

ANALISIS FAKTOR PENGARUH TERHADAP PENGHASILAN PROFESI DATA ENGINEER MENGGUNAKAN METODE REGRESI LINEAR BERGANDA

Noviana Riza¹, Maysa Zaidee Aulia², Paulina Bota Kolin³, Kiki Mustaqim⁴

^{1,2,3,4}Universitas Logistik dan Bisnis Internasional; Jl. Sariasih No.54, Sarijadi, Kec. Sukasari, Kota Bandung, Jawa Barat 40151.

Received: 24 Desember 2024

Accepted: 14 Januari 2025

Published: 20 Januari 2025

Keywords:

Data Engineer, Regresi Linear, Penghasilan, Faktor Pengaruh.

Correspondent Email:

novianariza@ulbi.ac.id

Abstrak. Penelitian ini dilakukan untuk melihat faktor pengaruh terhadap penghasilan profesi *data engineer*. Penghasilan seringkali dipengaruhi oleh beberapa faktor yang diantaranya seperti pengalaman kerja, keahlian *database*, dan jenis kelamin. Metode yang digunakan yaitu regresi linear berganda dengan sumber data yang digunakan berdasarkan kuisioner dari salah satu *website*. Hasil penelitian yaitu pengalaman kerja, keahlian *database*, dan jenis kelamin berpengaruh signifikan sebesar 9,8% secara simultan. Secara parsial pengalaman kerja dan keahlian *database* berpengaruh signifikan, sedangkan jenis kelamin tidak berpengaruh signifikan.

Abstract. *This study was conducted to examine the factors influencing the salary of data engineer professionals. Income is often affected by several factors such as work experience, database expertise, and gender. The method used is multiple linear regression with data sourced from a questionnaire on a website. The results of the study indicate that work experience, database expertise, and gender have a significant simultaneous effect of 9.8%. Partially, work experience and database expertise have a significant effect, while gender does not have a significant effect.*

1. PENDAHULUAN

Di era digital ini, data telah menjadi aset yang sangat berharga dalam berbagai sektor kehidupan, salah satunya dalam dunia bisnis [1]. Data perusahaan yang diterima setiap saat dalam jumlah yang besar membuat perusahaan memerlukan ahli data yang dapat menangani persoalan tersebut. Hal ini membuat peran data science sangat diperlukan dalam pemenuhan kebutuhan perusahaan akan ahli data [2].

Data science adalah suatu disiplin ilmu yang khusus mempelajari data baik struktur maupun yang tidak struktur serta memproses data, mulai dari pengumpulan data, analisis data, pengolahan data, pengarsipan, pengelompokan data, penyajian data, distribusi data, dan cara mengubah menjadi kesatuan informasi yang dapat dipahami semua orang [3].

Data engineer merupakan salah satu profesi penting dan paling dicari dari *data science* [4]. Seorang *data engineer* bertanggung jawab atas pengelolaan, pemrosesan, dan analisis data dalam skala besar, serta memastikan keandalan dan ketersediaan data yang diperlukan untuk keputusan bisnis.

Penghasilan merupakan salah satu faktor yang perlu diperhatikan ketika memilih berkarir di bidang *data engineer*. Penghasilan seorang *data engineer* diperkirakan puluhan sampai ratusan juta perbulan [4]. Penghasilan yang diperoleh seorang *data engineer* dapat dipengaruhi oleh beberapa faktor, termasuk pengalaman kerja, keahlian *database*, dan jenis kelamin. Pengalaman kerja memainkan peran penting dalam menentukan penghasilan seorang *data engineer*. Profesional dengan pengalaman

kerja yang lebih luas biasanya akan memperoleh penghasilan yang lebih tinggi dibandingkan dengan yang memiliki pengalaman lebih sedikit. Dalam suatu perusahaan, keahlian menggunakan database juga dapat mempengaruhi penghasilan seorang data engineer. Banyaknya *database* yang dikuasai seorang *data engineer* dapat membantu proses pengolahan data lebih efektif dan efisien. Sehingga semakin banyak *database* yang dikuasai maka akan semakin tinggi penghasilan yang diperoleh. Selain itu, jenis kelamin juga dapat mempengaruhi penghasilan. Secara universal produktivitas laki-laki lebih tinggi dibandingkan dengan perempuan. Hal ini dipengaruhi oleh faktor biologis pada perempuan seperti harus cuti melahirkan [8]. Hal ini tentunya akan berdampak pada penghasilan.

Dalam menganalisis data, model regresi adalah alat yang paling berharga [5]. Metode regresi linear berganda digunakan untuk mengidentifikasi pengaruh dari faktor independen terhadap faktor dependen. Dalam hal ini yang menjadi faktor dependen adalah penghasilan *data engineer*, sedangkan faktor independennya adalah pengalaman kerja, keahlian *database*, dan jenis kelamin. Penerapan regresi linear berganda digunakan untuk melihat pengaruh pengalaman kerja, keahlian *database*, dan jenis kelamin terhadap penghasilan *data engineer*.

Dalam penelitian sebelumnya terdapat perbandingan penggunaan Algoritma Regresi Linier Berganda dengan *tree-based train* untuk memprediksi faktor pengaruh terhadap penghasilan dengan penelitian berjudul *Research on the Effect of Age and Experience on Salary based on Linear Regression and Tree-Based Train*. Hasil dari regresi menunjukkan tidak adanya hubungan yang jelas baik antara umur atau tahun pengalaman dan pendapatan, sedangkan hasil dari *tree-based train* menunjukkan bahwa umur atau tahun pengalaman yang lebih tinggi sesuai dengan gaji yang lebih besar secara keseluruhan [6]. Pada penelitian lain dengan judul Pengaruh Modal, Pengalaman Kerja Dan Harga Jual Terhadap Pendapatan Petani Rumput Laut di Dusun Ponrang Kecamatan Ponrang Kabupaten Luwu, hasil menunjukkan bahwa secara bersama-sama variabel modal, harga jual, dan pengalaman kerja memiliki pengaruh positif

terhadap pendapatan. Secara parsial harga jual dan pengalaman kerja memiliki pengaruh positif, sedangkan modal memiliki pengaruh negatif terhadap pendapatan. [7].

