JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1177-1184

Analisis dan Prediksi Penghasilan Profesi IT Menggunakan Linear Regression dengan Library Python

Alfreza Routya Faizan[1], Arijal Pratama[2], Dila Kartika Putri[3], Fadhil Nata Pratama[4] , Reza Alifia Pratama [5], Verrel Aulia Rahman [6], Rahmawati [7]

[1234567]Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1]alfrezarf@gmail.com, [2]arijalpratama007@gmail.com, [3]dilakartikaputri1004@gmail.com, [4]fadhilpratama798@gmail.com, [5]rezaalifia1@gmail.com, [6]verrelaulia@gmail.com, [7]dosen02394@unpam.ac.id

Abstrak−Perkembangan teknologi informasi, khususnya pada bidang *Artificial Intelligence (AI)*, *Machine Learning (ML)*, dan *Data Science*, telah meningkatkan kebutuhan tenaga kerja profesional yang memiliki kompetensi tinggi. Informasi mengenai tingkat penghasilan pada berbagai profesi IT menjadi penting untuk memahami tren kompensasi serta membantu individu dalam menentukan perencanaan karier. Penelitian ini bertujuan untuk menganalisis distribusi gaji berdasarkan kategori profesi IT serta melakukan prediksi penghasilan menggunakan algoritma *Linear Regression* berbasis *Python*. Dataset yang digunakan adalah *AI Jobs Salary Dataset* yang diperoleh dari repositori GitHub foorilla/ai-jobs-net-salaries. Tahapan penelitian meliputi pengumpulan data, preprocessing, exploratory *data analysis* (EDA), encoding data, pelatihan model, evaluasi model, dan prediksi. Hasil penelitian menunjukkan bahwa kategori *Engineer* memiliki rata-rata gaji tertinggi sebesar USD 174.215 atau 18,85% dari total rata-rata gaji seluruh kategori profesi yang dianalisis. Hasil tersebut menunjukkan bahwa profesi yang berfokus pada pengembangan sistem dan teknologi tingkat lanjut cenderung memperoleh kompensasi yang lebih tinggi dibandingkan kategori profesi lainnya.

Kata Kunci: Profesi IT; *Linear Regression*; Analisis Gaji; *Python; Data Science*

Abstract−*The rapid growth of information technology, especially in Artificial Intelligence (AI), Machine Learning (ML), and Data Science, has increased the demand for highly skilled professionals. Salary information across IT professions is essential for understanding compensation trends and supporting career planning. This study aims to analyze salary distribution among IT professions and predict income using a Python-based Linear Regression algorithm. The dataset used in this study is the AI Jobs Salary Dataset obtained from the GitHub repository https://github.com/foorilla/ai-jobs-net-salaries. The research stages include data collection, preprocessing, exploratory data analysis (EDA), data encoding, model training, model evaluation, and prediction. The results indicate that the Engineer category has the highest average salary at USD 174,215 or 18.85% of the total average salary among all analyzed profession categories. These findings suggest that professions focused on advanced system development and emerging technologies tend to receive higher compensation than other profession categories.*

***Keywords**: IT Profession; Linear Regression; Salary Analysis; Python; Data Science*

1. PENDAHULUAN

Perkembangan teknologi informasi telah menciptakan berbagai peluang kerja baru yang berkaitan dengan *Artificial Intelligence (AI)*, *Machine Learning (ML)*, *Data Science*, dan *Software Engineering* (Géron, 2022). Meningkatnya transformasi digital di berbagai sektor industri menyebabkan kebutuhan terhadap tenaga kerja profesional di bidang teknologi terus mengalami peningkatan.

Selain kompetensi teknis, tingkat penghasilan menjadi salah satu faktor yang dipertimbangkan dalam menentukan pilihan karier. Informasi mengenai distribusi gaji pada berbagai profesi IT dapat memberikan gambaran mengenai kondisi pasar kerja dan membantu individu dalam merencanakan pengembangan kompetensi yang sesuai (VanderPlas, 2016).

Kemajuan teknologi pengolahan data memungkinkan dilakukan analisis terhadap berbagai faktor yang memengaruhi tingkat penghasilan. Salah satu metode yang umum digunakan adalah *Exploratory Data Analysis (EDA)* untuk memahami karakteristik data serta *Linear Regression* untuk memodelkan hubungan antar variabel dan melakukan prediksi nilai kontinu (James et al., 2023).

