TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1452-1459
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9193

Klasifikasi Rentang Gaji Lowongan Pekerjaan di Glints Wilayah Jabodetabek Menggunakan Regresi Logistik dan Random Forest Berbasis Web Scraping

Evander Banjarnahor[1*], Theodore Miracle Setiawan[2], Wellson Antonio Charlest[2], Ronald Belferik[3]

[1] Fakultas Sains dan Teknologi, Program Studi Matematika, Universitas Pelita Harapan, Tangerang, Indonesia
[2] Fakultas Artificial Intelligence and Data Science, Program Studi Sistem Informasi, Universitas Pelita Harapan, Tangerang, Indonesia
[3] Fakultas Artificial Intelligence and Data Science, Program Studi Informatika, Universitas Pelita Harapan, Tangerang, Indonesia
Email: [1,*]evander.banjarnahor@uph.edu, [2]01081240018@student.uph.edu, [3]01081240010@student.uph.edu, [4]ronald.belferik@uph.edu
Email Penulis Korespondensi: evander.banjarnahor@uph.edu

Abstrak–Transformasi digital telah mengubah pasar tenaga kerja, di mana platform daring seperti Glints berperan sebagai repositori data besar yang menghubungkan pencari kerja dengan perusahaan. Di wilayah Jabodetabek, informasi gaji menjadi faktor krusial dalam pengambilan keputusan karier, namun asimetri informasi terkait gaji masih menjadi tantangan utama. Penelitian ini diawali dengan analisis deskriptif terhadap 1.497 lowongan kerja yang dikumpulkan melalui teknik web scraping untuk memahami distribusi gaji berdasarkan lokasi dan status pekerjaan. Gaji lowongan kemudian diklasifikasikan ke dalam tiga kategori, yaitu gaji rendah (≤ 5 juta), gaji menengah (5–10 juta), dan gaji tinggi (≥ 10 juta). Hasil analisis menunjukkan bahwa mayoritas lowongan berada pada kategori gaji rendah (77,09%), diikuti oleh gaji menengah (21,37%), sementara gaji tinggi hanya mencakup 1,54% dari total data. Selanjutnya, penelitian ini bertujuan mengembangkan model klasifikasi untuk memprediksi kategori gaji dengan membandingkan dua metode pembelajaran mesin, yaitu Regresi Logistik dan Random Forest. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score pada beberapa skenario pembagian data latih dan uji. Hasil eksperimen menunjukkan bahwa Random Forest secara konsisten memberikan kinerja terbaik dengan akurasi tertinggi mencapai 98,00%, jauh melampaui Regresi Logistik yang mencapai sekitar 79%. Temuan ini mengindikasikan bahwa hubungan antara karakteristik pekerjaan dan kategori gaji bersifat kompleks dan non-linier, sehingga lebih efektif dimodelkan menggunakan algoritma ensemble berbasis pohon. Penelitian ini diharapkan dapat berkontribusi pada peningkatan transparansi gaji dan pengembangan sistem prediksi gaji berbasis data yang lebih akurat dan informatif.

Kata Kunci: Machine Learning; Klasifikasi Gaji; Random Forest; Regresi Logistik

Abstract–Digital transformation has reshaped the labor market, with online platforms such as Glints serving as large-scale data repositories that connect job seekers with employers. In the Greater Jakarta (Jabodetabek) region, salary information is a critical factor in career decision-making; however, salary-related information asymmetry remains a major challenge. This study begins with a descriptive analysis of 1,497 job vacancies collected through web scraping techniques to examine salary distributions across locations and employment statuses. The salaries were classified into three categories: low salary (≤ IDR 5 million), medium salary (IDR 5–10 million), and high salary (≥ IDR 10 million). The results indicate that the majority of job vacancies fall into the low-salary category (77.09%), followed by the medium-salary category (21.37%), while high-salary positions constitute only 1.54% of the total dataset. Subsequently, this study aims to develop salary category classification models by comparing two machine learning methods: Logistic Regression and Random Forest. Model performance was evaluated using accuracy, precision, recall, and F1-score under multiple training–testing split scenarios. The experimental results demonstrate that Random Forest consistently outperforms Logistic Regression, achieving a highest accuracy of 98.00%, compared to approximately 79% for Logistic Regression. These findings suggest that the relationship between job characteristics and salary categories is complex and non-linear, making it more effectively captured by ensemble-based, non-linear algorithms such as Random Forest. This study contributes to improving salary transparency and supports the development of more accurate and data-driven salary prediction systems.

Keywords: Machine Learning; Random Forest; Regresi Logistik; Salary Classification

1. PENDAHULUAN

Percepatan transformasi digital telah membawa perubahan mendasar pada struktur dan mekanisme pasar tenaga kerja global (Izzatul Mula & Auliya Ristiani, 2025). Digitalisasi tidak hanya mengubah cara organisasi beroperasi, tetapi juga mentransformasi proses rekrutmen tenaga kerja yang sebelumnya bergantung pada interaksi tatap muka dan media konvensional. Saat ini, proses pencarian dan penawaran kerja semakin bergeser ke platform rekrutmen daring yang terintegrasi, memungkinkan pertukaran informasi secara cepat, masif, dan lintas wilayah. Di Indonesia, fenomena ini tercermin dari meningkatnya penggunaan platform seperti Glints, LinkedIn, dan JobStreet yang menjadi penghubung utama antara pencari kerja dan pemberi kerja di berbagai sektor industri.