Tujuan dari penelitian ini adalah untuk mengetahui pengaruh pengalaman kerja, keahlian *database* dan jenis kelamin terhadap penghasilan seorang data engineer dan mengetahui penggunaan regresi linear berganda untuk melihat pengaruh pengalaman kerja, keahlian *database* dan jenis kelamin terhadap penghasilan seorang *data engineer* yang nantinya akan diimplementasikan menggunakan *Python*.

2. TINJAUAN PUSTAKA

2.1 Data Engineer

Data engineer adalah seseorang yang dapat membuat dan mengembangkan desain manajemen data dan memantau infrastrukturnya [4]. *Data engineer* adalah desainer, pembangun, dan manajer informasi yang menciptakan infrastruktur untuk menyimpan, memproses, dan menganalisis informasi [11].

2.2 Penghasilan

Penghasilan adalah jumlah uang yang diterima seseorang atas pekerjaan mereka selama periode waktu tertentu, baik dalam satuan harian, mingguan, bulanan, atau tahunan [10]. Penghasilan adalah sesuatu yang diterima seseorang atau rumah tangga melalui usaha atau bekerja [12].

2.3 Faktor Pengaruh Penghasilan

Produktivitas seseorang dapat mempengaruhi pengalaman kerja dan pengalaman kerja dapat mempengaruhi penghasilan yang akan didapat [9]. Keahlian dan keterampilan yang tinggi akan meningkatkan efektifitas dan efisiensi yang akan berdampak pada penghasilan [13]. Jenis kelamin berpengaruh pada penghasilan dan dapat menunjukkan tingkat produktivitasnya, yang berdampak pada penghasilannya [8].

2.4 Regresi Linear Berganda

Regresi linear berganda digunakan untuk menghubungkan dua variabel atau lebih dengan variabel bebas dan terikat dan digunakan dalam peramalan [14]. Model regresi linear berganda dapat dilihat pada persamaan berikut ini.

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n + e$$

Keterangan :

Y = Variabel terikat

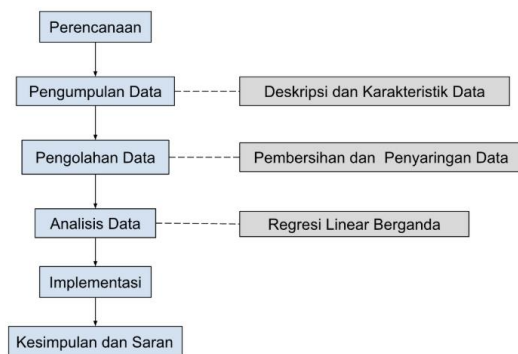
X = Variabel bebas

a = Konstanta

b = Koefisien estimasi

e = Nilai *error* (nilai – nilai lain yang tidak dimasukkan kedalam persamaan) [10].

3. METODE PENELITIAN



Gambar 1. Alur Penelitian

3.1 Perencanaan

Perencanaan penelitian merupakan perencanaan menyeluruh dari kegiatan penelitian yang akan dilakukan. Perencanaan ini dimulai dari pengumpulan data, menganalisis data, menguji segala hipotesis serta mengimplementasikannya. Selanjutnya data akan disimpulkan dan diberi saran.

3.2 Pengumpulan Data

Data yang akan digunakan dikumpulkan dari kuesioner yang diunggah melalui situs *brentozar.com*. Data yang telah diunduh kemudian diunggah ke *Google Spreadsheet* untuk diteliti lebih lanjut pada setiap data yang ada. Data yang diberikan memiliki beberapa kolom informasi yang nantinya akan diolah untuk melakukan penelitian ini. Terdapat 31 kolom dan 12.979 baris yang ada pada data mentah.

3.3 Pengolahan Data

Pembersihan dan penyaringan data dilakukan menggunakan *Google Spreadsheet*. Pembersihan data dilakukan dengan menghapus kolom yang tidak dibutuhkan, sedangkan penyaringan data dilakukan dengan menyaring nilai dari data-data yang akan digunakan sebagai bahan penelitian. Kolom yang akan

digunakan setelah pembersihan dan penyaringan data yaitu tahun survey, penghasilan (dalam USD), keahlian *database* (dalam satuan tahun), status pekerja, nama profesi, pengalaman bekerja di bidang profesi tersebut (dalam satuan tahun), dan jenis kelamin dengan keterangan sebagai berikut.

Y = Penghasilan

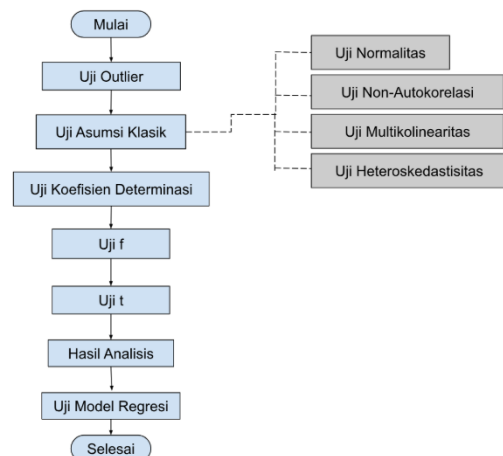
X1 = Pengalaman bekerja

X2 = Keahlian *database*

X3 = Jenis kelamin (1 = laki-laki, 2 = perempuan).

3.4 Analisis Data

Analisis data yang akan dilakukan adalah sebagai berikut.



Gambar 2. Alur Regresi Linear Berganda

3.5 Implementasi

Penelitian regresi linear berganda ini akan diimplementasikan menggunakan bahasa pemrograman *Python* untuk mengikuti perkembangan saat ini.

3.6 Kesimpulan dan Saran

Pada tahap ini akan disimpulkan semua hasil analisis yang telah dilakukan, serta akan diberikan saran-saran yang berhubungan dengan penelitian ini.

4. HASIL DAN PEMBAHASAN

4.1 Dataset

Dataset yang ada akan melalui proses *filtering* data dan juga perubahan tipe data untuk kolom *database* dan gender. Berikut merupakan data akhir yang telah melalui proses *filtering* dan memiliki 352 baris.

