Journal of Science and Social Research ISSN 2615 – 4307 (Print)
August 2026, IX (4): 6973 – 6978 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

KLASIFIKASI KESESUAIAN BIDANG PELATIHAN CALON TENAGA KERJA MENGGUNAKAN *INFORMATION GAIN* DAN *RANDOM FOREST*

Rika Nofitri[1], Suparmadi[2], Cecep Maulana[3]
Universitas Royal, Kisaran
e-mail: [1]nofitririka15@gmail.com

Abstract: *Selecting appropriate vocational training fields for prospective workers at Job Training Centers (BLK) is critical for optimizing skill development and ensuring job market alignment. However, manual selection processes and high-dimensional datasets containing irrelevant candidate attributes often lead to classification inaccuracies and computational noise. This study proposes an optimized classification model combining the Information Gain (IG) feature selection method with the Random Forest algorithm to enhance prediction accuracy for training field suitability. Using a dataset of 233 candidate records, feature weighting was executed via Information Gain to eliminate redundant demographic attributes. The feature selection identified four dominant predictors: Academic Potential Test (TPA) score, interview score, vocational skill score, and BLK training choice. Experimental evaluation was conducted using a 70:30 train-test split in RapidMiner Studio. The experimental results demonstrate that applying Information Gain reduced the feature space from nine to four attributes, increasing classification accuracy from 87.14% (baseline Random Forest) to 90.00% (+2.86%). Furthermore, the model achieved a notable precision enhancement, rising from 86.96% to 93.44% (+6.48%), and an improved F1-Score of 94.21%. These findings confirm that removing non-contributing demographic variables mitigates dataset noise and enables Random Forest to form more effective decision boundaries. The optimized model offers a robust, objective framework to support decision-making systems in candidate placement at vocational training institutions.*

Keywords: *Information Gain, Random Forest, Feature Selection, Training Field Classification, RapidMiner.*

Abstrak: Penentuan bidang pelatihan vokasi yang tepat bagi calon peserta di Balai Latihan Kerja (BLK) sangat krusial untuk mengoptimalkan pengembangan keterampilan dan kesesuaian dengan kebutuhan pasar kerja. Namun, proses seleksi manual dan tingginya dimensi dataset yang memuat atribut kandidat kurang relevan sering kali menyebabkan ketidakakuratan klasifikasi serta *noise* komputasi. Penelitian ini mengusulkan model klasifikasi teroptimasi yang menggabungkan metode seleksi fitur *Information Gain* (IG) dengan algoritma *Random Forest* untuk meningkatkan akurasi prediksi kesesuaian bidang pelatihan. Menggunakan dataset sebanyak 233 data calon peserta, pembobotan fitur dieksekusi melalui *Information Gain* untuk mengeliminasi atribut demografis yang redundan. Hasil seleksi fitur berhasil mengidentifikasi empat prediktor dominan: nilai Tes Potensi Akademik (TPA), nilai wawancara, nilai kejuruan, dan pilihan BLK. Evaluasi eksperimental dilakukan dengan pembagian data *train-test* sebesar 70:30 pada perangkat lunak RapidMiner Studio. Hasil pengujian membuktikan bahwa penerapan *Information Gain* yang mereduksi ruang fitur dari sembilan menjadi empat atribut mampu meningkatkan akurasi klasifikasi dari 87,14% (model awal *Random Forest*) menjadi 90,00% (+2,86%). Selain itu, model ini mencapai peningkatan presisi yang signifikan, naik dari 86,96% menjadi 93,44% (+6,48%), serta kenaikan *F1-Score* menjadi 94,21%. Temuan ini mengonfirmasi bahwa penghapusan variabel demografis yang tidak berkontribusi berhasil mengurangi *noise* data dan memungkinkan *Random Forest* membentuk pohon keputusan yang lebih efektif. Model teroptimasi ini

memberikan kerangka kerja yang objektif dan andal untuk mendukung sistem pengambilan keputusan dalam penempatan calon peserta di lembaga pelatihan vokasi.

