

Explainable Machine Learning Berbasis Simulasi Respons Untuk Analisis Kualitas Soal

Hidayattullah^{*1}, Abdul Madjid Hasanuddin², Murniati³, Frankgling Nusa⁴

^{1,2,3,4} Program Studi Magister Teknik Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta

e-mail: ^{*1}hidayattullah@students.amikom.ac.id, ²a.madjid@students.amikom.ac.id,

³murniati@students.amikom.ac.id, ⁴f.nusa@students.amikom.ac.id

^{*}Correspondence author

Abstrak

Analisis kualitas soal pilihan ganda merupakan tahap penting dalam penjaminan mutu asesmen pendidikan, tetapi analisis psikometrik konvensional bergantung pada data respons peserta yang sering belum tersedia pada tahap pengembangan butir. Penelitian ini mengembangkan kerangka explainable machine learning berbasis simulasi respons untuk menganalisis kualitas soal pilihan ganda menggunakan 10.962 butir CommonsenseQA. Dataset tersebut dipilih sebagai testbed komputasional karena memiliki format lima opsi yang konsisten, kunci jawaban pada split training dan validation, skala data besar, serta pengecoh yang dirancang untuk membedakan konsep commonsense; dataset tidak diposisikan sebagai pengganti bank soal pendidikan yang tervalidasi. Respons 200 responden virtual per butir disimulasikan secara probabilistik dengan pembobotan kunci jawaban dan plausibilitas permukaan pengecoh. Empat indikator psikometrik—indeks kesukaran, indeks daya beda, efektivitas pengecoh, dan entropi pengecoh—dikombinasikan dengan lima fitur linguistik permukaan untuk melatih XGBoost dalam klasifikasi kualitas butir menjadi kurang, sedang, dan baik. Validasi silang lima-lipat menghasilkan F1-score terbobot $0,9987 \pm 0,0011$ dan akurasi $0,9987 \pm 0,0011$. Feature importance dan SHAP menempatkan entropi pengecoh dan indeks daya beda sebagai prediktor paling dominan, sedangkan fitur linguistik permukaan berkontribusi minimal. Performa mendekati sempurna tidak ditafsirkan sebagai validitas prediktif eksternal karena target dan prediktor utama diturunkan dari mekanisme simulasi yang sama. Kerangka ini merupakan proof-of-concept transparan dan dapat diaudit untuk evaluasi awal bank soal sebelum uji coba empiris.

Kata kunci—explainable machine learning, XGBoost, SHAP, analisis butir soal, efektivitas pengecoh

1. PENDAHULUAN

Soal pilihan ganda masih menjadi format asesmen yang dominan karena efisien untuk mengukur capaian belajar dalam skala kelas, institusi, maupun tes terstandar. Meskipun demikian, kualitas butir pilihan ganda tidak hanya ditentukan oleh keberadaan satu jawaban benar. Butir yang baik perlu memiliki tingkat kesukaran yang sesuai, mampu membedakan peserta berkemampuan tinggi dan rendah, serta memiliki pengecoh yang cukup plausibel sehingga respons salah tidak terkonsentrasi pada satu pilihan yang terlalu lemah. Prinsip tersebut sejalan dengan pedoman penulisan butir pilihan ganda dan analisis butir dalam pengukuran pendidikan [1].

Classical Test Theory (CTT) menyediakan indikator praktis untuk mengevaluasi kualitas butir, seperti indeks kesukaran, indeks daya beda, dan analisis pilihan jawaban. Item Response Theory (IRT) menawarkan model yang lebih kaya melalui parameter kesukaran, daya beda, dan peluang menebak, tetapi estimasi parameter IRT memerlukan sampel respons yang relatif besar dan stabil [2]. Pada tahap penyusunan bank soal, data respons empiris sering belum tersedia. Kondisi ini menimbulkan kebutuhan terhadap pendekatan awal yang dapat membantu penulis soal menyeleksi dan merevisi butir sebelum uji coba lapangan dilakukan.

Perkembangan educational data mining, natural language processing, dan machine learning membuka peluang baru untuk analisis kualitas butir secara otomatis. Model seperti random forest, gradient boosting, dan model bahasa neural telah digunakan untuk memprediksi kesukaran soal, menilai kualitas jawaban, dan mendukung generasi soal atau pengecoh [3], [4],

[5]. Namun, model prediktif yang akurat belum tentu berguna secara pedagogis apabila alasan di balik prediksi tidak dapat dipahami oleh pengembang asesmen.

