Jurnal Komputasi

Vol 13 No. 2 , 2025

Klasifikasi Dokumen Akademik Berbasis XGBoost untuk Pemetaan Tujuan Pembangunan Berkelanjutan (SDGs) di Universitas Lampung

[1*]Rahman Taufik, [2]Aristoteles, [3]Candra Wijaya & [4]Rakhmat Herlambang

[1,2,3,4] Program Studi Ilmu Komputer, Fakultas MIPA, Universitas Lampung, Jl. Prof. Dr. Ir. Sumantri Brojonegoro No.1, Gedong Meneng,Kec. Rajabasa, Kota Bandar Lampung, Lampung, Indonesia

Abstrak —Pemetaan kontribusi institusi pendidikan tinggi terhadap Sustainable Development Goals merupakan tantangan krusial untuk akuntabilitas global dan capaian World Class University. Meskipun model-model canggih rentan terhadap overfitting dan menuntut sumber daya komputasi besar pada data yang tidak seimbang, penelitian ini mengeksplorasi algoritma XGBoost sebagai solusi efisien untuk klasifikasi SDGs pada dokumen akademik universitas. Penelitian ini menggunakan dataset sebanyak 148136 dokumen, diproses dengan TF−IDF, dan dioptimasi dengan hyperparameter tuning serta class sample weighting untuk mitigasi imbalance. Hasil evaluasi menunjukkan model yang stabil dengan accuracy 0.92, precision 0.92, recall 0.89, dan F1−score 0.90 pada dataset uji. Meskipun kinerja agregat tinggi, analisis log loss dan confusion matrix mengindikasikan adanya overfitting lokal pada kategori minoritas, yang menyebabkan recall rendah di kelas-kelas tersebut. Secara keseluruhan, model XGBoost terbukti valid sebagai alat ukur efektif untuk memetakan kontribusi universitas terhadap SDGs, sekaligus memberikan panduan strategis berbasis data untuk mengidentifikasi celah dan mendorong keseimbangan capaian WCU.

Kata Kunci: dokumen akademik; klasifikasi; sdgs; xgboost.

Abstract *— Mapping the contributions of higher education institutions to the Sustainable Development Goals (SDGs) is a crucial challenge for global accountability and achieving World Class University (WCU) status. Although advanced models are often prone to overfitting and require significant computational resources when dealing with unbalanced data, this study explores the XGBoost (eXtreme Gradient Boosting) algorithm as an efficient solution for the classification of SDGs in university academic documents. The research utilized a dataset of 148,136 documents, which were processed using TF-IDF feature extraction and optimized with hyperparameter tuning and class sample weighting to mitigate data imbalance. Evaluation results showed the model to be stable and robust, achieving an accuracy of 0.92, precision of 0.92, recall of 0.89, and an F1−score of 0.90 on the test dataset. Despite the high aggregate performance, analysis of the log loss and confusion matrix indicated local overfitting in minority categories, resulting in low recall for those specific classes. Overall, the XGBoost model proved valid as an effective measurement tool for mapping the university's SDG contributions, simultaneously providing data-driven strategic guidance to identify gaps and promote balanced achievement toward WCU goals.*

Keywords: academic documents; classification; sdgs; xgboost.

** Corresponding author :*
Rahman Taufik
Universitas Lampung, Bandar Lampung, Indonesia
rahman.taufik@fmipa.unila.ac.id

1. PENDAHULUAN

Pembangunan berkelanjutan telah bertransformasi menjadi kerangka kerja global yang fundamental. Hal ini didasarkan pada pengesahan Sustainable Development Goals (SDGs) oleh Perserikatan Bangsa-Bangsa (PBB) pada tahun 2015, yang menetapkan 17 tujuan universal untuk mencapai dunia yang lebih baik pada tahun 2030 [1]. Kerangka SDGs yang ambisius ini menuntut kolaborasi dari semua pemangku kepentingan, tidak terkecuali institusi pendidikan tinggi. Sebagai pusat pendidikan, penelitian, dan pengabdian, universitas berfungsi tidak hanya sebagai penghasil ilmu pengetahuan, tetapi juga sebagai agen perubahan sosial yang mampu berkontribusi langsung terhadap pencapaian target pembangunan

berkelanjutan [2]. Oleh karena itu, pengukuran dan pemetaan kontribusi universitas terhadap SDGs menjadi indikator penting dalam penilaian kinerja akademik global, seperti dalam sistem pemeringkatan QS Sustainability dan Times Higher Education Impact Rankings [3].

Namun, potensi besar dari luaran akademik, seperti publikasi ilmiah, laporan penelitian, laporan pengabdian kepada masyarakat, modul perkuliahan, serta program akademik lainnya, belum dimanfaatkan secara optimal dalam pemetaan dan dukungan terhadap agenda SDGs. Hal ini disebabkan oleh keterbatasan metode manual dalam menganalisis volume dokumen akademik yang sangat besar dan beragam. Keterbatasan ini menjadi kendala utama dalam mengidentifikasi secara akurat isu-isu keberlanjutan yang telah menjadi kontribusi nyata universitas.