Platform rekrutmen digital tersebut tidak lagi berfungsi sebatas papan pengumuman lowongan kerja, melainkan telah berkembang menjadi repositori Big Data yang dinamis. Data yang terkumpul mencakup berbagai atribut penting, seperti deskripsi pekerjaan, kualifikasi pendidikan, keterampilan teknis dan non-teknis, lokasi kerja, tingkat pengalaman, serta informasi kompensasi yang diperbarui secara real-time (Alsheyab et al., 2025). Skala data yang besar dan kecepatan pembaruan informasi ini membuka peluang luas bagi penelitian berbasis data untuk menganalisis dinamika pasar tenaga kerja secara lebih akurat dan komprehensif dibandingkan metode survei tradisional yang bersifat statis dan terbatas.

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1452-1459
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9193

Meskipun demikian, pemanfaatan data lowongan kerja daring menghadapi sejumlah tantangan metodologis, khususnya terkait dengan informasi gaji. Pada banyak kasus, gaji tidak disajikan dalam bentuk nilai numerik tunggal, melainkan dalam bentuk rentang yang sangat lebar, kategori tertentu, atau bahkan tidak dicantumkan sama sekali. Ketidaklengkapan dan ambiguitas data gaji ini menyebabkan pendekatan regresi konvensional menjadi kurang efektif. Memaksakan data rentang gaji menjadi satu nilai numerik dapat menghasilkan estimasi yang bias dan kurang merepresentasikan kondisi sebenarnya di pasar tenaga kerja. Oleh karena itu, diperlukan pendekatan pemodelan yang lebih fleksibel dan robust terhadap karakteristik data dunia nyata.

Berbagai penelitian terdahulu menunjukkan bahwa machine learning telah banyak diterapkan dalam domain sumber daya manusia, khususnya untuk analisis dan prediksi gaji. Machine learning dipahami sebagai pendekatan komputasi yang memanfaatkan data historis untuk mempelajari pola kompleks dan meningkatkan kinerja prediksi suatu sistem (Ramadhan et al., 2023). Sejumlah studi melaporkan bahwa penerapan model machine learning mampu menghasilkan tingkat akurasi yang relatif tinggi, bahkan mencapai 80–90%, dalam memprediksi gaji berdasarkan karakteristik individu dan pekerjaan (Liu, 2023). Fitur-fitur seperti pengalaman kerja, tingkat pendidikan, lokasi geografis, dan jenis pekerjaan umumnya digunakan sebagai prediktor utama dalam model-model tersebut linier (Maehendrayuga et al., 2024; Malaiarasan et al., 2025; ms, 2023).

Namun, sebagian besar penelitian sebelumnya masih berfokus pada pendekatan regresi, seperti regresi linear dan variannya, untuk memprediksi nilai numerik gaji (Tuah & Anyan, 2020; Das et al., 2020). Pendekatan ini relatif efektif ketika data bersifat terstruktur dan lengkap, tetapi menjadi kurang praktis ketika diterapkan pada data lowongan kerja daring yang memiliki banyak nilai hilang, variasi ekstrem, serta representasi gaji dalam bentuk rentang. Di Indonesia, penelitian terkait prediksi dan analisis gaji masih terbatas dan cenderung berfokus pada faktor makroekonomi seperti upah minimum regional atau pada prediksi gaji untuk profesi tertentu menggunakan model regresi linier (Maehendrayuga et al., 2024; ms, 2023). Dengan demikian, terdapat celah penelitian yang signifikan terkait eksplorasi pendekatan alternatif, khususnya pemodelan gaji sebagai masalah klasifikasi berbasis data rekrutmen daring.

Berdasarkan celah penelitian tersebut, penelitian ini mengusulkan pendekatan klasifikasi untuk memodelkan prediksi gaji. Dengan mengelompokkan gaji ke dalam kelas atau kategori tertentu, pendekatan ini diharapkan lebih sesuai dengan karakteristik data lowongan kerja daring yang ambigu dan tidak lengkap. Penelitian ini secara khusus membandingkan kinerja dua algoritma machine learning, yaitu *Logistic Regression* (Regresi Logistik) dan Random Forest. Regresi Logistik dipilih karena kesederhanaan model, efisiensi komputasi, serta kemudahan interpretasi koefisien sebagai representasi hubungan linier antar variabel. Sementara itu, Random Forest dipilih sebagai model ensemble non-linier yang mampu menangkap interaksi kompleks antar fitur, mengatasi overfitting, serta menyediakan mekanisme feature importance untuk mengidentifikasi faktor-faktor paling berpengaruh dalam penentuan gaji (Amin & Utami, 2025).

Tujuan utama penelitian ini adalah mengklasifikasikan kategori gaji berdasarkan data lowongan kerja pada platform Glints di wilayah Jabodetabek, kemudian mengevaluasi dan membandingkan kinerja Regresi Logistik dan Random Forest dalam konteks klasifikasi gaji.

