Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

Prediksi Promosi Pegawai dengan Stacking Ensemble dengan SMOTE-ENN dan SHAP

Andri Yudha Pratama[1,2*], Arief Hermawan[1], Donny Avianto[1]

[1]Magister Teknologi Informasi, Universitas Teknologi Yogyakarta, Sleman
Jl. Siliwangi, Jombor Lor, Sendangadi, Kec. Mlati, Kabupaten Sleman, Daerah Istimewa Yogyakarta, Indonesia
[2]Badan Kepegawaian Daerah Daerah Istimewa Yogyakarta, Yogyakarta
Jl. Kyai Mojo No.56, Bener, Kec. Tegalrejo, Kota Yogyakarta, Daerah Istimewa Yogyakarta, Indonesia
Email: [1,*]pratama.andriyudha@gmail.com, [2]ariefdb@uty.ac.id, [3]donny@uty.ac.id
Email Penulis Korespondensi: pratama.andriyudha@gmail.com
Submitted: **07/05/2026**; Accepted: **13/07/2026**; Published: **20/07/2026**

Abstrak−Paradigma manajemen sumber daya manusia di era digital menuntut proses promosi pegawai yang objektif dan berbasis data. Namun, ketidakseimbangan kelas yang ekstrem (rasio 10,74:1) berpotensi menimbulkan bias terhadap kelompok minoritas yang layak dipromosikan. Penelitian ini mengusulkan kerangka klasifikasi Stacking Ensemble yang terdiri atas Random Forest, XGBoost, dan LightGBM sebagai base learner, serta Logistic Regression sebagai meta-learner, dengan integrasi SMOTE-ENN dan interpretabilitas SHAP dua tingkat. Penelitian ini menunjukkan bahwa penerapan SMOTE-ENN sebelum cross-validation dapat menghasilkan estimasi performa yang bias hingga +110% pada F1-Score, sehingga diusulkan penggunaan imblearn.Pipeline yang membatasi resampling hanya pada training fold. Berdasarkan evaluasi menggunakan 10-fold stratified cross-validation yang bebas dari data leakage, model Stacking Ensemble memperoleh Accuracy sebesar 0,9022, Precision 0,4342, Recall 0,4889, F1-Score 0,4598, dan AUC-ROC 0,8053. Meskipun tidak menghasilkan F1-Score tertinggi, model ini mencapai nilai Recall terbaik di antara model kompetitif sehingga relevan untuk konteks yang sensitif terhadap kesalahan false negative. Analisis SHAP mengidentifikasi avg_training_score, usia, dan performance_index sebagai determinan utama keputusan promosi. Kerangka yang diusulkan memberikan kontribusi metodologis dalam evaluasi model pada data tidak seimbang sekaligus menawarkan sistem pendukung keputusan yang lebih transparan dan akuntabel untuk mendukung penerapan meritokrasi pada organisasi pemerintah maupun korporasi.

Kata Kunci: Stacking Ensemble; SMOTE-ENN; Explainable AI; SHAP; Data Mining;

Abstract−The paradigm of human resource management in the digital era demands an objective and data-driven employee promotion process. However, the extreme class imbalance (ratio 10.74:1) has the potential to introduce bias against minority groups that deserve promotion. This study proposes a stacking ensemble classification framework consisting of Random Forest, XGBoost, and LightGBM as base learners and Logistic Regression as a meta-learner, with the integration of SMOTE-ENN and two-level SHAP interpretability. This study shows that the application of SMOTE-ENN before cross-validation can result in a biassed performance estimate of up to +110% on the F1-Score; thus, the use of imblearn.Pipeline is proposed, which restricts resampling only to the training fold. Based on the evaluation using 10-fold stratified cross-validation free from data leakage, the stacking ensemble model achieved an accuracy of 0.9022, precision of 0.4342, recall of 0.4889, F1-score of 0.4598, and AUC-ROC of 0.8053. Although it did not achieve the highest F1 score, this model attained the best recall value among competitive models, making it relevant for contexts sensitive to false negative errors. SHAP analysis identifies avg_training_score, age, and performance_index as the main determinants of promotion decisions. The proposed framework provides a methodological contribution to model evaluation on imbalanced data while offering a more transparent and accountable decision support system to support the implementation of meritocracy in both government and corporate organisations.

Keywords: Stacking Ensemble; SMOTE-ENN; Explainable AI; SHAP; Data Mining;

1. PENDAHULUAN

Paradigma manajemen sumber daya manusia di era digital mengalami pergeseran besar dari pendekatan administratif konvensional menuju kerangka pengambilan keputusan yang lebih strategis dan berbasis data. Kemajuan ini didorong oleh penggunaan teknologi kecerdasan buatan dan analisis sumber daya manusia (HR), yang memungkinkan perusahaan untuk mengelola sumber daya manusia secara lebih akurat, adaptif, dan berbasis data [1]. Transformasi teknologi dalam manajemen ini sangat penting untuk mendukung penerapan sistem meritokrasi dalam pengelolaan Aparatur Sipil Negara (ASN), yang menekankan prinsip transparansi, keadilan, dan akuntabilitas dalam setiap proses manajemen kepegawaian [2]. Pendekatan berbasis data (data-driven decision making) telah menjadi keharusan dalam era transformasi digital, termasuk dalam manajemen sumber daya manusia [3]. Untuk menjamin keadilan dan kebenaran dalam manajemen kepegawaian, proses kenaikan pangkat Pegawai Negeri Sipil (PNS) harus dilakukan secara jujur, terbuka, dan bebas dari bias struktural dan praktik favoritisme karena proses ini merupakan evaluasi menyeluruh atas kinerja dan dedikasi pegawai [4], [5]. Teknik klasifikasi prediktif dalam machine learning menjadi strategi penting dalam manajemen sumber daya manusia untuk memastikan keputusan promosi yang adil dan berbasis data [6].

Bergantung pada penilaian kualitatif oleh atasan langsung dapat menyebabkan bias, yang dapat menyebabkan ketidaksetaraan dan ketidakkonsistenan dalam penerapan standar evaluasi kinerja [7]. Kegagalan sistem dalam praktik konvensional dapat menyebabkan pegawai yang memiliki kompetensi teknis yang layak

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

untuk dipromosikan terabaikan. Pada akhirnya, hal ini dapat mengurangi motivasi organisasi dan menurunkan kepercayaan pegawai terhadap proses pengambilan keputusan [1], [8].

Tantangan teknis yang turut memperburuk inefisiensi administratif tersebut adalah munculnya fenomena ketidakseimbangan kelas (class imbalance), yang secara inheren kerap ditemukan dalam data kepegawaian [9]. Hanya sekitar 8,5 persen dari populasi data dalam dataset HR Analytics: Employee Promotion Data mencakup pekerja minoritas yang mendapatkan promosi [4]. Kondisi ketidakseimbangan kelas ini telah berkembang menjadi masalah prediktif yang signifikan karena model pembelajaran mesin konvensional berfokus pada prediksi yang dibuat oleh kelas mayoritas untuk mengurangi kesalahan global. Akibatnya, model menjadi kurang peka untuk mengidentifikasi karakteristik khusus dari pegawai yang berprestasi dari kelas minoritas, yang dapat menyebabkan deteksi yang kurang akurat terhadap kandidat yang layak dipromosikan [6], [10]. Untuk menyelesaikan masalah ketidakseimbangan kelas, teknik interpolasi sampel sintetis kelas minoritas (SMOTE) digunakan untuk menyelesaikan distribusi data yang lebih seimbang dan sensitivitas model yang lebih tinggi [11].