Penelitian ini menggunakan *AI Jobs Salary Dataset* yang diperoleh dari repositori GitHub foorilla/ai-jobs-net-salaries (Foorilla, 2026). Dataset tersebut berisi informasi mengenai profesi,

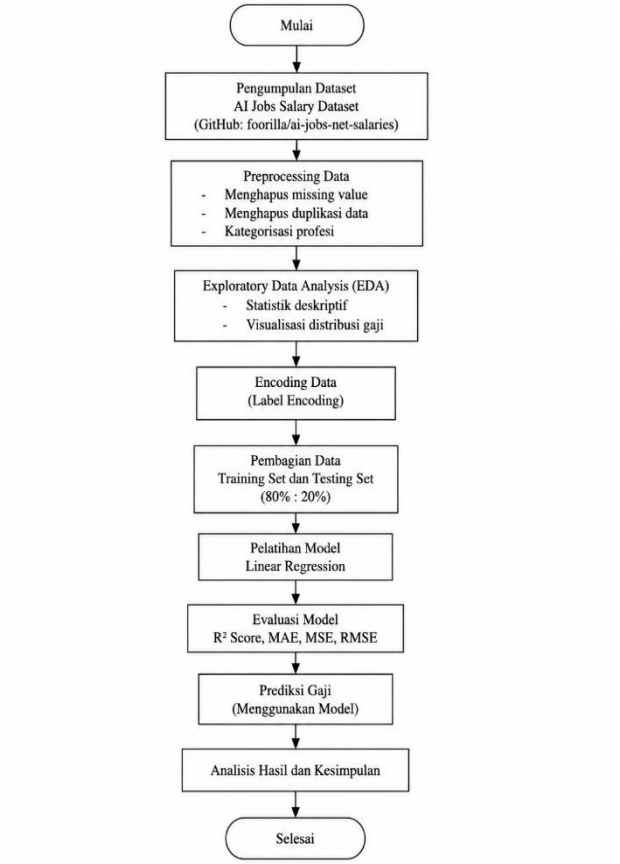
JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1177-1184

tingkat pengalaman, lokasi perusahaan, ukuran perusahaan, dan penghasilan dalam satuan USD. Melalui analisis ini diharapkan dapat diperoleh gambaran mengenai kategori profesi dengan tingkat penghasilan tertinggi serta pola distribusi gaji pada sektor teknologi informasi.

2. METODE

2.1 Metode Pengumpulan Data

Data yang digunakan berasal dari *AI Jobs Salary Dataset* yang tersedia secara publik pada repositori GitHub foorilla/ai-jobs-net-salaries (Foorilla, 2026). Dataset ini berisi informasi mengenai berbagai profesi pada bidang *AI, Machine Learning, Data Science*, dan teknologi informasi lainnya.



Gambar 1. Flowchart Tahapan Penelitian.

Gambar di atas menunjukkan tahapan penelitian yang dilakukan dalam menganalisis dan memprediksi penghasilan profesi IT menggunakan metode *Linear Regression*. Penelitian diawali dengan pengumpulan data dari *AI Jobs Salary Dataset* yang diperoleh melalui repositori GitHub. Selanjutnya dilakukan tahap *preprocessing* data yang meliputi pembersihan data, penghapusan nilai kosong, penghapusan data duplikat, dan pengelompokan kategori profesi. Setelah itu dilakukan *Exploratory Data Analysis (EDA)* untuk memahami karakteristik data dan melihat distribusi penghasilan pada setiap kategori profesi.

Data yang masih berbentuk kategorikal kemudian diubah ke bentuk numerik melalui proses encoding. Selanjutnya data dibagi menjadi data *training* dan *testing* untuk proses pelatihan model *Linear Regression*. Model yang telah dilatih dievaluasi menggunakan metrik R^2 Score, *Mean Absolute Error (MAE)*, Mean *Squared Error (MSE)*, dan *Root Mean Squared Error (RMSE)*. Tahap akhir penelitian adalah melakukan prediksi gaji berdasarkan model yang telah dibangun serta menyusun kesimpulan berdasarkan hasil analisis yang diperoleh.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1177-1184

2.2 Prepocessing Data

Tahapan preprocessing dilakukan untuk meningkatkan kualitas data sebelum dianalisis. Langkah-langkah yang dilakukan meliputi:

a) Menghapus data yang memiliki nilai kosong (*missing values*)
b) Menghapus data duplikat
c) Menyesuaikan format data numerik
d) Mengelompokkan jabatan ke dalam kategori profesi
e) Melakukan encoding terhadap variabel kategorikal

2.3 Linear Regression

Linear Regression digunakan untuk memodelkan hubungan antara variabel independen dan variabel dependen berupa penghasilan.

