



Integrasi Analisis Butir Soal dan Klasifikasi Kognitif Untuk Personalisasi Rekomendasi Remedial Adaptif Menggunakan K-Nearest Neighbor

Integration of Item Analysis and Cognitive Classification for Personalizing Adaptive Remedial Recommendations Using K-Nearest Neighbor

Karimah Agustin*, Alun Sujjada, Muhammad Ikhsan Thohir

Fakultas Teknik, Komputer dan Desain, Prodi Teknik Informatika, Universitas Nusa Putra, Sukabumi, Indonesia
Jl. Raya Cibolang Cisaat - Sukabhumi No.21, Cibolang Kaler, Kec. Cisaat, Kabupaten Sukabumi, Jawa Barat, Indonesia
Email: ¹*karimah.agustin_ti22@nusaputra.ac.id, ²Alun.sujjada@nusaputra.ac.id, ³Ikhsan.thohir@nusaputra.ac.id

Email Penulis Korespondensi: karimah.agustin_ti22@nusaputra.ac.id,

Received: 25/08/2026, Revised: 05/09/2026, Accepted: 09/09/2026, Published: 11/10/2026

Abstrak—Penyusunan soal ujian dibanyak sekolah menengah termasuk madrasah tsanawiyah masih dilakukan secara manual sehingga guru kesulitan memastikan tingkat kognitif soal sesuai taksonomi bloom sekaligus menilai kualitas butir soal. Penelitian ini mencoba menjawab persoalan itu dengan membangun sebuah platform ujian berbasis web yang secara otomatis mengklasifikasikan butir soal pilihan ganda, mengevaluasi kualitas psikometriknya, serta menghasilkan rekomendasi remedial yang disesuaikan dengan profil masing-masing siswa. Penandaan level kognitif (C1-C6) dilakukan melalui pencocokan kata kerja operasional menggunakan keyword matching. Sementara algoritma K-Nearest Neighbor yang dilatih pada representasi TF-IDF memisahkan soal ke dalam kategori berpikir tingkat tinggi atau Higher Order Thinking Skills (HOTS) dan tingkat rendah atau Lower Order Thinking Skills (LOTS). Kualitas butir soal dinilai memakai Classical Test Theory (CTT) untuk menghitung indeks kesukaran dan daya pembeda setiap soal secara otomatis dari pola jawaban siswa, dan materi remedial dipilihkan untuk setiap siswa melalui pendekatan content-based filtering berdasarkan kemiripan profil kelemahan kognitifnya. Sistem dibangun menggunakan model waterfall dan diuji pada 600 butir soal dari 3 mata Pelajaran. Model klasifikasi mencapai akurasi 86,67% pada k=11 dikonfirmasi melalui validasi 10-fold cross validation sebesar $81,83\% \pm 5,19\%$ sedangkan seluruh pengujian blackbox sekitar 23 skenario dinyatakan berhasil. Temuan ini menunjukkan bahwa penggabungan analisis butir soal dengan mekanisme remedial dapat diterapkan secara teknis pada konteks penilaian ditingkat madrasah. Kontribusi utama penelitian ini adalah penyatuan tiga proses yaitu klasifikasi kognitif otomatis, analisis psikometrik CTT, dan rekomendasi remedial berbasis content-based filtering ke dalam satu platform ujian yang berjalan secara real-time, sebuah kombinasi yang belum ditemukan pada penelitian-penelitian sejenis sebelumnya yang umumnya hanya membahas salah satu dari ketiga aspek tersebut secara terpisah.

Kata Kunci: Analisis Butir Soal; Klasifikasi Taksonomi Bloom; k-Nearest Neighbor; Rekomendasi Remedial; Ujian Berbasis Web

Abstract—In many secondary schools, classroom assessments still use tests that are written by hand, which makes it hard for teachers to ensure each question matches the right level of thinking from Bloom's taxonomy or to check if the questions are statistically reliable. This study tackles that issue by creating a web-based testing system for a madrasah tsanawiyah, which automatically sorts multiple-choice questions, checks their quality using psychometric methods, and offers custom learning suggestions to help students improve. Cognitive-level tags (C1 to C6) are assigned by matching keywords that describe actions, and a kNearest Neighbor classifier uses TF-IDF features to sort items into groups that show more or less complex thinking. Item quality is checked using Classical Test Theory module automatically computes the difficulty index and discrimination index of each item from students answer patterns, and materials that help improve learning are chosen based on each student's weak areas through content-based filtering. The system was developed using the Waterfall model and was tested on 600 items taken from three different subjects. The classifier achieved 86.67% accuracy when using k=11, which was verified through 10-fold cross-validation showing $81.83\% \pm 5.19\%$, and all 23 black-box test cases were successfully passed. The results show that using automatic item analysis along with adaptive remediation is possible to implement in assessment practices at the madrasah level. The main contribution of this study is the integration of automatic cognitive classification, CTT-based psychometric analysis, and content-based filtering remedial recommendation into a single real-time examination platform, a combination that has not been found in previous similar studies, which generally address only one of these three aspects separately.

Keywords: Item Analysis; Bloom Taxonomy Classification; k-Nearest Neighbor; Remedial Recommendation; Web-Based Examination

1. PENDAHULUAN

Evaluasi hasil belajar adalah bagian penting untuk mengetahui sejauh mana siswa telah mencapai kompetensi yang ditentukan oleh kurikulum. Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 21 Tahun 2022 tentang Standar Penilaian Pendidikan menyatakan bahwa penilaian nasional harus mampu mengukur kemampuan berpikir siswa mulai dari tingkat yang paling dasar hingga tingkat yang lebih tinggi[1]. Oleh karena itu, instrumen soal sebaiknya disusun dengan memperhatikan distribusi tingkat kognitif yang jelas, bukan hanya sekadar kumpulan soal berupa hafalan. Kerangka yang sering digunakan untuk menentukan tingkat kemampuan

berpikir adalah Taksonomi Bloom yang telah direvisi oleh Anderson dan Krathwohl. Kerangka ini membagi cara berpikir manusia menjadi enam tahap yang urut, yaitu mengingat, memahami, menerapkan, menganalisis, mengevaluasi, dan mencipta [2]. Panduan asesmen Kurikulum Merdeka justru menyebutkan bahwa penilaian sebaiknya memiliki tiga jenis yaitu diagnostik, formatif, dan sumatif secara bersamaan, agar guru bisa memahami secara lengkap perkembangan belajar siswanya [3].

Praktik di lapangan, terutama di tingkat Madrasah Tsanawiyah, sering kali tidak sesuai dengan kondisi yang ideal. Yandra dkk, yang memeriksa 275 soal Bahasa Indonesia dari Asesmen Sumatif Tengah Semester dan Asesmen Sumatif Akhir Semester di sebuah SMP negeri, menemukan bahwa 90,91% soal masih masuk dalam kategori Lower Order Thinking Skills (LOTS), sedangkan hanya 9,09% yang termasuk Higher Order Thinking Skills (HOTS) [4]. Cressa dan Mukhlis melaporkan pola yang sama, yaitu 56% soal Bahasa Indonesia berada pada tingkat mengingat dan memahami [5]. Kesenjangan ini biasanya terjadi karena guru membuat dan mengelompokkan soal secara manual berdasarkan pengalaman pribadi, sehingga cenderung bersifat subjektif dan kesulitannya dalam mempertahankan konsistensi dari satu waktu ujian ke waktu ujian berikutnya. Masalah tidak berhenti pada sebaran kognitif saja. Dari segi psikometri, kualitas soal juga dilihat dari tingkat kesulitannya dan kemampuannya membedakan jawaban peserta, dua hal yang dihitung dengan menggunakan teori uji klasik (CTT) [6][7]. Izzah dan teman-temannya menemukan bahwa sebagian besar soal Bahasa Indonesia di tingkat dasar memiliki kemampuan untuk membedakan jawaban yang rendah [8] dan kemungkinan besar kondisi serupa juga terjadi pada mata pelajaran lain seperti Akidah Akhlak dan IPS, meski belum banyak diteliti secara langsung, terutama di tingkat madrasah yang mengikuti kurikulum khusus di bawah Kementerian Agama [9]. Soal yang terlalu mudah atau tidak mampu membedakan antar peserta didik akan menghasilkan data evaluasi yang tidak mewakili secara baik, sehingga keputusan mengenai langkah pembelajaran yang diambil dari data tersebut, seperti menentukan peserta didik yang perlu mengikuti remedial serta materi apa yang perlu diulang, berisiko tidak tepat sasaran.