Kontribusi penelitian ini diharapkan dapat memperkaya literatur data science terapan, khususnya dalam studi pasar tenaga kerja berbasis Big Data di Indonesia. Selain memberikan kontribusi metodologis melalui penerapan pendekatan klasifikasi sebagai alternatif regresi, penelitian ini juga diharapkan memiliki implikasi praktis dalam mengurangi asimetri informasi antara pencari kerja dan pemberi kerja, serta mendukung pengambilan keputusan yang lebih transparan dan berbasis data di pasar tenaga kerja digital.

2. METODOLOGI PENELITIAN

2.1 Data Penelitian

Penelitian ini menggunakan data lowongan kerja yang diperoleh dari platform properti daring Glints. Data dikumpulkan melalui proses *web scraping* pada tanggal 15 Oktober 2025 dengan fokus pada lowongan di wilayah Jabodetabek. Dataset mentah yang diperoleh mencakup berbagai variabel yang berpotensi memengaruhi kompensasi, seperti judul pekerjaan, deskripsi pekerjaan, kualifikasi yang dibutuhkan, keterampilan, lokasi spesifik, dan informasi gaji. Tabel 1 menunjukkan informasi variabel dari dataset lowongan kerja yang digunakan dalam penelitian ini.

Tabel 1. Informasi Variabel dari Dataset Lowongan Kerja

Nama Variabel	Deskripsi	Tipe Data
Lowongan	Nama lowongan yang disediakan	Kategorikal
Status Pekerjaan	Status pekerjaan seperti: Freelance, PenuhWaktu, dan Kontrak	Kategorikal
Pengalaman	Pengalaman sudah bekerja lama karyawan bekerja	Ordinal
Ketentuan	Pendidikan harus ditempuh jika ingin melamar	Kategorikal
Soft_skill	Skill dibutuhkan untuk pekerjaan	Kategorikal
Perusahaan	Nama dari perusahaan	Kategorikal
Kecamatan	Kecamatan dari lowongan pekerjaan yang ditawarkan	Kategorikal
Kota	Kota dari lowongan pekerjaan yang ditawarkan	Kategorikal
Provinsi	Provinsi dari lowongan pekerjaan yang ditawarkan	Kategorikal

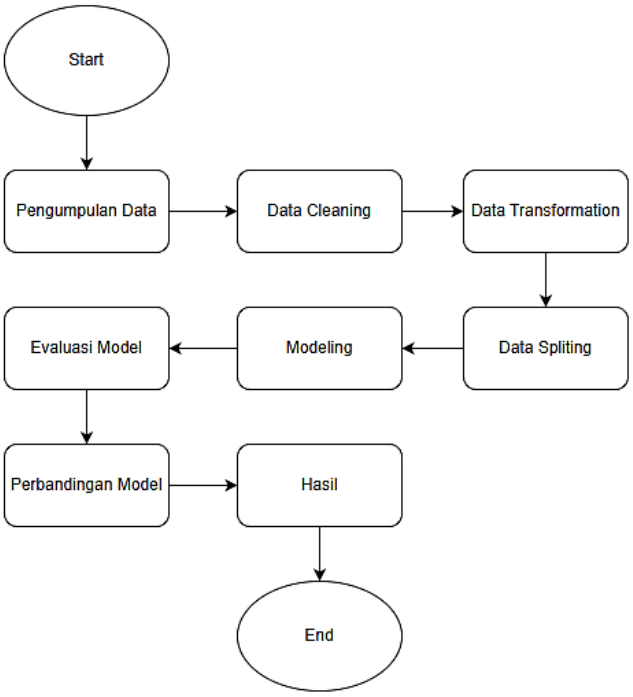
TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1452-1459
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9193

Nama Variabel	Deskripsi	Tipe Data
Rentang gaji	Range gaji yang ditentukan	Kategorikal

Setelah proses pengumpulan, data melalui tahap pembersihan untuk menangani nilai kosong, data duplikat, serta data pencilan yang tidak wajar agar hasil analisis menjadi lebih akurat. Variabel dalam dataset ini dibedakan menjadi variabel independen dan dependen. Variabel dependen adalah Kategori_Gaji yang merupakan variabel kategorikal yang dibentuk dari data gaji mentah. Variabel independen mencakup fitur terstruktur seperti lowongan, status pekerjaan, Lokasi dan lain sebagainya.

2.2 Alur Penelitian

Proses pengolahan data dimulai dengan tahap pembersihan data, di mana dataset diperiksa secara menyeluruh untuk mengidentifikasi dan menangani anomali. Langkah ini mencakup penghapusan entri duplikat yang dapat membiaskan hasil analisis, penanganan nilai kosong melalui teknik imputasi yang sesuai atau penghapusan baris jika diperlukan, serta konversi tipe data untuk memastikan konsistensi format di seluruh dataset (Kundu et al., 2020). Setelah data bersih, tahapan selanjutnya adalah fitur encoding. Variabel kategorikal status pekerjaan, diubah menjadi representasi numerik menggunakan metode One-Hot Encoding agar dapat diproses oleh algoritma pembelajaran mesin. Untuk variabel lainnya seperti lowongan, kecamatan, kota, provinsi dan lainnya) menggunakan label encoding, penggunaan label encoding dikarenakan lowongan yang dimunculkan ditiap kota memiliki perbedaan UMR yang berbeda-beda. Setelah proses fitur encoding dilakukan, dataset dipisah menjadi data latih dan data testing. Pemisahan data training dan testing dilakukan dengan 3 perbandingan yaitu 90:10, 80:20 dan 70:30. Pemisahan ini bertujuan untuk melatih model menggunakan sebagian besar data dan kemudian mengevaluasi kinerja serta kemampuan generalisasinya menggunakan data yang belum pernah dilihat sebelumnya. Gambar 1 menunjukkan alur penelitian yang diusulkan.