Explainable artificial intelligence (XAI) menjawab persoalan tersebut dengan menyediakan mekanisme interpretasi global dan lokal terhadap keputusan model. SHAP, misalnya, menguraikan kontribusi setiap fitur terhadap prediksi berdasarkan konsep Shapley value, sedangkan feature importance pada model berbasis pohon memberikan gambaran fitur yang paling sering dan paling kuat digunakan dalam proses pemisahan data [6], [7]. Dalam konteks analisis butir, explainability penting karena keluaran sistem seharusnya tidak berhenti pada label “baik” atau “kurang”, tetapi juga menunjukkan alasan psikometrik yang dapat ditindaklanjuti.

Penelitian automated item evaluation berkembang dari prediksi kesukaran berbasis fitur linguistik menuju representasi semantik yang lebih kaya. Benedetto et al. [8] menunjukkan bahwa model Transformer dapat mengungguli pendekatan terdahulu dalam estimasi kesukaran soal pilihan ganda berbasis teks, sedangkan Byrd dan Srivastava [9] memperluas prediksi hingga parameter kesukaran dan daya beda pada pertanyaan bahasa alami. Tinjauan sistematis AlKhuyaey et al. [10], [11] menegaskan bahwa supervised machine learning masih dominan, tetapi fitur sintaktik dan semantik yang lebih kaya semakin penting. Shared task BEA 2024 juga memperlihatkan meningkatnya perhatian pada prediksi otomatis kesukaran dan waktu respons menggunakan butir ujian berlabel empiris [11].

CommonsenseQA sendiri telah banyak digunakan sebagai benchmark untuk tugas pemilihan jawaban, bukan untuk klasifikasi kualitas butir. Pada publikasi awal, baseline BERT-large mencapai akurasi sekitar 56%, masih jauh di bawah performa manusia 89% [12]. Selanjutnya, QA-GNN yang menggabungkan language model dan knowledge graph mencapai akurasi 76,1% pada official test CommonsenseQA [13], sedangkan ACENet dengan attention-guided reasoning pada hybrid knowledge graph melaporkan 78,54% pada in-house development set dan 74,72% pada in-house test set [14]. Angka-angka tersebut tidak dapat dibandingkan langsung dengan F1-score penelitian ini karena tugasnya berbeda—answer prediction versus item-quality classification—namun menunjukkan bahwa CommonsenseQA merupakan benchmark publik berskala besar dengan struktur pilihan ganda yang stabil untuk pengujian pipeline komputasional.

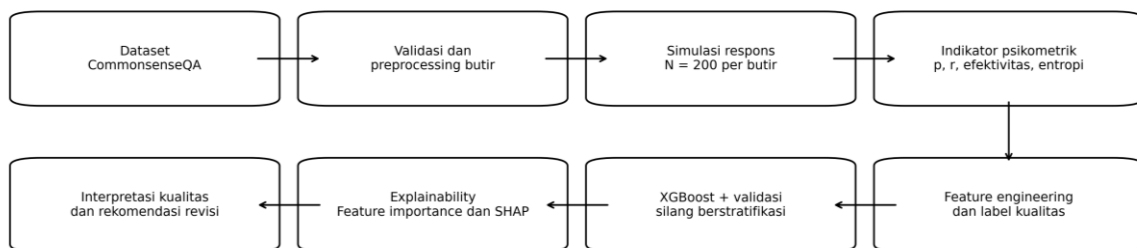
Dari sisi simulasi, studi Monte Carlo psikometrik klasik umumnya mengevaluasi pemulihan parameter di bawah asumsi model IRT yang telah ditentukan [15], sedangkan pendekatan mutakhir mulai menggunakan LLM untuk mensimulasikan respons siswa dengan tingkat kemampuan berbeda [16]. Keduanya memperlihatkan bahwa validitas simulasi sangat bergantung pada asumsi pembangkit respons. Kelemahan yang khusus pada desain penelitian ini adalah potensi feature–target coupling: label kualitas dibentuk dari indikator psikometrik yang juga digunakan sebagai fitur prediktor. Karena itu, kebaruan yang ditawarkan bukan klaim prediksi kualitas secara eksternal, melainkan integrasi yang transparan antara simulasi probabilistik, indikator CTT dan pengecoh, klasifikasi XGBoost, serta penjelasan SHAP untuk mengaudit konsistensi internal dan menghasilkan hipotesis revisi butir sebelum validasi dengan respons peserta nyata.

