# Predicting Life Expectancy of Population Using XGBoost Method [Prediksi Angka Harapan Hidup Penduduk Menggunakan Metode XGBoost]

Wildan Kurniawan*,1), Uce Indahyanti 2)

1)Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
2)Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
*Email Penulis Korespondensi: uceindahyanti@umsida.ac.id

***Abstract***. *This research aims to predict life expectancy in several Asian countries using the XGBoost Regressor algorithm. The data used is sourced from the UCI Machine Learning Repository. In this study, the researchers construct a predictive model using a machine learning approach and evaluate it based on accuracy and Mean Absolute Error (MAE). The research findings indicate that the XGBoost Regressor model achieves an accuracy rate of 96.8% in predicting life expectancy. The obtained MAE value is 0.97. These results highlight the potential of the XGBoost Regressor algorithm as an effective tool for predicting life expectancy in the Asian region. These findings could have positive implications for data-driven decision-making and welfare policy planning..*

***Keywords –*** *Asian Countries ; Life Expectancy Prediction; XGBoost*

***Abstrak***. *Penelitian ini bertujuan untuk memprediksi angka harapan hidup di beberapa negara wilayah Asia menggunakan algoritma XGBoost Regressor. Data yang digunakan berasal dari UCI Machine Learning Repository. Dalam penelitian ini, peneliti membangun model prediksi menggunakan pendekatan machine learning dan melakukan evaluasi berdasarkan tingkat akurasi dan Mean Absolute Error (MAE). Hasil penelitian menunjukkan bahwa model XGBoost Regressor memiliki tingkat akurasi sebesar 96,8% dalam memprediksi angka harapan hidup. Nilai MAE yang diperoleh adalah sebesar 0,97. Temuan ini menunjukkan potensi algoritma XGBoost Regressor sebagai alat yang efektif dalam memprediksi angka harapan hidup di wilayah Asia. Hasil ini dapat memiliki implikasi positif dalam pengambilan keputusan berbasis data serta perencanaan kebijakan kesejahteraan masyarakat.*

***Kata Kunci -*** *Negara-Negara Asia ; Prediksi Angka Harapan Hidup; XGBoost*

## I. PENDAHULUAN

Harapan hidup adalah jumlah tahun keseluruhan yang telah dilewati oleh seseorang setelah mencapai usia tertentu [1]. Ada beberapa faktor yang memengaruhi harapan hidup seseorang. Ini bisa digunakan sebagai cara untuk menilai sejauh mana pemerintah berhasil meningkatkan kesejahteraan masyarakat. Upaya untuk memperbaharui populasi negara harus disertai dengan rencana pembangunan kesehatan dan inisiatif sosial lainnya. Ini termasuk menjaga kebersihan lingkungan, memastikan asupan nutrisi dan kalori yang memadai, serta mengimplementasikan program untuk mengatasi masalah kemiskinan [2].

Berdasarkan data dari PBB yang mencakup periode tahun 1995 hingga 2015, angka harapan hidup global dihitung setiap interval lima tahun. Informasi ini diperoleh dari "Proyeksi populasi dunia: Database Kependudukan Revisi 2010" yang dikeluarkan oleh Perserikatan Bangsa-Bangsa. Data tersebut memberikan gambaran mengenai perubahan angka harapan hidup secara global dalam periode waktu yang dianalisis [3]. Badan Pusat Statistik (BPS) Indonesia menemukan bahwa dari tahun 2010 hingga 2015, Jepang menjadi negara dengan harapan hidup tertinggi di kawasan Asia, mencapai 83,5 tahun. Sementara itu, Hong Kong menempati peringkat kedua dengan harapan hidup sebesar 83,3 tahun dalam periode yang sama [4]. Indonesia berada di peringkat ke-14 dari 19 negara dalam hal angka harapan hidup, dengan rata-rata angka harapan hidup sebesar 70,1 tahun. Peringkat ini mengikuti beberapa negara Asia Tenggara lainnya, dengan Singapura memiliki angka harapan hidup tertinggi yaitu 82,2 tahun, diikuti oleh Vietnam dengan 75,9 tahun, Malaysia dengan 74,9 tahun, Thailand dengan 74,3 tahun, dan Kamboja dengan 71,6 tahun.