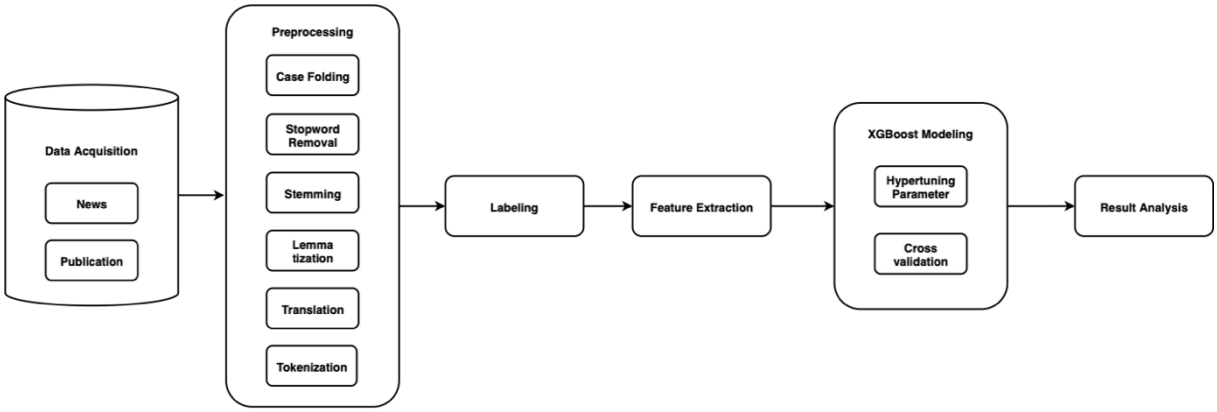
Untuk mengatasi permasalahan klasifikasi pada dokumen akademik dengan volume yang besar dan kompleks, Natural Language Processing (NLP) dan Machine Learning (ML) menawarkan solusi yang paling relevan dan terukur. Berbagai penelitian telah mengeksplorasi algoritma klasifikasi teks dalam beragam konteks. Li et al. (2022) meninjau evolusi model klasifikasi teks dari metode tradisional hingga deep learning, menegaskan efektivitas NLP dan ML dalam meningkatkan akurasi klasifikasi [4]. Secara spesifik, pendekatan K-Means (Kwale, 2013; Khan et al., 2024) sering digunakan karena efisiensi komputasinya, namun performanya terbukti sangat tergantung pada distribusi data dan pemilihan centroid awal, menjadikannya kurang optimal untuk data teks akademik yang kompleks dan tidak terstruktur [5][6]. Selanjutnya, model deep learning seperti CNN (Wang, 2023), BiLSTM (Huang et al., 2018; Duan et al., 2024), dan BERT (Montes et al., 2024) terbukti efektif dalam menangkap konteks semantik dan pola sekuensial teks yang mendalam. Akan tetapi, penggunaan model-model ini memerlukan sumber daya komputasi yang besar dan rentan terhadap overfitting ketika dihadapkan pada data berlabel yang tidak seimbang atau dalam jumlah terbatas [7][8][9].

Meskipun penelitian sebelumnya telah memberikan kontribusi penting terhadap klasifikasi teks secara umum, terdapat celah signifikan yang belum dieksplorasi, yakni evaluasi kinerja algoritma XGBoost (Extreme Gradient Boosting) dalam konteks klasifikasi kategori SDGs pada dokumen akademik universitas. Gap penelitian ini perlu dieksplorasi, mengingat XGBoost memiliki sejumlah keunggulan yang sangat relevan untuk skenario data akademik. Keunggulan tersebut meliputi efisiensi komputasi yang tinggi, kemampuan regularisasi untuk menghindari overfitting pada data yang rumit, serta penanganan otomatis terhadap data yang hilang [10]. Dengan karakteristik ini, XGBoost menawarkan alternatif yang ringan namun tetap akurat untuk klasifikasi berskala besar.

Oleh karena itu, studi ini bertujuan untuk mengembangkan model klasifikasi otomatis berbasis algoritma XGBoost guna memetakan dokumen akademik universitas ke dalam kategori SDGs. Melalui pendekatan ini, studi berfokus pada identifikasi fitur teks yang paling berpengaruh, evaluasi kinerja model secara komprehensif, serta analisis potensi penerapannya dalam mendukung agenda *World Class University* (WCU). Kontribusi yang diharapkan dari studi ini meliputi kajian ilmiah terkait eksplorasi model XGBoost dalam menangani volume dokumen akademik yang besar, serta penyediaan kerangka klasifikasi yang membantu universitas dalam memetakan, memonitor, dan mengevaluasi kontribusi akademiknya terhadap SDGs.