Penelitian ini mengembangkan kerangka explainable machine learning untuk mengevaluasi kualitas soal pilihan ganda berbasis indikator psikometrik dan efektivitas pengecoh. Secara khusus, penelitian bertujuan untuk: (1) menyusun pipeline simulasi respons pada dataset pilihan ganda yang tidak memiliki respons peserta empiris; (2) menghitung indikator psikometrik klasik dan fitur linguistik permukaan; (3) melatih XGBoost untuk mempelajari aturan klasifikasi kualitas butir dan membandingkannya dengan baseline aturan komposit; serta (4) menjelaskan prediksi secara global dan lokal menggunakan feature importance dan SHAP. Posisi kebaruan penelitian terletak pada integrasi simulasi respons, ukuran fungsi pengecoh termasuk entropi, klasifikasi gradient boosting, dan explainability dalam satu pipeline yang dapat ditelusuri. Berbeda dari studi simulasi psikometrik yang berfokus pada pemulihan parameter model dan studi prediksi kesukaran berbasis teks yang menggunakan target empiris atau penilaian ahli, penelitian ini sengaja diposisikan sebagai

proof-of-concept. Dengan demikian, hasil digunakan untuk menguji konsistensi logika fitur dan keterjelasan penjelasan model, bukan sebagai bukti bahwa kualitas butir dapat diprediksi secara valid pada peserta didik nyata.

2. METODE PENELITIAN

Penelitian ini menggunakan desain komputasional kuantitatif berbasis eksperimen simulasi. Unit analisis adalah butir soal pilihan ganda. Alur penelitian mencakup pemilihan dataset, validasi struktur butir, simulasi respons, perhitungan indikator psikometrik, rekayasa fitur dan pelabelan kualitas, pelatihan XGBoost serta baseline aturan komposit, evaluasi menggunakan precision, recall, F1-score, accuracy, confusion matrix, dan AUC-ROC, validasi silang berstratifikasi, serta interpretasi model dengan feature importance dan SHAP.



Gambar 1. Alur penelitian explainable machine learning untuk analisis kualitas butir

Dataset dan Unit Analisis

Dataset yang digunakan adalah CommonsenseQA, yaitu benchmark pertanyaan penalaran commonsense berbentuk pilihan ganda dengan lima kandidat jawaban berlabel A sampai E dan satu jawaban benar [12]. CommonsenseQA dipilih sebagai testbed komputasional karena format lima opsinya konsisten, kunci jawaban tersedia pada split training dan validation, jumlah butir cukup besar untuk eksperimen klasifikasi, dan opsi pengecohnya dibangun untuk membedakan konsep yang berelasi. Karakteristik tersebut mendukung pengujian pipeline secara terkontrol, tetapi tidak menjadikan CommonsenseQA ekuivalen dengan bank soal pendidikan yang telah divalidasi terhadap kurikulum atau populasi peserta didik. Penelitian ini menggunakan 9.741 butir dari split training dan 1.221 butir dari split validation, sehingga total butir yang dianalisis adalah 10.962. Split test tidak digunakan karena kunci jawaban tidak tersedia secara publik, sedangkan perhitungan indikator dalam penelitian ini membutuhkan identifikasi jawaban benar.

Sebelum analisis, setiap butir diperiksa untuk memastikan keberadaan stem, lima opsi jawaban, dan kunci jawaban. Butir yang tidak memenuhi struktur tersebut dikeluarkan. Posisi kunci jawaban tidak digunakan sebagai fitur prediktor agar model tidak belajar dari bias posisi pilihan jawaban.

Simulasi Respons Peserta

Karena data respons peserta empiris tidak tersedia, penelitian ini mensimulasikan respons 200 responden virtual untuk setiap butir. Kunci jawaban diberi bobot dasar dari distribusi uniform $U(4,0; 5,0)$ untuk merepresentasikan kecenderungan memilih jawaban benar. Setiap pengecoh memperoleh bobot berdasarkan panjang opsi sebagai proksi plausibilitas permukaan, kemudian diskalakan dengan distribusi uniform $U(0,5; 1,5)$. Semua bobot dinormalisasi menjadi probabilitas pilihan atas lima opsi, lalu respons virtual diambil dari distribusi multinomial tersebut. Mekanisme ini bersifat heuristik: tidak mengestimasi kemampuan laten peserta seperti pada IRT dan tidak melakukan penalaran semantik seperti simulasi berbasis LLM. Kesederhanaan tersebut dipilih agar mekanisme pembangkit respons

dapat diaudit, tetapi konsekuensinya hasil hanya merepresentasikan skenario simulasi yang didefinisikan oleh peneliti.

Strategi ini tidak dimaksudkan untuk menggantikan respons empiris, melainkan menyediakan data simulatif untuk menguji pipeline dan mekanisme explainability pada tahap awal. Oleh karena itu, distribusi respons virtual, indikator psikometrik yang diturunkan, dan label kualitas harus dipandang sebagai konstruksi komputasional. Generalisasi ke respons siswa nyata memerlukan validasi eksternal menggunakan data lapangan dan, idealnya, parameter CTT/IRT atau penilaian ahli yang dibentuk secara independen.

