Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2022) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Computer Science and Computational Intelligence 2022

Analisis performa ordinary least square dan random forest dalam prediksi gaji

Muhammad Rashya Chudlari[a], Reosta Bayu Pratama Pane[a], Sunarko Salim[a], Syarifah Diana Permai[a]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

Penelitian ini bertujuan untuk membandingkan performa metode Ordinary Least Square (OLS) dan Random Forest dalam konteks prediksi gaji. Penelitian ini menggunakan dataset yang dikumpulkan dari website Kaggle. Dataset tersebut terdiri dari 15 Variabel, tetapi hanya 5 Variabel yang digunakan dalam pengujian metode OLS dan Random Forest. Variabel yang digunakan meliputi C, EXP, EDU, EMP, dan Salary. Dalam penelitian ini hasil analisis korelasi menunjukkan adanya hubungan positif/negatif yang kuat antara beberapa variable independent dengan variable dependent. Meskipun terdapat beberapa variable yang tidak memiliki korelasi yang signifikan. Perbandingan performa metode OLS dengan metode Random Forest menggunakan metrik evaluasi seperti Root Mean Squared Error (RMSE), Mean Absolute Deviation (MAD) dan Mean Absolute Percentage Error (MAPE). Hasil perbandingan performa kedua metode tersebut akan memberikan gambaran tentang kemampuan keduanya dalam memprediksi gaji dan menangani kompleksitas data.

Peer-review under responsibility of the scientific committee of the 7th International Conference on Computer Science and Computational Intelligence 2022
*Keywords:*OLS; random forest; regresi linear; prediksi

1. Introduction

Prediksi gaji merupakan salah satu topik yang penting dalam bidang ekonomi, analisis data, dan sumber daya manusia. Manfaat dari kemampuan prediksi dapat memberikan dampak yang signifikan bagi organisasi ataupun individu. Dalam upaya mendapatkan hasil yang lebih baik, berbagai macam model dan metode statistik dikembangkan. Metode-metode ini dapat menganalisis berbagai faktor yang mempengaruhi gaji seseorang, misalnya faktor pendidikan, pengalaman, jabatan, jam kerja, lokasi, industry, dan lain sebagainya.

Teknik Prediksi dapat dilakukan dengan berbagai macam metode, terdapat dua metode yang sering digunakan yaitu Ordinary Least Square (OLS) dan Random Forest [1,2]. OLS adalah metode statistik yang telah banyak digunakan

sebagai teknik prediksi, dikarenakan hasilnya mudah untuk diinterpretasikan. Metode ini mencoba menemukan garis regresi terbaik dengan cara meminimalisir jumlah kuadrat error untuk menggambarkan hubungan antara variable independent dan dependent [3].

Selain itu, Random Forest adalah metode yang menggunakan konsep ensemble learning, yang dapat menghasilkan prediksi yang lebih akurat dengan cara menggabungkan beberapa pohon keputusan [4]. Metode ini juga dapat menangani hubungan yang non-linear, dan ketidakteraturan dalam data lebih baik dibandingkan dengan metode OLS [5]. Namun, meskipun metode Random Forest memiliki potensi yang lebih baik, tidak selalu menjamin metode ini akan memberikan hasil yang lebih baik dalam semua situasi. Oleh karena itu, perlu dilakukan penelitian untuk menganalisis perbandingan performa metode OLS dan Random Forest dalam konteks prediksi.

Penelitian ini bertujuan untuk membandingkan performa antara metode OLS dan Random Forest dalam konteks prediksi gaji. Kami akan menguji kedua metode tersebut dengan dataset yang dikumpulkan, dengan cara analisis kinerja keduanya menggunakan beberapa metrik seperti RMSE, MAD dan MAPE. Dengan begitu dapat digambarkan kehandalan kedua metode tersebut dalam memprediksi dan kemampuannya menangani kompleksitas data.

