Jurnal Sintak
Volume. 2, No. 2, Maret 2024
p-ISSN: 2963-3605 e-ISSN: 2963-3125

Life Expectancy Prediction Using Decision Tree, Random Forest, Gradient Boosting, and XGBoost Regressions

Ghevira Chairunisa[1], Mohamad K. Najib[2*], Sri Nurdiati[3], Salsabila F. Imni[4], Wardah Sanjaya[5], Rizka D. Andriani[6], Henriyansah[7], Renda S. P. Putri[8], Dhea Ekaputri[9]
[1,2,3,4,5,6,7,8,9]Departemen Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Kabupaten Bogor, Indonesia
mkhoirun_najib@apps.ipb.ac.id
*Corresponding author

A B S T R A K

Angka harapan hidup menggambarkan rata-rata lamanya waktu seseorang hidup sejak lahir di dunia. Angka harapan hidup menjadi salah satu aspek dalam menentukan indeks pembangunan manusia. Semakin tinggi Angka harapan hidup maka akan semakin tinggi nilai IPM. Tujuan penelitian ini adalah memprediksi angka harapan hidup melalui model yang paling akurat dengan menggunakan model *decision tree regression, random forest regression, gradient boosting regression,* dan *XGBoost regression*, serta analisis variabel penjelas yang paling mempengaruhi angka harapan hidup. Data yang digunakan dalam penelitian ini adalah dataset *Global Country Information Dataset* 2023. Data diperoleh dari situs Kaggle. Berdasarkan analisis diperoleh bahwa model *random forest regression* menunjukkan kinerja yang lebih unggul dalam memprediksi hasil, yang ditunjukkan dengan nilai RMSE yang lebih rendah dan nilai R^2 yang lebih tinggi. Kematian bayi dan rasio kematian ibu secara konsisten diidentifikasi sebagai prediktor yang signifikan di semua model, sedangkan populasi merupakan prediktor yang kurang memprengaruhi angka harapan hidup.

Kata Kunci: Harapan Hidup; D*ecision Tree; Gradient Boosting*

A B S T R A C T

Life expectancy describes the average length of time a person is expected to live from birth in the world. Life expectancy is one of the aspects in determining the Human Development Index. The higher the life expectancy, the higher the HDI value. The purpose of this research is to predict life expectancy through the most accurate model using decision tree regression, random forest regression, gradient boosting regression, and XGBoost regression models, as well as analyzing explanatory variables that most influence life expectancy. The data used in this research is the Global Country Information Dataset 2023. The data was obtained from the Kaggle website. Based on the analysis, it was found that the random forest regression model showed superior performance in predicting outcomes, as indicated by lower RMSE values and higher R^2 values. Infant mortality and maternal mortality ratios are consistently identified as significant predictors in all models, while population is a predictor that less influences life expectancy.

Keywords: *Decision Tree; Gradient Boosting; Life Expectancy*

1. PENDAHULUAN

Indikator untuk mendorong pertumbuhan pada faktor ekonomi dapat dilihat dari perkembangan pembangunan sumber daya manusia dengan memperhatikan indeks pembangunan manusia (IPM). *United Nations Development Program* (UNDP) bahwa indeks pembangunan manusia merupakan suatu acuan alat ukur dalam meraih kualitas hidup yang dimiliki suatu wilayah dengan mempunyai tiga komponen dasar penyusun yaitu dimensi pengetahuan, dimensi kesehatan, serta dimensi hidup yang layak (BPS 2014). Tingkatan derajat kesejahteraan masyarakat dijadikan sebagai indikator keberhasilan pada suatu program kesehatan dan pembangunan sosial ekonomi sehingga berkaitan dan memberikan kontribusi terhadap peningkatan Angka harapan hidup. Menurut Badan Pusat Statistik (2014), Indeks Pembangunan Manusia (IPM) dibagi menjadi empat kategori, yaitu rendah jika IPM<60, sedang 60<=IPM<70, tinggi 70<=IPM<80, dan sangat tinggi IPM>=80.

Salah satu komponen perhitungan Indeks Pembangunan Manusia adalah angka harapan hidup. Angka harapan hidup (AHH) merupakan rata-rata perkiraan banyak tahun yang dapat ditempuh oleh seseorang sejak lahir. Faktor-faktor kompleks, sepert kesehatan, ekonomi, dan pendidikan berperan dalam menentukan angka harapan hidup suatu negara. Angka harapan hidup merupakan indikator untuk mengevalusai kinerja pemerintah dalam meningkatkan kesejahteraan masyarakat. Dengan demikian, penggunaan metode *machine learning* menjadi langkah inovatif dalam upaya pemahaman dan perbaikan terhadap angka harapan hidup sebagai indikator kesejahteraan masyarakat.

WHO mencatat angka harapan hidup di Indonesia pada tahun 2016 adalah 69 tahun. BPS mencatat angka harapan hidup pada tahun 2020 meningkat sebesar 71,47 tahun. Pada tahun 2023, Indonesia memiliki angka harapan hidup selama 71,5 tahun. Hal ini menunjukkan bahwa angka harapan hidup Indonesia sedikit berada di bawah rata-rata global, yaitu 74.4 tahun. Oleh karena itu, dapat diartikan Indonesia tidak termasuk dalam kelompok negara dengan angka harapan hidup terendah dan masih ada ruang yang cukup besar untuk peningkatan dibandingkan dengan negara lainnya.

