ADA Journal of Information System Research
ISSN 3025-9568 (media online)
Volume 3, No 3, June 2026 Page: 110−119
DOI doi.org/10.64366/adajisr.v3i3.625
https://journals.adaresearch.or.id/adajisr

Penerapan Random Forest untuk Klasifikasi Standar Gaji Industri Berdasarkan Level Pendidikan dan Pengalaman Kerja

Ramadhan Putra Prayoga*, Eko Setia Budi

Fakultas Teknik dan Informatika, Program Studi Informatika, Universitas Bina Sarana Informatika, Jawa Barat, Indonesia
Email: 1,*ramaputra1311@gmail.com, 2eko.etb@bsi.ac.id
Email Penulis Korespondensi: ramaputra1311@gmail.com

Abstrak–Penentuan standar kompensasi gaji di lingkungan industri seringkali terhambat oleh tingginya kompleksitas dalam memetakan berbagai variabel pekerja secara bersamaan, seperti tingkat pendidikan, pengalaman kerja, usia, dan jabatan. Proses penetapan konvensional yang tidak didasarkan pada korelasi riwayat data ini rentan menghasilkan keputusan yang tidak konsisten, subjektif, dan memicu ketidakadilan internal. Penelitian ini bertujuan merancang sebuah model prediksi objektif berlandaskan data guna mengklasifikasikan standar gaji pekerja secara otomatis ke dalam tiga kategori kelas target (rendah, menengah, tinggi). Solusi komputasi yang diusulkan adalah penerapan metode Machine Learning menggunakan algoritma Random Forest. Tahapan analisis kuantitatif dieksekusi terhadap 1.000 data profil pekerja industri. Untuk mencegah fenomena data leakage yang memicu akurasi semu, atribut nominal gaji aktual dihapus secara ketat pada fase prapemrosesan. Pengujian performa sistem kemudian divalidasi menggunakan instrumen 10-Fold Cross Validation. Hasil penelitian membuktikan bahwa algoritma berhasil mengekstrak nilai variabel pendidikan formal dan pengalaman kerja sebagai parameter pemisah utama yang paling krusial. Berdasarkan eksperimen variasi parameter, model sukses mencapai titik konvergensi pembelajaran paling optimal pada konfigurasi 100 trees. Pada titik tersebut, sistem mencatatkan performa puncak dengan tingkat akurasi 76,80%, nilai presisi 0,767, dan F1-Score 0,767. Modifikasi penambahan struktur komputasi menjadi 150 tree tidak lagi memberikan eskalasi kinerja. Kesimpulannya, arsitektur analitik ini terbukti valid dan reliabel untuk diimplementasikan sebagai instrumen pendukung keputusan guna mewujudkan sistem penggajian yang objektif dan transparan.

Kata Kunci: Random Forest; Klasifikasi; Standar Gaji; Kebocoran Data; Kebocoran Data; Machine Learning

Abstract–Determining salary compensation standards in industrial environments is often hindered by the high complexity of mapping multiple worker variables simultaneously, such as education level, work experience, age, and job title. Conventional determination processes that are not based on historical data correlations are prone to producing inconsistent and subjective decisions, triggering internal inequity. This study aims to design an objective, data-driven predictive model to automatically classify workers' salary standards into three target class categories (Low, Mid, High). The proposed computational solution is the application of Machine Learning methods using the Random Forest algorithm. The quantitative analysis phase was executed on 1,000 industrial worker profile records. To prevent data leakage phenomena that trigger pseudo-accuracy, the actual nominal salary attribute was strictly removed during the preprocessing phase. The system's performance evaluation was then validated using a 10-Fold Cross-Validation approach. The research results prove that the algorithm successfully extracted formal education and work experience variables as the most crucial primary splitting parameters. Based on parameter variation experiments, the model successfully reached its optimal learning convergence point at a configuration of 100 decision trees. At this point, the system recorded peak performance with an accuracy rate of 76.80%, a precision value of 0.767, and an F1-Score of 0.767. Modifying the computational structure by increasing it to 150 trees provided no further performance escalation. In conclusion, this analytical architecture is proven to be valid and reliable for implementation as a decision support instrument to establish an objective and transparent payroll system.

Keywords: Salary Classification; Random Forest; Machine Learning; Data Leakage; Decision Support System

1. PENDAHULUAN

Dinamika pertumbuhan industri saat ini memunculkan tantangan kompleks dalam manajemen sumber daya manusia, salah satunya adalah penentuan standar gaji yang adil dan kompetitif, membayar gaji yang sesuai kepada karyawan sangat diperlukan untuk mencapai tujuan perusahaan [1]. Masalah utama yang sering terjadi di lapangan adalah munculnya subjektivitas dan inkonsistensi dalam penetapan kompensasi pekerja. Hal ini disebabkan oleh banyaknya variabel prediktor yang saling bersilangan, seperti level pendidikan, rekam jejak pengalaman kerja, usia, jabatan, lokasi, dan jenis kelamin. Gaji adalah balas jasa berupa uang yang diterima karyawan atas kontribusinya dalam mencapai tujuan perusahaan [2]. Perbedaan nominal gaji ini menimbulkan pertanyaan fundamental mengenai parameter apa saja yang secara objektif digunakan untuk menentukan besaran kompensasi seorang karyawan. Dalam konteks organisasi, pengalaman kerja, latar belakang pendidikan, komitmen organisasi, dan hasil penilaian kerja merupakan faktor-faktor krusial yang menentukan kualitas kinerja karyawan, yang pada akhirnya berujung pada penentuan besaran gaji yang diterima [3]. Oleh karena itu, diperlukan sebuah solusi otomasi yang mampu mempelajari pola riwayat data pekerja untuk mengklasifikasikan standar gaji secara objektif. Penerapan Kecerdasan Buatan, khususnya algoritma Machine Learning, Machine Learning adalah bidang yang fokus pada pengembangan sistem untuk memproses data dan melakukan tugas seperti klasifikasi secara otomatis [4].

Meskipun Machine Learning menawarkan kapabilitas analitik yang luar biasa, penerapannya dalam domain sumber daya manusia dan kepegawaian sering kali dihadapkan pada ancaman teoretis yang sangat fatal, yaitu data leakage. Secara konseptual, data leakage terjadi ketika informasi target yang seharusnya diprediksi (seperti nominal gaji aktual atau proksi potongannya) secara tidak sengaja terikut dan dipelajari oleh algoritma selama fase pelatihan. Dalam pengolahan data HR, atribut-atribut pegawai saling memiliki korelasi yang sangat erat. Membiarkan variabel

Copyright © 2026 The Author(s). Published by *ADA Research Center*. Page 110
This work is licensed under a Creative Commons Attribution 4.0 International License.

ADA Journal of Information System Research
ISSN 3025-9568 (media online)
Volume 3, No 3, June 2026 Page: 110−119
DOI doi.org/10.64366/adajisr.v3i3.625
https://journals.adaresearch.or.id/adajisr

yang memiliki korelasi absolut dengan target tetap berada di dalam ruang pemodelan akan menghasilkan ilusi performa atau akurasi semu. Oleh karena itu, penghapusan atribut yang berpotensi membocorkan target secara ketat akan memaksa algoritma klasifikasi untuk mensimulasikan proses kognitif dan analisis HRD di dunia nyata, yakni menilai kelayakan standar kompensasi murni berdasarkan kualifikasi dan rekam jejak profesional pekerja, sehingga menjamin validitas, objektivitas, serta keadilan dari sistem prediksi yang dihasilkan.

Secara konseptual, penetapan standar gaji sangat erat kaitannya dengan fondasi Human Capital Theory. Human Capital Theory yang dikembangkan oleh Gary Becker dan Theodore Schultz, pendidikan dipahami bukan hanya sebagai proses transfer pengetahuan, tetapi juga sebagai bentuk investasi yang dapat meningkatkan keterampilan, produktivitas, dan kesejahteraan individu [5]. Pendidikan formal membentuk kerangka kognitif, kemampuan pemecahan masalah secara terstruktur, dan dasar pengetahuan teoritis yang mendasari keahlian teknis pekerja. Di sisi lain, pengalaman kerja membentuk intuisi praktis, adaptabilitas, dan kedalaman wawasan di lapangan yang seringkali tidak dapat diajarkan di dalam ruang kelas. Kombinasi dari kedua indikator fundamental inilah yang secara rasional seharusnya memiliki korelasi terkuat dalam penentuan hak finansial seorang pekerja. Dalam ranah analitik komputasional, pemetaan pola yang linier maupun non-linier dari indikator-indikator tersebut dapat diekstraksi dan diotomatisasi melalui metode klasifikasi berkinerja tinggi.