2. Linear Regression

Model regresi linear mencoba untuk menghasilkan garis regresi antara variable independent (X) dan variable dependent (Y). Untuk menggambarkan hubungan dua variable tersebut, regresi linear menggunakan persamaan garis lurus [6]. Dengan menggunakan garis regresi model ini dapat melakukan prediksi nilai X dan Y. Disisi lain terdapat metode pemodelan regresi linear antara variable Y dengan dua atau lebih variable X, metode ini disebut dengan regresi linear ganda [3]. Rumus regresi linear ganda umumnya dinyatakan sebagai berikut:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + \varepsilon \quad (1)$$

$$Y = X\beta + \varepsilon \quad (2)$$

Keterangan,

- Y = variable dependent
- X = variable independent
- β = koefiesien regresi yang harus diestimasi untuk masing-masing variable independent
- ε = kesalahan acok yang mewakili faktor-faktor yang tidak dapat dijelaskan oleh variabel independent

OLS digunakan untuk mengestimasi parameter regresi (β) dalam model regresi linear. Dengan menerapkan rumus matematis di atas, metode OLS akan melakukan perhitungan koefisien regresi, untuk memberikan perhitungan terbaik untuk parameter regresi. Metode ini memiliki batasan yaitu asumsi yang harus dipenuhi seperti linieritas, independensi, normalitas, dan homoskedastisitas [7].

3. Random Forest

Random Forest adalah salah satu metode dari machine learning yang dapat digunakan untuk regresi maupun klasifikasi. Berbeda dengan metode OLS, Random Forest tidak menggunakan pendekatan yang linear, tetap memisahkan data (split) berdasarkan variable predictor [8].

Metode ini menggunakan teknik *boostrap aggregating* sampling untuk membuat setiap pohon keputusan, dimana menggunakan subset acak dari training data untuk membangun setiap pohon. Selain itu, metode ini juga menggunakan pemilihan variable acak. Setiap kali dilakukan split pada simpul dalam pohon keputusan, hanya sebagian variabel prediktor yang diambil untuk estimasi [9].

Jadi, rumus dari metode Random Forest adalah proses menghasilkan pohon keputusan dengan cara split data berdasarkan variable predictor dan penggabungan prediksi dari pohon-pohon yang berbeda.

4. Methodology

Penelitian ini menggunakan sumber data yang didapatkan dari website Kaggle [10]. Terdapat 15 Variabel di dalam dataset, namun hanya 5 variabel yang digunakan untuk pengujian metode OLS dan Random Forest, berikut tabel dari Variabel tersebut:

Table 1. Variabel Pengujian.

Variabel	Tipe	Karakteristik	Keterangan
C	Object	Ordinal	Career level such as manager, CEO, etc
EXP	Object	Ordinal	Experience required for applicants
EDU	Object	Ordinal	Education required for applicants
EMP	Object	Nominal	Full-time, part-time, or internship
Salary	Float	Ratio	Salary offered by the company

Berdasarkan tabel 1, penelitian ini menggunakan empat variable indenpendent (C, EXP, EDU, EMP) dan satu variable dependent (Y). Adapun keterangan variable sebagai berikut:

- C = Career Level
- EXP = Experience Level
- EDU = Education Level
- EMP = Employment Type
- Y = Salary

Adapun keterangan atribut dari 5 variabel pada tabel 1:

Table 2. Atribut Pengujian.

Variable	Atribut
C	Supervisor/Koordinator
C	Pegawai (non-manajemen & non-supervisor)
C	Manajer/Asisten Manajer
C	Lulusan baru/Pengalaman kerja kurang dari 1 tahun
C	CEO/GM/Direktur/Manajer Senior
EXP	4
EXP	5
EXP	2
EXP	1
EXP	3
EXP	10
EXP	7
EXP	8
EXP	15
EXP	6
EXP	18
EXP	20
EXP	17
EXP	12
EXP	9
EDU	S1
EDU	S1 - S2

EDU	D3 - S2
EDU	SMA/SMU
EDU	SMA/SMU - S1
EDU	Tidak terspesifikasi
EDU	D3 - S1
EDU	S1 - S3
EDU	S3
EDU	SMA/SMU - S2
EMP	Penuh Waktu
EMP	Kontrak
EMP	Paruh Waktu
EMP	Temporer
EMP	Penuh Waktu, Kontrak
EMP	Penuh Waktu, Paruh Waktu
EMP	Kontrak, Temporer

Data tersebut akan dianalisis menggunakan model regresi linear dengan metode OLS dan Random Forest. Penelitian ini dapat mengidentifikasi kelebihan dan kekurangan masing-masing metode menggunakan metrik evaluasi, meliputi RMSE, MAD dan, MAPE.