Machine learning merupakan studi ilmiah yang memfokuskan pada penelitian tentang algoritma dan model statistik yang digunakan oleh sistem computer yang pertama kali disampaikan pada tahun 1920-an oleh beberapa ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes dan Andrey Markov. Secara umum *machine learning* digunakan untuk melatih algoritma komputasi dalam memisahkan, mengelompokkan, dan mengubah kumpulan data dengan tujuan mencapai hasil yang maksimal pada kemampuan pengelempokkan atau penentuan pola dalam kumpulan data target yang digambarkan sebagai suatu bidang penelitian statistik [1]. Tujuan dari *machine learning* adalah untuk menjalankan tugas-tugas khusus pada sistem komputer tanpa memerlukan pemrograman eksplisit. Selain itu, *machine learning* memungkinkan sistem untuk memahami dan mengaplikasikan algoritma ini dalam berbagai aspek aplikasi sehari-hari, dengan akhir tujuan untuk meningkatkan efisiensi dalam pengelolaan data oleh mesin [2].

Salah satu jenis algoritma machine learning adalah *supervised learning*. *Supervised learning* merupakan pendekatan dalam *machine learning* dalam mempelajari model menggunakan data yang sudah diberi label (data *training*) untuk mengidentifikasi pola dan menghubungkan fitur input dengan label *output*, sehingga model dapat digunakan untuk memprediksi label untuk data baru [3]. Salah satu kategori dari *Supervised learning* adalah regresi. Regresi adalah teknik *machine learning* yang digunakan untuk memprediksi nilai variabel respon (*Y*) dengan memahami hubungannya dengan satu atau lebih variabel prediktor (*X*) serta melibatkan pembentukan model matematis untuk melakukan prediksi. Beberapa contoh penggunaan algoritma *supervised learning* kategori regresi di antaranya *decision tree*, *random forest, gradient boosting*, dan *XGBoost regressions*.

Model yang digunakan untuk penelitian ini yaitu *decision tree regression*, *random forest regression, gradient boosting regression*, dan *XGBoost regression* untuk memodelkan angka harapan hidup di suatu negara menggunakan beberapa prediktor, menentukan metode manakah yang paling akurat untuk memprediksi angka harapan hidup, serta menentukan variabel yang paling mempengaruhi angka harapan hidup pada model terpilih. Model *decision tree regression* merupakan suatu metode pengolahan data dengan membangun klasifikasi dan regresi model dan digambarkan ke dalam bentuk struktur pohon yang digunakan untuk memprediksi masa depan. Algortima diberikan meliputi cabang untuk mewakili tahapan pengambilan keputusan untuk memperoleh hasil yang efisien. Pemodelan visual menggunakan model ini membantu dalam memahami dari proses pembuatan keputusan yang bertahap, terstruktur, dan rasional [4]. Sedangkan model *random forest* merupakan algoritma dalam penentuan klasifikasi dengan penggabungan dari metode *classification* and *regression tree* yang dikelompokkan berdasarkan

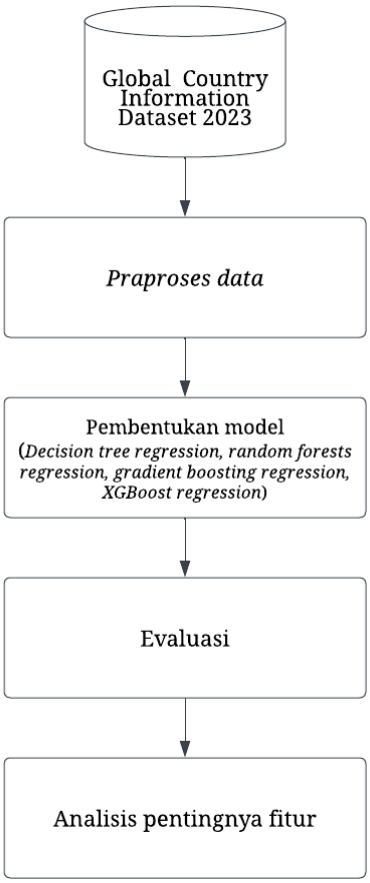
persamaan atau perbedaan [5]. Metode ini memiliki akurasi yang lebih tinggi, tidak adanya pemangkasan variabel, dan mengatasi data berskala besar secara efektif [6].

Model *gradient boosted tree regression* merupakan suatu model yang melakukan penggabungan dari berbagai *decision tree*. Model ini memiliki karakteristik fleksibel karena dari *regression tree* yang terdiri atas interpretabilitas yang tinggi, konsep yang sederhana, dan efisiensi komputasi dengan *boosting approach*, selain itu juga meningkatkan akurasi pada model faktor [7]. Akan tetapi, karena dilakukan secara pararel pada sampel *bootstrap* pada suatu kumpulan data asli sehingga *tree* dilakukan secara berurut. Adanya modifikasi dari *gradient descent* dengan *boosting algorithm* sehingga akurasi menjadi lebih meningkat [8]. Selanjutnya, metode *XGBoost* merupakan metode dengan algoritma yang melakukan penggabungan proses *boosting* dan *gradient boosted tree* untuk menghasilkan 10 kali lebih cepat, selain itu dimanfaatkan dalam menangani permasalahan ukuran besar dalam permasalahan *machine learning* sehingga dapat digunakan peramalan pada *time series* [9].

Penelitian ini memprediksi angka harapan hidup melaui model yang paling akurat dengan menggunakan metode *decision tree regression, random forest regression, gradient boosting regression*, dan *XGBoost regression*. Setelah diketahui metode yang paling akurat maka dapat diketahui terkait variabel penjelas yang paling mempengaruhi angka harapan hidup.