Gambar 1. Alur Penelitian

2.3 Regresi Logistik

Logistic Regression (Regresi Logistik) sebuah metode statistik yang digunakan untuk memodelkan hubungan antara variabel dependen dengan satu atau lebih variabel independen (Rianti & Andarsyah, 2024). Regresi Logistik adalah salah satu teknik statistik dan *machine learning* fundamental yang paling banyak digunakan untuk tugas klasifikasi biner atau multikelas (Yusuf et al., 2025; Gopal et al., 2021). Regresi logistik digunakan untuk menjelaskan hubungan antara variabel bebas X dan variabel tak bebas Y yang bersifat dikotomi/biner dengan outcome1 atau 0 (Reskiawati et al., 2025). Secara matematis, LR memodelkan probabilitas bahwa suatu *input* x termasuk dalam kelas 1 (misalnya, 'Gaji Tinggi') menggunakan fungsi sigmoid (σ):

$$P(y = 1|x) = \sigma(z) = \frac{1}{1+e^{-z}} \quad (1)$$

Di mana z adalah kombinasi linier dari fitur-fitur masukan x dan parameter bobot β :

$$z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \quad (2)$$

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1452-1459
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9193

Model ini pada intinya memodelkan *log-odds* (logit) dari probabilitas, yang memungkinkan interpretabilitas koefisien (β). Keunggulan utama LR terletak pada interpretabilitasnya. Koefisien yang dihasilkan dari model dapat diinterpretasikan untuk memahami bagaimana setiap fitur berkontribusi terhadap probabilitas output (Wewengkang et al., 2025). Dalam konteks analisis gaji di Indonesia, Regresi Logistik telah terbukti efektif untuk menganalisis faktor-faktor yang memengaruhi upah pekerja (Das et al., 2020). Dalam penelitian ini, LR berfungsi sebagai *baseline* model linier yang kuat untuk dibandingkan dengan model non-linier.

2.4 Random Forest

Random Forest merupakan salah satu model dalam Machine Learning yang mengumpulkan sejumlah besar Decision tree dari set data latihan, dan juga menggunakan alat yang disebut bagging untuk melakukan tugas klasifikasi dan regresi(Ananda Surya et al., 2025; Ramadhani, 2025). *Random Forest* adalah metode *ensemble learning* non-linier yang sangat efektif yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi, baik untuk tugas regresi maupun klasifikasi (Banjarnahor, Belferik, et al., 2025; Banjarnahor, Sibarani, et al., 2025). Algoritma ini beroperasi dengan membangun sejumlah besar pohon keputusan pada saat pelatihan (Gao et al., 2019).

Kekuatan metodologis algoritma ini berasal dari kemampuannya mengurangi varians tinggi yang sering menjadi masalah pada pohon keputusan tunggal sehingga dapat mencegah *overfitting*. Hal ini dicapai melalui dua teknik utama: *Bagging* di mana setiap pohon dilatih pada sebagian data yang diambil secara acak, dan *Random Subspace* di mana pada setiap titik pembelahan, model hanya mempertimbangkan sebagian fitur yang dipilih secara acak. Untuk tugas klasifikasi seperti dalam penelitian ini, prediksi akhir ($\hat{y}_{RF}$) ditentukan oleh *voting* mayoritas (modus) dari semua T:

$$\hat{y}_{RF} = mode\{\hat{y}_t(x)\}\, t = 1 \quad (3)$$

Di mana T adalah jumlah total pohon keputusan, dan $\hat{y}_t(x)$ adalah prediksi kelas dari pohon ke T untuk masukan x. Algoritma ini sangat relevan untuk data pasar tenaga kerja karena kemampuannya menangani interaksi yang kompleks dan non-linier antar fitur. Selain itu, metode ini juga menyediakan metrik tingkat kepentingan fitur yang berguna untuk mengidentifikasi prediktor gaji paling berpengaruh (Ismail & Hidayah, 2025).

3. HASIL DAN PEMBAHASAN

Analisis deskriptif dilakukan sebagai tahap awal untuk memahami karakteristik dataset yang digunakan dalam penelitian ini. Dataset terdiri dari 1.497 data lowongan kerja yang dikumpulkan dari platform Glints dan berfokus pada wilayah Jabodetabek. Analisis ini bertujuan untuk mengidentifikasi distribusi variabel target berupa kategori gaji, serta hubungan awal antara gaji dengan fitur-fitur utama seperti lokasi dan status pekerjaan. Pemahaman terhadap struktur data ini penting untuk memastikan kesesuaian pendekatan machine learning yang digunakan pada tahap selanjutnya. Distribusi gaji berdasarkan lokasi dan status pekerjaan disajikan pada Tabel 2.