Berikut adalah penjelasan dari metode yang telah diselesaikan.

Proses dalam memprediksi menggunakan metode OLS dan Random Forest

1 Import Data Dan Read Data
2 Cleaning Data
3 Discovering Data
4 Data Preprocessing
5 Pemodelan Data
6 Evaluasi

5. Experimentation

Tahap pertama dilakukan dalam proses experimentation adalah cleaning data yang bertujuan untuk membersihan data dari kesalahan, dan ketidakakuratan sehingga data dapat menjadi lebih reliabel untuk digunakan dalam pemodelan.

Berikut tahapan-tahapan yang dilakukan dalam cleaning data:

1. Menghapus kolom tabel yang tidak diperlukan untuk tujuan penelitian. Penelitian ini menggunakan 5 variabel untuk pengujian model regresi.

```
dataset = dataset.drop(labels=["id", "salary_currency", "job_benefits", "job_function", "job_description",
"company_process_time", "company_size", "company_industry", "location"], axis=1)
```

Fig. 1. Menghilangkan variable yang tidak diperlukan

2. Mengisi missing value untuk variabel Y dengan median dikarenakan terdapat outlier. Berikut gambar dari data outlier variable Y.

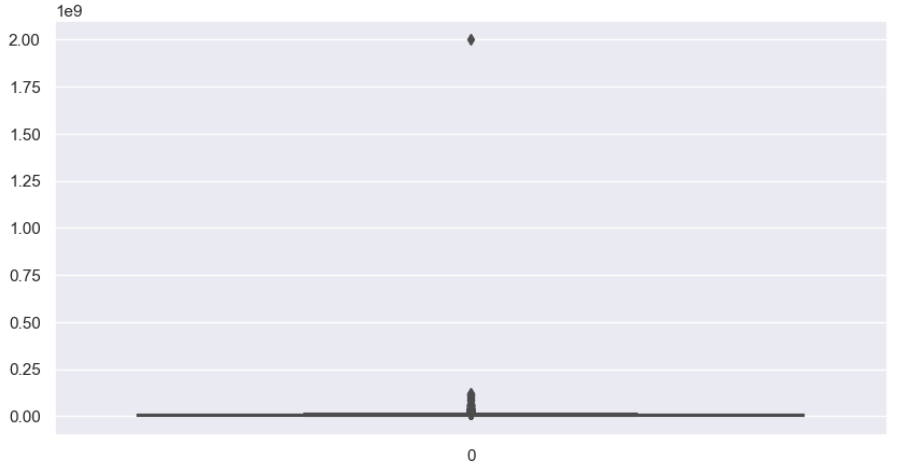

Fig. 2. Data outlier variable salary

3. Melakukan mapping data untuk variabel EDU agar nilainya tidak terlalu panjang. Metode untuk mapping tersebut menggunakan algorimat selection sederhana (IF-ELSE).
4. Menghapus duplikasi data untuk menghindari bias dalam melakukan pengujian model regresi dan agar menghindari estimasi yang tidak akurat.

```
print("Duplikasi Sebelum Di Drop: ", dataset.duplicated().sum())
dataset.drop_duplicates(keep='first', inplace=True)
print("Duplikasi Setelah Di Drop: ", dataset.duplicated().sum())
```

```
Duplikasi Sebelum Di Drop:  5105
Duplikasi Setelah Di Drop:  0
```

