Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi
https://ojs.stmik-banjarbaru.ac.id/index.php/jutisi/index
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0893

Analisis Perbandingan Algoritma *Linear Regression* dan *Polynomial Regression* dalam Memprediksi Durasi Rawat Inap Pasien Rumah Sakit

DOI: http://dx.doi.org/10.35889/jutisi.v14i3.3233

I Putu Abdiparta[1*], Ni Made Satvika Iswari[2], Nengah Widya Utami[3]
[1,2]Informatika, Universitas Primakara, Denpasar, Indonesia
[3]Sistem Informasi Akuntansi, Universitas Primakara, Denpasar, Indonesia
*e-mail *Corresponding Author*: iputuabdiparta@gmail.com

Abstract

The length of stay (LoS) of hospital patients is an essential indicator for measuring the efficiency of healthcare services. Accurate LoS prediction helps hospitals optimize resource management, estimate costs, and improve service quality. This study compares the performance of Linear Regression and Polynomial Regression algorithms in predicting patient LoS. The dataset, obtained from Kaggle, consists of 835 patient records that underwent preprocessing and transformation. The independent variables include gender, age, disease type, and type of service, while LoS serves as the dependent variable. The research applies the Knowledge Discovery from Data (KDD) approach, which includes the stages of selection, cleaning, transformation, data mining, and evaluation. Experiments were conducted using three data-splitting ratios (70:30, 80:20, and 90:10) with evaluation metrics MAE, MSE, RMSE, and R^2. The results show that Linear Regression performed slightly better, with average R^2 values ranging between 0.18 and 0.20, indicating its potential to support hospital management efficiency.
Keywords: *Length of Stay; Linear Regression; Polynomial Regression; Data Mining; Prediction.*

Abstrak

Durasi rawat inap pasien (*Length of Stay*/LoS) merupakan indikator penting dalam mengukur efisiensi pelayanan rumah sakit. Prediksi LoS yang akurat membantu rumah sakit dalam pengelolaan sumber daya, estimasi biaya, dan peningkatan mutu layanan. Penelitian ini membahas perbandingan kinerja algoritma Linear Regression dan Polynomial Regression dalam memprediksi LoS pasien. Data penelitian diperoleh dari Kaggle dengan total 835 data pasien yang melalui proses preprocessing dan transformasi. Variabel independen meliputi gender, umur, jenis penyakit, dan jenis service, sedangkan LoS menjadi variabel dependen. Metode penelitian menggunakan pendekatan *Knowledge Discovery from Data* (KDD) yang mencakup tahapan selection, cleaning, transformation, data mining, dan evaluation. Pengujian dilakukan pada tiga rasio pembagian data (70:30, 80:20, dan 90:10) menggunakan metrik MAE, MSE, RMSE, dan R^2. Hasil menunjukkan *Linear Regression* memiliki performa sedikit lebih unggul dengan rata-rata R^2 sebesar 0,18–0,20, yang dapat mendukung efisiensi manajemen rumah sakit.
Kata kunci: *Length of Stay; Linear Regression; Polynomial Regression; Data Mining; Prediksi.*

1. Pendahuluan

Pelayanan kesehatan yang efisien dan berkualitas merupakan faktor penting dalam meningkatkan kesejahteraan masyarakat. Salah satu indikator utama untuk menilai efisiensi pelayanan rumah sakit adalah durasi rawat inap atau *Length of Stay* (LoS) [1]. LoS yang terlalu panjang dapat meningkatkan risiko komplikasi medis, menurunkan rotasi tempat tidur, serta menambah beban biaya operasional bagi rumah sakit dan pasien [2],[3]. Oleh karena itu, kemampuan untuk memprediksi lama rawat inap pasien secara akurat menjadi aspek penting dalam mendukung perencanaan pelayanan dan efisiensi manajemen rumah sakit.

Saat ini, banyak rumah sakit masih menghadapi kendala dalam mengestimasi durasi rawat inap pasien secara tepat karena perbedaan karakteristik penyakit, usia, serta jenis pelayanan medis. Ketidakpastian ini menyebabkan pemborosan sumber daya, keterlambatan penanganan pasien baru, dan ketidakseimbangan kapasitas tempat tidur. Berdasarkan studi sebelumnya, faktor-faktor seperti usia, jenis penyakit, dan jenis layanan medis memiliki pengaruh terhadap lama rawat inap pasien [2],[4]. Namun, belum banyak penelitian yang mengkaji prediksi LoS dengan mempertimbangkan kombinasi variabel tersebut menggunakan metode regresi yang mudah diimplementasikan secara praktis di lingkungan rumah sakit.

Sebagai solusi terhadap permasalahan tersebut, penelitian ini mengusulkan pendekatan berbasis Data Mining menggunakan dua algoritma regresi, yaitu *Linear Regression* dan *Polynomial Regression* [5],[6]. Pendekatan ini dipilih karena mampu menganalisis hubungan antara variabel input dengan variabel output secara kuantitatif dan terukur. *Linear Regression* memiliki keunggulan dalam interpretasi hasil dan kesederhanaan model, sedangkan *Polynomial Regression* mampu menangkap hubungan nonlinier yang lebih kompleks antar variabel [7]. Penggunaan metode ini diharapkan dapat menghasilkan model prediksi yang akurat dan efisien dalam konteks data kesehatan.

Penelitian ini bertujuan untuk melakukan perbandingan kinerja antara algoritma *Linear Regression* dan *Polynomial Regression* dalam memprediksi durasi rawat inap pasien (*Length of Stay*). Hasil penelitian diharapkan dapat memberikan kontribusi bagi pihak rumah sakit dalam memperkirakan lama rawat inap secara lebih akurat, sehingga membantu pengambilan keputusan, perencanaan sumber daya, dan peningkatan efisiensi pelayanan kesehatan.