Kata kunci: *Information Gain*, *Random Forest*, Seleksi Fitur, Klasifikasi Bidang Pelatihan, RapidMiner.

PENDAHULUAN

Perkembangan teknologi informasi dan kemampuan komputasi telah mendorong pemanfaatan data sebagai aset strategis dalam proses pengambilan keputusan secara cepat, objektif, dan berbasis bukti. Di sektor ketenagakerjaan, Balai Latihan Kerja (BLK) sebagai unit pelaksana teknis di bawah Dinas Ketenagakerjaan memegang peranan krusial dalam meningkatkan kompetensi calon tenaga kerja melalui berbagai program pelatihan vokasi (Saputra & Sarnawa, 2022). Namun, penentuan kesesuaian bidang pelatihan bagi calon peserta umumnya masih dilakukan secara konvensional. Padahal, ketidaktepatan penempatan bidang pelatihan berpotensi menurunkan efektivitas proses pembelajaran, menghambat peningkatan kompetensi peserta, serta mengurangi peluang penyerapan di dunia kerja (Irwansya et al., 2024).

Dalam proses seleksi calon peserta pelatihan, BLK mengumpulkan berbagai atribut registrasi yang mencakup karakteristik demografis (jenis kelamin, usia, tingkat pendidikan, status perkawinan, dan pengalaman kerja) serta hasil evaluasi seleksi (nilai Tes Potensi Akademik, nilai wawancara, nilai kejuruan, dan pilihan bidang pelatihan). Kompleksitas atribut yang tinggi dan keberadaan variabel yang tidak relevan sering kali memicu penumpukan data (*high-dimensional data*) yang mengandung *noise* (gangguan). Jika data historis ini diolah tanpa eliminasi atribut redundan, model prediksi yang dihasilkan berpotensi mengalami penurunan performa komputasi dan ketidakakuratan klasifikasi.

Teknik *Data Mining*, khususnya algoritma klasifikasi *Random Forest*, dikenal luas memiliki keunggulan berupa tingkat akurasi yang tinggi, ketahanan terhadap *overfitting*, serta keandalan dalam mengolah dataset berdimensi besar (Amaliah et al., 2022). Meskipun demikian, beberapa penelitian terdahulu yang mengaplikasikan algoritma klasifikasi pada penentuan bidang pelatihan belum menerapkan tahapan seleksi fitur (*feature selection*) sebelum pemodelan (Ramadhan et al., 2025). Akibatnya, seluruh atribut demografis yang bersifat pasif tetap dilibatkan, sehingga dapat mengaburkan variabel-variabel penentu yang sebenarnya lebih dominan. Oleh karena itu, pengintegrasian metode seleksi fitur berbasis *Information Gain* menjadi sangat krusial untuk mengukur entropi dan bobot kontribusi informasi dari setiap atribut, sehingga hanya variabel-variabel paling relevan yang masuk ke dalam proses pemodelan (Firana et al., 2026).

Penelitian ini mengusulkan penerapan kombinasi metode *Information Gain* dan algoritma *Random Forest* untuk mengklasifikasikan kesesuaian bidang pelatihan calon tenaga kerja berdasarkan kerangka kerja CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Berada pada Tingkat Kesiapterapan Teknologi (TKT) 3 (*proof of concept*), penelitian ini memanfaatkan dataset historis sebanyak 233 sampel registrasi calon peserta di BLK. Rumusan masalah dan tujuan utama dari penelitian ini difokuskan pada tiga hal:

1. Menganalisis dan mengidentifikasi atribut-atribut dominan yang paling berpengaruh menggunakan seleksi fitur *Information Gain*.
2. Membangun model klasifikasi kesesuaian bidang pelatihan berbasis *Random Forest* menggunakan atribut hasil reduksi.

3. Mengevaluasi dan membandingkan peningkatan kinerja model sebelum dan sesudah seleksi fitur berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-score* menggunakan perangkat lunak RapidMiner.