Survey Year	EmploymentStatus	JobTitle	SalaryUSD	YearsWithThisTypeOfJob	Database	Gender
2024	Full time employee	Engineer	120.000	3	2	1
2024	Full time employee	Engineer	145.000	10	3	1
2024	Full time employee	Engineer	126.000	6	5	1
2024	Full time employee	Engineer	45.000	7	2	1
2024	Full time employee	Engineer	133.500	4	3	2
2024	Full time employee	Engineer	139.900	8	2	2
2024	Full time employee	Engineer	155.000	20	5	1
2024	Full time employee	Engineer	115.000	23	3	1
2024	Full time employee	Engineer	30.000	4	1	1
2024	Full time employee	Engineer	125.000	19	3	1
2024	Full time employee	Engineer	185.000	4	2	1
2024	Full time employee	Engineer	80.000	0	3	1
2024	Full time employee	Engineer	110.000	2	3	1
2024	Full time employee	Engineer	140.000	3	6	1
2024	Full time employee	Engineer	300.000	35	4	1
2024	Full time employee	Engineer	115.000	2	6	1
2024	Full time employee	Engineer	155.000	25	3	1
2024	Full time employee	Engineer	160.000	2018	2	1
2024	Full time employee	Engineer	125.000	3	1	1
2024	Full time employee	Engineer	81.000	2	2	1

Gambar 3. Dataset Filter

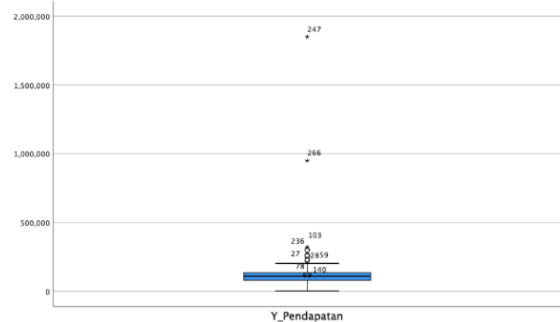
Survey Year	EmploymentStatus	JobTitle	SalaryUSD	YearsWithThisTypeOfJob	Database	Gender
2020	Full time employee	Engineer	120.000	3	2	1
2020	Full time employee	Engineer	117.600	5	1	1
2020	Full time employee	Engineer	105.617	1	2	1
2020	Full time employee	Engineer	92.000	6	2	1
2020	Full time employee	Engineer	90.000	11	3	1
2020	Full time employee	Engineer	95.000	12	1	1
2020	Full time employee	Engineer	135.000	25	2	1
2020	Full time employee	Engineer	100.000	20	4	1
2020	Full time employee	Engineer	85.000	10	1	1
2020	Full time employee	Engineer	78.000	2	1	1
2020	Full time employee	Engineer	145.000	25	4	1
2020	Full time employee	Engineer	103.000	22	1	1
2020	Full time employee	Engineer	148.000	15	4	1
2020	Full time employee	Engineer	139.000	3	4	1
2020	Full time employee	Engineer	65.800	2	3	1
2020	Full time employee	Engineer	90.000	7	3	1
2020	Full time employee	Engineer	105.000	20	4	1
2020	Full time employee	Engineer	135.000	1	2	1
2020	Full time employee	Engineer	150.000	20	2	1
2020	Full time employee	Engineer	75.000	3	1	1

Gambar 4. Dataset Filter Lanjutan

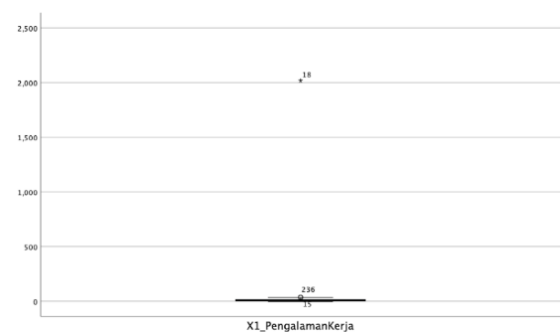
4.2 Analisis Data

Sebelum masuk ke analisis regresi linear berganda, terdapat uji asumsi analisis yang akan dilakukan seperti uji *outlier* dan uji asumsi klasik. Setelah uji asumsi analisis terpenuhi, maka akan dilanjutkan dengan uji analisis regresi linear berganda. Analisis data dilakukan dengan 3 faktor sebagai variabel X yaitu X1 adalah pengalaman kerja, X2 adalah keahlian database, dan X3 adalah jenis kelamin, serta 1 variabel Y yaitu penghasilan data engineer.

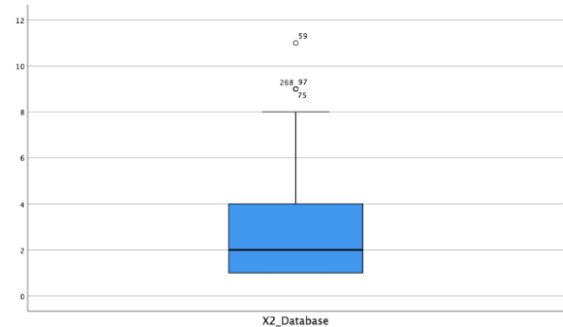
4.2.1 Uji Outlier



Gambar 5. Boxplot Penghasilan

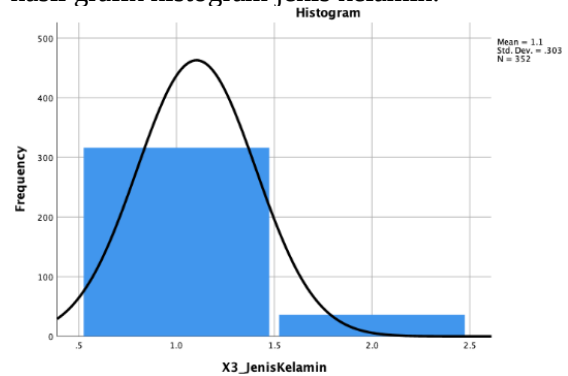


Gambar 6. Boxplot Pengalaman Kerja



Gambar 7. Boxplot Database

Berdasarkan Gambar 5, Gambar 6, dan Gambar 7, maka data yang akan dieliminasi yaitu baris 247, 266, 103, 236, 27, 28, 59, 78, 69, 140, 18, 15, 268, 97, dan 75, sehingga data yang akan digunakan akan berjumlah 337 baris. Untuk data jenis kelamin yang bersifat kategori, maka akan dilihat grafik histogramnya. Berikut hasil grafik histogram jenis kelamin.