2. Tinjauan Pustaka

Durasi rawat inap atau *Length of Stay* (LoS) merupakan salah satu indikator krusial dalam manajemen rumah sakit. LoS tidak hanya mencerminkan efisiensi operasional dan konsumsi sumber daya, tetapi juga dapat menjadi cerminan dari kompleksitas penanganan pasien. Oleh karena itu, penelitian dilakukan untuk memprediksi LoS untuk mengoptimalkan manajemen rumah sakit dan meningkatkan kualitas layanan [8]. Metode data mining dan *machine learning* seringkali menjadi pendekatan utama dalam upaya prediksi ini. Salah satu studi relevan berfokus pada prediksi LoS yang berkepanjangan pada pasien demam berdarah. Penelitian ini membandingkan model *Logistic Regression* dengan *Random Forest* dan menemukan bahwa *Logistic Regression* menunjukkan kinerja yang lebih baik dengan nilai AUC 0.75. Pendekatan ini relevan karena menggunakan teknik regresi untuk memprediksi LoS, meskipun berfokus pada satu jenis penyakit [9].

Studi lain juga menggunakan data mining untuk memprediksi LoS, namun pada pasien jantung. Mereka membandingkan *Artificial Neural Network* (ANN), *Support Vector Machines* (SVM), dan *Decision Tree*. Hasilnya menunjukkan bahwa SVM adalah metode terbaik untuk memprediksi LoS dalam konteks tersebut. Hal ini menyoroti bahwa pemilihan algoritma sangat bergantung pada karakteristik data dan jenis penyakit yang diteliti [10]. Selain itu, pada penelitian sejenis perbandingan antara model prediksi LoS klasik dengan model *real-time* pada pasien unit perawatan intensif (ICU). Mereka menyimpulkan bahwa model *real-time* lebih memadai karena mampu menyesuaikan prediksi seiring dengan perubahan kondisi pasien yang cepat di lingkungan ICU. Penelitian ini menegaskan pentingnya adaptasi model terhadap kondisi spesifik pasien dan lingkungan perawatan [11].

Dalam konteks algoritma yang digunakan, beberapa penelitian terdahulu secara khusus membandingkan *Linear Regression* dan *Polynomial Regression*. *Polynomial Regression* seringkali memberikan hasil prediksi yang lebih akurat dibandingkan *Linear Regression*, terutama ketika hubungan antar variabel tidak linier. Hal ini disebabkan kemampuan *Polynomial Regression* untuk menangkap hubungan nonlinier yang kompleks antar variabel, yang sering ditemukan dalam data medis. Pada penelitian tersebut menemukan bahwa *Polynomial Regression* memiliki akurasi yang lebih baik [7].

Riset ini hadir untuk mengatasi keterbatasan tersebut dengan menyajikan sebuah kebaruan (*novelty*). Alih-alih berfokus pada satu jenis penyakit, penelitian ini akan menganalisis prediksi durasi rawat inap berdasarkan berbagai jenis penyakit yang berbeda menggunakan dua algoritma regresi, yaitu *Linear Regression* dan *Polynomial Regression*. Dengan menganalisis beragam jenis penyakit dalam satu model, diharapkan dapat ditemukan pola yang lebih umum dan relevan untuk memprediksi LoS secara lebih komprehensif. Perbedaan utama penelitian ini dengan riset terdahulu adalah pada cakupan data dan metode analisisnya. Jika

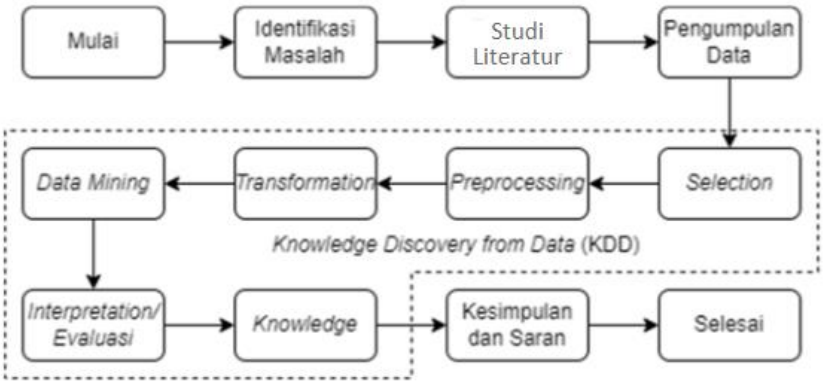
penelitian sebelumnya menggunakan pendekatan yang sangat spesifik, riset ini mencoba membangun model prediksi yang lebih umum dan aplikatif dengan mengintegrasikan data dari berbagai kasus penyakit. Dengan demikian, penelitian ini bertujuan untuk menyumbang pada pengembangan model prediksi LoS yang lebih fleksibel dan dapat diadaptasi di berbagai lingkungan klinis.

3. Metodologi

Metodologi penelitian ini berfokus pada kajian kinerja algoritma/metode komputasi dengan tujuan memprediksi *Length of Stay* (LoS) pasien. Prosedur penelitian ini mengikuti tahapan-tahapan *Knowledge Discovery from Data* (KDD) yang bersifat prosedural dan praktis, bukan teoritis.

3.1 Prosedur Penelitian

Penelitian ini menggunakan tahapan *Knowledge Discovery from Data* (KDD) untuk memproses data dan menghasilkan model prediksi. Alur penelitian secara keseluruhan digambarkan pada *Figure 1* dan dijelaskan secara rinci di bawah.



