

**ANALISIS ALGORITMA EXTREME GRADIENT BOOSTING
(XGBOOST) PADA METODE CONTENT-BASED FILTERING
UNTUK KLASIFIKASI RELEVANSI BUKU**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 - Informatika



disusun oleh

QUBUSSIMA FAEROZ FURQONA

22.11.5155

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

**ANALISIS ALGORITMA EXTREME GRADIENT BOOSTING
(XGBOOST) PADA METODE CONTENT-BASED FILTERING
UNTUK KLASIFIKASI RELEVANSI BUKU**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 - Informatika



disusun oleh

QUBUSSIMA FAEROZ FURQONA
22.11.5155

Kepada

FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

HALAMAN PERSETUJUAN

SKRIPSI

**ANALISIS ALGORITMA EXTREME GRADIENT BOOSTING
(XGBOOST) PADA METODE CONTENT-BASED FILTERING UNTUK
KLASIFIKASI RELEVANSI BUKU**

yang disusun dan diajukan oleh

Qubussima Faeroz Furqona

22.11.5155

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 21 Mei 2026

Dosen Pembimbing,



Ike Verawati, S.Kom., M.Kom.

NIK. 190302237

HALAMAN PENGESAHAN

SKRIPSI

**ANALISIS ALGORITMA EXTREME GRADIENT BOOSTING
(XGBOOST) PADA METODE CONTENT-BASED FILTERING UNTUK
KLASIFIKASI RELEVANSI BUKU**

yang disusun dan diajukan oleh

Qubussima Faeroz Furqona

22.11.5155

Telah dipertahankan di depan Dewan Penguji
pada tanggal 21 Mei 2026

Susunan Dewan Penguji

Nama Penguji

Arifivanto Hadinegoro, S.Kom., M.T.
NIK. 190302289

Theopilus Bayu Sasongko, S.Kom., M.Eng.
NIK. 190302375

Ike Verawati, S.Kom., M.Kom.
NIK. 190302237

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 21 Mei 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Qubussima Faeroz Furqona
NIM : 22.11.5155

Menyatakan bahwa Skripsi dengan judul berikut:

ANALISIS ALGORITMA EXTREME GRADIENT BOOSTING (XGBOOST) PADA METODE CONTENT-BASED FILTERING UNTUK KLASIFIKASI RELEVANSI BUKU

Dosen Pembimbing : Ike Verawati, S.Kom., M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 21 Mei 2026

Yang Menyatakan,



Qubussima Faeroz Furqona

HALAMAN PERSEMBAHAN

Skripsi ini dipersembahkan sebagai bentuk rasa syukur dan penghargaan penulis kepada pihak-pihak yang telah memberikan dukungan, doa, serta motivasi selama proses penyusunan skripsi ini.

Adapun skripsi ini dipersembahkan kepada:

1. Allah SWT atas segala limpahan rahmat, kesehatan, dan kemudahan yang diberikan kepada penulis dalam menyelesaikan skripsi ini.
2. Kedua orang tua tercinta yang senantiasa memberikan doa, kasih sayang, dukungan, serta pengorbanan yang tidak ternilai selama penulis menempuh pendidikan.
3. Dosen pembimbing skripsi yang telah memberikan arahan, bimbingan, dan masukan selama proses penyusunan skripsi.
4. Sahabat dan teman-teman seperjuangan yang telah memberikan bantuan, kebersamaan, serta dukungan selama masa perkuliahan.

KATA PENGANTAR

Puji Syukur penulis panjatkan kehadirat Allah SWT karena berkat Rahmat dan Hidayah-Nya sehingga penulis bisa menyelesaikan skripsi yang berjudul “Analisis Algoritma *Extreme Gradient Boosting* (XGBoost) Pada Metode *Content-Based Filtering* Untuk Klasifikasi Relevansi Buku”. Pada kesempatan ini, penulis menyampaikan ucapan terima kasih kepada berbagai pihak yang telah memberikan bantuan, dukungan, serta bimbingan dalam proses penyusunan skripsi ini, kepada:

1. Kedua orang tua dan keluarga tercinta yang selalu memberikan doa, dukungan, serta motivasi selama proses penyusunan skripsi
2. Ibu Ike Vewawati , S.Kom., M.Kom. selaku dosen pembimbing yang telah memberikan arahan, masukan, dan bimbingan selama proses penyusunan skripsi.
3. Bapak Arifiyanto Hadinegoro, S.Kom., M.T. selaku dosen penguji I yang telah memberikan kritik dan saran dalam penyempurnaan skripsi ini.
4. Bapak Theopilus Bayu Sasongko, S.Kom., M.Eng. selaku dosen penguji II yang telah memberikan kritik dan saran dalam penyempurnaan skripsi ini.
5. Prof. Dr. Kuswini, M.Kom selaku Dekan Fakultas Ilmu Komputer Universitas AMIKOM Yogyakarta
6. Teman-teman mahasiswa Program Studi S1 – Informatika khususnya kelas informatika 10 angkatan 22 serta seluruh pihak yang tidak dapat penulis sebutkan satu per satu, yang telah membantu kelancaran penyelesaian penulisan ini

Akhir kata, penulis berharap semoga skripsi ini dapat memberikan manfaat bagi pembaca serta dapat menjadi referensi bagi penelitian selanjutnya.

Yogyakarta, 8 Mei 2026

Penulis

DAFTAR ISI

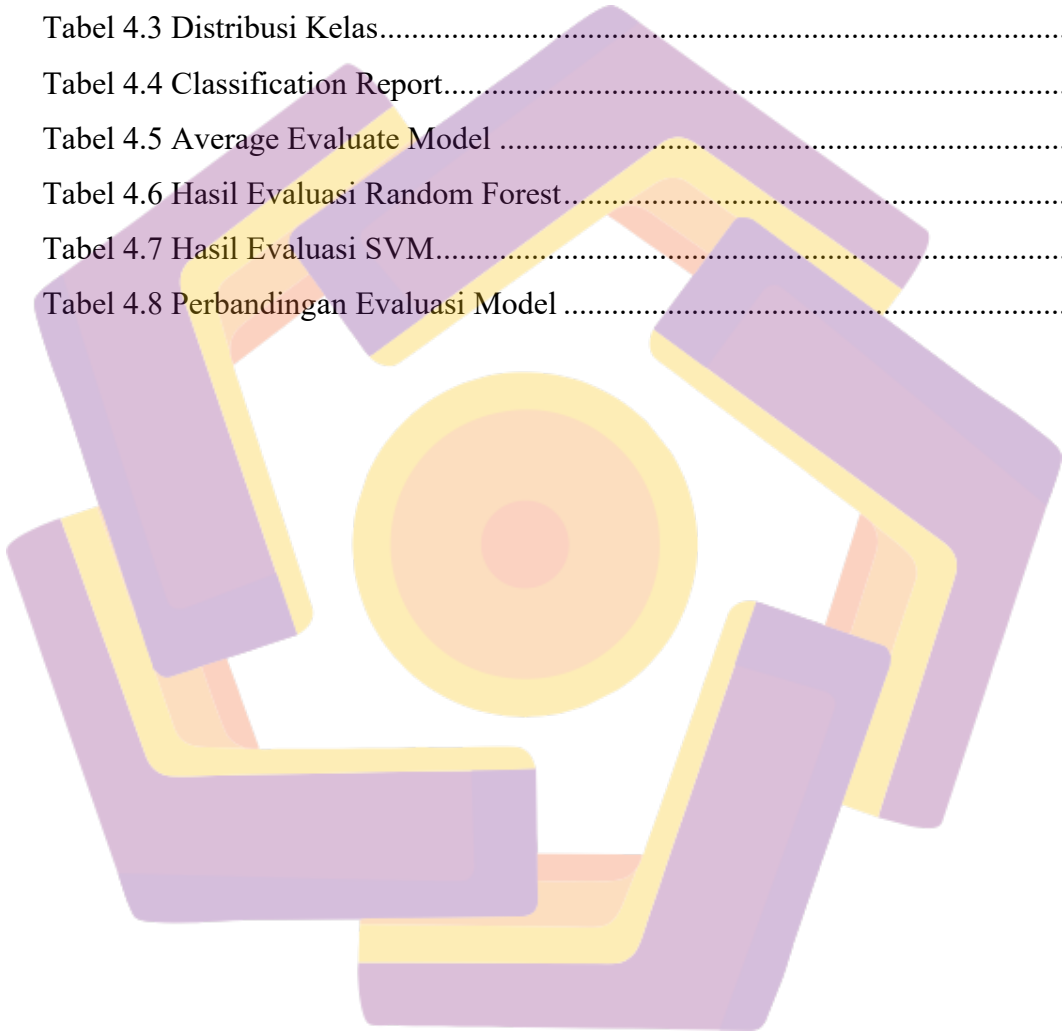
HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vi
DAFTAR ISI.....	vii
DAFTAR TABEL.....	x
DAFTAR GAMBAR.....	xi
DAFTAR LAMPIRAN.....	xii
DAFTAR LAMBANG DAN SINGKATAN	xiii
DAFTAR ISTILAH.....	xiv
INTISARI	xvii
<i>ABSTRACT</i>	xviii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah	2
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	3
1.6 Sistematika Penulisan	3
BAB II TINJAUAN PUSTAKA	5