2. METODOLOGI PENELITIAN

Bab ini menjelaskan tahapan metodologis yang digunakan dalam penelitian untuk membangun dan mengevaluasi model klasifikasi otomatis dokumen akademik universitas berdasarkan kategori Sustainable Development Goals (SDGs). Tahapan penelitian disajikan pada Gambar 1 dan dijelaskan secara sistematis untuk memastikan transparansi dan replikabilitas. Proses penelitian dimulai dari pengumpulan data, pra-pemrosesan, pelabelan manual, ekstraksi fitur, pelatihan model XGBoost, termasuk optimasi hyperparameter, dan evaluasi kinerja, lalu diakhiri dengan analisis hasilnya.



Gambar 1. Tahapan Penelitian

2.1. Akuisisi dan Sumber Data

Tahap awal penelitian ini adalah pengumpulan data teks dalam volume besar dari sumber-sumber akademik dan publikasi resmi Universitas Lampung. Data yang dikumpulkan merupakan luaran akademik seperti publikasi ilmiah, laporan penelitian, laporan pengabdian, dan berita yang mencerminkan kontribusi universitas terhadap pembangunan berkelanjutan. Sumber data utama meliputi SINTA (Science and Technology Index) untuk data publikasi ilmiah, Repositori Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Lampung untuk laporan penelitian dan pengabdian, dan berita resmi Universitas Lampung terkait informasi program dan kegiatan. Data mentah ini menjadi input utama dalam proses selanjutnya untuk dipetakan ke dalam kategori SDGs.

2.2. Pra-pemrosesan

Dilakukan pra-pemrosesan data teks yang telah diakuisisi guna memastikan kualitas, konsistensi, dan kesiapan data sebagai masukan model yang diusulkan. Teknik pra-pemrosesan ini meliputi case folding untuk menyeragamkan kapitalisasi dan stopword removal untuk menghilangkan kata-kata umum yang tidak signifikan secara semantik (misalnya, "yang", "dan", "adalah"). Dengan menghilangkan stopword, model dapat berfokus secara eksklusif pada istilah-istilah substantif yang benar-benar relevan dengan label, sehingga meningkatkan akurasi klasifikasi [11]. Selain itu, *stemming* dan *lemmatization* digunakan untuk mereduksi kata menjadi bentuk dasarnya. Normalisasi bentuk kata dasar ini penting guna meningkatkan *recall* model, serta frekuensi fitur yang signifikan akan terkumpul sehingga meningkatkan performa dan efisiensi komputasi model [12]. Selanjutnya, *translation* dilakukan pada frasa kunci yang mengandung bahasa inggris dan tokenisasi dilakukan untuk memecah teks menjadi unit-unit kata atau frasa diskrit.

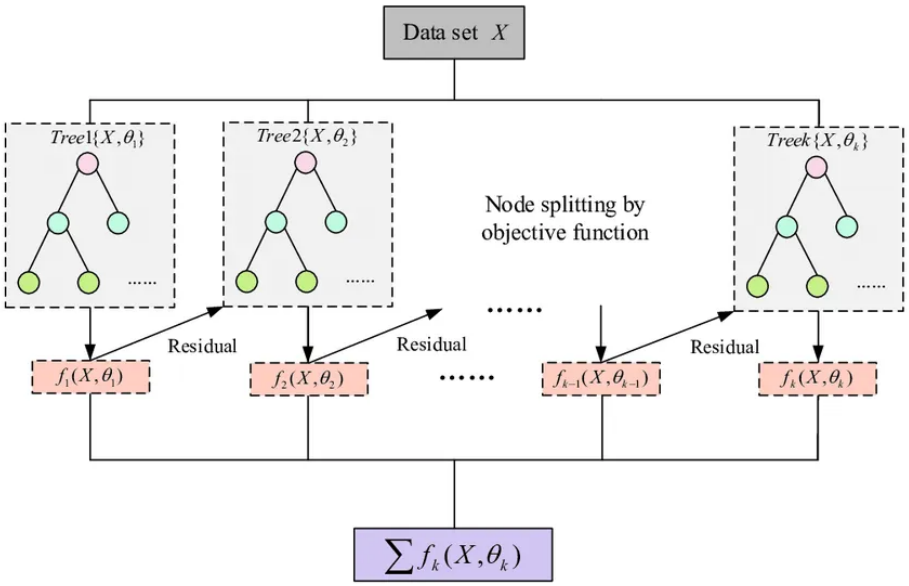
2.3. Pelabelan Manual

Proses pelabelan dokumen akademik dilakukan secara hati-hati untuk mengaitkan setiap dokumen dengan satu kategori SDGs yang relevan. Pelabelan dilakukan berdasarkan Pedoman Kata Kunci SDGs Penelitian dan Publikasi Ilmiah yang dikeluarkan oleh Vokasi Universitas Airlangga [13]. Panduan standar ini memastikan konsisten dan objektivitas pelabelan dengan menyediakan serangkaian kata kunci spesifik dan frasa yang dikelompokkan per tujuan SDG, dapat dilihat pada Gambar 2. Setiap dokumen di-*labeling* berdasarkan keberadaan kata kunci tersebut, menghasilkan dataset berlabel yang menjadi *ground truth* untuk pelatihan model.

