Prediksi Lama Rawat Inap (*Length of Stay*) Pasien Rumah Sakit Menggunakan Light Gradient Boosting Machine

Bagus Cipta Pratama
Muhammad Akmal Fauzan
Nasya Putri Raudhah Dahlan
Universitas Gadjah Mada
Yogyakarta, Indonesia

July 10, 2025

Abstract

Manajemen sumber daya rumah sakit yang efisien adalah kunci untuk perawatan pasien yang berkualitas dan keberlanjutan finansial. Lama Rawat Inap atau Length of Stay (LOS) menjadi metrik vital dalam manajemen ini. Laporan ini bertujuan untuk membangun model machine learning yang mampu memprediksi LOS secara akurat berdasarkan data historis pasien. Dengan menggunakan dataset yang berisi 100.000 catatan pasien, berbagai model regresi dievaluasi, termasuk Regresi Polinomial, Regresi Linear, Random Forest, SVR, dan LightGBM. Setelah melalui proses Analisis Data Eksplorasi (EDA) yang mendalam, rekayasa fitur, dan optimasi model menggunakan GridSearchCV, model LightGBM terpilih sebagai yang terbaik dengan performa R-squared (R^2) mencapai 0.97 pada data uji. Analisis juga mengidentifikasi bahwa kompleksitas klinis pasien (jumlah komorbiditas), riwayat kunjungan ulang, dan beberapa indikator laboratorium menjadi prediktor paling signifikan. Model ini menunjukkan potensi besar untuk diimplementasikan dalam sistem informasi rumah sakit guna membantu alokasi sumber daya yang lebih proaktif dan efisien.

1 Pendahuluan

1.1 Latar Belakang

Manajemen layanan kesehatan menempatkan Lama Rawat Inap atau Length of Stay (LOS) sebagai salah satu metrik kinerja utama keberhasilan perawatan pasien di rumah sakit [1]. LOS memiliki dampak signifikan terhadap biaya perawatan dan kualitas hasil klinis pasien.

Studi oleh Yuli Atomojo dkk. (2024) menyoroti dampak ekonomi dari sistem jaminan kesehatan terhadap praktik rumah sakit. Terdapat penurunan signifikan pada rata-rata lama rawat inap pasien dari 5,26 hari (2018) menjadi 4,67 hari (2022), angka yang lebih rendah dari durasi ideal. Hal ini sejalan dengan fakta bahwa 72,33 % pasien rawat inap pada 2022 merupakan peserta BPJS Kesehatan. Karena sistem klaim BPJS berbasis paket (INA-CBG's), rumah sakit secara strategis berupaya mempersingkat durasi perawatan untuk menjaga kesehatan finansial institusi, dengan memastikan pasien pulang dalam kondisi sembuh atau telah mendapat rujukan yang sesuai.

Rumah sakit dituntut efisien secara biaya tanpa mengorbankan keselamatan dan kualitas perawatan. Oleh karena itu, prediksi Length of Stay (LOS) yang akurat menjadi krusial untuk mendukung alokasi sumber daya seperti tempat tidur, staf, dan logistik. Dengan kemajuan teknologi, machine learning menawarkan solusi efektif melalui analisis data historis pasien untuk menghasilkan prediksi yang objektif dan andal.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah untuk penelitian ini adalah sebagi berikut:

1. Bagaimana membangun model machine learning yang mampu memprediksi lama rawat inap (Length of Stay) pasien secara akurat berdasarkan dataset yang tersedia?
2. Apa saja faktor-faktor yang menjadi prediktor paling signifikan dalam menentukan lama rawat inap seorang pasien berdasarkan model yang dikembangkan?

1.3 Tujuan

Sejalan dengan rumusan masalah di atas, tujuan dari penelitian ini adalah:

1. Melakukan pra-pemrosesan data dan feature engineering pada dataset.
2. Membangun dan mengevaluasi beberapa model machine learning untuk menemukan model dengan performa terbaik dalam memprediksi lama rawat inap.
3. Mengidentifikasi dan menganalisis variabel-variabel yang memiliki pengaruh terbesar terhadap hasil prediksi lama rawat inap untuk memberikan wawasan yang dapat ditindaklanjuti bagi manajemen rumah sakit.

2 Kajian Teori

2.1 Length Of Stay

Length of Stay (LOS) atau Lama Rawat Inap adalah sebuah metrik klinis fundamental dalam manajemen pelayanan kesehatan. Menurut Definitive Healthcare (2025), LOS secara definitif mengukur durasi atau rentang waktu yang dihabiskan oleh seorang pasien di rumah sakit, terhitung sejak pasien tersebut masuk (admittance) hingga saat dipulangkan (discharge). Terdapat berbagai faktor yang dapat memengaruhi durasi rawat inap seorang pasien, di antaranya adalah keberadaan penyakit penyerta (comorbidities), tingkat keparahan atau kegawatan kondisi pasien (patient acuity atau case mix), kinerja kualitas pelayanan rumah sakit, hingga rasio dan ketersediaan staf medis.