Persamaan Linear Regression:

$$Y = a + bX$$

Keterangan :
Y = nilai gaji
a = konstanta
b = koefisien regresi
X = variabel independen

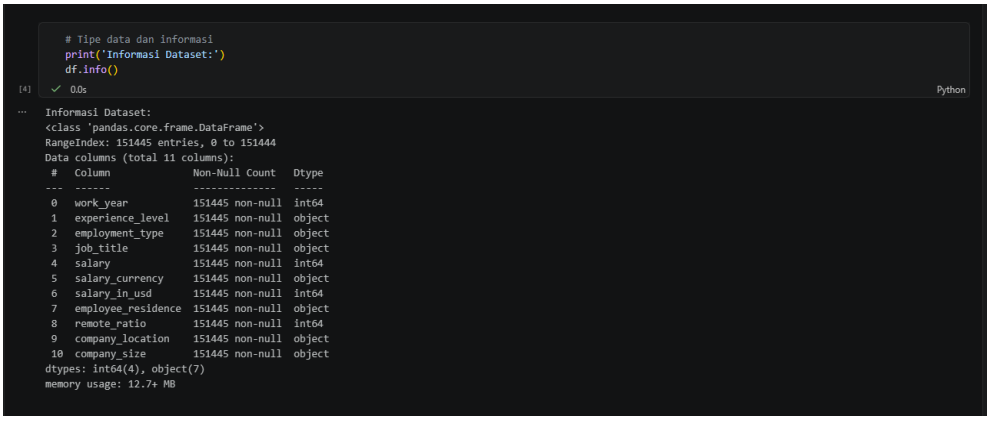
2.4 Evaluasi Model

Kinerja model dievaluasi menggunakan metrik *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan koefisien determinasi (R^2 Score) untuk mengetahui tingkat akurasi prediksi model.

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Dataset

Dataset yang digunakan terdiri dari berbagai profesi di bidang teknologi informasi dengan informasi terkait pengalaman kerja, lokasi perusahaan, ukuran perusahaan, dan penghasilan tahunan. Setelah dilakukan proses preprocessing, dataset siap digunakan untuk tahap analisis dan pelatihan model. Contoh baris data mentah yang digunakan dalam penelitian ini ditunjukkan pada Gambar dibawah ini.



```
# Tipe data dan informasi
print('Informasi Dataset:')
df.info()
```

```
[4]  ✓ 0.0s                                                    Python

Informasi Dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 151445 entries, 0 to 151444
Data columns (total 11 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   work_year           151445 non-null  int64
 1   experience_level    151445 non-null  object
 2   employment_type     151445 non-null  object
 3   job_title           151445 non-null  object
 4   salary              151445 non-null  int64
 5   salary_currency     151445 non-null  object
 6   salary_in_usd       151445 non-null  int64
 7   employee_residence  151445 non-null  object
 8   remote_ratio        151445 non-null  int64
 9   company_location    151445 non-null  object
 10  company_size        151445 non-null  object
dtypes: int64(4), object(7)
memory usage: 12.7+ MB
```

Gambar 2. Atribut dan Sampel Baris Data pada *AI Jobs Salary Dataset*

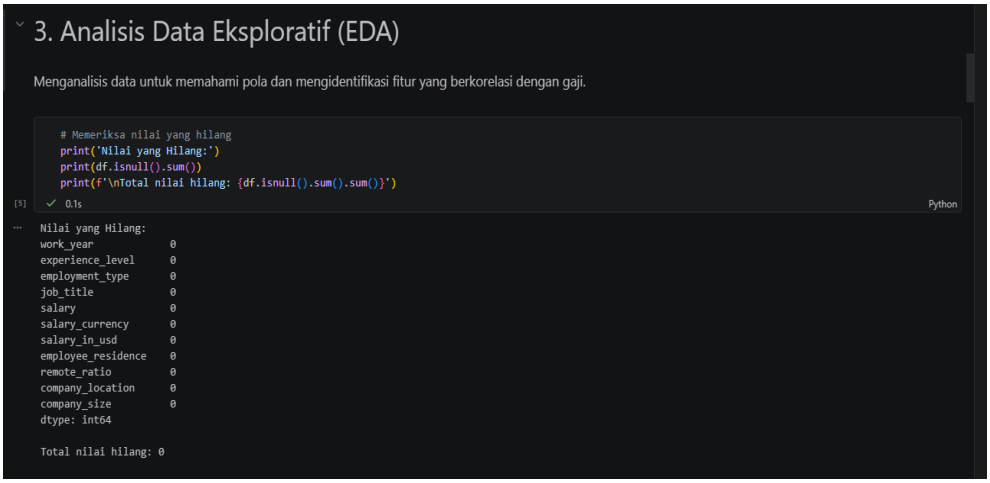
Berdasarkan Gambar diatas, setiap baris data merepresentasikan profil kompensasi riil dari pekerja teknologi di pasar global. Atribut utama yang menjadi prediktor dalam penelitian ini meliputi tahun komparasi data (*work_year*), tingkat kematangan profesi (*experience_level* seperti EN, MI, SE, EX), jenis ikatan kerja (*employment_type*), penamaan spesifik rumpun jabatan

