JCBD
JOURNAL OF COMPUTERS AND DIGITAL BUSINESS
E-ISSN : 2830 - 3121
Vol. 3, No.2, May, 2024, pp. 86-97

Komparasi *Multiple Linear Regression* dan *Random Forest Regression* Dalam Memprediksi Anggaran Biaya Manajemen Proyek Sistem Informasi

Farhanuddin[1*], Sarah Ennola Karina Sihombing[2], Yahfizham[3]
[1,3] Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[2] Universitas Brawijaya, Malang, Indonesia

[*] uddinfarhan144@gmail.com

DOI : 10.56427/jcbd.v3i2.408

INFO ARTIKEL

Histori Artikel
Diterima : 30 April 2024
Ditinjau : 11 Mei 2024
Disetujui : 29 Mei 2024

Kata Kunci
Machine Learning
Multiple Linear Regression
Random Forest Regression
Anggaran Biaya
Manajemen Proyek

Keywords
Machine Learning
Multiple Linear Regression
Random Forest Regression
Budget
Project management

ABSTRAK

Dalam dunia bisnis yang dinamis dengan persaingan yang semakin ketat, manajemen proyek kini menjadi kunci sukses, terutama dalam pengembangan sistem informasi. Memprediksi anggaran biaya merupakan aspek penting dalam manajemen proyek, dengan mengetahui perkiraan anggaran biaya yang akurat, perusahaan dapat membuat keputusan yang lebih tepat terkait dengan alokasi sumber daya dan pengelolaan keuangan proyek. Dengan kemajuan teknologi, *machine learning* menjadi solusi potensial untuk meningkatkan akurasi dan efisiensi pengelolaan anggaran. Penelitian ini bertujuan untuk menemukan model *machine learning* yang paling akurat dalam memprediksi anggaran biaya dengan menggunakan dataset dari *platform* kaggle.com. Membandingkan algoritma *machine learning multiple linear regression* (MLR) dan *random forest regression* (RFR) dilakukan sebagai langkah untuk mencapai tujuan tersebut. Pendekatan deskriptif kuantitatif digunakan dengan melakukan *Exploratory Data Analysis* (EDA) untuk mengidentifikasi pola dalam data. Hasilnya menunjukkan bahwa model *random forest regression* memiliki akurasi lebih tinggi, mencapai 81,6% dibandingkan *multiple linear regression*. Kesimpulannya, penggunaan *random forest regression* lebih efektif dalam memprediksi anggaran biaya proyek sistem informasi. Ini menandakan bahwa random forest regression dapat menjadi pilihan yang lebih baik untuk menghadapi kompleksitas dan ketidakpastian dalam manajemen proyek sistem informasi.

In the dynamic business world with increasingly fierce competition, project management has become the key to success, especially in information system development. Predicting project budget is a critical aspect of project management, especially in information system development. By knowing accurate budget estimates, companies can make better decisions regarding resource allocation and project financial management. With technological advancements, machine learning has emerged as a potential solution to improve the accuracy and efficiency of budget management. This study aims to find the most accurate machine learning model for predicting project budget using a dataset from the Kaggle.com platform. A comparison of the machine learning algorithms multiple linear regression (MLR) and random forest regression (RFR) is conducted as a step to achieve this objective. A quantitative descriptive approach is employed using Exploratory Data Analysis (EDA) to identify patterns in the data. The results show that the random forest regression model has higher accuracy, reaching 81.6% compared to multiple linear regression. In conclusion, using random forest regression is more effective in predicting the budget of information system projects. This indicates that random

forest regression can be a better choice for dealing with the complexity and uncertainty in information system project management.

1. Pendahuluan

Dunia bisnis saat ini ditandai dengan perubahan yang cepat dan persaingan yang semakin ketat, karena itu manajemen proyek menjadi kunci utama dalam memastikan kesuksesan pelaksanaan proyek, terutama dalam pengembangan sistem informasi [1]. Salah satu aspek yang sangat penting dalam manajemen proyek adalah pengelolaan anggaran biaya. Pengelolaan anggaran biaya menjadi landasan untuk mengendalikan sumber daya finansial yang tersedia dan mengantisipasi kemungkinan biaya tambahan yang mungkin timbul selama proses proyek. Akurasi dan efisiensi pengelolaan anggaran biaya sangatlah penting untuk memastikan kelancaran proyek dan mencapai tujuan yang telah ditetapkan.

Seiring perkembangan teknologi, machine learning telah muncul sebagai alat yang berpotensi kuat untuk meningkatkan akurasi dan efisiensi pengelolaan anggaran biaya. Machine learning adalah bidang ilmu komputer yang berfokus pada pengembangan algoritma yang memungkinkan komputer untuk belajar dari data tanpa diprogram secara eksplisit. Dengan menggunakan pendekatan machine learning yang memiliki kemampuan prediktif sangat tinggi, memungkinkan analisis data historis untuk menemukan pola dan tren yang tidak terlihat oleh metode analisis tradisional, sehingga menghasilkan prediksi biaya yang lebih akurat di masa depan [2]. Namun, di antara berbagai algoritma machine learning yang tersedia, pertanyaan muncul: algoritma mana yang paling efektif dalam memprediksi anggaran biaya pada manajemen proyek sistem informasi?

Penelitian ini bertujuan untuk menjawab pertanyaan tersebut dengan membandingkan dua algoritma yang populer, yaitu Multiple Linear Regression (MLR) dan Random Forest Regression (RFR). *Multiple linear regression* (MLR) adalah algoritma machine learning supervised learning yang digunakan untuk memprediksi nilai numerik berdasarkan satu atau lebih variabel independen. Algoritma ini bekerja dengan membangun model linier yang menghubungkan variabel independen dengan variabel dependen [3]. MLR dipilih karena kesederhanaannya dan kemampuannya dalam menangani hubungan linear. Sedangkan, *Random Forest Regression* (RFR) adalah algoritma machine learning ensemble learning yang digunakan untuk memprediksi nilai berdasarkan satu atau lebih variabel independen [4]. Algoritma ini bekerja dengan membangun hutan (forest) pohon keputusan (decision trees) yang acak, kemudian menggabungkan prediksi dari pohon-pohon tersebut untuk menghasilkan prediksi akhir yang lebih akurat dan stabil. RFR dipilih karena kemampuannya menangani hubungan non-linear dan kompleks, serta memberikan prediksi yang lebih robust dan akurat.