2. METODE

Penelitian ini dilakukan menggunakan bantuan bahasa pemrograman Python untuk membuat model. Tahapan penelitian dapat dilihat pada Gambar 1. Penelitian ini dimulai dengan pengambilan dataset. Secara garis besar, penelitian ini melewati empat tahapan, yakni praproses data, pembentukan model, proses evaluasi, dan analisis pentingnya fitur. Selain itu, korelasi antar variabel juga dievaluasi untuk mengidentifikasi sejauh mana keterkaitannya. Pada praproses data, dilakukan penghapusan data yang tidak memiliki nilai, pembagian data menjadi data latih dan data uji, dan diakhiri dengan melakukan normalisasi data. Setelah praproses data dilakukan, empat model dibuat kemudian model tersebut dilakukan *tuning* untuk mendapatkan hasil akurasi yang maksimal. Teknik evaluasi model yang digunakan adalah *Root Mean Square Error* (RMSE) serta koefisien determinasi (R^2).



Gambar 1. *Flowchart* penelitian

2.1. Decision Tree Regression

Decision tree regression merupakan sebuah struktur berbasis pohon yang digunakan untuk memprediksi hasil numerik dari varibel-variabel terkait. Cara kerja dari *decision tree regression* dimulai dari algoritma *decision tree* konvensional yang digunakan untuk membuat *tree*, *decision tree* ini menggunakan kriteria pemisahan yang meminimalkan variasi dari intra-subset dalam nilai kelas yang turun dari setiap cabang [10]. Pemilihan *root node* didasarkan kepada atribut yang dapat memaksimalkan pengurangan kesalahan yang diharapkan. Perhitungan standar deviasi sebagai berikut.

$$SDR = sd(T) - \sum_{i} \frac{|T_i|}{|T|} \times sd(T_i) \quad (1)$$

Setelah itu, pohon dipangkas dari setiap pohonnya dan dilanjutkan kepada proses pemulusan untuk mengkompensi diskontinuitas tajam .

2.2. Random Forest Regression

Random forest regression merupakan algoritma *supervised learning* yang membentuk suatu hutan (*forest*) menggunakan konsep pengulangan *decision tree*. Prediksi dari *multiple decision tree* yang dihasilkan akan dikombinasikan dan menjadi masukan (input) pada tahap pengujian [11]. *Random forest* adalah prosedur pemilihan kelas paling populer setelah sejumlah besar pohon dihasilkan [12]. Prosedur ini menghasilkan vektor acak Θ_k untuk pohon ke-k

dengan Θ_k tidak bergantung pada vektor acak $\Theta_1,\ldots,\Theta_{k-1}$. Pohon yang terbuat dari *training set* dan Θ_k akan menghasilkan h(x, Θ_k) dengan x merupakan vektor input. Misalnya, N merupakan banyaknya contoh *training set.* Pemilihan pemisahan acak Θ_k terdiri dari sejumlah bilangan bulat acak independen antara 1 dan K. Sifat dan dimensi dari Θ_k ditentukan oleh penggunaannya dalam mengkonstruksi pohon. Di bawah ini merupakan rumus untuk mencari kelas paling populer.

$$f(x) = \text{Rata-rata}(f_1(x), f_2(x), \ldots, f_n(x)) \quad (2)$$

Keterangan:
f(x) : hasil prediksi
$f_1(x)$, $f_2(x)$, …, $f_n(x)$: hasil prediksi dari setiap pohon keputusan
x : input

2.3. Gradient Boosting Regression

Algoritma g*radient boosting regression tree* merupakan pendekatan yang melibatkan penggabungan beberapa model untuk membuat prediksi yang lebih kuat di mana model peramalan yang kuat dibentuk dengan mengintegrasikan beberapa pohon regresi individu (pohon keputusan) yang disebut sebagai pembelajar yang lemah. Model yang dipelajari dengan lemah adalah model yang memiliki bias tinggi terkait dataset pelatihan, dengan varians dan keteraturan yang rendah, dan yang keluarannya dianggap hanya sedikit lebih baik jika dibandingkan dengan tebakan acak. Secara umum, algoritma *boosting* berisi tiga komponen, yaitu model aditif, *weak learner*, dan *loss function*. Algoritma ini dapat merepresentasikan hubungan non-linear seperti kurva tenaga angin dan menggunakan berbagai fungsi kerugian yang dapat dibedakan dan secara inheren dapat belajar selama iterasi antara fitur input [13]. *Gradient boosting regression* dapat didefinisikan sebagai penjumlahan dari n pohon-pohon regresi.

$$F_n(x_t) = \sum_{i=1}^{n} f_i(x_t) \quad (3)$$

Setiap $f_i(x_t)$ adalah sebuah pohon keputusan (pohon regresi). *Ensemble* dari pohon-pohon tersebut dibangun secara berurutan dengan mengestimasi pohon keputusan baru $f_{n+1}(x_t)$ dengan bantuan persamaan berikut.

$$\text{argmin} \sum_{t} L(y_t . F_n(x_t) + f_{n+1}(x_t)) \quad (4)$$

2.4. XGBoost Regression

XGBoost (eXtreme Gradient Boosting) regression adalah pengembangan lebih lanjut dari *gradient boosting*. XGBoost menggunakan model yang lebih teratur untuk membangun struktur pohon regresi, yang dapat memberikan performa lebih baik dan mengurangi kompleksitas model untuk menghindari *overfitting*. Hasil prediksi akhir dari XGBoost merupakan penjumlahan dari hasil prediksi setiap pohon regresi. Algoritma berbasis pohon keputusan bekerja dengan baik pada data dengan fitur kategorikal dan tidak mempengaruhi data dengan kelas tidak seimbang secara signifikan.