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1177-1184

(*job_title*), nilai upah nominal beserta mata uangnya (*salary dan salary_currency*), nominal gaji baku yang dikonversi ke mata uang global (*salary_in_usd*), domisili pekerja (*employee_residence*), persentase intensitas kerja jarak jauh (*remote_ratio*), basis lokasi operasional perusahaan (*company_location*), serta skala kapasitas internal korporasi (*company_size*).

3.2 Hasil Exploratory Data Analysis

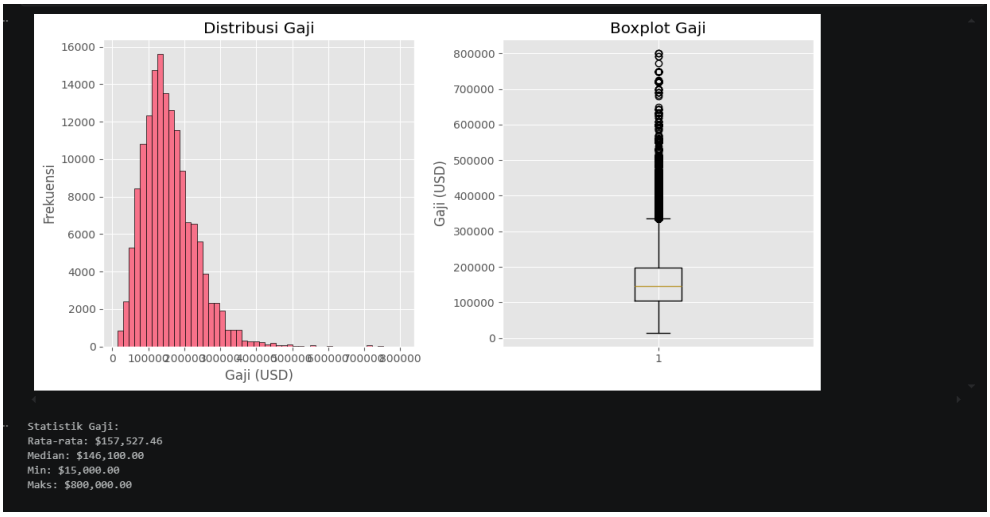
A. Pemeriksaan Missing Values



Gambar 3. Pemeriksaan *Missing Value*

Pemeriksaan *missing values* dilakukan untuk mengetahui kelengkapan data sebelum dilakukan proses analisis lebih lanjut. Berdasarkan hasil pemeriksaan, seluruh atribut memiliki jumlah nilai kosong sebesar 0 sehingga dataset dinyatakan lengkap dan tidak memerlukan proses imputasi data. Kondisi ini menunjukkan bahwa kualitas data sudah cukup baik untuk digunakan dalam proses pemodelan.