Sistem ujian digital sebenarnya sudah banyak digunakan oleh sekolah untuk menggantikan ujian di kertas, tapi sebagian besar hanya berfungsi sebagai alat administrasi create-read-update-delete (CRUD) [10] tanpa kemampuan untuk menganalisis kualitas soal atau memetakan kelemahan kognitif siswa secara otomatis. Akibatnya, program remedial yang diberikan lebih cenderung bersifat merata, mengulang semua materi tanpa memperhatikan tingkat kemampuan kognitif siswa yang sebenarnya masih belum dikuasai [11]. Kesenjangan ini jadi awal dari penelitian ini diperlukan satu platform yang bisa mengklasifikasikan tingkat kognitif soal secara otomatis, mengevaluasi kualitas soal berdasarkan CTT, serta memberi saran materi remedial yang disesuaikan dengan kebutuhan setiap siswa. Untuk memenuhi kebutuhan penelitian ini membuat dan mengembangkan sistem ujian online yang diberi nama Aksara E-Exam di MTs Al-Hapidhiyah.

Klasifikasi tingkat kognitif C1 hingga C6 dilakukan dengan mencocokkan kata kerja operasional Bloom pada teks soal, sedangkan pengelompokan soal ke dalam kategori HOTS atau LOTS menggunakan algoritma K-Nearest Neighbor (KNN) dengan representasi TF-IDF, sesuai dengan pendekatan yang sudah terbukti efektif dalam berbagai tugas klasifikasi teks [12] [13]. Beberapa penelitian sebelumnya sudah menggunakan KNN untuk mengklasifikasikan soal seperti Kurniawan dkk, yang membandingkan KNN dengan Random Forest dalam klasifikasi soal multilabel [14] atau Adam, Suhartono, dan Firmansyah yang mengklasifikasikan soal Bahasa Indonesia berdasarkan Taksonomi Bloom [15]. Namun, keduanya hanya berhenti di tahap klasifikasi dan belum menghubungkannya dengan analisis soal serta rekomendasi perbaikan dalam satu sistem yang berjalan secara real-time.

Berdasarkan uraian di atas, kesenjangan penelitian (research gap) yang mendasari penelitian ini dapat dirangkum dari empat kelompok penelitian terdahulu. Pertama, penelitian distribusi kognitif soal seperti Yandra dkk. [4] dan Cressa dan Mukhlis [5] hanya sebatas memetakan proporsi soal HOTS/LOTS secara deskriptif tanpa menawarkan mekanisme otomatis untuk mengklasifikasikan soal baru maupun memperbaiki kualitasnya. Kedua, penelitian analisis butir soal berbasis CTT seperti Izzah dkk. [8] berhenti pada pelaporan indeks kesukaran dan daya pembeda tanpa menghubungkannya dengan tingkat kognitif Taksonomi Bloom maupun tindak lanjut pembelajaran. Ketiga, penelitian klasifikasi soal berbasis KNN seperti Kurniawan dkk. [14] dan Adam dkk. [15] hanya berhenti pada tahap klasifikasi HOTS/LOTS tanpa mengaitkannya dengan evaluasi psikometrik maupun rekomendasi remedial. Keempat, penelitian sistem rekomendasi content-based filtering seperti Ardiansyah dkk. [16] dan Astuti dan Andriyani [17] diterapkan pada domain di luar penilaian kognitif siswa, yaitu rekomendasi buku dan pembimbing tugas akhir, sehingga pendekatan tersebut belum pernah dihubungkan dengan profil kelemahan kognitif hasil ujian. Dengan demikian, belum ditemukan penelitian yang menyatukan klasifikasi kognitif otomatis, evaluasi psikometrik CTT, dan rekomendasi remedial berbasis kemiripan profil siswa ke dalam satu sistem ujian yang berjalan secara real-time, khususnya pada konteks madrasah yang mengikuti kurikulum khusus di bawah Kementerian Agama. Kesenjangan inilah yang berusaha dijawab oleh penelitian ini.

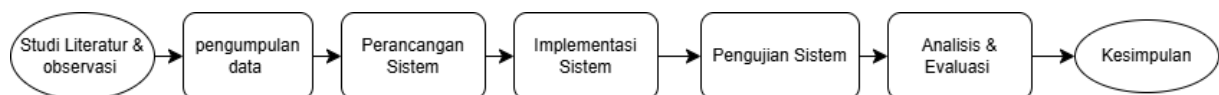
Rekomendasi materi remedial dalam penelitian ini menggunakan pendekatan content-based filtering yang menghubungkan profil kelemahan kognitif siswa dengan atribut materi remedial melalui cosine similarity, mengikuti pola yang digunakan oleh Ardiansyah dkk, dalam sistem rekomendasi buku perpustakaan [16] serta Astuti dan Andriyani dalam sistem rekomendasi pembimbing tugas akhir [17]. Kebaruan penelitian ini terletak pada penyatuan tiga bagian, yaitu pengklasifikasian otomatis berdasarkan kognitif, analisis CTT, dan rekomendasi remedial berbasis content-based filtering ke dalam satu proses yang berjalan sekaligus ketika guru memasukkan

soal atau saat siswa mengerjakan ujian. Selain itu, penelitian ini juga menggunakan data dari tiga mata pelajaran yaitu Bahasa Indonesia, Akidah Akhlak, dan IPS, yang berasal dari konteks madrasah, hal ini cukup langka dalam penelitian sejenis sebelumnya. Sistem dikembangkan menggunakan metode Software Development Life Cycle berbasis Waterfall [18] dan performanya dicek melalui confusion matrix untuk model klasifikasi serta uji fungsional black-box untuk memastikan semua fitur sistem berjalan dengan baik. Secara ringkas, kontribusi utama penelitian ini adalah: (1) menghasilkan platform ujian berbasis web yang menyatukan klasifikasi tingkat kognitif otomatis (KKO), pengelompokan HOTS/LOTS berbasis KNN, analisis psikometrik CTT, dan rekomendasi remedial berbasis content-based filtering ke dalam satu alur kerja yang berjalan secara real-time; (2) menyediakan bukti empiris penerapan pendekatan tersebut pada data soal madrasah dari tiga mata pelajaran (Bahasa Indonesia, Akidah Akhlak, dan IPS), konteks yang masih jarang diteliti pada penelitian sejenis sebelumnya; dan (3) menunjukkan bahwa hasil analisis kognitif dan psikometrik dapat ditampilkan secara terintegrasi kepada guru untuk mendukung pengambilan keputusan remedial yang lebih tepat sasaran dibandingkan sistem ujian digital konvensional yang hanya berfungsi sebagai alat administrasi. Bagian berikutnya dalam tulisan ini akan menjelaskan cara penelitian dilakukan, hasil uji coba sistem, serta kesimpulan yang diperoleh dari penelitian tersebut.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini memakai pendekatan kuantitatif eksperimental. Performa model klasifikasi diukur secara terukur menggunakan metrik seperti akurasi, presisi, recall, dan F1-score yang dihitung dari confusion matrix, sehingga kesimpulan yang dibuat didasarkan pada data numerik, bukan hanya pendapat subjektif. Pengembangan perangkat lunaknya menggunakan model Waterfall yang berjalan secara berurutan, dimulai dari analisis kebutuhan, perancangan, pembuatan, sampai pengujian, karena kebutuhan sistem dalam penelitian ini sudah jelas sejak awal.