Penelitian terdahulu telah mengeksplorasi penggunaan berbagai algoritma, termasuk Decision Tree dan Logistic Regression yang menunjukkan peningkatan akurasi. Namun, metode ini menghasilkan bias terhadap kelas mayoritas karena memiliki keterbatasan saat menangani masalah ketidakseimbangan kelas [12]. Teknik SMOTE telah terbukti mampu meningkatkan kualitas prediksi pada dataset HR Analytics, namun metode ini biasanya tidak menggunakan strategi ensemble stacking atau elemen explainability untuk menginterpretasikan hasil model [13]. Studi komparatif terhadap berbagai algoritma menunjukkan bahwa K-Nearest Neighbors (KNN) mampu mencapai nilai F1-Score tertinggi sebesar 0,8920 ketika dikombinasikan dengan SMOTE. namun terbatas pada satu model tanpa komponen interpretabilitas untuk menjelaskan hasil prediksi [6]. Penelitian lain menunjukkan bahwa algoritma Naive Bayes dengan pra-pemrosesan SMOTE masih memiliki masalah dengan kinerja. Ini karena asumsi independensi antar fitur tidak sesuai dengan kompleksitas dan atribut data HR Analytics [2]. Selain itu, penelitian sebelumnya menunjukkan bahwa algoritma Random Forest mampu mencapai nilai recall tertinggi sebesar 0,9466 ketika dikombinasikan dengan teknik SMOTE, dibandingkan dengan algoritma lain seperti Logistic Regression, XGBoost, dan Gradient Boosting. Namun demikian, pendekatan tersebut masih memiliki keterbatasan karena belum mengadopsi strategi stacking maupun analisis Explainable Artificial Intelligence (XAI) yang berfokus pada interpretasi individual terhadap hasil prediksi [14]. Penelitian yang lain membangun model Random Forest yang dikombinasikan dengan SMOTE hanya menghasilkan nilai recall sebesar 26,55% dan nilai F1-Score sebesar 40,55%. Hasil ini menunjukkan bahwa penggunaan model tunggal belum mampu mengatasi masalah ketidakseimbangan kelas tanpa bantuan strategi ensemble stacking atau teknik resampling hibrida [4]. Penelitian tambahan menyarankan pipeline Hybrid AutoML berbasis TPOT untuk pengaturan hyperparameter dan pemilihan model. Metode Hybrid AutoML dapat mencapai akurasi sebesar 0,97 dan F1-Score sebesar 0,94 dengan kombinasi teknik SMOTE+ENN. Ini menunjukkan kinerja prediktif yang kompetitif, tetapi metode ini tidak menyediakan local interpretability pada tingkat individu, elemen ini sangat penting untuk mendukung pengambilan keputusan yang adil, jelas, dan dapat dipertanggungjawabkan bagi setiap pekerja [8].

Berdasarkan tinjauan literatur, research gap pada domain ini dapat diidentifikasi dalam tiga dimensi yang saling berkaitan. Pertama, dimensi akurasi. Penelitian terdahulu yang menggunakan SMOTE tunggal maupun kombinasi model tunggal belum mampu menghasilkan keseimbangan precision dan recall secara optimal pada kondisi ketidakseimbangan kelas yang ekstrem (>10:1). Kedua, dimensi bias metodologis. Sebagian besar penelitian yang menerapkan teknik resampling pada dataset serupa belum secara eksplisit melaporkan pemisahan proses resampling dari skema cross-validation. Kondisi ini berpotensi menimbulkan data leakage yang dapat menginflasi nilai F1-Score. Namun, besarnya pengaruh tersebut pada penelitian terdahulu tidak dapat diverifikasi lebih lanjut karena tidak tersedia akses terhadap kode implementasi aslinya. Ketiga, dimensi interpretabilitas penelitian yang mengintegrasikan pendekatan XAI umumnya masih terbatas pada interpretasi model tunggal atau base learner dan belum mencakup interpretasi pada meta-learner dalam arsitektur stacking yang secara struktural menggabungkan keluaran dari beberapa model dasar.

Untuk mengatasi limitasi tersebut, penelitian ini mengintegrasikan tiga inovasi metodologis secara terpadu. Pertama, teknik hybrid resampling SMOTE-ENN yang menggabungkan oversampling dengan pembersihan data aktif melalui Edited Nearest Neighbors (ENN). Kedua, arsitektur Stacking Ensemble multi-level dengan tiga base learner heterogen (Random Forest, XGBoost, LightGBM) dan Logistic Regression sebagai meta-learner. Ketiga, integrasi SHAP (SHapley Additive exPlanations) sebagai mekanisme Explainable AI yang menghasilkan interpretasi global sekaligus lokal per individu pegawai. Transparansi dan akuntabilitas yang dihadirkan XAI sangat penting bagi auditor dan pengambil keputusan dalam memahami logika di balik prediksi model [15], [16] , [17]. Metode SHAP, yang berlandaskan teori permainan kooperatif Shapley, dinilai sebagai salah satu pendekatan interpretabilitas terbaik yang tersedia karena mampu mengukur kontribusi setiap fitur secara proporsional, aditif, dan konsisten [18], [19].

Untuk menjembatani celah (research gap) tersebut, penelitian ini bertujuan pertama untuk mengimplementasikan Synthetic Minority Oversampling Technique dan Edited Nearest Neighbor (SMOTE-ENN) sebagai teknik hybrid resampling yang menggabungkan oversampling dan data cleaning untuk menghasilkan distribusi data latih yang lebih bersih, kedua membangun Stacking Ensemble Classifier multi-level dengan tiga base learner heterogen Random Forest, XGBoost, dan LightGBM yang dikombinasikan melalui Logistic Regression sebagai meta-learner, ketiga melakukan rekayasa fitur melalui penciptaan empat fitur baru berbasis

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

domain knowledge HR, dan keempat mengintegrasikan SHAP sebagai mekanisme Explainable AI untuk menghasilkan interpretasi global sekaligus lokal per individu pegawai. Oleh karena itu, penelitian ini diharapkan menghasilkan kerangka metodologi terpadu yang tidak hanya memiliki kinerja prediktif yang lebih baik, tetapi juga dapat memenuhi kebutuhan praktis dalam manajemen SDM kontemporer, khususnya dalam menyediakan sistem pendukung keputusan yang dapat adil, transparan, dan dipertanggungjawabkan kontribusi nyata menuju transformasi birokrasi kepegawaian yang lebih meritokratis dan kredibel [4], [8].

2. METODOLOGI PENELITIAN

2.1 Deskripsi Dataset dan Fitur

Penelitian ini menggunakan dataset HR Analytics: Employee Promotion Data, yang disediakan oleh Arash Nic dari platform Kaggle [20]. Dataset ini menampilkan data pegawai dari sebuah perusahaan multinasional yang sedang melakukan proses untuk menemukan kandidat untuk promosi untuk kelompok jabatan tertentu. Dataset yang digunakan terdiri dari 54.808 baris data latihan (train) dengan 13 kolom dan 23.490 baris data uji (test) dengan 12 kolom (tanpa label target). Dua atribut memiliki nilai hilang (missing values), previous_year_rating sebanyak 4.124 baris (7,52%) dan education sebanyak 2.409 baris (4,40%). Selain itu, ketidakseimbangan yang sangat signifikan dalam distribusi variabel target. Rasio ketidakseimbangan sebesar 10,74:1 terjadi karena hanya 4.668 pegawai (8,52%) yang dipromosikan, sementara 50.140 pegawai (91,48%) tidak menerima promosi. Kondisi ini dapat dikategorikan sebagai imbalanced classification yang berat (sangat tidak seimbang), di mana model klasifikasi konvensional tanpa perawatan khusus cenderung mengalami penurunan kinerja yang signifikan, terutama dalam menemukan dan memprediksi kelas minoritas [21]. Berbagai variabel, termasuk metrik kuantitatif kinerja historis, demografi, dan latar belakang pendidikan, membentuk struktur multidimensional dataset ini. Variabel target (is_promoted) memiliki nilai biner 1 yang menunjukkan bahwa pegawai telah dipromosikan, dan nilai 0 menunjukkan bahwa tidak ada promosi. Dataset ini dipilih berdasarkan tingkat relevansinya terhadap tujuan penelitian dan kelengkapan variabel yang dapat secara sistematis mendukung proses analisis secara keseluruhan.

2.2 Tahapan Penelitian

Dengan menggunakan alur kerja data mining yang terstruktur, tujuan penelitian disusun secara sistematis. Ini memastikan bahwa seluruh proses mulai dari pengolahan data, pemodelan, hingga evaluasi berjalan dengan baik dan memiliki tingkat replikasi yang tinggi [22]. Ekplorasi data memerlukan perencanaan metodologi yang sistematis dalam klasifikasi berbasis machine learning agar model analitis yang dihasilkan dapat diinterpretasikan dengan jelas dan mencapai kinerja prediktif yang presisi [23]. Untuk memastikan bahwa model dapat mengidentifikasi pola di kelas minoritas secara akurat tanpa menimbulkan bias, diperlukan kombinasi metodologis antara teknik penyeimbangan data pada tahap pra-pemrosesan dan algoritma pemodelan yang kuat [24]. Gambar 1 menyajikan diagram alir tahapan penelitian yang menunjukkan mulai dari pengumpulan data hingga evaluasi sistematis hasil prediksi.