Gambar 8. Histogram Jenis Kelamin

Berdasarkan Gambar 8, terlihat bahwa ada perbedaan yang cukup jauh antara data jenis kelamin laki-laki dan perempuan.

4.2.2 Uji Normalitas

One-Sample Kolmogorov-Smirnov Test

		Unstandardized Residual
Normal Parameters ^{a,b}	Mean	.0000000
	Std. Deviation	35838.6350
Most Extreme Differences	Absolute	.036
	Positive	.023
	Negative	-.036
Test Statistic		.036
Asymp. Sig. (2-tailed)		.200 ^{c,d}

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

d. This is a lower bound of the true significance.

Gambar 9. Uji Normalitas

Berdasarkan Gambar 9, nilai probabilitas (error) atau *Asymp Sig.* adalah 0.200. Jika dibandingkan dengan tingkat signifikansi = 0,05 maka dapat disimpulkan bahwa nilai probabilitasnya lebih besar. Oleh karena itu, uji asumsi normalitas terpenuhi dan data berdistribusi normal.

4.2.3 Uji Multikolinearitas

Coefficients ^a							
Model	Unstandardized Coefficients B	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance VIF		
1	(Constant) 85870.428	8492.161	10.112	.000			
	X1_PengalamanKerja 1345.153	260.655	.270	5.161	.000	.990	1.010
	X2_Database 3133.281	1224.521	.135	2.559	.011	.980	1.020
	X3_JenisKelamin -637.932	6382.126	-.005	-.100	.920	.990	1.011

a. Dependent Variable: Y_Pendapatan

Gambar 10. Uji Multikolinearitas

Berdasarkan Gambar 10, nilai VIF X1 adalah 1.010, X2 adalah 1.020, dan X3 adalah 1.011 dimana nilainya lebih kecil dari 10. Berdasarkan hasil tersebut membuktikan bahwa tidak terjadi multikolinearitas.

4.2.4 Uji Autokorelasi

Model Summary ^b					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.313 ^a	.098	.090	35999.708	1.825

a. Predictors: (Constant), X3_JenisKelamin, X1_PengalamanKerja, X2_Database

b. Dependent Variable: Y_Pendapatan

Gambar 11. Uji Autokorelasi

Berdasarkan Gambar 11, nilai Durbin-Watson adalah 1.825. Berdasarkan hasil yang diperoleh disimpulkan bahwa data tidak terjadi autokorelasi sehingga uji asumsi non autokorelasi terpenuhi.

4.2.5 Uji Heteroskedastisitas

Coefficients ^a					
Model	Unstandardized Coefficients B	Standardized Coefficients Beta	t	Sig.	
1	(Constant) 1.710E+9	421905074	4.054	.000	
	X1_PengalamanKerja -177132.85	12949770.1	-.001	-.014	.989
	X2_Database -1.284E+8	60836316.6	-.116	-2.110	.036
	X3_JenisKelamin -83575293	317074918	-.014	-.264	.792

a. Dependent Variable: RES_2

Gambar 12. Uji Heteroskedastisitas

Berdasarkan Gambar 12, nilai *Sig.* untuk variabel X1 adalah 0.989, nilai *Sig.* untuk X2 adalah 0.036, dan nilai *Sig.* untuk X3 adalah 0.792. Nilai X1 dan X3 lebih besar dari tingkat signifikansi $\alpha = 0,05$ yang membuktikan bahwa data yang ada tidak menunjukkan gejala heteroskedastisitas, sedangkan nilai X2 lebih kecil dari tingkat signifikansi $\alpha = 0,05$ yang memungkinkan adanya indikasi heteroskedastisitas.

4.2.6 Uji Koefisien Determinasi

Model Summary ^b					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.313 ^a	.098	.090	35999.708	1.825

a. Predictors: (Constant), X3_JenisKelamin, X1_PengalamanKerja, X2_Database

b. Dependent Variable: Y_Pendapatan

Gambar 13. Uji Koefisien Determinasi

Berdasarkan Gambar 13, nilai *R square* menunjukkan nilai 0.098 yang berarti variabel independen berpengaruh sebesar 9,8% terhadap variabel dependen secara bersama-sama.

4.2.7 Uji F

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.698E+10	3	1.566E+10	12.084	.000 ^b
	Residual	4.316E+11	333	1.296E+9		
	Total	4.785E+11	336			

a. Dependent Variable: Y_Pendapatan

b. Predictors: (Constant), X3_JenisKelamin, X1_PengalamanKerja, X2_Database

Gambar 14. Uji F

Berdasarkan Gambar 14, nilai *Sig.* yang dihasilkan adalah 0.000. Nilai tersebut lebih kecil dari tingkat signifikansi $\alpha = 0,05$ yang membuktikan bahwa model yang diterapkan sudah cocok dan terdapat pengaruh signifikan yang dilakukan oleh variabel independen terhadap variabel dependen. Hal ini juga dapat dilihat dari nilai *f.hitung* yaitu 12.084, sedangkan nilai *f.tabel* dengan $df (k-1)(n-k) = (2)(334)$ dengan tingkat signifikansi 0,05 adalah 0,0513. Karena nilai *f.hitung* > *f.tabel* maka model yang diterapkan sudah cocok dan terdapat pengaruh signifikan yang dilakukan oleh variabel independen terhadap variabel dependen secara bersama-sama.