Pada metode ini diperlukan fungsi objektif yang berguna untuk menilai seberapa bagus model yang didapatkan sesuai dengan data latih. Karakteristik yang terpenting dari fungsi

objektif terdiri dari 2 bagian, yaitu nilai pelatihan yang hilang dan nilai regularisasi seperti pada persamaan (5) berikut ini.

$$Obj(\theta) = L(\theta) + \Omega(\theta) \quad (5)$$

dengan L adalah fungsi pelatihan yang hilang, dan Ω adalah fungsi regularisasi, dan θ adalah parameter model terkait. Fungsi pelatihan yang hilang secara umum dapat ditulis seperti pada persamaan (6) sebagai berikut.

$$L(\theta) = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i) \quad (6)$$

dengan y_i adalah nilai data sebenarnya yang dianggap benar dan $\hat{y}_i$ adalah hasil nilai prediksi dari model, sedangkan n adalah jumlah iterasi nilai dari model [14].

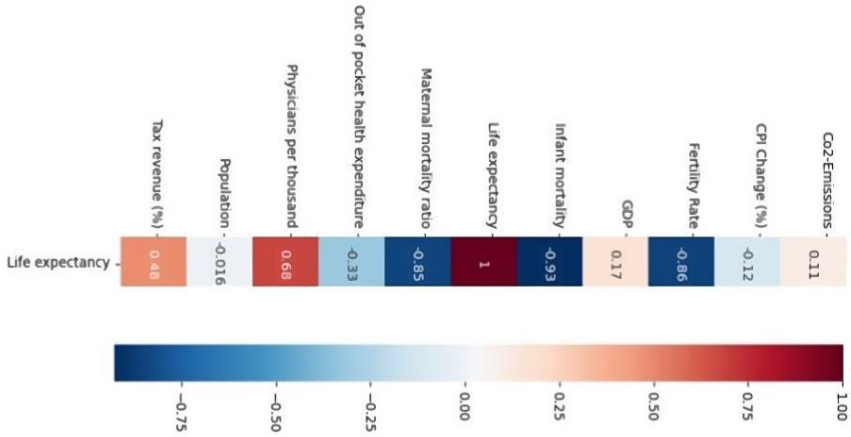
3. HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini adalah dataset *Global Country Information Dataset* 2023. Data diperolah dari situs Kaggle. Data terdiri dari 195 *record*. Penjelasan variabel dapat dilihat pada Tabel 1. Data ini dapat diunduh pada https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023. Berikut adalah variabel yang digunakan dalam penelitian ini.

Tabel 1. Deskripsi variabel

No	Variabel	Deskripsi
y	Angka harapan hidup	Rata-rata jumlah tahun harapan hidup bayi baru lahir
X_1	Emisi CO2	Emisi karbon dioksida dalam ton
X_2	Perubahan indeks harga konsumen	Persentase Perubahan indeks harga konsumen dibandingkan tahun sebelumnya
X_3	Tingkat fertilitas	Jumlah rata-rata anak yang dilahirkan oleh seorang wanita selama hidupnya
X_4	PDB	Total nilai barang dan jasa yang diproduksi di negara tersebut
X_5	Angka kematian bayi	Jumlah kematian per 1.000 kelahiran hidup sebelum mencapai usia satu tahun
X_6	Rasio kematian ibu	Jumlah kematian ibu per 100.000 kelahiran hidup
X_7	Pengeluaran kesehatan langsung	Persentase total pengeluaran kesehatan yang dibayarkan sendiri oleh individu
X_8	Jumlah dokter per seribu penduduk	Jumlah dokter per seribu orang
X_9	Populasi	Total populasi negara
X_{10}	Penerimaan pajak	Pendapatan pajak sebagai persentase terhadap PDB

Sebelum proses pemodelan dimulai, dilakukan tahap eksplorasi data untuk memahami keterkaitan antara variabel terikat dan variabel bebas. Untuk mengukur kekuatan dan arah hubungan antara variabel-variabel ini, digunakan penghitungan korelasi. Nilai korelasi antara variabel bebas dan variabel terikat dapat dilihat pada gambar 2. Warna dalam *heatmap* berkisar dari biru hingga merah, dengan biru menunjukkan korelasi negatif dan merah menunjukkan korelasi positif. Skala di sisi kanan menunjukkan intensitas korelasi dari -1 hingga 1.