Indikator Psikometrik dan Fitur

Empat indikator psikometrik dihitung dari distribusi respons simulatif. Indeks kesukaran (p) didefinisikan sebagai proporsi responden virtual yang memilih jawaban benar. Indeks daya beda (r) dihitung sebagai korelasi point-biserial antara skor butir biner dan proksi skor total. Efektivitas pengecoh dihitung sebagai proporsi pengecoh yang dipilih oleh sekurang-kurangnya 5% responden virtual. Entropi pengecoh dihitung menggunakan entropi Shannon pada distribusi respons salah di antara empat pengecoh; nilai mendekati $\log_2(4)$ atau 2 bit menunjukkan sebaran pengecoh yang lebih merata.

Rincian pembentukan proksi skor total perlu dinyatakan eksplisit agar perhitungan r dapat direplikasi. Untuk menjaga keterlacakan simulasi, proksi skor total untuk responden virtual ke- j pada butir ke- i dibentuk sebagai $T_{ij} = X_{ij} + \varepsilon_{ij}$, dengan $X_{ij} \in \{0,1\}$ merupakan skor respons pada butir yang sedang dianalisis dan $\varepsilon_{ij} \sim N(0; 0,28^2)$ merupakan perturbasi acak independen. Dengan konstruksi tersebut, skor butir yang sedang dianalisis tetap menjadi komponen proksi dan tidak dikeluarkan. Nilai r kemudian dihitung sebagai korelasi point-biserial antara X_{ij} dan T_{ij} . Proksi ini sengaja bersifat sintetis dan tidak diperlakukan sebagai pengganti corrected item-total correlation pada data tes empiris.

Selain empat indikator psikometrik, lima fitur linguistik permukaan diekstraksi dari stem soal, yaitu jumlah kata, jumlah karakter, kehadiran negasi, jumlah klausa, dan rata-rata panjang kata. Ringkasan fitur disajikan pada Tabel 1.

Tabel 1. Set fitur yang digunakan dalam klasifikasi kualitas soal

No	Fitur	Deskripsi	Tipe	Rata-rata
1	Indeks Kesukaran (p)	Proporsi jawaban benar dalam simulasi	Psikometrik	0,519
2	Indeks Daya Beda (r)	Korelasi point-biserial	Psikometrik	0,872
3	Efektivitas Pengecoh	Proporsi pengecoh efektif ($\geq 5\%$)	Psikometrik	0,983
4	Entropi Pengecoh	Entropi Shannon dari distribusi respons salah	Psikometrik	1,928
5	Jumlah Kata	Jumlah kata dalam stem soal	Linguistik	13,23
6	Jumlah Karakter	Total karakter dalam stem soal	Linguistik	69,31
7	Kehadiran Negasi	Indikator biner keberadaan kata negasi	Linguistik	0,07
8	Jumlah Klausa	Jumlah klausa berdasarkan tanda koma atau titik koma	Linguistik	1,63
9	Rata-rata Panjang Kata	Rata-rata jumlah karakter per kata	Linguistik	4,71

Pelabelan Kualitas Butir

Kualitas butir diklasifikasikan ke dalam tiga kategori: kurang, sedang, dan baik. Pelabelan dilakukan menggunakan skor komposit dari empat indikator psikometrik. Untuk indikator yang semakin besar semakin diharapkan, seperti daya beda, efektivitas pengecoh, dan entropi pengecoh, skor diberikan berdasarkan posisi relatif terhadap persentil ke-33 dan ke-67. Untuk indeks kesukaran, interpretasi diarahkan pada kedekatan terhadap tingkat kesukaran moderat, bukan semata-mata semakin tinggi semakin baik. Skor komposit memiliki rentang 0

sampai 8. Butir dengan skor 6 atau lebih dikategorikan baik, skor 3 sampai 5 dikategorikan sedang, dan skor di bawah 3 dikategorikan kurang.

Karena keempat indikator yang membentuk label juga digunakan sebagai fitur prediktor, skema ini secara sengaja menghasilkan feature–target coupling. Konsekuensinya, eksperimen XGBoost dievaluasi sebagai kemampuan model untuk mengaproksimasi aturan pelabelan internal, bukan sebagai prediksi kualitas eksternal yang independen.

Skema tersebut menghasilkan 1.088 butir kategori kurang (9,9%), 7.738 butir kategori sedang (70,6%), dan 2.136 butir kategori baik (19,5%). Distribusi ini menunjukkan dominasi butir kualitas sedang, suatu pola yang realistis dalam bank soal awal karena sebagian besar butir memerlukan penyempurnaan sebelum digunakan dalam asesmen berisiko tinggi.

