Prediction of Life Expectancy Based on Socioeconomic and Health Factors Using Random Forest Regressor Algorithm [Prediksi Angka Harapan Hidup Berdasarkan Faktor Sosioekonomi Dan Kesehatan Menggunakan Algoritma Random Forest Regressor]

Hamzah Zufarul Furqon[1)], Mochamad. Alfan Rosid [*,2)], Ade Eviyanti [*,3)], Yunianita Rahmawati [*,4)]

[1)]*Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia*
[2)] *Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia*
[3)] *Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia*
[4)] *Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia*
*Email Penulis Korespondensi: Adeeviyanti@umsida.ac.id

Abstract. *This research aims to develop a predictive model for life expectancy in various Asian countries using the Random Forest Regressor algorithm. The model is capable of predicting life expectancy with an accuracy rate of 96.5% and a Mean Absolute Error (MAE) of 0.99. Correlation analysis indicates that the "school_year," "HDI," and "BMI" features have a significant impact on life expectancy, highlighting the strong relationship between education, human development, dietary patterns, and societal well-being. The findings of this study can support policy efforts to enhance the welfare and quality of life of populations across different Asian countries, with a focus on education, human development, healthy eating habits, and an active lifestyle.*

Keywords *- Life expectancy prediction; Random Forest Regressor; Correlation analysis*

Abstrak. *Penelitian ini bertujuan untuk mengembangkan model prediktif angka harapan hidup di berbagai negara Asia menggunakan algoritma Random Forest Regressor. Model ini mampu memprediksi angka harapan hidup dengan tingkat akurasi mencapai 96,5%, dengan Mean Absolute Error (MAE) sebesar 0,99. Analisis korelasi menunjukkan bahwa fitur "school_year", "HDI", dan "BMI" memiliki pengaruh yang signifikan terhadap angka harapan hidup, menggambarkan hubungan yang kuat antara pendidikan, pembangunan manusia, pola makan, dan kesejahteraan masyarakat. Hasil penelitian ini dapat mendukung upaya kebijakan untuk meningkatkan kesejahteraan dan kualitas hidup masyarakat di berbagai negara Asia dengan fokus pada pendidikan, pembangunan manusia, pola makan sehat, dan gaya hidup yang aktif.*

Kata Kunci *- Prediksi Angka Harapan Hidup; Random Forest Regressor; Analisis Korelasi*

I. PENDAHULUAN

Di era globalisasi dan kemajuan teknologi saat ini, harapan hidup menjadi salah satu indikator utama yang mencerminkan kesejahteraan dan kualitas hidup suatu populasi [1]. Harapan hidup tidak hanya dipengaruhi oleh faktor kesehatan, tetapi juga oleh beragam faktor sosioekonomi yang saling terkait dan kompleks. Perbedaan signifikan dalam angka harapan hidup antar berbagai negara menunjukkan adanya disparitas yang besar dalam hal akses terhadap layanan kesehatan, kondisi ekonomi, pendidikan, dan lingkungan hidup [2]. Memahami dan menganalisis faktor-faktor ini penting tidak hanya untuk menggambarkan kondisi kesehatan populasi secara akurat, tetapi juga untuk merumuskan kebijakan dan strategi intervensi yang efektif [3].

Berdasarkan data yang dirilis oleh Perserikatan Bangsa-Bangsa dalam "*World Population Prospects: The 2010 Revision Population Database*", terdapat analisis mengenai tren harapan hidup di seluruh dunia dari tahun 1995 hingga 2015, diukur setiap lima tahun sekali. Data ini menunjukkan dinamika perubahan harapan hidup secara global selama periode tersebut [4]. Menurut data dari Badan Pusat Statistik Indonesia, pada periode 2010-2015, Jepang memimpin dalam daftar negara Asia dengan harapan hidup tertinggi, mencapai 83,5 tahun, diikuti oleh Hong Kong dengan 83,3 tahun [5]. Dalam konteks regional, Indonesia berada di posisi ke-14 dari 19 negara di Asia dari segi harapan hidup, dengan rata-rata mencapai 70,1 tahun. Di Asia Tenggara, Singapura memimpin dengan harapan hidup sebesar 82,2 tahun, disusul oleh Vietnam (75,9 tahun), Malaysia (74,9 tahun), Thailand (74,3 tahun), dan Kamboja (71,6 tahun).

Namun, kompleksitas dan volume data yang besar menjadi tantangan dalam analisis tradisional. Di sinilah peran teknologi canggih seperti machine learning menjadi sangat penting. *Machine learning*, cabang dari kecerdasan buatan, memungkinkan kita untuk mengolah dan menganalisis dataset besar dengan cara yang lebih efisien dan akurat. Salah satu teknik dalam *machine learning* yang menonjol adalah Random Forest, khususnya dalam konteks regresi.

