

Perbandingan Algoritma *Machine Learning* untuk Prediksi Gaji Pekerjaan Berdasarkan Umur, Pengalaman dan Pendidikan

Comparison of Machine Learning Algorithms for Job Salary Prediction Based on Age, Experience, and Education

Rizki Tazidi¹, Marisa Premitasari², Zahra Hisyam^{3*}

^{1,2,3} Program Studi Informatika, Fakultas Teknologi Industri, Institut Teknologi Nasional Bandung
Jl.Khp Hasan Mustopa No.23, Neglasari, Kec. Cibeunying Kaler, Kota Bandung 40124, Jawa Barat, Indonesia

Riwayat Artikel

Submitted :20/05/2026
Published :28/08/2026

* **Email Korespondensi:**
zahra.hisyam@mhs.itenas.ac.id



Abstrak: Penentuan gaji karyawan yang adil dan kompetitif masih menjadi persoalan dalam manajemen sumber daya manusia. Penelitian ini membandingkan lima metode machine learning, yaitu Gradient Boosting, Regresi Linear, Polynomial Regression, K-Nearest Neighbors (KNN), dan Decision Tree, untuk memprediksi gaji berdasarkan umur, jenis kelamin, tingkat pendidikan, jabatan, dan pengalaman kerja. Dataset awal berjumlah 375 data karyawan; setelah pembersihan data kosong/duplikat dan penghapusan satu observasi gaji \$350 yang teridentifikasi sebagai anomali, tersisa 323 sampel. Gender dan Job Title diproses dengan One-Hot Encoding, Education Level dipetakan secara ordinal, dan seluruh preprocessing dilakukan di dalam 5-fold cross-validation. Gradient Boosting memperoleh mean R^2 tertinggi sebesar $0,8934 \pm 0,0349$ dan MAPE $11,42\% \pm 1,42\%$ pada kondisi tanpa normalisasi, sedangkan KNN dengan Z-Score mencapai R^2 $0,8824 \pm 0,0418$. Aplikasi web berbasis Streamlit digunakan untuk implementasi model. Hasil ini menunjukkan bahwa Gradient Boosting berpotensi digunakan sebagai model pendukung estimasi gaji, dengan tetap mempertimbangkan keterbatasan ukuran dan keragaman dataset.

Kata kunci: machine learning, prediksi gaji, gradient boosting, cross-validation, sistem kompensasi.

Abstract: Determining fair and competitive employee salaries remains a challenge in human resource management. This study compares five machine learning methods—Gradient Boosting, Linear Regression, Polynomial Regression, K-Nearest Neighbors (KNN), and Decision Tree—to predict salary using age, gender, education level, job title, and years of experience. The dataset initially contained 375 employee records; after removing missing/duplicate records and one anomalous \$350 salary observation, 323 samples remained. Gender and Job Title were processed using One-Hot Encoding, Education Level was ordinally encoded, and all preprocessing was fitted within 5-fold cross-validation. Gradient Boosting achieved the highest mean R^2 of 0.8934 ± 0.0349 with a MAPE of $11.42\% \pm 1.42\%$ without normalization, while KNN with Z-Score reached an R^2 of 0.8824 ± 0.0418 . A Streamlit-based web application was used to implement the final model. These results indicate that Gradient Boosting has potential as a salary-estimation decision-support model while the limited size and diversity of the dataset should be considered.

Keywords: machine learning, salary prediction, gradient boosting, cross-validation, compensation system.

1. PENDAHULUAN

Manajemen sumber daya manusia (SDM) banyak berubah sejak teknologi digital dan kecerdasan buatan mulai dipakai di berbagai perusahaan [1], [2]. Persaingan bisnis yang ketat membuat organisasi perlu mengelola SDM dengan lebih baik agar kinerja dan daya saingnya terjaga [1]. Salah satu bagian penting dari pengelolaan SDM adalah sistem kompensasi, termasuk penentuan gaji karyawan berdasarkan faktor yang dapat diukur seperti umur, pengalaman kerja, dan tingkat pendidikan [2].

Dalam praktiknya, menentukan gaji yang tepat tidak mudah [1], [2]. Gaji yang tidak sesuai dengan kompetensi dan kontribusi karyawan dapat menurunkan motivasi dan produktivitas, bahkan mendorong karyawan untuk keluar [3].

Sebaliknya, kompensasi yang adil dan sesuai kinerja justru dapat meningkatkan semangat kerja [1], [2]. Karena itu, penentuan gaji sebaiknya tidak hanya mengandalkan pertimbangan subjektif, tetapi juga didukung analisis data.

Machine learning dapat dimanfaatkan untuk menjawab permasalahan tersebut [4]. Dengan mempelajari pola dari data historis, machine learning mampu menghasilkan prediksi yang lebih objektif dibanding perkiraan manual [5]. Di bidang SDM, penerapannya pun beragam, mulai dari rekrutmen sampai manajemen kinerja [4].

Sejumlah algoritma machine learning sudah pernah diterapkan pada data SDM, misalnya Random Forest, Support Vector Machine, Logistic Regression, Decision Tree, dan K-

Nearest Neighbors [6]. Namun, hasil penelitian perbandingan menunjukkan bahwa performa tiap algoritma tidak selalu sama karena bergantung pada karakteristik data dan tingkat kesulitan masalahnya [7].

Prediksi gaji karyawan dengan machine learning cukup banyak diteliti di Indonesia. Ramadhan dkk., misalnya, memakai analisis regresi linear multivariat dan menemukan bahwa usia dan lama bekerja di perusahaan memengaruhi pendapatan bulanan karyawan sebesar 90,9% [8]. Hasil ini memperlihatkan bahwa faktor demografis memang berkaitan erat dengan besaran gaji. Tuah dan Anyan juga menemukan bahwa gaji dan masa kerja berpengaruh positif dan signifikan terhadap kinerja karyawan [3]. Hal serupa dilaporkan Utomo dkk. yang mengkaji tingkat pendidikan, pengalaman usaha, dan umur, dengan hasil bahwa ketiganya berpengaruh signifikan terhadap penggunaan informasi akuntansi pada organisasi [9].