Pelatihan, Evaluasi, dan Explainability Model

Model yang digunakan adalah XGBoost Classifier, yaitu implementasi gradient boosting berbasis pohon keputusan yang efektif untuk data tabular [17]. Parameter utama meliputi $n_estimators = 300$, $max_depth = 5$, $learning_rate = 0,05$, $subsample = 0,8$, dan $colsample_bytree = 0,8$. Pertama, data dibagi menjadi train-test split 80:20 secara berstratifikasi. Kedua, stabilitas internal dinilai secara terpisah menggunakan Stratified 5-Fold Cross-Validation pada keseluruhan dataset. Metrik evaluasi meliputi accuracy, precision, recall, F1-score, confusion matrix, dan AUC-ROC dengan skema one-vs-rest. Sebagai kontrol logis, baseline aturan komposit menerapkan kembali persis aturan pelabelan.

Explainability dilakukan pada dua tingkat. Pertama, feature importance bawaan XGBoost digunakan untuk menilai kontribusi global setiap fitur terhadap struktur pohon. Kedua, nilai SHAP dihitung menggunakan TreeExplainer untuk menjelaskan kontribusi fitur terhadap prediksi model secara global dan lokal.

Lingkungan Komputasi dan Reprodusibilitas

Implementasi referensi pipeline distandarkan pada lingkungan Python 3.13.5 dengan NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0, XGBoost 3.1.3, dan SHAP 0.50.0. Pipeline bersifat CPU-only. Random seed 42 ditetapkan secara konsisten pada pembangkitan respons, perturbasi proksi skor total, train-test split, XGBoost ($random_state = 42$), serta StratifiedKFold dengan $shuffle = True$ dan $random_state = 42$.

eXtreme Gradient Boosting (XGBoost)

XGBoost adalah algoritma ensemble learning berbasis pohon keputusan yang mengimplementasikan teknik gradient boosting secara sangat efisien. Prinsip utama gradient boosting adalah membangun serangkaian model pohon secara berurutan, di mana setiap pohon baru dilatih untuk memprediksi dan mengoreksi kesalahan residual dari model gabungan sebelumnya.

Rumus Matematis: Untuk himpunan data dengan n observasi dan m fitur, model ensemble berbasis pohon direpresentasikan sebagai penjumlahan dari K fungsi aditif, ditunjukkan pada persamaan (1).

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (1)$$

Keterangan: K adalah Jumlah total pohon keputusan, $f_k(x_i)$ adalah skor atau prediksi yang diberikan oleh pohon ke- k untuk observasi x_i .

SHapley Additive exPlanations (SHAP)

SHAP, yang diperkenalkan oleh Lundberg dan Lee [6], adalah kerangka kerja Explainable AI (XAI) yang menyatukan berbagai metode atribusi fitur di bawah prinsip Shapley value dari teori permainan kooperatif. Metode ini mengkuantifikasi kontribusi marginal setiap

fitur terhadap prediksi tertentu, relatif terhadap output model yang diharapkan. Persamaan model penjelasannya ditunjukkan pada persamaan (2).

$$g(z') = \varphi_0 + \sum_{j=1}^M \varphi_j z'_j \quad (2)$$

3. HASIL DAN PEMBAHASAN

Statistik Deskriptif Indikator Psikometrik

Statistik deskriptif hasil simulasi disajikan pada Tabel 2. Rata-rata indeks kesukaran sebesar 0,519 menunjukkan bahwa sebagian besar butir berada pada rentang kesukaran moderat. Nilai indeks daya beda rata-rata sebesar 0,872 perlu ditafsirkan hati-hati karena nilainya lebih tinggi dibandingkan pola yang biasanya dijumpai pada respons empiris. Rata-rata efektivitas pengecoh sebesar 0,983 dan entropi pengecoh sebesar 1,928 menunjukkan bahwa pengecoh cenderung dipilih secara cukup merata dalam skenario simulasi.

Tabel 2. Statistik deskriptif indikator psikometrik hasil simulasi (n = 10.962)

Indikator	Min	Q1	Mean	Q3	Maks
Indeks Kesukaran (p)	0,220	0,480	0,519	0,560	0,720
Indeks Daya Beda (r)	0,825	0,866	0,872	0,878	0,905
Efektivitas Pengecoh	0,500	1,000	0,983	1,000	1,000
Entropi Pengecoh (bit)	1,045	1,901	1,928	1,970	2,000

Performa Klasifikasi Model

Model XGBoost menunjukkan performa klasifikasi internal yang sangat tinggi pada test set yang disisihkan. Seperti disajikan pada Tabel 3, precision, recall, dan F1-score pada seluruh kelas berada di atas 0,99. Hasil Stratified 5-Fold Cross-Validation pada Tabel 4 konsisten dengan F1-score terbobot dan accuracy rata-rata $0,9987 \pm 0,0011$.