4.2.8 Uji T

Coefficients ^a							
Model	Unstandardized Coefficients B	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance VIF		
1	(Constant) 85870.428	8492.161	10.112	.000			
	X1_PengalamanKerja 1345.153	260.655	.270	5.161	.000	.990	1.010
	X2_Database 3133.281	1224.521	.135	2.559	.011	.980	1.020
	X3_JenisKelamin -637.932	6382.126	-.005	-.100	.920	.990	1.011

a. Dependent Variable: Y_Pendapatan

Gambar 15. Uji T

Berdasarkan Gambar 15, nilai *Sig.* untuk variabel X1 adalah 0.000, nilai *Sig.* untuk X2 adalah 0.011, dan nilai *Sig.* untuk X3 adalah 0.920. Hal ini membuktikan bahwa variabel X1 dan X2 memiliki pengaruh signifikan karena nilai *Sig.* lebih kecil dari tingkat signifikansi $\alpha = 0,05$, sedangkan variabel X3 tidak memiliki pengaruh signifikan karena nilai *Sig.* lebih besar dari tingkat signifikansi $\alpha = 0,05$. Hal ini juga

dapat dilihat dari nilai t.hitung X1 yaitu 5.161, t.hitung X2 yaitu 2.559, dan t.hitung X3 yaitu $[-0.100]$, sedangkan nilai t.tabel dengan df $(n-k) = 334$ dengan tingkat signifikansi 0,05 adalah $[-1,6494]$. Hal ini membuktikan bahwa X1 dan X2 memiliki pengaruh signifikan karena lebih besar dari t.tabel, sedangkan X3 tidak memiliki pengaruh signifikan karena lebih kecil dari t.tabel. Persamaan regresi yang diperoleh adalah sebagai berikut.

$$Y = 85870.428 + 1345.153X_1 + 3133.281X_2 - 637.932X_3 + e.$$

4.2.9 Hasil Analisis

Dilihat dari persamaan yang dihasilkan, maka penjelasan yang dapat diberikan adalah sebagai berikut.

- Nilai konstanta sebesar 85870.428 yang berarti ketika variabel independen bernilai 0, maka variabel penghasilan akan bernilai 85870.428.
- Nilai koefisien regresi variabel X1 (pengalaman kerja) sebesar 1345.153 yang berarti ada pengaruh positif terhadap variabel penghasilan dan ketika ada kenaikan satuan pada variabel X1 maka variabel penghasilan akan ikut naik sebesar 1345.153.
- Nilai koefisien regresi variabel X2 (*database*) sebesar 3133.281 yang berarti ada pengaruh positif terhadap variabel penghasilan dan ketika ada kenaikan satuan pada variabel X2 maka variabel penghasilan akan ikut naik sebesar 3133.281.
- Nilai koefisien regresi variabel X3 (jenis kelamin) sebesar -637.932 yang berarti ada pengaruh negatif terhadap variabel penghasilan dan ketika ada kenaikan satuan pada variabel X3 maka variabel penghasilan akan turun sebesar 637.932.
- Berdasarkan nilai koefisien regresi variabel X3 (jenis kelamin) yang memiliki pengaruh negatif yaitu penurunan sebesar 637.932 setiap naik satu satuan, maka jika pekerja data engineer perempuan, penghasilan yang diterima mungkin lebih sedikit dibandingkan dengan pekerja laki-laki. Hal ini dibuktikan dengan pengubahan inisial jenis kelamin yaitu laki-laki sebagai 1 dan perempuan sebagai 2.

Namun hal ini tidak dapat dipastikan 100% karena adanya perbedaan jumlah data.

4.2.10 Uji Model Regresi

Berdasarkan analisis yang telah dilakukan, maka terbentuk persamaan regresi yang bisa dipakai untuk memprediksi penghasilan dengan nilai tertentu.

Pengalaman kerja 10 tahun, menguasai 3 database, laki-laki	
X1 =	10
X2 =	3
X3 =	1
Pendapatan =	108083,87

Pengalaman kerja 10 tahun, menguasai 4 database, laki-laki	
X1 =	10
X2 =	4
X3 =	1
Pendapatan =	111217,15

Gambar 16. Uji Model Regresi 1

Berdasarkan Gambar 16, ketika pekerja *data engineer* memiliki keahlian *database* yang lebih banyak, maka penghasilan yang akan diterima akan lebih besar 2,90% dari penghasilan *data engineer* yang memiliki keahlian *database* yang lebih sedikit.

Pengalaman kerja 3 tahun, menguasai 10 database, laki-laki	
X1 =	3
X2 =	10
X3 =	1
Pendapatan =	120600,76

Pengalaman kerja 4 tahun, menguasai 10 database, laki-laki	
X1 =	4
X2 =	10
X3 =	1
Pendapatan =	121945,92

Gambar 17. Uji Model Regresi 2

Berdasarkan Gambar 17, ketika pekerja *data engineer* memiliki pengalaman kerja yang lebih banyak, maka penghasilan yang akan diterima akan lebih besar 1,15% dari penghasilan *data engineer* yang memiliki pengalaman kerja yang lebih sedikit.

Pengalaman kerja 3 tahun, menguasai 3 database, laki-laki		
X1 =	3	
X2 =	3	
X3 =	1	Data Asli
Pendapatan =	98667,80	92.000

Pengalaman kerja 3 tahun, menguasai 3 database, perempuan		
X1 =	3	
X2 =	3	
X3 =	2	Data Asli
Pendapatan =	98029,87	85.000

Gambar 18. Uji Model Regresi 3

Berdasarkan Gambar 18, pekerja *data engineer* perempuan memiliki penghasilan

yang lebih sedikit dibandingkan dengan pekerja *data engineer* laki-laki, namun hasil prediksi dengan model regresi yang terbentuk tidak dapat dipastikan 100% karena memiliki adanya perbedaan jumlah data yang dianalisis dan perbedaan hasil dengan data asli.

4.3 Implementasi Metode

Implementasi metode yang akan dilakukan pada penelitian ini yaitu menggunakan *Python* sebagai alat uji. Berikut merupakan langkah-langkah yang dilakukan untuk melakukan uji analisis.