Pengetahuan mengenai angka harapan hidup penduduk dunia memiliki signifikansi yang besar, terutama di kawasan Asia. Ini memberikan pedoman dan referensi yang penting bagi negara-negara untuk menetapkan target angka harapan hidup serta merencanakan langkah-langkah strategis yang sesuai, khususnya untuk Indonesia. Dalam konteks ini, algoritma machine learning seperti XGBoost regressor memiliki potensi yang besar dalam melakukan prediksi terhadap situasi ini.

XGBoost regressor adalah alat yang efektif untuk memodelkan hubungan kompleks antara variabel-variabel yang mempengaruhi angka harapan hidup, seperti faktor sosial, ekonomi, kesehatan, dan lingkungan. Dengan menggunakan data historis dan informasi aktual, algoritma ini dapat membangun model yang mampu memprediksi angka harapan

hidup di masa depan. Hal ini dapat memberikan panduan berharga bagi pembuat kebijakan untuk mengambil langkah-langkah yang tepat dalam upaya meningkatkan kualitas hidup penduduk.

Pemanfaatan XGBoost regressor dalam memprediksi angka harapan hidup dapat membantu pemerintah dan lembaga terkait dalam menyusun rencana pembangunan jangka panjang, memperbaiki sistem kesehatan, serta mengidentifikasi area-area prioritas untuk intervensi dan perbaikan. Dengan demikian, pendekatan ini memiliki potensi untuk mengarahkan upaya menuju peningkatan harapan hidup dan kesejahteraan masyarakat di Asia, termasuk Indonesia [5].

Beberapa penelitian pernah dilakukan sebelumnya, seperti penelitian yang dilakukan oleh Teuku Afriiliansyah dan Zulfahmi pada tahun 2020 berjudul Memprediksi Harapan Hidup Masyarakat di Aceh Menggunakan Model Algoritma Cyclic Order. Tujuan dari penelitian ini adalah memprediksi angka harapan hidup masyarakat Aceh dengan menggunakan algoritma best-fit cyclic order weight/bias. Data survei yang digunakan adalah data angka harapan hidup 23 kabupaten/kota di Aceh yang disediakan oleh Badan Pusat Statistik (BPS). Model terbaik yang digunakan untuk prediksi adalah model 8-9-1 (8 input layer, 9 hidden layer neuron, 1 output) dengan akurasi 91% dengan nilai MSE training sebesar 0.0009907466, dan pada pengujian didapatkan MSE sebesar 0.0010800577. Hasil dari penelitian ini adalah proyeksi angka harapan hidup masyarakat Aceh dari tahun 2020 hingga tahun 2022 [1].

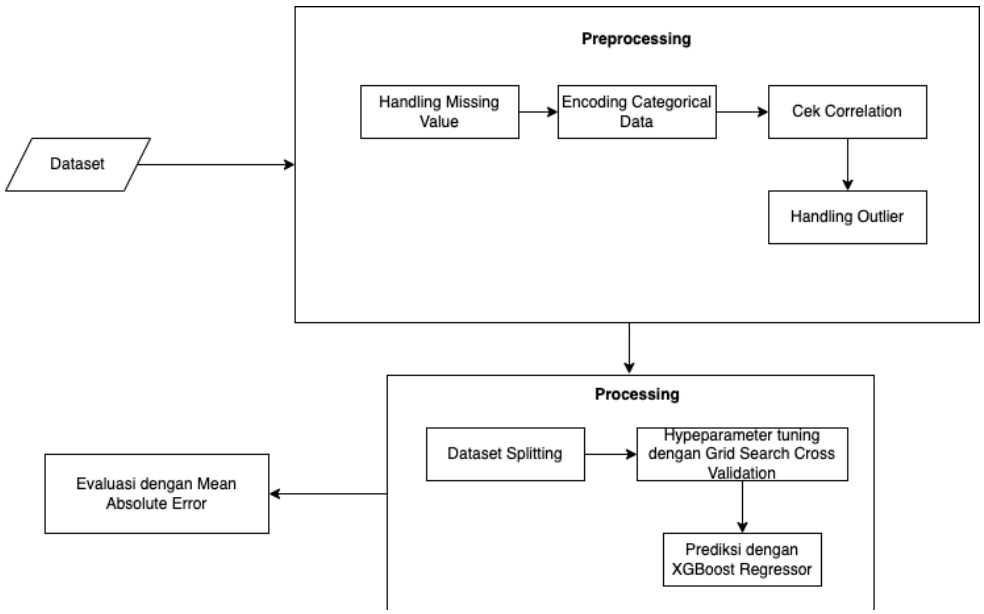
Penelitian selanjutnya dilakukan oleh Samuel Palentino Sinaga, Anjar Wanto dan Solikhun pada tahun 2019 yang berjudul Implementasi Jaringan Syaraf Tiruan Resilient Backpropagation dalam Memprediksi Angka Harapan Hidup Masyarakat Sumatera Utara. Penelitian ini menerapkan Jaringan Syaraf Tiruan Resilient Backpropagation untuk memprediksi angka harapan hidup masyarakat Sumatera Utara. Metode yang digunakan menghasilkan akurasi 88% [6].