B. Ringkasan Statistik dan Distribusi Gaji

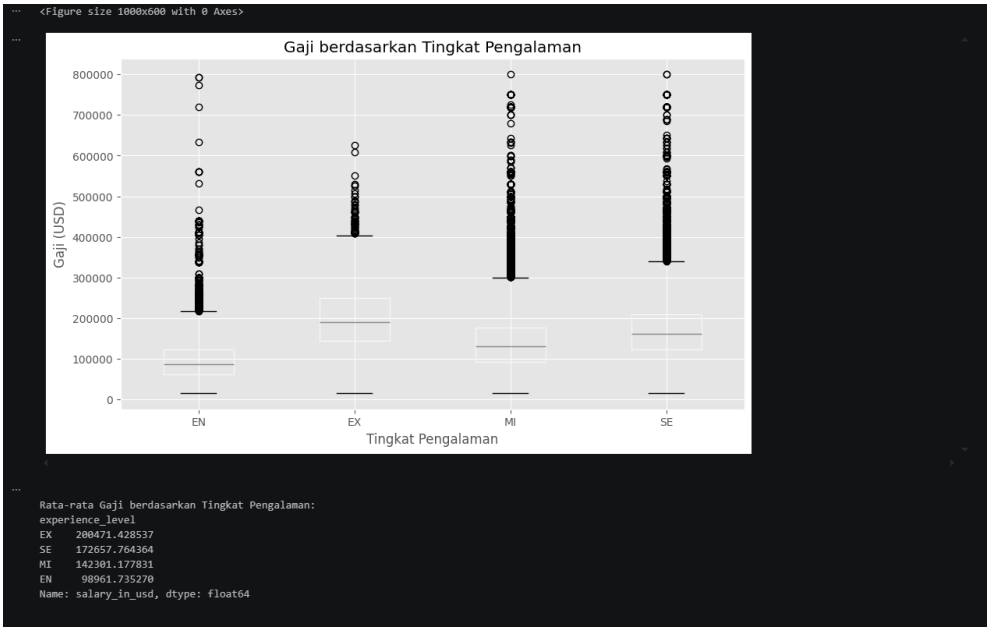


Gambar 4. Distribusi dan *Boxplot* Gaji

Hasil analisis statistik menunjukkan bahwa rata-rata gaji profesional IT sebesar USD 157.527 per tahun dengan nilai median sebesar USD 146.100. Nilai maksimum gaji mencapai USD 800.000 sedangkan nilai minimum sebesar USD 15.000. Histogram menunjukkan bahwa sebagian besar data berada pada rentang gaji menengah, sementara boxplot memperlihatkan adanya sejumlah outlier dengan nilai gaji yang jauh lebih tinggi dibandingkan mayoritas data.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1177-1184

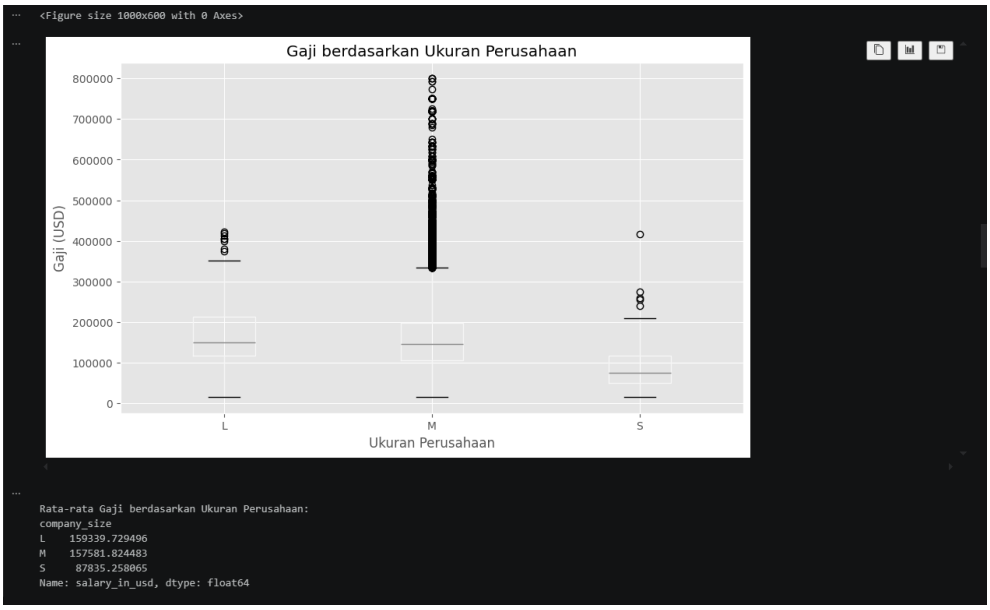
C. Pengaruh Tingkat Pengalaman terhadap Gaji



Gambar 5. Gaji Berdasarkan Tingkat Pengalaman

Tingkat pengalaman memiliki hubungan yang kuat terhadap besarnya penghasilan. Berdasarkan hasil analisis, kategori *Executive* (EX) memiliki rata-rata gaji tertinggi sebesar USD 200.471, diikuti *Senior* (SE) sebesar USD 172.658, *Mid-Level* (MI) sebesar USD 142.301, dan *Entry-Level* (EN) sebesar USD 98.962. Hasil ini menunjukkan bahwa peningkatan pengalaman kerja berkontribusi terhadap peningkatan kompensasi yang diterima pekerja di bidang teknologi informasi.