Gambar 2. Visualisasi Korelasi

Angka kematian bayi dan rasio kematian ibu menunjukkan korelasi negatif yang sangat kuat dengan angka harapan hidup, dengan koefisien korelasi masing-masing sebesar -0.93 dan -0.85. Hal ini menandakan bahwa peningkatan dalam angka kematian bayi dan ibu berkorelasi dengan penurunan dalam angka harapan hidup. Sementara itu, tingkat fertilitas terlihat memiliki korelasi negatif pada -0.86, menunjukkan bahwa tingkat kesuburan yang lebih tinggi mungkin berhubungan dengan penurunan angka harapan hidup. Jumlah dokter per seribu penduduk menampilkan korelasi positif yang kuat pada 0.68, mengindikasikan bahwa peningkatan jumlah dokter per seribu penduduk berpotensi meningkatkan angka harapan hidup. Penerimaan pajak juga menunjukkan korelasi positif pada 0.48, yang dapat diinterpretasikan bahwa penerimaan pajak yang lebih tinggi mungkin berkontribusi pada peningkatan angka harapan hidup, mungkin melalui alokasi dana yang lebih baik untuk layanan kesehatan dan infrastruktur sosial. Sedangkan PDB dan emisi CO2 memperlihatkan hubungan yang lebih lemah, dengan koefisien korelasi 0.17 dan 0.11 berturut-turut, menunjukkan bahwa faktor ekonomi makro dan dampak lingkungan mungkin tidak memiliki pengaruh langsung yang kuat terhadap angka harapan hidup seperti faktor kesehatan langsung. Variabel seperti populasi dan perubahan indeks harga konsumen tampaknya memiliki sedikit atau tidak ada hubungan dengan angka harapan hidup.

Dalam upaya untuk menentukan konfigurasi optimal model regresi, peneliti menggunakan metode *Grid Search Cross-Validation* untuk sistematis mengevaluasi dan membandingkan kombinasi *hyperparameter* yang berbeda. Proses ini melibatkan pencarian *best parameter* melalui *parameter* dengan *grid search value* yang telah ditentukan, di mana setiap kombinasi diuji untuk memastikan bahwa model yang dihasilkan tidak hanya cocok dengan baik pada data tetapi juga memiliki kemampuan generalisasi yang kuat pada data. Hasil dari pencarian ini tercantum dalam tabel 2, 3, dan 4, menyoroti nilai-nilai *hyperparameter* yang menyediakan kinerja terbaik untuk masing-masing model regresi.

Tabel 2. Hasil *tuning hyperparameter* model *decision tree regression.*

No	Parameter	*Grid Search Value*	*Best Parameter*
1	*Criterion*	mse, friedman_mse, mae	friedman_mse
2	*Max_depth*	10, 20, 30, 40, 50	10
3	*Min_samples_split*	10, 12, 14	12
4	*Min_samples_leaf*	1, 2, 4	4

Tuning hyperparameter untuk model *decision tree regression* mencakup empat *hyperparameter* utama. Setelah proses *tuning*, *criterion* yang memberikan hasil terbaik adalah 'friedman_mse', yang mengindikasikan bahwa variansi dari kesalahan kuadrat yang disesuaikan

oleh friedman memberikan pembagian pohon terbaik dibandingkan dengan mse atau mae. *Max_depth* terbaik adalah 10, menunjukkan bahwa pembatasan pohon pada kedalaman 10 membantu mencegah *overfitting* dan memberikan model yang lebih general. *Min_samples_split* optimal adalah 12, yang berarti setiap node harus memiliki setidaknya 12 sampel sebelum membagi menjadi node lebih lanjut, ini juga membantu model untuk tidak terlalu spesifik terhadap data latih. Terakhir, min_samples_leaf terbaik adalah 4, ini menjamin bahwa setiap daun pohon akhir memiliki setidaknya empat sampel, yang menambah stabilitas model.

Tabel 3. Hasil *tuning hyperparameter* model *random forest regression*

No	Parameter	*Grid Search Value*	*Best Parameter*
1	*n_estimators*	100, 200, 300	200
2	*Max_depth*	10, 20, 30	10
3	*Min_samples_split*	2, 5, 10	5
4	*Min_samples_leaf*	1, 2, 4	2

Pada *random forest regression*, hasil *tuning* menunjukkan bahwa 200 *n_estimators* memberikan hasil terbaik. Ini berarti bahwa ensemble dari 200 pohon keputusan bekerja paling baik dalam meminimalkan kesalahan prediksi. *Max_depth* optimal adalah 10, yang serupa dengan *decision tree*, memberikan kedalaman yang cukup untuk mengambil fitur penting tetapi tidak terlalu dalam sehingga mengakibatkan *overfitting*. *Min_samples_split* adalah 5, yang menunjukkan bahwa pembagian node lebih lanjut memerlukan setidaknya 5 sampel di node tersebut. Ini membantu menghindari pembagian yang berlebihan dan memastikan bahwa split memiliki cukup informasi. *Min_samples_leaf* yang optimal adalah 2, menegaskan bahwa pohon tidak tumbuh terlalu kompleks dengan memastikan setidaknya dua sampel per daun.

Tabel 4. Hasil *tuning hyperparameter* model *gradient boosting regression*

No	Parameter	*Grid Search Value*	*Best Parameter*
1	*n_estimators*	100, 200, 300	300
2	*Max_depth*	3, 5, 7	3
3	*Min_samples_split*	2, 5, 10	2
4	*Learning_rate*	0.01, 0.1, 0.2	0.01

Model *gradient boosting regression* menemukan *n_estimators* optimal pada 300, yang mengacu pada jumlah tahap *boosting* yang digunakan untuk membangun model. Jumlah yang lebih tinggi menunjukkan bahwa model memerlukan lebih banyak iterasi *boosting* untuk memperoleh hasil yang optimal. *Max_depth* terbaik adalah 3, yang mengindikasikan cara untuk menjaga model sederhana dan mencegah *overfitting*. *Min_samples_split* optimal adalah 2, ini menunjukkan bahwa model memungkinkan membagi node dengan jumlah sampel yang sangat sedikit. *Learning_rate* yang terbaik adalah 0.01, nilai yang sangat kecil menunjukkan bahwa proses pembelajaran dilakukan secara bertahap dan hati-hati untuk menghindari *overfitting* dan memungkinkan model untuk lebih baik menggeneralisasi dari data latih.