Random Forest merupakan metode yang kuat dalam memprediksi dan menganalisis data kompleks, berkat kemampuannya dalam menangani banyak variabel prediktor dan menangkap hubungan non-linear antara variabel-variabel tersebut [6]. Random Forest bekerja dengan cara menggabungkan sejumlah besar pohon keputusan (*decision trees*), di mana setiap pohon dibangun dari subset acak dari data dan variabel. Proses ini disebut sebagai 'bagging' atau 'Bootstrap Aggregating', di mana setiap pohon keputusan memberikan prediksi yang kemudian digabungkan untuk menghasilkan prediksi akhir yang lebih stabil dan akurat [7].

Keunggulan Random Forest terletak pada kemampuannya untuk mengurangi overfitting, yang sering terjadi pada model pohon keputusan tunggal. Selain itu, Random Forest juga memberikan informasi mengenai pentingnya masing-masing variabel dalam model, yang sangat berguna dalam mengidentifikasi faktor-faktor kunci yang mempengaruhi harapan hidup [8].

Dengan memanfaatkan Random Forest dalam analisis data tentang harapan hidup, penelitian ini bertujuan untuk mengembangkan model prediktif yang dapat mengungkapkan wawasan lebih dalam mengenai faktor-faktor yang mempengaruhi harapan hidup dan bagaimana interaksi antara faktor-faktor tersebut berkontribusi terhadap variasi harapan hidup di berbagai negara. Hasil dari penelitian ini diharapkan memberikan kontribusi signifikan bagi pembuat kebijakan dalam merancang strategi kesehatan dan sosioekonomi yang lebih efektif dan tepat sasaran.

Beberapa penelitian terdahulu yang relevan telah dilakukan sebelumnya. diantaranya adalah Penelitian oleh Riska Oktavia dkk dengan judul “Pengaruh Metode Algoritma K-means dalam Pengelompokan Angka Harapan Hidup Saat Lahir Menurut Provinsi” pada tahun 2020. Penelitian ini melakukan *clustering* setiap jumlah angka harapan hidup saat lahir dengan jumlah terendah sampai tertinggi menggunakan metode K-means Clustering. Data dikelompokkan berdasarkan nama provinsi yang memiliki jumlah angka harapan hidup saat lahir dari tahun 2015 sampai 2018. Hasilnya yaitu angka harapan hidup saat lahir yang terkelompok dari cluster terendah sampai tertinggi [9].

Selanjutnya penelitian yang dilakukan oleh Samuel Palentino Sinaga, Anjar Wanto dan S. Solikhun dengan judul “Implementasi Jaringan Syaraf Tiruan Resilient Backpropagation dalam Memprediksi Angka Harapan Hidup Masyarakat Sumatera Utara” pada tahun 2019. Penelitian ini menerapkan algoritma Resilient Backpropagation untuk memprediksi angka harapan hidup masyarakat di Sumatera Utara. Data yang digunakan yaitu data angka harapan hidup di Sumatera Utara yang terdiri dari 33 kabupaten / kota dari tahun 2013 sampai 2017 yan diperoleh dari BPS Sumatera Utara. Penelitian ini menghasilkan tingkat akurasi sebesar 88% dengan 22 epoch [10].

Penelitian lain yang relevan juga pernah dilakukan oleh Teuku Afriliansyah, Z Zulfahmi pada tahun 2020 dalam judul “Prediksi Angka Harapan Hidup Masyarakat Aceh dengan Model Terbaik Algoritma Cyclical Order”. Penelitian ini memprediksi harapan hidup masyarakat di Aceh dengan menggabungkan metode urutan siklik dengan akurasi 91% [11].

Penelitian selanjutnya dilakukan oleh Bofandra Muhammad pada tahun 2019 dengan judul “Implementasi Data Mining untuk Prediksi Standar Hidup Layak Berdasarkan Tingkat Kesehatan dan Pendidikan Masyarakat”. Penelitian ini menggunakan data mining untuk memprediksi standar hidup layak. Hasil yang didapatkan oleh metode tersebut adalah nilai akurasi sebesar 82,32% [12].

Terakhir, penelitian yang dilakukan oleh Samuel Fernando Manurung, Ahmad Andriansyah, Jaka Permana, dan Risky Pangestu pada tahun 2022 dengan judul "Pemanfaatan Algoritma JST untuk Menentukan Model Prediksi Umur Harapan Hidup Saat Lahir". Penelitian ini bertujuan untuk mengestimasi Umur Harapan Hidup di Sumatera Utara untuk membantu pemerintah setempat dalam menentukan kebijakan kesehatan. Menggunakan metode Jaringan Saraf Conjugate Gradient Fletcher-Reeves, penelitian ini menggunakan data dari Badan Pusat Statistik dan menghasilkan model arsitektur terbaik 4-10-1 dengan nilai MSE sebesar 0.00026375[13].

