LAPORAN EKSPERIMEN UTS

Machine Learning

Pembuatan Model Machine Learning dengan Algoritma Random Forest Regression untuk Memprediksi Biaya Tagih - oleh fasilitas kesehatan (provider) penyakit Tuberkolosis
Data studi kasus: BPJS Kesehatan Tahun 2022

Dibuat oleh:

Ardhan Anggana Prasetya 2310051

Mekatronika dan Kecerdasan Buatan

Universitas Pendidikan Indonesia

2025

I. Tujuan Eksperimen.

Tujuan eksperimen ini adalah untuk membuat model prediksi biaya tagih oleh fasilitas kesehatan terhadap pasien penderita tuberkulosis (TB) dengan memanfaatkan data Pasien BPJS di tahun 2021 (tb_2021_kepesertaan) dan data Fasilitas Kesehatan Rujukan Tingkat Lanjut (tb_fkrtl) berbentuk CSV. Eksperimen dimulai dengan proses pengambilan dua sumber data utama, yaitu tb_2021_kepesertaan dan tb_fkrtl, yang kemudian disatukan, dibersihkan, dan dilakukan rekayasa data untuk menghasilkan variabel yang relevan sebagai input pemodelan, seperti usia peserta dan durasi kunjungan. Pemilihan variabel pendukung dilakukan melalui analisis statistik, termasuk uji ANOVA, untuk mengidentifikasi faktor-faktor yang secara signifikan memengaruhi biaya tagih.

Selanjutnya, eksperimen difokuskan pada penerapan dan membandingkan berbagai algoritma regresi, meliputi Linear Regression, Ridge, Lasso, Decision Tree Regressor, Random Forest Regressor, dan Gradient Boosting Regressor. Setiap model dilatih dan dievaluasi menggunakan teknik validasi silang K-Fold untuk memperoleh performa prediksi yang optimal dan menghindari overfitting. Berbagai metrik evaluasi, seperti Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R^2), digunakan untuk menilai tingkat akurasi hasil prediksi model yang dibangun. Eksperimen juga mencakup analisis perbandingan nilai aktual dan nilai prediksi pada data uji, serta pengujian efek penggunaan teknik encoding pada variabel kategorikal.

Secara keseluruhan, tujuan eksperimen ini tidak hanya untuk menemukan model unggulan yang mampu memprediksi besaran biaya tagih secara akurat, tetapi juga untuk memberikan wawasan strategis mengenai variabel-variabel kunci yang berdampak signifikan terhadap biaya kesehatan. Hasil dari eksperimen ini diharapkan dapat digunakan oleh Badan Penyelenggara Jaminan Sosial Kesehatan untuk mendukung upaya efisiensi pembiayaan, penyusunan kebijakan, serta perencanaan manajemen layanan kesehatan berbasis data.

II. Metode Eksperimen

Pada eksperimen pemodelan prediksi biaya tagih fasilitas kesehatan untuk pasien TB, sejumlah tahapan kunci dilakukan untuk memastikan hasil prediksi yang akurat dan tepat. Proses ini dimulai dengan mengubah nama kolom sesuai dictionary yang diberikan kemudian menyatukan dua data (tb_2021_kepesertaan dan tb_fkrtl) dengan nama dataframe yaitu merged_df selanjutnya teknik preprocessing data yang meliputi data cleaning, Ekplorasi Data Analysis (EDA) serta Feature Engineering yang terdiri dari target encoding, one-hot encoding, label encoding, penambahan atribut usia peserta dan durasi kunjungan. Data yang telah melalui tahapan tersebut divalidasi sebelum digunakan untuk proses pemodelan lebih lanjut.

Metode pemilihan fitur dilakukan melalui eksplorasi statistik dengan uji ANOVA dan boxplot untuk variabel kategorikal sedangkan scatterplot dan heatmap digunakan untuk variabel numerik. Proses ini bertujuan untuk mengetahui fitur-fitur yang paling berpengaruh terhadap nilai biaya tagih, sehingga fitur yang digunakan dalam pemodelan didasarkan pada kombinasi hasil uji statistik dan analisis domain kesehatan, serta menghindari overfitting dengan mengurangi fitur yang kurang relevan.