2.2 LightGBM (Light Gradient Boosting Machine)

LightGBM adalah sebuah kerangka kerja gradient boosting yang menggunakan algoritma berbasis pohon keputusan (tree-based). Algoritma ini dirancang untuk menjadi sangat efisien dan dapat diskalakan. Beberapa fitur kunci yang membedakan LightGBM dari algoritma boosting tradisional lainnya adalah

- **Pertumbuhan pohon secara leaf-wise**: Alih-alih menumbuhkan pohon secara level-wise (rata kiri-kanan), LightGBM memilih cabang daun (leaf) dengan pengurangan loss terbesar untuk dikembangkan terlebih dahulu. Pendekatan ini sering kali menghasilkan penurunan error lebih tajam per iterasi (Soomro et al., 2024).
- **Gradient-based One-Side Sampling (GOSS)**: Untuk mempercepat pelatihan tanpa mengorbankan akurasi, LightGBM menyimpan semua sampel dengan gradien besar (yang paling "berinformasi") dan melakukan undersampling pada sampel dengan gradien kecil, sehingga jumlah data yang diproses per iterasi berkurang secara signifikan (Soomro et al., 2024).

- **Exclusive Feature Bundling**: Fitur-fitur yang jarang bersamaan aka "saling eksklusif" dapat digabungkan menjadi satu fitur gabungan, sehingga dimensionalitas fitur berkurang tanpa menimbulkan kehilangan informasi penting (Soomro et al., 2024).
- **Histogram-based algorithms**: LightGBM membagi rentang nilai fitur ke dalam bin histogram, sehingga menghitung gradien dan split point menjadi lebih cepat dan hemat memori dibandingkan cara tradisional yang mengulang melalui setiap nilai sampel (GeeksforGeeks, 2025).

Karena kombinasi inovasi-inovasi di atas, LightGBM mampu melatih model dengan cepat dan menggunakan lebih sedikit memori, sambil tetap mempertahankan (bahkan sering kali meningkatkan) ketepatan prediksi dibandingkan framework boosting lainnya. Hal ini menjadikannya pilihan populer untuk berbagai tugas regresi dan klasifikasi pada dataset skala besar.

3 Solusi Usulan

Solusi usulan yang dapat dilakukan berdasarkan Rumusan Masalah dan Tujuan yang sudah didefinisikan adalah sebagai berikut

1. **Pra-pemrosesan Data**: Membersihkan dan mempersiapkan dataset. Ini termasuk mengonversi kolom tanggal, membersihkan kolom `rcount` dari nilai '5+' menjadi 5, serta memetakan variabel kategorikal seperti `gender` dan `facid` ke dalam format numerik.
2. **Analisis Data Eksplorasi (EDA) dan Rekayasa Fitur**: Melakukan analisis mendalam untuk memahami pola data. Selama fase ini, sebuah fitur baru, `num_conditions`, direkayasa dengan menjumlahkan 11 kolom kondisi medis.
3. **Pemodelan dan Evaluasi**: Lima model regresi (Regresi Polinomial, Regresi Linear, LightGBM, Random Forest, dan SVR) dibangun dan dievaluasi menggunakan metrik R^2, MSE, dan RMSE untuk menemukan baseline kinerja.
4. **Optimasi Model**: Model dengan kinerja terbaik (LightGBM) dioptimasi menggunakan `GridSearchCV` untuk menemukan kombinasi hyperparameter yang optimal.

4 Hasil Eksperimen dan Pengujian

4.1 Feature Engineering

Tahap rekayasa fitur dilakukan untuk mengubah data mentah menjadi representasi yang lebih informatif dan sesuai untuk algoritma machine learning. Langkah-langkah ini tidak hanya membersihkan data, tetapi juga menciptakan variabel baru yang mampu menangkap pola-pola yang relevan secara lebih eksplisit.

Langkah rekayasa fitur utama adalah pembuatan variabel `num_conditions` untuk mengukur kompleksitas klinis pasien. Fitur ini terbukti memiliki korelasi kuat dengan lama rawat inap, seperti yang ditunjukkan pada analisis lanjutan. Selain itu, beberapa langkah rekayasa dan transformasi fitur lainnya juga dilakukan:

- **Agregasi Kondisi Medis (`num_conditions`)**: Fitur ini dibuat dengan menjumlahkan 11 kolom biner yang menandakan keberadaan penyakit penyerta, seperti `dialysisrenalendstage`, `asthma`, `pneum`, dan lainnya. Tujuannya adalah untuk menciptakan satu variabel tunggal yang merepresentasikan beban komorbiditas total pada seorang pasien.