Dari penelitian – penelitian terdahulu di atas, peneliti mengidentifikasi bahwa kebanyakan studi terfokus pada ruang lingkup geografis yang sempit dan tidak menyeluruh dalam mempertimbangkan variabel yang berkontribusi terhadap harapan hidup. Penelitian yang lebih spesifik geografisnya seperti pada Sumatera Utara atau Aceh memberikan kecenderungan hasil yang tidak dapat diaplikasikan pada konteks yang lebih luas. Metodologi yang digunakan sebelumnya, seperti K-means Clustering dan Resilient Backpropagation, meskipun efektif dalam penerapannya, terbatas dalam kemampuan mereka untuk memetakan hubungan yang kompleks antara faktor sosioekonomi dan kesehatan.

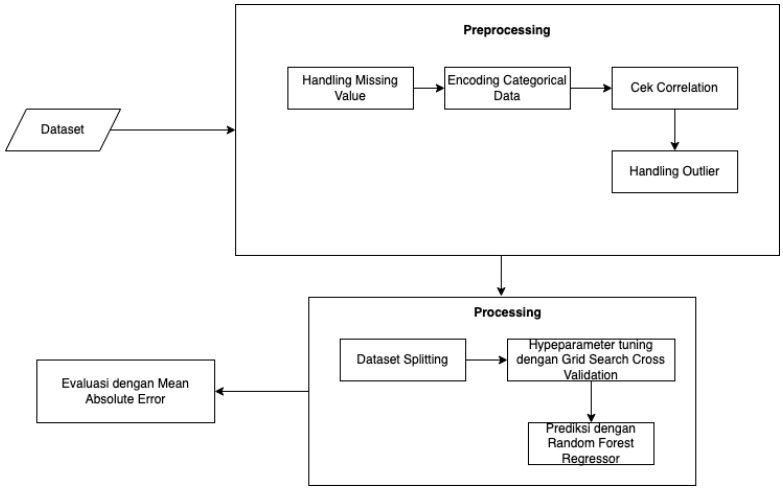
Penelitian yang dilakukan saat ini bermaksud untuk mengisi celah tersebut dengan menggunakan algoritma Random Forest Regressor yang kuat, diiringi dengan teknik GridSearchCV untuk hyperparameter tuning. Pendekatan ini akan memungkinkan peneliti untuk tidak hanya memilih parameter yang optimal untuk model tetapi juga untuk memahami bagaimana berbagai konfigurasi parameter dapat mempengaruhi kinerja model dalam memprediksi angka harapan hidup. GridSearchCV akan secara sistematis bekerja melalui berbagai kombinasi parameter dan menggunakan validasi silang untuk menentukan kombinasi yang menghasilkan kinerja terbaik.

Dari latar belakang dan tinjauan pustaka tersebut, penelitian ini tidak hanya penting dalam konteks akademis, tetapi juga memiliki implikasi praktis yang signifikan dalam bidang kesehatan publik dan pembangunan sosial. Penelitian ini direpresentasikan dalam skripsi dengan judul: " Prediksi Angka Harapan Hidup Berdasarkan Faktor Sosioekonomi dan Kesehatan Menggunakan Algoritma Random Forest Regressor".

II. METODE

A. Tahapan Penelitian

Tahapan penelitian merupakan gambaran umum terkait alur penelitian yang akan dilakukan dalam pengerjaan penelitian ini dari awal hingga akhir. Tahapan yang dilakukan dalam penelitian ini dapat dipaparkan melalui diagram alir seperti pada Gambar berikut



Gambar 1. Tahapan Penelitian

B. Data

Data yang digunakan pada penelitian ini diambil dari https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who yang terdiri dari 2938 record. Dataset ini terdiri dari 22 kolom dan 2928 baris yang diambil dari tahun 2000 hingga 2015 di 193 negara di Asia.

country	Negara
year	Tahun
status	Developed = Maju Developing = berkembang
life_expectancy	Tingkat harapan hidup (umur)
adult_mortality	Tingkat kematian orang dewasa (umur 15-60 tiap 1000 populasi)
infant_deaths	Jumlah kematian bayi tiap 1000 populasi
Alcohol	Konsumsi alkohol per kapita (umur 15+, liter)
percentage_expenditure	Persentase pengeluaran untuk kesehatan dari PDB per kapita
HepB	Persentase imunisasi Hepatitis B (HepB) untuk umur 1 tahun
BMI	Rata-rata Body Mass Index seluruh populasi
Measles	Jumlah kasus campak tiap 1000 populasi
u5_deaths	Jumlah kematian balita tiap 1000 populasi
Polio	Persentase imunisasi Polio (Pol3) untuk umur 1 tahun
total_expenditure	Persentase pengeluaran pemerintah Untuk kesehatan dari total pengeluaran pemerintah (%)
DPT	Persentase imunisasi dfiteri, pertusis, Dan tetanus (DPT3) untuk umur 1 tahun
HIV_AIDS	Kematian tiap 1000 kelahiran HIV/AIDS (umur 0-4 tahun)
GDP	GDP per kapita (USD)
population	Populasi negara
Thinness_10_19	Persentase kekurusan pada anak 10-19 tahun
Thinness_5_9	Persentase kekurusan pada anak 5-9 tahun
HDI	Indeks Pembangunan Manusia dalam hal Komposisi pendapatan sumber daya (0 sampai 1)
school_year	Jumlah tahun bersekolah