Berbagai algoritma regresi diuji, antara lain Linear Regression sebagai baseline, Ridge dan Lasso untuk mengatasi multikolinearitas serta regulasi model, Decision Tree dan Random Forest yang mampu menangkap hubungan non-linier, serta Gradient Boosting yang menawarkan akurasi lebih tinggi dengan teknik ensemble. Pemilihan Random Forest didasari oleh hasil dari metrik evaluasi dan kemampuannya yang kuat dalam menangani data yang

kompleks dan heterogen seperti data kesehatan. Random Forest efektif dalam mengurangi overfitting melalui teknik ensemble yang menggabungkan banyak pohon keputusan, sehingga menghasilkan model dengan prediksi yang stabil dan akurat. Selain itu, algoritma ini mampu menangani data dengan variabel numerik dan kategorikal secara bersamaan dan memberikan interpretasi yang baik terhadap kontribusi fitur dalam model.

Evaluasi performa model dilakukan secara komprehensif dengan menggunakan teknik validasi silang K-Fold yang berguna untuk menilai reliabilitas model pada data baru, serta berbagai metrik evaluasi seperti Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R^2) serta waktu eksekusi juga menjadi pertimbangan penting dalam pemilihan model. Pemanfaatan teknik evaluasi ini bertujuan menghasilkan model prediksi yang tidak hanya unggul secara akurasi tetapi juga robust dan siap diimplementasikan dalam studi lanjutan maupun aplikasi praktis.

III. Deskripsi Dataset dan Penyiapan Data

Data Training dan Testing

Data yang digunakan dalam analisis ini adalah data pasien TB tahun 2021 dan data Fasilitas Kesehatan Rujukan Tingkat Lanjut, masing-masing berasal dari file tb_2021_kepesertaan.csv dan tb_fkrtl.csv yang di-load sebagai DataFrame dengan Pandas. Pada tb_fkrtl saya mengambil data sampel sebanyak 0,1% (15.832 baris), Kolom yang terdapat pada kedua file tersebut di ubah namanya menyesuaikan dictionary dan metadata serta buku yang telah diberikan dan dilakukan pemilihan kolom prediktor dengan melihat hubungan dengan kolom target yaitu Biaya Tagih - oleh fasilitas kesehatan (provider) lalu diperoleh beberapa kolom yang akan digunakan untuk memprediksi sebagai berikut:

Nama Kolom	Tipe Data
Tanggal lahir peserta	Object (diubah ke datetime)
Tanggal datang kunjungan FKRTL	Object (diubah ke datetime)
Tanggal pulang kunjungan FKRTL	Object (diubah ke datetime)
Tingkat Pelayanan FKRTL	Object
Kelas iuran premi peserta saat akses layanan FKRTL	Object
Deskripsi kode INACBGs	Object
INACBGs - Tipe kelompok kasus atau case groups (Digit ke-2)	Object
INACBGs - Spesifikasi kelompok kasus (Digit ke-3)	Int64
INACBGs - Tingkat keparahan kelompok kasus(Digit ke-4)	Object
Tarif special procedures (SP)	Float64

Terdapat nilai yang hilang pada kolom Tarif special procedures (SP) sebanyak 604 baris sehingga harus ditangani dengan diisi nilai 0 karena kolom yang tidak memiliki nilai artinya tidak memiliki Tarif special procedures (SP). untuk kolom target yaitu Biaya Tagih - oleh fasilitas kesehatan (provider) memiliki distribusi miring ke kanan (right-skewed) sehingga harus diatasi dengan tranformasi logaritmik dan nantinya ketika data sudah di latih harus diubah kembali ke bentuk skala aslinya (inverse transform).

Penyiapan Data

Data mentah dari dua sumber (kepesertaan dan FKRTL) yang sudah di ubah nama kolomnya didigabungkan terlebih dahulu berdasarkan kolom yang sama yaitu Nomor peserta. Data mentah FKRTL terdiri dari 1.583.242 baris sehingga harus diambil sampel, saya mengambil 1% (15.832 baris). Lalu dilakukan eksplorasi hubungan dengan kolom target (Biaya Tagih - oleh fasilitas kesehatan (provider)) yang berbentuk numerik sehingga kolom kategorik harus dilakukan tes ANOVA sedangkan kolom numerik dilihat nilai korelasinya dengan heatmap sehingga terpilih beberapa kolom yang terdapat perbedaan yang signifikan secara statistik dan nilai korelasi tinggi yaitu Tanggal lahir peserta, Tanggal datang kunjungan FKRTL, Tanggal pulang kunjungan FKRTL, Tingkat Pelayanan FKRTL, Kelas iuran premi peserta saat akses layanan FKRTL, Deskripsi kode INACBGs, INACBGs - Tipe kelompok kasus atau case groups (Digit ke-2, INACBGs - Spesifikasi kelompok kasus (Digit ke-3), INACBGs - Tingkat keparahan kelompok kasus(Digit ke-4), Tarif special procedures (SP).