Fig. 3. Penghapusan data duplikasi

5. Menghapus data outlier untuk variabel Y menggunakan metode IQR

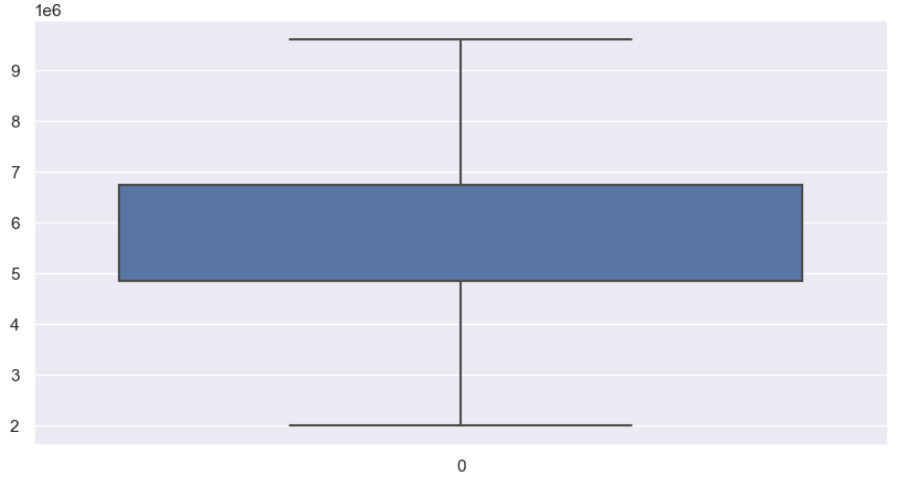

Fig. 4. Visualisasi data menghilangkan outlier menggunakan metode IQR.

Selanjutnya proses discovering data yang bertujuan mengelompokkan data kategorik dan numerik. Berikut hasil dari discovering data:

Table 3. Data kategorik dan numerik.

	C	EXP	EDU	EMP	Y
Kategorik	✓	✓	✓	✓	-
Numerik	-	-	-	-	✓

Kemudian tahap data preprocessing. Langkah pertama pada tahapan ini adalah melakukan split data yang bertujuan untuk membagi data menjadi dua yaitu data training dan data testing. Dengan melakukan split data, kita dapat menguji dan mengevaluasi kinerja dari model dan untuk mendapatkan estimasi yang akurat. Berdasarkan table 1, semua variabel indenpenden merupakan data kategorik. Maka dilakukan proses encoding data kategorik menjadi representasi biner, agar dapat digunakan dalam pemodelan machine learning. Hasil dari konversi data kategorik tersebut maka didapatkan 36 variabel independen.

```
enc_1 = OneHotEncoder(sparse_output=False)

X_train_career = enc_1.fit_transform(X_train[['career_level']])
X_test_career = enc_1.transform(X_test[['career_level']])
X_train_career
```

```
array([[0., 0., 0., 1., 0.],
       [0., 0., 0., 1., 0.],
       [0., 0., 0., 1., 0.],
       ...,
       [0., 0., 0., 1., 0.],
       [0., 0., 0., 1., 0.],
       [0., 0., 0., 0., 1.]])
```

Fig. 5. Proses encoding variable bersifat kategorik.

Setelah proses encoding variable dilakukan, maka semua variable tersebut digabungkan kembali menjadi satu table. Langkah pertama yang dilakukan adalah me-reset index semua variable agar memiliki posisi yang sama, kemduian menggabungkannya menggunakan function concat dan untuk variable independent yang sebelumnya memiliki tipe data string dapat dihilangkan.

```
X_train.reset_index(drop=True, inplace=True)
X_test.reset_index(drop=True, inplace=True)

y_train.reset_index(drop=True, inplace=True)
y_test.reset_index(drop=True, inplace=True)

X_train_new = pd.concat([X_train, X_train_career, X_train_experience, X_train_employment, X_train_education], axis=1)
X_train_new.drop(columns=['career_level', 'experience_level', 'employment_type', 'education_level'],inplace=True)

X_test_new = pd.concat([X_test, X_test_career, X_test_experience, X_test_employment, X_test_education], axis=1)
X_test_new.drop(columns=['career_level', 'experience_level', 'employment_type', 'education_level'],inplace=True)

X_train_new.head()
```

	career_level_CEO/GM /Direktur/Manajer Senior	career_level_Lulusan baru/Pengalaman kerja kurang dari 1 tahun	career_level_Manajer/Asisten Manajer	career_level_Pegawai (non-manajemen & non-supervisor)	career_level_Supervisor/Koordinator	e
0	0.0	0.0	0.0	1.0	0.0	
1	0.0	0.0	0.0	1.0	0.0	
2	0.0	0.0	0.0	1.0	0.0	
3	0.0	0.0	0.0	1.0	0.0	
4	0.0	0.0	0.0	1.0	0.0	

5 rows × 36 columns

Fig. 6. Penggabungan variable kategorik.