Kontribusi utama dari penelitian ini adalah menyajikan kerangka kerja klasifikasi yang teroptimasi melalui penghentian (*pruning*) atribut demografis yang redundan, sehingga mampu memberikan rekomendasi penempatan calon peserta pelatihan secara lebih akurat, efisien, dan bebas dari bias subjektif. Hasil penelitian ini diharapkan dapat menjadi rujukan ilmiah serta rekomendasi teknis bagi pihak pengelola BLK dalam memodernisasi sistem seleksi dan penempatan peserta pelatihan berbasis kecerdasan buatan.

METODE

Penelitian ini menggunakan pendekatan kuantitatif berbasis eksperimen *data mining* untuk mengklasifikasikan kesesuaian bidang pelatihan bagi calon peserta di Balai Latihan Kerja (BLK). Kerangka kerja penelitian secara umum dibagi menjadi empat tahapan utama, yaitu pengumpulan dataset, pra-pemrosesan data (*data preprocessing*), seleksi fitur menggunakan *Information Gain* (IG), serta pemodelan dan evaluasi menggunakan algoritma *Random Forest*.

Dataset yang digunakan dalam penelitian ini merupakan data sekunder pendaftaran calon peserta pelatihan vokasi sebanyak 233 sampel record. Dataset awal terdiri dari sembilan atribut masukan (*regular attributes*), yaitu Status Perkawinan, Usia, Pendidikan, Jenis Kelamin, Pengalaman, Nilai TPA, nilai wawancara, nilai kejuruan, dan pilihan BLK, serta satu atribut target (*label*) yaitu Bidang Pelatihan. Pada tahap pra-pemrosesan data, dilakukan pembersihan data (*data cleaning*) untuk memastikan tidak terdapat nilai yang hilang (*missing values*) maupun entitas data yang ganda (*duplicate data*).

Tahap berikutnya adalah seleksi fitur berbasis *Information Gain* untuk mengatasi masalah tingginya dimensi atribut yang dapat memicu timbulnya *noise* pada model. *Information Gain* mengukur tingkat entropi dan kontribusi informasi dari setiap atribut X terhadap atribut target Y. Persamaan entropi $H(Y)$ dan *Information Gain* $IG(Y,X)$ dihitung menggunakan rumus pada Persamaan (1) dan Persamaan (2):

$$H(Y) = -\sum_{y \in Y} P\,(y) \log_2 P\,(y)$$

$$IG(Y,X) = H(Y) - \sum_{x \in X} \frac{|Y_x|}{|Y|} H(Y_x)$$

Atribut yang memiliki nilai bobot *Information Gain* di bawah batas ambang (*threshold*) 0,010 dieliminasi (*pruned*), sehingga hanya menyisakan atribut-atribut dengan kontribusi informasi paling signifikan.

Tahap pemodelan dilakukan menggunakan algoritma *Random Forest* pada perangkat lunak RapidMiner Studio. *Random Forest* merupakan teknik *ensemble learning* berbasis *bagging* yang membangun sejumlah pohon keputusan (*decision trees*) secara acak. Parameter utama yang diterapkan pada model mencakup *Number of Trees* sebanyak 100 pohon, kriteria pemisahan nodus (*split criterion*) berupa *Gain Ratio*, dan *Max Depth* sebesar 10. Pembagian dataset dieksekusi menggunakan teknik *Split Data* dengan rasio 70% sebagai data *training* (163 sampel) untuk pembentukan model dan 30% sebagai data *testing* (70 sampel) untuk pengujian performa.