Melalui analisis data yang dilakukan, penelitian ini bertujuan untuk membangun model machine learning yang dapat memberikan prediksi yang akurat terkait dengan anggaran biaya proyek. Hasil dari penelitian ini diharapkan dapat memberikan wawasan yang berharga bagi para praktisi dan peneliti dalam memilih algoritma yang paling sesuai untuk memprediksi anggaran biaya proyek sistem informasi.

2. Metodologi Penelitian

Metode yang digunakan dalam penelitian ini adalah metode deskriptif kuantitatif. Dalam konteks ini, pendekatan kuantitatif digunakan untuk mengumpulkan dan menganalisis data terkait dengan anggaran biaya proyek sistem informasi, sehingga memungkinkan peneliti untuk memberikan pemahaman yang lebih mendalam tentang faktor-faktor yang mempengaruhi anggaran biaya proyek. Data yang digunakan dalam penelitian ini adalah data kuantitatif, Analisis deskriptif dengan pendekatan statistik diterapkan untuk mendeskripsikan karakteristik data, termasuk statistik deskriptif seperti mean, median, standar deviasi, serta visualisasi data melalui grafik dan plot untuk mengidentifikasi pola dan tren dalam data. yang digunakan untuk melakukan operasi matematika dan analisis statistik yang mendalam. Data ini diperoleh dari fakta yang ada dan memberikan dasar yang kuat untuk penelitian [5]. Penelitian ini memanfaatkan dataset yang tersedia di platform Kaggle.com, Penggunaan data dari kaggle.com memberikan keunggulan tambahan karena data tersebut telah melewati proses kurasi dan validasi, memastikan keandalannya untuk digunakan dalam penelitian [6].

Dengan menggunakan algoritma machine learning, model akan dilatih pada dataset untuk mempelajari hubungan antara variabel-variabel input dan anggaran biaya. Setelah model dilatih dan divalidasi, model ini akan digunakan untuk memprediksi anggaran biaya proyek di masa depan, khususnya untuk periode hingga 5 tahun ke depan. Pendekatan ini diharapkan dapat memberikan alat yang praktis dan akurat untuk membantu manajer proyek dalam perencanaan dan pengelolaan anggaran yang lebih efektif.

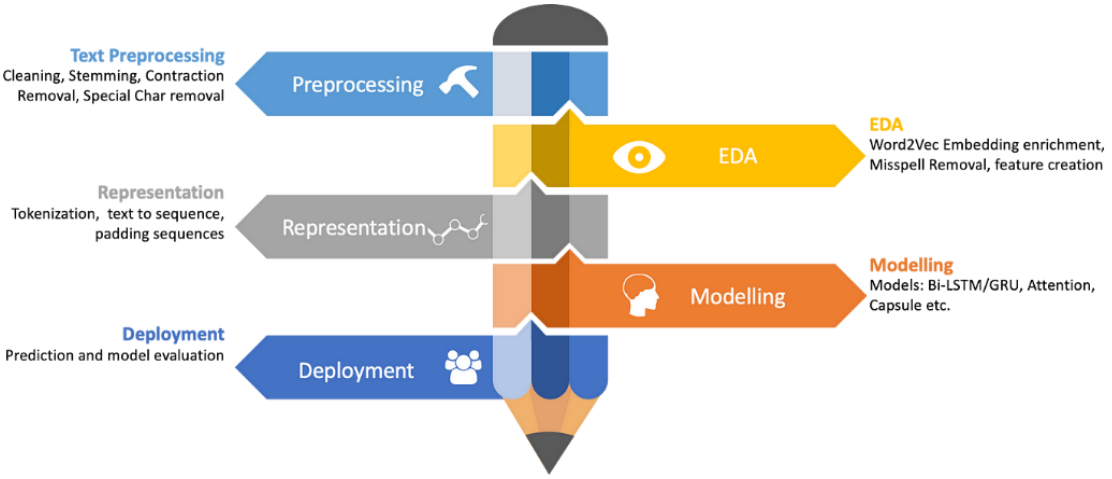
a. Teknik Pengumpulan Data

Pengumpulan data pada penelitian ini melalui beberapa tahap yaitu:

1) Pengambilan data
   Peneliti melakukan pengambilan data historis proyek sistem informasi yang diperoleh dari sumber data publik yang peneliti dapatkan melalui platform kaggle.com.
2) Studi Pustaka
   Sumber-sumber terpercaya dijadikan acuan dalam pelaksanaan metode ini, demi menjamin keandalan dan validitas data yang digunakan dalam laporan penelitian.
3) Dokumentasi
   Literatur menjadi sumber informasi dalam penelitian ini, peneliti juga memanfaatkannya untuk keperluan interpretasi dan bahkan prediksi. Dengan menganalisis dan mensintesis literatur yang relevan, peneliti dapat memperoleh pemahaman yang lebih dalam tentang konteks, teori, dan temuan terkait dengan topik penelitian.

b. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) adalah suatu pendekatan awal yang penting dalam analisis data yang bertujuan untuk menggali dan memahami struktur serta karakteristik data tanpa membuat asumsi sebelumnya [7]. Melalui proses EDA (Exploratory Data Analysis), peneliti mengeksplorasi data untuk mengidentifikasi pola, menemukan nilai-nilai yang berbeda dari biasanya (outlier), menguji hipotesis, dan memverifikasi asumsi yang mendasari analisis data. Tujuan utama dari EDA adalah untuk mendapatkan wawasan yang berharga dari data yang tersedia, sehingga memungkinkan pengguna untuk membuat keputusan yang lebih tepat dan akurat dalam analisis statistik [8].