Meskipun penelitian-penelitian sebelumnya telah berhasil memprediksi angka harapan hidup dengan menggunakan algoritma Cyclic Order dan Jaringan Syaraf Tiruan Resilient Backpropagation, namun belum ada penelitian yang secara khusus menerapkan algoritma XGBoost untuk prediksi angka harapan hidup. Penggunaan metode XGBoost, yang dikenal memiliki kemampuan prediksi yang kuat dalam berbagai bidang, dapat memberikan kontribusi baru dalam analisis dan prediksi angka harapan hidup penduduk. Selain itu, penelitian ini juga dapat membantu mengidentifikasi variabel-variabel penting yang berkontribusi pada prediksi angka harapan hidup, yang mungkin memiliki dampak signifikan terhadap perencanaan kebijakan kesehatan dan sosial di masa depan. Oleh karena itu, penelitian ini akan mengisi kesenjangan pengetahuan dengan mengaplikasikan metode XGBoost yang inovatif dan dapat memberikan wawasan baru dalam meramalkan angka harapan hidup penduduk.

## 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. Pengumpulan Data**

Pada penelitian ini, data yang digunakan diambil dari situs UCI Machine learning repository berupa data angka harapan hidup masyarakat di beberapa negara wilayah Asia. Data tersebut terdiri dari 2938 record dari tahun 2000 hingga 2015. Adapun atribut dari data tersebut adalah sebagai berikut.

**Table 1. Atribut Dataset**

| Attribut | Keterangan |
|---|---|
| Country | Negara |
| year | Tahun |
| status | Developed = Maju<br>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 |