Gambar 2. Dataset

C. Preprocessing

Data mentah tidak dapat diproses oleh mesin secara langsung, oleh sebab itu perlu dilakukan *preprocessing*. *Preprocessing* data menjadi tahap kritis karena kualitas dan kebersihan data yang baik dapat berdampak signifikan pada hasil akhir prediksi. Pada bagian ini peneliti akan membahas tahapan-tahapan penting dalam mempersiapkan data sebelum proses analisis dan prediksi menggunakan metode Random Forest Regressor.

1. ***Handling Missing Value***

   Pada tahapan awal *preprocessing*, data masukan dilakukan *handling missing value* terlebih dahulu. *Missing value* dapat terjadi karena kesalahan penginputan data atau memang data tersebut tidak ada. Karena algoritma *machine learning* tidak dapat memproses data yang terdapat *missing value*, maka sebelum dilakukan modelling harus dilakukan *handling missing value* terlebih dahulu. Peneliti menggunakan Teknik imputasi *mean*. Sehingga data yang terdapat *missing value* diisi dengan nilai rata-rata dari kolom tersebut.

2. ***Encoding Categorical Data***

   Tahapan selanjutnya yaitu encoding categorical data atau mengubah data yang bersifat kategorik menjadi numerik. Hal ini dikarenakan mesin hanya dapat membaca data berupa angka saja. Adapun Teknik encoding yang dilakukan adalah label encoder dan one hot encoder. Label encoder digunakan untuk mengubah data kategorik yang bersifat ordinal sedangkan one hot encoder digunakan untuk mengubah data kategorik yang bersifat nominal.

3. ***Handling Outlier***

   Dalam analisis data, keberadaan outlier dapat menjadi sumber ketidakakuratan dan interpretasi yang salah, serta memengaruhi performa model machine learning [14]. Salah satu metode yang umum digunakan untuk mendeteksi dan mengelola outlier adalah dengan menggunakan Z-Score. Teknik ini dapat digunakan untuk mengukur sejauh mana suatu nilai berbeda dari rata-rata dalam satuan deviasi standar. Dengan mengidentifikasi outlier berdasarkan ambang batas tertentu, Z-Score membantu memahami mana data yang mungkin tidak biasa atau merupakan anomali yang signifikan [15].

   Rumus Z-Score digunakan untuk mengukur seberapa jauh suatu nilai darirata-rata dalam satuan deviasi standar. Hal ini dapat membantu untuk mengidentifikasi nilai-nilai yang di luar kisaran normal dan potensial menjadi outlier. Rumus Z-Score adalah sebagai berikut:

$$Z = \frac{x - \mu}{\sigma}$$

   Di mana:

   - Z adalah Z-Score dari suatu nilai x.
   - x adalah nilai yang ingin dihitung Z-Score-nya.
   - μ adalah rata-rata dari seluruh data.
   - σ adalah deviasi standar dari seluruh data.

   penjelasan

   ✓ Ketika nilai x berada persis pada rata-rata μ, maka nilai Z-Score Z akan menjadi 0, menunjukkan bahwa nilai tersebut berada pada posisi rata-rata dalam distribusi data.

   ✓ Jika nilai x lebih besar dari μ, maka Z-Score Z akan positif, menunjukkan bahwa nilai tersebut berada di atas rata-rata.

   ✓ Jika nilai x lebih kecil dari μ,, maka Z-Score Z akan negatif, menunjukkan bahwa nilai tersebut berada di bawah rata-rata.

   Pada penelitian ini, peneliti menerapkan nilai dengan Z-Score lebih besar dari 3 atau lebih kecil dari -3 dianggap sebagai outlier.

4. ***Cek Correlation***

   Tahap selanjutnya yaitu, peneliti akan melakukan pengecekan korelasi data. Koefisien korelasi data yang lemah akan dihapus sehingga menyisakan korelasi data cukup, kuat dan sempurna. Adapun table koefisien korelasi data adalah sebagai berikut:

Tabel 1. Koefisien Korelasi

Koefisien	Keterangan

0	Tidak ada korelasi antara dua variable
>0 – 0,25	Korelasi sangat lemah
>0,25 - 0,5	Korelasi cukup
>0,5 – 0,75	Korelasi kuat
1	Korelasi hubungan sempurna positif
-1	Korelasi hubungan sempurna negatif

D. Processing

Setelah melalui tahap *preprocessing*, selanjutnya adalah tahap *processing* atau pemrosesan data dengan metode Random Forest Regressor. Adapun tahapan processing pada penelitian ini adalah sebagai berikut.