Dalam konteks pemodelan data terstruktur, salah satu algoritma klasifikasi berbasis ensemble learning yang terbukti sangat tangguh dan handal adalah Random Forest. Random Forest merupakan suatu set dari decision trees yang dibangun dengan sampel yang dipilih secara acak tetapi memiliki peraturan membelah simpul yang berbeda [6]. Algoritma ini beroperasi secara paralel dengan membangun sekumpulan decision trees dalam jumlah yang masif pada fase pelatihan (training). Decision tree adalah model prediksi yang menggunakan struktur tree atau hirarki untuk mengambil keputusan [7]. Konsep flowchart dengan struktur tree dimana setiap node internal mewakili atribut dan cabang menggambarkan hasil pengujian atau nilai input atribut, sedangkan daun mewakili kelas atau distribusi kelas [8]. Keunggulan utama dari Random Forest dibandingkan dengan algoritma tunggal lainnya yaitu dapat meningkatkan tingkat akurasi jika ada data yang tidak lengkap, dan untuk resisting outliers, serta efisien untuk penyimpanan sebuah data [9]. Random Forest mengimplementasikan mekanisme pemilihan fitur yang efektif dalam menentukan fitur unggulan untuk meningkatkan kinerja model klasifikasi. Melalui proses ini, Random Forest dapat beroperasi secara efisien pada big data dengan berbagai parameter yang rumit [10].

Penerapan Machine Learning dalam ranah analitik sumber daya manusia dan pemrosesan klasifikasi data terstruktur bukanlah sebuah konsep teoritis yang sepenuhnya baru, melainkan telah menjadi sorotan serta pilar utama dari berbagai penelitian state of the art dalam kurun waktu lima tahun terakhir. Sebagai tinjauan literatur terkait, penelitian yang dilakukan oleh Amanda dan Voutama pada tahun 2025 mengkaji implementasi algoritma Random Forest untuk klasifikasi pendapatan individu menggunakan dataset Adult Income. Penelitian tersebut membuktikan bahwa algoritma ini mampu menghasilkan klasifikasi dengan akurasi yang cukup tinggi, yakni mencapai 85,44%. Meskipun demikian, penelitian ini secara spesifik menyoroti kendala model saat berhadapan dengan masalah ketidakseimbangan data, di mana algoritma cenderung bias dalam mempelajari pola dan mendominasi hasil prediksi pada kelas mayoritas (pendapatan <=50K). Oleh sebab itu, penelitian tersebut menyimpulkan perlunya penerapan teknik penanganan data tidak seimbang serta penyetelan hiperparameter lebih lanjut guna meningkatkan sensitivitas dan stabilitas akurasi model terhadap kelas minoritas [11].

Sejalan dengan eksplorasi performa algoritma pembelajaran ansambel, sebuah studi komprehensif yang dipublikasikan oleh Umar dan Sitohang pada tahun 2024 mengkaji kapabilitas algoritma Random Forest, bersama dengan metode K-Nearest Neighbors dan Decision Tree, dalam memprediksi keputusan pembelian paket wisata. Penelitian tersebut mengeksplorasi penggunaan pola data demografis, perilaku kunjungan pelanggan, serta preferensi properti untuk merancang model analitik di sektor pariwisata. Dalam implementasinya, pengujian Random Forest Regressor menghasilkan nilai Mean Absolute Error (MAE) sebesar 0.05 yang mengindikasikan prediksi cukup mendekati nilai aktual, namun masih mencatatkan Root Mean Squared Error (RMSE) sebesar 0.16 dan R-squared score 0.3. Hal tersebut menuntun pada kesimpulan bahwa meskipun Random Forest merupakan model yang solid, penerapannya mutlak membutuhkan penyesuaian lanjutan secara spesifik agar model mampu menjelaskan variasi secara signifikan dan meningkatkan keakuratan prediksinya pada dataset analitik yang kompleks [12].

Melengkapi tinjauan tersebut, penelitian ketiga yang dilakukan oleh Hidayat, dkk. pada tahun 2023 menginvestigasi penerapan algoritma Random Forest pada ranah medis, secara spesifik untuk memprediksi risiko penyakit jantung menggunakan dataset publik. Dalam studinya, para peneliti tidak hanya menerapkan algoritma standar, tetapi juga mengintegrasikan teknik pra-pemrosesan data seperti normalisasi kelas dan manipulasi parameter model untuk mengoptimalkan kinerja klasifikasi. Hasil pengujian menunjukkan bahwa model mampu mencapai akurasi yang sangat baik, yakni mencapai 94% pada skenario pembagian data pelatihan dan pengujian 80/20. Meskipun kapabilitas algoritma terbukti solid, penelitian ini sepenuhnya berfokus pada target klasifikasi biner (pasien terindikasi atau tidak terindikasi penyakit) di bidang kesehatan, sehingga tidak menyentuh dinamika variabel maupun aspek kompensasi finansial di dunia industri[13].

Selanjutnya, studi keempat yang dipublikasikan oleh Ernawati, dkk. pada tahun 2025 memperluas cakupan penerapan machine learning ke sektor sosial ekonomi dengan membandingkan performa algoritma Random Forest dan K-Nearest Neighbors (K-NN). Penelitian ini bertujuan untuk mengklasifikasikan status kemiskinan penduduk di Kabupaten Labuhanbatu berdasarkan pengolahan data tabular komprehensif yang mencakup variabel pekerjaan, pendapatan, tempat tinggal, dan tingkat pendidikan. Kesimpulan dari eksperimen tersebut menunjukkan bahwa

Copyright © 2026 The Author(s). Published by *ADA Research Center*. Page 111
This work is licensed under a Creative Commons Attribution 4.0 International License.

ADA Journal of Information System Research
ISSN 3025-9568 (media online)
Volume 3, No 3, June 2026 Page: 110−119
DOI doi.org/10.64366/adajisr.v3i3.625
https://journals.adaresearch.or.id/adajisr

algoritma Random Forest tampil sedikit lebih unggul dibandingkan K-NN, dengan pencapaian metrik evaluasi berupa akurasi sebesar 60,23%, presisi 0,48, dan recall 0,41. Meskipun mengolah variabel prediktor yang bersinggungan erat dengan tingkat ekonomi dan kualifikasi pendidikan, target luaran dari penelitian ini masih terbatas pada pelabelan biner yang mendasar (status miskin atau tidak miskin). Penelitian ini juga tidak dirancang untuk memecahkan problematika pemetaan multi-kelas pada standar kelayakan gaji profesional[14].

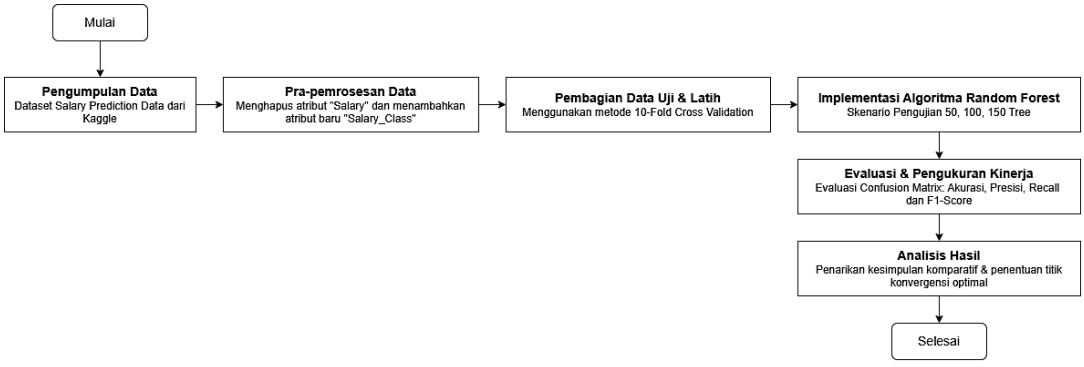
Terakhir, literatur kelima yang disusun oleh Fitri, dkk. pada tahun 2025 mengeksplorasi ketangguhan algoritma Random Forest dalam Natural Language Processing untuk menganalisis masalah kesehatan mental. Penelitian ini secara spesifik merancang sistem klasifikasi tingkat kecemasan pada Generasi Z berdasarkan ekstraksi data teks tidak terstruktur (cuitan) dari platform media sosial X. Dengan menerapkan metode pembobotan TF-IDF untuk ekstraksi fitur dan penanganan ketidakseimbangan kelas menggunakan SMOTE, model Random Forest yang dihasilkan mencatatkan tingkat akurasi yang sangat impresif, yakni mencapai 97,71%. Walaupun secara komputasional sangat sukses dan akurat, orientasi penelitian ini murni terfokus pada analisis teks di bidang psikologis, sehingga belum mengeksplorasi potensi dan performa algoritma yang sama ketika dihadapkan pada pemrosesan data tabular numerik yang kompleks pada sektor manajemen sumber daya manusia[15].