Gambar 1. Tahapan penelitian menggunakan EDA

Analisis eksplorasi data (EDA) melibatkan penerapan berbagai teknik untuk menyelidiki data. Beberapa teknik tersebut antara lain:

1) Pre-Processing
Pre-processing adalah langkah penting dalam analisis data untuk memastikan data yang digunakan akurat, lengkap, dan terstruktur. Langkah-langkah pre-processing ini bertujuan untuk memastikan keakuratan, kebersihan, dan konsistensi data sehingga hasil analisis yang dihasilkan dapat menjadi lebih relevan dan dapat dipercaya.

2) Exploratory Data Analysis
Melalui Exploratory Data Analysis (EDA), kita dapat memahami karakteristik data dan menentukan langkah pengolahan data yang sesuai. Tahap ini berfokus pada pembersihan data dengan memeriksa data kosong, menghapus data duplikat, dan mengonversi kategori data. Langkah-langkah ini penting untuk memastikan kebersihan dan kualitas data sebelum melakukan analisis lebih lanjut.

3) Representation
Representasi data merujuk pada cara visual atau grafis untuk menyajikan informasi dari dataset. Representasi data sangat penting dalam EDA karena membantu peneliti untuk memahami struktur, pola, dan karakteristik dari data yang sedang diteliti dengan lebih baik.

4) Modelling

Modelling merupakan proses membangun model algoritma untuk menganalisis data yang telah disiapkan, dengan tujuan untuk mendapatkan kesimpulan atau prediksi. Melalui modelling, peneliti menggunakan teknik statistik atau algoritma untuk mengidentifikasi pola, hubungan, atau tren yang mungkin terdapat dalam data, dengan tujuan untuk memperoleh pemahaman yang lebih mendalam tentang fenomena yang diamati [9].

5) Deployment / Evaluation

Tahap penarikan kesimpulan memanfaatkan hasil data mining dan berbagai hipotesis yang dihasilkan untuk merumuskan kesimpulan akhir. Hasil analisis dan eksplorasi data membantu untuk mengonfirmasi atau menolak hipotesis yang diajukan, serta memberikan wawasan yang berharga tentang pola atau tren yang terdapat dalam data yang diamati.

3. Hasil dan Pembahasan

Penelitian ini memerlukan data historis proyek sistem informasi. Data ini dapat diperoleh dari platform Kaggle.com, platform berbagi data dan kode yang populer bagi para peneliti dan praktisi data. Langkah pertama dalam proses pengambilan data adalah menentukan kebutuhan data, termasuk jenis data, periode data, dan format data. Jenis data yang diperlukan dalam penelitian ini adalah data historis proyek sistem informasi, yang meliputi informasi seperti nama proyek, jenis proyek, durasi proyek, anggaran proyek, dan biaya aktual proyek. Format data yang diinginkan adalah CSV. Kata kunci yang digunakan dalam pencarian adalah "proyek sistem informasi", "anggaran biaya", "manajemen proyek", "historis proyek", "data historis". dataset yang ditemukan dibaca dengan cermat untuk memahami sumber data, variabel yang tersedia, format data, dan informasi lainnya. Memastikan dataset tersebut memenuhi kebutuhan data yang telah ditentukan sebelumnya. Tabel 1 merupakan dataset yang peniliti dapatkan dari platform Kaggle.com.

Tabel 1. Dataset dari Kaggle.com

No	Nama Project	Anggaran Biaya	Durasi Project	Kesulitan	Tim pengembang	Platform
1	Sistem Informasi Rumah Sakit	Rp. 175.000.000,00	213 Hari	Sedang	6 Orang	Web
2	Aplikasi Mobile Pembelajaran Interaktif	Rp. 200.000.000,00	243 Hari	Tinggi	7 Orang	Mobile
3	Sistem Manajemen Inventaris Sekolah	Rp. 150.000.000,00	152 Hari	Rendah	2 Orang	Desktop
4	Sistem E-Commerce	Rp. 500.000.000,00	304 Hari	Tinggi	10 orang	Web & Mobile
5	Aplikasi Pengelolaan Proyek Konstruksi	Rp. 400.000.000,00	213 Hari	Sedang	5 Orang	Web

Metode Analisis yang digunakan dalam penelitian ini adalah Exploratory Data Analysis (EDA), untuk mengeksplorasi dan memahami pola serta tren yang ada di dalam dataset [10]. Dalam Exploratory Data Analysis, visualisasi seperti histogram, box plot, dan violin plot digunakan untuk melihat distribusi data, menemukan pola dan tren, dan mendeteksi anomali. Berikut adalah proses penelitian yang menggunakan EDA.

a. Import Library

Sebelum memulai, kita perlu mengimpor library yang diperlukan untuk mengakses dan memanfaatkan modul-modul dalam bahasa Python. Rincian library yang digunakan tercantum dalam Gambar 2.

```
import pandas as pd
import matplotlib.pyplot as pit
import numpy as np
import seaborn as sns

from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_eror
from sklearn.model_selection import train_test_split
```

Gambar 2. Import Library

Berikut adalah penjelasan mengenai fungsi library yang diimpor dalam bahasa pemrograman Python pada Gambar 2:

a. Pandas adalah library yang digunakan untuk manipulasi dan analisis data.
b. Numpy berfungsi untuk melakukan operasi matematika dan manipulasi array dengan efisien dan mudah.
c. Matplotlib merupakan library Python untuk membuat visualisasi data yang informatif dan menarik.
d. Seaborn adalah library yang memperluas kemampuan visualisasi data dengan memanfaatkan struktur data dari Pandas dan plotting engine dari Matplotlib.
e. Sklearn menyediakan berbagai macam algoritma machine learning yang dapat digunakan untuk berbagai macam tugas. [11]

b. Data *Pre-processing*

Tahap ini berfokus pada penyaringan dataset proyek sistem informasi. Tujuannya adalah untuk membuang data yang tidak berguna, seperti data yang tidak valid, tidak konsisten, dan tidak relevan. Pada Tabel 1, terdapat data nomor yang tidak digunakan, sehingga diperlukan langkah pre-processing untuk menghapus data nomor agar menjadi seperti Tabel 2.