D. Pengaruh Ukuran Perusahaan terhadap Gaji



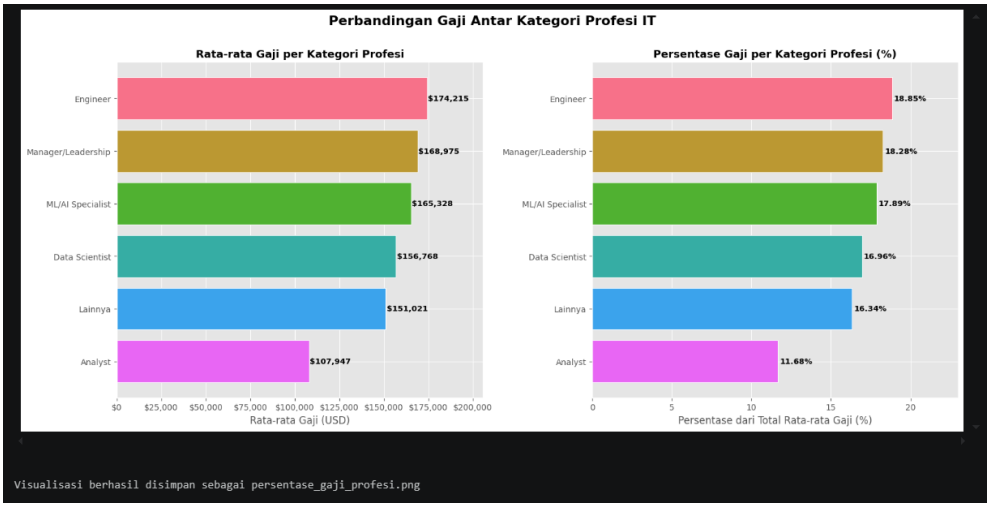
Gambar 6. Gaji Berdasarkan Ukuran Perusahaan

Hasil analisis menunjukkan bahwa ukuran perusahaan turut memengaruhi tingkat penghasilan. Perusahaan besar memiliki rata-rata gaji sebesar USD 159.340, perusahaan menengah

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1177-1184

sebesar USD 157.581, sedangkan perusahaan kecil sebesar USD 87.835. Temuan ini mengindikasikan bahwa perusahaan dengan skala lebih besar cenderung menawarkan kompensasi yang lebih tinggi**.**

E. Perbandingan Gaji Antar Kategori Profesi



Gambar 7. Perbandingan Gaji Antar Kategori Profesi IT

Untuk mempermudah analisis, berbagai jabatan dikelompokkan ke dalam enam kategori profesi utama yaitu *Engineer*, *Manager/Leadership*, *ML/AI Specialist*, *Data Scientist*, *Analyst*, dan Lainnya. Hasil analisis menunjukkan bahwa kategori *Engineer* memiliki rata-rata gaji tertinggi sebesar USD 174.215 atau 18,85% dari total rata-rata gaji seluruh kategori profesi. Posisi berikutnya ditempati oleh *Manager/Leadership* sebesar 18,28%, *ML/AI Specialist* sebesar 17,89%, *Data Scientist* sebesar 16,96%, Lainnya sebesar 16,34%, dan *Analyst* sebesar 11,68%.

3.3 Pelatihan Model Linear Regression

Sebelum proses pelatihan model dilakukan, data kategorikal terlebih dahulu diubah ke bentuk numerik menggunakan teknik encoding. Selanjutnya dataset dibagi menjadi data latih sebesar 80% dan data uji sebesar 20%. Model Linear Regression kemudian dilatih menggunakan 121.156 data latih dan diuji menggunakan 30.289 data uji untuk memprediksi nilai salary_in_usd.