1. *Dataset Splitting*

Pada awal tahap processing, data dibagi menjadi dua bagian yaitu fitur dan target. Fitur merupakan seluruh kolom kecuali kolom "*life_expectancy"* sedangkan target yaitu kolom "*life_expectancy".* Dari data tersebut kemudian dibagi lagi menjadi dua bagian yaitu data uji dan data latih dengan prosentase masing-masing sebesar 80% untuk data latih dan 20% untuk data uji. Data latih ini nantinya digunakan untuk proses pembelajaran mesin sedangkan data uji digunakan untuk pengujian model yang dihasilkan.

2. *Hyperparameter Tuning* dengan *Grid Search Cross Validation*

Hyperparameter tuning adalah proses menemukan kombinasi terbaik dari hyperparameter untuk model machine learning guna mencapai kinerja optimal [16]. Hyperparameter adalah parameter yang tidak dipelajari oleh model selama pelatihan, tetapi mereka mengontrol bagaimana model tersebut belajar dan beroperasi. Salah satu metode populer untuk melakukan hyperparameter tuning adalah dengan menggunakan Grid Search Cross Validation.

Grid Search Cross Validation adalah sebuah metode untuk mencari konfigurasi terbaik dari hyperparameter dalam model machine learning. Hyperparameter adalah pengaturan yang tidak dipelajari oleh model, tetapi memengaruhi bagaimana model belajar [17]. Dalam Grid Search, peneliti membuat "grid" dari semua kombinasi kemungkinan nilai hyperparameter yang telah ditentukan sebelumnya. Adapun hyperparameter yang digunakan pada penelitian ini diambil dari hyperparameter yang disediakan oleh *library* jcopml. Berikut merupakan parameternya:

Table 1. Hyperparameter Tuning

Hyperparameter	Value
n_estimators	[100, 150, 200]
max_depth	[20, 50, 80]
max_features	[0.3, 0.6, 0.8]
min_samples_leaf	[1, 5, 10]

3. Pemodelan dengan Random Forest Regressor

Penelitian ini menggunakan metode Random Forest regressor untuk memprediksi angka harapan hidup penduduk di beberapa wilayah Asia. Random Forest Regressor adalah sebuah algoritma *machine learning* yang digunakan untuk melakukan prediksi terhadap variabel target yang bersifat numerik atau kontinu dalam konteks penelitian atau analisis data [18].

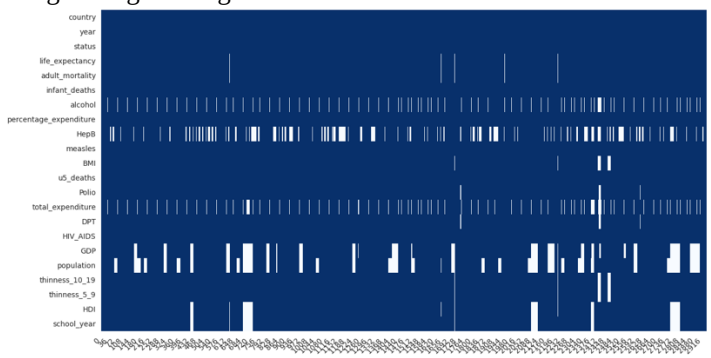
E. Evaluasi

Tahap evaluasi dalam konteks *machine learning* adalah proses mengukur kinerja dan keefektifan model yang telah dibangun. Evaluasi model penting untuk memahami sejauh mana model tersebut mampu melakukan prediksi yang akurat dan berguna dalam praktiknya. Peneliti menggunakan *Mean Absolute Error* (MAE) dimana metode ini digunakan untuk mengukur sejauh mana model prediksi numerik mendekati nilai sebenarnya dalam data pengujian. MAE menghitung rata-rata dari selisih absolut antara prediksi model dan nilai sebenarnya [19].

III. HASIL DAN PEMBAHASAN

A. Preprocessing

Tahapan ini menghasilkan data yang telah dibersihkan dan diubah. Visualisasi dari kelengkapan data ditunjukkan pada Gambar 2, dimana area berwarna putih menunjukkan keberadaan nilai yang hilang atau missing value pada kolom tertentu. Dari Gambar 2, dapat dilihat bahwa beberapa kolom memiliki missing value dalam jumlah yang beragam. Khusus untuk kolom target, yaitu kolom ekspektasi hidup yang mengalami missing value, pendekatan yang diambil adalah dengan menghapus baris yang mengandung missing value pada kolom tersebut. Untuk kolom lain, penulis menggunakan metode imputasi rata-rata berdasarkan negara dan status pembangunan negara tersebut (baik itu negara berkembang atau maju), dikarenakan adanya perbedaan signifikan dalam angka harapan hidup antara negara-negara dengan status berbeda tersebut.