Gambar 1. Alur Penelitian

Berikut merupakan penjelasan dari alur penelitian yang dilakukan:

1) Identifikasi Masalah
   Tahapan ini menguraikan permasalahan yang dijumpai dalm penelitian serta cara permasalahn tersebut diidentifikasi, diukur, dan dikaitkan dengan tahapan prosedur penelitian yang dilakukan. Untuk mendukung proses tersebut, penelitian ini menerapkan metode *Knowledge Discovery from Data*.
2) Studi Literatur
   Peneliti memperoleh data dari sumber-sumber yang sudah ada sebelumnya atau tidak dikumpulkan oleh peneliti dalam penelitiannya. Penelitian ini menggunakan dataset yang berasal dari Kaggle.
3) Pengumpulan Data
   Pengumpulan data dilakukan dengan mengambil dataset dari Kaggle. Data tersebut kemudian diolah untuk menghasilkan model yang dapat memprediksi durasi rawat inap pasien.
4) *Selection*
   Pada langkah ini dilakukan pemilahan data, dimana dataset yang digunakan berasal dari Kaggle dengan total 835 data.
5) *Prepocessing (Cleaning*)
   Langkah ini mencakup penanganan data hilang, menghapus data duplikat, dan koreksi data. Pada tahap ini akan diproses menggunakan Microsoft Excel.
6) *Transformation*
   Langkah ini adalah proses mengubah atau mentransformasi data yang telah dibersihkan ke dalam format yang sesuai untuk dianalisa. Proses ini akan dilakukan menggunakan Microsoft Excel.
7) *Data Mining*

Tahap ini merupakan proses penggalian data dengan menerapkan algoritma yang telah ditetapkan. Pada penelitian ini digunakan dua algoritma, yaitu *Linear Regression* dan *Polynomial Regression*. Proses data mining tersebut dilakukan dengan bantuan perangkat lunak Orange.

8) *Interpretation/Evaluation*
Tahapan ini bertujuan untuk mengevaluasi pola atau model yang telah dihasilkan guna mengetahui tingkat kegunaan dan kinerjanya. Pada tahap ini dilakukan penilaian terhadap performa masing-masing algoritma dengan mengukur tingkat akurasi menggunakan metrik MAE, MSE, MRSE, dan R^2.
9) *Knowledge*
Langkah ini akan menghasilkan pengetahuan berdasarkan hasil prediksi yang dihasilkan oleh model. Pengetahuan dalam bentuk rumus inilah yang nanti bisa dijadikan bahan pertimbangan bagi pihak Rumah Sakit dan Pasien.
10) Kesimpulan dan Saran
Kesimpulan dari penelitian ini diharapkan dapat membantu pihak Rumah Sakit dan pasien dalam mengambil keputusan.

3.1.1 Pengumpulan dan Pemilihan Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh melalui studi literatur. Sumber data berasal dari dataset publik di Kaggle (https://www.kaggle.com/datasets/staniherstaniher/data-patients-los-in-a-semi-urban-hospital?resource=download&select=mon_dataset_ok.csv). Dataset ini berisi 835 data pasien dari sebuah rumah sakit semi-urban. Data yang dipilih mencakup atribut *Gender, Age, Disease, Service*, dan LOS sebagai variabel target.

3.1.2 Pra-proses Data (*Cleaning* & *Transformation*)

Tahap ini mencakup pembersihan dan transformasi data. Langkah-langkah yang dilakukan:

1) Pembersihan Data (*Cleaning*): Mengidentifikasi dan menangani data yang hilang (missing values), menghapus data duplikat, dan mengoreksi data yang salah jika ditemukan.
2) Transformasi Data (*Transformation*): Mengubah data ke dalam format yang sesuai untuk analisis komputasi. Data kualitatif seperti jenis penyakit (*Disease*) akan diubah menjadi format numerik. Proses ini akan dilakukan menggunakan perangkat lunak Microsoft Excel.

3.1.3 Evaluasi Kinerja Algoritma

Kinerja kedua algoritma akan divalidasi dan diuji menggunakan metrik evaluasi standar untuk regresi, yaitu:

1) *Mean Absolute Error* (MAE)
Mean Absolute Error (MAE) adalah sebuah metrik yang menghitung besarnya rata-rata kesalahan absolut antara nilai prediksi dan nilai sebenernya [12]. MAE tidak peduli walaupun kesalahan itu bernilai besar atau kecil. Oleh karena itu, MAE tidak sensitif terhadap *outlier*, sehingga metrik ini berguna untuk model dimana *outlier* tidak menjadi perhatian khusus. Rumus dari MAE adalah:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{..... (1)}$$

n: Jumlah data (jumlah observasi).
y_i: Nilai aktual (nilai yang sebenarnya).
$\hat{y}_i$: Nilai prediksi (nilai yang diprediksi oleh model).
$|y_i - \hat{y}_i|$: Selisih absolut antara nilai aktual dan nilai prediksi.

2) *Mean Squared Error* (MSE)
Mean Squared Error (MSE) adalah sebuah metrik yang menghitung rata-rata perbedaan kuadrat antara nilai sentimen yang diprediksi dan nilai sentimen sebenarnya

[13]. Dengan kata lain, MSE menghitung rata-rata kuadrat nilai prediksi dan nilai sebenarnya. Karena kesalahan dihitung dalam nilai kuadrat, ini berarti MSE lebih sensitif terhadap kesalahan bernilai besar/outlier.