Tabel 2. Data setelah pre-processing

Nama Project	Anggaran Biaya	Durasi Project	Kesulitan	Tim pengembang	Platform
Sistem Informasi Rumah Sakit	Rp. 175.000.000,00	213 Hari	Sedang	6 Orang	Web
Aplikasi Mobile Pembelajaran Interaktif	Rp. 200.000.000,00	243 Hari	Tinggi	7 Orang	Mobile
Sistem Manajemen Inventaris Sekolah	Rp. 150.000.000,00	152 Hari	Rendah	2 Orang	Desktop
Sistem E-Commerce	Rp. 500.000.000,00	304 Hari	Tinggi	10 orang	Web & Mobile
Aplikasi Pengelolaan Proyek Konstruksi	Rp. 400.000.000,00	213 Hari	Sedang	5 Orang	Web

c. Exploratory Data Analysis

Tahap ini bertujuan untuk memastikan data bersih dan terstruktur dengan baik, dengan cara memeriksa data kosong, menghapus duplikat, dan mengonversi kategori data jika diperlukan. setelah data dicek dan dibersihkan, langkah berikutnya adalah melakukan analisis statistik deskriptif untuk menjelaskan karakteristik data secara lebih mendalam.

Tabel 3. Konversi kategori data

Nama Project	Anggaran Biaya	Durasi Project	Kesulitan	Tim pengembang	Platform
Sistem Informasi Rumah Sakit	175000000	213	2	6	1
Aplikasi Mobile Pembelajaran Interaktif	200000000	243	3	7	2
Sistem Manajemen Inventaris Sekolah	150000000	152	1	2	3
Sistem E-Commerce	500000000	304	3	10	4
Aplikasi Pengelolaan Proyek Konstruksi	400000000	213	2	5	1

Pada tabel 3 dilakukan *Label Encoding* untuk mengubah variabel kategorikal menjadi representasi numerik berdasarkan penomoran kategori secara berurutan, di mana setiap kategori diberi nilai numerik yang unik [12]. Tahapan selanjutnya yaitu melakukan statistik deskriptif, dapat dilihat pada tabel 4.

Tabel 4. Deskripsi data

	Anggaran Biaya	Durasi Project	Kesulitan	Tim pengembang	Platform
Count	1.010000e+03	1010	1010	1010	1010
Mean	1.044161e+07	307.9	2	2.5	5.7
Std	5.378287e+06	115.5	0.8	1.1	2.8
Min	1.013317e+06	100	1	1	1
25%	5.800576e+06	209	1	2	3
50%	1.044834e+07	319	2	2	6
75%	1.488371e+07	410	3	3	8
Max	1.999123e+07	499	3	4	10

Informasi statistik data dirangkum dalam Tabel 4, dengan mencantumkan jumlah data (count), rata-rata (mean), standar deviasi (std), nilai terkecil (min), dan nilai terbesar (max).

Tabel 5. Informasi data

Kolom	Hitungan Bukan-Nol	Tipe Data
Nama Project	1010 non-null	Object
Anggaran Biaya	1010 non-null	Int64
Durasi Project	1010 non-null	Int64
Kesulitan	1010 non-null	Int64
Tim Pengembang	1010 non-null	Int64
Platform	1010 non-null	In64

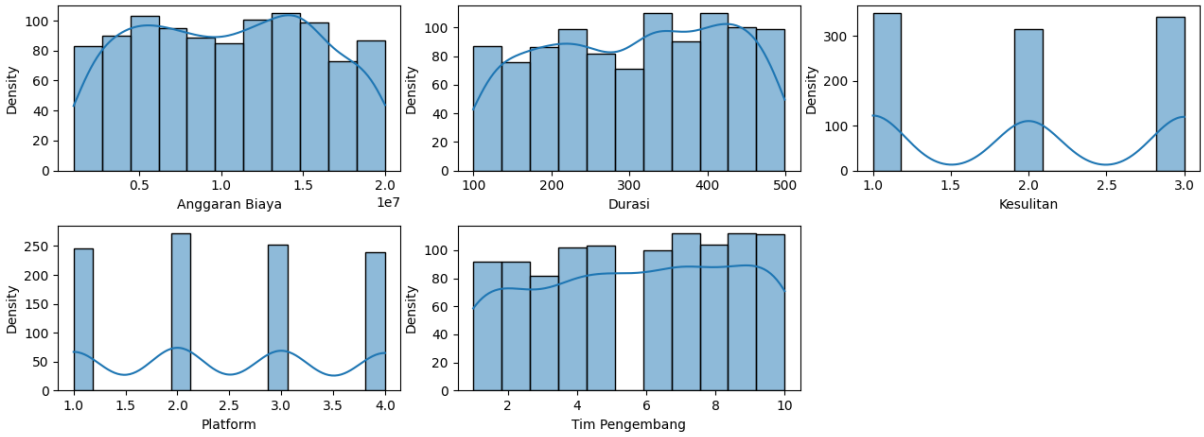
Tabel 5 menampilkan informasi tentang tipe data yang terdiri dari 1010 record. Nama proyek memiliki tipe data objek, sementara anggaran biaya, durasi project, tingkat kesulitan, dan platform memiliki tipe data integer. Selanjutnya, melakukan pengecekan terhadao data yang kosong. Jika tidak terdapat data yang kosong, proses dapat dilanjutkan ke tahap selanjutnya.