Tabel 5. Hasil *tuning hyperparameter* model *XGBoost regression*

No	Parameter	*Grid Search Value*	*Best Parameter*
1	*n_estimators*	50, 100, 150	150
2	*Max_depth*	3, 5, 7	5
3	*Min_chld_weight*	1, 2, 3	2
4	*Learning_rate*	0.01, 0.1, 0.2	0.2

Model *XGBoost Regression* menemukan bahwa *n_estimators* optimal adalah 150, menunjukkan jumlah pohon yang harus dibangun dalam proses *boosting*. *Max_depth* terbaik

adalah 5, yang cukup untuk menangkap hubungan kompleks dalam data tanpa menjadi terlalu spesifik. *Min_child_weight* optimal adalah 2, parameter ini mirip dengan *min_samples_leaf* tetapi khusus untuk XGBoost, digunakan untuk mengontrol jumlah observasi yang diperlukan dalam bobot *child* sebelum model membuat keputusan untuk membagi lebih lanjut. *Learning_rate* terbaik adalah 0.2, yang lebih tinggi dibandingkan dengan *gradient boosting*, menunjukkan bahwa model *XGBoost* mungkin mengkonvergensi lebih cepat dan memerlukan langkah pembelajaran yang lebih besar.

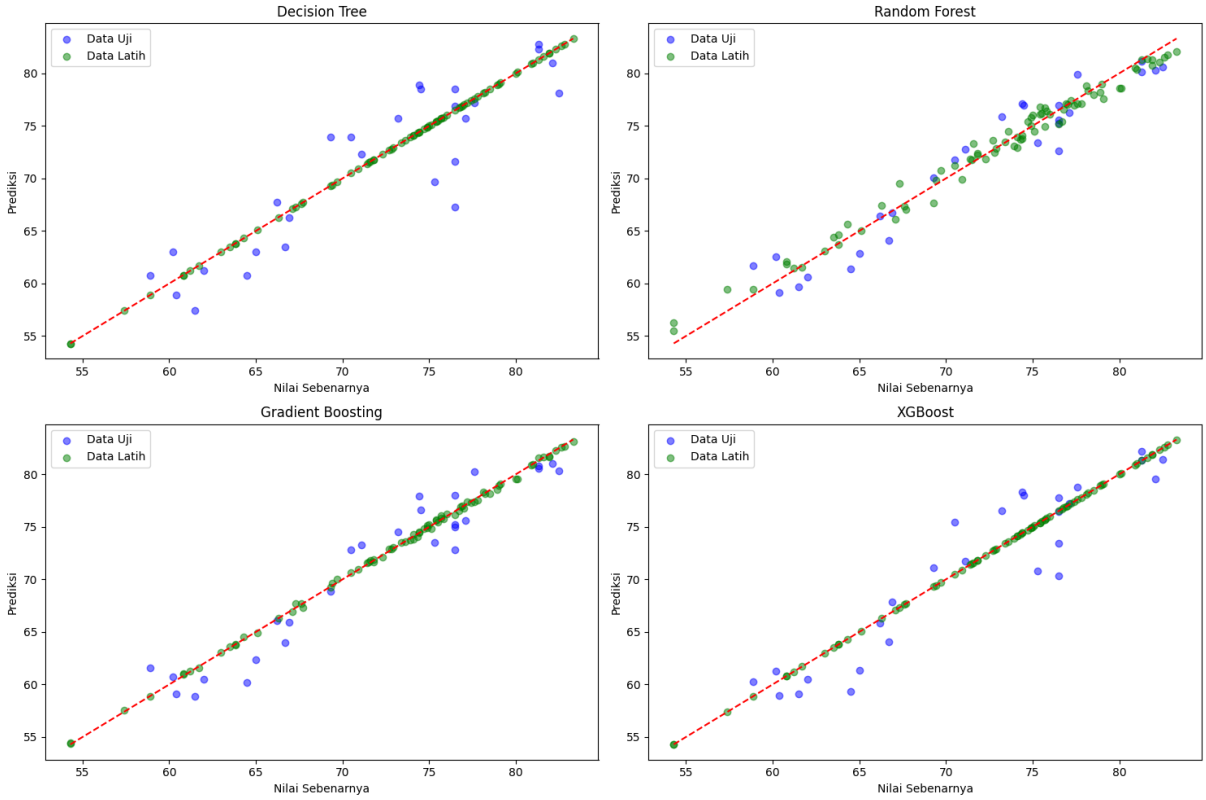
Dalam evaluasi kinerja model regresi terkait prediksi angka harapan hidup, empat model yang berbeda dianalisis baik sebelum dan setelah proses tuning hyperparameter, sebagaimana ditunjukkan dalam tabel 6. Evaluasi ini dilakukan dengan mengukur *Root Mean Square Error* (RMSE) dan koefisien determinasi (R^2) untuk setiap model. RMSE memberikan ukuran seberapa jauh prediksi model berada dari nilai sebenarnya, dengan nilai yang lebih rendah menunjukkan kesalahan prediksi yang lebih kecil, sementara R^2 mengukur proporsi varians dalam variabel terikat yang dapat dijelaskan oleh variabel bebas dalam model, dengan nilai mendekati 1 menunjukkan penjelasan yang lebih baik.