2.1	Studi Literatur	5
2.2	Dasar Teori.....	10
2.2.1	Sistem Rekomendasi.....	10
2.2.2	Content-Based Filtering	10
2.2.3	TF-IDF (Term Frequency-Inverse Document Frequency)	11
2.2.4	Cosine Similarity	11
2.2.5	Pseudo-Labeling	12
2.2.6	Machine Learning	12
2.2.7	Pengumpulan Data	13
2.2.8	Preprocessing	13
2.2.9	Klasifikasi	14
2.2.10	XGBoost (Extreme Gradient Boosting).....	14
2.2.11	Evaluasi Model	16
BAB III METODE PENELITIAN		18
3.1	Objek Penelitian.....	18
3.2	Alur Penelitian	18
3.3	Alat dan Bahan.....	25
3.3.1	Data Penelitian	25
3.3.2	Alat dan Instrumen Penelitian.....	26
BAB IV HASIL DAN PEMBAHASAN		27
4.1	Akuisisi Data.....	27
4.2	EDA (Exploratory Data Analysis)	28

4.3	Attribute Selection	32
4.4	Preprocessing Data.....	34
4.5	Testing Query.....	39
4.6	Cosine Similarity	40
4.7	Pseudo-Labeling	40
4.8	Data Splitting	42
4.9	Modeling.....	42
4.10	Evaluasi Model	43
4.11	Deployment.....	51
4.12	Pengujian Sistem.....	51
BAB V PENUTUP		53
5.1	Kesimpulan	53
5.2	Saran	53
REFERENSI		55
LAMPIRAN.....		58

DAFTAR TABEL

Tabel 2.1 Keaslian Penelitian	7
Tabel 3.1 Alat dan Instrumen Penelitian.....	26
Tabel 4. 1 Atribut Dataset Buku	27
Tabel 4.2 Analisis Pemilihan Atribut.....	32
Tabel 4.3 Distribusi Kelas.....	41
Tabel 4.4 Classification Report.....	46
Tabel 4.5 Average Evaluate Model	47
Tabel 4.6 Hasil Evaluasi Random Forest.....	49
Tabel 4.7 Hasil Evaluasi SVM.....	49
Tabel 4.8 Perbandingan Evaluasi Model	50