Tabel 2. Data Distribusi Gaji Berdasarkan Lokasi dan Status Pekerjaan

Kategori	<= 5 juta	5 - 10 juta	>= 10 juta	Grand Total
Lokasi				
Bekasi	109	18		127
Bogor	23	4	2	29
Depok	141	5		146
JakartaBarat	78	42	2	122
JakartaPusat	24	25	2	51
JakartaSelatan	159	89	12	260
JakartaTimur	80	2	3	85
JakartaUtara	103	40	2	145
Kab.Bekasi	63	23		86
Kab.Bogor	75	3		78
Kab.KepulauanSeribu	2			2
Kab.Tangerang	120	32		152
Tangerang	79	27		106
TangerangSelatan	98	10		108
Total	1154	320	23	1497
Status Pekerjaan				
Freelance	123	2		125
Harian	2			2
Hybrid	41			41
Kontrak	26	75	10	111
Magang	19			19
ParuhWaktu	158	12		170
PenuhWaktu	756	231	13	1000

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1452-1459
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9193

Kategori	<= 5 juta	5 - 10 juta	>= 10 juta	Grand Total
Remote/Darirumah	29			29
Grand Total	1154	320	23	1497

Secara umum, terlihat bahwa distribusi kategori gaji sangat tidak seimbang (*imbalanced*). Dari total 1.497 lowongan kerja, sebanyak 1.154 lowongan (77,09%) berada pada kategori gaji ≤ 5 juta rupiah. Kategori gaji menengah, yaitu 5–10 juta rupiah, mencakup 320 lowongan (21,37%), sedangkan kategori gaji tinggi ≥ 10 juta rupiah hanya berjumlah 23 lowongan (1,54%). Ketimpangan ini menunjukkan bahwa mayoritas lowongan kerja yang dipublikasikan secara terbuka pada platform Glints masih berada pada level gaji rendah, yang kemungkinan besar merepresentasikan posisi entry-level atau pekerjaan dengan kebutuhan keterampilan dasar.

Berdasarkan distribusi geografis, Jakarta Selatan tercatat sebagai wilayah dengan jumlah lowongan kerja terbanyak, yaitu 260 lowongan, diikuti oleh Kabupaten Tangerang (152) dan Depok (146). Jakarta Selatan juga memiliki konsentrasi lowongan dengan kategori gaji tinggi yang paling besar, yaitu 12 dari total 23 lowongan bergaji ≥ 10 juta. Temuan ini mengonfirmasi posisi Jakarta Selatan sebagai pusat bisnis, keuangan, dan industri kreatif di Jabodetabek, yang umumnya menawarkan kompensasi lebih kompetitif dibandingkan wilayah penyangga lainnya. Sebaliknya, wilayah seperti Bekasi, Bogor, dan Kabupaten Bogor didominasi oleh lowongan dengan kategori gaji ≤ 5 juta. Hal ini mengindikasikan adanya kesenjangan geografis dalam distribusi gaji, yang kemungkinan dipengaruhi oleh tingkat urbanisasi, konsentrasi industri, dan daya beli regional. Temuan ini sejalan dengan literatur yang menyatakan bahwa lokasi geografis merupakan salah satu determinan utama dalam penentuan tingkat gaji, khususnya di kawasan metropolitan dengan heterogenitas ekonomi yang tinggi.

Dari perspektif status pekerjaan, tipe pekerjaan Penuh Waktu (Full-Time) mendominasi dataset dengan total 1.000 lowongan (66,8%). Sebagian besar lowongan penuh waktu berada pada kategori gaji ≤ 5 juta, yaitu sebanyak 756 lowongan (75,6%). Hal ini menunjukkan bahwa meskipun pekerjaan penuh waktu memberikan stabilitas kerja, kompensasi yang ditawarkan masih relatif rendah, terutama untuk posisi non-manajerial. Menariknya, status pekerjaan Kontrak menunjukkan proporsi gaji menengah yang relatif tinggi. Dari 111 lowongan kontrak, sebanyak 75 lowongan berada pada kategori gaji 5–10 juta. Pola ini mengindikasikan bahwa pekerjaan kontrak sering kali digunakan untuk peran-peran spesialis atau berbasis proyek yang membutuhkan keterampilan tertentu dengan kompensasi yang lebih tinggi. Sementara itu, pekerjaan Remote/Dari Rumah masih sangat terbatas jumlahnya dan seluruhnya berada pada kategori gaji terendah, yang mengindikasikan bahwa fleksibilitas kerja jarak jauh di Indonesia masih didominasi oleh pekerjaan dengan nilai tambah yang relatif rendah.

Setelah analisis deskriptif, tahap selanjutnya adalah implementasi model machine learning untuk memprediksi kategori gaji. Proses ini dimulai dengan preprocessing data, yang meliputi pembersihan data, penanganan nilai hilang, serta encoding variabel kategorikal seperti lokasi dan status pekerjaan. Variabel target berupa kategori gaji diklasifikasikan ke dalam tiga kelas: ≤ 5 juta, 5–10 juta, dan ≥ 10 juta.

Dua algoritma machine learning digunakan dalam penelitian ini, yaitu Regresi Logistik dan Random Forest. Regresi Logistik digunakan sebagai baseline model linier yang umum digunakan dalam tugas klasifikasi, sementara Random Forest digunakan sebagai model ensemble non-linier yang mampu menangkap interaksi kompleks antar fitur. Untuk memastikan evaluasi yang adil dan komprehensif, data dibagi ke dalam tiga skenario pembagian data latih dan uji, yaitu 90:10, 80:20, dan 70:30. Evaluasi kinerja model dilakukan menggunakan beberapa metrik, yaitu Accuracy, Precision, Recall, dan F1-Score. Penggunaan berbagai metrik ini penting mengingat distribusi kelas yang tidak seimbang, sehingga akurasi saja tidak cukup untuk menggambarkan kinerja model secara menyeluruh. Hasil evaluasi kinerja kedua model disajikan pada Tabel 3.