Gambar 2. Kategori tujuan SDGs

2.4. Ekstraksi Fitur

Setelah data bersih, data teks diubah menjadi representasi numerik menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF dipilih karena efektivitasnya dalam membedakan signifikansi sebuah kata (*term*) dalam satu dokumen relatif terhadap seluruh korpus. Teknik ini tidak hanya menghitung frekuensi kemunculan kata, tetapi juga memberikan bobot yang lebih tinggi pada kata-kata yang jarang muncul di banyak dokumen, menjadikannya sangat informatif untuk memisahkan kategori [14]. Dalam konteks SDGs, TF-IDF memastikan bahwa istilah-istilah teknis atau spesifik yang terkait erat dengan satu atau dua tujuan SDG tertentu (misalnya, "ekosistem pesisir" untuk SDG 14) akan mendapatkan bobot yang lebih tinggi dibandingkan kata-kata umum, sehingga memperkuat diferensiasi antar-kelas.



Gambar 3. Arsitektur XGBoost

2.5. XGBoost Modeling

Pada studi ini, model yang diusulkan adalah XGBoost. XGBoost (eXtreme Gradient Boosting) adalah algoritma ensemble yang membangun model secara sekuensial menggunakan pohon keputusan (*Decision Tree*) sebagai pembelajar dasarnya (*base learner*). Pada Gambar 3, dapat dilihat inti dari arsitektur XGBoost adalah *boosting adaptive*, di mana setiap pohon yang baru ditambahkan dilatih untuk memperbaiki kesalahan (*residual*) dari gabungan semua pohon sebelumnya, dengan prediksi akhir berupa penjumlahan bobot semua pohon.

Untuk mendapatkan performa optimal pada pemodelan XGBoost, dilakukan optimasi hyperparameter dengan nilai-nilai yang telah di-*tuning*. Hyperparameter *tuning* ini meliputi colsample_bytree 1.0, learning_rate 0.1, max_depth 4, n_estimators 300, dan subsample 1.0. Nilai hyperparameter max_depth 4 dan n_estimators 300 dipilih untuk menyeimbangkan bias dan varians, sementara colsample_bytree 1.0 dan subsample 1.0 dipertahankan untuk memanfaatkan kekayaan informasi dari volume dataset yang besar [15]. Selain itu, penelitian ini secara khusus menguji coba pembobotan sampel kelas (*class sample weighting*) dalam XGBoost. Pembobotan ini bertujuan untuk memitigasi bias yang timbul akibat ketidakseimbangan kategori atau label yang dominan dalam dataset SDGs.

2.6. Evaluasi Kinerja Model dan Analisis Hasil

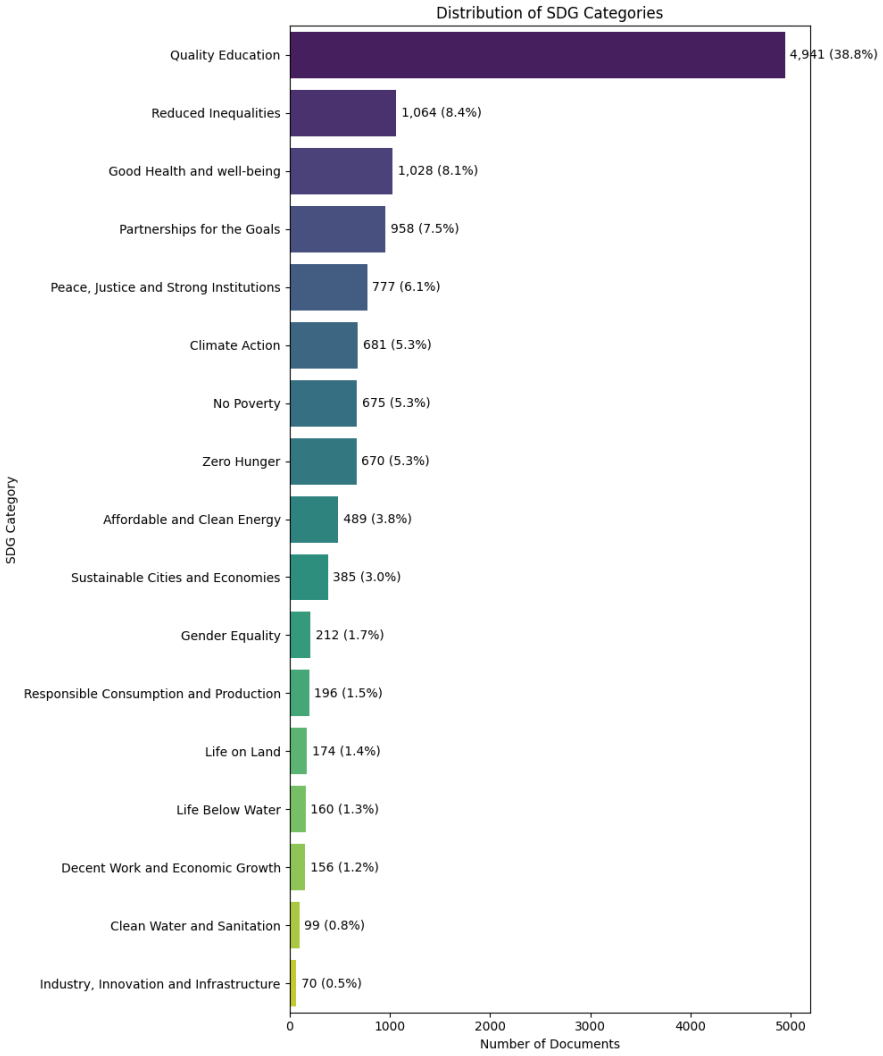
Evaluasi kinerja model dilakukan menggunakan 10-Fold Cross-Validation. Dataset dibagi menjadi tiga bagian, yaitu data *train* (pelatihan), data *evaluation* (validasi untuk hyperparameter tuning), dan data *test* (pengujian akhir) untuk mengukur kemampuan generalisasi model. Metrik evaluasi yang digunakan meliputi *accuracy*, *precision*, *recall*, dan F1-*score*. Analisis hasil kinerja tidak hanya berfokus pada skor agregat, tetapi juga mencakup analisis terhadap pemetaan klasifikasi SDG, analisis ini dapat memberikan gambaran komprehensif mengenai kemampuan model dalam mengidentifikasi kontribusi untuk setiap kategori SDG.