Gambar 3. Sebelum Imputasi *Missing Value*

Hasil imputasi dapat dilihat pada Gambar 6, yang menunjukkan bahwa seluruh data telah terisi lengkap tanpa adanya missing value. Visualisasi pada Gambar 6 menunjukkan warna yang seragam, yang mengindikasikan bahwa nilai-nilai yang hilang telah berhasil diimputasi.

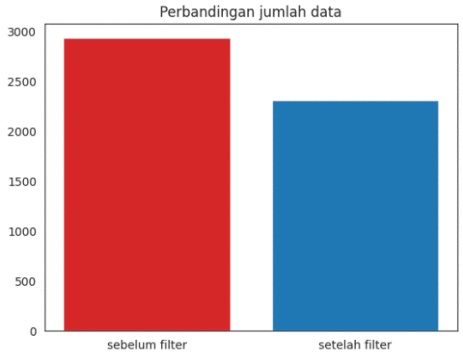
Gambar 4. Setelah Imputasi *Missing Value*

Setelah memastikan bahwa setiap kolom dalam dataset tidak ada *missing value*, langkah berikutnya adalah melakukan pemeriksaan korelasi. Hal ini bertujuan untuk menentukan kolom mana yang akan digunakan dalam pelatihan model. Peneliti memfokuskan pada korelasi dengan tingkat dari sedang hingga sangat kuat, sehingga kolom dengan korelasi sangat rendah atau rendah akan dihilangkan. Seperti yang diuraikan sebelumnya, untuk memudahkan dalam menentukan korelasi antara fitur dengan target, peneliti memanfaatkan fungsi `corr` dari pandas dan hasilnya divisualisasikan pada Gambar 8.

Gambar 5. Visualisasi Korelasi

Berdasarkan Gambar 4, peneliti kemudian melanjutkan dengan proses pemilihan fitur, atau yang dikenal sebagai feature selection. Dalam proses ini, kolom dengan tingkat korelasi sangat rendah (dalam rentang 0,00 hingga 0,199) dan rendah (dalam rentang 0,20 hingga 0,399) dihilangkan. Akibatnya, terpilih 6 kolom yaitu 'year', 'population', 'measles', 'percentage_expenditure', 'infant_deaths', dan 'u5_deaths', sehingga total terdapat 16 kolom yang digunakan untuk pelatihan model.

Tahap berikutnya adalah penyaringan outlier menggunakan metode z-score, seperti yang telah dijelaskan sebelumnya. Hasil dari penyaringan outlier ini adalah pengurangan jumlah data sekitar 600 baris. Untuk memvisualisasikan perbedaan jumlah data sebelum dan sesudah penyaringan outlier, visualisasi tertentu disajikan.



Gambar 6. Perbandingan Jumlah Data

B. *Processing* dan Evaluasi

Setelah melalui tahapan *training* dengan *grid search cross validation,* selanjutnya didapatkan *hyperparameter* terbaik sebagai berikut.

Table 2. Hasil *Hyperparameter Tuning*

Hyperparameter	Value
n_estimators	150
max_depth	20
max_features	0.6
min_samples_leaf	1

Berdasarkan parameter optimal yang diperoleh dari *cross validation* menggunakan pencarian grid, yang tercantum dalam Tabel 4, skor pelatihan yang diperoleh adalah 99,4%, sementara skor pengujian adalah 96,5%. Selain itu, nilai mean absolute error (MAE) yang tercapai adalah 0.99. Nilai MAE yang lebih dekat ke 0 menandakan bahwa model yang dikembangkan memiliki kinerja yang lebih baik.

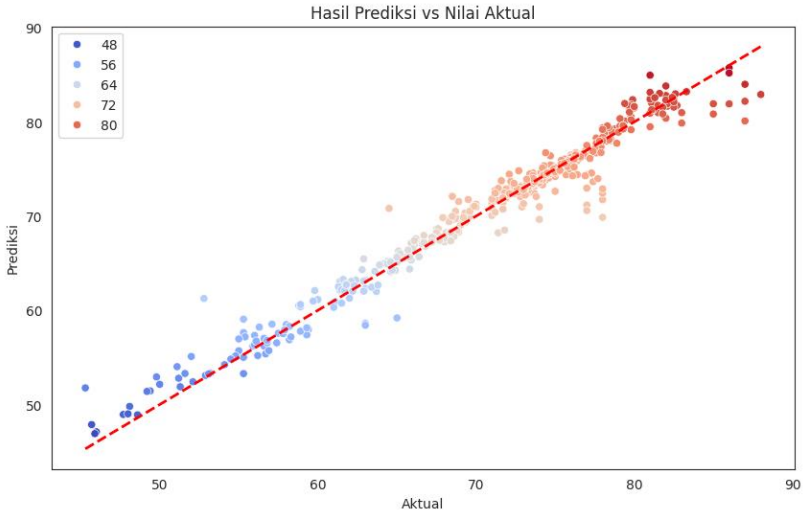
C. Uji Coba Model

Setelah model dibangun, tahapan selanjutnya yaitu pengujian model pada data tes. Berikut merupakan sampel hasil pengujian.