setelah diperiksa nilai yang hilang terdapat pada kolom Tarif special procedures (SP) berjumlah 604 baris karena nilai yang hilang memiliki makna yang sama dengan 0 sehingga diisi dengan 0. Hal tersebut membuat kolom Tarif special procedures (SP) harus diatasi dengan tranformasi logaritmik. Transformasi logaritmik pun dilakukan pada kolom target untuk mengatasi kemiringan distribusi (right-skewed). Kolom-kolom yang berkaitan dengan tanggal ("Tanggal lahir peserta", "Tanggal datang kunjungan FKRTL", "Tanggal pulang kunjungan FKRTL") dikonversi ke format datetime untuk keperluan kalkulasi, lalu dibuat kolom baru "Durasi Kunjungan FKRTL" dan “Usia peserta”. Setelah itu, kolom-kolom tanggal yang asli dihapus karena hanya dibutuhkan fitur durasi atau umur, tidak tanggal mentah

IV. Hasil Eksperimen dan Analisisnya

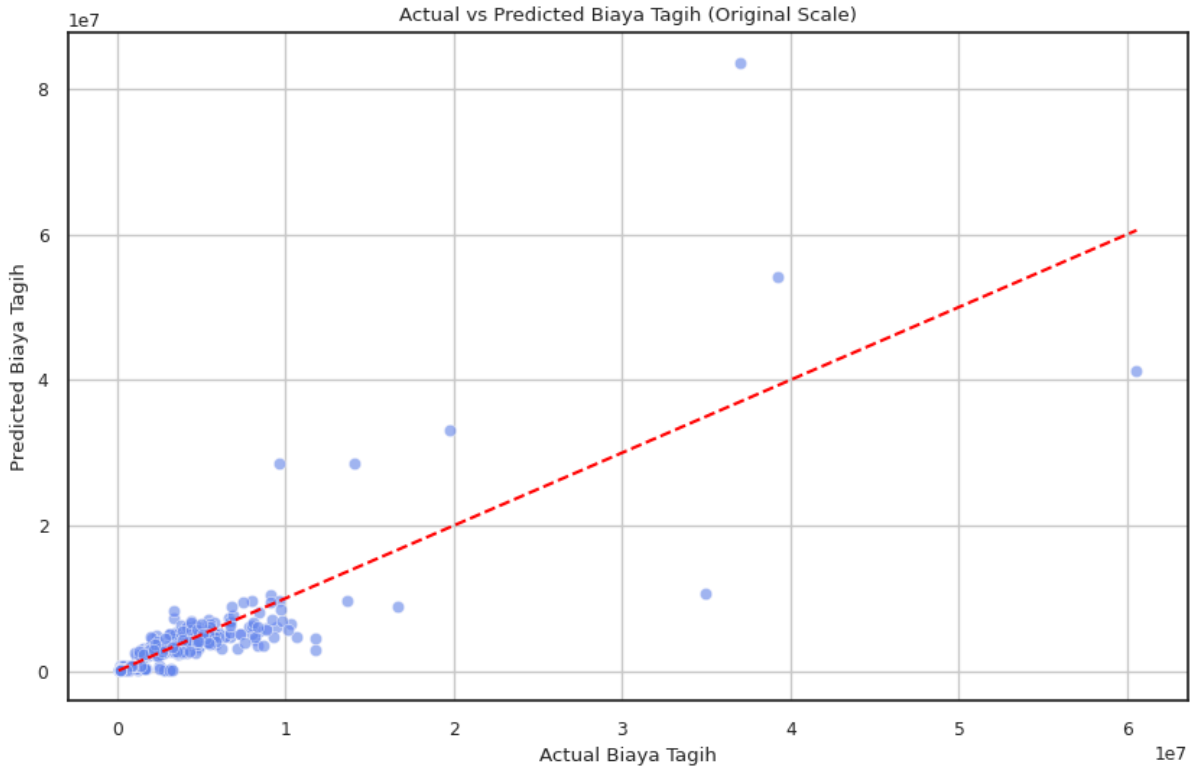
Pada eksperimen ini saya menggunakan model - model regresi yang umum digunakan. Berikut hasil eksperimen pada pemodelan prediksi "Biaya Tagih oleh fasilitas kesehatan provider", beserta perbandingan akurasi dan waktu eksekusi pada berbagai algoritma:

Nama model	***Metrik evaluasi Train/test split***	***K-fold cross validation***
Linear Regression	*Mean Absolute Error (MAE): 0.2416* *Mean Squared Error (MSE): 0.1684* *Root Mean Squared Error (RMSE): 0.4104* *R-squared (R2): 0.8170* *Execution Time: 0.0142 seconds*	*Average MAE: 0.2474* *Average MSE: 0.1739* *Average RMSE: 0.4170* *Average R-squared: 0.8284* *Average Execution Time: 0.0758 seconds*

Ridge Regression	*Mean Absolute Error (MAE): 0.2419* *Mean Squared Error (MSE): 0.1682* *Root Mean Squared Error (RMSE): 0.4102* *R-squared (R2): 0.8172* *Execution Time: 0.1196 seconds*	*Average MAE: 0.2480* *Average MSE: 0.1745* *Average RMSE: 0.4176* *Average R-squared: 0.8279* *Average Execution Time: 0.0943 seconds*
Lasso Regression	*Mean Absolute Error (MAE): 0.4822* *Mean Squared Error (MSE): 0.6443* *Root Mean Squared Error (RMSE): 0.8027* *R-squared (R2): 0.3000* *Execution Time: 0.1218 seconds*	*Average MAE: 0.5192* *Average MSE: 0.7306* *Average RMSE: 0.8543* *Average R-squared: 0.2809* *Average Execution Time: 0.0516 seconds*
Decission Tree Regression	*Mean Absolute Error (MAE): 0.1504* *Mean Squared Error (MSE): 0.0923* *Root Mean Squared Error (RMSE): 0.3038* *R-squared (R2): 0.8998* *Execution Time: 0.0378 seconds*	*Average MAE: 0.1545* *Average MSE: 0.0927* *Average RMSE: 0.3044* *Average R-squared: 0.9084* *Average Execution Time: 0.0530 seconds*
Random Forest Regression	*Mean Absolute Error (MAE): 0.1431* *Mean Squared Error (MSE): 0.0801* *Root Mean Squared Error (RMSE): 0.2829* *R-squared (R2): 0.9130* *Execution Time: 1.8730 seconds*	*Average MAE: 0.1459* *Average MSE: 0.0797* *Average RMSE: 0.2821* *Average R-squared: 0.9213* *Average Execution Time: 1.4978 seconds*
Gradien Boosting Regression	*Mean Absolute Error (MAE): 0.1669* *Mean Squared Error (MSE): 0.0891* *Root Mean Squared Error (RMSE): 0.2986* *R-squared (R2): 0.9032* *Execution Time: 1.0688 seconds*	*Average MAE: 0.1699* *Average MSE: 0.0891* *Average RMSE: 0.2984* *Average R-squared: 0.9121* *Average Execution Time: 0.6799 seconds*

Berdasarkan hasil evaluasi metrik train/test split dan K-fold cross validation di atas model Random Forest Regression menunjukkan performa terbaik dengan nilai Mean Absolute Error (MAE) rata-rata sebesar 0.1459 dan Root Mean Squared Error (RMSE) rata-rata sebesar 0.2821, serta nilai R-squared tertinggi sebesar 0.9213 pada uji k-fold cross-validation. Meskipun waktu eksekusinya lebih lama dibandingkan model linear, keunggulan akurasi dan kestabilan model ini sangat signifikan. Model lain seperti Gradient Boosting dan Decision Tree Regressor juga memberikan hasil yang cukup baik, sedangkan model linear seperti Linear Regression dan Ridge Regression memiliki performa yang lebih rendah, dan Lasso Regression menunjukkan performa paling buruk. Penggunaan Random

Forest yang mampu menangkap interaksi kompleks antar fitur dan model non-linear menjadikannya sangat cocok untuk data ini. Dengan bukti eksperimen yang lengkap, dapat disimpulkan bahwa Random Forest adalah algoritma terbaik untuk memprediksi biaya tagih di fasilitas kesehatan pada dataset ini, menggabungkan akurasi tinggi dengan kemampuan generalisasi yang baik. Hal ini didukung dengan hasil visualisasi di bawah yang menunjukkan sebaran antara nilai aktual dan nilai prediksi biaya tagih, di mana titik-titik banyak terkonsentrasi di sekitar garis identitas yang menunjukkan prediksi yang mendekati nilai aktual.