6. Result and Discussions

Perbandingan performa antara metode OLS dan Random Forest dalam prediksi, diawali dengan melakukan pemodelan prediksi menggunakan metode OLS dengan library statsmodels.api. Berikut tabel dari pemodelan regresi linear berganda menggunakan metode OLS.

Table 4. Regresi linear menggunakan metode OLS.

Variable	Estimate	T value	p-value
const	3.872e+06	32.825	0.000
C_CEO/GM/Direktur/Manajer Senior	2.316e+06	8.479	0.000
C_Lulusan baru/Pengalaman kerja kurang dari 1 tahun	-1.031e+06	-14.184	0.000
C_Manajer/Asisten Manajer	1.661e+06	22.226	0.000
C_Pegawai (non-manajemen & non-supervisor)	-3.3e+05	-5.232	0.000
C_Supervisor/Koordinator	1.257e+06	19.786	0.000
EXP_1	2.208e+05	1.907	0.057
EXP_10	7.404e+05	4.290	0.000
EXP_12	-3.161e+05	-0.539	0.590
EXP_15	3.01e+05	0.623	0.533
EXP_17	-2.412e+06	-2.938	0.003
EXP_18	2.784e+05	0.339	0.734
EXP_2	3.418e+05	2.953	0.003
EXP_20	9.921e+05	2.050	0.040
EXP_3	4.681e+05	4.034	0.000
EXP_4	4.749e+05	3.894	0.000
EXP_5	5.647e+05	4.794	0.000
EXP_6	3.965e+05	2.358	0.018
EXP_7	4.757e+05	2.654	0.008
EXP_8	7.841e+05	3.986	0.000
EXP_9	5.611e+05	0.684	0.494
EMP_Kontrak	7.607e+05	7.601	0.000
EMP_Paruh Waktu	5.339e+05	4.213	0.000
EMP_Penuh Waktu	7.802e+05	7.871	0.000
EMP_Penuh Waktu, Kontrak	7.914e+05	4.424	0.000
EMP_Penuh Waktu, Paruh Waktu	2.376e+05	0.452	0.651
EMP_Temporer	7.682e+05	5.643	0.000
EDU_D3 - S1	1.807e+05	3.629	0.000
EDU_D3 - S2	2.996e+05	6.721	0.000
EDU_S1	4.181e+05	9.486	0.000
EDU_S1 - S2	4.711e+05	9.338	0.000
EDU_S1 - S3	6.044e+05	3.980	0.000
EDU_S3	9.922e+05	3.535	0.000
EDU_SMA/SMU	-2.44e+04	-0.491	0.623
EDU_SMA/SMU - S1	1.388e+05	2.985	0.003
EDU_SMA/SMU - S2	5.42e+05	2.253	0.024
EDU_Tidak terspesifikasi	2.495e+05	5.446	0.000

Berdasarkan metode OLS diatas menunjukkan dampak yang signifikan dan kurang signifikan dari setiap variable independent terhadap variable terikat. Terdapat 7 variable yang kurang signifikan berdasarkan p-value, yaitu X6, X8, X9, C1, EXP0, EXP5 dan EDU3. Jadi, hasil dari regresi model berganda dari OLS sebagai berikut:

Y = (3.872e+06) + (2.316e+06)C_CEO/GM/Direktur/Manajer Senior + (-1.031e+06)C_Lulusan baru/Pengalaman kerja kurang dari 1 tahun + (1.661e+06)C_Manajer/Asisten Manajer + (-3.3e+05)C_Pegawai (non-manajemen & non-supervisor) + (1.257e+06)C_Supervisor/Koordinator + (2.208e+05)EXP_1 + (7.404e+05)EXP_10 + (-3.161e+05)EXP_12 + (3.01e+05)EXP_15 + (-2.412e+06)EXP_17 + (2.784e+05)EXP_18 + (3.418e+05)EXP_2 + (9.921e+05)EXP_20 + (4.681e+05)EXP_3 + (4.749e+05)EXP_4 + (5.647e+05)EXP_5 + (3.965e+05)EXP_6 + (4.757e+05)EXP_7 + (7.841e+05)EXP_8 + (5.611e+05)EXP_9 + (7.607e+05)EMP_Kontrak + (5.339e+05)EMP_Paruh Waktu + (7.802e+05)EMP_Penuh Waktu + (7.914e+05)EMP_Penuh Waktu, Kontrak + (2.376e+05)EMP_Penuh Waktu, Paruh Waktu + (7.682e+05)EMP_Temporer + (1.807e+05)EDU_D3 - S1 + (2.996e+05)EDU_D3 - S2 + (4.181e+05)EDU_S1 + (4.711e+05)EDU_S1 - S2 + (6.044e+05)EDU_S1 - S3 + (9.922e+05)EDU_S3 + (-2.44e+04)EDU_SMA/SMU + (1.388e+05)EDU_SMA/SMU - S1 + (5.42e+05)EDU_SMA/SMU - S2 + (2.495e+05)EDU_Tidak terspesifikasi

