menunjukkan bahwa untuk subgrup female_old, female_young, dan male_old pada metriks evaluasi TPR, ACC, PPV, dan STP berada dalam rentang yang dianggap wajar (0,8-1,25). Namun, terdapat kecenderungan false positive (FPR) yang lebih tinggi dari batas yang diperbolehkan pada subgrup female_old.

Skema Pass vs Fail (gender dan disability) juga tidak bisa dikatan adil karena memiliki bias pada salah satu metriksnya. Hasil menunjukkan bahwa setiap supgrub pada metriks evaluasi TPR, ACC, PPV, dan STP berada dalam rentang yang dianggap wajar. Namun, terdapat kecenderungan false positive yang lebih tinggi dari batas yang diperbolehkan pada subgrup female_disability dan female_normal.

Skema Distinction vs Fail (gender dan age_band) tidak bisa dikatakan adil karena memiliki bias pada kedua metriksnya. Hasil menunjukkan bahwa setiap subgrub pada metriks evaluasi TPR, ACC, dan PPV berada dalam rentang yang dianggap wajar. Namun, terdapat kecenderungan nilai FPR yang lebih rendah dan lebih tinggi dari batas yang diperbolehkan. Selain itu, hasil pada metriks STP dengan subgrub female_old menunjukkan kecenderungan nilai yang lebih rendah dari batas yang diperbolehkan.

Skema Distinction vs Fail (gender dan disability) juga tidak bisa dikatakan adil karena memiliki bias pada salah satu metriksnya. Hasil menunjukkan bahwa setiap subgrub pada metriks evaluasi TPR, ACC, PPV, dan STP berada dalam rentang yang dianggap wajar. Namun, terdapat kecenderungan false positive yang lebih tinggi dari batas yang diperbolehkan pada subgrub female_normal dan male_disability. Selain itu, pada subgrub female_disability menunjukkan hasil *Not a Number* (NaN) pada metriks FPRnya.

Berdasarkan hasil analisa diatas, seluruh skema klasifikasi dengan kombinasi atributnya masih belum bisa dikatakan adil. Hal ini dikarenakan pada setiap deteksi bias yang dilakukan masih terdapat bias dalam beberapa metriksnya. Selain itu, peneliti menemukan adanya NaN pada salah satu metriks evaluasi. Munculnya nilai NaN terjadi ketika tidak ada kasus yang diklasifikasikan sebagai negatif sebenarnya (True Negative + False Positive = 0) dalam confusion matrics suatu subgrup, sehingga pembaginya menjadi nol. Hal tersebut bisa terjadi dikarenakan kurangnya variasi dalam data, kesalahan implementasi, atau ketidakcocokan antara algoritma perhitungan dengan karakteristik data tertentu

4. Kesimpulan

Meskipun model *machine learning* yang dibangun untuk memprediksi kelulusan mahasiswa berdasarkan aktivitas virtual learning environment menunjukkan performa yang sangat baik. Hasil analisis menemukan adanya bias dalam hasil evaluasi model yang digunakan. Walaupun metrik-metrik tersebut menunjukkan performa yang cukup baik dengan sebagian besar nilai berada dalam rentang yang dianggap wajar. Hasil deteksi bias menunjukkan bahwa ada ketidakseimbangan dalam prediksi model terhadap beberapa subgrub. Terdapat kecenderungan untuk mengklasifikasikan beberapa kelompok tertentu yang melebihi batas yang diperbolehkan. Selain itu, terdapat juga hasil NaN pada beberapa metrik yang menunjukkan bahwa ada subgrub yang tidak mendapat prediksi false positive dan true negative sama sekali.

Berdasarkan penelitian ini, terdapat beberapa peningkatan yang dapat dilakukan pada penelitian berikutnya. Untuk menghasilkan penelitian yang lebih inklusif dan relevan, penelitian berikutnya dapat memperdalam analisis terhadap ketidakseimbangan dalam prediksi model terhadap berbagai subkelompok mahasiswa. Selain itu, langkah-langkah baru dalam pengujian keadilan model dapat diadopsi guna memperbaiki kinerja dan keadilan model. Diperlukan juga pengumpulan data yang lebih representatif dari berbagai subkelompok serta penerapan strategi mitigasi bias yang sesuai. Langkah-langkah ini penting untuk memastikan model dapat memberikan prediksi yang adil dan akurat untuk seluruh populasi mahasiswa.

Daftar Notasi

TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative
ε : Nilai epsilon yang digunakan
$metric_i$: Metriks ke-i

Referensi

[1] A. H. A Rashid, N. A. Shukor, Z. Tasir, and S. N. Kew, "Teachers' perceptions and readiness toward the implementation of virtual learning environment," IJERE (*International Journal of Evaluation and Research in Education*), vol. 10, no. 1, pp. 209–214, 2021, doi: 10.11591/ijere.v10i1.21014.

[2] R. Yusny and G. I. Yasa, "Mengembangkan (Pembelajaran) blended learning dengan sistem lingkungan pembelajaran virtual (Vle) Di ptkin," Jurnal Ilmiah Islam Futura, vol. 19 no. 1, pp. 103–127, 2019, doi: 10.22373/jiif.v19i1.3707.

[3] Y. A. Hambali, R. R. Putra, and Wahyudin, "Implementasi Metaverse menggunakan Aplikasi gather town untuk Pendidikan Jarak Jauh dengan Pendekatan Virtual Learning Environment," Information System For Educators and Professionals: Journal of Information System, vol. 7, no.2, pp. 163–172, 2023, doi: 10.51211/isbi.v7i2.2039.