Di antara berbagai metode, regresi linear termasuk yang paling sering dipakai untuk prediksi gaji di Indonesia. Tuah dan Anyan menerapkan regresi linear sederhana untuk memprediksi gaji dari lama bekerja dan menemukan adanya hubungan yang positif dan signifikan antara variabel tersebut dengan kinerja karyawan [3]. Pada konteks lain, Aroginanto dkk. memakai regresi linear berganda dengan bantuan IBM SPSS dan berhasil mengidentifikasi faktor-faktor yang berpengaruh signifikan terhadap kualitas laporan keuangan UMKM [10].

Selain regresi linear, hubungan antarvariabel juga dapat bersifat tidak linear sehingga diperlukan metode yang mampu menangkap pola tersebut. Polynomial regression merupakan pengembangan dari regresi linear yang memodelkan hubungan antarvariabel sebagai polinomial derajat n , sehingga dapat menangani hubungan yang tidak linear. Agar model semacam ini dapat dipercaya, pemilihan metrik evaluasi menjadi penting. Shihab dan Napiah, misalnya, mengevaluasi model regresi mereka dengan root mean squared error (RMSE) dan memperoleh nilai error yang rendah, sehingga modelnya dinilai andal untuk memprediksi persentase perokok [11].

Algoritma K-Nearest Neighbors (KNN) juga sering dipakai pada penelitian prediksi di Indonesia. Tafonao dan Sibero menerapkannya untuk memprediksi kelulusan mahasiswa dan memperoleh akurasi tertinggi 95% pada $k = 3$, yang antara lain dipengaruhi oleh penanganan missing value pada indeks prestasi semester [12]. Nugroho dkk. membandingkan KNN dengan Support Vector Machine (SVM) untuk memprediksi kejang demam pada anak usia 6 bulan hingga 5 tahun, dengan hasil KNN mencapai akurasi 95%, sedangkan SVM 77% [13].

Decision Tree juga terbukti bekerja baik pada berbagai masalah prediksi. Wardhana dkk. melaporkan bahwa pemodelan Decision Tree mampu mengklasifikasikan tingkat kasus penyakit dengan akurasi yang sangat baik [14]. Shafa dkk. membandingkan Random Forest, Decision Tree, dan Naïve Bayes untuk prediksi kanker paru, dengan hasil Random Forest dan Decision Tree mencapai akurasi 100%, sedangkan Naïve Bayes 86% [15].

Gradient Boosting, termasuk varian populernya XGBoost (Extreme Gradient Boosting), banyak dipakai pada kompetisi maupun penelitian prediksi. Mardiah dkk. menggunakannya untuk memprediksi produksi beras di Sumatera dengan feature engineering dan data split 90:10, sehingga model memperoleh R^2 sebesar 0,976, RMSE 0,109, dan MAE 0,076 [16]. Hasil tersebut memperlihatkan bahwa feature engineering dan porsi data latih yang besar dapat meningkatkan akurasi prediksi [16].

Meski begitu, penerapan algoritma-algoritma tersebut tidak lepas dari kendala. Asbullah dan Samsudin mencatat bahwa keterbatasan data historis dan volatilitas pasar masih menjadi penghambat, sehingga penambahan variabel atau perbaikan prapengolahan data kerap diperlukan [17]. Tidak ada pula satu algoritma yang selalu unggul di semua kasus karena pemilihannya bergantung pada karakteristik dataset dan tingkat kesulitan masalah. Karena itu, evaluasi model sebaiknya memakai beberapa metrik sekaligus seperti MAE, MSE, RMSE, dan R^2 , bukan hanya satu [18].

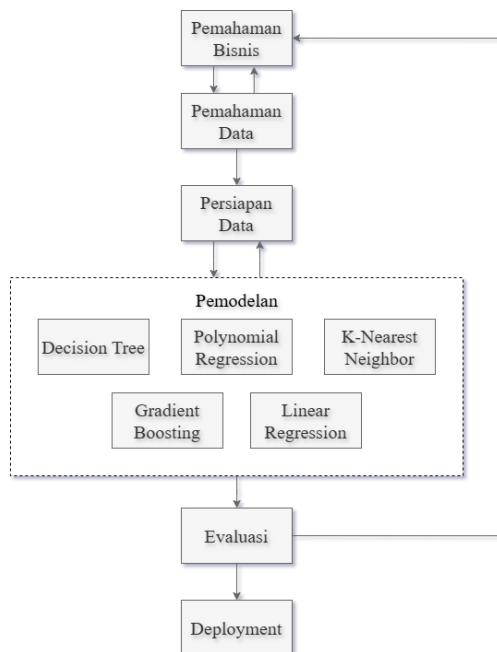
Berdasarkan hal tersebut, penelitian ini membangun sekaligus membandingkan beberapa metode machine learning untuk memprediksi gaji karyawan berdasarkan umur, pengalaman kerja, dan tingkat pendidikan, dengan tujuan menemukan algoritma yang paling akurat untuk kasus ini. Penelitian ini bertujuan untuk menganalisis hubungan antara faktor demografis dengan tingkat gaji karyawan serta membandingkan performa berbagai algoritma machine learning dalam prediksi gaji. Penelitian ini juga diharapkan dapat menjadi alat bantu pengambilan keputusan dalam penentuan kompensasi karyawan.

Hasil penelitian diharapkan dapat memberikan referensi bagi praktisi HR dalam menyusun strategi kompensasi yang lebih objektif serta menambah referensi penerapan machine learning di bidang SDM dalam konteks Indonesia.

2. BAHAN DAN METODE PENELITIAN

2.1. Kerangka Penelitian

Penelitian ini mengadopsi metodologi Cross Industry Standard Process for Data Mining (CRISP-DM) yang telah terbukti efektif dalam berbagai penelitian *machine learning* di Indonesia [19].