Gambar 1. Tahapan Penelitian

2.2 Pengumpulan dan Pelabelan Data

Data dikumpulkan melalui dua cara yang saling melengkapi, yaitu dengan observasi langsung di MTs Al-Hapidhiyah, Kecamatan Bantargadung, Kabupaten Sukabumi, sesuai dengan kurikulum madrasah yang sedang berlaku pada tahun ajaran [19] dengan memahami cara penyusunan dan pelaksanaan ujian yang sedang dilakukan saat ini, serta melalui studi literatur terhadap jurnal yang terakreditasi dan bukubuku yang diterbitkan dalam lima tahun terakhir yang membahas tentang algoritma KNN, analisis CTT, sistem rekomendasi berbasis content-based filtering, dan pengembangan sistem informasi. Dataset utama terdiri dari 600 soal pilihan ganda, terbagi menjadi 300 soal kategori HOTS dan 300 soal kategori LOTS. Soal-soal tersebut diambil dari arsip soal Asesmen Sumatif Tengah Semester dan Asesmen Sumatif Akhir Semester di tiga mata pelajaran, yaitu Bahasa Indonesia, Akidah Akhlak, dan IPS. Setiap soal diberi label level kognitifnya (C1 hingga C6) secara manual oleh guru yang mengajar mata pelajaran tersebut, berdasarkan kata kerja operasional dari Taksonomi Bloom dan panduan penulisan soal HOTS yang dikeluarkan Kemendikbud [3]. Selanjutnya, soal tersebut divalidasi silang oleh satu guru untuk mengurangi kemungkinan bias subjektif. Dataset berikutnya dibagi menjadi data latihan yang berisi 80% atau 480 soal dan data uji yang berisi 20% atau 120 soal, digunakan untuk melatih serta mengevaluasi model KNN.

2.3 Arsitektur Sistem dan Algoritma

Modul pengklasifikasi kognitif dalam sistem terdiri dari dua bagian yang saling melengkapi. Bagian pertama mengklasifikasi level Taksonomi Bloom (C1-C6) dengan mencocokkan kata kunci KKO, sedangkan bagian kedua mengklasifikasi HOTS/LOTS menggunakan algoritma KNN. Saat guru memasukkan teks soal yang baru, sistem akan membandingkan katakata di dalam soal tersebut dengan kamus KKO Bloom, mulai dari tingkat tertinggi (C6) hingga tingkat terendah (C1). Level pertama yang tepat ditetapkan sebagai label Bloom untuk soal tersebut. Untuk mengklasifikasikan soal-soal HOTS/LOTS teks soal terlebih dahulu diubah menjadi vektor menggunakan metode TF-IDF dengan 500 fitur, lalu jarak antara vektor soal baru dengan semua vektor soal latihan dihitung dengan menggunakan jarak Euclidean, seperti yang terlihat pada Rumus (1).

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

Jarak antara dua titik x dan y dihitung dengan mengambil akar kuadrat dari jumlah kuadrat selisih setiap komponen mereka. Kelas yang paling banyak digunakan oleh tetangga terdekat kemudian digunakan untuk memprediksi kategori HOTS atau LOTS. Nilai k terbaik ditentukan dengan membandingkan hasil model pada lima jenis nilai k, yaitu 3, 5, 7, 9, dan 11, menggunakan data uji. Setelah ujian selesai, sistem menghitung dua

parameter psikometri berdasarkan teori uji klasik untuk setiap soal yang diberikan. Tingkat kesukaran (P) dihitung menurut Rumus (2).

$$P = \frac{B}{N} \quad (2)$$

P merupakan proporsi atau indeks tingkat kesukaran, B merupakan jumlah siswa yang menjawab benar, dan N merupakan jumlah peserta ujian. Dengan B adalah jumlah orang yang menjawab dengan benar dan N adalah jumlah total orang yang mengerjakan soal tersebut. Daya pembeda (D) dihitung menurut Persamaan (3).

$$D = \frac{BA}{NA} - \frac{BB}{NB} \quad (3)$$

D merupakan discrimination indeks atau indeks daya pembeda, BA merupakan jumlah kelompok atas yang menjawab benar, BB merupakan jumlah kelompok bawah yang menjawab benar, sedangkan NA dan NB masing-masing merupakan jumlah anggota kelompok atas dan kelompok bawah.

BA dan NA masing-masing adalah jumlah jawaban benar dan jumlah peserta di kelompok atas, sedangkan BB dan NB adalah jumlah jawaban benar dan jumlah peserta di kelompok bawah. Kelompok atas dan bawah masing-masing diambil dari 27% skor tertinggi dan 27% skor terendah sesuai dengan rekomendasi Arikunto. Rekomendasi materi remedial dihasilkan dengan cara pendekatan content-based filtering yang relevan. Profil kelemahan kognitif setiap siswa ditunjukkan sebagai vektor berdasarkan tingkat Bloom yang belum dikuasainya, lalu tingkat kesamaannya dengan vektor atribut setiap materi remedial dihitung menggunakan cosine similarity.

Materi yang memiliki nilai kemirisan minimal 0,70 terhadap profil siswa sebaiknya menjadi prioritas utama. Dari segi penerapan, sistem dibuat menggunakan PHP native dan MySQL untuk bagian aplikasi web, sedangkan modul KKO dan KNN diimplementasikan dengan Python (Scikit-learn) yang diakses sebagai REST API melalui Flask, sehingga proses pengklasifikasian bisa dijalankan secara realtime dari antarmuka guru.

2.4 Skenario Pengujian

Performa model KNN diuji dengan dua cara. Pertama, data uji (20% dari dataset) dinilai menggunakan confusion matrix untuk menghitung nilai accuracy, precision, recall, dan F1-score sesuai dengan Persamaan (5) sampai (8).

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{Recall}}{\text{precision} + \text{Recall}} \quad (7)$$

True Positif (TP) merupakan jumlah soal HOTS yang diprediksi benar sebagai HOTS, True Negatif (TN) merupakan jumlah soal LOTS yang diprediksi benar sebagai LOTS, False Positif (FP) merupakan jumlah soal LOTS yang keliru diprediksi sebagai HOTS, sedangkan False Negatif (FN) merupakan jumlah soal HOTS yang keliru diprediksi sebagai LOTS. Perhitungan keempat metrik tersebut didasarkan pada susunan confusion matrix yang disajikan pada Tabel 1 berikut.

Tabel 1. Confusion matrix klasifikasi

	Prediksi: HOTS	Prediksi: LOTS
Aktual: HOTS	TP (True Positive)	FN (False Negative)
Aktual: LOTS	FP (False Positive)	TN (True Negative)

Sistem diimplementasikan dengan PHP native dan MySQL untuk lapisan aplikasi dan basis data, sementara modul KKO, KNN, dan CTT diimplementasikan dengan Python (scikit-learn) yang diekspose sebagai REST API melalui Flask, dan dijalankan pada server lokal berbasis XAMPP. Pengujian sistem terdiri atas dua jenis. Pertama, pengujian kuantitatif terhadap model KNN menggunakan confusion matrix untuk menghitung accuracy, precision, recall, dan F1-score pada data uji, dengan target minimal akurasi 70%, serta validasi tambahan menggunakan 10-fold cross validation. Kedua, pengujian fungsional black-box menggunakan teknik equivalence partitioning dan boundary value analysis untuk memverifikasi bahwa setiap kebutuhan fungsional sistem (tujuh kebutuhan fungsional utama) berjalan sebagaimana mestinya, dengan melibatkan guru dan siswa MTs Al-Hapidhiyah sebagai responden pengujian.

3. HASIL DAN PEMBAHASAN

3.1 Implementasi Sistem

Sistem Aksara E-Exam diimplementasikan pada lingkungan server lokal berbasis XAMPP dengan konfigurasi Apache sebagai web server dan MYSQL sebagai sistem manajemen basis data. Bahasa pemrograman yang digunakan adalah PHP 8.x untuk back-end, MYSQL untuk menyimpan data, serta HTML, CSS, dan JavaScript untuk front-end antarmuka pengguna. Sistem diakses melalui browser pada alamat localhost/Aksara_Exam.