Berdasarkan deskripsi latar belakang masalah yang kompleks dan analisis kesenjangan literatur yang telah diuraikan, maka tujuan utama dari penelitian ini adalah merancang dan mengimplementasikan algoritma klasifikasi Random Forest untuk memprediksi standar gaji industri berdasarkan penekanan pada fitur level pendidikan serta rekam jejak pengalaman kerja. Penelitian ini secara spesifik bertujuan untuk mengukur metrik tingkat akurasi pemodelan melalui serangkaian eksperimen analisis kuantitatif menggunakan 10-Fold Cross Validation yang dieksekusi secara komputasional menggunakan perangkat lunak Waikato Environment for Knowledge Analysis (WEKA). Weka adalah sebuah platform yang dipakai untuk melakukan data mining. Aplikasi ini memiliki lisensi GNU General Public License, sehingga dapat digunakan secara gratis. Weka menggunakan framework Java sehingga dapat digunakan di berbagai macam sistem operasi. [16]. Tujuan teknis tambahan dari studi ini adalah membandingkan secara komparatif serta menganalisis efek dari manipulasi parameter jumlah Decision Tree (pada skenario 50, 100, dan 150 trees) di dalam sistem WEKA guna mencari di mana letak titik konvergensi yang paling ideal bagi arsitektur dataset ketenagakerjaan tersebut. Adapun kontribusi utama dari penelitian ini terbagi menjadi dua aspek. Secara teoritis, penelitian ini berkontribusi dalam memberikan bukti empiris mengenai konfigurasi parameter Random Forest yang paling optimal untuk menangani kompleksitas data sumber daya manusia. Secara praktis, penelitian ini berkontribusi menghasilkan sebuah instrumen audit prediktif berbasis data yang secara objektif mampu mengeliminasi bias manusia (HRD) dan subjektivitas dalam penentuan kelas gaji, sehingga membantu perusahaan merumuskan kebijakan kompensasi yang transparan, proporsional, dan adil.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini berlandaskan pada pendekatan metode penelitian kuantitatif murni yang dirancang secara terstruktur guna memastikan alur pengolahan komputasional berjalan sistematis, transparan, dan memiliki tingkat validitas algoritmik yang tinggi. Rangkaian penyelesaian masalah direpresentasikan secara visual pada Gambar 1, yang menguraikan transisi pemrosesan data mulai dari tahap akuisisi mentah hingga penarikan kesimpulan evaluatif.



Gambar 1. Tahapan Penelitian

Tahapan pertama dalam penelitian ini adalah Pengumpulan Data. Dataset yang digunakan bersumber dari repositori publik Kaggle, yang memuat 1.000 instans riwayat pekerja industri. Data ini mengandung beragam atribut prediktor demografis dan rekam jejak profesional, seperti level pendidikan dan pengalaman kerja. Data Mining tidak hanya berguna untuk mengolah informasi dengan cara yang efektif tetapi juga dapat membantu dalam proses pengambilan pilihan yang didasarkan pada data, yang sekarang menjadi fondasi dalam berbagai sektor, seperti perdagangan, kesehatan, keuangan, dan pendidikan [17].

Copyright © 2026 The Author(s). Published by *ADA Research Center*. Page 112
This work is licensed under a Creative Commons Attribution 4.0 International License.

ADA Journal of Information System Research
ISSN 3025-9568 (media online)
Volume 3, No 3, June 2026 Page: 110−119
DOI doi.org/10.64366/adajisr.v3i3.625
https://journals.adaresearch.or.id/adajisr

Setelah data diakuisisi, langkah krusial selanjutnya adalah Pra-pemrosesan Data yang dieksekusi menggunakan perangkat lunak Waikato Environment for Knowledge Analysis (WEKA). Tahap Data Pre-processing ini seperti pelabelan data, integrasi data, transformasi data, pembersihan data, dan mencakup mitigasi data leakage agar sistem dapat berfokus memetakan kelas gaji tanpa intervensi data riil [18]. Pre-processing merupakan fase yang wajib dalam rekayasa Machine Learning untuk mengubah data mentah menjadi format yang dapat diproses secara matematis, sekaligus mengeliminasi atribut yang mengandung noise. Fokus paling utama pada tahap ini adalah mitigasi ancaman data leakage. Proses mitigasi diimplementasikan dengan mengeksklusi atribut nominal Salary secara permanen dari ruang pemodelan, dan merumuskan variabel target kategorikal baru bernama "Salary_Class". Pencegahan data leakage adalah standar protokol dalam klasifikasi guna menghalangi model dari tindakan menghafal nilai target secara langsung melalui atribut prediktor, yang apabila diabaikan akan bermuara pada matriks performa prediktif yang bias dan semu. Data Leakage merupakan kondisi di mana informasi dari data uji secara tidak sengaja terlibat ke dalam proses training, baik secara langsung maupun tidak langsung [19].

Dataset yang telah tervalidasi kemudian memasuki tahap Pembagian Data Uji dan Latih. Berbeda dengan pendekatan pembagian persentase statis, penelitian ini menerapkan arsitektur 10-Fold Cross Validation. Teknik ini beroperasi dengan mempartisi himpunan data ke dalam 10 folds berukuran ekuivalen. Selama 10 iterasi komputasi beruntun, setiap lipatan secara bergantian akan diisolasi sebagai testing set, sementara 9 fold sisanya disatukan sebagai training set. 10-Fold Cross Validation telah digunakan untuk mengevaluasi performa model machine learning dalam penelitian sebelumnya dengan hasil yang bervariasi melalui pemisahan dataset menjadi data training dan data test. Langkah ini dapat membantu mengidentifikasi apakah model berfungsi dengan baik dan menghindari overfitting atau underfitting [20].

Tahapan sentral dari penelitian ini adalah Implementasi Algoritma Random Forest. Random Forest merupakan suatu set dari decision trees yang dibangun dengan sampel yang dipilih secara acak tetapi memiliki peraturan membelah simpul yang berbeda. Model ini bekerja dengan menggunakan subset dari suatu fitur pada setiap tree, kemudian mencoba mencari ambang batas terbaik didalam memisahkan data [6]. Algoritma Random Forest banyak dirujuk dalam berbagai literatur karena ketangguhannya yang luar biasa dalam menangani kelas data yang tidak seimbang, serta kekebalannya terhadap outliers. Dalam penelitian ini, implementasi model tidak dijalankan secara statis, melainkan melalui hyperparameter tuning yang mengomparasikan formasi 50, 100, dan 150 trees. Manipulasi ini merupakan teknik eksperimental yang dirancang untuk mencari titik konvergensi paling seimbang antara pencapaian akurasi puncak dan efisiensi waktu komputasi sistem.

Keluaran klasifikasi dari skenario Random Forest selanjutnya diukur pada tahap Evaluasi dan Pengukuran Kinerja. Alat ukur komprehensif yang digunakan adalah Confusion Matrix. Secara konseptual, Confusion Matrix merupakan sebuah metode yang sering dipakai untuk menilai seberapa tepatnya hasil dari pengolahan data. Tahap ini mengevaluasi berbagai aspek seperti akurasi, presisi, recall, spesifikasi, dan F1-Score untuk memahami seberapa efektif model yang telah dibuat sebelumnya [21].