3. HASIL DAN PEMBAHASAN

Bab ini menyajikan temuan empiris dan analisis mendalam dari pengembangan model klasifikasi XGBoost untuk memetakan dokumen akademik Universitas Lampung ke dalam kategori SDGs. Hasil yang diuraikan meliputi distribusi label SDGs, kinerja kuantitatif model yang diusulkan, serta pembahasan kritis terhadap tantangan utama, yakni isu ketidakseimbangan kelas (imbalance) pada kategori SDGs. Pembahasan ini bertujuan untuk menginterpretasikan angka kinerja model dalam konteks kontribusi nyata universitas terhadap agenda pembangunan berkelanjutan, sekaligus memberikan rekomendasi area fokus untuk mendukung target WCU.

3.1. Distribusi Label SDGs

Tahap awal penyajian hasil berfokus pada karakterisasi data termasuk distribusinya. Dataset penelitian ini terdiri dari total 148.136 dokumen yang berhasil dikumpulkan dari sumber-sumber utama universitas, antara lain SINTA, Repositori LPPM, dan Berita Resmi Universitas Lampung. Hasil menunjukkan bahwa dataset tersebut memiliki ketidakseimbangan kelas (*imbalance*) yang signifikan, hal ini dapat dilihat pada Gambar 4. Beberapa kategori SDGs, misalnya *quality education* (pendidikan berkualitas) memiliki jumlah data sekitar 4941 atau sekitar 38,8% dari seluruh data, persentase ini sangat dominan terhadap kategori lainnya. Sementara itu, kategori *industry, innovation, and infrastructure* (industri, inovasi, dan infrastruktur) memiliki jumlah data yang sangat sedikit, yaitu 70 atau sekitar 0,5%. Ketidakseimbangan ini merupakan cerminan langsung dari distribusi aktivitas dan fokus riset di lingkungan universitas Lampung saat ini, yang secara inheren memprioritaskan beberapa isu keberlanjutan tertentu. Kondisi *imbalance* ini menjadi tantangan utama yang harus diatasi dalam proses pemodelan untuk memastikan kinerja yang adil di seluruh 17 tujuan.



Gambar 4. Distribusi label SDGs

Test Accuracy: 0.9183098591549296
Test Report:

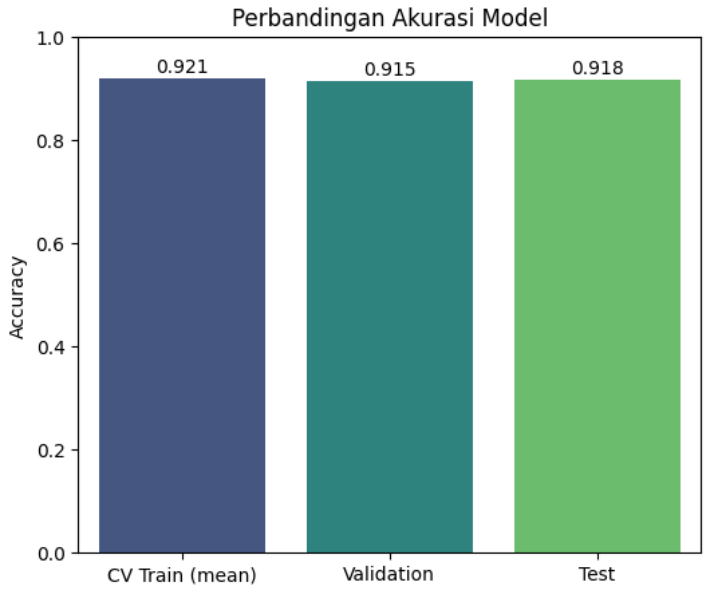
	precision	recall	f1-score	support
1	0.89	0.94	0.91	17
2	0.99	0.88	0.93	83
3	0.84	0.92	0.88	177
4	0.99	0.96	0.98	267
5	0.75	0.50	0.60	6
6	0.88	0.88	0.88	49
7	1.00	0.97	0.99	34
8	0.87	0.94	0.90	80
9	0.89	0.91	0.90	64
10	1.00	1.00	1.00	1
11	0.89	0.90	0.89	61
12	0.98	0.92	0.95	66
13	0.95	0.83	0.89	24
14	0.79	0.75	0.77	36
15	0.87	0.95	0.91	41
16	0.94	0.96	0.95	48
17	1.00	0.82	0.90	11
accuracy			0.92	1065
macro avg	0.91	0.88	0.90	1065
weighted avg	0.92	0.92	0.92	1065

Gambar 5. Laporan kinerja klasifikasi pada data *test*

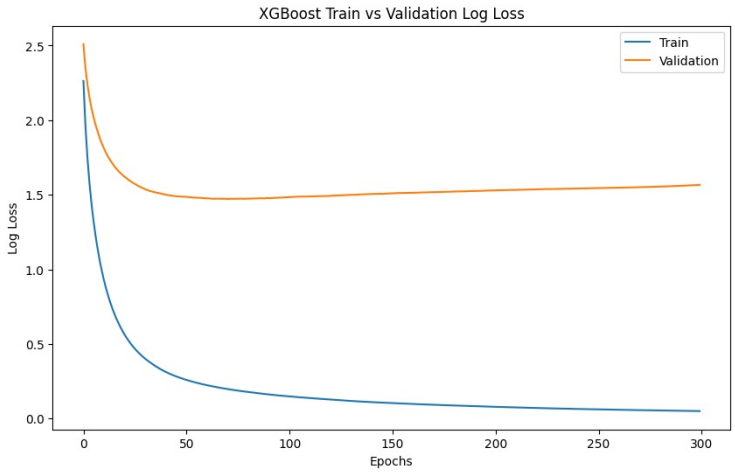
3.2. Kinerja Model XGBoost

Kinerja model XGBoost dalam klasifikasi otomatis dokumen akademik berdasarkan kategori SDGs dievaluasi menggunakan metrik agregat dan analisis visual yang mendalam. Model klasifikasi XGBoost yang diusulkan dengan optimasi menunjukkan kinerja prediktif yang kuat. Kinerja model pada dataset uji mencapai *accuracy* sebesar 0.92, *precision* sebesar 0.91, *recall* sebesar 0.88, dan F1−*score* 0.90. Secara keseluruhan, hasil kinerja ini dapat dilihat pada laporan klasifikasi yang ditampilkan pada Gambar 5, hasil kinerja ini menegaskan validitas model XGBoost sebagai alat yang efisien dan akurat untuk memetakan luaran akademik ke SDGs.

Kualitas kinerja model pada data uji, kemudian divalidasi oleh konsistensi kinerja antar set data yang dapat dilihat pada Gambar 6. Kinerja ini meliputi *accuracy* pada data *train* sebesar 0.921, validasi sebesar 0.915, dan pengujian sebesar 0.918. Stabilitas skor yang tinggi ini menegaskan bahwa model memiliki kemampuan *generalization* yang kuat dan dapat diandalkan. Hal ini membuktikan bahwa model tidak hanya menghafal data pelatihan, tetapi juga mampu menerapkan pola yang dipelajari untuk membuat prediksi akurat untuk kategori SDG pada dokumen baru.