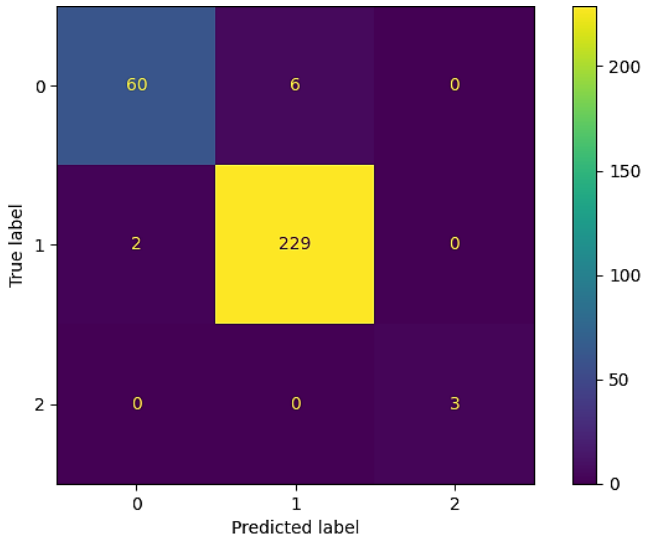
Tabel 3. Tabel Perbandingan Kinerja Model Logistik dan Random Forest

Trani/Test	Model	Accuracy	Precision	Recall	F1 Score
90:10:00	Regresi Logistik	78.00%	74.39%	78.00%	75.3%
	Random Forest	97.33%	97.33%	97.3%	97.3%
80:20:00	Regresi Logistik	79.33%	75.76%	79.33%	75.9%
	Random Forest	98.00%	98.03%	98.01%	98.01%
70:30:00	Regresi Logistik	80.22%	76.88%	80.22%	76.4%
	Random Forest	96.44%	96.5%	96.4%	96.4%

Berdasarkan Tabel 3, terlihat pola kinerja yang kontras antara kedua model yang diuji. Random Forest secara konsisten menunjukkan kinerja yang lebih tinggi dibandingkan Regresi Logistik pada seluruh skenario pengujian. Kinerja terbaik Random Forest diperoleh pada rasio pembagian data latih dan uji 80:20, dengan nilai akurasi sebesar 98,00% dan F1-Score sebesar 98,01%. Sebaliknya, Regresi Logistik menunjukkan performa yang relatif moderat dengan tingkat akurasi berkisar antara 78% hingga 80%. Meskipun demikian, nilai akurasi yang sangat tinggi pada model Random Forest perlu diinterpretasikan secara hati-hati mengingat karakteristik dataset yang digunakan. Distribusi kelas gaji yang sangat tidak seimbang, di mana kategori gaji rendah mendominasi lebih dari 77% total data, berpotensi menyebabkan metrik akurasi memberikan gambaran yang terlalu optimistis terhadap kinerja model. Dalam kondisi data yang imbalanced seperti ini, model dapat mencapai akurasi tinggi meskipun performanya terhadap kelas minoritas belum optimal.

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1452-1459
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9193

Analisis tren pada variasi rasio data latih dan uji menunjukkan bahwa Random Forest memiliki tingkat stabilitas yang relatif baik. Meskipun proporsi data latih dikurangi dari skenario 90:10 menjadi 70:30, nilai akurasi Random Forest tetap berada di atas 96%. Hal ini mengindikasikan bahwa model ensemble ini memiliki kemampuan generalisasi yang baik terhadap data uji. Di sisi lain, Regresi Logistik menunjukkan peningkatan akurasi yang relatif kecil seiring bertambahnya proporsi data uji, dari 78% menjadi 80,22%, namun secara keseluruhan tetap tertinggal jauh dari Random Forest. Perbedaan kinerja yang cukup signifikan antara kedua model (sekitar 18–20%) mengindikasikan keterbatasan pendekatan linier dalam menangkap hubungan non-linier dan interaksi kompleks antar fitur pada data gaji yang bersifat heterogen. Temuan ini sejalan dengan hasil penelitian Wewengkang et al. (2025), yang melaporkan bahwa algoritma berbasis tree atau ensemble learning cenderung lebih unggul dibandingkan model linier pada dataset dengan kompleksitas tinggi. Untuk memperoleh gambaran kinerja model yang lebih komprehensif, evaluasi lebih lanjut dilakukan menggunakan confusion matrix yang ditunjukkan pada Gambar 2.