Rumus dari MSE adalah:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{..... (2)}$$

n: Jumlah data (jumlah observasi).
y_i: Nilai aktual (nilai yang sebenarnya).
$\hat{y}_i$: Nilai prediksi (nilai yang diprediksi oleh model).
$(y_i - \hat{y}_i)^2$: Selisih kuadrat antara nilai aktual dan nilai prediksi.

3) *Root Mean Squared Error* (RMSE)
Root Mean Squared Error (RMSE) didefinisikan sebagai ukuran perbedaan antara nilai yang diprediksi oleh model dan nilai yang sebenarnya diamati [14]. Secara sederhana, RMSE adalah akar kuadrat dari MSE. RMSE memberikan ukuran kesalahan dalam satuan yang sama dengan data asli. Rumus dari RMSE adalah:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{..... (3)}$$

n: Jumlah data (jumlah observasi).
y_i: Nilai aktual (nilai yang sebenarnya).
$\hat{y}_i$: Nilai prediksi (nilai yang diprediksi oleh model).
$(y_i - \hat{y}_i)^2$: Selisih kuadrat antara nilai aktual dan nilai prediksi.

4) *Coefficient of Determination* (R2)
Semakin tinggi nilai R2, semakin bagus model yang dibentuk. Namun perlu diperhatikan, apabila terjadinya overfitting, maka nilai R2 akan menjadi tinggi, sehingga nilai R2 yang sangat tinggi bukan selalu berarti bagus [15]. Rumus dari R2 adalah:

$$R^2 = 1 - \frac{RSS}{TSS} \quad \text{..... (4)}$$

RSS: Residual Sum of Squares adalah jumlah kuadrat selisih antara nilai sebenarnya dan nilai prediksi yang dihasilkan oleh model. RSS mengukur kesalahan model (variabilitas data yang tidak dapat dijelaskan oleh model). Rumus RSS:

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{..... (5)}$$

n: Jumlah data (jumlah observasi).
y_i: Nilai aktual pada titik ke-i
$\hat{y}_i$: Nilai prediksi pada titik ke-i
TSS: Total Sum of Squares adalah jumlah kuadrat dari selisih antara setiap nilai sebenarnya dan rata-rata nilai sebenernya. TSS mengukur variabilitas total dalam data asli. Rumus TSS:

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{..... (6)}$$

y_i: Nilai aktual pada titik ke-i
$\bar{y}$: Rata-rata dari nilai aktual

Hasil dari setiap metrik akan dianalisis untuk membandingkan akurasi antara *Linear Regression* dan *Polynomial Regression*. Perangkat lunak Orange Data Mining akan digunakan untuk melakukan perhitungan metrik ini.

4. Hasil dan Pembahasan

4.1 Pengumpulan dan Pemilihan Data

Selection data pada penelitian ini menggunakan data pasien Rumah Sakit kota Butembo, provinsi Kivu Utara, Republik Demokratik Kongo, yang berasal dari Kaggle. Data pasien diambil dari tahun 2010 hingga tahun 2021 dengan total 835 data.
Berikut merupakan dataset pasien rumah sakit:

Tabel 1. Data Pasien Rumah Sakit

No	Gender	Age	*Disease*	Service	LOS
1	0	12	Malaria	0	5
2	0	30	Diabetes	1	6
3	1	6	Hernia	0	7
4	1	23	*Infection and Diabetes*	1	8
5	1	22	*Right Ovarian Cyst*	2	9
6	0	23	*Malaria*	1	9
7	1	23	*Emotional trauma*	1	10
8	0	25	*Uterine infection*	1	11
9	1	45	*Digestive tract*	1	5
10	1	30	*Chronic valvular disease*	1	3
…					
835	0	0.01	*Macrosomia*	3	21

4.2 Pembersihan Data (*Cleaning*)

Pada tahap ini dilakukan pemeriksaan data menggunakan Orange Data Mining untuk mengidentifikasi keberadaan *missing value* data duplikat yang berpotensi memengaruhi tingkat akurasi hasil pada proses data mining. Pada data ini, tidak ditemukan *missing value* dan *duplicate* data. Namun terdapat 11 data *outlier*, yang mana didefinisikan sebagai data yang memiliki lebih dari 1.000 LOS. Sehingga setelah proses ini, maka jumlah data menjadi 824 data. Berikut dataset pasien rumah sakit setelah dilakukan proses cleaning:

Tabel 2. Proses *Cleaning*

No	Gender	Age	*Disease*	Service	LOS
1	0	12	Malaria	0	5
2	0	30	Diabetes	1	6
3	1	6	Hernia	0	7
4	1	23	*Infection and Diabetes*	1	8
5	1	22	*Right Ovarian Cyst*	2	9
6	0	23	*Malaria*	1	9
7	1	23	*Emotional trauma*	1	10
8	0	25	*Uterine infection*	1	11
9	1	45	*Digestive tract*	1	5
10	1	30	*Chronic valvular disease*	1	3
…					
824	0	0.01	*Macrosomia*	3	21

4.3 Transformasi Data (*Transformation*)

Pada tahap *transformation* dilakukan proses pengubahan format data, yaitu dengan menyamakan seluruh variabel ke dalam bentuk numerik. Penjelasan lebih rinci mengenai proses transformasi dataset disajikan pada tabel berikut:

Tabel 3. Tabel Atribut

Atribut	Variabel Awal	Transformasi
Jenis Kelamin	L, Laki-laki, Laki-Laki	0
	P, Perempuan, perempuan	1
Kategori Penyakit	Neonatal dan Pediatrics	0
	Infeksius/penyakit menular	1
	Penyakit psikologis, saraf, dan umum	2
	Ginekologis (penyakit reproduksi, kehamilan)	3
	Cedera, Trauma, bedah, dan penyakit fisik lainnya	4
Kategori Service	*Pediatrics*	0
	Hospitalisation	1
	Gynecology	2
	Neonatology	3