Tabel 6. Cek missing value

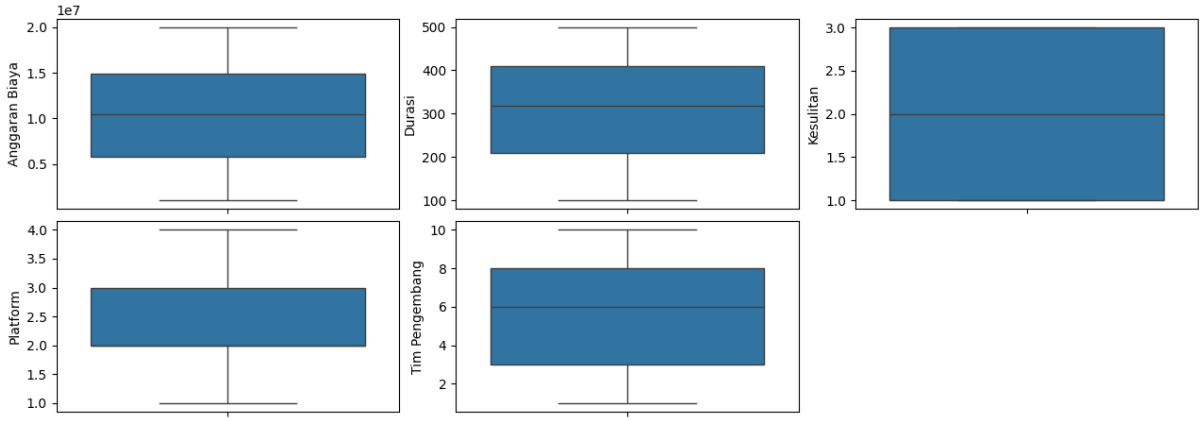
Column	#
Nama Project	0
Anggaran Biaya	0
Durasi Project	0
Kesulitan	0
Tim Pengembang	0
Platform	0
Dtype : int64	

d. Exploratory Data Analysis

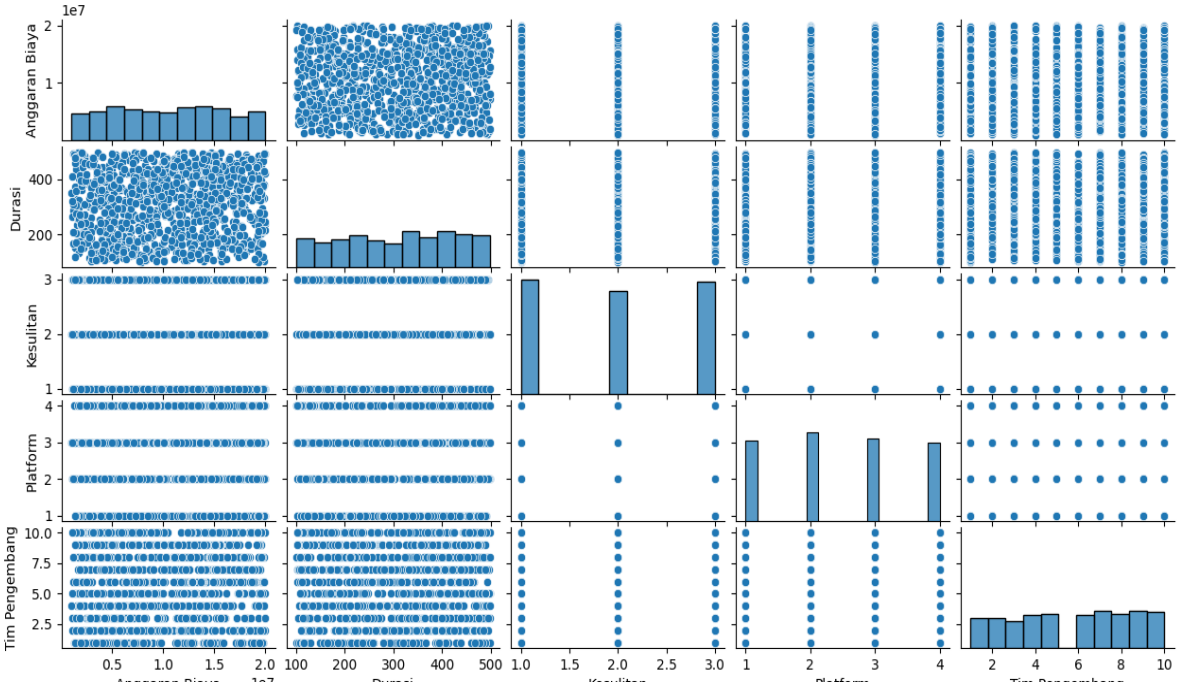
Pada tahap ini, dilakukan visualisasi data proyek untuk melihat persebaran data. Gambar 3, 4, dan 5 menampilkan visualisasi yang dimaksud.



Gambar 3. Dataset project (displot)



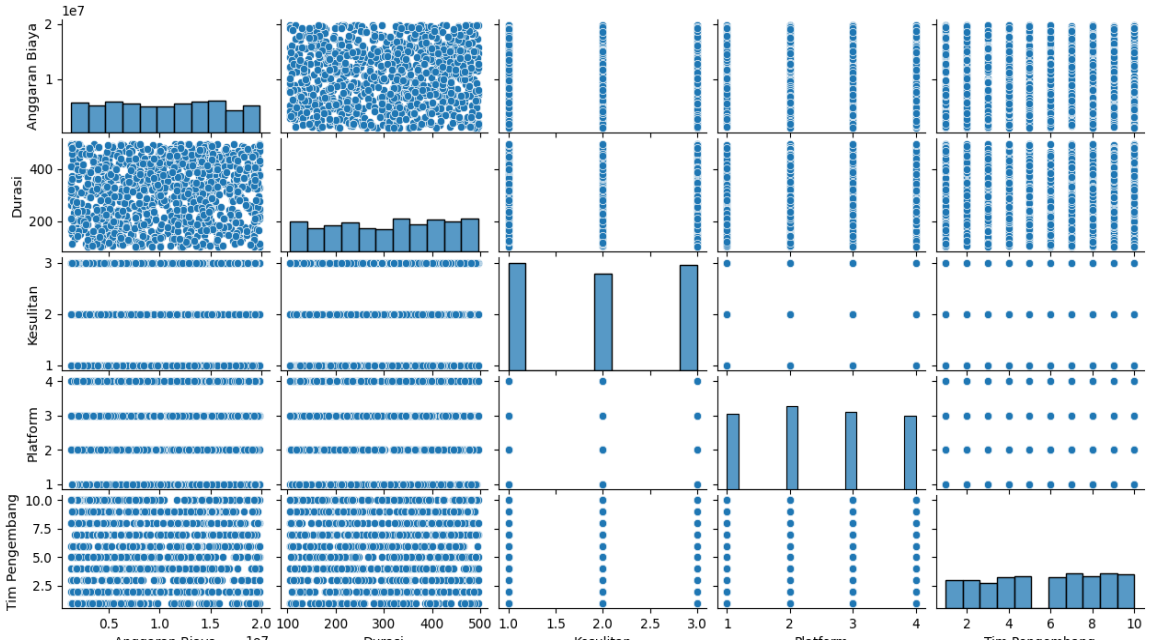
Gambar 4. Dataset project (boxplot)