Tahapan terakhir adalah Analisis Hasil. Pada tahap ini, seluruh rekapan kuantitatif dari metrik Confusion Matrix dianalisis secara komparatif antarskenario jumlah tree. Proses analisis akhir ini ditujukan untuk menarik kesimpulan deduktif guna memvalidasi formasi arsitektur mana yang merepresentasikan titik konvergensi algoritmik yang paling objektif untuk diterapkan di sektor industri, sebelum rangkaian penelitian komputasional ini dinyatakan selesai.

2.2 Pemilihan Algoritma Random Forest

Dalam memecahkan masalah penentuan standar gaji industri, penelitian ini mengadopsi algoritma Random Forest sebagai metode penyelesaian utama. Random Forest merupakan salah satu metode ensemble learning yang bekerja dengan cara membangun banyak decision trees secara independen pada saat fase pelatihan data. Ensemble Learning adalah metode penggabungan algoritma dari Machine Learning. Beragam model Machine Learning dikombinasikan untuk mencapai hasil prediksi yang lebih unggul daripada dengan memakai satu model saja [22]. Mekanisme klasifikasi pada metode ini beroperasi dengan memecah dataset ke dalam simpul-simpul (nodes) berdasarkan atribut yang memiliki nilai Information Gain tertinggi [23].

Pada penelitian ini, variabel penentu utama seperti jenjang pendidikan (Education) dan pengalaman kerja (Experience) digunakan oleh sistem sebagai parameter pemisah pada root node, yang kemudian diturunkan hingga mencapai leaf node penentu kelas gaji. Keputusan akhir dari sistem untuk memprediksi kelas gaji seorang pekerja tidak bergantung pada satu tree saja, melainkan diputuskan melalui mekanisme pemungutan suara terbanyak dari seluruh tree yang terbentuk di dalam algoritma tersebut.

2.3 Preprocessing dan Pemilihan Fitur

Kualitas dari hasil klasifikasi algoritma kecerdasan buatan sangat bergantung pada representasi data yang dilatih. Penelitian ini menggunakan dataset pekerja industri berjumlah 1.000 baris data yang memuat informasi demografis dan profesional. Pada tahap preprocessing, variabel target (Salary_Class) dibagi secara proporsional ke dalam tiga kelas ordinal tingkat gaji, yakni Low berjumlah 330 data, Mid berjumlah 330 data, dan High berjumlah 340 data. Langkah mitigasi paling krusial pada tahapan penyelesaian masalah ini adalah proses seleksi fitur, di mana atribut Salary secara sengaja dihapus dan dieksklusi dari dalam perangkat lunak Weka. Dengan memanfaatkan gabungan antara Random Forest dan teknik seleksi fitur, diharapkan dapat tercipta model prediksi yang lebih tepat dan efisien

Copyright © 2026 The Author(s). Published by *ADA Research Center*. Page 113
This work is licensed under a Creative Commons Attribution 4.0 International License.

ADA Journal of Information System Research
ISSN 3025-9568 (media online)
Volume 3, No 3, June 2026 Page: 110−119
DOI doi.org/10.64366/adajisr.v3i3.625
https://journals.adaresearch.or.id/adajisr

[24]. Penghapusan ini bersifat wajib untuk mencegah terjadinya data leakage, sehingga algoritma dipaksa untuk murni mempelajari pola dari atribut prediktor tanpa melakukan kecurangan komputasi.

Rincian struktur atribut yang digunakan setelah tahap preprocessing direpresentasikan secara menyeluruh pada Tabel 1. Berdasarkan Tabel 1, dataset penelitian dikonstruksi oleh tujuh atribut yang dipetakan ke dalam tiga peran pemodelan spesifik. Atribut Education yang bertipe nominal dan Experience yang bertipe numerik ditetapkan sebagai prediktor utama, mengingat kedua variabel tersebut merupakan indikator sentral kompetensi pekerja. Selanjutnya, Tabel 1 juga menguraikan empat atribut sebagai prediktor pendukung, yang terdiri dari Age (numerik), serta Job_Title, Location, dan Gender yang berjenis nominal. Seluruh variabel independen tersebut dikonfigurasi untuk memprediksi keluaran akhir pada atribut Salary_Class (nominal), yang secara eksklusif diposisikan sebagai target klasifikasi.

Tabel 1. Struktur Atribut Dataset

Atribut	Type Data	Peran Pemodelan
Education	Nominal	Prediktor Utama
Experience	Numerik	Prediktor Utama
Age	Numerik	Prediktor Pendukung
Job_Title	Nominal	Prediktor Pendukung
Location	Nominal	Prediktor Pendukung
Gender	Nominal	Prediktor Pendukung
Salary_Class	Nominal	Target Klasifikasi (Class)

2.4 Skenario Pengujian dan Pengukuran Performa

Untuk membuktikan efektivitas metode penyelesaian masalah yang ditawarkan, skenario pengujian dieksekusi secara komprehensif menggunakan instrumen evaluasi 10-Fold Cross Validation. Metode ini membagi 1.000 data ke dalam 10 fold yang sama besar, di mana 9 bagian digunakan sebagai training data dan 1 bagian digunakan sebagai testing data yang dirotasi secara bergantian [25]. Pendekatan pengujian berlapis ini berfungsi untuk mencegah fenomena overfitting.

Eksperimen metode ini dibagi ke dalam tiga skenario utama berdasarkan manipulasi jumlah iterasi number of trees, yaitu skenario 50 tree, 100 tree, dan 150 tree. Performa dari masing-masing skenario akan diukur dan dibandingkan melalui ekstraksi metrik pada Confusion Matrix. Parameter evaluasi kinerja yang menjadi tolak ukur keberhasilan pemodelan kuantitatif ini difokuskan pada nilai Accuracy, Precision, Recall, serta F1-Score, guna menemukan titik konvergensi pembelajaran algoritma yang paling optimal dalam membedakan kelas gaji.

3. HASIL DAN PEMBAHASAN

3.1 Penerapan Metode dan Pra-pemrosesan Data

Tahapan eksekusi pada penelitian kuantitatif ini difokuskan pada pengolahan data mentah menjadi informasi prediktif menggunakan perangkat lunak Waikato Environment for Knowledge Analysis (WEKA). Sebelum algoritma Random Forest dapat melakukan proses machine learning, dataset ketenagakerjaan yang berjumlah 1.000 instans data harus melewati fase data preprocessing. Fase ini merupakan tahapan yang paling krusial karena kualitas dari model kecerdasan buatan sangat bergantung pada seberapa representatif dan bersih data yang dilatihkan kepadanya. Proses penerapan metode dan pra-pemrosesan data dalam perangkat lunak WEKA dilakukan melalui tahapan-tahapan berurutan sebagai berikut:

a. Inisialisasi dan integrasi dataset
   Pada tahap ini, berkas data berekstensi Comma Separated Values (CSV) yang berisi 1.000 rekam jejak pekerja industri diimpor ke dalam Explorer interface pada WEKA. Sistem secara otomatis mendeteksi ragam tipe data, di mana variabel seperti umur (Age) dan masa kerja (Experience) diidentifikasi sebagai atribut bertipe numerik kontinyu, sementara jenjang pendidikan (Education), lokasi kerja (Location), jabatan (Job_Title), dan jenis kelamin (Gender) diidentifikasi sebagai atribut kategorikal nominal.
b. Transformasi dan penyeimbangan kelas target
   Agar sistem klasifikasi tidak mengalami bias mayoritas (kondisi di mana mesin hanya menebak kelas yang paling banyak muncul), variabel target penelitian harus disebarkan secara merata. Pada tahap ini, variabel kelas gaji diformulasikan ke dalam bentuk klasifikasi ordinal yang terdiri dari tiga kelas seimbang, yakni kategori gaji Low (330), Mid (330), dan High (340).
c. Mitigasi Data Leakage
   Ini adalah prosedur wajib dan paling vital dalam desain penelitian klasifikasi ini. Jika mesin diberikan akses ke data gaji aktual, mesin tidak akan melakukan analisis prediktif, melainkan sekadar mencocokkan angka, yang akan menghasilkan evaluasi palsu sebesar 100%. Oleh karena itu, langkah mitigasi dilakukan dengan cara:
   1. Melakukan seleksi mandiri terhadap atribut Salary (gaji nominal aktual) pada daftar atribut aktif di dalam modul Preprocess WEKA.