Hasil diatas menunjukkan adanya ketidaknormalan dalam data residu, dan model memiliki VIF lebih dari 1, yang menunjukkan adanya multikolinieritas dalam model regresi. Berdasarkan hasil tersebut, metode OLS tidak memenuhi asumsi dari regresi linear pada saat di implementasi pada model regresi linear berganda. Oleh karena itu, dapat dilanjutkan dengan menggunakan metode regresi lain yaitu Random Forest.

Table 5. Regresi linear menggunakan metode Random Forest.

Variable	Estimate
C_CEO/GM/Direktur/Manajer Senior	2.2619e-03
C_Lulusan baru/Pengalaman kerja kurang dari 1 tahun	2.2530e-02
C_Manajer/Asisten Manajer	1.1911e-01
C_Pegawai (non-manajemen & non-supervisor)	2.7523e-03
C_Supervisor/Koordinator	7.4665e-01
EXP_1	2.1224e-02
EXP_10	1.1419e-03
EXP_12	2.2316e-04
EXP_15	2.0076e-06
EXP_17	6.9214e-04
EXP_18	1.8336e-06
EXP_2	9.9795e-03
EXP_20	2.7125e-04
EXP_3	3.3713e-03
EXP_4	2.8944e-03
EXP_5	3.1220e-03
EXP_6	1.2034e-04
EXP_7	9.1522e-04
EXP_8	8.9109e-04
EXP_9	3.2688e-07
EMP_Kontrak	5.9797e-03
EMP_Paruh Waktu	1.5481e-03
EMP_Penuh Waktu	5.2655e-03

EMP_Penuh Waktu, Kontrak	9.5872e-05
EMP_Penuh Waktu, Paruh Waktu	6.1242e-05
EMP_Temporer	9.6294e-04
EDU_D3 - S1	2.6937e-03
EDU_D3 - S2	3.9693e-03
EDU_S1	1.1453e-02
EDU_S1 - S2	6.7512e-03
EDU_S1 - S3	1.1998e-03
EDU_S3	6.8067e-04
EDU_SMA/SMU	8.7703e-03
EDU_SMA/SMU - S1	6.0996e-03
EDU_SMA/SMU - S2	8.4119e-04
EDU_Tidak terspesifikasi	5.4738e-03

Jadi, hasil dari regresi model berganda dari Random Forest sebagai berikut:

Y = (2.2619e-03)C_CEO/GM/Direktur/Manajer Senior + (2.2530e-02)C_Lulusan baru/Pengalaman kerja kurang dari 1 tahun + (1.1911e-01)C_Manajer/Asisten Manajer + (2.7523e-03)C_Pegawai (non-manajemen & non-supervisor) + (7.4665e-01)C_Supervisor/Koordinator + (2.1224e-02)EXP_1 + (1.1419e-03)EXP_10 + (2.2316e-04)EXP_12 + (2.0076e-06)EXP_15 + (6.9214e-04)EXP_17 + (1.8336e-06)EXP_18 + (9.9795e-03)EXP_2 + (2.7125e-04)EXP_20 + (3.3713e-03)EXP_3 + (2.8944e-03)EXP_4 + (3.1220e-03)EXP_5 + (1.2034e-04)EXP_6 + (9.1522e-04)EXP_7 + (8.9109e-04)EXP_8 + (3.2688e-07)EXP_9 + (5.9797e-03)EMP_Kontrak + (1.5481e-03)EMP_Paruh Waktu + (5.2655e-03)EMP_Penuh Waktu + (9.5872e-05)EMP_Penuh Waktu, Kontrak + (6.1242e-05)EMP_Penuh Waktu, Paruh Waktu + (9.6294e-04)EMP_Temporer + (2.6937e-03)EDU_D3 - S1 + (3.9693e-03)EDU_D3 - S2 + (1.1453e-02)EDU_S1 + (6.7512e-03)EDU_S1 - S2 + (1.1998e-03)EDU_S1 - S3 + (6.8067e-04)EDU_S3 + (8.7703e-03)EDU_SMA/SMU + (6.0996e-03)EDU_SMA/SMU - S1 + (8.4119e-04)EDU_SMA/SMU - S2 + (5.4738e-03)EDU_Tidak terspesifikasi