Gambar 2. Confusion matriks pada model Random Forest

Confusion matrix tersebut menunjukkan bahwa model Random Forest memiliki kemampuan yang sangat baik dalam mengklasifikasikan kelas mayoritas, yaitu kategori gaji rendah dan menengah. Pada kelas gaji rendah, sebagian besar data berhasil diprediksi dengan benar, dengan sebagian kecil kesalahan klasifikasi ke kelas gaji menengah. Demikian pula pada kelas gaji menengah, hampir seluruh data diklasifikasikan secara tepat, dengan tingkat kesalahan yang sangat minimal. Namun, performa model pada kelas gaji tinggi perlu ditafsirkan secara lebih kritis. Meskipun seluruh data pada kelas gaji tinggi berhasil diprediksi dengan benar, jumlah data pada kelas ini sangat terbatas dibandingkan kelas lainnya. Kondisi ini menunjukkan bahwa keberhasilan prediksi pada kelas gaji tinggi belum dapat sepenuhnya mencerminkan kemampuan generalisasi model terhadap kelas minoritas, melainkan lebih dipengaruhi oleh ukuran sampel yang kecil. Dengan demikian, hasil confusion matrix mengonfirmasi bahwa nilai akurasi yang tinggi pada model Random Forest sebagian besar didorong oleh dominasi kelas mayoritas dalam dataset.

Berdasarkan temuan ini, penelitian selanjutnya disarankan untuk memperluas sumber dan jumlah dataset guna meningkatkan keberagaman serta representativitas data, khususnya pada kategori gaji tinggi yang memiliki proporsi sampel lebih sedikit. Selain itu, untuk mengatasi permasalahan data tidak seimbang, penerapan Synthetic Minority Over-sampling Technique (SMOTE) menjadi pendekatan yang relevan, karena metode ini mampu menghasilkan sampel sintetis pada kelas minoritas sehingga distribusi data menjadi lebih seimbang dan mengurangi bias model terhadap kelas mayoritas. Penggunaan SMOTE diharapkan dapat meningkatkan kemampuan model dalam mengenali pola pada kelas gaji tinggi serta memperbaiki kinerja prediksi secara menyeluruh. Lebih lanjut, eksplorasi algoritma ensemble learning lain seperti XGBoost, CatBoost, atau LightGBM juga dapat menjadi alternatif yang menjanjikan karena memiliki kemampuan yang lebih adaptif dalam menangani data yang kompleks dan tidak seimbang, sehingga berpotensi menghasilkan performa yang lebih optimal dibandingkan Random Forest.

4. KESIMPULAN

Penelitian ini mengidentifikasi adanya ketimpangan distribusi gaji pada lowongan kerja di wilayah Jabodetabek berdasarkan lokasi dan status pekerjaan. Mayoritas lowongan kerja yang dipublikasikan pada platform Glints berada pada kategori gaji rendah (≤ 5 juta rupiah) dengan proporsi sekitar 77%, sementara lowongan dengan gaji tinggi (≥ 10 juta rupiah) hanya mencakup sebagian kecil data dan cenderung terkonsentrasi di wilayah pusat aktivitas ekonomi, seperti Jakarta Selatan. Dari sisi status pekerjaan, tipe Penuh Waktu (Full-Time) mendominasi pasar kerja daring dan sebagian besar berada pada kategori gaji rendah, yang mencerminkan kuatnya karakteristik pekerjaan entry-level dalam ekosistem lowongan kerja digital yang dipublikasikan secara terbuka. Dalam aspek pemodelan, pendekatan klasifikasi

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1452-1459
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9193

gaji berbasis pembelajaran mesin terbukti mampu menangani data lowongan kerja digital yang bersifat heterogen, kompleks, dan mengandung noise. Random Forest secara konsisten menunjukkan kinerja yang lebih baik dibandingkan Regresi Logistik pada seluruh skenario pengujian, dengan akurasi tertinggi mencapai 98%. Namun demikian, hasil evaluasi juga menunjukkan bahwa tingginya nilai akurasi tersebut tidak terlepas dari kondisi dataset yang tidak seimbang, di mana kelas gaji rendah mendominasi data. Analisis confusion matrix memperlihatkan bahwa meskipun Random Forest mampu mengklasifikasikan kelas mayoritas dengan sangat baik, performa pada kelas minoritas perlu ditinjau secara lebih mendalam melalui metrik evaluasi berbasis kelas. Temuan ini menegaskan bahwa model ensemble non-linier lebih sesuai dibandingkan model linier dalam memodelkan data gaji dari platform digital, namun tetap memerlukan strategi penanganan ketidakseimbangan data. Untuk penelitian selanjutnya, disarankan penerapan teknik seperti *Synthetic Minority Over-sampling Technique* (SMOTE) atau pendekatan cost-sensitive learning guna meningkatkan kemampuan model dalam mengenali kelas minoritas dan menghasilkan evaluasi kinerja yang lebih adil serta andal.

REFERENCES

Alsheyab, A. R., Alkhasawneh, M., & Shahin, N. (2025). *Job Market Cheat Codes: Prototyping Salary Prediction and Job Grouping with Synthetic Job Listings*. Arxiv. http://arxiv.org/abs/2506.15879

Amin, R., & Utami, A. S. F. (2025). Prediksi Nilai Ujian Berdasarkan Kebiasaan Siswa Menggunakan Algoritma Random Forest Regressor. *Information System For Educators And Professionals : Journal of Information System*, *10*(2), 149. https://doi.org/10.51211/isbi.v10i2.3722

Ananda Surya, A., Rizki Darmawan, D., & Solichin, A. (2025). Prediksi Kapabilitas Calon Debitur Menggunakan Analisis Data Machine Learning Dengan Metode Random Forest. *Jurnal Algoritma*, *22*(1), 777–788. https://doi.org/10.33364/algoritma/v.22-1.1929