Gambar 5. Dataset project (pairplot)

Dapat dilihat pada gambar 3, 4, dan 5, terdapat visualisasi distribusi dan persebaran data project. Perhatikan bahwa dalam visualisasi tersebut, terlihat adanya beberapa titik data yang terpencil (outlier), yang dapat memengaruhi analisis secara keseluruhan [13]. Oleh karena itu, perlu dilakukan pembersihan terhadap outlier tersebut agar data menjadi lebih bersih dan representatif.

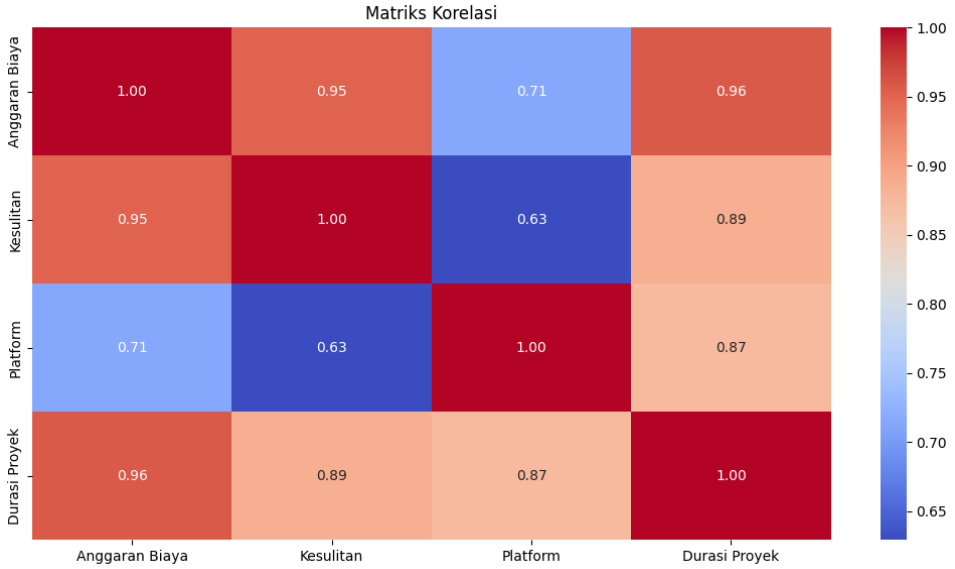
Untuk menjaga kualitas data, akan dilakukan pembersihan outlier yang hanya mencakup 1% dari total data project. Dengan demikian, dari jumlah data awal yang mencakup 100%, setelah pembersihan, akan tersisa 99% dari total data. Hasil dari pembersihan outlier ini dapat dilihat dalam Gambar 6 yang disajikan di bawah ini.



Gambar 6. Dataset project (pairplot) setelah dibersihkan

e. Matriks Korelasi

Matriks korelasi berfungsi untuk menunjukkan tingkat hubungan linier antara dua atau lebih variabel dalam sebuah dataset [14]. Kuat lemahnya hubungan antar variabel diukur dengan koefisien korelasi Pearson. Nilai koefisien ini berkisar antara -1 hingga +1.



Gambar 7. Matriks Korelasi

Dilihat pada gambar 7, Nilai koefisien korelasi berwarna merah menunjukkan hubungan linear positif yang kuat antara variabel-variabel yang terkait. Ini berarti ketika nilai satu variabel naik, kemungkinan besar nilai variabel lainnya juga akan naik, dan sebaliknya. Sedangkan, warna biru pada menandakan hubungan linear yang negatif. Hal ini mengindikasikan bahwa ketika nilai satu variabel meningkat, kemungkinan besar nilai variabel lainnya akan menurun, dan sebaliknya.

f. Multiple Linear Regression

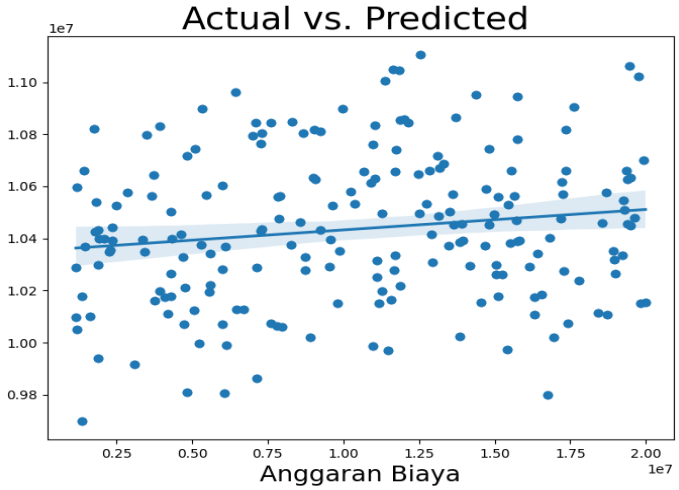
Pembagian dataset 1010 menjadi 80% data latih dan 20% data tes dilakukan di awal perhitungan dengan model multiple linear regression, untuk membangun model regresi pada sklearn. Setelah itu, variabel dependen (y) ditetapkan sebagai anggaran biaya dan variabel independen (x) ditetapkan sebagai durasi, tingkat kesulitan, tim pengembang, dan platform [15]. Proses regresi linear multipel dilakukan seperti yang ditunjukkan pada Tabel 6.