Gambar 1. Kerangka Penelitian

Seperti terlihat pada Gambar 1, CRISP-DM terdiri dari enam tahapan: pemahaman bisnis

(*business understanding*), pemahaman data (*data understanding*), persiapan data (*data preparation*), pemodelan (*modeling*), evaluasi (*evaluation*), dan *deployment* [20]. Metodologi ini dipilih karena tahapannya jelas dan sudah umum dipakai pada penelitian *machine learning* [21].

Pada penelitian ini, tahap pertama CRISP-DM dilakukan dengan mengidentifikasi permasalahan penentuan gaji yang dihadapi organisasi, yaitu bagaimana menentukan kompensasi yang adil berdasarkan faktor yang terukur. Dari sana ditetapkan tujuan penelitian serta variabel-variabel yang akan dipakai dalam pemodelan.

2.2 Pengumpulan Data

Dataset penelitian diambil dari situs Kaggle dan berisi 375 sampel data karyawan dengan 6 variabel: umur (Age), jenis kelamin (Gender), tingkat pendidikan (Education Level), jabatan (Job Title), pengalaman kerja (Years of Experience), dan gaji (Salary). Cuplikan isinya dapat dilihat pada Tabel 1.

Tabel 1. Dataset

Age	Gender	Education Level	Job Title	Years of Experience	Salary
26.0	Male	Bachelor's	Junior Accountant	1.0	35000.0
41.0	Male	Bachelor's	Senior Product Manager	14.0	150000.0
40.0	Female	Bachelor's	Office Manager	15.0	65000.0
42.0	Female	PhD	Senior Marketing Manager	18.0	140000.0

Dataset mencakup rentang umur 23–53 tahun dan pengalaman kerja 0–25 tahun. Audit data menemukan satu observasi Salary sebesar \$350 pada karyawan berusia 29 tahun dengan Job Title Junior Business Operations Analyst. Nilai tersebut merupakan nilai anomali terhadap distribusi gaji pada dataset dan menghasilkan distorsi sangat besar pada MAPE. Analisis sensitivitas menunjukkan MAPE Gradient Boosting berubah dari 46,15% menjadi 11,42% setelah observasi tersebut dikeluarkan, sehingga observasi \$350 dihapus sebelum pemodelan final.

2.3 Preprocessing Data

Sebelum pemodelan, dilakukan pembersihan data (*data cleaning*) dengan menghapus data kosong (*missing values*) dan data ganda (*duplikat*). Dari 375 data awal, 51 baris dihapus pada tahap ini sehingga tersisa 324 sampel. Setelah audit dan analisis sensitivitas terhadap observasi Salary = \$350, satu baris anomali tersebut dikeluarkan sehingga dataset final yang

digunakan pada eksperimen berjumlah 323 sampel.

Transformasi fitur kategorikal dilakukan sesuai jenis datanya. Gender dan Job Title yang bersifat nominal diubah menggunakan One-Hot Encoding dengan parameter *drop=None* dan *handle_unknown='ignore'*. Education Level diperlakukan sebagai variabel ordinal dengan pemetaan Bachelor's = 0, Master's = 1, dan PhD = 2. Encoder hanya di-fit pada data latih di setiap fold agar informasi kategori dari data validasi tidak masuk ke proses pelatihan.

Fitur masukan model terdiri atas Age, Gender, Education Level, Job Title, dan Years of Experience, sedangkan Salary menjadi target. Pada Polynomial Regression, ekspansi PolynomialFeatures derajat 4 hanya diterapkan pada Age dan Years of Experience karena keduanya merupakan variabel kontinu; Education Level tetap sebagai fitur ordinal dan fitur hasil One-Hot Encoding tidak dipolinomialkan.

Lima kondisi preprocessing dibandingkan, yaitu tanpa normalisasi, StandardScaler (Z-

Score), MinMaxScaler, Normalizer L1, dan Normalizer L2. Normalisasi hanya diterapkan pada fitur numerik, sedangkan kolom hasil One-Hot Encoding dipertahankan dalam bentuk biner. Pendekatan ini mencegah standarisasi kategori biner yang jarang muncul dan menjaga interpretasi representasi kategorikal.

Evaluasi dilakukan menggunakan 5-Fold Cross Validation dengan KFold($n_splits=5$, $shuffle=True$, $random_state=90$). Pada setiap fold, encoder dan scaler di-fit hanya menggunakan data latih lalu diterapkan pada data validasi. Dengan demikian, setiap sampel digunakan sebagai data validasi satu kali dan seluruh proses preprocessing berada di dalam fold untuk meminimalkan data leakage.

2.4 Algoritma Regresi

Algoritma regresi mempelajari hubungan antara variabel independen (fitur) dan variabel dependen (target) untuk memprediksi nilai kontinu. Lima algoritma yang dibandingkan pada penelitian ini adalah Regresi Linear, Regresi Polinomial, Decision Tree, K-Nearest Neighbors (KNN), dan Gradient Boosting.

Regresi Linear memodelkan hubungan linear antara fitur dan target menggunakan ordinary least squares. Polynomial Regression memperluas hubungan tersebut dengan PolynomialFeatures derajat 4 yang diterapkan hanya pada Age dan Years of Experience, kemudian koefisien diestimasi menggunakan sklearn LinearRegression berbasis least-squares solver. Pembatasan ekspansi pada dua fitur kontinu dilakukan untuk menghindari interaksi polinomial yang tidak bermakna pada fitur ordinal dan dummy hasil encoding.

Decision Tree membagi ruang fitur secara rekursif berdasarkan threshold yang meminimalkan error pada setiap node. K-Nearest Neighbors Regressor (KNN) menggunakan $k = 5$ dan memprediksi target dari tetangga terdekat berdasarkan jarak Euclidean. Gradient Boosting membangun ensemble secara bertahap dengan mengoreksi residual model sebelumnya; pada penelitian ini digunakan $n_estimators = 100$.