Gambar 1. Diagram Alir Tahapan Penelitian Prediksi Klasifikasi Promosi Pegawai

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

Berdasarkan gambar 1, metodologi penelitian ini disusun melalui tahapan yang sistematis dengan menekankan integritas proses evaluasi prediktif. Setelah tahapan Exploratory Data Analysis (EDA), prapemrosesan data, dan feature engineering untuk menghasilkan empat atribut baru, proses penyeimbangan data tidak dilakukan secara global. Untuk menghindari terjadinya data leakage pada pemodelan data tidak seimbang, teknik hybrid resampling SMOTE-ENN diintegrasikan ke dalam imblearn.Pipeline. Melalui skema ini, pada setiap iterasi 10-Fold Stratified Cross-Validation, proses resampling hanya diterapkan pada training fold, sedangkan validation fold tetap mempertahankan distribusi kelas aslinya dengan rasio ketidakseimbangan sebesar 10,74:1. Model prediktif dibangun menggunakan arsitektur Stacking Ensemble bertingkat yang terdiri atas Random Forest, XGBoost, dan LightGBM sebagai base learners, serta Logistic Regression sebagai meta-learner. Setelah performa generalisasi model dievaluasi secara objektif dan terbebas dari bias akibat data leakage, pendekatan Explainable AI berbasis SHAP diterapkan pada dua tingkat interpretasi. TreeExplainer digunakan untuk mengidentifikasi kontribusi fitur pada masing-masing base learner, sedangkan KernelExplainer diterapkan pada meta-learner untuk menginterpretasikan mekanisme kalibrasi prediksi akhir. Tahapan penelitian diakhiri dengan pelatihan model final menggunakan seluruh data latih dan penerapannya untuk memprediksi kelayakan promosi pada 23.490 pegawai dalam test set.

2.3 Preprocessing dan Feature Engineering

Tiga proses utama terlibat dalam tahap preprocessing penelitian ini. Pertama, penanganan nilai hilang atribut pendidikan dengan 4,40% data hilang diimputasi dengan nilai modus (Bachelor's), sedangkan atribut tahun sebelumnya dengan 7,52% data hilang diimputasi dengan nilai median (3,0). Metode median dinilai lebih kuat dalam menjaga representasi data karena didasarkan pada karakteristik distribusi data yang tidak normal dan sensitivitas terhadap outlier [25]. Kedua, variabel kategori (departemen, wilayah, pendidikan, gender, dan kanal rekrutmen) dikodekan menggunakan label encoding. Ketiga, untuk memastikan bahwa semua fitur memiliki skala yang sebanding, StandardScaler digunakan untuk melakukan standardisasi fitur numerik. Tahap ini sangat penting untuk mendukung kinerja meta-learner Logistic Regression, yang sensitif terhadap perbedaan skala antara fitur [26]. Proses feature engineering merupakan salah satu kontribusi kebaruan (novelty) dalam penelitian ini. Empat fitur baru dirancang berdasarkan domain knowledge pada bidang Human Resources (HR), yang secara sistematis disajikan pada Tabel 1.

Tabel 1. Fitur Rekayasa Baru (Feature Engineering)

Nama Fitur	Formula	Justifikasi
performance_index	avg_training_score × previous_year_rating	Mengukur produktivitas kumulatif pegawai secara holistik
experience_score	length_of_service × no_of_trainings	Akumulasi pembelajaran dari durasi kerja dan intensitas pelatihan
awards_training_combo	awards_won × avg_training_score	Sinergi antara pengakuan formal dan kompetensi teknis
age_service_ratio	age / (length_of_service + 1)	Efisiensi karir: kecepatan pencapaian pengalaman per usia

Tabel 1 menunjukkan empat fitur rekayasa yang dibuat menggunakan pengetahuan domain dalam bidang manajemen SDM. Dengan menggunakan interaksi multiplikatif antara skor pelatihan dan hasil penilaian kinerja tahun sebelumnya, fitur kinerja indeks dimaksudkan untuk menunjukkan produktivitas pegawai secara keseluruhan. Selain itu, skor pengalaman dimaksudkan untuk menggambarkan kombinasi kemampuan pegawai yang terdiri dari masa kerja dan tingkat partisipasi dalam program pelatihan. Untuk menggabungkan pengakuan formal atas prestasi dengan kemampuan teknis yang dimiliki, fitur awards_training_combo dirancang. Namun, fitur age_service_ratio menunjukkan efisiensi perjalanan karir pegawai, terutama tentang seberapa besar pengalaman kerja dapat meningkat secara proporsional terhadap usia. Selanjutnya, keempat fitur tersebut diintegrasikan ke dalam proses pelatihan model bersama dengan fitur asli dataset.

2.4 SMOTE-ENN Hybrid Resampling dan Pencegahan Data Leakage

SMOTE-ENN adalah teknik resampling hibrida yang menggabungkan dua metode yang saling melengkapi Synthetic Minority Over-sampling Technique (SMOTE) dan Edited Nearest Neighbors (ENN). Berbeda dengan metode sebelumnya, yang biasanya hanya menggunakan SMOTE konvensional dalam penelitian, SMOTE-ENN bekerja melalui dua tahapan utama. Pada tahap pertama, sampel sintetis dari kelas minoritas dihasilkan melalui oversampling menggunakan SMOTE dengan parameter k = 5 dari tetangga terdekat. Pada tahap kedua, sampel yang tidak konsisten atau salah diklasifikasikan dari tetangga terdekat, baik dari kelas mayoritas maupun minoritas, dihapus menggunakan ENN [27]. Hasil penggunaan SMOTE-ENN pada data latihan menunjukkan peningkatan yang signifikan dalam keseimbangan distribusi kelas. Sebelum proses resampling, terdapat 50.140 sampel kelas 0 yang tidak dipromosikan dan 4.668 sampel kelas 1 yang dipromosikan. Setelah penerapan SMOTE-ENN, distribusi berubah menjadi 37.433 sampel kelas 0 dan 46.108 sampel kelas 1, sehingga totalnya 83.541 sampel. SMOTE-ENN tidak hanya meningkatkan jumlah sampel untuk kelas minoritas, tetapi juga menghilangkan

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

beberapa sampel dari kelas mayoritas yang dikenal sebagai suara. Oleh karena itu, dataset yang dibuat tidak hanya lebih seimbang, tetapi juga berkualitas tinggi untuk mendukung proses pelatihan model [28]. Evaluasi performa model menerapkan SMOTE-ENN yang diintegrasikan dalam imblearn.Pipeline bersama setiap algoritma klasifikasi. Dengan pendekatan ini, proses resampling hanya dilakukan pada training fold di setiap iterasi 10-Fold Stratified Cross-Validation, sedangkan fold validasi tetap menggunakan distribusi kelas asli dan tidak pernah menerima sampel sintetis hasil resampling. Pemisahan data latih dan validasi dilakukan terlebih dahulu pada data asli, kemudian SMOTE-ENN diterapkan hanya pada data latih. Dengan demikian, data validasi tetap mempertahankan rasio ketidakseimbangan asli sebesar 10,74:1 sehingga lebih merepresentasikan kondisi operasional model di dunia nyata. Sebaliknya, pada tahap pelatihan model final untuk prediksi test set, penggunaan seluruh data hasil resampling tidak menimbulkan data leakage karena test set bersifat independen dan tidak pernah terlibat dalam proses resampling maupun pelatihan sebelumnya.

2.5 Stacking Ensemble Classifier

Stacking Ensemble yang diusulkan dalam penelitian ini menggunakan arsitektur dua level. Pada level pertama, terdapat tiga base learner heterogen yang dilatih secara paralel, yaitu Random Forest dengan 200 decision tree, max_depth = 10, dan min_samples_leaf = 4. XGBoost dengan 200 estimator, learning_rate = 0,05, subsample = 0,8, dan colsample_bytree = 0,8; serta (3) LightGBM dengan 200 estimator, num_leaves = 63, learning_rate = 0,05, dan subsample = 0,8. Pemilihan ketiga model tersebut didasarkan pada karakteristiknya yang saling melengkapi. Mekanisme bagging, yang menggabungkan prediksi dari berbagai pohon keputusan secara bersamaan, memberikan stabilitas dan ketahanan terhadap overfitting. Tingkat presisi yang tinggi dimungkinkan oleh XGBoost melalui pendekatan boosting sekuensial dengan fungsi objektif yang terregularisasi. Sementara itu, algoritma berbasis histogram LightGBM meningkatkan efisiensi komputasi dan mampu mengurangi kebutuhan memori secara signifikan, khususnya untuk dataset berskala besar [29], [30].