- **Encoding Variabel Kategorikal**: Untuk memungkinkan model memproses data non-numerik, variabel kategorikal diubah menjadi angka. Kolom `gender` dipetakan dari 'F' dan 'M' menjadi 1 dan 0 , sementara `facid` dipetakan dari 'A' hingga 'E' menjadi nilai integer 1 hingga 5.
- **Pembersihan dan Konversi Numerik (`rcount`)**: Kolom `rcount`, yang merepresentasikan riwayat kunjungan ulang, memiliki nilai '5+'. Nilai ini dibersihkan menjadi '5' agar seluruh kolom dapat diubah menjadi tipe data integer, sehingga dapat digunakan sebagai fitur numerik yang mengindikasikan riwayat kesehatan pasien.
- **Binning Variabel Target (untuk Visualisasi)**: Khusus untuk keperluan analisis visual pada *pair plot*, variabel target kontinu `lengthofstay` dikelompokkan menjadi tiga kategori: 'Pendek', 'Menengah', dan 'Panjang'. Langkah ini tidak digunakan untuk melatih model regresi, tetapi sangat membantu dalam mengidentifikasi bagaimana sebaran data klinis berbeda antar kelompok pasien.

4.2 Analisis Data Eksplorasi (Exploratory Data Analysis)

Tahap Analisis Data Eksplorasi (EDA) dilakukan untuk menginvestigasi karakteristik fundamental dataset, mengidentifikasi pola, serta memvalidasi asumsi-asumsi awal. Tujuan dari tahap ini adalah untuk mengekstraksi wawasan statistik yang dapat menginformasikan proses rekayasa fitur dan pemilihan model. Analisis disajikan secara sistematis, dimulai dari pemeriksaan variabel target hingga eksplorasi hubungan multivariat yang kompleks.

Analisis Distribusi Variabel Target Investigasi awal difokuskan pada variabel dependen, yaitu `lengthofstay`. Visualisasi distribusi variabel ini, seperti yang disajikan pada Gambar 1, menunjukkan adanya kemiringan positif yang signifikan (*positive skewness*). Mayoritas pasien memiliki durasi rawat inap yang pendek, dengan frekuensi yang menurun secara eksponensial seiring bertambahnya durasi. Observasi ini mengindikasikan bahwa sebagian besar kasus perawatan di fasilitas ini bersifat jangka pendek. Berdasarkan distribusi ini, investigasi selanjutnya bertujuan untuk mengidentifikasi faktor-faktor determinan yang membedakan antara kelompok pasien dengan durasi rawat inap yang berbeda.

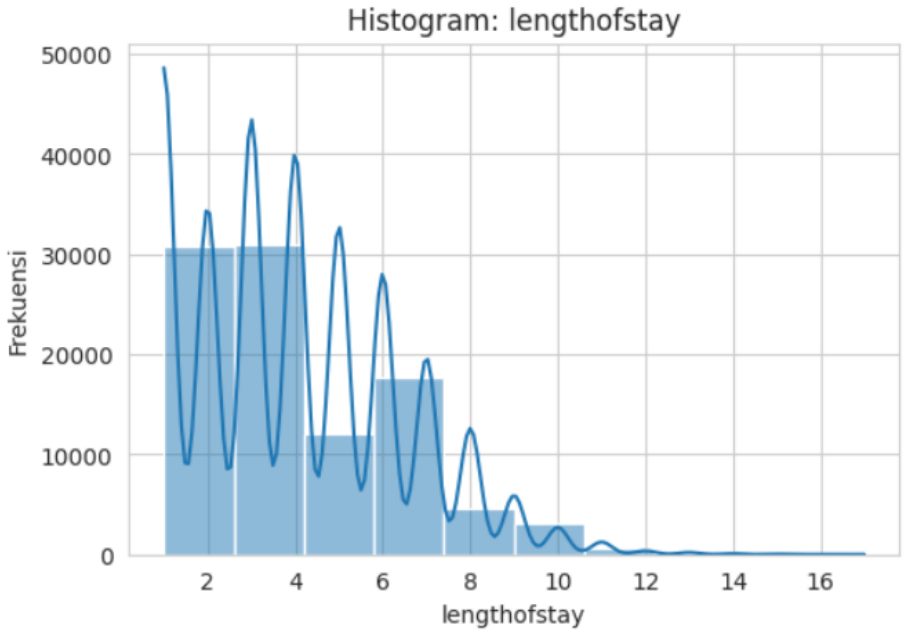

Figure 1: Histogram dari `lengthofstay`, menunjukkan distribusi yang condong ke kanan.