2.5 Strategi Validasi dan Evaluasi Model

Performa model dievaluasi pada setiap fold menggunakan Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), dan R-squared (R^2). MSE digunakan untuk mengukur rata-rata kuadrat kesalahan

prediksi, sedangkan RMSE menyatakan besar kesalahan prediksi dalam satuan yang sama dengan target dan memberikan penalti lebih besar terhadap error yang besar. MAE menunjukkan rata-rata kesalahan absolut antara nilai aktual dan hasil prediksi, sementara MAPE menyajikan kesalahan prediksi dalam bentuk persentase. R^2 digunakan untuk mengukur proporsi variasi target yang dapat dijelaskan oleh model. Untuk mengurangi ketergantungan hasil terhadap satu pembagian data, evaluasi dilakukan menggunakan 5-fold cross-validation. Seluruh metrik dihitung pada setiap fold dan kemudian dirangkum berdasarkan nilai rata-rata dari lima fold. Pada penyajian hasil, MSE, RMSE, MAE, dan MAPE ditampilkan dalam bentuk nilai rata-rata, sedangkan R^2 disajikan sebagai mean $\pm$ standar deviasi untuk menunjukkan variasi dan kestabilan performa model antar-fold. Standar deviasi pada metrik error lainnya tetap dihitung dan dibahas secara deskriptif pada bagian hasil dan pembahasan.

2.6 Implementasi dan Perangkat

Penelitian ini diimplementasikan dengan bahasa pemrograman Python, menggunakan pustaka scikit-learn untuk algoritma *machine learning*, pandas untuk pengolahan data, dan NumPy untuk operasi numerik. Python dipilih karena banyak dipakai pada penelitian *machine learning* di Indonesia dan pustakanya lengkap. Seluruh kode dijalankan pada *environment* yang sama dengan versi pustaka yang dicatat, sehingga hasil penelitian dapat direproduksi.

3. HASIL DAN PEMBAHASAN

Kelima algoritma dievaluasi pada 323 sampel menggunakan 5-Fold Cross Validation. Fitur yang digunakan meliputi umur, jenis kelamin, tingkat pendidikan, jabatan, dan pengalaman kerja. Setiap algoritma diuji pada kondisi tanpa normalisasi, Z-Score, Min-Max, L1, dan L2. Nilai MSE, RMSE, MAE, dan MAPE pada Tabel 2–6 disajikan sebagai nilai rata-rata dari lima fold, sedangkan R^2 disajikan dalam bentuk mean $\pm$ standar deviasi untuk menunjukkan variasi performa antar-fold.

3.1 Performa Model Tanpa Normalisasi

Pengujian pertama dilakukan pada data asli tanpa normalisasi untuk melihat performa dasar tiap algoritma sebelum diberi perlakuan apa pun. Hasil evaluasi kelima model pada kondisi ini disajikan pada Tabel 2.

Tabel 2. Tanpa Normalisasi

Metode	MSE ($\times 10^6$)	RMSE	MAE	MAPE	R^2 (mean $\pm$ std)
Decision Tree	462,47	21.090	13.454	13,65%	0,7961 $\pm$ 0,0744
Gradient Boosting	244,31	15.380	10.788	11,42%	0,8934 $\pm$ 0,0349

Metode	MSE ($\times 10^6$)	RMSE	MAE	MAPE	R ² (mean $\pm$ std)
Polynomial Regression	360,53	18.749	13.663	16,37%	0,8427 $\pm$ 0,0467
KNN	298,67	17.159	12.220	12,67%	0,8708 $\pm$ 0,0226
Regresi Linear	363,77	18.793	13.330	15,98%	0,8423 $\pm$ 0,0485

Tanpa normalisasi (Tabel 2), Gradient Boosting mencatatkan error paling kecil sekaligus R² tertinggi: MSE 244.306.425, RMSE 15.380, MAE 10.788, MAPE 11,42%, dan R² 0,8934. KNN mengikuti di posisi kedua dengan R² 0,8708, disusul Polynomial Regression (0,8427) dan Regresi Linear (0,8423). Adapun Decision Tree menjadi model terlemah pada skenario ini dengan R² 0,7961. Standar deviasi R² antar fold pada skenario ini tergolong kecil, berkisar dari 0,0226 (KNN) hingga 0,0744 (Decision Tree), yang menandakan hasil cukup konsisten.

Selain variasi R², kestabilan metrik error juga diamati dari hasil lima fold. Pada Gradient

Boosting sebagai model dengan mean R² tertinggi, standar deviasi MSE sebesar $89,91 \times 10^6$, RMSE sebesar 2.783, MAE sebesar 1.180, dan MAPE sebesar 1,42%. Nilai tersebut menunjukkan adanya variasi error antar-fold yang tetap perlu dipertimbangkan dalam mengevaluasi kemampuan generalisasi model.

3.2 Normalisasi dengan Z-Score

Selanjutnya, kelima model diuji pada data yang telah distandardisasi dengan Z-Score, yaitu dengan mengubah tiap fitur agar memiliki rata-rata 0 dan standar deviasi 1. Hasilnya dapat dilihat pada Tabel 3.

Tabel 3. Normalisasi Z-Score

Metode	MSE ($\times 10^6$)	RMSE	MAE	MAPE	R ² (mean $\pm$ std)
Decision Tree	463,25	21.107	13.486	13,71%	0,7957 $\pm$ 0,0748
Gradient Boosting	244,70	15.395	10.794	11,42%	0,8932 $\pm$ 0,0347
Polynomial Regression	360,53	18.749	13.663	16,37%	0,8427 $\pm$ 0,0467
KNN	270,20	16.139	11.284	12,13%	0,8824 $\pm$ 0,0418
Regresi Linear	363,77	18.793	13.330	15,98%	0,8423 $\pm$ 0,0485

Hasil dengan normalisasi Z-Score (Tabel 3) tidak banyak mengubah urutan model. Gradient Boosting tetap di posisi teratas dengan R² 0,8932 (std 0,0347), diikuti KNN yang justru terbantu oleh normalisasi ini: R²-nya naik dari 0,8708 menjadi 0,8824 (std 0,0418), karena algoritma berbasis jarak memang sensitif terhadap skala fitur. Menariknya, hasil Regresi Linear dan Polynomial Regression pada Z-Score praktis identik dengan kondisi tanpa normalisasi, karena solusi minimum-norm kedua model linear tersebut tidak dipengaruhi oleh penskalaan kolom.