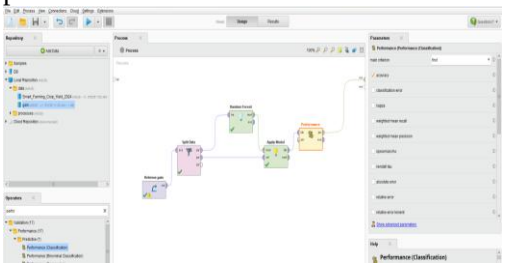
Evaluasi kinerja model dilakukan dengan membandingkan dua skenario eksperimen, yaitu *Random Forest* tanpa seleksi fitur (9 atribut) dan *Random Forest* dengan seleksi fitur *Information Gain* (4 atribut). Pengukuran performa dilakukan menggunakan matriks evaluasi *Confusion Matrix* berdasarkan empat

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
August 2026, IX (4): 1 – 6 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

metrik utama: *Accuracy*, *Precision*, *Recall*, dan *F1-Score*.

HASIL DAN PEMBAHASAN

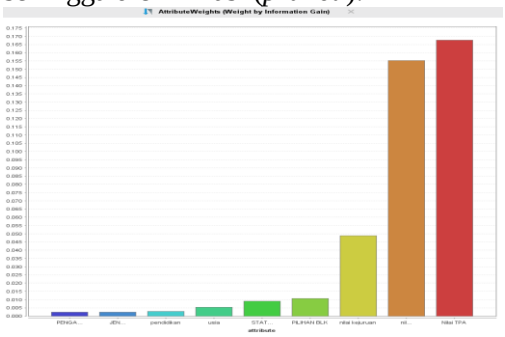
Pengujian model klasifikasi kesesuaian bidang pelatihan bagi calon peserta Balai Latihan Kerja (BLK) dilakukan dengan membandingkan dua skenario eksperimen pada perangkat lunak RapidMiner Studio. Skenario pertama menerapkan algoritma *Random Forest* menggunakan seluruh atribut awal (9 *regular attributes*), sedangkan skenario kedua menerapkan algoritma *Random Forest* yang dikombinasikan dengan metode seleksi fitur *Information Gain*. Pembagian data untuk kedua skenario menggunakan metode *Split Data* dengan rasio 70% sebagai data *training* (163 data) dan 30% sebagai data *testing* (70 data). Alur proses pemodelan dan pengujian pada RapidMiner Studio ditunjukkan pada Gambar 1.



Gambar 1 Alur Proses Pemodelan *Information Gain* dan *Random Forest* pada RapidMiner Studio

Seleksi Fitur Menggunakan Information Gain

Penerapan *Information Gain* bertujuan untuk mengukur kontribusi informasi setiap variabel terhadap variabel target (Bidang Pelatihan). Berdasarkan hasil kalkulasi pembobotan fitur, diperoleh urutan bobot variabel sebagai berikut: Nilai TPA (0,168), nilai wawancara (0,155), nilai kejuruan (0,049), dan PILIHAN BLK (0,011). Empat atribut ini terpilih untuk masuk ke dalam proses pemodelan. Sementara itu, lima atribut lainnya—termasuk variabel demografis seperti Status Perkawinan (0,008), Usia (0,007), Pendidikan (0,005), Jenis Kelamin (0,002), dan Pengalaman (0,002)—memiliki nilai pembobotan di bawah ambang batas (di bawah 0,010) sehingga dieliminasi (*pruned*).



Gambar 2 Bobot Fitur Hasil Perhitungan *Information Gain*

Perbandingan Performa Model Klasifikasi

Hasil pengujian klasifikasi sebelum dan sesudah penerapan seleksi fitur *Information Gain* dievaluasi menggunakan *Confusion Matrix* dengan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score*, yang dirangkum pada Tabel 1.