Table 3. Sampel Hasil Pengujian

No	Y_test	Y_predict
1	77,3	76,586
2	75,8	75,90011852
3	81,7	81,393
4	47,7	48,94666667
5	79,8	81,38066667
6	73,6	73,33817308
7	71	71,35915641
8	77,6	78,284
9	72,6	72,57580346
10	61,4	63,056

Dari Tabel 3 terlihat bahwa model prediksi memiliki tingkat akurasi yang cukup baik karena hasil prediksi (Y_predict) cukup dekat dengan nilai sebenarnya (Y_test). Gambar 7 menampilkan visual dari pengujian model yang telah dikembangkan. Garis putus-putus melambangkan nilai usia yang sebenarnya, sedangkan titik-titik yang tersebar mewakili prediksi usia yang dihasilkan oleh model.



Gambar 7. Visualisasi Uji Coba Model

Grafik yang disajikan pada gambar 7 menunjukkan hasil pemodelan dengan membandingkan nilai aktual dengan prediksi. Dari visualisasi tersebut, titik-titik data tampaknya tersebar di sekitar garis diagonal (ditunjukkan oleh garis putus-putus merah), yang menunjukkan hubungan antara nilai prediksi dan aktual. Seiring dengan meningkatnya nilai aktual, nilai prediksi juga meningkat, yang mengindikasikan bahwa model memiliki kapasitas yang baik dalam memprediksi nilai yang benar. Warna dari titik-titik tersebut bervariasi dari biru muda ke merah, merepresentasikan densitas dari data yang berbeda pada nilai harapan hidup, dengan biru muda menunjukkan nilai yang lebih rendah dan merah menunjukkan nilai yang lebih tinggi. Secara umum, ketepatan prediksi model terlihat baik, mengingat sebagian besar titik-titik berada dekat dengan garis yang sesuai dengan nilai aktual.

VII. SIMPULAN

Model yang telah dikembangkan dalam penelitian ini menunjukkan kemampuan yang sangat baik dalam memprediksi angka harapan hidup, dengan tingkat akurasi yang tercatat sebesar 96,5%. Dari hasil analisis yang dilakukan, tercatat nilai Mean Absolute Error (MAE) sebesar 0,99, yang merefleksikan rata-rata deviasi prediksi dari nilai aktual. Temuan ini memperkuat kemampuan model Random Forest Regressor sebagai instrumen prediktif yang canggih untuk memperkirakan angka harapan hidup di berbagai negara Asia.

Selain itu, berdasarkan hasil dari korelasi fitur-fitur terhadap target, terlihat bahwa fitur "school_year", "HDI", dan "BMI" memiliki korelasi atau pengaruh yang cukup tinggi terhadap angka harapan hidup. Hal ini bisa dikarenakan hubungan yang kuat antara "school_year", "HDI", dan "BMI" dengan kesejahteraan masyarakat dan akses terhadap layanan kesehatan di berbagai negara Asia. Indeks Pembangunan Manusia (HDI) mencerminkan tingkat kesejahteraan dan distribusi pendapatan, sedangkan jumlah tahun bersekolah (school_year) berkaitan dengan pengetahuan masyarakat tentang gaya hidup sehat, dan rata-rata Body Mass Index (BMI) mencerminkan pola makan dan gaya hidup. Hubungan positif antara HDI, jumlah tahun bersekolah, dan BMI dengan angka harapan hidup menunjukkan bahwa model Random Forest Regressor dapat menjadi alat yang canggih untuk memprediksi angka harapan hidup dengan tingkat akurasi yang tinggi. Temuan ini dapat mendukung upaya-upaya kebijakan untuk meningkatkan kesejahteraan dan kualitas hidup masyarakat di berbagai negara Asia dengan fokus pada pendidikan, pembangunan manusia, pola makan sehat, dan gaya hidup yang aktif.

UCAPAN TERIMA KASIH

Terimakasih peneliti ucapkan kepada pihak-pihak terkait yang membantu dalam berjalannya penelitian ini. Terimakasih juga peneliti ucapkan kepada GHO sebagai penyedia data terbuka sehingga dapat peneliti gunakan untuk penelitian ini**.**

REFERENSI

[1] R. Muda, R. Koleangan, and J. Bintang Kalangi, “PENGARUH ANGKA HARAPAN HIDUP, TINGKAT PENDIDIKAN DAN PENGELUARAN PERKAPITA TERHADAP PERTUMBUHAN EKONOMI DI SULAWESI UTARA PADA TAHUN 2003-2017,” *Jurnal Berkala Ilmiah Efisiensi*, 2019.