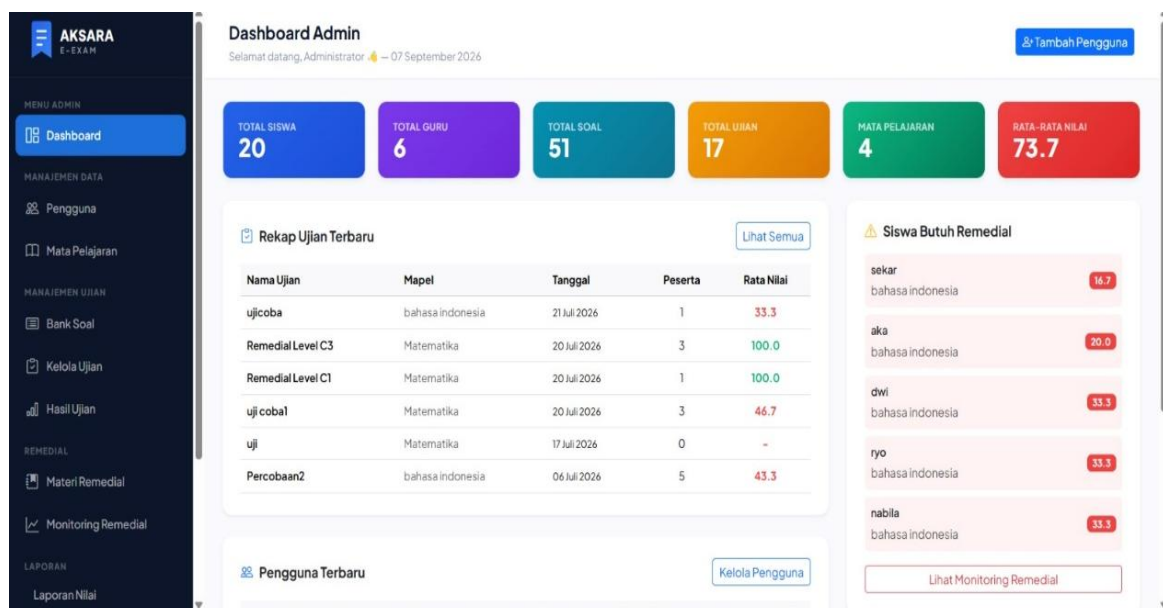
Sistem terdiri dari tiga modul utama yang berjalan secara terintegrasi. Pertama, modul keyword matching kko yang mengklasifikasikan level kognitif taksonomi bloom (C1-C6) secara real-time saat guru menginput soal. Kedua, modul KNN ($k=11$) yang mengklasifikasikan soal kedalam kategori HOTS atau LOTS berdasarkan representasi TF-IDF. Ketiga, modul ctt yang menghitung tingkat kesukaran dan daya pembeda secara otomatis setelah sesi ujian selesai.

3.1 Implementasi Antarmuka Pengguna

Implementasi antarmuka pengguna terdiri dari 3 halaman sesuai peran pengguna, yaitu halaman admin, halaman guru dan halaman siswa. Seluruh antarmuka dibangun dengan prinsip role-based access control sehingga setiap pengguna dapat mengakses fitur sesuai dengan perannya masing-masing.

3.1.1 Halaman Dashboard Admin

Halaman dashboard Administrator menampilkan statistik global sistem secara real-time. Berdasarkan data sistem per tanggal 30 Juni 2026, dashboard admin menampilkan total 20 siswa, 5 guru, 52 soal, 6 ujian, 5 mata pelajaran, dan rata-rata nilai siswa sebesar 81,0. Panel "Siswa Butuh Remedial" menampilkan daftar siswa dengan nilai di bawah KKM secara otomatis. Administrator juga dapat mengakses rekap ujian terbaru dan mengelola seluruh akun pengguna. Tampilan halaman dashboard administrator secara keseluruhan disajikan pada Gambar 2.



Gambar 2. Halaman Admin

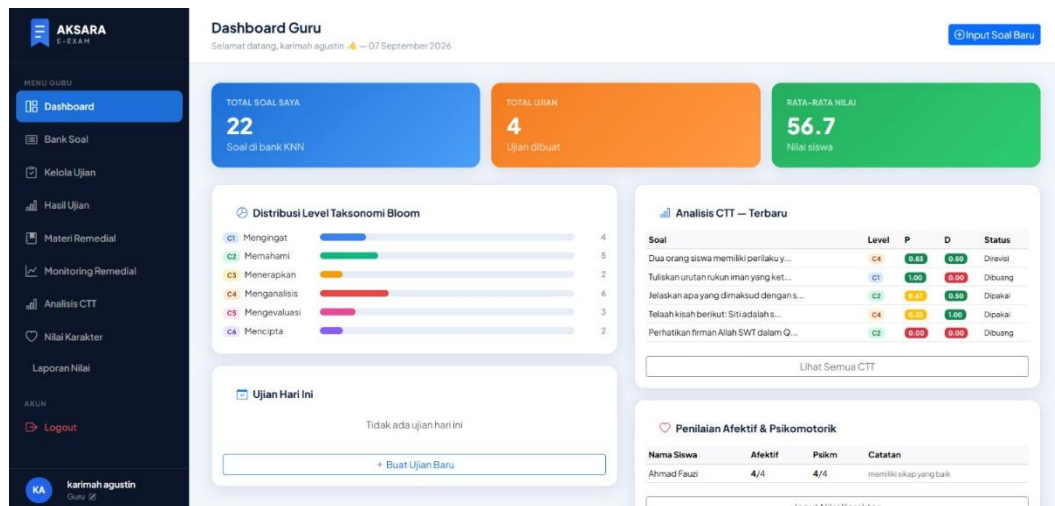
3.1.2 Halaman Dashboard Guru

Dashboard guru menampilkan ringkasan total soal yang dimiliki guru bersangkutan (31 soal), jumlah sesi ujian yang telah dibuat (2 ujian), serta distribusi level Taksonomi Bloom dalam bentuk diagram batang. Panel Analisis CTT terbaru dan panel penilaian karakter siswa juga tersedia pada halaman ini. Tabel 2 menyajikan distribusi level Bloom dari 31 soal yang tersimpan. Distribusi level Bloom dari 31 soal yang tersimpan disajikan pada Tabel 2.

Tabel 2. Distribusi level taksonomi bloom dari 31 soal

Level	Nama Level	Jumlah Soal	Kategori
C1	Mengingat	7	LOTS
C2	Memahami	11	LOTS
C3	Menerapkan	2	LOTS
C4	Menganalisis	6	HOTS
C5	Mengevaluasi	3	HOTS
C6	Mencipta	2	HOTS

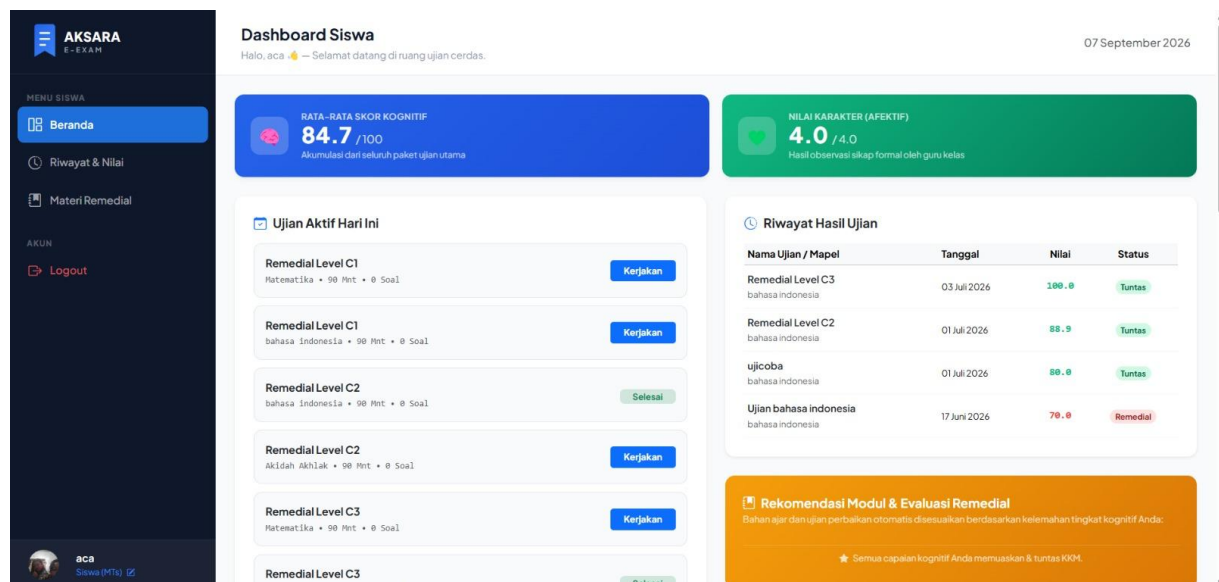
Tabel 2 menunjukkan bahwa soal dengan kategori LOTS mendominasi bank soal, yaitu sebanyak 20 soal yang terdiri atas level C1 (Mengingat) sebanyak 7 soal, C2 (Memahami) sebanyak 11 soal, dan C3 (Menerapkan) sebanyak 2 soal. Sementara itu, soal dengan kategori HOTS berjumlah 11 soal, meliputi level C4 (Menganalisis) sebanyak 6 soal, C5 (Mengevaluasi) sebanyak 3 soal, dan C6 (Mencipta) sebanyak 2 soal. Komposisi tersebut menggambarkan proporsi soal HOTS yang masih lebih rendah dibandingkan soal LOTS pada bank soal yang tersimpan. Tampilan halaman dashboard guru secara keseluruhan, termasuk diagram distribusi level Bloom dan panel Analisis CTT, disajikan pada Gambar 3.