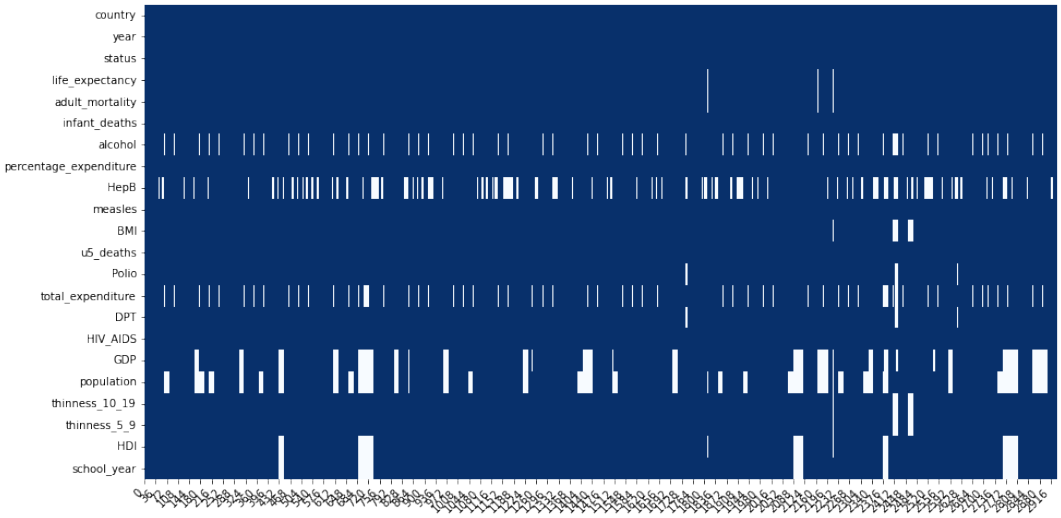
**C. *Preprocessing***

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

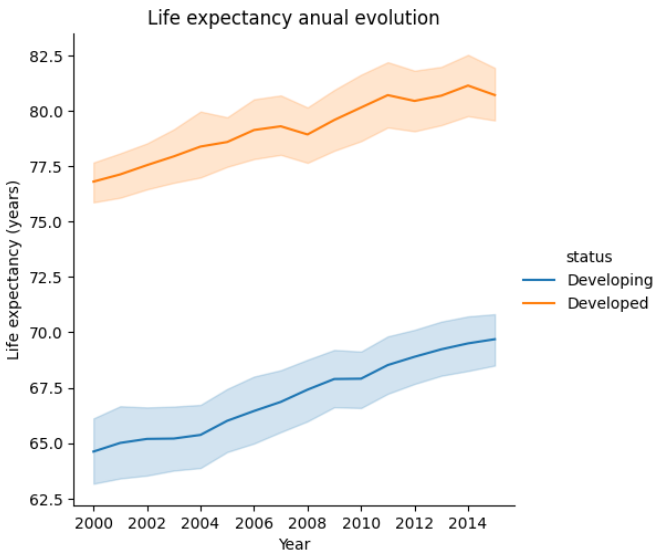
signifikan pada hasil akhir prediksi. Pada bagian ini peneliti akan membahas tahapan-tahapan penting dalam mempersiapkan data sebelum proses analisis dan prediksi menggunakan metode XGBoost.

**1. *Handling Missing Value***

Gambar 2 merupakan visualisasi dari kelengkapan jumlah data. Warna putih menandakan adanya *missing value* atau data kosong pada suatu kolom. pada gambar 2 terlihat dengan jelas bahwa beberapa kolom memiliki *missing value* dengan jumlah yang bervariasi. Karena pada kolom target yaitu kolom life expectancy mengalami *missing value,* maka teknik yang dilakukan adalah *drop missing value* atau pengahapusan baris pada kolom tersebut. Sedangkan pada kolom lainnya peneliti menggunakan teknik imputasi mean berdasarkan negara dan status negara tersebut (berkembang atau maju). alasannya adalah karena pada negara yang berstatus maju dan berkembang mengalami ketimpangan angka harapan hidup atau usia penduduk di negara dengan status tersebut. pada gambar 3 terlihat jelas bahwa angka harapan hidup penduduk di negara berkembang dan maju mengalami perbedaan yang signifikan dari tahun 2000 sampai dengan 2015. Dimana pada negara maju angka harapan hidupnya lebih tinggi daripada negara berkembang.




**Gambar 2 Visualisasi Missing Value**




**Gambar 3 Visualisasi Usia Penduduk Berdasarkan Status**

**2. *Encoding Categorical Data***

Pada deskripsi data yang terdapat pada gambar 4 terlihat bahwa kolom-kolom yang digunakan pada penelitian ini memiliki 3 tipe data yang berbeda salah satunya adalah tipe data object yang artinya kolom tersebut bersifat kategorik. Adapun kolom yang bersifat kategorik adalah kolom ‘country’ dan ‘status’.