Pada awalnya data tersebut berisi 68 nama penyakit dengan kode masing-masing. Kemudian 68 Nama Penyakit ini ditransformasi menjadi 5 Kategori penyakit seperti yang disebutkan pada Tabel Atribut. Berikut merupakan data setelah di transformasi:

Tabel 4. Tabel Transformasi

No	Gender	Age	Disease	Service	LOS
1	0	12	1	0	5
2	0	30	2	1	6
3	1	6	3	0	7
4	1	23	1	1	8
5	1	22	3	2	9
6	0	23	1	1	9
7	1	23	2	1	10
8	0	25	3	1	11
9	1	45	4	1	5
10	1	30	2	1	3
…					
824	0	0,01917	0	3	21

4.4 Eksperimen dan Pengujian Data

1) *Evaluation*

Pada tahap Interpretation/Evaluation dilakukan evaluasi kinerja algoritma *Linear Regression* dan *Polynomial Regression* dengan menggunakan metrik MSE, RMSE, MAE, dan R^2. Pengujian dilakukan melalui tiga kali percobaan. Proses evaluasi ini didasarkan pada skenario perbandingan data yang telah ditetapkan sebelumnya pada perangkat lunak Orange untuk menjalankan proses data mining menggunakan algoritma *Linear Regression* dan *Polynomial Regression* pada data pasien rawat inap rumah sakit.

Pada percobaan pertama digunakan perbandingan data sebesar 70:30, yaitu 70% data sebagai data pelatihan (*training*) sebanyak 577 data dan 30% data sebagai data pengujian (testing) sebanyak 247 data. Hasil dari percobaan dengan perbandingan pertama tersebut disajikan pada tabel berikut.

Tabel 5. *Split Validation* 70:30

Model	MSE	RMSE	MAE	R2
Linear Regression	2139.124	46.251	31.955	0.197
Polynomial Regression	2612.291	51.111	36.943	0.019

Berdasarkan tabel di atas, pada skema Split Validation 70:30, algoritma *Linear Regression* memperoleh nilai MSE sebesar 2139.124, RMSE sebesar 46.251, MAE sebesar 31.955, serta nilai R^2 sebesar 0.197. Sementara itu, algoritma *Polynomial Regression* menghasilkan nilai MSE sebesar 2612.291, RMSE sebesar 51.111, MAE sebesar 36.943, dan R^2 sebesar 0.019.

Selanjutnya, pada percobaan kedua digunakan perbandingan data sebesar 80:20, yaitu 80% data sebagai data pelatihan (training) dengan jumlah 660 data dan 20% data sebagai data pengujian (testing) sebanyak 164 data. Hasil dari percobaan dengan perbandingan kedua tersebut disajikan pada tabel berikut.

Tabel 6. *Split Validation* 80:20

Model	MSE	RMSE	MAE	R2
Linear Regression	2283.499	47.786	33.540	0.200
Polynomial Regression	2780.005	52.726	38.850	0.027

Berdasarkan tabel di atas, pada skema Split Validation 80:20, algoritma *Linear Regression* memperoleh nilai MSE sebesar 2283.499, RMSE sebesar 47.786, MAE sebesar 33.540, serta nilai R^2 sebesar 0.200. Sementara itu, algoritma *Polynomial Regression* menghasilkan nilai MSE sebesar 2780.005, RMSE sebesar 52.726, MAE sebesar 38.850, dan R^2 sebesar 0.027.

Selanjutnya, pada percobaan ketiga digunakan perbandingan data sebesar 90:10, yaitu 90% data sebagai data pelatihan (*training*) sebanyak 742 data dan 10% data sebagai data pengujian (*testing*) sejumlah 82 data. Hasil dari percobaan dengan perbandingan tersebut disajikan pada tabel berikut.

Tabel 7. *Split Validation* 90:10

Model	MSE	RMSE	MAE	R2
Linear Regression	2228.533	47.207	33.230	0.189
Polynomial Regression	2681.310	51.781	37.928	0.025

Berdasarkan tabel di atas, pada skema Split Validation 90:10, algoritma *Linear Regression* memperoleh nilai MSE sebesar 2228.533, RMSE sebesar 47.207, MAE sebesar 33.230, serta nilai R^2 sebesar 0.189. Sementara itu, algoritma *Polynomial Regression* menghasilkan nilai MSE sebesar 2681.310, RMSE sebesar 51.781, MAE sebesar 37.928, dan R^2 sebesar 0.025.

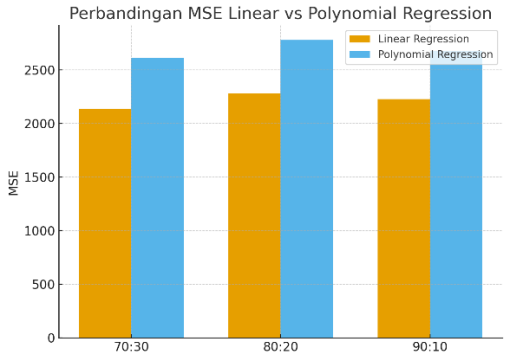
Selanjutnya, disajikan tabel perbandingan hasil evaluasi dari ketiga skema *Split Validation* yang telah dilakukan, yaitu 70:30, 80:20, dan 90:10, sebagaimana ditampilkan pada tabel berikut.