Langkah selanjutnya membandingkan antara hasil dari metode OLS dan Random Forest menggunakan parameter estimasi:

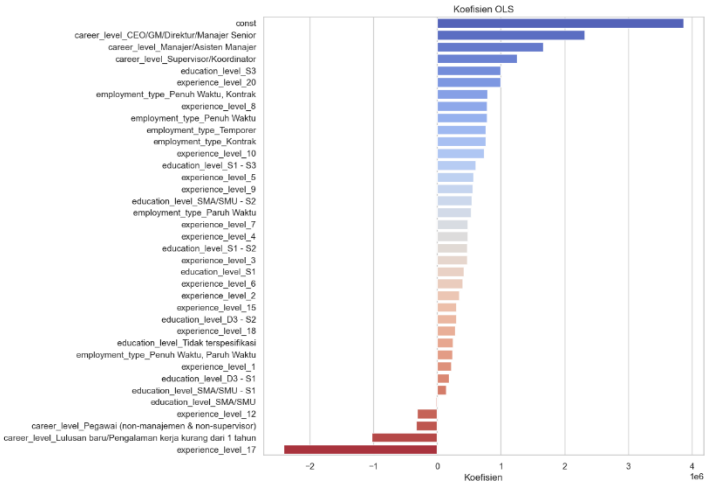

Fig. 7. Parameter estimasi metode OLS

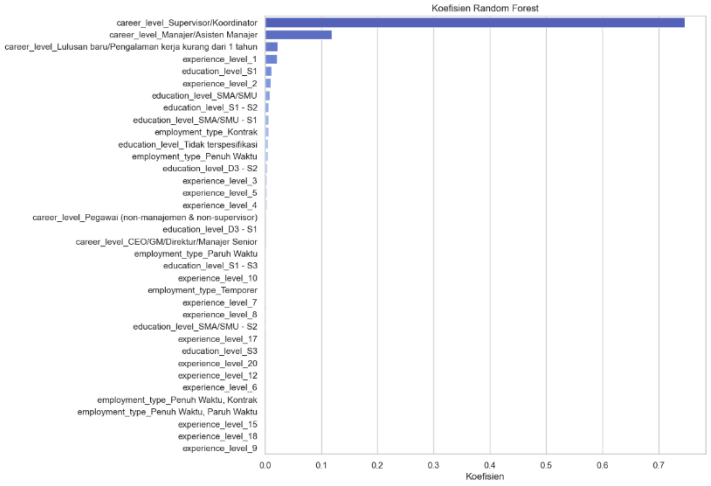

Fig. 8. Parameter estimasi metode Random Forest

Jika dibandingkan dengan metode OLS, semua parameter estimasi dari metode Random Forest menunjukkan hasil yang positif. Hal tersebut dikarenakan adanya kompleksitas yang berbeda antara model Random Forest dan OLS, serta metode estimasi yang digunakan oleh masing-masing model. Random Forest menggunakan kumpulan pohon keputusan yang independen untuk melakukan prediksi. Dengan cara kerja setiap pohon secara individu memperhitungkan sejumlah fitur yang acak dan memberikan kontribusi terhadap prediksi. Sementara, OLS menggunakan pendekatan yang linear untuk memperkirakan parameter yang optimal.