Ketiga pendekatan pembelajaran base learner metode split tepat dibandingkan dengan histogram dan pendekatan bagging dibandingkan dengan boosting memungkinkan model untuk menangkap pola-pola yang saling melengkapi. Pada akhirnya, ini menghasilkan prediksi agregat yang lebih kuat dan dapat diandalkan daripada prediksi yang dibuat secara terpisah oleh masing-masing model [31]. Pada level kedua, Logistic Regression digunakan sebagai meta-learner dengan parameter C=1,0. Model ini menerima keluaran probabilitas dari ketiga base learner (total 6 nilai probabilitas, yaitu 2 per model × 3 model), dan kemudian menggunakan mekanisme passthrough=True untuk menggabungkan semua fitur asli. Konfigurasi ini menghasilkan tambahan 15 fitur asli, sehingga meta-learner menggunakan total 21 fitur. Dengan menggunakan passthrough=True, meta-learner dapat bergantung pada prediksi base learner dan memiliki akses langsung ke informasi asli dalam dataset [30]. Ini memungkinkan untuk menemukan pola atau hubungan yang tidak terakomodasi pada level pertama. Untuk menjaga validitas model dan mencegah data hilang antara level pertama dan kedua, proses stacking dilakukan dengan cross-validation internal menggunakan skema 5-Fold. Metode ini memastikan bahwa meta-learner dilatih menggunakan prediksi di luar kotak, sehingga representasi yang dipelajari mencerminkan performa generalisasi model secara lebih realistis [31].

2.6 Metrik Evaluasi

Agar menghindari interpretasi kinerja model yang menyesatkan, sangat penting untuk memilih metrik evaluasi yang tepat mengingat tingkat ketidakseimbangan kelas yang sangat tinggi pada dataset ini (rasio 10,74:1). Satu metrik, seperti ketepatan, tidak cukup representatif dalam kondisi klasifikasi yang tidak seimbang karena cenderung bias terhadap kelas mayoritas. Oleh karena itu, lima metrik evaluasi tambahan digunakan dalam penelitian ini yaitu Precision, recall, F1-Score, akurasi, dan AUC-ROC (kemampuan model untuk membedakan kelas pada berbagai ambang keputusan). Metode ini memungkinkan penilaian yang lebih menyeluruh terhadap kinerja model, terutama dalam identifikasi kelas minoritas dengan benar. F1-Score dianggap sebagai metrik utama karena kemampuan untuk menggabungkan ketepatan dan recall ke dalam satu nilai yang lebih representatif. Ini relatif tidak terpengaruh oleh dominasi kelas mayoritas, sehingga dinilai lebih akurat daripada ketepatan dalam kondisi klasifikasi yang tidak seimbang [32]. Untuk menggambarkan kemampuan diskriminasi model secara keseluruhan, AUC-ROC melengkapi F1-Score dengan memberikan evaluasi yang bersifat threshold-independent. Namun, dalam kondisi ketidakseimbangan kelas yang ekstrim, nilai AUC-ROC dapat tampak tinggi meskipun kinerja model terhadap kelas minoritas sebenarnya kurang baik. Oleh karena itu, interpretasi AUC-ROC harus dilakukan bersamaan dengan F1-Score untuk menghasilkan penilaian yang lebih akurat [33].

Untuk memastikan hasil yang stabil, tidak bias, dan memiliki kemampuan generalisasi yang baik terhadap data yang belum pernah dilihat sebelumnya, metode cross-validation stratified 10-fold digunakan untuk menghitung semua metrik evaluasi. Pendekatan Stratified K-Fold secara khusus menjaga proporsi kelas target tetap konsisten pada setiap fold evaluasi, meskipun data telah melalui proses resampling menggunakan SMOTE-ENN. Hal ini penting untuk menghindari terjadinya distributional shift yang dapat menghasilkan estimasi kinerja model yang tidak realistis, sehingga evaluasi yang diperoleh lebih merepresentasikan performa model secara aktual [34].

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

3. HASIL DAN PEMBAHASAN

3.1 Hasil Eksplorasi Data (EDA)

Analisis eksploratif terhadap dataset latih mengungkap karakteristik distribusi yang krusial untuk pemahaman mendalam tentang fenomena promosi pegawai. Distribusi variabel target merupakan temuan awal dan paling penting, sebagaimana diilustrasikan pada Gambar 2.

Gambar 2. Distribusi Kelas Target Promosi Pegawai

Gambar 2 memperlihatkan ketidakseimbangan yang sangat ekstrem pada distribusi variabel target hanya 4.668 pegawai (8,52%) yang mendapat promosi dari total 54.808 pegawai, dengan rasio imbalance 10,74:1. Kondisi ini jauh melebihi ambang batas imbalance moderat (1:3 sampai 1:5) yang umum dibahas dalam literatur, sehingga memerlukan penanganan khusus yang lebih agresif dibandingkan sekadar oversampling sederhana.

Pegawai yang dipromosikan memiliki nilai rata-rata avg_training_score yang jauh lebih tinggi (sekitar ±76) dibandingkan dengan pegawai yang tidak dipromosikan. Perbedaan ini tidak hanya signifikan secara statistik, tetapi juga sesuai dengan pemahaman konseptual bahwa kemampuan teknis adalah salah satu syarat utama untuk promosi. Namun, distribusi usia (age) menunjukkan pola bimodal, dengan dua puncak utama pada rentang usia 25-30 tahun dan 35-42 tahun. Proporsi promosi yang lebih tinggi ditemukan di kelompok usia menengah, terutama di rentang usia 30-40 tahun. Hal ini menunjukkan bahwa pengalaman kerja dan tingkat kematangan profesional sangat memengaruhi peluang promosi pegawai.

Analisis korelasi menunjukkan bahwa variabel target is_promoted memiliki hubungan positif tertinggi dengan variabel avg_training_score (r = 0,157) dan previous_year_rating (r = 0,111). Mengingat jumlah sampel yang besar, nilai korelasi ini masih signifikan dan dapat menunjukkan variabel yang berpengaruh secara awal. Atribut awards_won? menghasilkan korelasi linear yang relatif rendah (r = 0,07), tetapi menunjukkan kekuatan diskriminatif yang signifikan ketika dikombinasikan dengan fitur lain dalam model non-linear, terutama dalam pendekatan tree-based ensemble. Temuan ini kemudian diperkuat melalui analisis SHAP, yang menunjukkan kontribusi signifikan atribut tersebut dalam membentuk prediksi model secara keseluruhan.

Untuk variabel previous_year_rating ada 7,52% data yang hilang (missing values), ini menunjukkan bahwa pegawai baru tidak memiliki catatan penilaian kinerja tahunan. Oleh karena itu, imputasi dengan nilai median (3,0) dipilih karena distribusi data cenderung skewed ke kiri dan ada beberapa outlier pada nilai rendah, sehingga pendekatan ini lebih robust dalam menjaga representasi data. Ada 4,40% data yang hilang dari variabel pendidikan, yang diimputasi dengan nilai modus (Bachelor), yang sesuai dengan proporsi dominan tingkat pendidikan dalam dataset.

3.2 Hasil SMOTE-ENN Hybrid Resampling

Penerapan SMOTE-ENN menghasilkan transformasi distribusi kelas yang signifikan sebagaimana disajikan pada Tabel 2. Perbedaan mendasar antara SMOTE-ENN dan SMOTE konvensional terletak pada peran fase ENN, yang tidak hanya melengkapi proses oversampling, tetapi juga secara aktif melakukan pembersihan data dengan mengeliminasi sampel yang dianggap ambigu atau noisy dari kedua kelas.