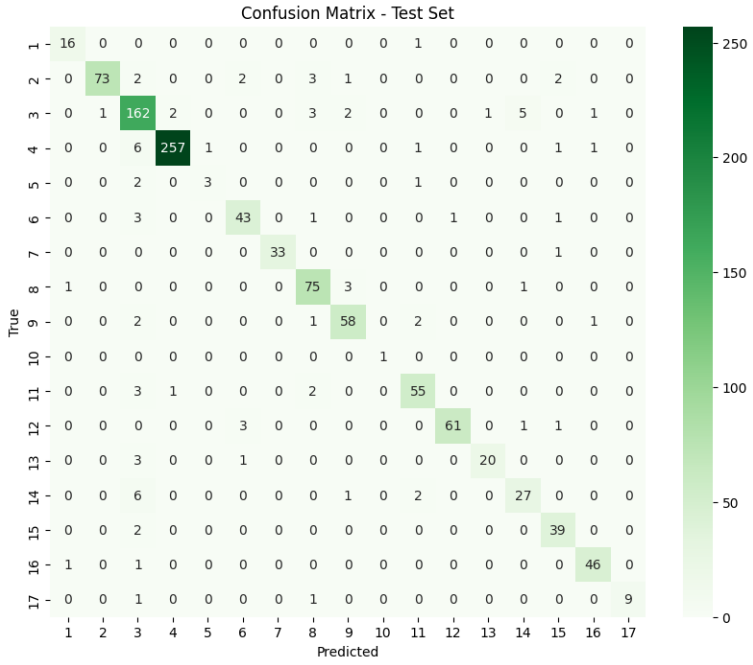
Gambar 6. Perbandingan akurasi model antar *dataset*



Gambar 7. Perbandingan nilai *log loss* pada data *train* dan *validation*

Meskipun model menunjukkan kinerja yang andal ditandai dengan skor metrik yang tinggi, terdapat kontras saat analisis diperluas ke metrik berbasis *loss* dan metrik per-kelas. Indikasi *overfitting* terlihat pada kurva *log loss* yang ditampilkan pada Gambar 7 dan *confusion matrix* pada Gambar 8. Kurva *log loss* menunjukkan perbandingan antara *train loss* dan *validation loss* yang melebar, mengindikasikan bahwa model mulai mengalami *overfitting* pada data *train* karena *validation loss* tidak lagi menurun dan cenderung naik atau mendatar, sementara *train loss* terus berkurang.

Selain itu, kinerja agregat yang tinggi sebagian besar dipengaruhi oleh keberhasilan model memprediksi kategori dengan jumlah banyak, hal ini dapat dilihat berdasarkan analisis *confusion matrix* yang ditujukan Gambar 8. Diagonal utama pada *confusion matrix* set uji menunjukkan bahwa sebagian besar prediksi yang benar terpusat pada kategori-kategori SDGs yang paling banyak muncul dalam data, seperti SDG 3 (*good wealth and well-being*) yang memiliki jumlah total data sekitar 1028 dan terprediksi benar di data test yaitu sekitar 162, dan SDG 4 (*quality education*) yang memiliki jumlah total data sekitar 4941 dan terprediksi benar di data test yaitu sekitar 257. Namun, anomali terjadi pada SDG 10 (*reduced inequalities*), secara distribusi total data sekitar 1064, tergolong kategori mayoritas, tetapi temuan pada matriks justru menunjukkan hasil sebaliknya, jumlah prediksi benarnya tergolong kecil yaitu hanya 1. Hal ini mengindikasikan bahwa meskipun sampelnya banyak, fitur teks (kata kunci TF-IDF) yang terkait dengan SDG 10 tidak cukup *discriminative* atau terlalu banyak tumpang tindih dengan kategori SDG lainnya. Akibatnya, model yang diusulkan kesulitan membangun batas keputusan yang jelas. Model gagal memanfaatkan ketersediaan sampel tersebut untuk mencapai kinerja prediksi yang andal, menyebabkan banyak sampel SDG 10 terklasifikasi sebagai *false negatives* atau salah prediksi ke kategori lain. Kesulitan ini semakin memperkuat kesimpulan bahwa skor agregat yang tinggi tidak menutupi tantangan kinerja yang mendalam pada tingkat per-kelas yang disebabkan oleh isu ketidakseimbangan data dan kompleksitas semantik antar kategori SDGs.



Gambar 8. *Confusion matrix* pada data uji

3.3. Analisis Kinerja

Analisis hasil kinerja model XGBoost mengungkapkan adanya *trade−off* antara *precision* (0.91) dan *recall* (0.88) dipengaruhi oleh distribusi kelas yang tidak seimbang. Tingginya nilai *precision* menunjukkan bahwa ketika model memprediksi sebuah dokumen termasuk SDG tertentu, klaim tersebut probabilitas benarnya besar. Namun, penurunan kecil pada *recall* mengindikasikan bahwa model gagal

menangkap seluruh dokumen yang relevan (banyak *false negatives*). Kesulitan ini tidak hanya disebabkan oleh sedikitnya sampel data minoritas, tetapi juga oleh kualitas fitur teks (kata kunci TF-IDF) yang tumpang tindih antar kelas. Misalnya, kata kunci terkait "*no poverty*" (SDG 1) seringkali juga muncul dalam konteks "*reduce inequalities*" (SDG 10) atau "*quality education*" (SDG 4). Meskipun *class weighting* telah diterapkan, ambiguitas semantik fitur pada kategori minoritas ini membuat model XGBoost membuat prediksi yang salah. Selain itu, nilai F1−*score* sebesar 0.90 menunjukan ukuran yang menyeimbangkan kecenderungan model ini, memberikan gambaran yang akurat mengenai kinerja optimal dalam menghadapi kompleksitas data akademik berbasis SDGs.