Uji Hipotesis Pengaruh Komorbiditas terhadap Lama Rawat Inap Hipotesis pertama yang diuji adalah adanya hubungan positif antara beban komorbiditas pasien dengan lama rawat inap. Beban komorbiditas diukur melalui agregasi jumlah kondisi medis penyerta (`num_conditions`). Analisis menggunakan diagram kotak (boxplot), sebagaimana diilustrasikan pada Gambar 2, menunjukkan korelasi positif yang kuat dan monoton. Terlihat bahwa nilai median `lengthofstay` meningkat secara konsisten seiring dengan bertambahnya jumlah total kondisi medis. Temuan ini memberikan bukti empiris yang kuat bahwa **kompleksitas klinis pasien merupakan prediktor signifikan terhadap durasi rawat inap**.

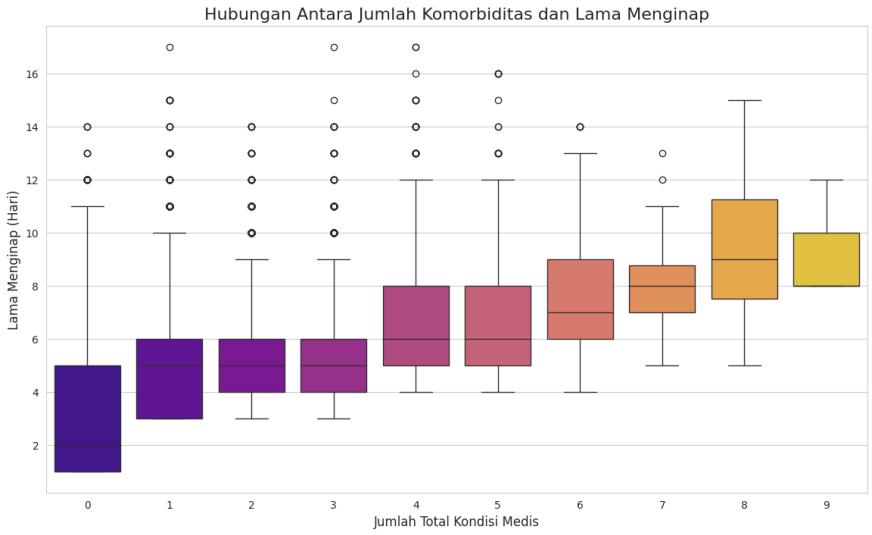

Figure 2: Hubungan Antara Jumlah Komorbiditas dan Lama Menginap.

Analisis Pengaruh Riwayat Kunjungan terhadap Lama Rawat Inap Selanjutnya, dianalisis pengaruh riwayat kesehatan pasien, yang direpresentasikan oleh variabel jumlah kunjungan ulang (`rcount`), terhadap lama rawat inap saat ini. Variabel ini dapat berfungsi sebagai proksi untuk kondisi kronis atau keparahan penyakit yang persisten. Hasil analisis pada Gambar 3 mengonfirmasi adanya tren positif yang serupa. Pasien dengan frekuensi kunjungan ulang yang lebih tinggi menunjukkan kecenderungan untuk memiliki durasi rawat inap yang lebih panjang secara statistik. Hal ini menguatkan kesimpulan sebelumnya bahwa **faktor historis pasien, selain kondisi akut saat ini, memiliki nilai prediktif yang esensial**.

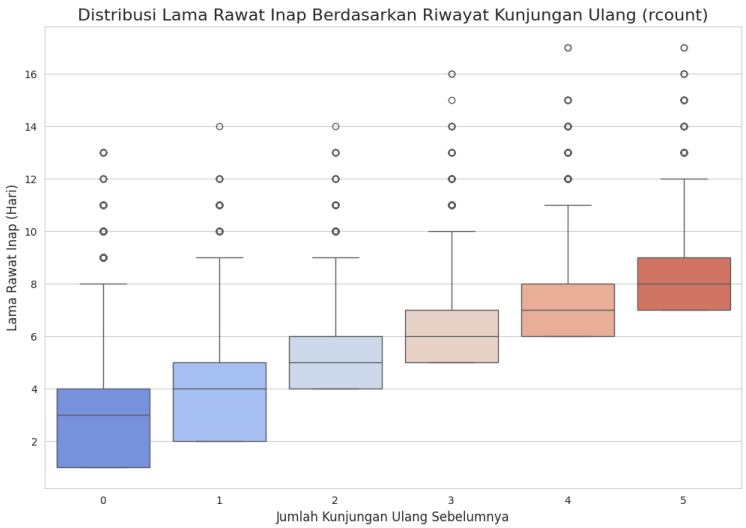

Figure 3: Distribusi Lama Rawat Inap Berdasarkan Riwayat Kunjungan Ulang (`rcount`).

Eksplorasi Hubungan Multivariat Antar Variabel Klinis Untuk mengkaji interaksi yang lebih kompleks, dilakukan analisis *pair plot* terhadap beberapa variabel klinis kunci seperti `hematocrit`, `glucose`, `bmi`, dan `pulse` (Gambar 4). Meskipun terdapat sedikit perbedaan dalam distribusi marginal (diagonal) antar kategori `lengthofstay`, plot sebar (di luar diagonal) menunjukkan tumpang tindih yang substansial. Tidak ditemukan adanya hubungan linear atau pemisahan kelas yang jelas dari kombinasi dua variabel mana pun. Observasi ini mengimplikasikan bahwa hubungan antara indikator-indikator klinis dengan lama rawat inap bersifat **non-linear dan multivariat**, sehingga pemodelan menggunakan algoritma sederhana kemungkinan tidak akan menghasilkan performa yang optimal.