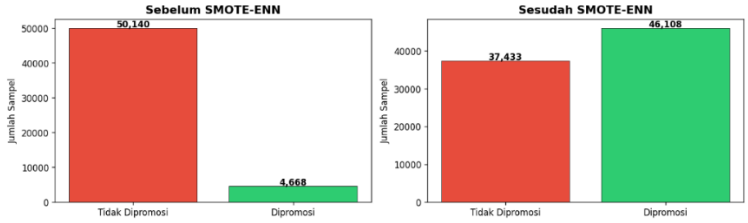
Tabel 2. Perbandingan Distribusi Kelas Sebelum dan Sesudah SMOTE-ENN

Kondisi	Kelas 0 (Tidak Promosi)	Kelas 1 (Dipromosi)	Total Sampel	Rasio Imbalance
Sebelum SMOTE-ENN	50.140 (91,48%)	4.668 (8,52%)	54.808	10,74:1
Sesudah SMOTE-ENN	37.433 (44,81%)	46.108 (55,19%)	83.541	≈ 0,81:1

Tabel 2 menunjukkan bahwa penerapan SMOTE-ENN tidak hanya meningkatkan jumlah sampel pada kelas minoritas dari 4.668 menjadi 46.108 (peningkatan sebesar +887%), tetapi juga mengurangi jumlah sampel pada kelas mayoritas dari 50.140 menjadi 37.433 (penurunan sebesar -25,4%). Pengurangan kelas mayoritas tersebut disebabkan oleh fase ENN, yang secara selektif menghapus 12.707 sampel kelas 0 yang berada di sekitar

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

batas keputusan dengan kelas 1 dan berpotensi menyebabkan ambiguitas selama proses pelatihan model. Hasilnya, distribusi kelas menjadi lebih seimbang, dengan rasio hampir 0,81:1, yang jauh lebih baik untuk proses pembelajaran model daripada rasio awal 10,74:1. Gambar 3 menunjukkan perbedaan distribusi kelas secara visual sebelum dan sesudah intervensi SMOTE-ENN.



Gambar 3. Perbandingan Distribusi Kelas Sebelum dan Sesudah SMOTE-ENN

Gambar 3 menunjukkan perubahan distribusi kelas setelah penerapan SMOTE-ENN. Terlihat bahwa distribusi kelas 0 (tidak dipromosikan) yang dominan sebelumnya direstorasi menjadi hampir seimbang, dan distribusi kelas 1 (dipromosikan) menjadi sedikit lebih dominan sebesar 55,19% setelah proses resampling. Pola ini menunjukkan bahwa SMOTE-ENN berhasil menangani masalah ketidakseimbangan kelas yang ekstrim. Selain itu, distribusi yang lebih seimbang menciptakan kondisi pembelajaran yang lebih baik bagi model Stacking Ensemble untuk mempelajari pola pada kelas minoritas dan menghasilkan hasil prediktif yang lebih stabil.

3.3 Perbandingan Performa Model

Evaluasi komprehensif dilakukan terhadap delapan model menggunakan skema 10-Fold Stratified Cross-Validation berbasis imblearn.Pipeline, sehingga SMOTE-ENN hanya diterapkan pada training fold di setiap iterasi guna mencegah data leakage. Hasil evaluasi 10-Fold Cross-Validation pada data SMOTE-ENN disajikan secara lengkap pada Tabel 3.

Tabel 3. Perbandingan Performa Model 10-Fold Cross-Validation + SMOTE-ENN

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Naïve Bayes	0,8793	0,2967	0,3042	0,3003	0,7156
Decision Tree	0,8392	0,258	0,473	0,3338	0,6732
KNN	0,7083	0,1597	0,5688	0,2494	0,6786
Logistic Regression	0,61	0,1427	0,7146	0,2379	0,7355
Random Forest	0,8916	0,386	0,4631	0,421	0,7815
XGBoost	0,9211	0,5469	0,4325	0,4829	0,8075
LightGBM	0,917	0,5148	0,4402	0,4745	0,8098
Stacking Ensemble	0,9022	0,4342	0,4889	0,4598	0,8053

Tabel 3 menunjukkan XGBoost mencapai F1-Score tertinggi (0,4829), diikuti LightGBM (0,4745), sementara Stacking Ensemble yang diusulkan menempati peringkat ketiga (0,4598). Meskipun tidak unggul pada F1-Score agregat, Stacking Ensemble mencapai Recall tertinggi di antara model berbasis pohon/ensemble yang kompetitif (0,4889, dibandingkan 0,4325–0,4631 pada XGBoost/LightGBM/Random Forest). Trade-off ini relevan secara strategis pada konteks promosi pegawai, kegagalan mengidentifikasi kandidat yang sebenarnya layak dipromosikan (False Negative) berpotensi menimbulkan kerugian organisasional berupa hilangnya talenta dan penurunan motivasi kerja, sehingga model dengan sensitivitas lebih tinggi terhadap kelas minoritas meskipun dengan precision yang sedikit lebih rendah (0,4342) dapat dinilai lebih sesuai untuk kasus penggunaan yang sensitif terhadap FN, sekalipun tidak optimal secara F1-Score murni. Logistic Regression mencapai Recall tertinggi secara keseluruhan (0,7146) namun dengan Precision sangat rendah (0,1427) yang mengindikasikan model tersebut terlalu agresif memprediksi kelas positif akibat sifatnya yang linear dan kurang mampu menangkap batas keputusan non-linear pada data HR. Kontribusi utama pendekatan stacking pada penelitian ini terletak pada keseimbangan trade-off recall-precision serta fleksibilitas interpretasi dua-level (base learner dan meta-learner) yang tidak tersedia pada model tunggal.

3.3.1 Bukti Empiris Dampak Data Leakage

Untuk mengonfirmasi secara kuantitatif potensi bias metodologis pada prosedur resampling-cross-validation, dilakukan eksperimen komparatif antara pendekatan naive, yaitu SMOTE-ENN diterapkan pada seluruh data latih sebelum cross-validation, dan pendekatan pipeline, yaitu SMOTE-ENN diterapkan hanya pada training fold. Kedua pendekatan dievaluasi menggunakan model Stacking Ensemble yang identik disajikan pada Tabel 4.

Tabel 4. Perbandingan Dampak Data Leakage Pada Hasil Evaluasi

Pendekatan	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Naive (Data Leakage)	0,9640	0,9790	0,9553	0,9670	0,9886

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

Pendekatan	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Pipeline (Bebas Leakage)	0,9022	0,4342	0,4889	0,4598	0,8053

Tabel 4 menunjukkan bahwa pendekatan naive menghasilkan performa yang secara sistematis lebih tinggi pada seluruh metrik, dengan selisih F1-Score mencapai 0,5072 poin atau setara peningkatan 110,31% dibandingkan pendekatan pipeline. Peningkatan performa yang tidak wajar ini mengindikasikan bahwa estimasi pada pendekatan naive tidak merepresentasikan kemampuan generalisasi model yang sesungguhnya, melainkan artefak dari kebocoran informasi antar-fold akibat resampling yang dilakukan sebelum pemisahan data latih dan validasi. Temuan ini menjadi bukti empiris yang mengonfirmasi kekhawatiran metodologis mengenai risiko data leakage pada skema resampling sebelum cross-validation, menjadi dasar penetapan pendekatan pipeline sebagai skema evaluasi utama pada penelitian ini.

3.4 Analisis Confusion Matrix

Analisis confusion matrix dilakukan menggunakan skema hold-out yang menerapkan pemisahan data latih dan validasi pada data asli sebelum proses resampling. SMOTE-ENN kemudian diterapkan hanya pada data latih, sehingga jumlah sampel meningkat dari 43.846 menjadi 66.796 setelah resampling. Sementara itu, data validasi yang terdiri atas 10.962 sampel tidak pernah mengalami resampling dan tetap mempertahankan rasio ketidakseimbangan asli sebesar 10,74:1. Confusion matrix dari model Stacking Ensemble disajikan pada Tabel 5.

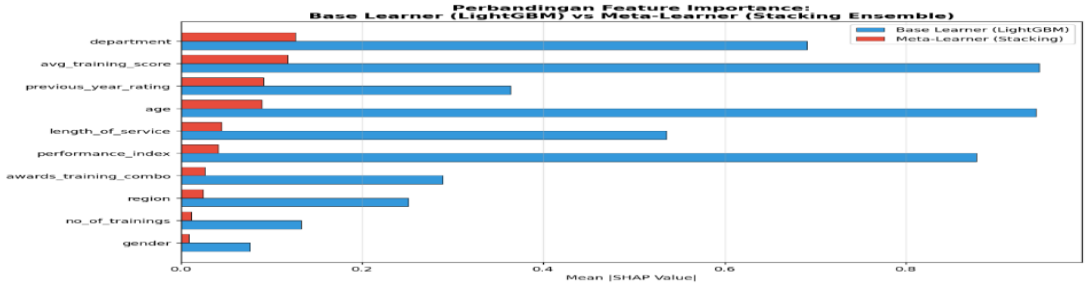
Tabel 5. Confusion Matrix dan Classification Report Stacking Ensemble