4.3.1 Uji Outlier

```
[ ] def remove_outliers_iqr(data, column):
    Q1 = data[column].quantile(0.25)
    Q3 = data[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    cleaned_data = data[(data[column] >= lower_bound) & (data[column] <= upper_bound)]
    return cleaned_data

# Menghapus outlier dari setiap variabel independen
data = remove_outliers_iqr(data, 'YearsWithThisTypeOfJob')
data = remove_outliers_iqr(data, 'Database')
data = remove_outliers_iqr(data, 'SalaryUSD')
```

Gambar 19. Uji Outlier (Python)

Berdasarkan Gambar 19, proses penghapusan data *outlier* yang dilakukan dengan menghilangkan data yang merupakan *outlier* pada kedua variabel numerik.

4.3.2 Uji Normalitas

```
[313] residuals = model.resid

#menggunakan Kolmogorov smirnov
statistic, pvalue = kstest(residuals, 'norm', args=(residuals.mean(), residuals.std()))
print(f'nilai pvalue sebesar: {:.4f}'.format(pvalue))
```

nilai pvalue sebesar: 0.7462

Gambar 20. Uji Outlier (Python)

Hasil Gambar 20, menunjukkan bahwa nilai *p-value* adalah 0.7462. Jika dibandingkan dengan tingkat signifikansi = 0,05 dapat disimpulkan bahwa nilai *p-value* lebih besar. Berdasarkan hasil tersebut, maka data berdistribusi normal.

4.3.3 Uji Multikolinearitas

```
[314] VIF = [variance_inflation_factor(X, i) for i in range(X.shape[1])]
print('constant :', round(VIF[0], 4))
print('YearsWithThisTypeOfJob :', round(VIF[1], 4))
print('Database :', round(VIF[2], 4))
print('Gender :', round(VIF[3], 4))

constant : 18.7529
YearsWithThisTypeOfJob : 1.0098
Database : 1.0203
Gender : 1.0106
```

Gambar 21. Uji Multikolinearitas (Python)

Hasil Gambar 21, menunjukkan bahwa nilai VIF X1 adalah 1.0098, X2 adalah 1.0203, dan X3 adalah 1.0106 dimana nilai ini lebih kecil dari 10. Berdasarkan hasil tersebut, maka data tidak terjadi multikolinearitas.

4.3.4 Uji Autokorelasi

```
[315] dw_statistic = durbin_watson(model.resid)
print(f'Durbin-Watson statistic: {dw_statistic:.4f}')
```

Durbin-Watson statistic: 1.8252

Gambar 22. Uji Autokorelasi (Python)

Hasil Gambar 22, menunjukkan bahwa nilai *durbin-watson* adalah 1.8252 dimana nilai ini mendekati angka dua. Berdasarkan hasil tersebut, maka data tidak terjadi autokorelasi.

4.3.5 Uji Heteroskedastisitas

```
[316] BP, pvalue, _, _ = het_breuschpagan(residuals, X)
print('nilai pvalue sebesar: {:.4f}'.format(pvalue))
```

nilai pvalue sebesar: 0.2125

Gambar 23. Uji Heteroskedastisitas (Python)

Hasil Gambar 23, menunjukkan bahwa nilai *p-value* adalah 0.2125. Jika dibandingkan dengan tingkat signifikansi = 0,05 dapat disimpulkan bahwa nilai *p-value* lebih besar. Berdasarkan hasil tersebut, maka data tidak terjadi heteroskedastisitas.

4.3.6 Uji Koefisien determinasi

```
[317] # Menampilkan R-squared dalam bentuk persentase
r_squared = model.rsquared
r_squared_percentage = r_squared * 100
print(f'Koefisien Determinasi (R-squared) dalam bentuk persentase: {r_squared_percentage:.2f}%')
```

Koefisien Determinasi (R-squared) dalam bentuk persentase: 9.82%

Gambar 24. Uji Koefisien Determinasi (Python)

Hasil Gambar 24, menunjukkan bahwa variabel independen berpengaruh sebesar 9.82% terhadap variabel dependen secara bersama-sama.

4.3.7 Uji F

```
[318] f_statistic = model.fvalue
f_pvalue = model.f_pvalue
print(f'\nUji F untuk keseluruhan model:')
print(f'F-statistic: {f_statistic:.4f}')
print(f'p-value: {f_pvalue:.4f}')
```

Uji F untuk keseluruhan model:
F-statistic: 12.0838
p-value: 0.0000

Gambar 25. Uji F (Python)

Hasil Gambar 25, menunjukkan bahwa nilai *f.hitung* yaitu 12.0838, sedangkan nilai *f.tabel* dengan $df (k-1)(n-k) = (2)(334)$ dengan tingkat signifikansi 0,05 adalah 0,0513, sedangkan *p-value* yang dihasilkan adalah 0.0000. Karena nilai *f.hitung* > *f.tabel* dan nilai *p-value* < tingkat signifikansi $\alpha = 0,05$ maka model yang diterapkan sudah cocok dan terdapat pengaruh signifikan yang dilakukan

oleh variabel independen terhadap variabel dependen secara bersama-sama.

4.3.8 Uji T

```
[319] # Menampilkan hasil uji T secara spesifik untuk masing-masing variabel independen
print("\nUji T untuk koefisien regresi:")
print(f"Intercept: t-value = {model.tvalues[0]:.4f}, p-value = {model.pvalues[0]:.4f}")
print(f"YearsWithThisTypeOfJob: t-value = {model.tvalues[1]:.4f}, p-value = {model.pvalues[1]:.4f}")
print(f"Database: t-value = {model.tvalues[2]:.4f}, p-value = {model.pvalues[2]:.4f}")
print(f"Gender: t-value = {model.tvalues[3]:.4f}, p-value = {model.pvalues[3]:.4f}")
```

```
Uji T untuk koefisien regresi:
Intercept: t-value = 10.1117, p-value = 0.0000
YearsWithThisTypeOfJob: t-value = 5.1607, p-value = 0.0000
Database: t-value = 2.5588, p-value = 0.0109
Gender: t-value = -0.1000, p-value = 0.9204
```

Gambar 26. Uji T (Python)

Hasil Gambar 26, menunjukkan bahwa nilai t_{hitung} untuk X_1 adalah 5.1607 dengan $p-value$ 0.0000, t_{hitung} untuk X_2 adalah 2.5588 dengan $p-value$ 0.0109, dan t_{hitung} untuk X_3 adalah $|-0.1000|$ dengan $p-value$ 0.9204. Nilai t_{tabel} dengan df ($n-k$) = 334 dengan tingkat signifikansi 0,05 adalah $|-1,6494|$. Karena $t_{hitung} X_1$ dan $X_2 > t_{tabel}$ dan $p-value <$ tingkat signifikansi 0,05, maka X_1 dan X_2 memiliki pengaruh signifikan terhadap variabel dependen, sedangkan karena $t_{hitung} X_3 < t_{tabel}$ dan $p-value >$ tingkat signifikansi 0,05, maka X_3 tidak memiliki pengaruh signifikan terhadap variabel dependen.