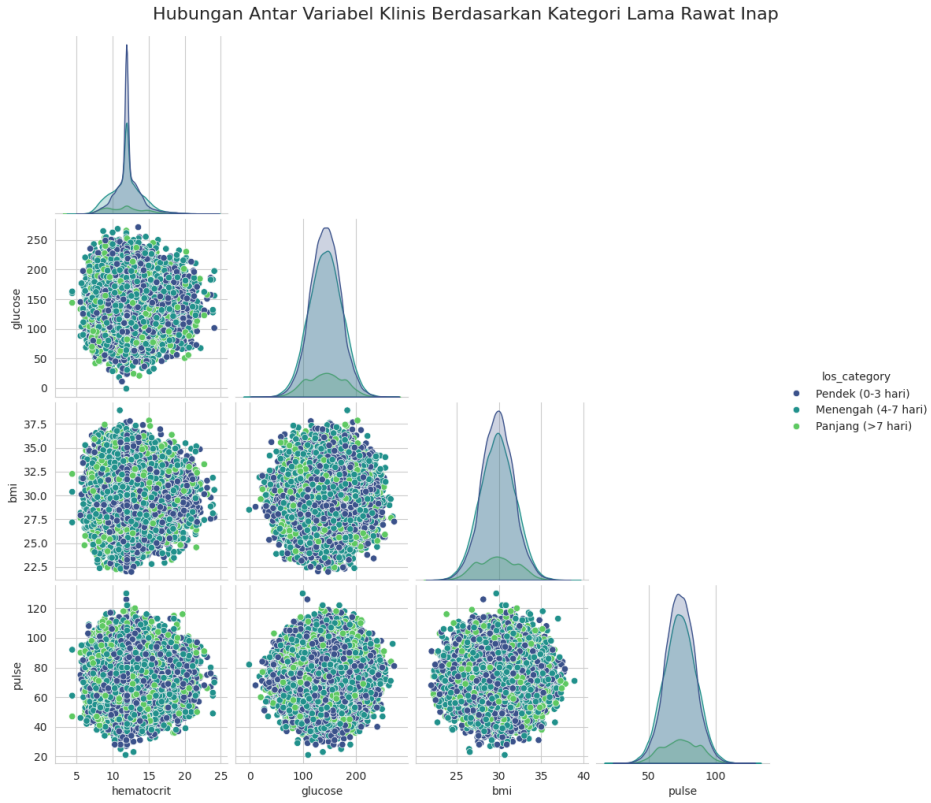

Figure 4: Hubungan Antar Variabel Klinis Berdasarkan Kategori Lama Rawat Inap.

Analisis Stabilitas Data Berbasis Waktu Sebagai tahap akhir EDA, dilakukan analisis deret waktu terhadap jumlah admisi pasien per bulan untuk mengevaluasi stabilitas operasional selama periode pengumpulan data. Grafik pada Gambar 5 menunjukkan bahwa volume pasien masuk relatif stasioner sepanjang tahun 2012, tanpa fluktuasi musiman yang signifikan. Hal ini memberikan keyakinan bahwa dataset berasal dari lingkungan operasional yang konsisten, sehingga valid untuk pemodelan. Namun, teridentifikasi adanya penurunan admisi yang drastis pada awal tahun 2013, yang mengindikasikan bahwa data pada periode tersebut tidak lengkap dan perlu ditangani sebelum tahap pemodelan.

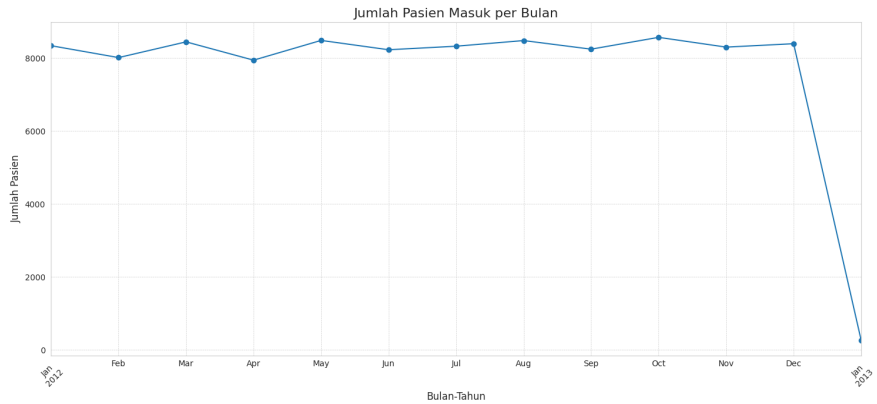

Figure 5: Jumlah Pasien Masuk per Bulan.