Variasi hasil KNN antar-fold ditunjukkan oleh standar deviasi MSE sebesar $108,53 \times 10^6$,

RMSE sebesar 3.119, MAE sebesar 1.224, dan MAPE sebesar 1,32%. Nilai tersebut menunjukkan bahwa performa KNN pada Z-Score tetap memiliki variasi pada masing-masing fold meskipun menghasilkan nilai rata-rata yang lebih baik dibandingkan kondisi tanpa normalisasi.

3.3 Normalisasi dengan Min-Max

Pada skenario ketiga, fitur dinormalisasi dengan Min-Max sehingga seluruh nilainya terpetakan ke rentang [0, 1]. Hasil evaluasi kelima model dengan teknik ini ditampilkan pada Tabel 4.

Tabel 4. Normalisasi Min-Max

Metode	MSE ($\times 10^6$)	RMSE	MAE	MAPE	R ² (mean $\pm$ std)
Decision Tree	464,48	21.138	13.485	13,67%	0,7952 $\pm$ 0,0743
Gradient Boosting	244,31	15.380	10.788	11,42%	0,8934 $\pm$ 0,0349
Polynomial Regression	360,53	18.749	13.663	16,37%	0,8427 $\pm$ 0,0467
KNN	274,83	16.280	11.450	12,28%	0,8784 $\pm$ 0,0431
Regresi Linear	363,77	18.793	13.330	15,98%	0,8423 $\pm$ 0,0485

Pola yang sama terlihat pada normalisasi Min-Max (Tabel 4). Gradient Boosting kembali unggul dengan R^2 0,8934 (std 0,0349), disusul KNN dengan R^2 0,8784 (std 0,0431). Dengan kata lain, baik tanpa normalisasi maupun dengan Z-Score dan Min-Max, urutan performa model praktis tidak berubah; normalisasi per-kolom hanya sedikit memperbaiki konsistensi hasil pada beberapa metrik.

Pada normalisasi Min-Max, Gradient Boosting menghasilkan variasi metrik error yang hampir sama dengan kondisi tanpa normalisasi, yaitu

standar deviasi MSE sebesar $89,91 \times 10^6$, RMSE sebesar 2.783, MAE sebesar 1.180, dan MAPE sebesar 1,42%. Sementara itu, KNN memiliki standar deviasi MSE sebesar $102,48 \times 10^6$, RMSE 3.129, MAE 1.483, dan MAPE 1,62%.

3.4 Normalisasi dengan L2

Skenario berikutnya memakai Normalizer L2, yang mengubah tiap sampel menjadi vektor satuan berdasarkan norma L2-nya. Hasil pengujian pada kondisi ini disajikan pada Tabel 5.

Tabel 5. Normalisasi L2

Metode	MSE ($\times 10^6$)	RMSE	MAE	MAPE	R^2 (mean $\pm$ std)
Decision Tree	512,03	22.472	14.807	15,76%	0,7726 $\pm$ 0,0567
Gradient Boosting	310,49	17.380	11.557	12,19%	0,8645 $\pm$ 0,0408
Polynomial Regression	605,96	23.518	15.535	19,90%	0,7396 $\pm$ 0,1548
KNN	338,75	18.259	12.446	13,12%	0,8504 $\pm$ 0,0387
Regresi Linear	441,07	20.700	14.478	16,93%	0,8087 $\pm$ 0,0593

Normalisasi L2 memberikan gambaran yang agak berbeda (Tabel 5). Gradient Boosting memang masih terdepan dengan R^2 0,8645 (std 0,0408), tetapi performanya turun dibanding skenario sebelumnya, dan KNN menyusul dengan R^2 0,8504. Penurunan paling tajam terjadi pada Polynomial Regression: R^2 -nya anjlok ke 0,7396 dengan standar deviasi yang jauh lebih besar (0,1548), sehingga menjadi model terburuk pada skenario ini. Hal ini wajar karena normalisasi L2 menormalkan tiap sampel per-baris, sehingga mendistorsi fitur polinomial yang memiliki rentang skala sangat berbeda antar derajat.

Pada metrik error, Gradient Boosting memiliki standar deviasi MSE sebesar $103,31 \times 10^6$, RMSE sebesar 2.900, MAE sebesar 1.097, dan MAPE sebesar 1,13%. Sementara itu, Polynomial Regression memiliki standar deviasi MSE sebesar $402,96 \times 10^6$, RMSE 7.268, MAE 3.063, dan MAPE 5,79%, yang menunjukkan variasi error antar-fold lebih besar dibandingkan Gradient Boosting.

3.5 Normalisasi dengan L1

Terakhir, kelima model diuji dengan normalisasi L1 yang menormalkan tiap sampel berdasarkan jumlah nilai absolut fiturnya. Hasilnya dapat dilihat pada Tabel 6.

Tabel 6. Normalisasi L1

Metode	MSE ($\times 10^6$)	RMSE	MAE	MAPE	R^2 (mean $\pm$ std)
Decision Tree	451,76	20.990	13.219	13,39%	0,8031 $\pm$ 0,0550
Gradient Boosting	311,99	17.442	11.680	12,26%	0,8641 $\pm$ 0,0386
Polynomial Regression	625,53	23.926	15.672	20,37%	0,7307 $\pm$ 0,1576
KNN	324,20	17.885	12.343	13,11%	0,8569 $\pm$ 0,0340
Regresi Linear	450,26	20.915	14.499	16,69%	0,8044 $\pm$ 0,0610

Hasil yang menarik ditunjukkan Decision Tree pada normalisasi L1 (Tabel 6), dengan R^2 meningkat dari 0,7961 tanpa normalisasi menjadi 0,8031 $\pm$ 0,0550. Meskipun demikian, peningkatan tersebut relatif kecil dan bersifat spesifik pada dataset penelitian. Gradient Boosting tetap memperoleh mean R^2 tertinggi pada skenario L1 sebesar 0,8641 $\pm$ 0,0386.