4.5 Evaluasi Performa Algoritma

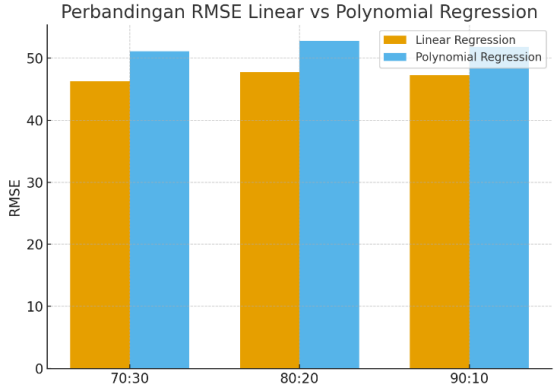
Tabel 8. Perbandingan *Split Validation*

Model	Rasio	MSE	RMSE	MAE	R2
Linear Regression	70:30	2139.124	46.251	31.955	0.197
	80:20	2283.499	47.786	33.540	0.200
	90:10	2228.533	47.207	33.230	0.189
Polynomial Regression	70:30	2612.291	51.111	36.943	0.019
	80:20	2780.005	52.726	38.850	0.027
	90:10	2681.310	51.781	37.928	0.025

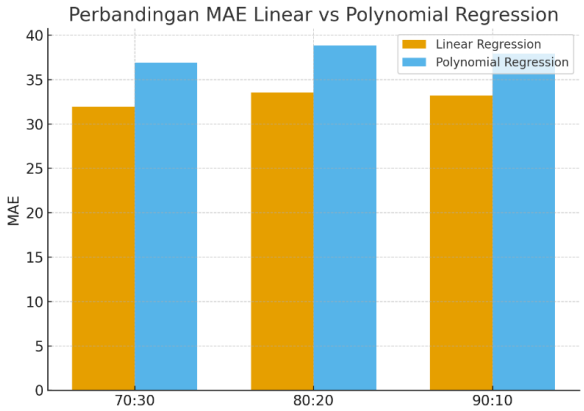
Berikut adalah tampilan dalam bentuk grafik,



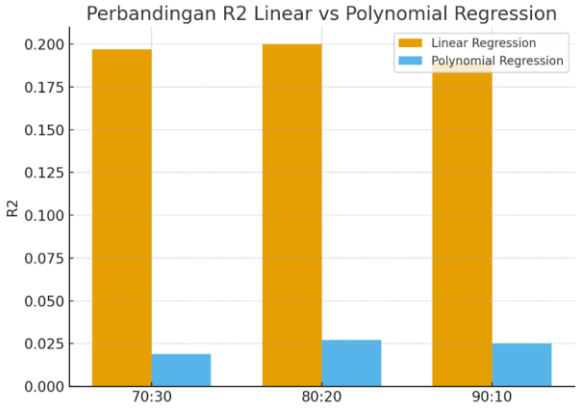
Gambar 2. Perbandingan MSE pada *Linear Regression* dengan *Polynomial Regression*



Gambar 3. Perbandingan RMSE pada *Linear Regression* dengan *Polynomial Regression*



Gambar 4. Perbandingan MAE pada *Linear Regression* dengan *Polynomial Regression*



Gambar 5. Perbandingan R2 pada *Linear Regression* dengan *Polynomial Regression*

Berdasarkan tabel diatas dan juga grafik, dapat diketahui bahwa *Linear Regression* lebih unggul dibanding *Polynomial Regression* dalam memprediksi durasi rawat inap pasien. Hal ini dibuktikan dengan nilai MSE, RMSE, dan MAE yang lebih rendah serta nilai R^2 yang lebih tinggi pada semua skenario split data.

Berdasarkan *Table 8* diatas dapat diketahui bahwa *Linear Regression* lebih unggul dibanding *Polynomial Regression* dalam memprediksi durasi rawat inap pasien. Hal ini dibuktikan dengan nilai MSE, RMSE, dan MAE yang lebih rendah serta nilai R^2 yang lebih tinggi pada semua skenario split data. Selanjutnya dilakukan analisis terhadap hasil prediksi yang dihasilkan oleh algoritma Linear Regression dan Polynomial Regression dengan menggunakan data pengujian (testing) sebesar 30% dari total dataset, yaitu sebanyak 247 data. Tabel berikut menyajikan contoh hasil prediksi dari kedua algoritma tersebut:

Tabel 9. Hasil Prediksi Data Testing

No	LOS Category	*Polynomial Regression*	*Linear Regression*
1	6	56,1878	78,2434
2	44	44,8919	43,4061
3	16	42,0271	34,0379
4	44	44,8919	50,7459
5	11	55,2022	37,142
6	6	42,0271	34,0379
7	22	43,9023	29,4554
8	3	47,873	19,3255
9	168	56,1084	109,155
10	22	43,8938	39,3388
...			
247	169	54,9642	75,0785

4.6 Pembahasan

Menggunakan algoritma *Linear Regression* dan *Polynomial Regression* untuk mengklasifikasi 30% data pada tahap pengujian, diatas merupakan 247 data yang telah diprediksi. Berdasarkan penelitian yang telah dilakukan, pengetahuan yang diperoleh adalah sebagai berikut:

1) Performa *Linear Regression* lebih konsisten dibanding *Polynomial Regression*. Pada semua skenario split data, *Linear Regression* menunjukkan nilai MSE, RMSE, dan MAE yang lebih rendah dibanding *Polynomial Regression*. Hal ini menandakan bahwa *Linear Regression* lebih mampu meminimalkan kesalahan prediksi dibanding *Polynomial Regression*.
2) Nilai R^2 *Linear Regression* lebih tinggi. *Linear Regression* secara konsisten memberikan nilai R^2 sekitar 0.18–0.20, sedangkan *Polynomial Regression* hanya menghasilkan nilai R^2 sekitar 0.019–0.027. Dengan demikian, kemampuan *Linear Regression* dalam menjelaskan variasi data target lebih baik daripada *Polynomial Regression.*