Secara sintesis, tahapan EDA ini telah menghasilkan beberapa kesimpulan fundamental. Pertama, mayoritas kasus

adalah rawat inap jangka pendek, dengan prediktor utama perpanjangan durasi adalah kompleksitas klinis dan riwayat kunjungan pasien. Kedua, interaksi antar variabel klinis bersifat kompleks dan non-linear, yang membenarkan penggunaan model machine learning yang lebih canggih. Ketiga, data yang akan digunakan untuk pemodelan terbukti stabil sepanjang periode observasi utama. Dengan fondasi pemahaman data yang solid ini, penelitian dapat dilanjutkan ke tahap pengembangan model prediktif.

5 Analisis Hasil

Bagian ini menyajikan hasil kuantitatif dari perbandingan model.

Table 1: Perbandingan Kinerja Awal Model Regresi pada Data Validasi.

MODEL	MSE	RMSE	R^2 SCORE
Regresi Polinomial (deg=2)	0.73	0.85	0.87
Regresi Linear	1.36	1.17	0.76
LightGBM	**0.21**	**0.46**	**0.96**
Random Forest	0.44	0.66	0.92
SVR	0.55	0.74	0.90

Tabel 1 menunjukkan bahwa model LightGBM memberikan kinerja terbaik secara signifikan dibandingkan empat model lainnya, dengan R^2 Score mencapai 0.96 pada evaluasi awal. Kinerjanya yang superior menjadikannya pilihan utama untuk optimasi lebih lanjut.

Model LightGBM terbaik, setelah dioptimalkan melalui GridSearchCV, dievaluasi pada data uji untuk mengukur kemampuannya dalam generalisasi. Model akhir ini berhasil mencapai **R^2 Score sebesar 0.97**, MSE 0.19, dan RMSE 0.44.

Untuk menganalisis perilaku model lebih lanjut, kurva pembelajaran (Gambar 6) di-plot. Kurva ini menunjukkan bahwa seiring bertambahnya data pelatihan, skor validasi (hijau) terus meningkat dan mendekati skor pelatihan (biru). Celah yang kecil antara kedua kurva menandakan bahwa model memiliki varians yang rendah dan tidak mengalami overfitting yang signifikan.

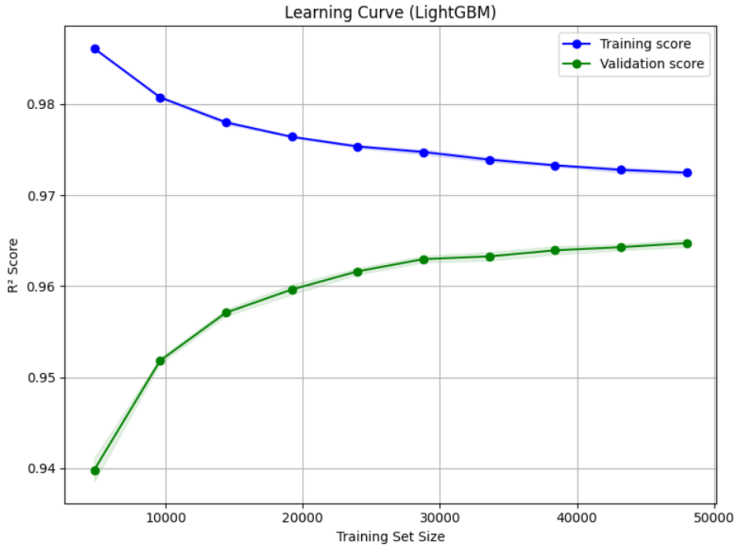

Figure 6: Kurva Pembelajaran (Learning Curve) dari Model LightGBM Terbaik.

Kesesuaian prediksi dengan nilai aktual divisualisasikan pada Gambar 7. Titik-titik data terlihat mengelompok sangat rapat di sekitar garis diagonal (garis ideal di mana prediksi sama dengan aktual). Hal ini secara visual mengonfirmasi akurasi tinggi dari model di seluruh rentang nilai 'lengthofstay'.