3.4 Evaluasi Model

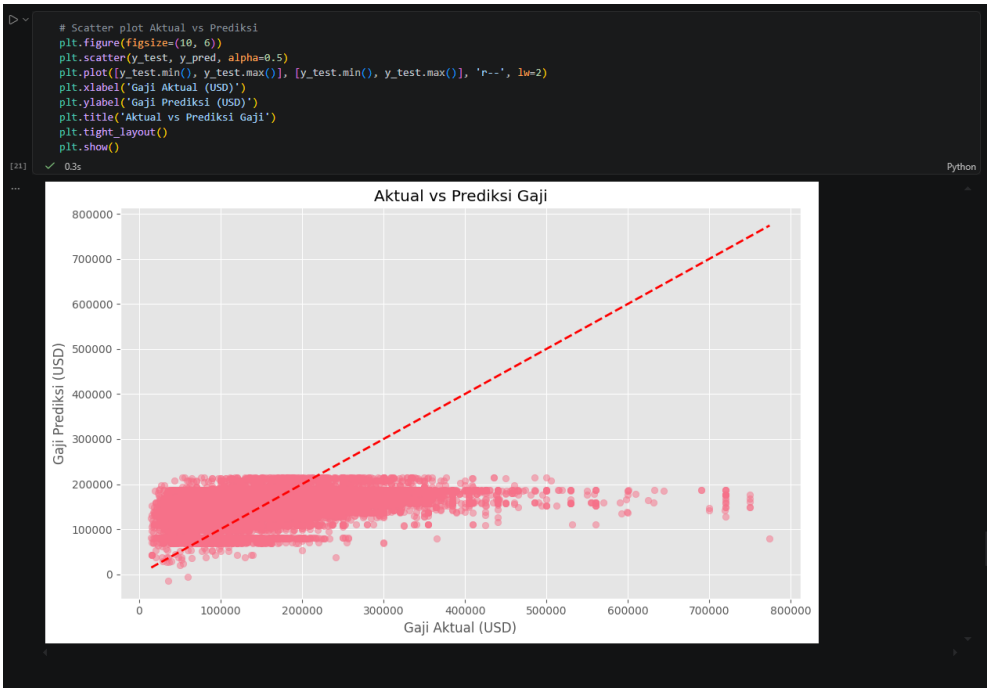
Tabel 1. Hasil Evaluasi Model

Metrik	Nilai
R^2 Score	0,1617
MAE	50.429,95
MSE	4.656.419.436,53
RMSE	68.237,96

Berdasarkan hasil evaluasi, model memperoleh nilai R^2 sebesar 0,1617 yang menunjukkan bahwa model mampu menjelaskan sekitar 16,17% variasi data penghasilan. Nilai *Mean Absolute Error (MAE)* sebesar USD 50.429,95 menunjukkan rata-rata selisih antara hasil prediksi dan data aktual. Sementara itu nilai Root *Mean Squared Error (RMSE)* sebesar USD 68.237,96 menunjukkan bahwa masih terdapat variasi kesalahan prediksi yang cukup besar sehingga model dapat dikategorikan memiliki performa sedang.

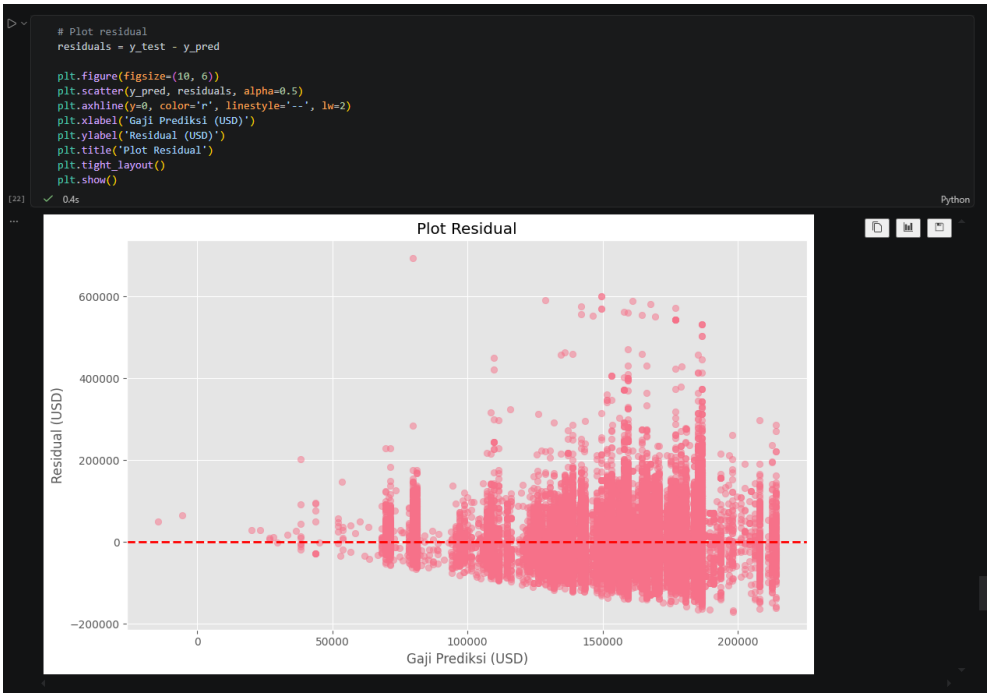
JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1177-1184