Kedua kolom tersebut tidak dapat diproses oleh mesin karena mesin tidak dapat memproses data yang bersifat kategorik, maka peneliti melakukan *encoding categorical data* menggunakan teknik label encoder dimana nantinya setiap kategori diberi label berupa angka bilangan bulat. Selanjutnya, agar kolom-kolom pada penelitian ini memiliki satu tipe data yang sama, maka semua kolom diubah menjadi tipe data float agar tidak menghilangkan nilai desimal dari kolom yang awalnya sudah bernilai float.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   country                 2938 non-null   object
 1   year                    2938 non-null   int64
 2   status                  2938 non-null   object
 3   life_expectancy         2928 non-null   float64
 4   adult_mortality         2928 non-null   float64
 5   infant_deaths           2938 non-null   int64
 6   alcohol                 2744 non-null   float64
 7   percentage_expenditure  2938 non-null   float64
 8   HepB                    2385 non-null   float64
 9   measles                 2938 non-null   int64
 10  BMI                     2904 non-null   float64
 11  u5_deaths               2938 non-null   int64
 12  Polio                   2919 non-null   float64
 13  total_expenditure       2712 non-null   float64
 14  DPT                     2919 non-null   float64
 15  HIV_AIDS                2938 non-null   float64
 16  GDP                     2490 non-null   float64
 17  population              2286 non-null   float64
 18  thinness_10_19          2904 non-null   float64
 19  thinness_5_9            2904 non-null   float64
 20  HDI                     2771 non-null   float64
 21  school_year             2775 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

**Gambar 4 Deskripsi Data**

**3. *Correlation dan Feature Selection***

Korelasi data merupakan hubungan antara fitur (X) dengan target (y). korelasi dapat dilihat hubungannya dari negatif lemah sampai korelasi kuat. Korelasi lemah menandakan jika fitur tidak mempengaruhi target. Menurut penelitian yang dilakukan oleh sugiyono pada tahun 2007 yang berisi pedoman untuk memberikan interpretasi koefisien korelasi adalah sebagai berikut.

**Table 2 Interpretasi Koefisien Korelasi**

| Interval Koefisien | Tingkat Hubungan |
|---|---|
| 0,00 – 0,199 | Sangat rendah |
| 0,20 – 0,399 | Rendah |
| 0,40 – 0,599 | Sedang |
| 0,60 – 0,799 | Kuat |
| 0,80 – 1,00 | Sangat Kuat |

Berdasarkan tabel 2 di atas, peneliti hanya mengambil *feature* yang memiliki tingkat hubungan sedang, kuat dan sangat kuat terhadap target. Untuk menentukan koefisien tiap *feature* terhadap target, peneliti menggunakan *function* milik pandas yaitu *corr* yang selanjutnya divisualisasikan dengan library plotly.

**4. *Handling Outlier***

Dalam analisis data, keberadaan outlier dapat menjadi sumber ketidakakuratan dan interpretasi yang salah, serta memengaruhi performa model machine learning [7]. 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 [8].

Rumus Z-Score digunakan untuk mengukur seberapa jauh suatu nilai dari rata-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:

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

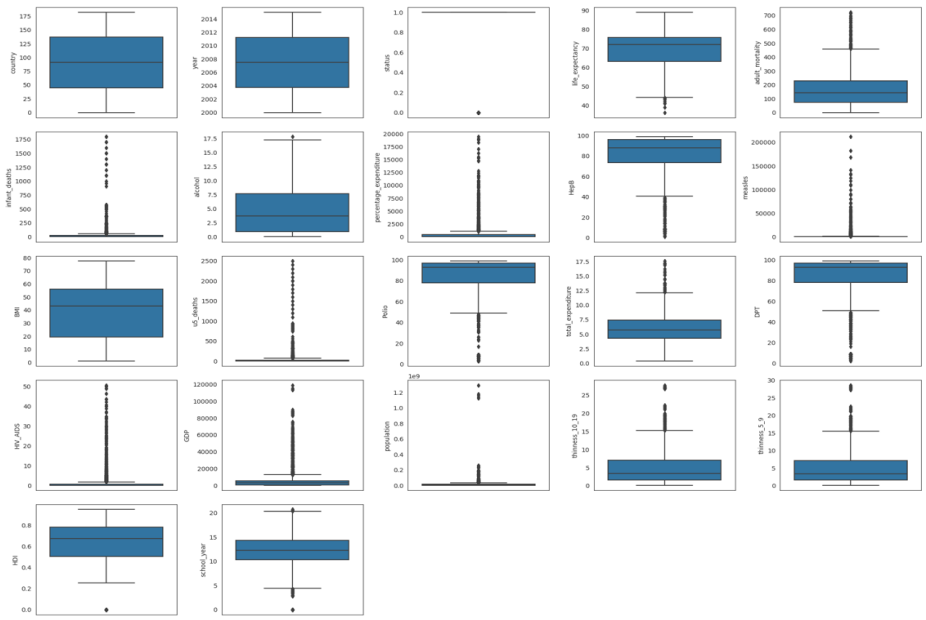
**Di mana:**

- $Z$ adalah Z-Score dari suatu nilai $x$.
- $x$ adalah nilai yang ingin dihitung Z-Score-nya.
- $\mu$ adalah rata-rata dari seluruh data.
- $\sigma$ adalah deviasi standar dari seluruh data.

**penjelasan**

- ✓ Ketika nilai $x$ berada persis pada rata-rata $\mu$, 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 $\mu$, maka Z-Score $Z$ akan positif, menunjukkan bahwa nilai tersebut berada di atas rata-rata.
- ✓ Jika nilai $x$ lebih kecil dari $\mu$,, 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.




**Gambar 5 Visualisasi Outlier**

Gambar 5 merupakan visualisasi data tiap kolom dalam bentuk boxplot dimana pada gambar tersebut terlihat bahwa beberapa kolom memiliki data yang bersifat outlier.

## D. *Processing*

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

### 1. *Dataset Splitting*

Pada awal tahap processing, data dibagi 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 [9]. 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 [10]. Dalam Grid Search, peneliti membuat "grid" dari semua kombinasi kemungkinan nilai hyperparameter yang telah ditentukan sebelumnya. Adapun hyperparameter yang digunakan pada penelitian ini adalah sebagai berikut.

**Table 3 Hyperparameter Tuning**

| Hyperparameter | Value | Fungsi |
|---|---|---|
| Min_child_weight | 4, 6, 8 | Parameter ini mengontrol minimum jumlah sampel yang diperlukan di setiap node daun (leaf node) dalam pohon. Mengatur nilai ini dapat membantu mencegah pembentukan cabang-cabang kecil yang terlalu spesifik, yang dapat mengurangi overfitting. |
| Max_depth | 8, 10, 12 | Parameter ini mengontrol kedalaman maksimum pohon yang akan dibangun. Ini membatasi kompleksitas model dan membantu menghindari overfitting. Semakin rendah nilai max_depth, semakin dangkal pohon dan semakin sederhana modelnya. |
| eta | 0.3, 0.03 | Eta adalah laju pembelajaran, yaitu seberapa besar kontribusi dari setiap pohon dalam model. Nilai yang lebih rendah membuat model belajar lebih lambat dan bisa membantu menghindari overfitting, sementara nilai yang lebih tinggi dapat memberikan model kecepatan pembelajaran yang lebih cepat. |
| Learning_rate | 0.01, 0.1 | Ini mirip dengan eta. Ini adalah faktor dengan nilai antara 0 dan 1 yang mengontrol seberapa besar kontribusi setiap pohon ke dalam model. Semakin rendah learning rate, semakin kecil pengaruh setiap pohon, dan pembelajaran model menjadi lebih lambat dan stabil. |
| reg_alpha | 0.1, 1, 3 | Ini adalah parameter regularisasi yang mengendalikan kompleksitas model dengan menambahkan biaya terhadap berat koefisien dalam model. Ini membantu menghindari overfitting dengan membatasi bobot yang lebih besar. |
| reg_lambda | 0.1, 1, 2, 3 | Ini adalah parameter regularisasi tambahan yang mirip dengan reg_alpha, tetapi lebih fokus pada norma L2 dari bobot koefisien. Ini juga membantu dalam mengontrol kompleksitas model dan mengurangi risiko overfitting. |

**3. Pemodelan dengan XGBoost**

Penelitian ini menggunakan metode XGBoost regressor untuk memprediksi angka harapan hidup penduduk di beberapa wilayah Asia. XGBoost 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. Metode ini merupakan implementasi yang sangat efektif dalam pengembangan model prediktif berbasis pohon keputusan yang ditingkatkan dengan gradient boosting. Cara kerja XGBoost Regressor secara formal adalah sebagai berikut:

**1. Inisialisasi Model Awal**

Algoritma ini dimulai dengan model awal yang sederhana, seringkali berupa rata-rata dari variabel target yang ingin diprediksi. Sebagai contoh, jika variabel target adalah harga rumah, model awal mungkin adalah rata-rata harga rumah dalam dataset.