Tabel 6. Evaluasi Model

No	Model	RMSE		R^2	
		Sebelum *Tuning*	Setelah *Tuning*	Sebelum *Tuning*	Setelah *Tuning*
1	*Decision Tree Regression*	3.38	3.07	0.77	0.81
2	*Random Forest Regression*	2.04	1.94	0.91	0.92
3	*Gradient Boosting Regression*	2.10	2.02	0.91	0.92
4	*XGBoost Regression*	2.77	2.57	0.85	0.87

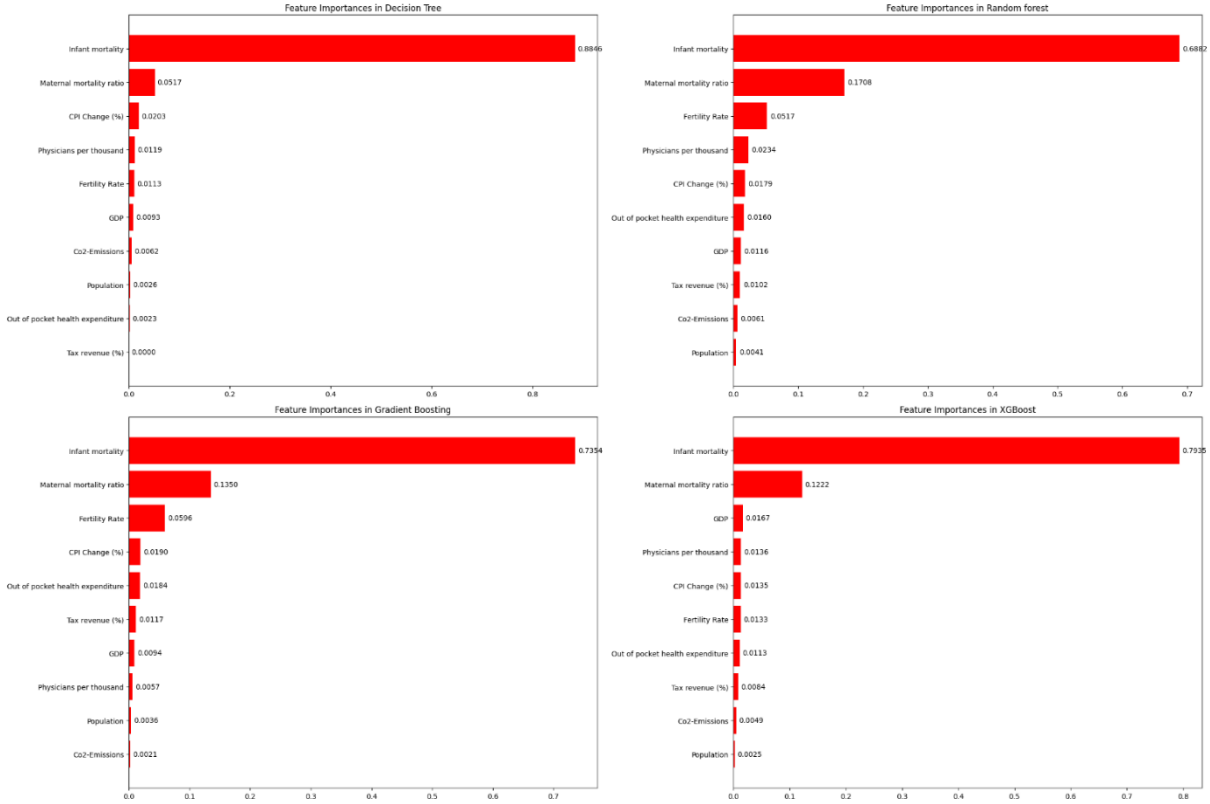
Sebelum *tuning*, model *random forest regression* menunjukkan performa yang paling unggul dengan RMSE terendah di 2.04 dan R^2 tertinggi di 0.91. Setelah *tuning*, model ini bahkan meningkatkan kinerjanya menjadi RMSE 1.94 dan R^2 0.92, menegaskan posisinya sebagai model dengan prediksi yang paling akurat dan kemampuan menjelaskan variabilitas data yang paling tinggi di antara model-model yang diuji. Sementara itu, model *decision tree regression* dan *gradient boosting regression* juga menunjukkan peningkatan kinerja setelah *tuning*. *Decision tree regression* menurunkan RMSE-nya dari 3.38 menjadi 3.07 dan meningkatkan R^2 dari 0.77 menjadi 0.81, sedangkan *gradient boosting regression* menunjukkan penurunan RMSE dari 2.10 menjadi 2.02 dengan R^2 yang tetap di 0.91. Model *XGBoost regression* mengalami peningkatan setelah *tuning*, dengan RMSE berkurang dari 2.77 menjadi 2.57 dan R^2 meningkat dari 0.85 menjadi 0.87. Peningkatan ini menandakan bahwa *tuning hyperparameter* memiliki dampak positif signifikan terhadap kemampuan prediksi dan penjelasan model.

Selain dievaluasi menggunakan *Root Mean Square Error* (RMSE) dan koefisien determinasi (R^2), kinerja model regresi juga dievaluasi menggunakan *scatter plot* nilai sebenarnya dan prediksi, sebagaimana ditunjukkan dalam gambar 3. Garis putus-putus merah merepresentasikan garis identitas di mana prediksi sempurna akan terletak, hal ini merupakan situasi di mana nilai prediksi tepat sama dengan nilai sebenarnya. Titik-titik yang dekat dengan garis ini menunjukkan prediksi yang akurat, sedangkan titik-titik yang jauh dari garis menunjukkan prediksi yang kurang akurat.