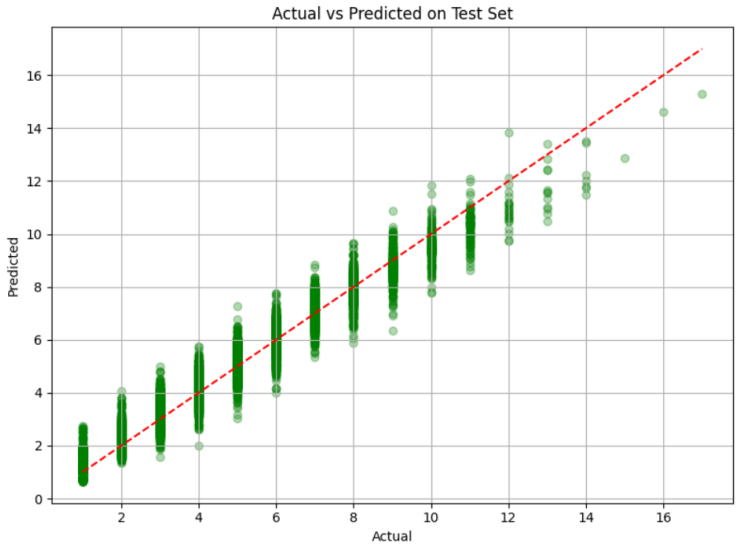

Figure 7: Plot Sebar (Scatter Plot) Nilai Aktual vs. Prediksi pada Data Uji.

Analisis Kepentingan Fitur (Feature Importance) Untuk melengkapi analisis dan menjawab rumusan masalah kedua, dilakukan evaluasi terhadap kontribusi relatif setiap variabel dalam proses pengambilan keputusan model. Gambar 8 menyajikan peringkat kepentingan fitur (*feature importance*) yang diekstraksi dari model LightGBM. Hasil ini memberikan wawasan yang jelas mengenai faktor-faktor yang paling dominan dalam memprediksi lama rawat inap.

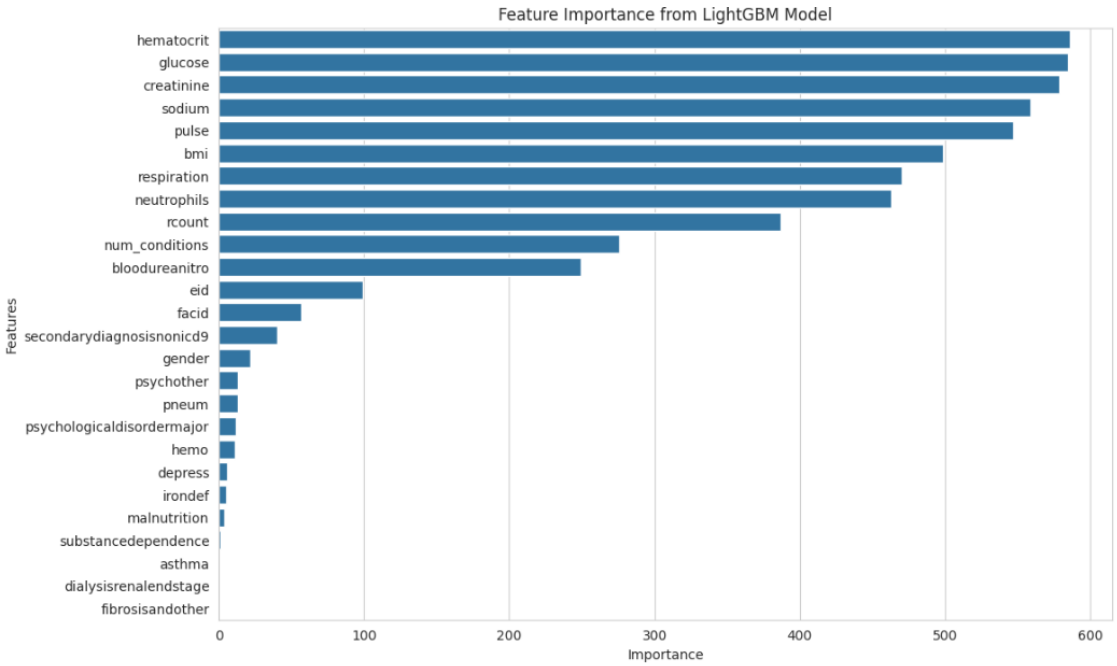

Figure 8: Peringkat Kepentingan Fitur dari Model LightGBM.

Secara konsisten, hasil laboratorium dan tanda-tanda vital pasien saat admisi menjadi prediktor dengan bobot tertinggi. Variabel seperti `hematocrit`, `glucose`, `creatinine`, `sodium`, dan `pulse` mendominasi peringkat teratas. Hal ini mengonfirmasi bahwa kondisi fisiologis dan metabolik akut pasien merupakan determinan utama yang menggerakkan prediksi model. Tingkat keparahan penyakit, yang tercermin secara kuantitatif dalam hasil lab ini, terbukti lebih informatif dibandingkan sekadar diagnosis kategorikal.