Pada metrik error, Gradient Boosting memiliki standar deviasi MSE sebesar $100,51 \times 10^6$, RMSE sebesar 2.785, MAE sebesar 1.028, dan MAPE sebesar 0,99%. Polynomial Regression menunjukkan variasi yang lebih besar dengan standar deviasi MSE sebesar $409,90 \times 10^6$, RMSE 7.284, MAE 3.200, dan MAPE 6,15%.

3.6 Keunggulan Gradient Boosting dalam Prediksi Gaji

Berdasarkan lima skenario preprocessing, Gradient Boosting memperoleh mean R^2 tertinggi pada seluruh pengujian. Nilai terbaik dicapai pada kondisi tanpa normalisasi dan Min-Max dengan R^2 $0,8934 \pm 0,0349$ dan MAPE $11,42\% \pm 1,42\%$. Z-Score menghasilkan performa yang hampir identik (R^2 $0,8932 \pm 0,0347$), sedangkan L1 dan L2 menurunkan R^2 menjadi sekitar 0,864. Stabilitas pada tanpa normalisasi, Z-Score, dan Min-Max konsisten dengan karakteristik model berbasis pohon yang tidak bergantung pada skala absolut fitur untuk menentukan split. Dengan demikian, kondisi tanpa normalisasi dipilih sebagai konfigurasi final karena memberikan performa tertinggi tanpa menambah tahap penskalaan yang tidak diperlukan.

3.7 Pengaruh Normalisasi terhadap Stabilitas Model

Pengaruh normalisasi paling jelas terlihat pada KNN. Jarak Euclidean $d(x_i, x_j) = \sqrt{\sum (x_{ik} - x_{jk})^2}$ membuat fitur numerik dengan skala berbeda dapat memberikan kontribusi jarak yang tidak seimbang. Z-Score dan Min-Max bekerja secara feature-wise sehingga mengurangi ketidakseimbangan skala; pada penelitian ini Z-Score memberikan hasil KNN terbaik (R^2 $0,8824 \pm 0,0418$). Sebaliknya, L1 dan L2 melakukan normalisasi per-sampel, $x'_i = x_i / ||x_i||$, sehingga magnitude absolut tiap observasi dihilangkan dan hubungan jarak antarsampel berubah. Pada Decision Tree, Z-Score dan Min-Max hampir tidak mengubah R^2 karena keduanya merupakan transformasi monotonik per-fitur. L1 dapat sedikit mengubah performa karena pembagi setiap baris bergantung pada seluruh fitur, sehingga urutan nilai dan threshold split dapat berubah; efek peningkatan yang terlihat pada penelitian ini tidak dapat digeneralisasi sebagai keunggulan L1 untuk Decision Tree.

3.8 Keterbatasan dan Tantangan Model

Penelitian ini memiliki beberapa keterbatasan. Dataset final hanya berjumlah 323 sampel dan Job Title memiliki kardinalitas tinggi, sehingga pada setiap fold terdapat sekitar 19–29 jabatan pada data validasi yang tidak muncul pada data latih. Kondisi ini membatasi kemampuan model untuk mempelajari pola pada jabatan yang sangat jarang dan dapat menurunkan generalisasi pada kategori pekerjaan baru. Selain itu, model belum memasukkan faktor

seperti lokasi geografis, jenis industri, ukuran perusahaan, kondisi ekonomi, maupun inflasi, dan belum divalidasi pada dataset eksternal. Karena itu, hasil penelitian lebih tepat dipandang sebagai proof-of-concept dan model pendukung estimasi, bukan dasar tunggal penetapan gaji.

3.9 Pengembangan Antarmuka Pengguna Web

Aplikasi prediksi gaji karyawan dibangun menggunakan *framework* Streamlit.