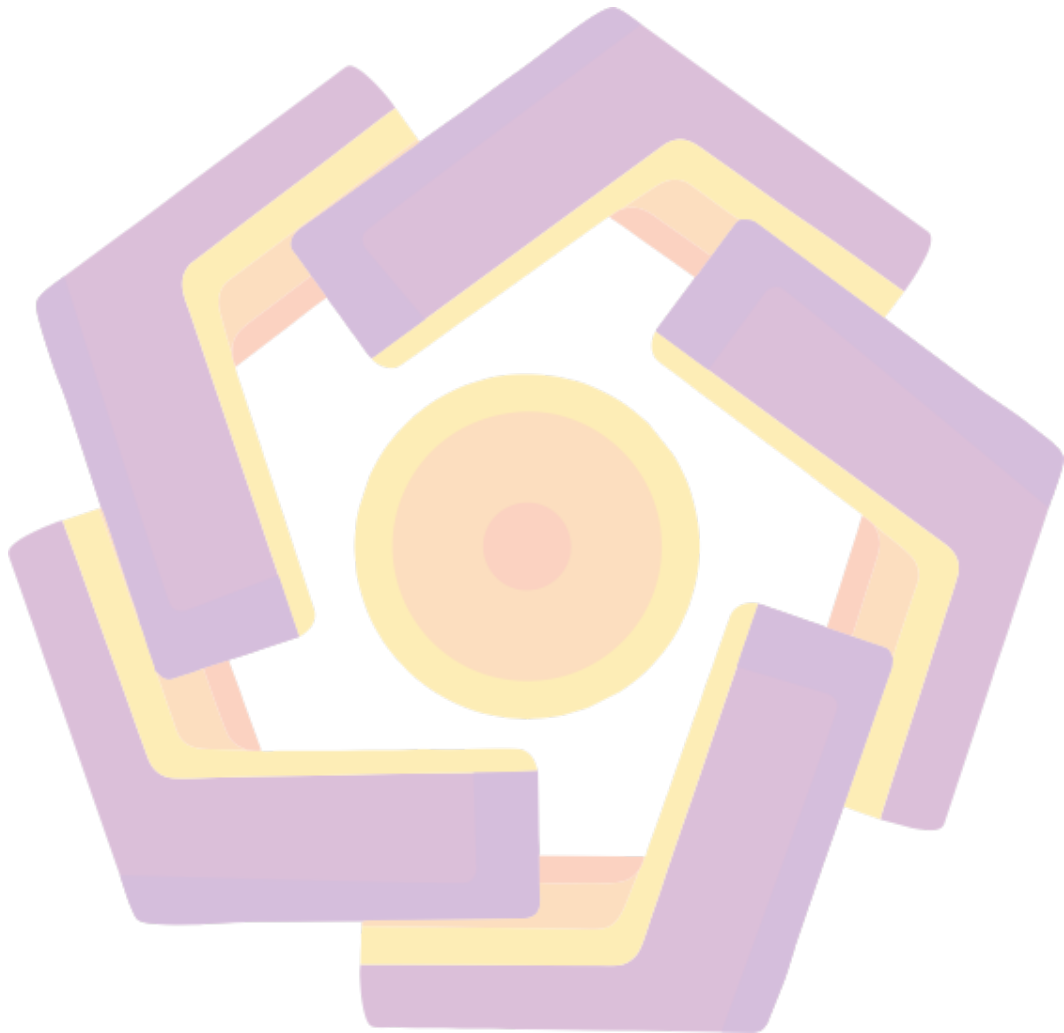


DAFTAR GAMBAR

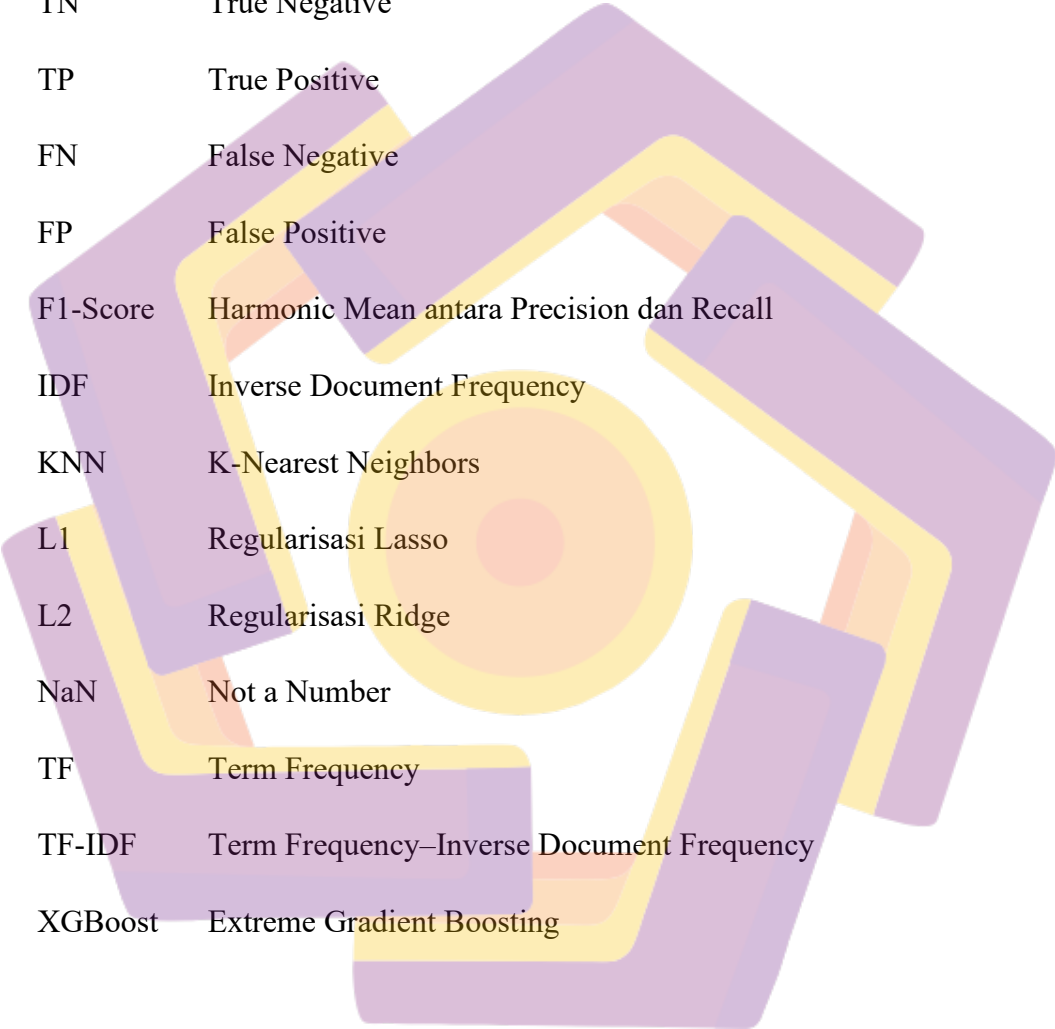
Gambar 3.1 Alur Penelitian	19
Gambar 3.2 Sample Dataset.....	26
Gambar 4.1 Dataset Buku	28
Gambar 4.2 Data Duplikat	28
Gambar 4.3 Nilai Null.....	29
Gambar 4.4 Placeholder.....	30
Gambar 4.5 Distribusi Panjang Deskripsi.....	30
Gambar 4.6 Distribusi Genre	31
Gambar 4.7 Wordcloud Deskripsi Buku.....	32
Gambar 4.8 Attribute Selection	34
Gambar 4.9 Penghapusan Duplikat.....	34
Gambar 4.10 Pembersihan Deskripsi.....	35
Gambar 4.11 Filtering Deskripsi.....	35
Gambar 4.12 Genre dengan data Kurang dari 5.....	36
Gambar 4.13 Pengelompokan Genre	37
Gambar 4.14 Combined Text.....	38
Gambar 4.15 TF-IDF	38
Gambar 4.16 Testing Query.....	39
Gambar 4.17 Cosine Similarity.....	40
Gambar 4.18 Visualisasi Distribusi Kelas.....	41
Gambar 4.19 Data Splitting	42
Gambar 4.20 Model Xgboost.....	43
Gambar 4.21 Confusion Matrix	44
Gambar 4.22 Overall Confusion Matrix	48
Gambar 4.23 Prototype	51
Gambar 4.24 Hasil Pengujian Sistem	52