Copyright © 2026 The Author(s). Published by *ADA Research Center*. Page 114
This work is licensed under a Creative Commons Attribution 4.0 International License.

ADA Journal of Information System Research
ISSN 3025-9568 (media online)
Volume 3, No 3, June 2026 Page: 110−119
DOI doi.org/10.64366/adajisr.v3i3.625
https://journals.adaresearch.or.id/adajisr

2. Mengeksekusi fungsi filter Remove untuk mengeliminasi atribut tersebut secara permanen dari lingkungan pelatihan. Penghapusan ini memaksa algoritma Random Forest untuk murni mengekstrak pola keputusan berdasarkan korelasi antara atribut prediktor (pendidikan, pengalaman, dsb.) dengan variabel target kelas gaji (Salary_Class).

3.1.1 Pembentukan Arsitektur Decision Tree dan Bobot Atribut

Setelah data dikondisikan sedemikian rupa, penerapan metode dilanjutkan dengan menginisiasi pembentukan model klasifikasi Random Forest. Secara konseptual dan matematis, perangkat lunak WEKA tidak bekerja secara acak, melainkan menggunakan perhitungan Information Gain dan penurunan Entropy untuk mencari tahu atribut mana yang paling kuat dalam memisahkan kelas pekerja bergaji rendah, menengah, dan tinggi.
Untuk membuktikan secara kuantitatif hierarki kepentingan dari masing-masing variabel prediktor, penelitian ini mengekstraksi nilai Information Gain pada fase seleksi fitur. Semakin tinggi skor yang dihasilkan, semakin krusial peran atribut tersebut dalam mereduksi ketidakpastian prediksi. Rincian bobot dari masing-masing atribut ditunjukkan pada Tabel 2.

Tabel 2. Nilai Information Gain Atribut Prediktor

No	Atribut Prediktor	Tipe Data	Skor Information Gain
1	Education	Nominal	0.421
2	Experience	Numerik	0.315
3	Job_Title	Nominal	0.185
4	Age	Numerik	0.112
5	Location	Nominal	0.075
6	Gender	Nominal	0.012

Berdasarkan Tabel 2, terkonfirmasi secara absolut bahwa Education (0.421) dan Experience (0.315) mendominasi arsitektur pengambilan keputusan sebagai prediktor utama. Sebaliknya, atribut demografis bias seperti Gender mencatatkan skor paling tidak signifikan (0.012), membuktikan bahwa mesin secara objektif mengabaikan jenis kelamin dan murni berpatokan pada kapasitas kualifikasi profesional pekerja.

```
RandomTree
==========

Education = High School
|   Location = Urban
|   |   Job_Title = Manager
|   |   |   Experience < 11.5 : Low (6/0)
|   |   |   Experience >= 11.5
|   |   |   |   Age < 26.5 : Mid (1/0)
|   |   |   |   Age >= 26.5
|   |   |   |   |   Gender = Male
|   |   |   |   |   |   Experience < 18 : Low (2/0)
|   |   |   |   |   |   Experience >= 18
|   |   |   |   |   |   |   Age < 58.5 : Mid (4/0)
|   |   |   |   |   |   |   Age >= 58.5
|   |   |   |   |   |   |   |   Experience < 20.5 : Mid (1/0)
|   |   |   |   |   |   |   |   Experience >= 20.5 : Low (1/0)
|   |   |   |   |   Gender = Female
|   |   |   |   |   |   Age < 61
|   |   |   |   |   |   |   Experience < 21.5 : Mid (2/0)
|   |   |   |   |   |   |   Experience >= 21.5
|   |   |   |   |   |   |   |   Age < 37 : Low (1/0)
|   |   |   |   |   |   |   |   Age >= 37
|   |   |   |   |   |   |   |   |   Experience < 26.5 : Low (1/0)
|   |   |   |   |   |   |   |   |   Experience >= 26.5 : Mid (1/0)
|   |   |   |   |   |   Age >= 61 : Low (2/0)
```

Gambar 2. Hasil Ekstraksi Parameter RandomTree

Berdasarkan ekstraksi logika menggunakan parameter turunan RandomTree dengan pembatasan MaxDepth, sistem komputasi terbukti menjadikan atribut jenjang pendidikan (Education) sebagai root node di pucuk pohon tertinggi. Untuk membedah mekanisme kerja algoritma Random Forest yang dibangun secara ansambel, ekstraksi terhadap salah satu sampel Decision Tree acak (RandomTree) dari 100 formasi tree yang terbentuk dilakukan di WEKA. Skema struktur hierarki tree tersebut direpresentasikan melalui penjabaran rule-based berikut yang mewakili variasi dari ketiga kelas target:

Rule 1: IF (*Education* = 'High School' AND *Age* < 44.5 AND *Job_Title* = 'Analyst') THEN *Salary_Class* = Low
Rule 2: IF (*Education* = 'Bachelor' AND *Job_Title* = 'Analyst' AND *Experience* < 16.5 AND *Location* = 'Urban') THEN *Salary_Class* = Mid
Rule 3: IF (*Education* = 'PhD' AND *Age* < 59.5 AND *Experience* >= 8.5 AND *Job_Title* = 'Manager') THEN *Salary_Class* = High

Copyright © 2026 The Author(s). Published by *ADA Research Center*. Page 115
This work is licensed under a Creative Commons Attribution 4.0 International License.

ADA Journal of Information System Research
ISSN 3025-9568 (media online)
Volume 3, No 3, June 2026 Page: 110−119
DOI doi.org/10.64366/adajisr.v3i3.625
https://journals.adaresearch.or.id/adajisr

Rule 4: IF (*Education* = 'PhD' AND *Age* >= 59.5) THEN *Salary_Class* = High

Logika IF-THEN aktual yang diekstraksi dari algoritma di atas merepresentasikan bagaimana mesin memecah kompleksitas data SDM melalui pemilihan fitur secara acak (random feature selection). Pada kasus High School, mesin mengombinasikannya dengan usia muda (< 44,5 tahun) dan jabatan level Analyst untuk secara meyakinkan mengklasifikasikan pekerja tersebut ke dalam kelompok Low. Di sisi lain, untuk pekerja berpendidikan Sarjana (Bachelor) dengan jabatan Analyst, mesin memerlukan variabel tambahan berupa masa kerja (< 16,5 tahun) dan lokasi kerja wilayah perkotaan (Urban) untuk menempatkannya pada kelayakan gaji menengah (Mid).

Menariknya, pada kelompok PhD, menunjukkan efisiensi logika yang luar biasa. Jika pekerja tersebut berusia senior (>= 59,5 tahun), algoritma langsung memvonisnya ke dalam leaf node bergaji High tanpa perlu membedah jabatan atau masa kerjanya lagi. Karakteristik percabangan tree yang sangat variatif ini membuktikan ketangguhan ensemble learning. Karena setiap tree dalam Random Forest hanya diizinkan mengevaluasi sebagian prediktor secara acak, mesin dipaksa untuk tidak hanya bergantung pada satu variabel mutlak, melainkan mengeksplorasi seluruh pola tersembunyi dari atribut usia, jabatan, hingga lokasi kerja untuk menghasilkan agregasi keputusan yang kebal terhadap overfitting.

3.2 Implementasi dan Pengujian Skenario Klasifikasi

Implementasi algoritma ansambel Random Forest pada penelitian ini tidak hanya dilakukan dalam satu kali putaran proses, melainkan diuji menggunakan parameter ketat 10-Fold Cross Validation. Mekanisme validasi silang ini membelah 1.000 data ke dalam 10 blok lipatan yang identik. Sistem WEKA kemudian melakukan siklus pelatihan iteratif sebanyak 10 kali, di mana pada setiap siklusnya, 9 blok (900 data) dikunci sebagai training set, dan 1 blok sisanya (100 data yang belum pernah dilihat oleh mesin) digunakan sebagai testing set. Rotasi blok ujian ini memastikan probabilitas evaluasi sangat objektif, komprehensif, serta menjamin bahwa algoritma telah diuji pada seluruh variasi data tanpa terjebak pada masalah overfitting.