[2] F. Winston, N. 1∗, M. D. Pangastuti, and Y. R. S. S. S. B. Utami, “Analisis determinan faktor penentu usia harapan hidup di Provinsi Nusa Tenggara Timur,” *Jurnal Ekonomi, Keuangan dan Manajemen*, vol. 18, no. 3, p. 459, 2022, doi: 10.29264/jinv.v18i3.10813.

[3] I. Arofah and S. Rohimah, “ANALISIS JALUR UNTUK PENGARUH ANGKA HARAPAN HIDUP, HARAPAN LAMA SEKOLAH, RATA-RATA LAMA SEKOLAH TERHADAP INDEKS PEMBANGUNAN MANUSIA MELALUI PENGELUARAN RIIL PER KAPITA DI PROVINSI NUSA TENGGARA TIMUR,” *JURNAL SAINTIKA UNPAM*, 2019.

[4] P. Parulian *et al.*, “Analysis of Sequential Order Incremental Methods in Predicting the Number of Victims Affected by Disasters,” in *Journal of Physics: Conference Series*, 2019. doi: 10.1088/1742-6596/1255/1/012033.

[5] A. Wanto *et al.*, “Forecasting the Export and Import Volume of Crude Oil, Oil Products and Gas Using ANN,” in *Journal of Physics: Conference Series*, 2019. doi: 10.1088/1742-6596/1255/1/012016.

[6] A. Sekulić, M. Kilibarda, G. B. M. Heuvelink, M. Nikolić, and B. Bajat, “Random forest spatial interpolation,” *Remote Sens (Basel)*, vol. 12, no. 10, 2020, doi: 10.3390/rs12101687.

[7] M. Schonlau and R. Y. Zou, “The random forest algorithm for statistical learning,” *Stata Journal*, vol. 20, no. 1, 2020, doi: 10.1177/1536867X20909688.

[8] M. Erkamim, S. Suswadi, M. Z. Subarkah, and E. Widarti, “Komparasi Algoritme Random Forest dan XGBoosting dalam Klasifikasi Performa UMKM,” *Jurnal Sistem Informasi Bisnis*, vol. 13, no. 2, pp. 127–134, Oct. 2023, doi: 10.21456/vol13iss2pp127-134.

[9] R. Oktavia *et al.*, “Penerapan Metode Algoritma K-means Dalam Pengelompokan Angka Harapan Hidup Saat Lahir Menurut Provinsi.” [Online]. Available: www.bps.go.id/

[10] S. P. Sinaga *et al.*, “Implementasi Jaringan Syaraf Tiruan Resilient Backpropagation dalam Memprediksi Angka Harapan Hidup Masyarakat Sumatera Utara,” *Jurnal Infomedia*, vol. 4, no. 2, 2019.

[11] T. Afriliansyah and Z. Zulfahmi, “Prediksi Angka Harapan Hidup Masyarakat Aceh dengan Model Terbaik Algoritma Cyclical Order,” *Prosiding Seminar Nasional Riset Dan Information Science (SENARIS)*, vol. 2, pp. 441–449, 2020.

[12] J. S. Komputer, K. Buatan, and B. Muhammad, “Implementasi Data Mining untuk Prediksi Standar Hidup Layak Berdasarkan Tingkat Kesehatan dan Pendidikan Masyarakat,” *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, vol. 2, no. 2, 2019.

[13] S. F. Manurung, A. Andriansya, J. Permana, and R. Pangestu, “Pemanfaatan Algoritma JST untuk Menentukan Model Prediksi Umur Harapan Hidup Saat Lahir,” *Hello World Jurnal Ilmu Komputer*, vol. 1, no. 1, pp. 19–35, May 2022, doi: 10.56211/helloworld.v1i1.9.

[14] P. R. , , D. A. Sihombing, S. Suryadiningrat, and Y. P. A. C. Yuda, . “Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya,” *Jurnal Ekonomi dan Statistik Indonesia*, pp. 307–316, 2022.

[15] C. Nkikabahizi, W. Cheruiyot, and A. Kibe, "Chaining Zscore and feature scaling methods to improve neural networks for classification[Formula presented]," *Appl Soft Comput*, vol. 123, 2022, doi: 10.1016/j.asoc.2022.108908.
[16] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, 2021, doi: 10.3390/informatics8040079.
[17] D. A. Anggoro and N. A. Afdallah, "Grid Search CV Implementation in Random Forest Algorithm to Improve Accuracy of Breast Cancer Data," *Int J Adv Sci Eng Inf Technol*, vol. 12, no. 2, 2022, doi: 10.18517/ijaseit.12.2.15487.
[18] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int J Distrib Sens Netw*, vol. 18, no. 6, 2022, doi: 10.1177/15501329221106935.
[19] D. S. K. Karunasingha, "Root mean square error or mean absolute error? Use their ratio as well," *Inf Sci (N Y)*, vol. 585, 2022, doi: 10.1016/j.ins.2021.11.036.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.