3) *Polynomial Regression* tidak menunjukkan keunggulan. Alih-alih meningkatkan akurasi, penggunaan *Polynomial Regression* justru meningkatkan error (MSE, RMSE, MAE) dan menurunkan nilai R^2. Hal ini mengindikasikan bahwa model polinomial yang digunakan mungkin mengalami *overfitting* pada data *training* atau kurang sesuai dengan karakteristik data.
4) Stabilitas model terhadap variasi split data. *Linear Regression* menunjukkan kestabilan performa meskipun dilakukan variasi split data (70:30, 80:20, dan 90:10). Nilai MSE, RMSE, MAE, dan R^2 tidak mengalami perubahan yang signifikan, menandakan model lebih *robust* terhadap proporsi data training dan testing.

Berdasarkan hasil eksperimen yang telah dilakukan menggunakan algoritma *Linear Regression* dan *Polynomial Regression* dengan proporsi data pengujian sebesar 30% (247 data testing), diperoleh sejumlah temuan penting terkait kinerja kedua algoritma dalam memprediksi durasi rawat inap pasien (*Length of Stay*/LoS). Pembahasan ini tidak hanya mengulas hasil penelitian secara internal, tetapi juga mengaitkannya dengan hasil-hasil penelitian terdahulu yang relevan, guna memperkuat integrasi temuan ke dalam pengembangan bidang ilmu data mining di bidang kesehatan. Hasil penelitian menunjukkan bahwa *Linear Regression* memiliki performa yang lebih konsisten dan unggul dibandingkan *Polynomial Regression* pada seluruh skenario pembagian data (70:30, 80:20, dan 90:10). *Linear Regression* secara konsisten menghasilkan nilai MSE, RMSE, dan MAE yang lebih rendah, serta nilai koefisien determinasi (R^2) yang lebih tinggi, yaitu berada pada kisaran 0,18–0,20. Sebaliknya, Polynomial Regression menunjukkan nilai error yang lebih besar dan nilai R^2 yang relatif rendah, yaitu hanya berkisar antara 0,019–0,027. Temuan ini memperkuat hasil penelitian terdahulu yang menyatakan bahwa model regresi sederhana sering kali lebih efektif pada data medis yang bersifat heterogen dan memiliki hubungan variabel yang relatif linear. Penelitian Eskandari et al. [2] serta Ba-Aoum et al. [4] menunjukkan bahwa pendekatan regresi linear mampu memberikan performa yang stabil dalam memodelkan *Length of Stay*, khususnya ketika variabel independen terdiri dari faktor demografis dan kategori layanan medis. Dengan demikian, hasil penelitian ini sejalan dengan temuan sebelumnya yang menunjukkan bahwa kesederhanaan model Linear Regression menjadi keunggulan tersendiri ketika diterapkan pada data rumah sakit yang bersifat kompleks dan memiliki tingkat variasi yang tinggi.

Selain itu, stabilitas *Linear Regression* terhadap variasi rasio data training dan testing juga sejalan dengan temuan Stone et al. [8], yang menekankan bahwa model prediksi LoS yang baik harus memiliki performa yang konsisten pada berbagai kondisi pembagian data. Nilai error dan R^2 yang relatif tidak mengalami fluktuasi signifikan pada penelitian ini menunjukkan bahwa *Linear Regression* lebih *robust* dan memiliki generalisasi yang lebih baik dibandingkan *Polynomial Regression*. Lebih lanjut, rendahnya nilai R^2 pada kedua model (terutama *Polynomial Regression*) juga sejalan dengan temuan Ansari et al. [9], yang menyatakan bahwa prediksi Length of Stay merupakan permasalahan yang kompleks dan sulit dijelaskan sepenuhnya hanya dengan variabel demografis dan administratif. Hal ini mengindikasikan bahwa masih terdapat faktor-faktor lain yang berpengaruh terhadap LoS, seperti kondisi klinis pasien, tingkat keparahan penyakit, serta tindakan medis yang diterima, yang belum dimasukkan dalam model.

Dengan demikian, kontribusi utama penelitian ini terletak pada penguatan bukti empiris bahwa *Linear Regression* tetap relevan dan efektif untuk memprediksi durasi rawat inap pasien pada dataset rumah sakit dengan karakteristik heterogen dan variabel terbatas. Penelitian ini juga memberikan klarifikasi bahwa *Polynomial Regression* tidak selalu unggul, khususnya ketika data tidak memiliki pola nonlinier yang kuat. Temuan ini memperkaya kumpulan penelitian terdahulu dengan menunjukkan bahwa pemilihan algoritma prediksi LoS harus disesuaikan dengan karakteristik data, bukan semata-mata berdasarkan kompleksitas model. Secara keseluruhan, hasil penelitian ini berkontribusi dalam mengintegrasikan temuan-temuan sebelumnya dan memberikan dasar empiris bagi pengembangan model prediksi *Length of Stay* yang lebih sederhana, stabil, dan mudah diimplementasikan dalam lingkungan rumah sakit, khususnya di negara berkembang dengan keterbatasan data klinis yang detail.