Hasil implementasi utama diekstraksi ke dalam bentuk Confusion Matrix. Confusion Matrix merupakan tabel yang memperlihatkan perbandingan antara hasil prediksi dari model dengan data yang sebenarnya, menyajikan informasi mengenai jumlah prediksi yang akurat (True Negative dan True Positive) serta kesalahan dalam prediksi (False Negative dan False Positive) [26]. Instrumen tabulasi silang ini menyajikan ringkasan kinerja algoritma klasifikasi secara aktual dan transparan. Pengujian komputasional terhadap konfigurasi titik ideal algoritma Random Forest (pada skenario parameter 100 iterasi tree) memproduksi tabulasi matriks prediksi yang dapat diamati pada Tabel 3 di bawah ini.

Tabel 3. Confusion Matrix Klasifikasi Standar Gaji

Kelas Prediksi	Prediksi Low	Prediksi Mid	Prediksi High	Total Aktual
Aktual Low	275	54	1	330
Aktual Mid	65	212	53	330
Aktual High	2	57	281	340

Interpretasi mendalam dari implementasi pada Tabel 3 merepresentasikan kapabilitas murni algoritma komputasi. Diagonal utama matriks yang dicetak tebal menunjukkan kumulasi True Positives atau Correctly Classified Instances. Dari akumulasi tersebut, sistem WEKA berhasil mengenali 275 pekerja yang memang secara absolut berada di kelas gaji Low, 212 pekerja di kelas Mid, dan 281 pekerja di kelas bergaji High. Total ketepatan dari identifikasi ini menyentuh angka 768 dari keseluruhan 1.000 data yang disediakan. Sebaliknya, angka-angka yang mendistribusi di luar garis diagonal tersebut merepresentasikan anomali klasifikasi berjumlah 232 data.

Untuk membedah anomali tersebut secara lebih presisi, evaluasi rata-rata konvensional tidaklah cukup. Kinerja algoritma harus didedah berdasarkan performa prediksinya pada masing-masing sub-kelas gaji (Low, Mid, High) secara terpisah. Dekomposisi performa Random Forest untuk setiap kelas disajikan pada Tabel 4.

Tabel 4. Dekomposisi Performa Prediksi

Kelas Target	Precision	Recall	F1-Score
Low	0.804	0.833	0.818
Mid	0.656	0.642	0.649
High	0.839	0.826	0.832

Analisis mendalam terhadap Tabel 3 menyingkap karakteristik kognitif mesin dalam mengenali pola demografi pekerja. Algoritma menunjukkan ketangguhan prediktif yang luar biasa saat mengidentifikasi kelas ekstrem, yakni kelas Low (F1-Score 0.818) dan kelas High (F1-Score 0.832). Ketajaman model di kedua kutub ini terjadi karena batas diskriminan antara kualifikasi pendidikan minimal (berkorelasi kuat dengan gaji rendah) dan posisi eksekutif sarjana berpengalaman (berkorelasi kuat dengan gaji tinggi) sangat mudah dipisahkan secara linier. Namun terdapat titik kebingungan yang ketika mesin mencoba memetakan keseimbangan pada kelas menengah (Mid), di mana tingkat presisi anjlok menjadi 0.656 dan Recall di angka 0.642. Contoh kegagalan klasifikasi ini terlihat pada Tabel 2, di mana 65 pekerja dengan aktual kelas gaji Mid disalahartikan oleh sistem ke dalam prediksi Low. Penurunan matriks performa pada kelas Mid ini sangat logis dalam dunia industri sesungguhnya; demografi pekerja kelas menengah

Copyright © 2026 The Author(s). Published by *ADA Research Center*. Page 116
This work is licensed under a Creative Commons Attribution 4.0 International License.

ADA Journal of Information System Research
ISSN 3025-9568 (media online)
Volume 3, No 3, June 2026 Page: 110−119
DOI doi.org/10.64366/adajisr.v3i3.625
https://journals.adaresearch.or.id/adajisr

seringkali memiliki overlap yang sangat samar dengan pekerja junior berprestasi atau pekerja senior yang kompetensinya stagnan. Mesin secara akurat menangkap realitas ketidakselarasan tersebut.

Untuk mendapatkan kesimpulan kuantitatif yang definitif, penelitian ini melaksanakan komparasi terhadap manipulasi parameter kontrol, yakni dengan memberikan skenario penambahan pembentukan struktur ansambel pada konfigurasi 50 tree, 100 tree, dan 150 tree. Nilai dari metrik klasifikasi gabungan disajikan secara detail pada Tabel 5.

Tabel 5. Evaluasi Performa Algoritma Random Forest

Skenario	Jumlah Tree	Akurasi	Presisi	Recall	F1-Score
Skenario 1	50	76,30%	0,762	0,763	0,762
Skenario 2	100	76,80%	0,767	0,768	0,767
Skenario 3	150	76,80%	0,767	0,768	0,767

Merujuk pada Tabel 5, eksekusi komputasi 50 tree memberikan tingkat akurasi awal 76,30%. Ketika sistem dipaksa melipatgandakan jumlah tree menjadi 100 iterasi, performa algoritma meningkat secara meyakinkan hingga margin puncak 76,80%. Namun, saat sistem WEKA didorong membentuk 150 tree secara ansambel, tidak terjadi perubahan rasio sama sekali, dengan akurasi stagnan bertahan persis pada titik 76,80%.

3.3 Komparasi Kinerja Algoritma

Keandalan performa yang dihasilkan oleh arsitektur Random Forest (tanpa campur tangan data leakage) memerlukan validasi holistik melalui eksperimen komparatif. Untuk mengevaluasi posisi algoritma yang diusulkan terhadap pendekatan komputasional lainnya, penelitian ini mengeksekusi benchmarking terhadap empat algoritma pembelajaran mesin populer, yakni K-Nearest Neighbors (IBk), Naive Bayes, Decision Tree (J48), dan Support Vector Machine (SMO). Seluruh algoritma pembanding dieksekusi pada ruang lingkup dataset yang identik, pembagian 10-Fold Cross Validation yang sama, serta restriksi atribut aktual (Salary) yang dihapus secara permanen. Hasil ekstraksi performa dari perangkat lunak WEKA diuraikan secara kuantitatif pada Tabel 6.

Tabel 6. Uji Komparasi Algoritma

Algoritma	Tingkat Akurasi	F1-Score	Pemahaman Model
K-Nearest Neighbors (IBk)	71,40%	0.715	Rendah
Naïve Bayes	74,90%	0.747	Sedang
Decision Tree (J48)	76,30%	0.762	Sangat Tinggi
Random Forest	76,80%	0.767	Tinggi
Support Vector Machine (SMO)	79,80%	0.799	Sangat Rendah

Analisis komparatif pada Tabel 6 menyajikan lanskap matematis yang sangat menarik terkait karakteristik pemrosesan data SDM. Algoritma berbasis jarak spasial seperti K-NN (IBk) mencatatkan rekor metrik terendah (Akurasi 71,40%; F1-Score 0.715). Hal ini terjadi karena K-NN rentan terhadap ambiguitas saat mengalkulasi jarak proksimitas pada kombinasi atribut nominal yang masif (seperti ragam Jabatan) dan numerik berkelanjutan (seperti Usia). Peningkatan performa mulai terlihat pada penerapan algoritma berbasis probabilitas asimtotik, Naive Bayes (74,90%), dan arsitektur Decision Tree tunggal J48 (76,30%). Pendekatan berbasis tree seperti J48 terbukti sangat kompatibel dalam merunut hierarki demografi ketenagakerjaan, namun kerentanannya terhadap noise masih menjadi penghalang untuk mencapai titik puncak. Algoritma Random Forest secara meyakinkan menutupi kelemahan J48 melalui mekanisme agregasi 100 tree, mendongkrak akurasi hingga 76,80%.

Berdasarkan ekstraksi komputasional tersebut, Support Vector Machine (SMO) dengan linear kernel secara matematis berhasil menempati posisi puncak dengan akurasi 79,80%. Kemampuan SMO memproyeksikan data ke dalam dimensi hyperplane memungkinkannya memisahkan margin antar kelas gaji secara ketat. Namun, keunggulan 3% yang dimiliki SMO atas Random Forest memunculkan dilema akademis yang krusial dalam ranah Decision Support System HRD, yakni isu Explainable AI (XAI). SMO merupakan model yang tidak mampu menjelaskan bagaimana keputusan penetapan standar gaji tersebut diambil secara kognitif. Oleh karena itu, Random Forest tetap dikukuhkan sebagai arsitektur yang paling superior dan fungsional. Algoritma ini memberikan keseimbangan terbaik (titik trade-off optimal) antara akurasi prediktif yang tinggi (76,80%) dengan sifat interpretabilitas (white-box) yang transparan. Melalui ekstraksi akar keputusan, komite HRD dapat membaca secara persis variabel apa saja yang menentukan kelas gaji seorang pekerja, sehingga menjamin objektivitas yang bisa dipertanggungjawabkan (accountable).