DAFTAR LAMPIRAN

Lampiran 1. Dataset Penelitian	58
Lampiran 2. Repository	58
Lampiran 3. Streamlit	58

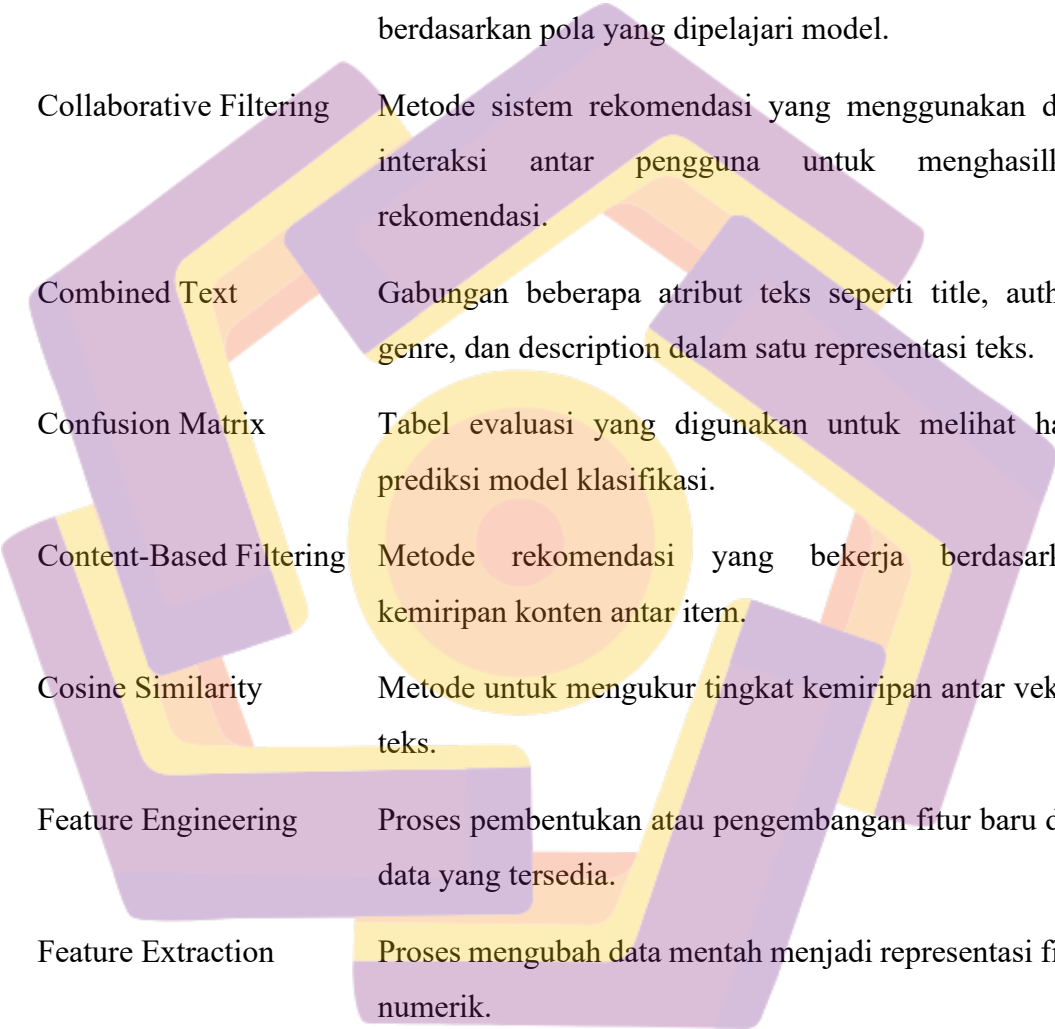


DAFTAR LAMBANG DAN SINGKATAN

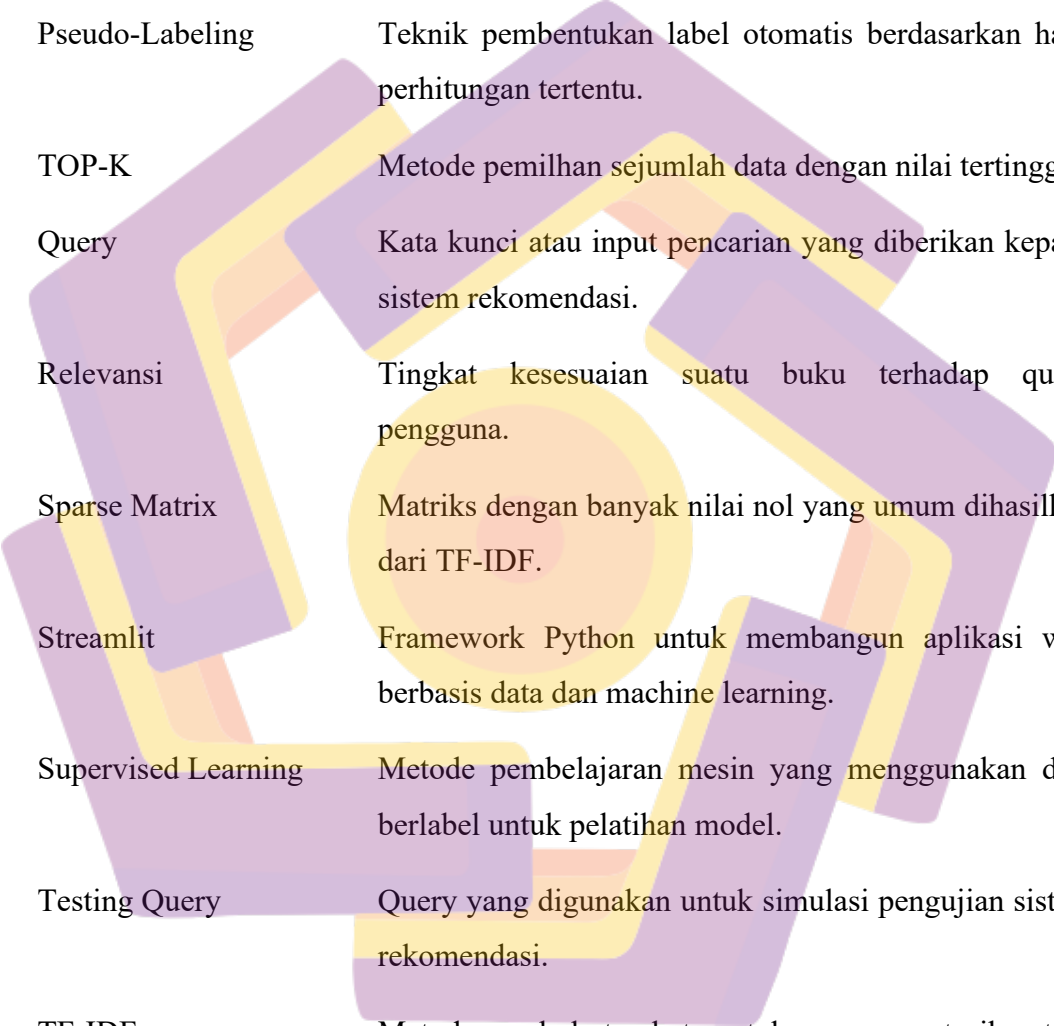


CBF	Content-Based Filtering
EDA	Exploratory Data Analysis
TN	True Negative
TP	True Positive
FN	False Negative
FP	False Positive
F1-Score	Harmonic Mean antara Precision dan Recall
IDF	Inverse Document Frequency
KNN	K-Nearest Neighbors
L1	Regularisasi Lasso
L2	Regularisasi Ridge
NaN	Not a Number
TF	Term Frequency
TF-IDF	Term Frequency–Inverse Document Frequency
XGBoost	Extreme Gradient Boosting