Gambar 3. Halaman Guru

3.1.3 Halaman Dashboard Siswa

Dashboard siswa menampilkan rata-rata skor kognitif, daftar ujian aktif yang dapat dikerjakan, riwayat hasil ujian, dan panel rekomendasi materi remedial berbasis Content-Based Filtering. Ketika nilai siswa berada di bawah KKM (75), sistem menampilkan notifikasi pendampingan remedial secara otomatis beserta materi yang direkomendasikan berdasarkan kelemahan kognitif per level Bloom. Tampilan halaman dashboard siswa secara keseluruhan disajikan pada Gambar 4.



Gambar 4. Halaman Siswa

3.2 Evaluasi Akurasi Model KNN

3.2.1 Dataset dan Pembagian Data

Model KNN dilatih dengan menggunakan 600 dataset soal (300 HOTS dan 300 LOTS) dari tiga mata pelajaran dengan rasio pembagian data 80:20. Sebanyak 480 soal (80%) digunakan sebagai data latih dan 120 soal (20%) sebagai data uji. Representasi teks soal dilakukan menggunakan TF-IDF Vectorizer dengan 500 fitur. Berikut disajikan simulasi perhitungan manual untuk salah satu soal dari 120 data uji. Misalkan soal uji berbunyi “Peserta

didik mampu menganalisis dampak perubahan iklim terhadap kehidupan sosial masyarakat sekitar” dengan label aktual HOTS (C4 –menganalisis). Untuk mempermudah perhitungan manual, representasi TF-IDF disederhanakan menjadi 4 fitur kata kunci yang paling relevan, yaitu “menganalisis”, “iklim”, “menjelaskan”, dan “menyebutkan”, sehingga diperoleh vektor TF-IDF soal uji sebagai berikut:

$$X_{uji} = [0,48 ; 0,10 ; 0,00 ; 0,00]$$

Jarak Euclidean antara data uji dengan tiga data latih dihitung menggunakan Persamaan (3) sebagai berikut. Data latih A berlabel HOTS (soal “menganalisis penyebab terjadinya perubahan iklim global”) memiliki vektor $XA = [0,45 ; 0,08 ; 0,00 ; 0,02]$, data latih D berlabel HOTS (soal “menganalisis dampak pemanasan global bagi ekosistem”) memiliki vektor $XD = [0,40 ; 0,12 ; 0,00 ; 0,05]$, dan data latih B berlabel LOTS (soal “menjelaskan pengertian iklim dan cuaca”) memiliki vektor $XB = [0,00 ; 0,15 ; 0,40 ; 0,05]$.

$$d(uji,A) = \sqrt{[(0,48-0,45)^2+(0,10-0,08)^2+(0-0)^2+(0-0,02)^2]} = \sqrt{0,0017} = 0,041$$

$$d(uji,D) = \sqrt{[(0,48-0,40)^2+(0,10-0,12)^2+(0-0)^2+(0-0,05)^2]} = \sqrt{0,0093} = 0,096$$

$$d(uji,B) = \sqrt{[(0,48-0,00)^2+(0,10-0,15)^2+(0-0,40)^2+(0-0,05)^2]} = \sqrt{0,3954} = 0,629$$

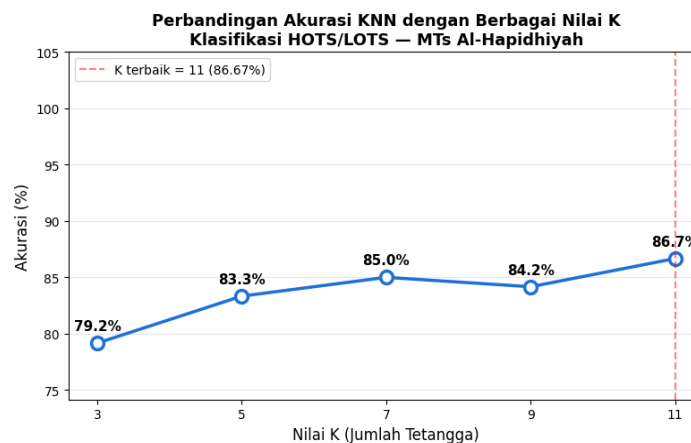
Ketiga jarak tersebut diurutkan dari yang terkecil, yaitu A (0,041), D (0,096), dan B (0,629). Dengan mengambil $k=3$ tetangga terdekat sebagai ilustrasi, diperoleh mayoritas label HOTS (2 dari 3 tetangga, yaitu A dan D), sehingga soal uji tersebut diklasifikasikan sebagai HOTS, sesuai dengan label aktualnya. Perlu dicatat bahwa sistem yang sesungguhnya menggunakan $k=11$ dengan representasi TF-IDF berdimensi 500 fitur.

3.2.2 Penentuan Nilai K Optimal

Tabel 3 dan Gambar 5 menyajikan perbandingan akurasi model KNN pada lima variasi nilai k.

Tabel 3. Perbandingan akurasi knn pada berbagai nilai k

Nilai K	Akurasi
K = 3	79,17 %
K = 5	83,33 %
K = 7	85,00 %
K = 9	84,17 %
K = 11 (terbaik)	86,67 %



Gambar 5. Penentuan Nilai K

Akurasi terbaik dicapai saat $k=11$ yaitu 86,67%, sehingga nilai ini dipilih sebagai k optimal untuk model akhir. Pola ini sesuai dengan sifat algoritma KNN umumnya, yaitu jika nilai k terlalu kecil, model mudah terpengaruh oleh gangguan pada data latih, sedangkan nilai k terlalu besar bisa menyembunyikan pola lokal yang sebenarnya membedakan kedua kelas [20].

3.2.3 Evaluasi Klasifikasi HOTS/LOTS

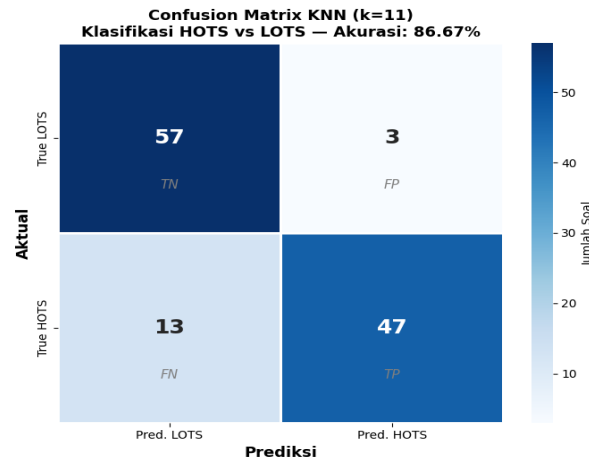
Model KNN final ($k=11$) diuji pada 120 data uji yang terdiri dari 60 soal HOTS dan 60 soal LOTS. Hasilnya disajikan pada Tabel 4 dan Gambar 6.

Tabel 4. Classification report knn

Kelas	Precision	Recall	F1-score	Support
LOTS	0,94	0,78	0,85	60
HOTS	0,81	0,95	0,88	60

Macro Average	0,88	0,87	0,87	120
Akurasi Keseluruhan			86,67 %	

Model Mendapatkan akurasi keseluruhan sebesar 86,67%, precision rata-rata 87,71%, recall 86,67% dan f1-score 86,57% yang jauh melebihi batas minimal 70% yang ditentukan pada tahap perancangan kebutuhan non-fungsional. Recall untuk kategori HOTS 0,95 jauh lebih tinggi dibandingkan kategori LOTS 0,78 yang menunjukkan bahwa model jarang menganggao soal HOTS sebagai LOTS meskipun sebaliknya precision untuk LOTS lebih tinggi. Pola ini masuk akal karena soal HOTS biasanya menggunakan kata kerja operasional yang lebih khas, seperti menganalisis, mengevaluasi, atau merancang sehingga lebih mudah dibedakan melalui representasi TF-IDF dibandingkan dengan soal LOTS yang cenderung menggunakan kata-kata umum.