Kelas	Precision	Recall	F1-Score	Support (n)
Tidak Dipromosi (0)	0,9525	0,9347	0,9435	10.028
Dipromosi (1)	0,4162	0,5	0,4543	934
Accuracy	-	-	0,8976	10.962
Macro Avg	0,6844	0,7173	0,6989	10.962
Weighted Avg	0,9068	0,8976	0,9018	10.962
ROC-AUC Score		0,801		10.962

Tabel 5 memperlihatkan pada kelas minoritas (Dipromosi), model mencapai Recall sebesar 0,5000, dari 934 pegawai yang secara aktual layak dipromosikan, model berhasil mengidentifikasi 467 pegawai secara benar, sedangkan 467 pegawai lainnya diklasifikasikan sebagai False Negative. Nilai ini jauh lebih rendah dibandingkan hasil pada validasi yang mengalami data leakage (Recall = 0,9537) dan lebih merepresentasikan tingkat kesulitan yang sebenarnya dalam memprediksi kelas minoritas pada kondisi ketidakseimbangan ekstrem tanpa kontaminasi informasi dari data validasi. Di sisi lain, nilai Precision pada kelas promosi sebesar 0,4162 menunjukkan bahwa hanya sekitar 41,62% dari seluruh pegawai yang diprediksi layak dipromosikan benar-benar termasuk dalam kelompok yang dipromosikan pada distribusi data asli. Kondisi ini mencerminkan trade-off yang wajar karena model dirancang untuk meningkatkan sensitivitas terhadap kelas minoritas pada rasio ketidakseimbangan yang sangat tinggi (10,74:1).

3.5 Analisis SHAP Explainable AI

Analisis SHAP dilakukan menggunakan TreeExplainer pada base learner LightGBM, yang dipilih sebagai model proksi karena memiliki performa individual terbaik serta efisiensi komputasi yang lebih tinggi dibandingkan KernelExplainer pada model non-tree. Tabel berikut menyajikan sepuluh fitur teratas berdasarkan nilai Mean SHAP Value. Urutan fitur tersebut berbeda dari versi awal manuskrip dan telah diperbarui berdasarkan hasil komputasi ulang menggunakan skema evaluasi yang telah dikoreksi. Meta-learner berupa Logistic Regression menggabungkan probabilitas keluaran dari ketiga base learner bersama fitur asli melalui mekanisme passthrough. Untuk menginterpretasikan model akhir tersebut, analisis SHAP diterapkan pada fungsi predict_proba dari Stacking Ensemble menggunakan KernelExplainer. Pendekatan ini dipilih karena KernelExplainer bersifat model-agnostic dan mampu memberikan interpretasi yang valid terhadap arsitektur ensemble yang kompleks, termasuk model stacking. Gambar 4 menyajikan perbandingan feature importance base learner (LightGBM) dengan Meta-Learner (Stacking Esemble)



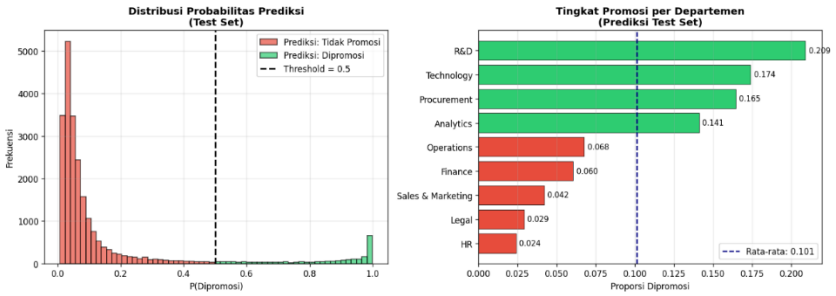
Gambar 4. Perbandingan feature importance base learner dengan Meta-Learner

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

Gambar 4 membandingkan kontribusi fitur pada dua tingkat model, yaitu base learner (LightGBM) dan meta-learner (Stacking Ensemble). Nilai Mean SHAP pada base learner berada pada rentang yang lebih tinggi (0,08–0,95) dibandingkan meta-learner (0,01–0,13). Perbedaan ini terjadi karena meta-learner tidak memproses fitur asli secara langsung, tetapi lebih mengandalkan probabilitas prediksi yang dihasilkan oleh base learner, sedangkan fitur asli hanya berfungsi sebagai sinyal pelengkap dengan bobot yang relatif kecil. Oleh karena itu, interpretasi yang lebih relevan adalah membandingkan urutan peringkat fitur, bukan besaran nilai SHAP-nya. Tiga fitur utama, yaitu avg_training_score, age, dan performance_index, secara konsisten menempati posisi teratas pada kedua tingkat model, yang diperkuat oleh nilai korelasi Spearman sebesar 0,9071. Namun, fitur department menunjukkan pola yang berbeda, yaitu meningkat dari peringkat keempat pada base learner menjadi peringkat pertama pada meta-learner. Temuan ini mengindikasikan bahwa meta-learner memberikan bobot yang lebih besar pada faktor departemen ketika mengalibrasi keputusan akhir. Pola tersebut hanya dapat diidentifikasi melalui analisis SHAP yang diterapkan secara langsung pada meta-learner, bukan melalui inferensi berdasarkan hasil interpretasi base learner.

3.6 Analisis Prediksi Test Set

Untuk menguji kelayakan dan konsistensi model final pada data yang sepenuhnya independen, dilakukan analisis terhadap distribusi probabilitas prediksi serta pola proporsi promosi antar-departemen pada test set, sebagaimana disajikan pada Gambar 5.



Gambar 5. Grafik Distribusi Prediksi Promosi Pegawai

Gambar 5 menunjukkan bahwa distribusi probabilitas prediksi sangat terkonsentrasi pada nilai rendah, dengan sebagian besar pegawai memiliki probabilitas promosi (P(Dipromosi)) di bawah 0,2. Pola ini mencerminkan rendahnya base rate kelas minoritas pada distribusi data asli. Menariknya, pada bagian ekor kanan distribusi (mendekati 1,0) terlihat akumulasi frekuensi yang jelas, bukan sekadar sebaran acak di sekitar ambang keputusan 0,5. Temuan ini mengindikasikan bahwa model tidak hanya menghasilkan keputusan yang bersifat borderline, tetapi juga mampu mengidentifikasi sebagian pegawai dengan sinyal kelayakan promosi yang kuat. Dengan demikian, distribusi probabilitas tersebut secara kualitatif mendukung validitas kalibrasi model, meskipun kinerja pada kelas minoritas yang diukur melalui F1-Score masih relatif terbatas. Panel kanan menunjukkan adanya variasi yang cukup besar pada proporsi prediksi promosi antar-departemen, mulai dari 2,4% pada departemen HR hingga 20,9% pada departemen R&D, atau hampir sembilan kali lipat dari rata-rata keseluruhan sebesar 10,1%. Temuan ini konsisten dengan hasil analisis SHAP pada Bagian 3.5.1 yang mengidentifikasi departmen sebagai fitur paling berpengaruh pada tingkat meta-learner. Variasi tersebut dapat mencerminkan perbedaan nyata dalam distribusi kinerja, intensitas pelatihan, maupun struktur karier antar-departemen. Namun, kondisi ini juga mengindikasikan kemungkinan bahwa model turut mempelajari pola historis yang berpotensi mengandung bias struktural. Oleh karena itu, temuan ini semakin menegaskan pentingnya pelaksanaan audit fairness secara eksplisit terhadap atribut department dan region sebelum model dipertimbangkan untuk implementasi operasional.