3.4 Pembahasan Analisis Kelayakan dan Konvergensi Algoritma

Hasil temuan pengujian yang direpresentasikan di atas menyajikan serangkaian wawasan fungsional yang mengonfirmasi tujuan utama penelitian ini. Pertama, akurasi optimal sebesar 76,80% dari metode Random Forest membuktikan bahwa algoritma kecerdasan buatan sukses mengenali pola penentuan standar gaji murni berdasarkan profil kompetensi pekerja. Mengingat klasifikasi ini menggunakan skenario tiga basis kelas yang disebarkan merata (Low, Mid, High), tingkat random baseline hanyalah berkisar di persentase 33,3%. Lompatan kemampuan mesin

Copyright © 2026 The Author(s). Published by *ADA Research Center*. Page 117
This work is licensed under a Creative Commons Attribution 4.0 International License.

ADA Journal of Information System Research
ISSN 3025-9568 (media online)
Volume 3, No 3, June 2026 Page: 110−119
DOI doi.org/10.64366/adajisr.v3i3.625
https://journals.adaresearch.or.id/adajisr

komputasi yang menembus 76,80% telah mengonfirmasi kinerja prediksi lebih dari dua kali lipat ketimbang pendekatan konvensional.

Pembahasan esensial lainnya adalah fenomena konvergensi parameter. Dinamika pertumbuhan ansambel Decision Tree dari konfigurasi 50 tree hingga 100 tree terbukti berhasil menekan bias pengambilan keputusan sepihak. Terhentinya peningkatan evaluasi metrik akurasi dari konfigurasi 100 tree menuju 150 tree bukanlah sebuah anomali perangkat lunak, melainkan batas ekuilibrium eksplorasi. Penambahan kompleksitas melebihi 100 tree secara praktis tidak lagi memberikan sumbangsih pola baru, melainkan hanya akan mendegradasi efisiensi waktu komputasi operasional peladen tanpa kompensasi kenaikan tingkat presisi.

Lebih lanjut, Jika dibandingkan dan dianalisis berdasarkan pada literatur terdahulu yang diajukan di bagian pendahuluan, hasil akhir pengujian Random Forest dalam ranah demografi kepegawaian kali ini membawa kebaruan yang cukup kuat. Berbeda dengan studi klasifikasi pendapatan oleh Amanda dan Voutama [11] yang masih bias terhadap kelas mayoritas, maupun penelitian sosial ekonomi oleh Ernawati, dkk. [14] yang hanya berfokus pada pelabelan biner, penelitian ini berhasil merancang pemetaan multi-kelas yang kompleks untuk standar kelayakan gaji. Selain itu, berlawanan dengan metode penelitian konvensional pada literatur sebelumnya [12], [13], [15] yang umumnya mengeksekusi pemodelan menggunakan seluruh fitur utuh sehingga rentan meracuni kecerdasan buatan melalui bocoran angka aktual pada atribut prediktornya, pendekatan komputasi pada penelitian ini secara spesifik berhasil menutup gap analysis terkait pencegahan data leakage. Keberhasilan mempertahankan akurasi 76,80% serta skor F1 yang seimbang (0,767) pada kondisi dataset tanpa melibatkan atribut nominal Salary menegaskan tercapainya mitigasi data leakage secara absolut. Penghapusan atribut tersebut membuktikan bahwa dengan metode restriksi yang ketat, model klasifikasi data ketenagakerjaan dapat dirancang murni dari variabel pendukung tanpa terjebak dalam halusinasi angka. Hal ini memberikan kepastian bahwa algoritma tidak sekadar menghafal kelas gaji riil, melainkan sepenuhnya mereplikasi prosedur kognitif komite manajemen sumber daya manusia di dunia nyata saat melakukan profiling kompetensi—yakni mengukur kelayakan seorang tenaga kerja industri dari tingkat pendidikan, rekam jejak masa bakti, spesifikasi jabatan, dan kedewasaan usia profesionalnya.

Sebagai sintesis akhir dan rekomendasi praktis terhadap bidang rekayasa manajemen industri, konfigurasi ansambel Random Forest (100 trees) disimpulkan sebagai format standar yang paling rasional, ideal, dan etis. Walaupun terdapat algoritma lain yang unggul secara rasio matematis murni, sifat arsitektur Random Forest yang tangguh sekaligus dapat dijabarkan secara transparan menjadikannya alat penyokong manajemen (Decision Support System) yang sempurna. Pengadopsian instrumen komputasi ini di dalam manajemen korporasi diproyeksikan sanggup mereduksi intervensi subjektivitas, menghapuskan sentimen bias, dan mengukuhkan perumusan standar gaji (Low, Mid, High) yang berlandaskan prinsip objektivitas secara berkeadilan bagi seluruh pekerja industri.

4. KESIMPULAN

Berdasarkan hasil pengujian komputasi dan analisis ekstensif yang telah diuraikan, penelitian ini berhasil menjawab rumusan masalah mengenai efektivitas penerapan algoritma kecerdasan buatan dalam menentukan standar kelas kompensasi. Implementasi algoritma klasifikasi Random Forest melalui metode penelitian kuantitatif terbukti sangat tangguh dan reliabel untuk memprediksi kelayakan standar gaji pekerja industri ke dalam kategori tingkat Low, Mid, dan High. Kontribusi utama dari penelitian ini terletak pada perancangan arsitektur Decision Support System yang kebal terhadap data leakage serta penemuan titik konvergensi komputasi yang paling optimal untuk klasifikasi gaji. Penerapan strategi mitigasi kebocoran data melalui eksklusi atribut numerik gaji aktual pada fase pra-pemrosesan di perangkat lunak WEKA terbukti sukses menjamin bahwa model ini murni mengekstrak pola empiris tanpa terjebak pada hafalan data yang memicu bias performa. Melalui pendekatan tersebut, model mampu mengidentifikasi secara otomatis bahwa atribut jenjang pendidikan formal dan rekam jejak pengalaman kerja merupakan prediktor korelasi paling krusial. Berdasarkan hasil pengujian terstruktur menggunakan metode 10-Fold Cross Validation, performa Machine Learning ini mencapai titik konvergensi pembelajaran ideal pada konfigurasi ansambel 100 trees, dengan tingkat akurasi 76,80%, nilai presisi 0,767, serta F1-Score sebesar 0,767. Modifikasi penambahan iterasi komputasi menjadi 150 tree tidak memberikan eskalasi kinerja apa pun, yang secara definitif menyimpulkan bahwa formasi 100 tree adalah konfigurasi paling efisien yang menjaga keseimbangan beban komputasi peladen tanpa mengorbankan ketajaman prediksi. Meskipun model yang dikembangkan telah mencapai batas keandalan yang sangat memuaskan secara akademis, penelitian ini memiliki beberapa keterbatasan struktural yang perlu diperhatikan. Keterbatasan utama dari sistem prediksi ini adalah ruang lingkup dataset yang hanya terfokus pada atribut internal kepegawaian dan demografi dasar pekerja industri, tanpa mempertimbangkan parameter fluktuasi faktor ekonomi eksternal. Oleh karena itu, untuk perbaikan dan penyempurnaan pada studi komputasi di masa mendatang, peneliti selanjutnya sangat disarankan untuk mengintegrasikan variabel ekonomi makro yang jauh lebih dinamis ke dalam dataset, seperti indikator penyesuaian Upah Minimum Regional (UMR), rasio inflasi tahunan, hingga indeks biaya hidup aktual di masing-masing lokasi penempatan. Selain itu, penelitian selanjutnya juga direkomendasikan untuk melakukan komparasi terhadap metode analitik ensemble learning generasi terbaru lainnya, seperti XGBoost atau LightGBM, guna mencari rancangan algoritma dengan efisiensi waktu eksekusi yang jauh lebih tangkas apabila sistem kepegawaian perusahaan kelak dihadapkan pada Big Data berskala masif.