4. KESIMPULAN

Studi ini berhasil membangun dan memvalidasi model klasifikasi XGBoost untuk memetakan dokumen akademik ke dalam kategori SDGs. Kinerja yang andal terbukti melalui nilai metrik *accuracy*, *precision*, *recall*, dan F1−*score*. Namun, hasil analisis mengungkap adanya tantangan kinerja mendalam yang disebabkan oleh isu ketidakseimbangan kelas (*imbalance*), di mana skor agregat yang tinggi didominasi oleh prediksi kategori mayoritas. Tantangan ini menghasilkan *recall* yang rendah dan *overfitting* lokal pada kategori minoritas yang memiliki tumpang tindih semantik fitur. Secara implisit, model ini memberikan implikasi strategis bagi universitas, berfungsi sebagai pemetaan SDGs yang tidak hanya mengonfirmasi kontribusi yang sudah ada, tetapi juga mengidentifikasi celah kontribusi pada SDG lainnya. Sebagai saran untuk penelitian mendatang, keterbatasan studi ini meliputi penanganan data imbalance, serta eksplorasi fitur ekstraksi dan model lainnya yang lebih canggih untuk menangkap konteks semantik yang lebih kaya, demi meningkatkan kinerja klasifikasi pada kategori SDGs yang langka.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih atas dukungan finansial yang diterima dari DIPA BLU Universitas Lampung Tahun Anggaran 2025. Bantuan dana ini dialokasikan melalui Skema Penelitian Dasar dengan perjanjian kontrak 687/UN26.21/PN/2025, yang secara signifikan mendukung penyediaan sumber daya komputasi dan akuisisi data yang krusial untuk keberhasilan pelaksanaan studi ini.

DAFTAR PUSTAKA

[1] United Nations General Assembly, "Transforming our world: the 2030 Agenda for Sustainable Development (A/RES/70/1)," 2015. [Online]. Available: https://sustainabledevelopment.un.org/post2015/transformingourworld/publication

[2] W. L. Filho, J. Sierra, E. Price, J. H. P. P. Eustachio, A. Novikau, M. Kirrane, and A. L. Salvia, "The role of universities in accelerating the sustainable development goals in Europe," Scientific Reports, vol. 14, no. 1, p. 15464, 2024.

[3] E. De la Poza, P. Merello, A. Barberá, and A. Celani, "Universities' reporting on SDGs: Using the impact rankings to model and measure their contribution to sustainability," Sustainability, vol. 13, no. 4, p. 2038, 2021. [4]

[4] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, and L. He, "A survey on text classification: From traditional to deep learning," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 13, no. 2, pp. 1-41, 2022.

[5] F. M. Kwale, "A critical review of k means text clustering algorithms," International Journal of Advanced Research in Computer Science, vol. 4, no. 9, pp. 1-9, 2013.

[6] A. A. Khan, M. S. Bashir, A. Batool, M. S. Raza, and M. A. Bashir, "K-Means Centroids Initialization Based on Differentiation Between Instances Attributes," International Journal of Intelligent Systems, p. 7086878, 2024.

[7] L. Wang, "Text sentiment analysis method based on support vector machine and long short-term memory network," in Proc. 2023 4th Int. Conf. Computing, Networks and Internet of Things, 2023, pp. 87-91.

[8] Y. Huang, Y. Jiang, T. Hasan, Q. Jiang, and C. Li, "A topic BiLSTM model for sentiment classification," in Proc. 2nd Int. Conf. Innovation in Artificial Intelligence, 2018, pp. 143-147.

[9] A. Duan and R. C. Raga, "BiLSTM model with Attention mechanism for multi-label news text classification," in 2024 4th International Conference on Neural Networks, Information and Communication (NNICE), 2024, pp. 566-569.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[11] H. Schütze, C. D. Manning, and P. Raghavan, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge University Press, 2008.

[12] Z. Abidin and A. Junaidi, "Text stemming and lemmatization of regional languages in Indonesia: a systematic literature review," Journal of Information Systems Engineering and Business Intelligence, vol. 10, no. 2, pp. 217-231, 2024.

[13] Pusat Informasi dan Humas, Fakultas Vokasi Universitas Airlangga, "Pedoman kata kunci SDGs penelitian dan publikasi ilmiah," Universitas Airlangga, 2025. [Online]. Available: https://vokasi.unair.ac.id/wp-content/uploads/2025/05/Pedoman-Kata-Kunci-SDGs-Penelitian-dan-Publikasi-Ilmiah-1_opt.pdf

[14] S. Qaiser and R. Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," International Journal of Computer Applications, vol. 181, no. 1, pp. 25-29, 2018.

[15] M. Ester, H. P. Kriegel, and X. J. G. A. Xu, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794, 2016.