V. Prediksi Atribut Target dari Data Baru

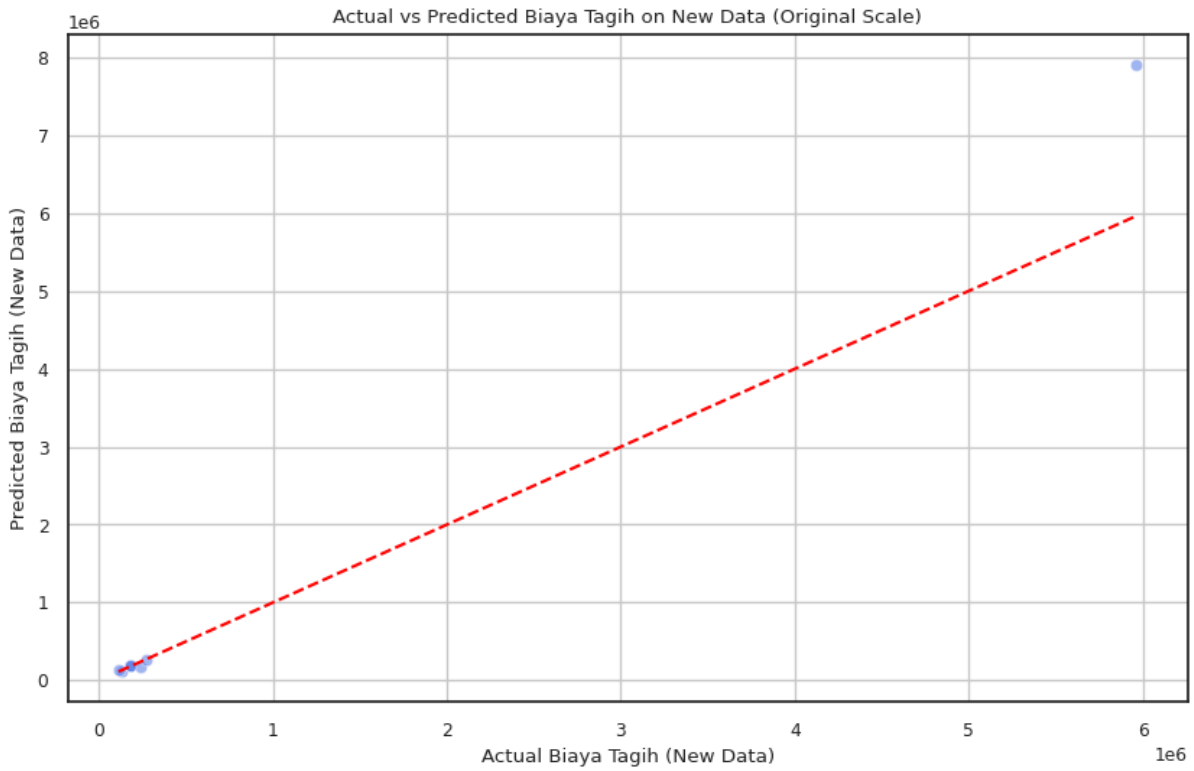
Prediksi atribut target dari data baru dilakukan dengan memanfaatkan model yang telah dilatih sebelumnya menggunakan data pelatihan. Prosesnya meliputi beberapa tahap utama. Pertama, data baru yang akan diprediksi dipersiapkan dengan memilih atribut fitur yang relevan berdasarkan hasil analisis dan pemilihan fitur dari data pelatihan. Selanjutnya, data baru tersebut diproses melalui tahap pra-pemrosesan yang sama seperti data latihan, termasuk penyesuaian tipe data, penanganan nilai hilang, transformasi fitur seperti pengkodean kategori menggunakan teknik target encoding, serta pembuatan fitur hasil rekayasa.