Selain itu, penting untuk dicatat bahwa fitur hasil rekayasa, `num_conditions`, dan fitur yang telah dipra-pemrosesan, `rcount`, juga menempati posisi yang signifikan. Ini memvalidasi hipotesis awal dari tahap EDA bahwa kompleksitas klinis (jumlah komorbiditas) dan riwayat kunjungan pasien adalah faktor prediktif yang krusial. Sebaliknya, beberapa kondisi medis spesifik secara individual seperti `asthma` atau `fibrosisandother` memiliki kepentingan yang lebih rendah. Hal ini kemungkinan besar terjadi karena dampak dari kondisi tersebut telah terwakili secara lebih efektif oleh gabungan fitur `num_conditions` dan indikator laboratorium yang lebih umum.

Analisis kepentingan fitur ini tidak hanya menegaskan validitas model, tetapi juga memberikan justifikasi klinis yang kuat terhadap hasil prediksi. Model tidak berfungsi sebagai 'kotak hitam' (*black box*), melainkan berhasil mengidentifikasi variabel-variabel yang secara medis relevan dalam penentuan lama rawat inap.

6 Kesimpulan dan Saran

Penelitian ini berhasil mencapai tujuannya dengan mengembangkan sebuah model machine learning berbasis LightGBM yang menunjukkan performa sangat tinggi dalam memprediksi lama rawat inap (Length of Stay) pasien, dengan perolehan R^2 Score mencapai 0.97 pada data uji. Kinerja ini secara signifikan melampaui model-model lain yang dievaluasi, seperti Random Forest dan Regresi Polinomial, mengukuhkan LightGBM sebagai solusi yang paling efektif untuk dataset ini.

Analisis mendalam terhadap model mengidentifikasi bahwa prediktor paling signifikan adalah serangkaian indikator klinis dan laboratorium objektif, seperti `hematocrit`, `glucose`, dan `creatinine`. Selain itu, fitur yang di-

rekayasa selama penelitian, yaitu `num_conditions` (yang mengukur beban komorbiditas) dan `rcount` (riwayat kunjungan ulang), juga terbukti memiliki bobot prediksi yang tinggi. Temuan ini menegaskan bahwa model tidak hanya akurat, tetapi juga mendasarkan prediksinya pada faktor-faktor yang relevan secara medis, yaitu kondisi fisiologis akut dan riwayat kesehatan kronis pasien.

Secara praktis, model prediktif ini memiliki potensi besar untuk diintegrasikan ke dalam sistem informasi manajemen rumah sakit. Dengan kemampuan memberikan estimasi LOS yang akurat sejak awal pasien masuk, pihak manajemen dapat melakukan alokasi sumber daya—seperti ketersediaan tempat tidur, penjadwalan staf medis, dan perencanaan logistik—secara lebih proaktif dan efisien. Hal ini dapat berujung pada optimalisasi alur pasien, pengurangan biaya operasional, dan pada akhirnya peningkatan kualitas pelayanan kesehatan.

Untuk pengembangan di masa depan, beberapa langkah disarankan. Pertama, melakukan **validasi prospektif** dengan menguji coba model secara langsung di lingkungan klinis (pilot implementation) untuk mengukur dampak dan keandalannya di dunia nyata. Kedua, **memperkaya dataset** dengan variabel-variabel yang lebih granular, seperti data diagnosis terstruktur menggunakan kode internasional (misalnya, ICD-10), catatan medis tekstual yang dapat dianalisis dengan teknik Natural Language Processing (NLP), serta data pengobatan pasien. Ketiga, penting untuk melakukan **analisis bias dan keadilan** (fairness analysis) untuk memastikan bahwa model memberikan prediksi yang akurat secara merata di seluruh kelompok demografis pasien dan tidak memperburuk disparitas kesehatan yang ada.

Acknowledgments

Kami mengucapkan terima kasih kepada penyelenggara Datathon Ristek Fasilkom UI dan semua pihak yang telah membantu serta mendukung pelaksanaan proyek ini.

References

[1] Yuli Atmojo, C., Elasari, Y., Adi Nugroho, T., et al. (n.d.). FAKTOR-FAKTOR YANG BERHUBUNGAN DENGAN LAMA HARI RAWAT PASIEN BPJS KESEHATAN DI RUMAH SAKIT MARDI WALUYO KOTA METRO. *Journal of Nursing Invention*. https://doi.org/10.33859/jni

[2] Definitive Healthcare. (2025). *Length of Stay Definition*. Diakses pada 5 Juli 2025, dari https://www.definitivehc.com/resources/glossary/length-of-stay

[3]Ahmed Soomro, A., Akmar Mokhtar, A., B Hussin, H., Lashari, N., Lekan Oladosu, T., Muslim Jameel, S., Inayat, M. (2024). Analysis of machine learning models and data sources to forecast burst pressure of petroleum corroded pipelines: A comprehensive review. Engineering Failure Analysis, 155. https://doi.org/10.1016/j.engfailanal.2023.107747

[4] LightGBM (Light Gradient Boosting Machine) - GeeksforGeeks. (n.d.). Retrieved July 8, 2025, from https://www.geeksforgeeks.org/machine-learning/lightgbm-light-gradient-boosting-machine/