Tabel 1 Perbandingan Performa Model Random Forest Sebelum dan Sesudah Information Gain

Metrik Evaluasi	Baseline (9 Atribut)	Random Forest + IG (4 Atribut)	Selisih Kenaikan
Akurasi (*Accuracy*)	87,14%	90,00%	+2,86%
Presisi (*Precision*)	86,96%	93,44%	+6,48%
Recall	98,36%	95,08%	-3,28%
F1-Score	93,02%	94,21%	+1,19%

Berdasarkan Tabel 1, terjadi peningkatan kinerja model secara signifikan pada skenario kedua. Nilai akurasi keseluruhan mengalami kenaikan dari 87,14% menjadi **90,00%** (meningkat sebesar 2,86%). Peningkatan paling tajam terlihat pada nilai presisi kelas target (*Cocok*), yaitu naik dari 86,96% menjadi 93,44% (meningkat sebesar 6,48%). Nilai *F1-Score* yang mencerminkan keseimbangan antara presisi dan *recall* juga mengalami kenaikan dari 93,02%

menjadi 94,21%.

Analisis Pengaruh Fitur Dominan dan Implikasi Praktis

Peningkatan performa model klasifikasi setelah diterapkannya *Information Gain* memberikan bukti empiris bahwa tidak semua variabel dalam data registrasi calon peserta pelatihan memberikan dampak positif terhadap keputusan penempatan. Variabel-variabel demografis seperti usia, jenis kelamin, dan status perkawinan terbukti bersifat *noise* (gangguan) dalam proses pencarian pola keputusan. Ketika variabel-variabel yang tidak relevan ini dieliminasi, algoritma *Random Forest* mampu membentuk pohon-pohon keputusan (*decision trees*) yang lebih dangkal, efisien, dan fokus pada kriteria kompetensi utama.

Atribut Nilai TPA, nilai wawancara, dan nilai kejuruan menjadi prediktor paling dominan karena secara langsung merepresentasikan kemampuan kognitif, motivasi, dan kesiapan teknis calon peserta. Secara praktis, temuan ini memberikan panduan obyektif bagi pengelola lembaga pelatihan vokasi (BLK) untuk mentransformasi sistem seleksi manual. Dengan memfokuskan kriteria kelulusan pada bobot penilaian akademis dan wawancara teknis, proses klasifikasi tidak hanya menjadi lebih akurat (90,00%), tetapi juga meminimalkan potensi bias subjektif demografis dalam penentuan bidang pelatihan kerja.

Uji Hipotesis Pertama

Pengujian hipotesis pertama diawali dengan membuktikan pengaruh seleksi fitur *Information Gain* terhadap tingkat akurasi algoritma *Random Forest*. Hipotesis nol (H_0) menyatakan bahwa penerapan seleksi fitur tidak memberikan peningkatan akurasi secara signifikan, sedangkan hipotesis kerja (H_1) menyatakan bahwa *Information Gain* mampu meningkatkan akurasi klasifikasi. Berdasarkan hasil eksperimen pada perangkat lunak RapidMiner Studio, nilai akurasi sebelum seleksi fitur menggunakan sembilan atribut awal adalah sebesar 87,14%. Setelah dilakukan seleksi fitur dan hanya menyisakan empat atribut terpilih, akurasi klasifikasi meningkat menjadi 90,00%, atau terjadi kenaikan sebesar 2,86%. Kenaikan angka akurasi ini secara ilmiah membuktikan bahwa H_0 ditolak dan H_1 diterima, yang berarti seleksi fitur *Information Gain* terbukti secara signifikan meningkatkan akurasi model *Random Forest*.

Uji Hipotesis Kedua

Pengujian dilanjutkan pada hipotesis kedua untuk menilai dampak eliminasi atribut redundan terhadap presisi prediksi model. Hipotesis nol (H_0) mengasumsikan bahwa pereduksian atribut tidak berpengaruh pada presisi kelas target, sementara hipotesis kerja (H_2) mengemukakan bahwa penyiangan atribut redundan dapat meningkatkan presisi prediksi. Pengujian menunjukkan bahwa nilai presisi pada kelas target mengalami kenaikan signifikan sebesar 6,48%, yaitu dari 86,96% pada model tanpa seleksi fitur menjadi 93,44% pada model yang dioptimasi dengan *Information Gain*. Hasil ini secara konsisten menolak H_0 dan menerima H_2, yang mengonfirmasi bahwa penghapusan variabel demografis yang bersifat *noise* berhasil meminimalkan kesalahan prediksi *false positive* dan mempertajam tingkat presisi model.