3.7 Pembahasan

Secara empiris, kerangka Stacking Ensemble yang diusulkan menghasilkan F1-Score sebesar 0,4598 dan Recall sebesar 0,4889 pada kondisi ketidakseimbangan kelas yang ekstrem. Meskipun secara numerik terlihat moderat, capaian ini memberikan perspektif evaluasi yang lebih konservatif dibandingkan tren literatur terdahulu. Beberapa penelitian sebelumnya melaporkan kinerja yang jauh lebih tinggi, seperti Shafie et al. dengan F1-Score 0,8750, Nuraeni et al. sebesar 0,8920, Xiong dan Lu dengan AUC-ROC 0,9740, serta Abbour et al. yang mencapai F1-Score 0,9320 [13], [6], [4], [14], [8]. Namun, sebagian besar penelitian tersebut tidak menjelaskan secara eksplisit prosedur pemisahan cross-validation dan oversampling. Kondisi ini berpotensi menimbulkan data leakage (kebocoran data) yang dapat menginflasi F1-Score hingga 110,31%. Oleh karena itu, kinerja yang diperoleh dalam penelitian ini lebih merepresentasikan kemampuan generalisasi yang sesungguhnya karena dievaluasi menggunakan skema imblearn.Pipeline yang bebas dari kontaminasi informasi antar-fold. Meskipun Stacking Ensemble tidak melampaui F1-Score XGBoost secara individual (0,4829), model ini menghasilkan Recall yang

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

lebih tinggi dan lebih relevan pada konteks tata kelola SDM yang sensitif terhadap kesalahan False Negative. Selain itu, dibandingkan penelitian Kurniawan et al. yang hanya memperoleh F1-Score 0,4050 dan Recall 0,2655, kerangka yang diusulkan menunjukkan peningkatan sensitivitas deteksi kelas minoritas secara proporsional tanpa mengorbankan validitas metodologis [4]. Kontribusi kebaruan penelitian ini juga terletak pada penerapan SHAP dua tingkat yang mampu membedah mekanisme internal arsitektur ensemble yang bersifat black-box. Pada tingkat base learner (LightGBM), avg_training_score dan age muncul sebagai prediktor utama kelayakan promosi. Namun, pada tingkat meta-learner (Logistic Regression), variabel department menjadi faktor paling berpengaruh dalam menentukan probabilitas promosi akhir. Pergeseran signifikansi ini mengindikasikan bahwa peluang promosi tidak hanya dipengaruhi oleh merit individu, tetapi juga oleh dinamika struktural antar-departemen. Meskipun pola tersebut dapat merefleksikan perbedaan kebutuhan formasi dan jalur karier, dominasi variabel department juga berpotensi menunjukkan adanya bias historis yang memerlukan audit fairness lebih lanjut. Dengan demikian, sistem yang dibangun berpotensi dimanfaatkan sebagai instrumen audit investigatif SDM untuk mendukung implementasi kebijakan meritokrasi yang lebih transparan, akuntabel, dan minim bias laten.

4. KESIMPULAN

Penelitian ini mengusulkan dan mengevaluasi kerangka prediksi promosi pegawai yang mengintegrasikan SMOTE-ENN, Stacking Ensemble multi-level, dan Explainable Artificial Intelligence (XAI) dua tingkat pada base learner dan meta-learner. Kontribusi utama penelitian ini terletak pada aspek metodologis. Melalui eksperimen komparatif, penelitian ini menunjukkan bahwa penerapan resampling sebelum pemisahan data latih dan validasi dapat menginflasi F1-Score hingga lebih dari 110%, suatu risiko yang masih jarang diverifikasi secara eksplisit dalam literatur klasifikasi berbasis resampling. Setelah bias tersebut dikoreksi melalui skema imblearn.Pipeline yang memastikan resampling hanya dilakukan pada training fold, model Stacking Ensemble memperoleh F1-Score sebesar 0,4598 dan AUC-ROC sebesar 0,8053. Meskipun tidak melampaui kinerja base learner XGBoost secara individual dari sisi F1-Score, model yang diusulkan menghasilkan nilai Recall tertinggi di antara model kompetitif. Temuan ini menunjukkan adanya trade-off yang relevan secara strategis, terutama pada konteks organisasi yang sensitif terhadap kesalahan False Negative dalam proses identifikasi pegawai yang layak dipromosikan. Lessons learned utama dari penelitian ini menunjukkan bahwa pemilihan teknik resampling yang tepat, seperti SMOTE-ENN yang memiliki kemampuan pembersihan noise secara aktif melalui ENN, tidak secara otomatis menjamin validitas hasil evaluasi. Validitas metodologis juga sangat bergantung pada penerapan prosedur cross-validation yang benar. Oleh karena itu, pemilihan teknik resampling dan perancangan eksperimen perlu diperlakukan sebagai dua aspek yang berbeda, tetapi sama-sama memerlukan perhatian yang cermat. Dari perspektif interpretabilitas, analisis SHAP pada tingkat meta-learner menunjukkan korelasi Spearman sebesar 0,9071 terhadap hasil pada base learner. Temuan ini menegaskan bahwa determinan utama promosi, yaitu avg_training_score, usia, dan performance_index, tetap konsisten pada kedua tingkat analisis. Konsistensi tersebut memberikan landasan empiris yang lebih kuat dan kredibel bagi organisasi dalam melakukan audit serta membangun sistem pendukung keputusan sumber daya manusia yang lebih transparan dan berbasis data. Penelitian ini memiliki beberapa keterbatasan yang perlu diungkapkan secara eksplisit. Pertama, performa model setelah dikoreksi dari data leakage (F1-Score = 0,4598) menunjukkan bahwa permasalahan klasifikasi pada kelas minoritas ekstrem (rasio 10,74:1) masih belum sepenuhnya teratasi. Kedua, penggunaan dataset sintetis dari Kaggle, bukan data operasional riil suatu instansi, membatasi generalisasi temuan pada konteks birokrasi atau organisasi lain yang memiliki karakteristik data berbeda. Penelitian ini memiliki sejumlah keterbatasan yang membuka arah pengembangan lebih lanjut. Pertama, menguji sensitivitas hasil terhadap teknik penanganan ketidakseimbangan kelas lainnya, seperti cost-sensitive learning atau threshold tuning berbasis analisis biaya False Negative-False Positive, baik sebagai alternatif maupun pelengkap SMOTE-ENN pada kondisi imbalance yang ekstrem. Kedua, melakukan audit fairness secara eksplisit terhadap atribut demografis, seperti gender dan region, menggunakan metrik demographic parity dan equalized odds, mengingat analisis SHAP mengidentifikasi department dan region sebagai fitur berpengaruh yang berpotensi menjadi proksi bias struktural. Ketiga, melakukan validasi eksternal menggunakan dataset operasional riil dari instansi pemerintah maupun korporasi untuk menguji kemampuan generalisasi lintas domain. Kerangka analitik yang mengedepankan transparansi ini sangat relevan untuk diintegrasikan secara langsung ke dalam platform manajemen informasi kepegawaian di instansi pemerintahan daerah. Implementasi algoritma ini pada sistem pengelolaan ASN dapat mendukung pelaksanaan audit objektivitas secara berkala, memastikan bahwa proses promosi bebas dari bias struktural, serta menjadi katalis terwujudnya sistem meritokrasi yang kredibel. Keempat mereplikasi prosedur anti-leakage yang didemonstrasikan dalam penelitian ini pada studi-studi terdahulu yang berpotensi terpengaruh bias serupa, sebagai kontribusi audit metodologis bagi pengembangan penelitian machine learning pada bidang analitik sumber daya manusia secara lebih luas.

REFERENCES

[1] A. Benabou and F. Touhami, “Optimizing employee promotion predictions using machine learning,” International Journal of Production Management and Engineering, vol. 14, no. 1, pp. 16–33, 2026, doi: 10.4995/ijpme.2025.23364.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