Banjarnahor, E., Belferik, R., Cendana, W., & Abraham, Y. A. S. (2025). Analisis Implementasi Support Vector Machine dan Random Forest untuk Prediksi Kategori Indeks Kualitas Udara Jakarta. *Jurnal INSTEK (Informatika Sains Dan Teknologi)*, *10*(1), 175–184. https://doi.org/10.24252/instek.v10i1.56477

Banjarnahor, E., Sibarani, D. P., Wibawanta, B., Sihotang, D. A. G., & Abraham, Y. A. S. (2025). A Machine Learning Approach to Predicting Student Success Through Data Mining of LMS Moodle Activity Data. *2025 4th International Conference on Electronics Representation and Algorithm (ICERA)*, 233–238. https://doi.org/10.1109/ICERA66156.2025.11086633

Das, S., Barik, R., & Mukherjee, A. (2020). Salary Prediction Using Regression Techniques. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3526707

Gao, X., Wen, J., & Zhang, C. (2019). An Improved Random Forest Algorithm for Predicting Employee Turnover. *Mathematical Problems in Engineering*, *2019*(1). https://doi.org/10.1155/2019/4140707

Gopal, K., Singh, A., & Sagar, S. (2021). *Salary Prediction Using Machine Learning*.

Ismail, & Hidayah, A. (2025). *Implementasi Machine Learning Dengan Metode Regresi Linear Untuk Prediksi Gaji Karyawan Berdasarkan Masa Kerja*. Jurnal Rister, *2*(1), 1–7. https://doi.org/10.25126/Rister

Izzatul Mula, & Auliya Ristiani. (2025). Transformasi Struktur Pekerjaan dan Kebutuhan Keterampilan di Era Teknologi AI dan Otomatisasi di Pasar Global. *Nian Tana Sikka : Jurnal Ilmiah Mahasiswa*, *3*(1), 155–167. https://doi.org/10.59603/niantanasikka.v3i1.665

Kundu, Souren, Mikhalev, O., Handerson, S., Bailey, Y. R., Peters, A., & Kundu, S. (2020). *Machine Learning for Salary Estimation: Insights from Logistic Regression*. https://www.researchgate.net/publication/391735353

Liu, X. (2023). Salary Grades Prediction Using Machine Learning. *Applied and Computational Engineering*, *8*(1), 248–255. https://doi.org/10.54254/2755-2721/8/20230152

Maehendrayuga, A., Setyanto, A., & Kusnawi. (2024). Analisa Prediksi Turnover Karyawan menggunakan Machine Learning. *Bit-Tech*, *7*(2), 648–659. https://doi.org/10.32877/bt.v7i2.1999

Malaiarasan, M. S., Ameer Riyaz, M., & Appadurai, M. (2025). Salary Prediction Using Machine Learning. *International Journal of Scientific Research and Engineering Development*, *8*(2)

ms, G. (2023). Salary Prediction System using Machine Learning. *Interantional Journal Of Scientific Research In Engineering And Management*, *07*(05). https://doi.org/10.55041/IJSREM22822

Ramadhan, B., Firdaus, D., & Adiningrum, N. T. R. (2023). Analisis Data Pegawai Untuk Memprediksi Gaji Berdasarkan Faktor-Faktor Spesifik Dengan Pendekatan Machine Learning. *Naratif : Jurnal Nasional Riset, Aplikasi Dan Teknik Informatika*, *5*(2), 131–139. https://doi.org/10.53580/naratif.v5i2.205

Reskiawati, Somayasa, W., & Adi Wibawa, G. (2025). Pemodelan Data Pemberian Asi Eksklusif Ibu Melahirkan Di Kelurahan Matabubu Dengan Analisis Regresi Logistik Biner. *Jurnal Matematika Komputasi Dan Statistika*, *5*(2). https://doi.org/10.33772/jmks.v5i2.145

Rianti, R., & Andarsyah, R. (2024). Memprediksi Tingkat Atrisi Karyawan Menggunakan Machine Learning. *Jurnal Tekno Insentif*, *18*(1), 39–52. https://doi.org/10.36787/jti.v18i1.1263

Tuah, Y. A. E., & Anyan, A. (2020). Implementasi Model Regresi Linear Sederhana Untuk Prediksi Gaji Berdasarkan Pengalaman Lama Bekerja. *JUTECH : Journal Education and Technology*, *1*(2), 56–70. https://doi.org/10.31932/jutech.v1i2.1289

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1452-1459
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9193

Wewengkang, R. C., Tirta Nugraha, Z., & Armera, A. M. (2025). *Prediksi Gaji Karyawan dengan Machine Learning Menggunakan Teknik Linear Regression dan Decision Tree*. Prosiding Seminar Nasional Penelitian LPPM UMJ, 2025, https://jurnal.umj.ac.id/index.php/semnaslit/article/view/29361

Yusuf, D., Razi, F., Arman, S. A., Terisia, V., & Nurjayanti, R. (2025). *Prediksi Risiko Stunting pada Balita menggunakan Algoritma Logistic Regression dan Decision Tree berbasis Data Terbuka*. Prosiding Semnastek, 2025, https://jurnal.umj.ac.id/index.php/semnastek/article/view/27662