[4] F. N. Andryas, A. Irmarahayu, and N. Bustamam, "Virtual learning environment and learning approach among pre-clinical undergraduate medical students during COVID-19 pandemic," The Indonesian Journal of Medical Education, vol. 11, no. 1, pp. 10–21, 2022, doi: 10.22146/jpki.63975.

[5] P. A. Petare, M. Shamim, T. Gupta, R. Verma, and G. Singh, "Exploring the impact of virtual learning environments on student engagement and academic achievement," Journal of Survey in Fisheries Sciences, vol. 10, no. 1S, pp. 5912–5923, 2023, doi: 10.13140/RG.2.2.23223.91040.

[6] B. Mandasari, "The impact of online learning toward students' academic performance on business correspondence course," EDUTEC (*Journal of Education and Technology*), vol. 4, no. 1, pp. 98–110, 2020, doi: 10.29062/edu.v4i1.74.
[7] P. Muljana and T. Luo, "Factors contributing to student retention in online learning and recommended strategies for improvement: A systematic literature review," Journal of Information Technology Education: Research, vol. 18, pp. 19–57, 2019, doi: 10.28945/4182.
[8] A. Al-Azawei and M. A. Al-Masoudy, "Predicting learners' performance in virtual learning environment (VLE) based on demographic, behavioral and engagement antecedents," International Journal of Emerging Technologies in Learning (iJET), vol. 15, no. 09, pp. 60–75, 2020, doi: 10.3991/ijet.v15i09.12691.
[9] M. S. Ahmad, A. H. Asad, and A. Mohammed, "A machine learning based approach for student performance evaluation in educational data mining," in 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), 2021, pp. 187–192, doi: 10.1109/miucc52538.2021.9447602.
[10] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," Computers & Education, vol. 143, no. 103676, pp. 1–18, 2020, doi: 10.1016/j.compedu.2019.103676.
[11] H. Waheed, S. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," Computers in Human Behavior, vol. 104, no. 106189, 2020, doi: 10.1016/j.chb.2019.106189.
[12] S. Rizvi, B. Rienties, and S. A. Khoja, "The role of demographics in online learning; A decision tree based approach," Computers & Education, vol. 137, pp. 32–47, 2019, doi: 10.1016/j.compedu.2019.04.001.
[13] A. Al-Zawqari, D. Peumans, and G. Vandersteen, "A flexible feature selection approach for predicting students' academic performance in online courses," Computers and Education: Artificial Intelligence, vol. 3, no. 100103, 2022, doi: 10.1016/j.caeai.2022.100103.
[14] R. S. Baker and A. Hawn, "Algorithmic bias in education," International Journal of ArtificialIntelligence in Education, vol. 32, no. 4, pp. 1052–1092, 2021, doi: 10.35542/osf.io/pbmvz.
[15] R. Fu, Y. Huang, and P. V. Singh, "AI and algorithmic bias: Source, detection, mitigation and implications," SSRN Electronic Journal, vol. 65, pp. 39–63, 2020, doi: 10.2139/ssrn.3681517.
[16] K. Cachel, E. Rundensteiner, and L. Harrison, "MANI-rank: Multiple attribute and intersectional group fairness for consensus ranking," in 2022 IEEE 38th International Conference on Data Engineering (ICDE), 2022, pp. 1124–1137, doi: 10.1109/icde53745.2022.00089.
[17] S. Alelyani, "Detection and evaluation of machine learning bias," Applied Sciences, vol. 11, no. 14, pp. 6271, 2021, doi: 10.3390/app11146271.
[18] J. Gardner, C. Brooks, and R. Baker, "Evaluating the fairness of predictive student models through slicing analysis," in Proceedings of the 9th International Conference on Learning Analytics & Knowledge, 2019, pp. 225–234, doi: 10.1145/3303772.3303791.
[19] S. Riazy, K. Simbeck, and V. Schreck, "Fairness in learning analytics: Student at-risk prediction in virtual learning environments," in Proceedings of the 12th International Conference on Computer Supported Education, 2020, pp. 15–25, doi: 10.5220/0009324100150025.
[20] M. Verger, S. Lalle, F. Bouchet, and V. Luengo, "Is your model 'MADD'? A novel metric to evaluate algorithmic fairness for predictive student models," in Proceedings of the 16th International Conference on Educational Data Mining (EDM 2023), 2023, doi: 10.5281/zenodo.8115786.
[21] S. Riazy and K. Simbeck, "Predictive algorithms in learning analytics and their fairness," Lecture Notes in Informatics (LNI) - Proceedings, 2019, pp. 223–228, doi: 10.18420/delfi2019_305.
[22] J. Wiśniewski and P. Biecek, "Fairmodels: A flexible tool for bias detection, visualization, and mitigation in binary classification models," The R Journal, vol. 14, no. 1, pp. 227–243, 2022, doi: 10.32614/rj-2022-019.
[23] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," Sci. Data, vol. 4, No. 170171, pp. 1–8 , 2017, doi: 10.1038/sdata.2017.171.

[24] H. Baniecki, W. Kretowicz, P. Piatyszek, J. Wisniewski, and P. Biecek, "dalex: Responsible Machine Learning with Interactive," Journal of Machine Learning Research, vol. 22, pp. 1–7, 2021, doi: 10.48550/arXiv.2012.14406.
[25] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," IEEE Access, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.
[26] S. Alija, E. Beqiri, A. S. Gaafar, and A. K. Hamoud, "Predicting students performance using supervised machine learning based on imbalanced dataset and wrapper feature selection," Informatica, vol. 47, no. 1, pp. 11–19, 2023, doi: 10.31449/inf.v47i1.4519.