3.5 Hasil Prediksi



Gambar 8. Aktual vs Prediksi Gaji

Grafik aktual dan prediksi digunakan untuk membandingkan hasil prediksi model dengan data sebenarnya. Titik-titik yang mendekati garis diagonal menunjukkan prediksi yang semakin akurat. Hasil visualisasi menunjukkan bahwa model mampu mengikuti pola umum data meskipun masih terdapat penyimpangan pada nilai gaji yang sangat tinggi.



Gambar 9. Residual Plot

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1177-1184

Residual plot digunakan untuk mengevaluasi kesalahan prediksi model. Sebagian besar residual tersebar di sekitar garis nol, namun masih terdapat beberapa residual yang cukup besar pada rentang prediksi tertentu. Hal ini menunjukkan bahwa model Linear Regression mampu menangkap pola umum data, tetapi belum sepenuhnya mampu menjelaskan seluruh variasi penghasilan profesional IT.

4. KESIMPULAN

Berdasarkan hasil analisis menggunakan *AI Jobs Salary Dataset*, penelitian ini berhasil mengidentifikasi faktor-faktor yang berpengaruh terhadap tingkat penghasilan profesional IT. Hasil *Exploratory Data Analysis (EDA)* menunjukkan bahwa dataset memiliki kualitas yang baik karena tidak ditemukan nilai kosong (*missing values*). Selain itu, distribusi gaji menunjukkan adanya variasi penghasilan yang cukup besar pada berbagai profesi di bidang teknologi informasi.

Analisis lebih lanjut menunjukkan bahwa tingkat pengalaman kerja memiliki hubungan positif terhadap besarnya penghasilan. *Kategori Executive (EX)* memiliki rata-rata gaji tertinggi dibandingkan kategori pengalaman lainnya. Dari sisi profesi, kategori ***Engineer*** memperoleh rata-rata gaji tertinggi sebesar USD 174.215 atau sekitar 18,85% dari total rata-rata gaji seluruh kategori profesi yang dianalisis. Hasil ini menunjukkan bahwa profesi yang berfokus pada pengembangan sistem dan teknologi cenderung memiliki tingkat kompensasi yang lebih tinggi.

Model Linear Regression yang dibangun mampu melakukan prediksi penghasilan berdasarkan beberapa variabel yang digunakan dalam penelitian. Hasil evaluasi menunjukkan nilai R^2 sebesar 0,1617 dengan MAE sebesar USD 50.429,95 dan RMSE sebesar USD 68.237,96. Meskipun model belum mampu menjelaskan seluruh variasi data penghasilan, hasil yang diperoleh sudah dapat memberikan gambaran mengenai pola penghasilan profesional IT serta faktor-faktor yang memengaruhinya.

Untuk penelitian selanjutnya, disarankan menambahkan variabel lain yang berpotensi memengaruhi tingkat penghasilan, seperti tingkat pendidikan, sertifikasi profesional, keterampilan teknis, serta lokasi geografis yang lebih spesifik agar hasil prediksi menjadi lebih akurat. Selain itu, penggunaan metode *machine learning* lain seperti *Random Forest Regression*, *Decision Tree Regression*, atau *XGBoost* dapat dipertimbangkan untuk memperoleh performa prediksi yang lebih baik dibandingkan Linear Regression. Pengembangan penelitian ke dalam bentuk aplikasi atau dashboard interaktif juga dapat dilakukan agar hasil analisis dan prediksi gaji dapat dimanfaatkan secara lebih luas oleh mahasiswa, pencari kerja, maupun perusahaan dalam mendukung pengambilan keputusan terkait karier dan kompensasi di bidang teknologi informasi.

REFERENCES

Foorilla. (2026). AI Jobs Salary Dataset. GitHub. https://github.com/foorilla/ai-jobs-net-salaries. [Accessed: Jun. 2026].

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An introduction to statistical learning: With applications in Python. Springer.

VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. O'Reilly Media.