[2] D. S. Angreni and M. Susanti, "Implementasi Data Mining Untuk Rekomendasi Kenaikan Pangkat Pegawai Negeri Sipil Menggunakan Algoritma Naïve Bayes Pada Biro Administrasi Pimpinan Sekretariat Daerah Provinsi Sulawesi Tengah," Innovative: Journal Of Social Science Research, vol. 4, no. 1, pp. 9661–9674, Feb. 2024, doi: 10.31004/innovative.v4i1.9016.
[3] A. Wang, "Enhancing HR management through HRIS and data analytics," Applied and Computational Engineering, vol. 64, no. 1, pp. 222–228, 2024, doi: 10.54254/2755-2721/64/20241394.
[4] S. Kurniawan, A. Nugroho, and Suherman, "Analisis Faktor yang Mempengaruhi Promosi Karyawan Menggunakan Random Forest pada Dataset Employee Promotion," Jurnal Pustaka AI, vol. 5, no. 2, pp. 177–187, Aug. 2025, doi: 10.55382/jurnalpustakaai.v5i2.1115.
[5] M. Isra, "Behavior Analysis and Prediction of Civil Services Staff in Occupational Functional Positions Using C4.5 Algorithm," Jurnal Informasi dan Teknologi (jidt), vol. 4, no. 1, pp. 58–63, Feb. 2022, doi: 10.37034/jidt.v4i1.186.
[6] F. Nuraeni, D. Kurniadi, and M. H. Diazki, "Algoritma K-Nearest Neighbor pada Kasus Dataset Imbalanced untuk Klasifikasi Kinerja Karyawan Perusahaan," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 11, no. 3, pp. 557–568, Jul. 2024, doi: 10.25126/jtiik.938144.
[7] R. Ningsih, M. Devan Basunanda, and W. Erawati, "Implementasi Algoritma C4.5 Untuk Klasifikasi Tingkat Kenaikan Jabatan Pada PT Siprama Cakrawala Depok," Jurnal Infortech, vol. 5, no. 1, 2025, doi: 10.31294/infortech.v7i1.25469.
[8] F. Z. Abbour, E. M. Lghaouch, K. Mahjoubi, S. Ardchir, S. Ounacer, and M. Azzouazi, "Enhancing recruitment strategies with a hybrid AutoML approach to employee promotion prediction," International Journal of Information Management Data Insights, vol. 6, no. 1, pp. 1–10, 2026, doi: 10.1016/j.jjimei.2026.100399.
[9] A. I. Al-Alawi and M. S. Albuainain, "Machine Learning in Human Resource Analytics: Promotion Classification using Data Balancing Techniques," in 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), IEEE, Jan. 2024, pp. 1–5. doi: 10.1109/ICETSIS61505.2024.10459566.
[10] Y. Li, "Research on Faculty Manpower Management of vocational undergraduate based on Decision tree algorithm," in Proceedings of the 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023), Atlantis Press, 2023, pp. 31–38. doi: 10.2991/978-94-6463-264-4_5.
[11] M. B. Zhanuzakov, "Prediction of Employee Promotion Based on Ratings Using Machine-Learning Algorithms," Bulletin of Abai KazNPU. Series of Physical and Mathematical sciences, vol. 77, no. 1, pp. 106–111, Mar. 2022, doi: 10.51889/2022-1.1728-7901.14.
[12] E. D. Anggara, A. Widjaja, and B. R. Suteja, "Prediksi Kinerja Pegawai sebagai Rekomendasi Kenaikan Golongan dengan Metode Decision Tree dan Regresi Logistik," Jurnal Teknik Informatika dan Sistem Informasi, vol. 8, no. 1, Apr. 2022, doi: 10.28932/jutisi.v8i1.4479.
[13] S. Shafie, S. P. Ooi, and K. W. Khaw, "Prediction of Employee Promotion Using Hybrid Sampling Method with Machine Learning Architecture," Malaysian Journal of Computing, vol. 8, no. 1, pp. 1264–1286, 2023, doi: 10.24191/mjoc.v8i1.18456.
[14] Z. Xiong and W. Lu, "Employee Promotion Prediction: Machine Learning in HR Management," Advances in Economics, Management and Political Sciences, vol. 241, no. 1, pp. 157–169, 2025, doi: 10.54254/2754-1169/2025.BJ30202.
[15] N. E. Vijayan, "Automated Feature Engineering for Predictive HR Analytics Using Cloud-Based ETL and ML Pipelines," Journal of Artificial Intelligence, Machine Learning and Data Science, vol. 1, no. 4, pp. 1–9, Nov. 2023, doi: 10.51219/JAIMLD/naveen-edapurath-vijayan/344.
[16] M. Liu and X. Liu, "Design and Implementation of an Intelligent Employee Promotion Strategy Optimization System," in Proceedings of the 2024 4th International Conference on Big Data, Artificial Intelligence and Risk Management, New York, NY, USA: ACM, Apr. 2024, pp. 515–520. doi: 10.1145/3718751.3718834.
[17] J. and V.-R. Y. Tusell-Rey Claudia C. and Pino-Gómez, "Evaluative Customized Naïve Associative Classifier: Promoting Equity in AI for the Selection and Promotion of Human Resources," in Intelligent Data Engineering and Automated Learning – IDEAL 2024, D. and Y. H. and A. J. M. and N. V. B. and N. P. and T.-B. A. Julian Vicente and Camacho, Ed., Cham: Springer Nature Switzerland, 2025, pp. 275–286. doi: 10.1007/978-3-031-77738-7_23.
[18] P. Ardehkhani, S. R. Moslemi, and H. Hooshmand, "Enhancing Employee Promotion Prediction with a Novel Hybrid Model Integrating Convolutional Neural Networks and Random Forest," in 2023 14th International Conference on Information and Knowledge Technology (IKT), IEEE, Dec. 2023, pp. 62–68. doi: 10.1109/IKT62039.2023.10433023.
[19] P. Agrawal, S. Goyal, and A. Jandwani, "A novel agile based framework for employee promotion," International Journal of Information Technology, vol. 16, no. 7, pp. 4481–4487, Jul. 2024, doi: 10.1007/s41870-024-02071-x.
[20] Möbius, "HR Analytics: Employee Promotion Data." Accessed: Mar. 31, 2026. [Online]. Available: https://www.kaggle.com/datasets/arashnic/hr-ana
[21] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," Mach. Learn., vol. 113, no. 7, pp. 4903–4923, 2024, doi: 10.1007/s10994-022-06296-4.
[22] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," SAGE Publications, vol. 18, no. 6, 2022, doi: 10.1177/15501329221106935.
[23] M. Todorovic, N. Stanisic, M. Zivkovic, N. Bacanin, V. Simic, and E. B. Tirkolaee, "Improving audit opinion prediction accuracy using metaheuristics-tuned XGBoost algorithm with interpretable results through SHAP value analysis," Appl. Soft Comput., vol. 149, p. 110955, 2023, doi: https://doi.org/10.1016/j.asoc.2023.110955.
[24] R. Andespa, K. Sadik, C. Suhaeni, and A. M. Soleh, "Evaluating Random Forest and XGBoost for Bank Customer Churn Prediction on Imbalanced Data Using SMOTE and SMOTE-ENN," Media Statistika, vol. 18, no. 1, pp. 25–36, 2025, doi: 10.14710/medstat.18.1.25-36.
[25] S. Alam, M. S. Ayub, S. Arora, and M. A. Khan, "An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity," Decision Analytics Journal, vol. 9, p. 100341, 2023, doi: 10.1016/j.dajour.2023.100341.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1076−1087
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.9879

[26] N. M. N. Mathivanan, E. F. Z. Xian, D. F. Yong Xi, and C. H. Kiat, "Impact of Feature Standardization on Heart Disease Prediction: A Comparative Analysis of Logistic Regression and Support Vector Machine Models," Malaysian Journal of Computing, vol. 10, no. 2, pp. 2159–2175, 2025, doi: 10.24191/mjoc.v10i1.6835.
[27] Y. D. Nugroho H, F. Zakiyabarsi, and A. J. Paramita, "Implementasi SMOTE-ENN dan Borderline SMOTE terhadap Performa LightGBM pada Imbalanced Class," Rabit : Jurnal Teknologi dan Sistem Informasi Univrab, vol. 10, no. 1, pp. 51–59, 2025, doi: 10.36341/rabit.v10i1.5436.
[28] G. Husain et al., "SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models," Algorithms, vol. 18, no. 1, pp. 1–16, 2025, doi: 10.3390/a18010037.
[29] A. F. Salsabilah, B. Rahmat, and E. Y. Puspaningrum, "Stacking Ensemble of XGBoost, LightGBM, and CatBoost for Green Economy Index Prediction," bit-Tech, vol. 8, no. 1, pp. 327–339, 2025, doi: 10.32877/bt.v8i1.2530.
[30] J. Lei, J. Zhai, Y. Zhang, J. Qi, and C. Sun, "Supervised Machine Learning Models for Predicting Sepsis-Associated Liver Injury in Patients With Sepsis: Development and Validation Study Based on a Multicenter Cohort Study," J Med Internet Res, vol. 27, p. e66733, 2025, doi: 10.2196/66733.
[31] S. Q. Sultan, N. Javaid, N. Alrajeh, and M. Aslam, "Machine Learning-Based Stacking Ensemble Model for Prediction of Heart Disease with Explainable AI and K-Fold Cross-Validation: A Symmetric Approach," Symmetry (Basel)., vol. 17, no. 2, 2025, doi: 10.3390/sym17020185.
[32] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," BioData Min., vol. 16, no. 1, p. 4, 2023, doi: 10.1186/s13040-023-00322-4.
[33] M. Imani, M. Joudaki, A. Bagheri, and H. R. Arabnia, "Why ROC-AUC Is Misleading for Highly Imbalanced Data: In-Depth Evaluation of MCC, F2-Score, H-Measure, and AUC-Based Metrics Across Diverse Classifiers," Technologies (Basel)., vol. 14, no. 1, 2026, doi: 10.3390/technologies14010054.
[34] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," Sensors, vol. 23, no. 4, 2023, doi: 10.3390/s23042333.