Gambar 6. Confusion Matrix

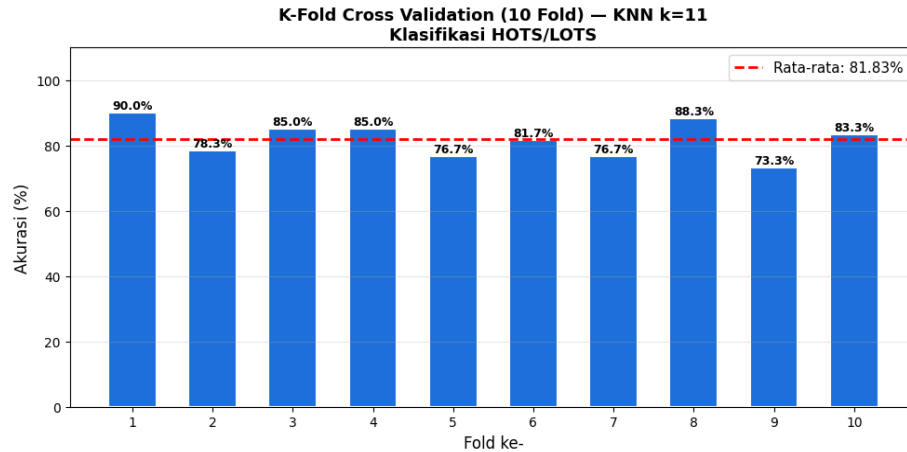
3.2.4 Validasi K-Fold Cross Validation

Untuk memastikan model tidak overfitting terhadap satu pembagian data digunakan metode validasi 10- fold cross validation. Hasilnya menunjukkan akurasi sebesar 81,83% dengan standar deviasi 5,19% yang lebih sedikit rendah dibanding akurasi pada skema split Tunggal 80:20. Penurunan biasa terjadi karena validasi silang menguji model pada berbagai bagian data secara bergantian bukan hanya pada satu bagian data uji saja sehingga hasilnya cenderung memberi perkiraan performa yang lebih hati-hati namun lebih mencerminkan kemampuan model dalam generalisasi. Hasil akurasi pada setiap fold disajikan pada Tabel 5.

Tabel 5. Hasil K-Fold Cross Validation

Fold	Akurasi (%)
Fold 1	90,00 %
Fold 2	78,33 %
Fold 3	85,00 %
Fold 4	85,00 %
Fold 5	76,67 %
Fold 6	81,67 %
Fold 7	76,67 %
Fold 8	88,33 %
Fold 9	73,33 %
Fold 10	83,33 %
Rata-rata	81,83 % ± 5,19%

Perbandingan akurasi antar fold beserta rata-ratanya secara visual ditunjukkan pada Gambar 7.



Gambar 7. K-Fold Cross Validation

3.3 Evaluasi Analisis Butir Soal berbasis CTT

Uji coba analisis CTT dilakukan pada instrument soal akidah akhlak kelas 9 yang terdiri dari 30 soal pilihan ganda yang diberikan kepada 38 siswa. Kelompok atas dan kelompok bawah masing-masing diambil dari 27% skor tertinggi dan terendah [21]. Disajikan pada Tabel 6 dan Tabel 7

Tabel 6. Rekapitulasi tingkat kesukaran (P)

Kategori	Jumlah	Persentase
Mudah ($P > 0,70$)	26	86,67 %
Sedang ($0,30 \leq P \leq 0,70$)	4	13,33 %
Sukar ($P < 0,30$)	0	0%

Tabel 7. Rekapitulasi status soal berdasarkan daya pembeda (D)

Status	Jumlah	Persentase
Diterima	17	56,67 %
Direvisi	7	23,33 %
Dibuang	6	20,00 %

Dominasi soal dengan kategori mudah sebesar 86,67% menunjukkan bahwa instrument yang biasa digunakan guru masih perlu diperkaya dengan soal yang memiliki tingkat kesukaran sedang sehingga sukar agar dapat membedakan kemampuan siswa secara jelas. Dari segi daya pembeda, lebih dari separuh soal 56,67% dianggap layak dipertahankan sementara 20% dianggap perlu di buang karena kemampuan membedakan terlalu rendah untuk digunakan mengukur perbedaan kemampuan siswa[22]. Dengan menampilkan hasil CTT di dashboard guru system memberi guru pijakan untuk memperbaiki kualitas instrument soal. Dan akan sulit jika perhitungan CTT dilakukan secara manual satu per satu.

3.4 Pengaruh Fitur KKO Terhadap Akurasi KNN

Sebagai pembandingan KNN diuji juga pada dataset yang tidak menggunakan KKO hanya menggunakan representasi TF-IDF dari teks soal mentah berjumlah 1.065 soal dengan jarak $K=3$. Perbandingannya di Tabel 8.

Tabel 8. Perbandingan KNN dengan tanpa fitur KKO

Metrik	KNN+KKO (k=11)	KNN tanpa KKO (k=3)
Akurasi	86,67 %	93,43 %
Precision	87,71 %	93,55 %
Recall	86,67 %	93,43 %
F1- Score	86,57 %	93,41 %
K-Fold rata-rata	81,83 %	93,90 %

Skenario tanpa menggunakan kko memang menunjukkan akurasi dan kestabilan yang lebih baik, tapi perbedaan ini harus dilihat dengan hati hati karena jumlah data pada kedua scenario tidak sama. Jadi, Sebagian besar perbedaan akurasi kemungkinan disebabkan oleh perbedaan data latih bukan karena ada atau tidak adanya kko. Yang perlu diperhatikan bahwa kko dan knn memiliki fungsi yang berbeda namun saling melengkapi. KKO digunakan untuk menentukan Tingkat kognitif dari C1-C6 dan KNN berfungsi meklasifikasikan HOTS/LOTS sehingga keduanya tetap bisa digunakan bersamaan[23].

3.5 Pengujian Fungsional Sistem

Pengujian black-box dilakukan pada 23 skenario yang meliputi modul autentikasi, bank soal, pengelolaan ujian, proses mengerjakan ujian oleh siswa, analisis CTT, serta input nilai karakter. Semua scenario dinyatakan berhasil termasuk deteksi label bloom secara langsung saat guru mengetik soal dengan KKO seperti kata “sebutkan” yang terdeteksi sebagai C1, “analisis” sebagai C4, dan “rancang” sebagai C6. Selain itu, sistem juga melakukan validasi input soal serta menghitung nilai p dan d secara otomatis setelah ujian selesai. Hasil menunjukkan bahwa kebutuhan fungsional yang sudah dibuat saat perancangan berjalan lancar dan sistem siap digunakan dalam lingkungan kerja dimadrasah.