DAFTAR ISTILAH

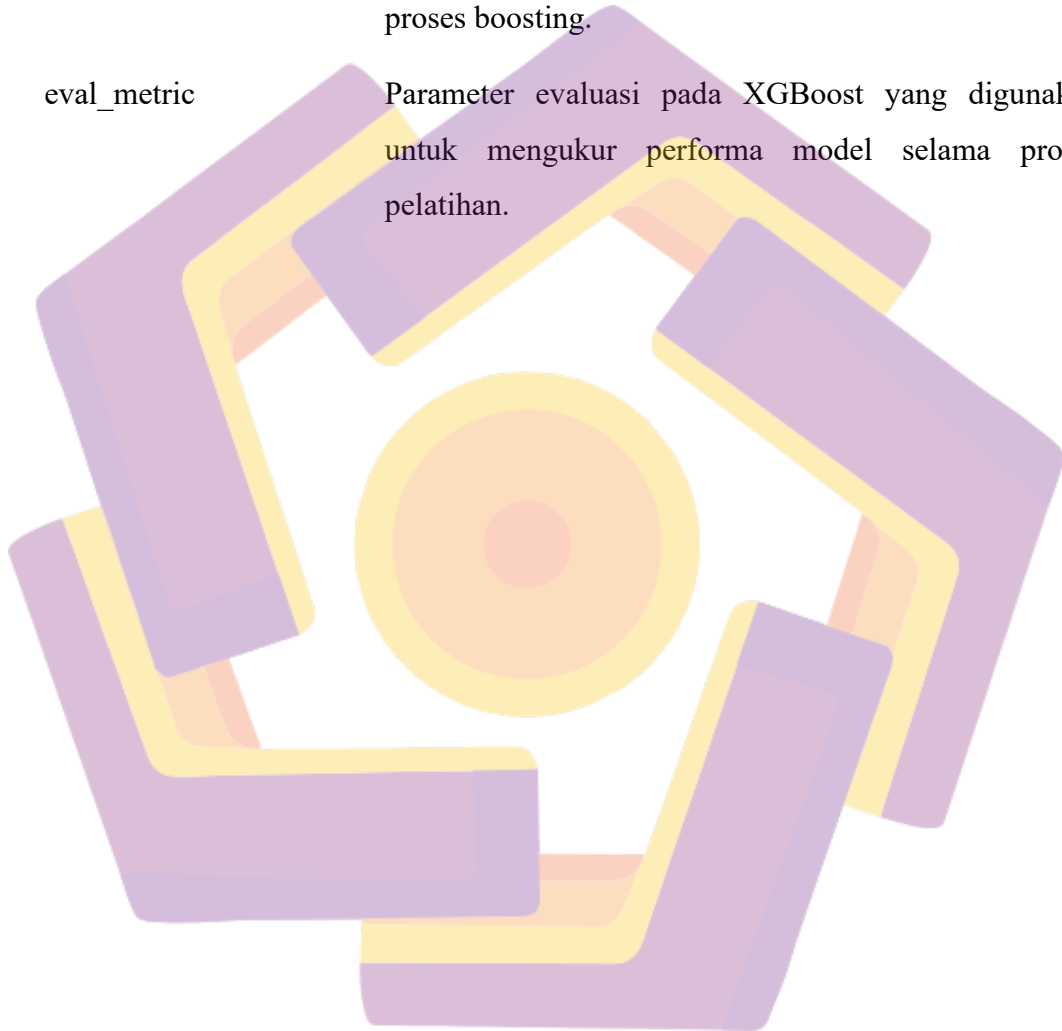


Attribute Selection	Proses pemilihan atribut yang relevan untuk digunakan dalam penelitian.
Classification	Proses pengelompokan data ke dalam kategori tertentu berdasarkan pola yang dipelajari model.
Collaborative Filtering	Metode sistem rekomendasi yang menggunakan data interaksi antar pengguna untuk menghasilkan rekomendasi.
Combined Text	Gabungan beberapa atribut teks seperti title, author, genre, dan description dalam satu representasi teks.
Confusion Matrix	Tabel evaluasi yang digunakan untuk melihat hasil prediksi model klasifikasi.
Content-Based Filtering	Metode rekomendasi yang bekerja berdasarkan kemiripan konten antar item.
Cosine Similarity	Metode untuk mengukur tingkat kemiripan antar vektor teks.
Feature Engineering	Proses pembentukan atau pengembangan fitur baru dari data yang tersedia.
Feature Extraction	Proses mengubah data mentah menjadi representasi fitur numerik.
Gradient Boosting	Teknik ensemble learning yang membangun model secara bertahap untuk memperbaiki kesalahan prediksi sebelumnya.
Machine Learning	Cabang kecerdasan buatan yang memungkinkan sistem belajar dari data.



Overfitting	Kondisi ketika model terlalu menyesuaikan diri dengan data latih sehingga performanya menurun pada data baru.
Prototype	Bentuk awal sistem yang digunakan untuk menunjukkan fungsi utama aplikasi.
Pseudo-Labeling	Teknik pembentukan label otomatis berdasarkan hasil perhitungan tertentu.
TOP-K	Metode pemilihan sejumlah data dengan nilai tertinggi.
Query	Kata kunci atau input pencarian yang diberikan kepada sistem rekomendasi.
Relevansi	Tingkat kesesuaian suatu buku terhadap query pengguna.
Sparse Matrix	Matriks dengan banyak nilai nol yang umum dihasilkan dari TF-IDF.
Streamlit	Framework Python untuk membangun aplikasi web berbasis data dan machine learning.
Supervised Learning	Metode pembelajaran mesin yang menggunakan data berlabel untuk pelatihan model.
Testing Query	Query yang digunakan untuk simulasi pengujian sistem rekomendasi.
TF-IDF	Metode pembobotan kata untuk merepresentasikan teks dalam bentuk numerik.
Vectorization	Proses mengubah data teks menjadi bentuk vektor numerik.
XGBoost	Algoritma machine learning berbasis gradient boosting yang digunakan untuk klasifikasi dan prediksi.

random_state	Parameter pada model machine learning yang digunakan untuk mengatur seed acak agar hasil pelatihan dapat direproduksi secara konsisten.