Tabel 3. Classification report pada test set

Metrik	Kurang	Sedang	Baik	Rata-rata Terbobot
Precision	0,9954	0,9981	1,0000	0,9982
Recall	1,0000	0,9994	0,9930	0,9982
F1-score	0,9977	0,9987	0,9965	0,9982
Support (n)	218	1.548	427	2.193

Tabel 4. Hasil Stratified 5-Fold Cross-Validation

Fold	F1 (terbobot)	Accuracy	n Train	n Val
1	0,9995	0,9995	8.769	2.193
2	0,9977	0,9977	8.769	2.193
3	1,0000	1,0000	8.769	2.193
4	0,9991	0,9991	8.770	2.192
5	0,9973	0,9973	8.770	2.192
Mean $\pm$ SD	0,9987 $\pm$ 0,0011	0,9987 $\pm$ 0,0011	-	-

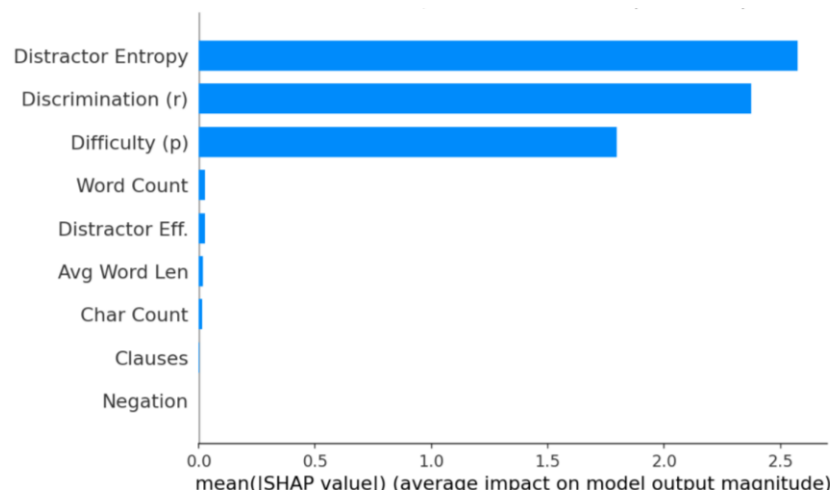
Meskipun angka performa tersebut mendekati sempurna, pola ini lebih tepat dijelaskan sebagai structural feature–target coupling daripada sebagai bukti generalisasi yang luar biasa. Target kualitas dibentuk secara deterministik dari empat indikator psikometrik, dan indikator yang sama menjadi prediktor utama model.

Interpretasi Global: Feature Importance dan SHAP

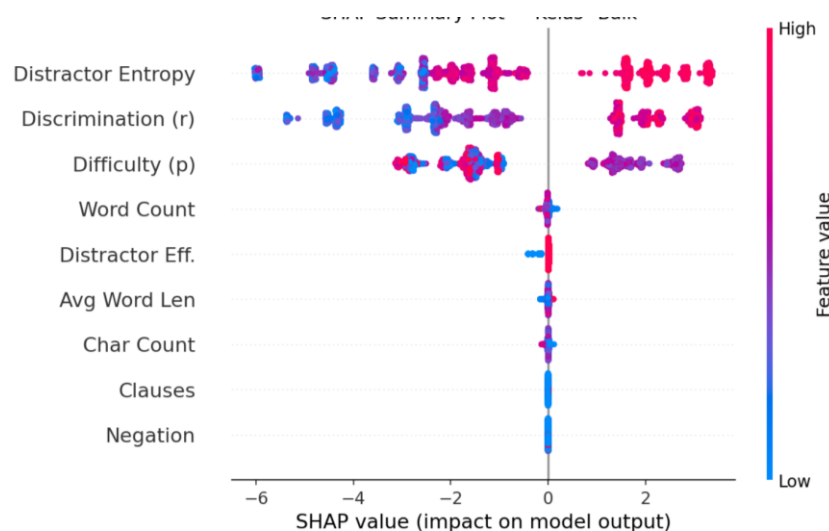
Feature importance bawaan XGBoost menempatkan entropi pengecoh sebagai kontributor terbesar (36,4%), diikuti indeks daya beda (27,9%), efektivitas pengecoh (17,6%), dan indeks kesukaran (15,7%). Lima fitur linguistik secara kolektif menyumbang kurang dari

2,5% dari total importance. Pola ini menunjukkan bahwa kualitas butir dalam pipeline ini lebih ditentukan oleh karakteristik distribusi respons dan pengecoh daripada oleh ciri permukaan teks.