Gambar 3. *Scatter plot* nilai sebenarnya dan prediksi

Analisis tentang pentingnya fitur dalam empat model regresi yang berbeda menunjukkan keseragaman dan variasi dalam bagaimana model-model tersebut mengevaluasi dan menggunakan informasi yang diberikan. Dalam semua model,kematian bayi muncul sebagai variabel yang paling berpengaruh terhadap prediksi, menandakan bahwa ini mungkin merupakan indikator kunci dalam kumpulan data yang sedang dianalisis. Selain itu, rasio kematian ibu juga secara konsisten diidentifikasi sebagai variabel penting di semua model, meskipun dengan tingkat pengaruh yang berbeda. Dalam model *random forest regression*, variabel seperti jumlah dokter per seribu penduduk dan tingkat fertilitas juga muncul sebagai penting, sementara *gradient boosting regression* memberikan bobot yang lebih signifikan pada perubahan indeks harga konsumen dan pengeluaran kesehatan langsung. Untuk model *XGBoost regression*, distribusi *feature importance* menunjukkan bahwa selain kematian bayi, fitur seperti PDB, jumlah dokter per seribu penduduk, dan perubahan indeks harga konsumen memiliki peran yang lebih seimbang dalam kontribusinya terhadap prediksi model. Analisis ini mengungkapkan bahwa terdapat beberapa variabel dianggap penting secara universal, model tertentu mungkin mengaitkan fitur penting yang berbeda terhadap variabel tertentu, yang mencerminkan bagaimana algoritma yang berbeda memproses dan mengevaluasi informasi dalam membuat prediksi.



Gambar 4. Diagram *Feature Importances*

4. KESIMPULAN

Secara keseluruhan, model *random forest regression* menunjukkan kinerja yang lebih unggul dalam memprediksi hasil, yang ditunjukkan dengan nilai RMSE yang lebih rendah dan nilai R^2 yang lebih tinggi. Kematian bayi dan rasio kematian ibu secara konsisten diidentifikasi sebagai prediktor yang signifikan di semua model, sedangkan populasi merupakan prediktor yang kurang mempengaruhi angka harapan hidup.

5. REFERENSI

[1] K. M. Barasi and I. Nurhaida, "Klasifikasi Jenis Tensimeter Pada Instansi Kesehatan Di Indonesia," *J. Soc. Sci. Res.*, vol. 3, no. 2, pp. 9388–9396, 2023.

[2] B. Mahesh, "Machine learning algorithms-a review," *Int. J. Sci. Res.*, pp. 381–386, 2020.

[3] E. Retnoningsih and R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised dan Unsupervised Learning Menggunakan Python," *Bina Insa. Ict J.*, vol. 7, no. 2, pp. 156–165, 2020.

[4] E. Mardiani *et al.*, "Analisis Prediksi Pendapatan Penduduk dengan Metode K-Nearest Neighbor, Decision Tree, Naive Bayes, Ensemble Methods, dan Linear Regression," *J. Soc. Sci. Res.*, vol. 3, no. 4, pp. 8667–8679, 2023.

[5] O. Saputra and I. Ismail, "Klasifikasi Pada Literasi Membaca Buku Oleh Mahasiswa Menggunakan Algoritma Random Forest Di Perguruan Tinggi Lampung," *J. Ilmudata.org*, vol. 2, no. 11, pp. 1–15, 2022.

[6] A. Ramadhan, B. Susetyo, and Indahwati, "Penerapan Metode Klasifikasi Random Forest Dalam Mengidentifikasi Faktor Penting Penilaian Mutu Pendidikan," *J. Pendidik. dan Kebud.*, vol. 4, no. 2, pp. 169–182, 2019, doi: 10.24832/jpnk.v4i2.1327.

[7] Y. Shin, "Application of boosting regression trees to preliminary cost estimation in building construction projects," *Comput. Intell. Neurosci.*, vol. 2015, no. 9, pp. 1–9, 2015, doi: 10.1155/2015/149702.
[8] E. Fitri, "Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 58–64, 2023, doi: 10.52158/jacost.v4i1.491.
[9] M. D. A. Alhamdi, Herman, and W. Astuti, "Peramalan Kebutuhan Obat Menggunakan XGBoost Studi Kasus pada Rumah Sakit XYZ," *Indones. J. Comput. Sci.*, vol. 12, no. 5, 2023, doi: 10.33022/ijcs.v12i5.3344.
[10] S. S. Rathore and S. Kumar, "A Decision Tree Regression based Approach for the Number of Software Faults Prediction," *ACM SIGSOFT Softw. Eng. Notes*, vol. 41, no. 1, pp. 1–6, 2016, doi: 10.1145/2853073.2853083.
[11] S. Saadah and H. Salsabila, "Prediksi Harga Bitcoin Menggunakan Metode Random Forest," *J. Komput. Terap.*, vol. 7, no. 1, pp. 24–32, 2021, doi: 10.35143/jkt.v7i1.4618.
[12] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[13] U. Singh, M. Rizwan, M. Alaraj, and I. Alsaidan, "A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments," *Energies*, vol. 14, no. 16, 2021, doi: 10.3390/en14165196.
[14] S. E. H. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *J. Math. Theory Appl.*, vol. 4, no. 1, pp. 21–26, 2022, doi: 10.31605/jomta.v4i1.1792.