2. **Iterasi Bertahap**
   XGBoost melakukan iterasi untuk memperbaiki model awal tersebut. Ini dilakukan dengan menambahkan pohon keputusan (decision tree) ke dalam model secara bertahap. Setiap pohon keputusan yang ditambahkan berusaha untuk mengurangi kesalahan prediksi yang dihasilkan oleh pohon-pohon sebelumnya.
3. **Penekanan pada Kesalahan**
   XGBoost memberikan bobot lebih besar pada data yang sulit diprediksi. Data yang telah salah diperkirakan oleh pohon-pohon sebelumnya diberi perhatian lebih besar sehingga pohon-pohon berikutnya fokus untuk memperbaiki kesalahan tersebut.
4. **Penggabungan Hasil Pohon**
   Hasil dari seluruh pohon keputusan yang digunakan dijumlahkan bersama untuk menghasilkan prediksi akhir. Masing-masing pohon memberikan kontribusi dalam membuat prediksi akhir ini, dengan pohon-pohon yang lebih baik dalam memprediksi data memberikan pengaruh yang lebih besar.
5. **Regularisasi**
   XGBoost juga menerapkan teknik regularisasi untuk menghindari overfitting, yaitu kondisi saat model terlalu sesuai dengan data pelatihan dan tidak dapat menggeneralisasi dengan baik pada data baru.

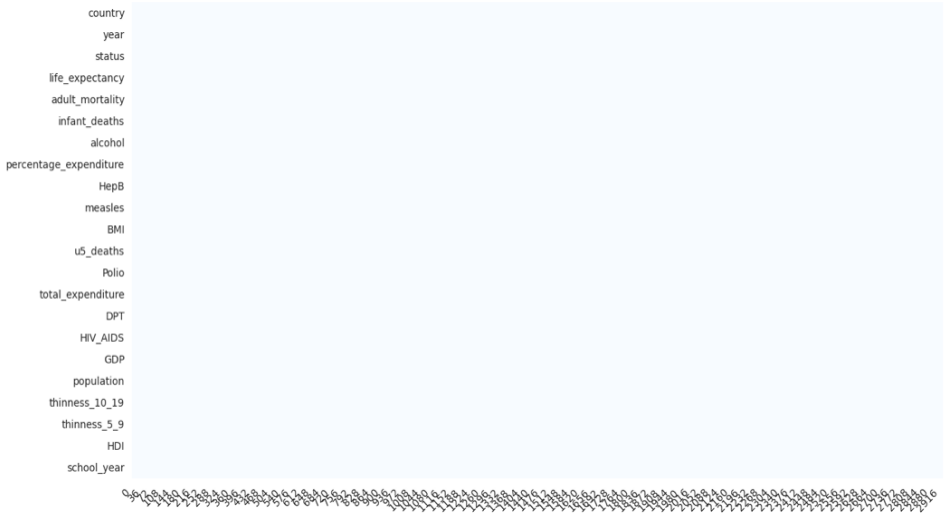
### E. Evaluasi

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 (seperti XGBoost Regressor) mendekati nilai sebenarnya dalam data pengujian. MAE menghitung rata-rata dari selisih absolut antara prediksi model dan nilai sebenarnya [11].

## III. HASIL DAN PEMBAHASAN

### A. Preprocessing

Hasil dari tahapan ini adalah data yang telah dibersihkan dan dimodifikasi. Pada gambar 6 terlihat bahwa data telah terisi secara keseluruhan (tidak ada *missing value*). Visualisasi pada gambar 6 menampilkan warna yang sama. Hal ini menandakan bahwa *missing value* telah ter-*impute*.