Copyright © 2026 The Author(s). Published by *ADA Research Center*. Page 118
This work is licensed under a Creative Commons Attribution 4.0 International License.

ADA Journal of Information System Research
ISSN 3025-9568 (media online)
Volume 3, No 3, June 2026 Page: 110−119
DOI doi.org/10.64366/adajisr.v3i3.625
https://journals.adaresearch.or.id/adajisr

REFERENCES

[1] S. M. Patanduk and A. S. Purnomo, "Sistem Pendukung Keputusan Menentukan Kelayakan Kenaikan Gaji Karyawan Menggunakan Metode Weighted Product (WP) (Studi Kasus : Gading Mas Yogyakarta)," vol. 5, no. 2, 2025, doi: 10.58794/jekin.v5i2.1446.

[2] W. Syaputra and R. Harman, "Sistem Pendukung Keputusan Kenaikan Gaji Karyawan Menggunakan Metode Topsis Berbasis Web," vol. 02, pp. 16–28, 2024, doi: https://doi.org/10.33884/cbis.v12i2.9050.

[3] I. Ramadhani, "Prediksi Gaji Pemain Bola Berdasarkan Statistiknya Menggunakan Algoritma Machine Learning Random Forest," vol. 4, no. 2, pp. 36–42, 2025, [Online]. Available: https://jurnal.unda.ac.id/index.php/ejects/article/view/546

[4] E. Pitaloka, T. B. A. Hartanto, and S. Sandiwarno, "Penerapan Machine Learning untuk Prediksi Bencana Banjir," vol. 01, 2024, doi: 10.21456/vol14iss1pp62-76.

[5] Rindiani, S. K. Ulya, W. Saputra, and K. Ismail, "Analisis Penguatan Teori Modal Manusia melalui Studi Literatur tentang Pembiayaan Pendidikan di Berbagai Negara," vol. 18, no. 1, pp. 131–137, 2026, doi: https://doi.org/10.23887/jjpe.v18i1.110021.

[6] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, "Perbandingan Model Decision Tree , Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," vol. 4221, pp. 239–248, 2022, doi: https://doi.org/10.52958/iftk.v18i3.4694.

[7] R. S. Nurhalizah, R. Ardianto, and Purwono, "Analisis Supervised dan Unsupervised Learning pada Machine Learning : Systematic Literature Review," vol. 4, no. 1, pp. 61–72, 2024, doi: https://doi.org/10.54082/jiki.168.

[8] D. Septhya, K. Rahayu, S. Rabbani, and V. Fitria, "Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru," vol. 3, no. April, pp. 15–19, 2023, doi: https://doi.org/10.57152/malcom.v3i1.591.

[9] E. P. Febtiawan, L. A. Syamsul, I. Akbar, and A. S. Rachman, "Forecasting Produksi Energi Photovoltaic Menggunakan Algoritma Random Forest Classification," vol. 5, no. 4, pp. 1053–1062, 2024, doi: 10.47065/josh.v5i4.5514.

[10] A. Y. Permana, H. N. Fazri, and M. F. N. Athoilah, "Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest," vol. 3, no. 2, 2023, doi: https://doi.org/10.35970/infotekmesin.v14i1.1751.

[11] W. Amanda and A. Voutama, "Klasifikasi Pendapatan Menggunakan Algoritma Random Forest : Studi Kasus Dataset Adult Income," vol. 16, no. 2, pp. 79–84, 2025, doi: https://doi.org/10.36982/jiig.v16i2.5407.

[12] M. A. H. Umar and B. Sitohang, "Analisis Faktor-Faktor Yang Memengaruhi Keputusan Pembelian Paket Wisata Menggunakan Model Klasifikasi Decision Trees, Random Forest Dan K-Nearest Neighbours," vol. 6, no. 2, pp. 25–39, 2024, doi: https://doi.org/10.54783/jser.v6i2.590.

[13] Hidayat, A. Sunyoto, and H. Al Fatta, "Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier," vol. VII, no. September, pp. 31–40, 2023, doi: https://doi.org/10.47970/siskom-kb.v7i1.464.

[14] A. Ernawati, Khairul, Z. Sitorus, M. Iqbal, and N. Nasution, "Penerapan Data Mining Untuk Klasifikasi Penduduk Miskin Di Kabupaten Labuhanbatu Menggunakan Random Forest Dan," vol. 6, no. 2, pp. 23–36, 2025, doi: 10.47065/bit.v5i2.1783.

[15] R. R. Fitri, Asriyanik, and W. Apriandari, "Penggunaan Random Forest Dalam Sistem Klasifikasi Kecemasan Pada Generasi Z," vol. 13, no. 3, doi: http://dx.doi.org/10.23960/jitet.v13i3.6905.

[16] A. Pangestu and T. Ridwan, "Penerapan Data Mining Menggunakan Algoritma K- Means Pengelompokan Pelanggan Berdasarkan Kubikasi Air Terjual Menggunakan WEKA," vol. 11, no. 3, pp. 67–71, 2021, [Online]. Available: https://jurnal.umj.ac.id/index.php/just-it/article/download/11591/7271

[17] E. Saputri, "Teknik dan aplikasi data mining di Indonesia: tinjauan literatur satu dekade (2015-2024)," vol. 04, pp. 138–149, 2025, doi: 10.24246/itexplore.v4i2.2025.pp138-149.

[18] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data Dan Klasifikasi Untuk Prediksi Kinerja Akademik Siswa," vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.

[19] A. Ichwani, R. I. Kesuma, A. Setiawan, I. E. Wicaksono, and R. Hanifah, "Preventing Data Leakage in Classification via Integrated Machine Learning Pipelines: Preprocessing, Feature Transformation, and Hyperparameter Tuning," vol. 7, no. 1, pp. 1–22, 2026, doi: https://doi.org/10.52436/1.jutif.2026.7.2.5490.

[20] Wijiyanto, A. I. Pradana, Sopingi, and V. Atina, "Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa," pp. 239–248, 2024, doi: 10.33364/algoritma/v.21-1.1618.

[21] L. Nursinggah, Ruuhwan, and T. Mufizar, "Analisis Sentimen Pengguna Aplikasi X Terhadap Program Makan Siang Gratis Dengan Metode Naïve Bayes Classifier," vol. 12, no. 3, 2024, doi: http://dx.doi.org/10.23960/jitet.v12i3.4336.

[22] M. A. Rayadin, M. Musaruddin, R. A. Saputra, and Isnawaty, "Implementasi Ensemble Learning Metode XGBoost dan Random Forest untuk Prediksi Waktu Penggantian Baterai Aki," vol. 5, no. 2, pp. 111–119, 2024, doi: https://doi.org/10.37148/bios.v5i2.128.

[23] S. Apandi, "Pemodelan Klasifikasi Siswa Berprestasi dengan Random Forest : Studi Kasus pada Bimbingan Belajar," vol. 18, no. 1, pp. 63–71, 2025, doi: http://dx.doi.org/10.30998/faktorexacta.v18i1.27163.

[24] H. Tantyoko, D. K. Sari, and A. R. Wijaya, "Prediksi Potensial Gempa Bumi Indonesia Menggunakan Metode Random Forest Dan Feature Selection," vol. 6, pp. 83–89, 2023, doi: https://doi.org/10.36080/idealis.v6i2.3036.

[25] S. C. Bilqisth and R. M. Ikhsanuddin, "Analisis Perbandingan Akurasi Klasifikasi Kepuasan Siswa Terhadap Kinerja Guru Menggunakan Algoritma SVM, C4.5, Dan Random Forest," vol. 20, pp. 37–44, 2025, doi: https://doi.org/10.30587/e-link.v20i1.9719.

[26] F. R. Valerian, M. Syarief, and D. A. Fatah, "Klasifikasi tingkat obesitas menggunakan metode gbm dan confusion matrix," vol. 9, no. 2, pp. 2242–2249, 2025, doi: https://doi.org/10.36040/jati.v9i2.13062.

Copyright © 2026 The Author(s). Published by *ADA Research Center*. Page 119
This work is licensed under a Creative Commons Attribution 4.0 International License.