4.3.9 Hasil Analisis

```
[320] print("Koefisien Regresi:")
print(f"Konstanta: {model.params[0]:.4f}")
print(f"YearsWithThisTypeOfJob: {model.params[1]:.4f}")
print(f"Database: {model.params[2]:.4f}")
print(f"Gender: {model.params[3]:.4f}")
```

```
Koefisien Regresi:
Konstanta: 85870.4276
YearsWithThisTypeOfJob: 1345.1528
Database: 3133.2811
Gender: -637.9315
```

```
[321] a = model.params[0]
b1 = model.params[1]
b2 = model.params[2]
b3 = model.params[3]
a, b1, b2, b3
```

```
(85870.42763112953, 1345.152758423222, 3133.281058644627, -637.9315397456236)
```

Gambar 27. Hasil Analisis (Python)

Dilihat dari persamaan yang dihasilkan, maka penjelasan yang dapat diberikan adalah sebagai berikut.

- Nilai konstanta sebesar 85870.4276 yang berarti ketika variabel independen bernilai 0, maka variabel penghasilan akan bernilai 85870.4276.
- Nilai koefisien regresi variabel X_1 (pengalaman kerja) sebesar 1345.1528 yang berarti ada pengaruh positif terhadap variabel penghasilan dan ketika ada kenaikan satuan pada variabel X_1 maka variabel penghasilan akan ikut naik sebesar 1345.1528.
- Nilai koefisien regresi variabel X_2 (*database*) sebesar 3133.2811 yang

berarti ada pengaruh positif terhadap variabel penghasilan dan ketika ada kenaikan satuan pada variabel X_2 maka variabel penghasilan akan ikut naik sebesar 3133.2811.

- Nilai koefisien regresi variabel X_3 (jenis kelamin) sebesar -637.9315 yang berarti ada pengaruh negatif terhadap variabel penghasilan dan ketika ada kenaikan satuan pada variabel X_3 maka variabel penghasilan akan turun sebesar 637.9315.
- Berdasarkan nilai koefisien regresi variabel X_3 (jenis kelamin) yang memiliki pengaruh negatif yaitu penurunan sebesar 637.9315 setiap naik satu satuan, maka jika pekerja data engineer perempuan, penghasilan yang diterima mungkin lebih sedikit dibandingkan dengan pekerja laki-laki. Hal ini dibuktikan dengan pengubahan inisial jenis kelamin yaitu laki-laki sebagai 1 dan perempuan sebagai 2. Namun hal ini tidak dapat dipastikan 100% karena adanya perbedaan jumlah data.

4.3.10 Uji Model Regresi

Berdasarkan analisis yang telah dilakukan, maka terbentuk persamaan regresi yang bisa dipakai untuk memprediksi penghasilan dengan nilai tertentu.

```
[26] #Pengalaman kerja 10 tahun, menguasai 3 database, laki-laki
x1 = 10
x2 = 3
x3 = 1
predicted_Y = a + b1*x1 + b2*x2 + b3*x3
predicted_Y
```

```
108083.8668257411
```

```
[27] #Pengalaman kerja 10 tahun, menguasai 4 database, laki-laki
x1 = 10
x2 = 4
x3 = 1
predicted_Y = a + b1*x1 + b2*x2 + b3*x3
predicted_Y
```

```
111217.14788438573
```

Gambar 28. Uji Model Regresi 1 (Python)

Berdasarkan Gambar 28, ketika pekerja *data engineer* memiliki keahlian *database* yang lebih banyak, maka penghasilan yang akan diterima akan lebih besar 2,90% dari penghasilan *data engineer* yang memiliki keahlian *database* yang lebih sedikit.

```
[ ] #Pengalaman kerja 3 tahun, menguasai 10 database, laki-laki
x1 = 3
x2 = 10
x3 = 1
predicted_Y = a + b1*x1 + b2*x2 + b3*x3
predicted_Y

120600.76494535716

[ ] #Pengalaman kerja 4 tahun, menguasai 10 database, laki-laki
x1 = 4
x2 = 10
x3 = 1
predicted_Y = a + b1*x1 + b2*x2 + b3*x3
predicted_Y

121945.9177011995
```

Gambar 29. Uji Model Regresi 2 (Python)

Berdasarkan Gambar 29, ketika pekerja *data engineer* memiliki pengalaman kerja yang lebih banyak, maka penghasilan yang akan diterima akan lebih besar 1,15% dari penghasilan *data engineer* yang memiliki pengalaman kerja yang lebih sedikit.

```
[ ] #Pengalaman kerja 3 tahun, menguasai 3 database, laki-laki
x1 = 3
x2 = 3
x3 = 1
predicted_Y = a + b1*x1 + b2*x2 + b3*x3
predicted_Y

98667.79753484478

[ ] #Pengalaman kerja 3 tahun, menguasai 3 database, perempuan
x1 = 3
x2 = 3
x3 = 2
predicted_Y = a + b1*x1 + b2*x2 + b3*x3
predicted_Y

98029.86599509916
```

Gambar 30. Uji Model Regresi 3 (Python)

Berdasarkan Gambar 30, pekerja *data engineer* perempuan memiliki penghasilan yang lebih sedikit dibandingkan dengan pekerja *data engineer* laki-laki, namun hasil prediksi dengan model regresi yang terbentuk tidak dapat dipastikan 100% karena memiliki adanya perbedaan jumlah data yang dianalisis dan perbedaan hasil dengan data asli.