3.6 Pembahasan

Secara keseluruhan, hasil-hasil di atas menunjukkan bahwa pendekatan campuran yang menggabungkan metode berbasis aturan (Keyword Matching KKO) yang jelas dan mudah dijelaskan kepada guru dengan model machine learning yang sederhana (KNN) yang akurat untuk klasifikasi dua kategori, terbukti bisa diterapkan di lingkungan sekolah dengan sumber daya terbatas tanpa membutuhkan perangkat keras komputasi yang besar. Model KNN dengan nilai $k=11$ mencapai tingkat akurasi sebesar 86,67%, yang jauh melebihi target minimal 70% yang ditentukan dalam tahap analisis kebutuhan. Rata-rata akurasi pada validasi 10-fold cross validation adalah 81,83%, yang menunjukkan bahwa model tidak mengalami overfitting. Dibandingkan dengan penelitian Kurniawan dkk. yang mengklasifikasikan soal dengan cara multi-label menggunakan gabungan KNN dan Random Forest serta mendapatkan akurasi sebesar 66,18% [14] hasil penelitian ini menunjukkan performa yang lebih baik. Namun, perbedaan dalam skema klasifikasi (multi-label pada penelitian sebelumnya dan biner HOTS/LOTS pada penelitian ini) juga memengaruhi tingkat kesulitan dalam tugas klasifikasi, sehingga perbandingan ini perlu dilihat secara seimbang. Pada bagian analisis butir soal, hasil CTT pada instrumen Akidah Akhlak kelas IX menunjukkan bahwa 86,67% soal dianggap mudah, sedangkan hanya 56,67% soal memiliki daya pembeda yang memadai. Temuan ini memperkuat hasil penelitian Izzah dkk. yang juga menemukan bahwa sebagian besar soal Bahasa Indonesia di tingkat dasar memiliki tingkat kesukaran yang rendah dan daya pembeda yang belum cukup baik [8]. Hal ini menunjukkan bahwa kecenderungan membuat soal yang mudah tidak hanya terjadi pada satu mata pelajaran atau satu tingkat pendidikan, melainkan mungkin menjadi pola umum dalam instrumen ujian sekolah. Yang membedakan penelitian ini dari penelitian CTT sebelumnya adalah hasilnya tidak hanya berupa laporan statis, melainkan langsung dibandingkan dengan label Bloom dari KKO dan kategori HOTS/LOTS dari KNN dalam satu tampilan. Dengan demikian, guru bisa mendapat petunjuk awal untuk meninjau kembali soal-soal yang secara teks menggunakan kata kerja operasional level tinggi, tetapi ternyata memiliki tingkat kesukaran yang rendah. Ini merupakan jenis pemeriksaan silang yang sulit dilakukan secara manual pada sistem ujian konvensional yang hanya berfungsi sebagai alat administrasi untuk menginput, menyimpan, mengubah, dan menghapus data [10]. Dari segi fungsional, uji black-box pada 23 skenario menunjukkan hasil yang berhasil 100%, yang membuktikan bahwa ketujuh kebutuhan fungsional sistem sudah terpenuhi. Keputusan menampilkan badge Bloom dan HOTS/LOTS secara langsung di halaman Tambah Soal memberikan manfaat nyata, guru bisa langsung melihat seberapa banyak soal yang termasuk level rendah dan tinggi sebelum soal disimpan, tanpa perlu menunggu hasil penilaian di akhir semester. Selain itu, notifikasi remedial otomatis yang diberikan ke siswa juga membantu mempercepat tindak lanjut pembelajaran dibandingkan metode yang sering digunakan sebelumnya, yaitu pendekatan satu ukuran untuk semua.

Meskipun demikian, beberapa keterbatasan perlu dicermati. Pertama, uji coba CTT baru dilakukan pada satu mata pelajaran dengan jumlah siswa yang tidak terlalu banyak (38 siswa), sehingga hasilnya perlu diperhatikan secara hati-hati sebelum diterapkan ke mata pelajaran lain. Kedua, belum dilakukan perbandingan akurasi metode KNN dengan dan tanpa fitur KKO menggunakan dataset yang ukurannya sama, sehingga kontribusi sebenarnya dari fitur KKO terhadap akurasi klasifikasi masih bersifat menunjukkan saja. Ketiga, penentuan level kognitif dalam dataset pelatihan sepenuhnya didasarkan pada penilaian manual oleh guru sebagai ahli di bidangnya, sehingga faktor subjektivitas pada tahap pelabelan awal masih bisa memengaruhi batas maksimal akurasi yang berhasil dicapai oleh model. Secara umum, menggabungkan klasifikasi kognitif, analisis CTT, dan rekomendasi remedial berdasarkan Content-Based Filtering dalam satu platform ternyata bisa memberikan dasar data yang baik bagi guru dalam menyusun kebijakan remedial yang lebih tepat sasaran. Selain itu, hal ini juga memberi kesempatan untuk melanjutkan penelitian dengan memperbesar dataset dan memperluas mata pelajaran yang dievaluasi menggunakan CTT.

4. KESIMPULAN

Penelitian ini menghasilkan sistem ujian berbasis web bernama Aksara E-Exam yang menyatukan pengklasifikasian kognitif otomatis, analisis kualitas butir soal, dan pemberian saran remedial dalam satu platform yang dipakai langsung oleh guru MTs Al-Hapidhiyah. Kontribusi penelitian ini adalah membuktikan bahwa klasifikasi kognitif berbasis KKO dan KNN, analisis psikometrik CTT, dan rekomendasi remedial berbasis content-based filtering dapat disatukan ke dalam satu sistem ujian yang berjalan secara real-time pada konteks madrasah, sebuah kombinasi yang belum ditawarkan oleh penelitian-penelitian terdahulu yang cenderung

membahas ketiga aspek tersebut secara terpisah. Penandaan level Bloom (C1-C6) dilakukan dengan mencocokkan kata kerja operasional pada teks soal begitu guru memasukkannya ke sistem, sedangkan pengelompokan HOTS/LOTS memakai algoritma KNN dengan $k=11$ yang mencapai akurasi 86,67%, precision 87,71%, recall 86,67%, dan F1-score 86,57%, dan angka ini dikonfirmasi lewat validasi 10-fold cross validation sebesar $81,83\% \pm 5,19\%$. Pengujian tambahan tanpa fitur KKO memang menghasilkan akurasi lebih tinggi yaitu 93,43%, tetapi selisih ini lebih disebabkan oleh perbedaan jumlah data pelatihan dibanding peran fitur KKO itu sendiri, sehingga kedua pendekatan tetap dianggap saling melengkapi dan bukan saling menggantikan. Dari sisi analisis CTT, evaluasi terhadap butir soal Akidah Akhlak menunjukkan 86,67% soal berada pada kategori mudah, 56,67% soal memiliki daya beda yang layak dipertahankan, dan 20% soal disarankan untuk direvisi atau dibuang karena daya bedanya rendah, temuan yang mempertegas bahwa sebagian besar bank soal di madrasah ini masih perlu perbaikan dari sisi tingkat kesukaran maupun kemampuan membedakan siswa berkemampuan tinggi dan rendah. Pengujian black-box terhadap 23 skenario fungsional seluruhnya berjalan sesuai rancangan, menandakan sistem sudah layak dipakai pada kondisi operasional sehari-hari. Secara praktis, platform ini memberi guru madrasah alat bantu yang mempersingkat waktu penelaahan soal secara manual, sekaligus menyediakan data objektif mengenai soal mana yang perlu direvisi dan siswa mana yang perlu mengikuti remedial pada topik tertentu, sehingga keputusan pembelajaran lanjutan tidak lagi hanya mengandalkan perkiraan guru. Kombinasi klasifikasi kognitif, analisis CTT, dan rekomendasi remedial berbasis content-based filtering ini juga bisa dipertimbangkan sekolah lain yang ingin beralih dari ujian digital yang sekadar administratif menuju sistem yang benar-benar membantu perbaikan instrumen pengajaran. Untuk pengembangan selanjutnya, penelitian berikutnya disarankan memperbesar ukuran dataset dan cakupan mata pelajaran, menjajal model klasifikasi berbasis contextual embedding seperti BERT atau IndoBERT sebagai pembanding representasi TF-IDF yang dipakai saat ini, serta melakukan uji penerimaan pengguna secara resmi yang melibatkan lebih banyak guru dan siswa agar kelayakan sistem ini punya bukti empiris yang lebih kuat.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: Karimah Agustin, Alun Sujjada, dan M. Ikhsan Thohir;

Metodologi: Karimah Agustin, Alun Sujjada, dan M. Ikhsan Thohir;

Perangkat Lunak: Karimah Agustin;

Validasi: Karimah Agustin, Alun Sujjada, dan M. Ikhsan Thohir;

Analisis Formal: Karimah Agustin, Alun Sujjada, dan M. Ikhsan Thohir;

Investigasi: Karimah Agustin, Alun Sujjada, dan M. Ikhsan Thohir;

Sumber Daya: Karimah Agustin, Alun Sujjada, dan M. Ikhsan Thohir;

Kurasi Data: Karimah Agustin, Alun Sujjada, dan M. Ikhsan Thohir;

Penulisan Draf Awal: Karimah Agustin, Alun Sujjada, dan M. Ikhsan Thohir;

Penulisan, Peninjauan, dan Penyuntingan: Karimah Agustin, Alun Sujjada, dan M. Ikhsan Thohir;

Visualisasi: Karimah Agustin, Alun Sujjada, dan M. Ikhsan Thohir;

Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis pertama.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan eksternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