Prediction Model 🏠 🔥

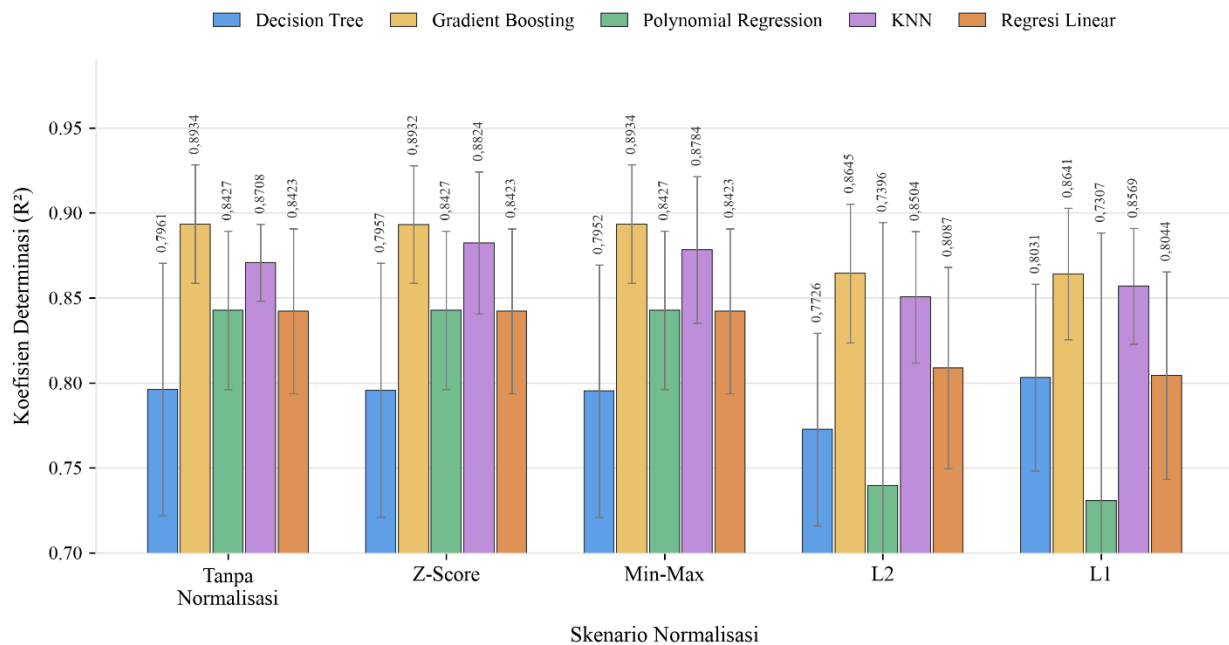
Gambar 2. Halaman Prediksi

Gambar 2 memperlihatkan halaman prediksi sebagai fungsi utama aplikasi. Pengguna cukup mengisi umur, pengalaman kerja, tingkat pendidikan, dan jabatan pada formulir yang tersedia; variabel kategorikal otomatis di-*encode* oleh sistem, lalu diteruskan ke model Gradient Boosting yang sudah dilatih untuk menghasilkan estimasi gaji.

Gambar 3. Hasil Prediksi

Gambar 3 menampilkan hasil akhir prediksi dari contoh masukan pada Gambar 2, yaitu estimasi gaji sebesar \$54,710.

Sebagai penutup pembahasan, Gambar 4 merangkum nilai R^2 kelima model pada seluruh skenario normalisasi yang telah diuji.



Gambar 4. Perbandingan nilai R^2 kelima model pada tiap skenario normalisasi

Gambar 4 menunjukkan bahwa Gradient Boosting memiliki mean R^2 tertinggi pada seluruh skenario, dengan nilai maksimum 0,8934 pada kondisi tanpa normalisasi dan Min-Max. KNN menjadi pesaing terdekat dan memperoleh hasil terbaik pada Z-Score (R^2 0,8824). Regresi Linear dan Polynomial Regression stabil pada tanpa normalisasi, Z-Score, dan Min-Max dengan R^2 sekitar 0,842, sedangkan Decision Tree berada pada kisaran 0,80. Normalizer L1 dan L2 cenderung menurunkan performa, terutama pada Polynomial Regression. Berdasarkan kombinasi mean R^2 tertinggi, MAPE terendah, serta kebutuhan preprocessing yang sederhana, Gradient Boosting tanpa normalisasi dipilih sebagai model final untuk implementasi Streamlit.

4. KESIMPULAN

Penelitian ini membandingkan lima metode machine learning untuk memprediksi gaji berdasarkan karakteristik karyawan menggunakan 5-Fold Cross Validation. Gradient Boosting memperoleh mean R^2 tertinggi sebesar $0,8934 \pm 0,0349$ dengan MAPE $11,42\% \pm 1,42\%$ pada kondisi tanpa normalisasi. KNN dengan Z-Score menjadi model terdekat dengan R^2 $0,8824 \pm 0,0418$. Hasil menunjukkan bahwa feature-wise scaling terutama bermanfaat bagi KNN, sementara Gradient Boosting relatif stabil pada kondisi tanpa normalisasi, Z-Score, dan Min-Max.

Dengan mempertimbangkan ukuran dataset dan tingginya kardinalitas Job Title, model yang dihasilkan lebih tepat digunakan sebagai alat bantu estimasi dan eksplorasi struktur kompensasi, bukan sebagai dasar tunggal penentuan gaji. Penggunaan One-Hot Encoding untuk fitur nominal, encoding ordinal untuk

pendidikan, penghapusan observasi anomal, serta validasi silang menghasilkan evaluasi yang lebih representatif terhadap kemampuan generalisasi model.

Penelitian selanjutnya dapat menggunakan dataset yang lebih besar dan lebih beragam, mengelompokkan kategori Job Title yang sangat jarang secara terukur, serta menambahkan variabel seperti lokasi, industri, ukuran perusahaan, inflasi, dan kondisi ekonomi. External validation pada dataset independen dan optimasi hyperparameter yang dilakukan di dalam kerangka validasi yang tepat juga diperlukan untuk menilai ketahanan model sebelum digunakan pada konteks operasional yang lebih luas.

DAFTAR PUSTAKA

- [1] Taryono, S. Efendi, and Hasanudin, "Strategi inovasi dalam manajemen sumber daya manusia untuk meningkatkan kinerja organisasi," *Co-Creation: Jurnal Ilmiah Ekonomi Manajemen Akuntansi dan Bisnis*, vol. 3, no. 1, pp. 48–55, Jun. 2024, doi: 10.55904/cocreation.v3i1.1278.
- [2] Nurfitriani, U. Kulsum, Sunarto, and Heriyanto, "Tinjauan mendalam tentang pengelolaan sumber daya manusia: Tantangan, strategi dan inovasi di PT Tirta Mahakam Resource Tbk," *Jurnal Ilmiah M-Progress*, vol. 14, no. 2, pp. 191–199, Jun. 2024, doi: 10.35968/mp.v14i2.1218.
- [3] Y. A. E. Tuah and A. Anyan, "Implementasi model regresi linear sederhana untuk prediksi gaji berdasarkan pengalaman lama bekerja," *JUTECH: Journal Education and Technology*, vol. 1, no. 2, pp. 56–70, Dec. 2020, doi: 10.31932/jutech.v1i2.1289.
- [4] R. Rathore and S. P. S. Rathore, "Machine