5. KESIMPULAN

- Hasil analisis berdasarkan nilai R-squared menunjukan bahwa variabel X1 (pengalaman kerja), variabel X2 (keahlian *database*), dan variabel X3 (jenis kelamin) memberikan pengaruh sebesar 9.82% secara bersama-sama.
- Hasil analisis berdasarkan persamaan regresi yang dihasilkan menunjukan bahwa nilai konstanta sebesar 85870.4276 yang berarti ketika variabel independen bernilai

0, maka variabel penghasilan akan bernilai 85870.4276. Nilai koefisien regresi variabel X1 (pengalaman kerja) sebesar 1345.1528, koefisien regresi variabel X2 (keahlian *database*) sebesar 3133.2811, koefisien regresi variabel X3 (jenis kelamin) sebesar -637.9315. Hal ini menunjukkan bahwa kenaikan ataupun penurunan pada penghasilan *data engineer* dipengaruhi oleh nilai satuan dari variabel X1, X2, dan X3 yang dikalikan dengan koefisien regresinya.

- Hasil penelitian ini dapat dijadikan bahan rujukan untuk penelitian selanjutnya dengan meneliti variabel lain yang mungkin memiliki pengaruh yang lebih besar atau metode lain dengan data yang lebih seimbang.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada para dosen program studi Sains Data Universitas Logistik dan Bisnis Internasional yang telah memberikan motivasi serta dukungan sehingga artikel penelitian ini terbit.

DAFTAR PUSTAKA

- [1] B. Siregar et al., "Pengenalan Data Science dan Profesi Data Scientist di SMA Pramita Tangerang," 2022. [Online]. Available: <https://journal.formosapublisher.org/index.php/jpmb>.
- [2] M. Syamsu and Widodo, "Peran Data Science dan Data Scientist Untuk Mentransformasi Data Dalam Industri 4.0," 2021. [Online]. Available: <https://ojs.itb-ad.ac.id/index.php/JUTECH/article/view/1540>
- [3] M. Adhisyanda Aditya, R. Dicky Mulyana, I. Putu Eka, and S. Rheno Widiyanto, "Penggabungan Teknologi Untuk Analisa Data Berbasis Data Science," Seminar Nasional Teknologi Komputer & Sains (SAINTEKS), 2020. [Online]. Available: <https://prosiding.seminar-id.com/index.php/sainteks/article/view/404>
- [4] I. Setiawan, "PERBEDAAN DATA ENGINEER, DATA SCIENTIST DAN DATA ANALYST," Jurnal Kajian Pendidikan FKIP Universitas Dwijendra, vol. 12, no. 2, 2021, [Online]. Available:

- <http://ejournal.undwi.ac.id/index.php/widyaaccarya/index>
- [5] H. T. Nugroho and S. Hidayat, "Implementasi Intelegen Bisnis dengan Visualisasi Data Gaji dan Algoritma Linear Regresion," 2024. [Online]. Available: <https://journal.stmiki.ac.id>.
 - [6] H. Gao, "Research on the Effect of Age and Experience on Salary based on Linear Regression and Tree-Based Train," 2024. [Online]. Available: <https://drpress.org/ojs/index.php/HBEM/article/view/16482/16008>
 - [7] Darmayanti, M. Yusuf Qamaruddin, and W. Rajiman, "Pengaruh Modal, Pengalaman Kerja Dan Harga Jual Terhadap Pendapatan Petani Rumput Laut di Dusun Ponrang Kecamatan Ponrang Kabupaten Luwu," SEIKO: Journal of Management & Business, vol. 6, no. 1, pp. 830–838, 2022. [Online]. Available: <https://www.journal.stieamkop.ac.id/index.php/seiko/article/view/4060/0>
 - [8] S. H. Uge, T. H. Amaliah, and M. Mahmud, "Pengaruh Modal Usaha, Biaya Dan Jenis Kelamin Terhadap Pendapatan Pedagang Pasar Tradisional," Jambura Accounting Review, vol. 3, no. 1, pp. 37–51, 2022. [Online]. Available: <https://jar.fe.ung.ac.id/index.php/jar/article/view/46>
 - [9] R. Madingo, U. Moonti, R. Hafid, F. Bumulo, and M. Mahmud, "Analisis Faktor-Faktor Yang Mempengaruhi Pendapatan Nelayan Tangkap," JOURNAL of ECONOMIC and BUSINESS EDUCATION, vol. 1, no. 1, 2023. [Online]. Available: <https://ejurnal.ung.ac.id/index.php/JEBE/index>
 - [10] U. Usman and M. Yanti, "ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI PENDAPATAN PETANI PADI WANITA DI KECAMATAN SAMUDERA KABUPATEN ACEH UTARA," Jurnal Ekonomi Pertanian Unimal, vol. 03, no. 01, 2020. [Online]. Available: <https://ojs.unimal.ac.id/index.php/JEPU/article/view/3175>
 - [11] Bella, Laura La, Becoming a Data Engineer. The Rosen Publishing Group, Inc, 2017, 2017.
 - [12] R. Pangesti, "PENGARUH MODAL, TINGKAT PENDIDIKAN DAN TEKNOLOGI TERHADAP PENDAPATAN USAHA MIKRO KECIL DAN MENENGAH (UMKM)," 2022.
 - [13] S. Hidayah, "PENGARUH MODAL, LAMA USAHA, DAN KOMPETENSI WIRAUSAHA TERHADAP PENDAPATAN PEDAGANG DI PASAR CEGER KOTA TANGERANG SELATAN," 2022.
 - [14] T. Hidayat, R. Darnis, and D. Hidayatussa'adah, "ALGORITMA REGRESI LINIER BERGANDA UNTUK ANALISIS EFISIENSI STOK PRODUK DI PT. MADU PRAMUKA BATANG," Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 3, pp. 2747–2754, Aug. 2024, doi: 10.23960/jitet.v12i3.4899.