Tabel 7. Proses model multiple linear regression

1	X = df[['Kesulitan', 'Platform', 'Durasi', 'Tim Pengembang']] y = df['Anggaran Biaya']	Menetapkan variabel dependen (y) dan variabel independen (x)
2	X_latih, X_tes, y_latih, y_tes = train_test_split(X, y, test_size=0.2, random_state=2)	Pembagian data menjadi dua, yaitu data latih dan data tes.
3	lin_reg = LinearRegression() lin_reg.fit(X_latih, y_latih)	Membuat dan melatih model *linear regression*
4	y_pred = lin_reg.predict(X_tes) r_squared = r2_score(y_test, y_pred) mse = mean_squared_error(y_tes, y_pred) mrse = mse ** 0.5	Menghitung nilai prediksi, koefisien determinan, MSE, dan MRSE

5	print("Coefficient of Determination: ", r_squared) print("Mean Squared Error (MSE): ", mse) print("Mean Root Squared Error (MRSE): ", mrse)	Coefficient of Determination: 0.6291315109431107 MSE: 1.433269785249304e+16 MRSE: 119719245.95691806
6	vis.figure(figsize=(8,6)) vis.title("Actual vs. Predicted ", fontsize=25) vis.xlabel("Predicted", fontsize=18) vis.scatter(x=y_tes, y=y_pred) sns.regplot(x=y_tes, y=y_pred) vis.show()	Memvisualisasikan anggaran biaya awal dan prediksi

Gambar 8 dibawah merupakan representasi grafis dari performa model *multiple linear regression*.



Gambar 8. Visualisasi data *multiple linear regression*

g. Random Forest Regression

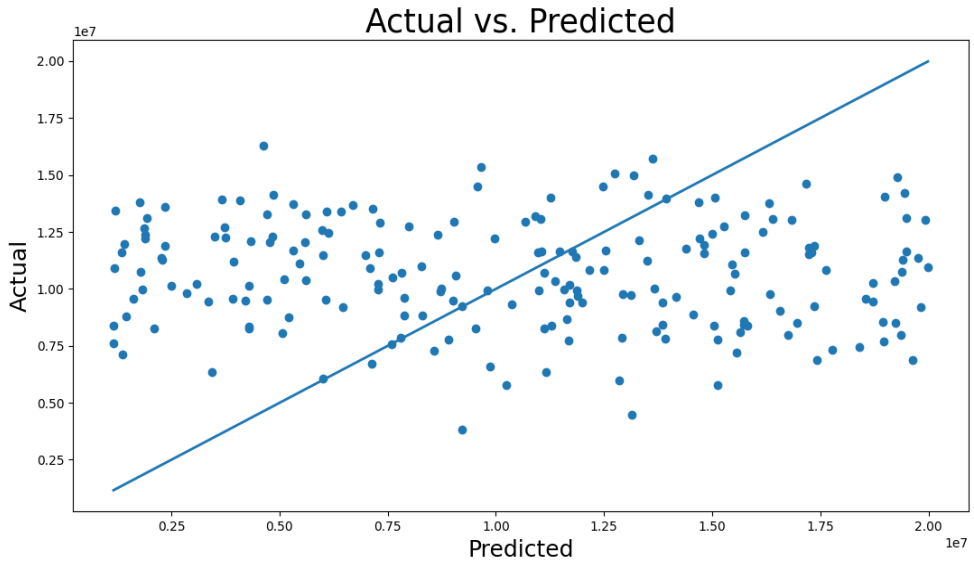
Proses perhitungan dengan algoritma Random dilakukan pada dataset yang terdiri dari 1010 data, dengan pembagian 80% data latih dan 20% data tes Variabel dependen (y) ditetapkan sebagai kolom anggaran biaya, sedangkan variabel independen (x) terdiri dari kolom durasi proyek, tingkat kesulitan, tim pengembang, platform yang digunakan. Langkah-langkah meliputi pembuatan model, pelatihan, evaluasi, dan pengujian model untuk memprediksi anggaran biaya berdasarkan atribut-atribut tersebut [16]. Proses random forest dapat dilihat seperti pada tabel 7.

Tabel 8. Proses model random forest regression

1	X = df[['Kesulitan', 'Platform', 'Durasi', 'Tim Pengembang']] y = df['Anggaran Biaya']	Menetapkan variabel dependen (y) dan variabel independen (x)
2	X_latih, X_tes, y_latih, y_tes = train_test_split(X, y, test_size=0.2, random_state=2)	Pembagian data menjadi dua, yaitu data latih dan data tes.
3	rf_reg = RandomForestRegressor(n_estimators=100, random_state=42) rf_reg.fit(X_train, y_train)	Membuat dan melatih model *Random Forest Regression*
4	y_pred = rf_reg.predict(X_test) r_squared = r2_score(y_test, y_pred) mse = mean_squared_error(y_test, y_pred) mrse = mse ** 0.5	Menghitung nilai prediksi, koefisien determinan, MSE, dan MRSE

5	print("Coefficient of Determination (R-squared):", r_squared) print("Mean Squared Error (MSE):", mse) print("Mean Root Squared Error (MRSE):", mrse)	Coefficient of Determination: 0.8153413940794764 MSE: 1.6926039873179744e+16 MRSE: 130100114.80848026
6	vis.figure(figsize=(8, 6)) vis.title("Actual vs. Predicted", fontsize=25) vis.xlabel("Predicted", fontsize=18) vis.scatter(x=y_tes, y=y_pred) vis.plot([y_tes.min(), y_tes.max()], [y_tes.min(), y_tes.max()], lw=2) vis.ylabel("Actual", fontsize=18) vis.show()	Memvisualisasikan anggaran biaya awal dan prediksi

Gambar 9 dibawah merupakan representasi grafis dari performa model *random forest regression*.



Gambar 9. Visualisasi data *random forest regression*

h. Deployment / Evaluation

Pada proses perhitungan menggunakan algoritma multiple linear regression menghasilkan Coefficient of Determination sebesar 0.6291315109431107, maka model MLR menghasikanl akurasi sebesar 62,9%, sedangkan algoritma random forest regression menghasilkan Coefficient of Determination mencapai mencapai 0.8153413940794764, dengan akurasi 81,5%, model RFR menunjukkan performa yang lebih baik dengan selisih 18,6% [17].