Uji Hipotesis Ketiga

Pengujian hipotesis ketiga difokuskan untuk mengidentifikasi tingkat dominansi atribut berbasis penilaian akademis dan teknis dibandingkan variabel demografis. Hipotesis nol (H_0) menyatakan atribut penilaian tidak memiliki kontribusi yang lebih dominan, sedangkan hipotesis kerja (H_3) menyatakan bahwa atribut penilaian menjadi variabel paling dominan dalam penentuan kesesuaian bidang pelatihan. Pengukuran bobot *Information Gain* membuktikan bahwa atribut Nilai TPA dengan bobot 0,168, nilai wawancara

dengan bobot 0,155, dan nilai kejuruan dengan bobot 0,049 menempati posisi teratas. Sebaliknya, atribut demografis seperti Jenis Kelamin dan Pengalaman hanya memiliki bobot yang sangat rendah sebesar 0,002. Dengan demikian, H_0 ditolak dan H_3 diterima, yang menegaskan bahwa kriteria kompetensi dan ujian teknis merupakan faktor penentu utama dalam klasifikasi kelayakan calon peserta pelatihan.

SIMPULAN

Penelitian ini berhasil membuktikan bahwa pengintegrasian metode seleksi fitur *Information Gain* pada algoritma *Random Forest* mampu meningkatkan performa klasifikasi kesesuaian bidang pelatihan bagi calon peserta Balai Latihan Kerja (BLK). Penerapan *Information Gain* terbukti efektif mereduksi dimensi data dari sembilan atribut menjadi empat atribut dominan, yaitu Nilai TPA, nilai wawancara, nilai kejuruan, dan PILIHAN BLK, dengan mengeliminasi atribut demografis yang bersifat *noise*. Reduksi fitur ini berhasil meningkatkan akurasi klasifikasi *Random Forest* dari 87,14% menjadi 90,00% (+2,86%), presisi dari 86,96% menjadi 93,44% (+6,48%), serta *F1-Score* menjadi 94,21%. Temuan ini mengonfirmasi bahwa kriteria penilaian akademis dan uji kompetensi merupakan prediktor utama dalam menentukan kelayakan calon peserta, sehingga model teroptimasi ini dapat dimanfaatkan oleh pengelola BLK sebagai dasar sistem pendukung keputusan yang objektif, akurat, dan berbasis data.

DAFTAR PUSTAKA

Amaliah, S., Nusrang, M., & Aswi, A. (2022). Penerapan metode Random Forest untuk klasifikasi varian minuman kopi di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its Application in Teaching and Research*, 4(3), 121–127. https://doi.org/10.35580/variansiunm31

Firana, N. A., Afrianty, I., Novriyanto, N., & Yanto, F. (2026). Information Gain and Random Forest for sex classification based on craniometric measurements. *Journal of Artificial Intelligence and Software Engineering*, 6(2), 118–129. https://doi.org/10.30811/jaise.v6i2.9409

Irwansya, I., Ramadhan, S., & Lorosae, T. A. (2024). Perancangan sistem informasi pencari kerja pada Dinas Tenaga Kerja dan Transmigrasi Kabupaten Kep. Sula Propinsi Maluku Utara menggunakan Visual Basic.Net. *Jurnal Mediatik*, 7(2), 107–113. https://doi.org/10.59562/mediatik.v7i2.1304

Ramadhan, S., et al. (2025). Penerapan metode Random Forest dalam mengklasifikasi. *Jurnal Penelitian Komputer*, 9(6), 9433–9440.

Saputra, I. E., & Sarnawa, B. (2022). Peran Dinas Tenaga Kerja dalam perlindungan terhadap hak-hak atas upah pekerja. *Media Law Sharia*, 3(4), 284–300. https://doi.org/10.18196/mls.v3i4.14330