Langkah berikutnya adalah melakukan analisis terhadap performa antara metode OLS dan Random Forest dalam konteks prediksi menggunakan kriteria RMSE, MAD, dan MAPE.

Table 6. Perbandingan training data antara metode OLS dan Random Forest.

Metode	RMSE	MAD	MAPE
OLS	868964.6646	475091.8862	9.73%
Random Forest	849614.0182	462353.3993	9.49%

Table 7. Perbandingan testing data antara metode OLS dan Random Forest.

Metode	RMSE	MAD	MAPE
OLS	890542.7937	490964.6607	10.07%
Random Forest	900999.7926	490678.4223	10.03%

Pada tabel 6 dan 7 menunjukkan bahwa kriteria RMSE, MAD, dan MAPE dari metode Random Forest menghasilkan prediksi yang lebih baik dibandingkan dengan metode OLS. Meskipun hasil dari kedua metode tersebut tidak terlalu jauh berbeda, namun metode Random Forest memiliki kinerja yang lebih baik dalam memprediksi nilai terikat dalam kasus prediksi gaji. Dikarenakan terdapat beberapa variabel yang kurang signifikan, maka dilakukan tes ulang menggunakan variable yang signifikan. Berikut hasil dari tes ulang pemodelan OLS dan Random Forest.

Table 8. Tes ulang training data antara metode OLS dan Random Forest.

Metode	RMSE	MAD	MAPE
OLS	887711.2936	427037.1831	8.87%
Random Forest	887711.4940	427237.3657	8.87%

Table 9. Tes ulang testing data antara metode OLS dan Random Forest.

Metode	RMSE	MAD	MAPE
OLS	910457.7893	441240.1834	9.17%
Random Forest	910485.1872	441435.2614	9.17%

Pada tabel 8 dan 9 menunjukkan bahwa hasil dari tes ulang dari metode OLS dan Random Forest hampir sama. Dengan begitu, dapat disimpulkan bahwa metode Random Forest mampu memprediksi lebih akurat data yang kompleks dibandingan metode OLS.

7. Conclusion

Dalam perbandingan performa antara metode OLS dan Random Forest untuk kasus prediksi gaji, metode Random Forest memberikan hasil yang sedikit lebih baik untuk data yang kompleks berdasarkan kriteria RMSE, MAD, dan MAPE. Meskipun terdapat variable independent yang kurang signifikan, metode Random Forest dapat memprediksi lebih akurat walaupun tidak menggunakan pendekatan yang linear seperti OLS. Dan dapat disimpulkan bahwa model regresi linear menggunakan Random Forest lebih baik dari pada metode OLS.

Acknowledgements

Penulis ingin berterima kasih kepada Universitas Bina Nusantara, Jakarta, Indonesia, atas dukungan dalam penelitian ini.

References

1. Frost J. Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models. 1st ed. Statistics By Jim Publishing; 2019.
2. Williams AS. Data Analytics for Beginners: Introduction to Data Analytics. Anthony S. Williams; 2021.
3. Id ID. Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python. 1st ed. UR PRESS; 2021.
4. Sullivan W. Python Machine Learning Illustrated Guide For Beginners & Intermediates: The Future Is Here! CreateSpace Independent Publishing Platform; 2018.
5. Luminous T. Machine Learning For Beginners Guide Algorithms: Supervised & Unsupervsied Learning. Decision Tree & Random Forest Introduction. Healthy Pragmatic Solutions Inc; 2017.
6. Venkatesan R, Wilcox RT, Farris PW. Marketing Analytics: Essential Tools for Data-Driven Decisions. University of Virginia Press; 2021.
7. Gilchrist A. Machine Learning: Adaptive Behaviour Through Experience: Thinking Machines. Alasdair Gilchrist; 2017.
8. Serrano LG. Grokking Machine Learning. Manning Publications; 2021.
9. Bari A, Chaouchi M, Jung T. Predictive Analytics For Dummies, 2nd Edition. 2nd ed. John Wiley & Sons, Inc; 2017.
10. Wibowo CP. Jog Description and Salary in Indonesia [Internet]. 2022 [cited 2023 Jun 10]. Available from: https://www.kaggle.com/datasets/canggih/jog-description-and-salary-in-indonesia