- learning applications in human resource management: Predicting employee turnover and performance,” *International Journal for Global Academic & Scientific Research*, vol. 3, no. 2, pp. 48–59, Jul. 2024, doi: 10.55938/ijgasr.v3i2.77.
- [5] J. Cui and Y. Gu, “Application of machine learning in digital human resource management,” *SHS Web of Conferences*, vol. 170, Art. no. 01002, 2023, doi: 10.1051/shsconf/202317001002.
- [6] N. Afiatuddin, M. T. Wicaksono, V. R. Akbar, R. Rahmadden, and D. Wulandari, “Komparasi algoritma machine learning dalam klasifikasi kanker payudara,” *Jurnal Media Informatika Budidarma*, vol. 8, no. 2, pp. 889–899, Apr. 2024, doi: 10.30865/mib.v8i2.7457.
- [7] R. Pramudita, S. Muis, N. Safitri, and F. Shafirawati, “Optimasi algoritma machine learning menggunakan teknik bagging pada klasifikasi diagnosis kanker payudara,” *TEMATIK*, vol. 11, no. 1, pp. 128–134, Jun. 2024, doi: 10.38204/tematik.v11i1.1928.
- [8] B. Ramadhan, D. Firdaus, and N. T. R. Adiningrum, “Analisis data pegawai untuk memprediksi gaji berdasarkan faktor-faktor spesifik dengan pendekatan machine learning,” *Naratif: Jurnal Nasional Riset, Aplikasi dan Teknik Informatika*, vol. 5, no. 2, pp. 131–139, Dec. 2023, doi: 10.53580/naratif.v5i2.205.
- [9] M. P. Utomo, C. L. Sembiring, and D. Astriani, “Pengaruh tingkat pendidikan, pengetahuan akuntansi, umur usaha, pengalaman usaha terhadap penggunaan informasi akuntansi,” *Jurnal Pendidikan Ekonomi (JURKAMI)*, vol. 8, no. 3, pp. 627–638, Dec. 2023, doi: 10.31932/jpe.v8i3.2817.
- [10] Aroginanto, M. F. Triatmaja, D. E. Purnomo, and Yohani, “Pengaruh tingkat pendidikan, pemahaman akuntansi, ukuran usaha dan lama usaha terhadap kualitas laporan keuangan UMKM batik binaan Dinkop UMKM dan Naker Kabupaten Pekalongan,” *Neraca*, vol. 19, no. 2, pp. 18–28, Dec. 2023, doi: 10.48144/neraca.v19i2.1683.
- [11] M. A. R. Shihab and M. Napiah, “Linear regression analysis to predict the percentage of smoking in the population age 15 years and over,” *Informatics and Software Engineering*, vol. 2, no. 2, pp. 76–88, Nov. 2024, doi: 10.58777/ise.v2i2.328.
- [12] I. P. J. Tafonao and A. F. K. Sibero, “Teknik klasifikasi prediksi kelulusan mahasiswa Sistem Informasi Universitas Sari Mutiara Indonesia menggunakan K-Nearest Neighbors,” *Jurnal Mahajana Informasi*, vol. 7, no. 1, pp. 83–90, 2022, doi: 10.51544/jurnalmi.v7i1.2956.
- [13] A. Nugroho, D. Rosmani, and I. N. Abdullah, “Analisa perbandingan algoritma K-Nearest Neighbors dan Support Vector Machine untuk prediksi kejang demam pada anak usia 6 bulan hingga 5 tahun,” *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 4, no. 12, pp. 639–655, Dec. 2024, doi: 10.52436/1.jpti.506.
- [14] R. G. Wardhana, G. Wang, and F. Sibuea, “Penerapan machine learning dalam prediksi tingkat kasus penyakit di Indonesia,” *Journal of Information System Management (JOISM)*, vol. 5, no. 1, pp. 40–45, Jul. 2023, doi: 10.24076/joism.2023v5i1.1136.
- [15] B. Shafa, H. H. Handayani, S. A. P. Lestari, and Y. Cahyana, “Prediksi kanker paru dengan normalisasi menggunakan perbandingan algoritma Random Forest, Decision Tree dan Naïve Bayes,” *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 4, no. 3, pp. 1057–1070, Nov. 2024, doi: 10.51454/decode.v4i3.779.
- [16] R. S. Mardiah, A. Suhendra, F. Fitrianiingsih, and D. K. Putri, “Advancing prediction of rice production in Sumatra: Utilizing eXtreme gradient boosting algorithm for accurate forecasts,” in *Proc. 2024 Ninth International Conference on Informatics and Computing (ICIC)*, Oct. 2024, pp. 1–6, doi: 10.1109/ICIC64337.2024.10957232.
- [17] J. Asbullah and S. Samsudin, “Prediksi harga cryptocurrency Binance berdasarkan informasi blockchain dengan menggunakan algoritma Random Forest,” *Jurnal Media Informatika Budidarma*, vol. 8, no. 1, pp. 260–271, Jan. 2024, doi: 10.30865/mib.v8i1.7100.
- [18] A. Syahreza, N. K. Ningrum, and M. A. Syahrazy, “Perbandingan kinerja model prediksi cuaca: Random Forest, Support Vector Regression, dan XGBoost,” *Edumatic: Jurnal Pendidikan Informatika*, vol. 8, no. 2, pp. 526–534, Dec. 2024, doi: 10.29408/edumatic.v8i2.27640.
- [19] E. W. Murti, “Analisis dan perbandingan algoritma prediksi dalam mengetahui perkiraan peningkatan jumlah kasus COVID-19 di Kabupaten Boyolali dengan metodologi CRISP-DM,” *JIKES: Jurnal Ilmu Kesehatan*, vol. 3, no. 1, pp. 24–34, Jul. 2024, doi: 10.71456/jik.v3i1.933.
- [20] N. N. Hakim, “Implementasi machine learning pada sistem prediksi kejadian dan lokasi patah rel kereta api di Indonesia,” *Jurnal Sistem Cerdas*, vol. 3, no. 1, pp. 25–35, May 2020, doi: 10.37396/jsc.v3i1.58.
- [21] M. F. Narendra, “Forecasting manpower planning using the CRISP-DM method and machine learning algorithm: A case study of Tiki Jalur Nugraha Ekakurir (JNE) company,” *Journal of Information Systems Engineering and Management*, vol. 10, no. 19s, pp. 371–378, Mar. 2025, doi: 10.52783/jisem.v10i19s.3040.