5. Simpulan

Berdasarkan hasil penerapan metode regresi linear berganda pada dataset yang digunakan dalam penelitian ini, maka dapat disimpulkan *Linear Regression* lebih unggul dibanding *Polynomial Regression* dalam memprediksi durasi rawat inap pasien. Hal ini

dibuktikan dengan nilai MSE, RMSE, dan MAE yang lebih rendah serta nilai R^2 yang lebih tinggi pada semua skenario split data. *Polynomial Regression* tidak memberikan hasil yang lebih baik. Model ini justru menghasilkan *error* yang lebih besar dan nilai R^2 yang lebih rendah, sehingga kurang efektif digunakan untuk memprediksi durasi rawat inap pasien. Kinerja *Linear Regression* relatif stabil pada berbagai skenario pembagian data (70:30, 80:20, dan 90:10), menunjukkan bahwa model ini lebih sesuai dengan karakteristik dataset yang digunakan.

Berdasarkan hasil penelitian, ada beberapa hal yang bisa dilakukan untuk mengembangkan penelitian ini ke depannya, dapat mengunakan data yang lebih banyak dan bervariasi, misalnya menambah jumlah data dan variasinya akan membantu model membuat prediksi yang lebih akurat dan tidak mudah salah. Tambahkan faktor-faktor lain yang relevan, misalnya data medis yang lebih lengkap atau informasi demografis pasien, agar model punya lebih banyak bahan pertimbangan untuk memprediksi. Bandingkan dengan metode lain, misalnya mencoba gunakan metode prediksi lain seperti *Random Forest* atau *Gradient Boosting*, lalu lihat apakah hasilnya lebih baik dari regresi linear berganda.

Daftar Referensi

[1] A. Rezaianzadeh, M. Dastoorpoor, M. Sanaei, C. Salehnasab, M. Mohammadi, J. M, dan A. Mousavizadeh, "Predictors of length of stay in the coronary care unit in patient with acute coronary syndrome based on data mining methods", *Clinical Epidemiology and Global Health,* vol. 8, no. 2, pp.383–388, Juni 2020.

[2] M. Eskandari, A. H. Alizadeh Bahmani, H. A. Mardani-Fard, I. Karimzadeh, N. Omidifar, dan P. Peymani, "Evaluation of factors that influenced the length of hospital stay using data mining techniques", *BMC Med Inform Decis Mak*, vol. 22, pp.1, Des 2022.

[3] T. Kelley *dkk.*, *Interventions To Decrease Hospital Length of Stay Technical Brief Number 40 R*, vol. 21-EHC015. Rockville: AHRQ, 2021. Diakses: 9 Oktober 2024. [Daring]. Tersedia pada: https://www.ncbi.nlm.nih.gov/books/NBK574435/pdf/Bookshelf_NBK574435.pdf

[4] M. Ba-Aoum, N. Hosseinichimeh, K. P. Triantis, K. Pasupathy, M. Sir, dan D. Nestler, "Statistical analysis of factors influencing patient length of stay in emergency departments," *International Journal of Industrial Engineering and Operations Management*, vol. 5, no. 3, pp.220–239, Agu 2023, doi: 10.1108/ijieom-10-2022-0056.

[5] S. Lam *dkk.*, "*Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit*," 2006. Diakses: 9 Oktober 2024. [Daring]. Tersedia pada: https://www.sciencedirect.com/science/article/pii/S1470211824006717

[6] W. Chango, J.A. Lara, R. Cerezo, C. Romero, "A review on data fusion in multimodal learning analyticsand educational data mining*"*, *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 4, pp.1-19, April 2022

[7] B. Wibisana, Y. Hendra Pratama, dan H. Purnomo, "Forecasting of Rice Production Using a Linear Regression and Polynomial Regression", *Journal of Informatics and Communications Technology*, vol. 6, no. 1, pp.33–41, Juni 2024, doi: 10.52661.

[8] K. I. Stone, R. I. Zwiggelaar, P. Jones, dan N. Mac Parthalá in, "A systematic review of the prediction of hospital length of stay: Towards a unified framework", *PLOS Digital Health*, vol. 1, no. 4, pp.1-38, Apr 2022, doi: 10.1371/JOURNAL.PDIG.0000017.

[9] M. Shahid Ansari dkk., "Identification of predictors and model for predicting prolonged length of stay in dengue patients", *Health Care Management Science*, vol. 24, no. 4, pp.786–798, Des 2021, doi: 10.1007/s10729-021-09571-3.

[10] K. Alghatani, N. Ammar, A. Rezgui, A. S. Nejad, "Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation", *JMIR Publications Advancing Digital Health & Open Science*, vol. 9, no. 5, pp.1–23, May 2023, doi: 10.2196/21347

[11] A. Mehran, A. Eltahawi, M.H.A. Elaziz, M.N.A. Elwhab, "Predicting length of stay in hospitals intensive care unit using general admission features", *Ain Shams Engineering Journal*, vol. 12, no. 4, pp.3691-3702, 2021

[12] G. Xia, Y. Bi, dan C. Wang, "Optimization design of passive residual heat removal system based on improved genetic algorithm," *Annals of Nuclear Energy*, vol. 189, Sep 2023, doi: 10.1016/j.anucene.2023.109859.

[13] S. Lai, X. Hu, H. Xu, Z. Ren, dan Z. Liu, "Multimodal sentiment analysis: A survey", Displays, vol. 80, Des 2023, doi: 10.1016/j.displa.2023.102563.

[14] D. K. Sharma, M. Chatterjee, G. Kaur, dan S. Vavilala, “Deep learning applications for disease diagnosis”, *Deep Learning for Medical Applications with Unique Data*, hlm. 31–51, Jan 2022, doi: 10.1016/B978-0-12-824145-5.00005-8.
[15] G. Romeo, “Data analysis for business and economics”, *Elements of Numerical Mathematical Economics with Excel*, pp. 695–761, 2020, doi: 10.1016/B978-0-12-817648-1.00013-X.