Hasil SHAP global memperkuat temuan tersebut. Entropi pengecoh dan indeks daya beda kembali menjadi fitur paling dominan, sebagaimana ditunjukkan pada Gambar 2 melalui nilai kepentingan fitur secara global. Konsistensi antara feature importance dan SHAP penting karena kedua metode memiliki dasar perhitungan berbeda. Feature importance mengukur kontribusi fitur terhadap struktur pohon, sedangkan SHAP mengukur kontribusi marginal fitur terhadap output prediksi. Distribusi nilai SHAP beserta arah pengaruh setiap fitur terhadap prediksi kelas kualitas “Baik” ditunjukkan lebih rinci pada Gambar 3. Selanjutnya, untuk memberikan gambaran empiris mengenai karakteristik butir pada masing-masing kelas kualitas, Tabel 5 menyajikan profil psikometrik beberapa butir yang representatif.



Gambar 2. SHAP feature importance global pada model XGBoost



Gambar 3. SHAP summary plot untuk kelas kualitas “Baik”.

Interpretasi Lokal dan Studi Kasus Butir

Selain interpretasi global, SHAP juga memungkinkan penjelasan lokal pada tingkat butir. Pada butir kategori baik, kontribusi positif terbesar berasal dari entropi pengecoh dan indeks daya beda. Pada butir

kategori kurang, kedua indikator tersebut memberikan kontribusi negatif, yang mengindikasikan bahwa distribusi pengecoh atau kemampuan butir membedakan responden belum optimal.

Tabel 5. Profil psikometrik butir representatif berdasarkan kelas kualitas

Kelas Kualitas	Pernyataan Soal (dipotong)	P	r	Efektivitas Pengecoh	Entropi
Baik	If a student is staring at the ceiling while bored, where are they likely to be?	0,505	0,878	1,000	1,931
Sedang	Where would you put an eraser if you are at home and need to go to class?	0,390	0,864	1,000	1,934
Kurang	Where are all employees likely to carry a weapon?	0,565	0,865	1,000	1,908

Pembahasan

Temuan utama penelitian ini adalah dominasi indikator psikometrik berbasis distribusi respons terutama entropi pengecoh dan indeks daya beda—dalam klasifikasi kualitas butir pada skenario simulasi. Literatur empiris mendukung pentingnya fungsi pengecoh, tetapi juga menunjukkan hubungan yang tidak selalu deterministik. Puthiaram dan Rahman [18] menemukan bahwa distractor efficiency berkaitan dengan indeks psikometrik, tetapi tidak menghasilkan satu pola sempurna yang berlaku pada semua butir. Rezigalla et al. [19] juga melaporkan keterkaitan distractor efficiency dengan difficulty index dan discrimination power pada data ujian aktual.

Kontribusi lima fitur linguistik permukaan yang sangat kecil juga perlu dipahami secara hati-hati. Temuan ini tidak membuktikan bahwa karakteristik teks tidak relevan, melainkan menunjukkan bahwa word count, character count, negation, clause count, dan average word length terlalu dangkal dibandingkan indikator yang secara langsung membentuk target. Studi berbasis Transformer telah menunjukkan manfaat representasi teks yang lebih kaya untuk estimasi kesukaran [8], sedangkan tinjauan sistematis AlKhuzayy et al. [10] menempatkan fitur sintaktik dan semantik sebagai komponen penting. Shared task Yaneva et al. [11] dan studi terbaru Razavi dan Powers [20] juga memperkuat arah menuju pemodelan kesukaran menggunakan sinyal linguistik/kognitif yang dikaitkan dengan target empiris.

Secara praktis, pipeline dapat dipertimbangkan sebagai alat prapenyaringan bank soal: sistem menandai butir yang perlu diperiksa, menunjukkan indikator yang dominan, dan membantu penulis memprioritaskan revisi sebelum uji coba lapangan.

4. KESIMPULAN

Penelitian ini menghasilkan proof-of-concept explainable machine learning berbasis simulasi respons untuk mengaudit kualitas butir pilihan ganda sebelum data peserta empiris tersedia. Pada skenario yang dibangun, entropi pengecoh dan indeks daya beda menjadi pendorong utama prediksi XGBoost, sedangkan fitur linguistik permukaan memberi kontribusi terbatas. Implikasi teoretisnya adalah bahwa pemerataan respons salah melalui entropi berpotensi melengkapi ukuran distractor efficiency berbasis ambang; implikasi praktisnya adalah bahwa SHAP dapat mengubah keluaran klasifikasi menjadi petunjuk diagnostik yang lebih mudah ditelusuri oleh pengembang bank soal. Namun, performa internal yang sangat tinggi terutama mencerminkan keterkaitan struktural antara fitur dan label, bukan bukti validitas prediktif eksternal.

5. SARAN

Validasi berikutnya perlu diprioritaskan pada penggunaan respons siswa nyata, pembentukan target kualitas yang independen, dan integrasi parameter CTT/IRT sehingga kualitas model dapat dibandingkan dengan konstruk psikometrik empiris. Penelitian lanjutan

juga perlu membandingkan beberapa mekanisme simulasi respons, melakukan stress test dengan synthetic noise dan distribution shift, menambahkan representasi semantik berbasis Transformer/LLM untuk stem dan pengecoh, serta menguji generalisasi lintas domain.