n_estimators	Parameter pada XGBoost yang menentukan jumlah pohon keputusan (decision tree) yang digunakan dalam proses boosting.
eval_metric	Parameter evaluasi pada XGBoost yang digunakan untuk mengukur performa model selama proses pelatihan.



INTISARI

Banyaknya pilihan buku digital di berbagai platform sering membuat pengguna kesulitan menemukan bacaan yang sesuai dengan minat dan kebutuhannya. Untuk mengatasi hal tersebut, dibutuhkan sistem yang mampu memberikan rekomendasi buku secara otomatis berdasarkan kesamaan konten. Penelitian ini bertujuan untuk menganalisis penerapan algoritma Extreme Gradient Boosting (XGBoost) pada metode Content-Based Filtering dalam klasifikasi relevansi buku. Dataset yang digunakan memuat informasi seperti judul, penulis, genre, dan deskripsi buku. Teks dari atribut tersebut direpresentasikan menggunakan metode TF-IDF (Term Frequency–Inverse Document Frequency), kemudian tingkat kemiripan antar buku dihitung menggunakan Cosine Similarity. Hasil pengukuran ini menjadi dasar dalam proses klasifikasi menggunakan XGBoost. Melalui pendekatan ini diharapkan diperoleh hasil analisis yang akurat dan dapat menjadi acuan bagi pengembangan sistem rekomendasi berbasis konten di masa mendatang.

Kata kunci: XGBoost, Content-Based Filtering, TF-IDF, Cosine Similarity, Klasifikasi

ABSTRACT

The vast selection of digital books available across various platforms often makes it difficult for users to find reading material that aligns with their interests and needs. To address this challenge, a system capable of automatically recommending books based on content similarity is required. This study aims to analyze the application of the Extreme Gradient Boosting (XGBoost) algorithm within the Content-Based Filtering method for book relevance classification. The dataset used includes information such as book titles, authors, genres, and descriptions. Text from these attributes is represented using the TF-IDF (Term Frequency–Inverse Document Frequency) method, and the similarity between books is calculated using Cosine Similarity. The results of these similarity measurements serve as the foundation for the classification process using XGBoost. Through this approach, it is expected that accurate analytical results can be achieved, providing a useful reference for the future development of content-based recommendation systems.

Keyword: *XGBoost, Content-Based Filtering, TF-IDF, Cosine Similarity, Classification*