4. Kesimpulan

Dalam penelitian ini, kami menjalankan pengujian terhadap dua model algoritma, yaitu *multiple linear regression* dan *random forest regression*, untuk memprediksi anggaran biaya pada manajemen proyek sistem informasi. Hasil pengujian menunjukkan bahwa model *random forest regression* memberikan nilai akurasi yang lebih tinggi, mencapai 81,5%, dibandingkan dengan nilai akurasi 62,9% yang diperoleh oleh model *multiple linear regression*. Random forest regression lebih fleksibel dan dapat menangani hubungan non-linear antara fitur dan variabel target. Ini berarti bahwa RFR dapat menangkap pola yang lebih kompleks dalam data daripada MLR, yang hanya mampu menangani hubungan linier. Hal ini menunjukkan bahwa *random forest regression* lebih efektif dalam memprediksi anggaran biaya proyek sistem informasi berdasarkan data yang telah ada. Kesimpulan ini menunjukkan bahwa penggunaan metode *random forest regression* dapat menjadi pilihan yang lebih baik dalam menghadapi tantangan kompleksitas dan ketidakpastian yang sering dihadapi dalam manajemen proyek sistem informasi.

Ucapan Terima Kasih
Ucapan terima kasih kepada pihak-pihak yang telah memberikan dukungan terhadap penelitian.

Referensi
[1] d. Hadion Wijoyo, SISTEM INFORMASI MANAJEMEN, Kapalo Koto: INSAN CENDEKIA MANDIRI, 2021.
[2] T. H. Salsabila, T. M. Indrawati dan R. A. Fitrie, "Meningkatkan Efisiensi Pengambilan Keputusan Publik melalui Kecerdasan Buatan," Journal of Internet and Software Engineering, vol. 1, no. 2, pp. 1-21, 2024.
[3] A. C. Handoko dan Hendry, "PERBANDINGAN METODE SUPERVISED LEARNING UNTUK PREDIKSI DIABETES GESTASIONAL," Jurnal Ilmiah Penelitian dan Pembelajaran Informatika, vol. 8, no. 4, pp. 1238-1247, 2023.
[4] J. M. A. Soraya Dachi dan P. Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam (JURRIMIPA), vol. 2, no. 2, pp. 87-103, 2023.
[5] M. R. Fahlepi dan A. Widjaja, "Penerapan Metode Multiple Linear Regression Untuk Prediksi Harga Sewa Kamar Kost," Jurnal Strategi, vol. 1, no. 2, pp. 615-629, 2019.
[6] Farhanuddin, "Perancangan Sistem Pendukung Keputusan Pemilihan Investasi Menggunakan Metode AHP Pada DPMPTSP Kota Medan," JOURNAL OF COMPUTERS AND DIGITAL BUSINESS, vol. 3, no. 1, pp. 26-30, 2024.
[7] M. Radhi, A. Amalia, D. R. H. Sitompul, S. H. Sinurat dan E. Indra, "ANALISIS BIG DATA DENGAN METODE EXPLORATORY DATA ANALYSIS (EDA) DAN METODE VISUALISASI MENGGUNAKAN JUPYTER NOTEBOOK," JUSIKOM PRIMA, vol. 4, no. 2, pp. 23 -27, 2022.
[8] M. Sholeh, S. Suraya dan D. Andayati, "Machine Linear untuk Analisis Regresi Linier Biaya Asuransi Kesehatan dengan Menggunakan Python Jupyter Notebook," Jurnal Edukasi dan Penelitian Informatika, vol. 8, no. 1, pp. 20-27, 2022.
[9] A. N. Fadhilah, A. F. Boy dan R. Syahputra, "Implementasi Data Mining Untuk Pengelompokan Data Penjualan Berdasarkan Pola Pembelian Menggunakan Algoritma K-Means Clustering Pada Toko Syihan," Jurnal CyberTech, vol. 2, no. 5, 2019.
[10] S. Jesika, S. Ramadhani dan Y. P. Putri, "Implementasi Model Machine Learning dalam Mengklasifikasi Kualitas Air," Jurnal Ilmiah Dan Karya Mahasiswa, vol. 1, no. 6, pp. 382-396, 2023.
[11] C. Haryanto, N. Rahaningsih dan F. M. Basysyar, "KOMPARASIALGORITMA MACHINE LEARNINGDALAM MEMPREDIKSI HARGA RUMAH," Jurnal Mahasiswa Teknik Informatika, vol. 7, no. 1, pp. 533-539, 2023.
[12] D. Novaliendry, Deep Learning Untuk Pemula Jilid 1, Purwodadi: CV. SARNU UNTUNG, 2023.
[13] A. A. Munawar dan Hasanuddin, Analisis Data Multivariat Menggunakan The Unscrambler X, Banda Aceh: Syiah Kuala University Press, 2020.
[14] M. D. H. Kusuma dan S. Hidayat, "Penerapan Model Regresi Linier dalam Prediksi Harga Mobil Bekas di India dan Visualisasi dengan Menggunakan Power BI," Jurnal Indonesia: Manajemen Informatika dan Komunikasi, vol. 5, no. 2, pp. 1097-1110, 2024.
[15] L. M. Ginting, M. MT.Sigiro, E. D. Manurung dan J. J. P. Sinurat, "Perbandingan Metode Algoritma Support Vector Regression dan Multiple Linear Regression Untuk Memprediksi Stok Obat," Journal of Applied Technology and Informatics, vol. 1, no. 2, pp. 29-34, 2021.
[16] A. Primajaya dan B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), vol. 1, no. 1, pp. 27-31, 2018.
[17] M. Yafi, Urrochman, E. Setyati dan Y. Kristian, "Prediksi Timing Financial Distress Pada Bank Perkreditan Rakyat di Indonesia Menggunakan Machine Learning," Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi, vol. 12, no. 2, pp. 576-584, 2023.