Setelah data baru siap dan sudah dalam format yang sesuai, model yang sudah terlatih kemudian digunakan untuk melakukan prediksi terhadap atribut target, dalam hal ini "Biaya Tagih - oleh fasilitas kesehatan (provider)". Prediksi ini menghasilkan estimasi nilai biaya yang akan dibebankan berdasarkan fitur-fitur input pada data baru tersebut. Untuk memastikan hasil prediksi berada pada skala yang sesuai, transformasi logaritmik yang sebelumnya dilakukan pada target di data latih juga diterapkan pada hasil prediksi dan kemudian dikembalikan ke skala semula menggunakan transformasi invers. Proses ini

memungkinkan prediksi atribut target yang akurat berdasarkan pola yang telah dipelajari oleh model dari data pelatihan, sehingga dapat memberikan estimasi biaya tagih pada data pasien baru secara andal dan sistematis.

Proses prediksi atribut target dari data baru tidak hanya menghasilkan estimasi nilai biaya tagih, namun juga memungkinkan evaluasi tingkat akurasi model dengan membandingkan nilai aktual dan nilai prediksi pada data yang belum pernah dilihat oleh model. Tabel berikut memperlihatkan hasil prediksi pada sepuluh contoh data baru, di mana kolom pertama menunjukkan nilai aktual biaya tagih (dalam Rupiah) dan kolom kedua memperlihatkan nilai yang diprediksi oleh model:

Nilai asli (Rupiah)	**Nilai prediksi (Rupiah)**
Rp 183.300,00	Rp 196.570,03
Rp 130.600,00	Rp 125.378,09
Rp 183.300,00	Rp 193.697,32
Rp 183.500,00	Rp 194.631,26
Rp 183.300,00	Rp 199.468,49
Rp 5.965.000,00	Rp 7.900.323,90
Rp 241.900,00	Rp 182.446,38
Rp 113.100,00	Rp 143.224,42
Rp 183.000,00	Rp 198.025,03
Rp 275.100,00	Rp 273.079,80



Perbandingan ini memperlihatkan seberapa dekat hasil prediksi model dengan nilai sebenarnya, yang juga divisualisasikan melalui grafik hubungan antara nilai aktual dan nilai prediksi (scatter plot dan garis referensi merah). Mayoritas hasil prediksi berada cukup dekat dengan nilai aktual, walaupun terdapat beberapa deviasi yang cukup besar pada kasus tertentu (misalnya pada biaya yang sangat tinggi).

Dengan visualisasi dan data numerik tersebut, proses prediksi atribut target dari data baru dapat dievaluasi secara kuantitatif dan visual, sehingga memperkuat keyakinan bahwa model mampu memberikan estimasi yang relatif akurat sesuai kebutuhan aplikasi di dunia nyata.

VI. Kesimpulan

Model Random Forest Regression telah berhasil dikembangkan dan diimplementasikan untuk memprediksi biaya tagih oleh fasilitas kesehatan terhadap pasien penyakit Tuberkulosis dengan menggunakan data BPJS Kesehatan tahun 2021. Berdasarkan evaluasi dengan metrik MAE, MSE, RMSE, dan R^2 melalui validasi silang K-Fold, model Random Forest menunjukkan performa terbaik dibandingkan algoritma regresi lain seperti Linear Regression, Ridge, Lasso, Decision Tree, dan Gradient Boosting. Model ini mampu menangkap pola kompleks dan interaksi antar fitur secara efektif sehingga menghasilkan prediksi yang akurat dan stabil meskipun waktu eksekusinya lebih lama. Hasil prediksi dari data baru juga menunjukkan tingkat kesesuaian yang baik dengan nilai aktual biaya tagih. Dengan demikian, Random Forest Regression merupakan pilihan terbaik untuk aplikasi prediksi biaya kesehatan pada dataset ini, yang dapat mendukung perencanaan, pengelolaan, dan pengambilan keputusan berbasis data di sektor kesehatan.

Lampiran

Link google collab:

prediksi Biaya Tagih - oleh fasilitas kesehatan (provider)penyakit TB.ipynb

Seluruh file yang di yang digunakan:

Pembuatan Model Machine Learning dengan Algoritma Random Forest Regression untuk Mem…