6. UCAPAN TERIMA KASIH

Penelitian ini menggunakan dataset publik dan respons peserta yang disimulasikan secara komputasional; tidak ada subjek manusia aktual yang direkrut dalam penelitian ini. Dataset CommonsenseQA tersedia secara publik melalui repositori resmi pengembang dataset.

DAFTAR PUSTAKA

- [1] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment," *Applied Measurement in Education*, vol. 15, no. 3, pp. 309–333, Jul. 2002, doi: 10.1207/S15324818AME1503_5.
- [2] F. B. Baker and S.-H. Kim, *The Basics of Item Response Theory Using R*. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-54205-8.
- [3] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A Systematic Review of Automatic Question Generation for Educational Purposes," *Int. J. Artif. Intell. Educ.*, vol. 30, no. 1, pp. 121–124, Mar. 2020, Accessed: Sep. 30, 2026. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1560429226006967>
- [4] B. Settles, G. T. LaFlair, and M. Hagiwara, "Machine Learning–Driven Language Assessment," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 247–263, Dec. 2020, doi: 10.1162/tacl_a_00310.
- [5] L. A. Ha, V. Yaneva, P. Baldwin, and J. Mee, "Predicting the Difficulty of Multiple Choice Questions in a High-stakes Medical Exam," in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 11–20. doi: 10.18653/v1/W19-4402.
- [6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in *NIPS'17*. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 4768–4777.
- [7] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.
- [8] L. Benedetto, G. Aradelli, P. Cremonesi, A. Cappelli, A. Giussani, and R. Turrin, "On the application of Transformers for estimating the difficulty of Multiple-Choice Questions from text," in *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, J. Burstein, A. Horbach, E. Kochmar, R. Laarmann-Quante, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, and T. Zesch, Eds., Online: Association for Computational Linguistics, Apr. 2021, pp. 147–157. [Online]. Available: <https://aclanthology.org/2021.bea-1.16/>
- [9] M. Byrd and S. Srivastava, "Predicting Difficulty and Discrimination of Natural Language Questions," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*,

- Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 119–130. doi: 10.18653/v1/2022.acl-short.15.
- [10] S. AlKhuyaey, F. Grasso, T. R. Payne, and V. Tamma, “Text-based Question Difficulty Prediction: A Systematic Review of Automatic Approaches,” *Int. J. Artif. Intell. Educ.*, vol. 34, no. 3, pp. 862–914, Sep. 2024, doi: 10.1007/s40593-023-00362-1.
 - [11] V. Yaneva et al., “Findings from the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple-Choice Questions,” in *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 470–482. [Online]. Available: <https://aclanthology.org/2024.bea-1.39/>
 - [12] A. Talmor, J. Herzig, N. Lourie, and J. Berant, “CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4149–4158. doi: 10.18653/v1/N19-1421.
 - [13] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, “QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 535–546. doi: 10.18653/v1/2021.naacl-main.45.
 - [14] C. Hao, M. Xie, and P. Zhang, “ACENet: Attention Guided Commonsense Reasoning on Hybrid Knowledge Graph,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 8461–8471. doi: 10.18653/v1/2022.emnlp-main.579.
 - [15] M. Harwell, C. A. Stone, T.-C. Hsu, and L. Kirisci, “Monte Carlo Studies in Item Response Theory,” *Appl. Psychol. Meas.*, vol. 20, no. 2, pp. 101–125, Jun. 1996, doi: 10.1177/014662169602000201.
 - [16] L. Benedetto, G. Aradelli, A. Donvito, A. Lucchetti, A. Cappelli, and P. Buttery, “Using LLMs to simulate students’ responses to exam questions,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 11351–11368. doi: 10.18653/v1/2024.findings-emnlp.663.
 - [17] T. Chen and C. Guestrin, “XGBoost,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
 - [18] T. Puthiarampil and M. Rahman, “How important is distractor efficiency for grading Best Answer Questions?,” *BMC Med. Educ.*, vol. 21, no. 1, p. 29, Dec. 2021, doi: 10.1186/s12909-020-02463-0.
 - [19] A. A. Rezigalla et al., “Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items,” *BMC Med. Educ.*, vol. 24, no. 1, p. 445, Apr. 2024, doi: 10.1186/s12909-024-05433-y.
 - [20] P. Razavi and S. Powers, “Estimating item difficulty using large language models and tree-based machine learning algorithms,” *Int. J. Artif. Intell. Educ.*, vol. 36, no. 3, p. 100015, Sep. 2026, doi: 10.1016/j.ijaiied.2026.100015.