- [1] K. R. dan T. R. I. Kementerian Pendidikan, “Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 21 Tahun 2022 tentang Standar Penilaian Pendidikan pada Pendidikan Anak Usia Dini, Jenjang Pendidikan Dasar, dan Jenjang Pendidikan Menengah,” Jakarta, 2022.
- [2] Suyit Ratno et al., “Analisis Penerapan Taksonomi Bloom dalam Pembelajaran Siswa Kelas VI SD Negeri 105293 Medan Estate,” *Sinar Dunia: Jurnal Riset Sosial Humaniora dan Ilmu Pendidikan*, vol. 3, no. 4, pp. 127–136, Dec. 2024, doi: 10.58192/sidu.v3i4.2726.
- [3] Pusat Kurikulum dan Pembelajaran, “Panduan Pembelajaran dan Asesmen Kurikulum Merdeka,” Jakarta, 2022.
- [4] B. R. Yandra, F. Meliasanti, and D. Huri, “Tinjauan Tipe HOTS dan LOTS dalam Soal ASTS serta ASAS Bahasa Indonesia Tahun 2022–2024,” *Jurnal Ilmiah FONEMA*, vol. 7, no. 1, pp. 61–77, Aug. 2024, doi: 10.25139/fn.v7i1.8633.
- [5] J. Cressa and M. Mukhlis, “Level Kognitif Taksonomi Bloom pada Soal Mata Pelajaran Bahasa Indonesia,” *J-LELC: Journal of Language Education, Linguistics, and Culture*, vol. 3, no. 1, pp. 55–62, Feb. 2023, doi: 10.25299/j-lelc.2023.12094.
- [6] F. Y. Hermanto, A. D. Anggraeni, and S. Sutirman, “Analisis Butir Soal dengan Classical Test Theory untuk Mengukur Kemampuan Kognitif Siswa,” *Efisiensi : Kajian Ilmu Administrasi*, vol. 20, no. 2, pp. 159–170, Jan. 2025, doi: 10.21831/efisiensi.v20i2.80225.
- [7] Mu’ah Sharoh Suprihatin and Fina Fakhriyah, “Analisis Butir Soal Materi Rantai Makanan Meliputi Validitas, Reliabilitas, Tingkat Kesukaran, dan Daya Beda,” *Jurnal Yudistira : Publikasi Riset Ilmu Pendidikan dan Bahasa*, vol. 4, no. 1, pp. 77–91, Jan. 2026, doi: 10.61132/yudistira.v4i1.2615.
- [8] A. N. Izzah, S. S. L. Azizah, R. Ekawati, and I. S. Rufiana, “Eksplorasi Analisis Butir Soal Bahasa Indonesia di Kelas II Sekolah Dasar Berdasarkan Taksonomi Bloom,” *Jurnal Pendidikan dan Pembelajaran Indonesia (JPPI)*, vol. 5, no. 2, pp. 1031–1043, Apr. 2025, doi: 10.53299/jppi.v5i2.1064.
- [9] Kementerian Agama Republik Indonesia, “Keputusan Menteri Agama Republik Indonesia Nomor 450 Tahun 2024 tentang Pedoman Implementasi Kurikulum pada Raudhatul Athfal, Madrasah Ibtidaiyah, Madrasah Tsanawiyah, Madrasah Aliyah, dan Madrasah Aliyah Kejuruan,” Jakarta, 2024.
- [10] H. Lahza, T. G. Smith, and H. Khosravi, “Beyond item analysis: Connecting student behaviour and performance using e-assessment logs,” *British Journal of Educational Technology*, vol. 54, no. 1, pp. 335–354, Jan. 2023, doi: 10.1111/bjet.13270.
- [11] N. I. Rosyida, S. W. Nugraha, U. Agustin, C. Rochman, and A. Y. Nuryantini, “OPTIMALISASI PROGRAM REMEDIAL BERDASARKAN INSTRUMEN AKREDITASI SATUAN PENDIDIKAN DALAM PEMBELAJARAN DIFERENSIASI KURIKULUM MERDEKA,” *Jurnal Evaluasi Pendidikan*, vol. 14, no. 1, pp. 12–16, Oct. 2023, doi: 10.21009/jep.v14i1.36763.
- [12] M. R. Arrohman, N. Khoiriyah, A. Fawaida, and D. L. Fithri, “ANALISIS KLASIFIKASI KELULUSAN SISWA MENGGUNAKAN METODE KNN (K-NEAREST NEIGHBOR) RANDOM FOREST PADA DATASET ‘STUDENTS PERFORMANCE.’” *Jurnal Mahasiswa Sistem Informasi (JMSI)*, vol. 6, no. 2, pp. 299–305, 2025, doi: 10.24127/jmsi.v6i2.9102.
- [13] V. Alvian, D. Hidayatullah, A. Nilogiri, H. Azizah, and A. Faruq, “Klasifikasi Siswa Berprestasi Menggunakan Metode K-Nearest Neighbor (KNN) Pada SMA Negeri 2 Situbondo Classification Of Achieving Students Using K-Nearest Neighbor (KNN) Method At SMA Negeri 2 Situbondo,” *Jurnal Smart Teknologi*, vol. 3, no. 6, 2022, <https://jurnal.unmuhjembar.ac.id/index.php/JST/article/view/8484/0>
- [14] M. R. Kurniawan, O. N. Pratiwi, and F. Hamami, “KLASIFIKASI SOAL MENGGUNAKAN MULTI-LABEL PROBLEM TRANSFORMATION DENGAN METODE RANDOM FOREST DAN K-NEAREST NEIGHBOR,” *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 1, pp. 367–380, Jan. 2025, doi: 10.29100/jipi.v10i1.5910.
- [15] B. P. Adam, C. Suhartono, and D. Firmansyah, “Klasifikasi Teks Soal Ujian Bahasa Indonesia Berdasarkan Taksonomi Bloom,” *Jurnal Informatika*, vol. 9, no. 1, p. 15–26, 2022.
- [16] R. Ardiansyah, M. Ari Bianto, and B. D. Saputra, “Sistem Rekomendasi Buku Perpustakaan Sekolah menggunakan Metode Content-Based Filtering,” *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 4, no. 2, pp. 510–518, Oct. 2023, doi: 10.37859/coscitech.v4i2.5131.
- [17] F. D. Astuti and W. Andriyani, “Pengembangan Sistem Rekomendasi Pembimbing Tugas Akhir Menggunakan Teknik Content Based Filtering,” *JIKO (Jurnal Informatika dan Komputer)*, vol. 9, no. 2, p. 474, Jun. 2025, doi: 10.26798/jiko.v9i2.1599.
- [18] N. Jannah, “ANALISIS PENGEMBANGAN SISTEM INFORMASI BATIK MADURA MENGGUNAKAN METODE WATERFALL DAN BLACK BOX TESTING,” *JURNAL SATYA INFORMATIKA*, vol. 8, no. 01, pp. 64–81, Jun. 2023, doi: 10.59134/jsk.v8i01.239.
- [19] Tim Penyusun Kurikulum MTs Al-Hapidhiyah, “Kurikulum Operasional Madrasah MTs Al-Hapidhiyah Tahun Ajaran 2025/2026,” 2025.
- [20] I. Listiowarni and N. Puspa Dewi, “Pemanfaatan Klasifikasi Soal Biologi Cognitive Domain Bloom’s Taxonomy Menggunakan KNN Chi-Square Sebagai Penyusunan Naskah Soal,” *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, vol. 11, no. 2, pp. 186–197, Nov. 2020, doi: 10.31849/digitalzone.v11i2.4798.
- [21] S. Nurjanah et al., “Psychometric quality of multiple-choice tests under classical test theory (CTT): AnBuso, Iteman, and R,” *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 28, no. 2, pp. 161–172, Dec. 2024, doi: 10.21831/pep.v28i2.71542.
- [22] R. SHEPTIAN, I. SARIFAH, and R. RIYADI, “CLASSICAL TEST THEORY ANALYSIS USING ANATES: A STUDY OF MATHEMATICS READINESS TEST FOR ELEMENTARY SCHOOL STUDENTS,” *SCIENCE : Jurnal Inovasi Pendidikan Matematika dan IPA*, vol. 5, no. 1, pp. 20–29, Feb. 2025, doi: 10.51878/science.v5i1.3863.



- [23] N. F. Munazhif, G. J. Yanris, and M. N. S. Hasibuan, “Implementation of the K-Nearest Neighbor (kNN) Method to Determine Outstanding Student Classes,” SinkrOn, vol. 8, no. 2, pp. 719–732, Apr. 2023, doi: 10.33395/sinkron.v8i2.12227.