**Gambar 6 Visualisasi Setelah Handling Missing Value**

Selanjutnya yaitu pada gambar 7 terlihat bahwa semua kolom memiliki satu tipe data yang sama yaitu float. Hal ini dikarenakan data yang bersifat kategorik telah dilakukan encoding dengan teknik label encoder seperti yang telah dijelaskan pada sub bab sebelumnya. Setelah dilakukan encoding, seluruh kolom diubah tipe datanya menjadi float.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2928 entries, 0 to 2927
Data columns (total 22 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   country                 2928 non-null   float64
 1   year                    2928 non-null   float64
 2   status                  2928 non-null   float64
 3   life_expectancy         2928 non-null   float64
 4   adult_mortality         2928 non-null   float64
 5   infant_deaths           2928 non-null   float64
 6   alcohol                 2928 non-null   float64
 7   percentage_expenditure  2928 non-null   float64
 8   HepB                    2928 non-null   float64
 9   measles                 2928 non-null   float64
 10  BMI                     2928 non-null   float64
 11  u5_deaths               2928 non-null   float64
 12  Polio                   2928 non-null   float64
 13  total_expenditure       2928 non-null   float64
 14  DPT                     2928 non-null   float64
 15  HIV_AIDS                2928 non-null   float64
 16  GDP                     2928 non-null   float64
 17  population              2928 non-null   float64
 18  thinness_10_19          2928 non-null   float64
 19  thinness_5_9            2928 non-null   float64
 20  HDI                     2928 non-null   float64
 21  school_year             2928 non-null   float64
dtypes: float64(22)
memory usage: 503.4 KB
```

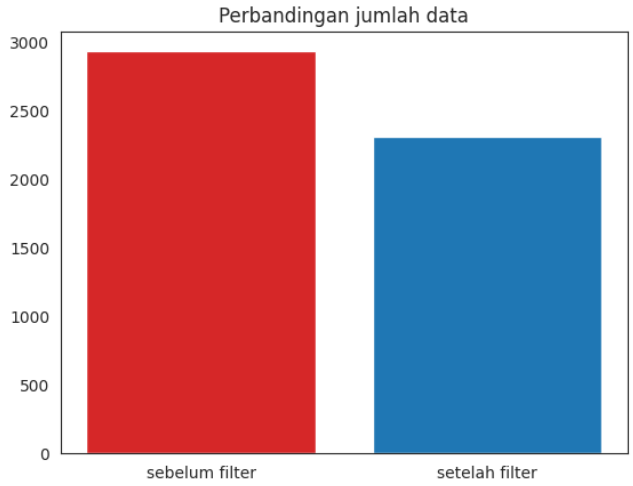
**Gambar 7 Deskripsi Data**

setelah semua kolom memiliki satu tipe data yang sama, selanjutnya dilakukan pengecekan korelasi sehingga dapat ditentukan mana saja kolom yang akan dipakai untuk training model. adapun tingkat hubungan yang digunakan peneliti yaitu dari tingkat hubungan sedang sampai sangat kuat. sehingga kolom yang memiliki tingkat hubungan sangat rendah dan rendah akan dihapus. Seperti yang dijelaskan pada sub bab sebelumnya, untuk memudahkan peneliti dalam menentukan tingkat hubungan antara fitur terhadap target, peneliti menggunakan *function corr* milik pandas dan divisualisasikan pada gambar 8.


**Gambar 8 Visualisasi Korelasi**

Dari gambar 8 selanjutnya dilakukan pemilihan fitur atau biasa disebut dengan *feature selection* dimana kolom yang tingkat hubungannya sangat rendah (interval 0,00-1,99) dan rendah (interval 0,20-0,399) dihapus. Maka didapatkan 6 kolom yaitu 'year','population','measles','percentage_expenditure','infant_deaths','u5_deaths'. Sehingga jumlah kolom yang digunakan untuk training model terdapat 16 kolom.

Tahapan selanjutnya yaitu filter *outliers* dengan menggunakan zscore seperti yang telah dijelaskan sebelumnya. Hasil dari filter *outliers* ini adalah jumlah data berkurang sekitar 600an baris. Adapun visualisasi dari perbedaan jumlah data sebelum dan sesudah dilakukan filter outliers adalah sebagai berikut.




**Gambar 9 Perbandingan Jumlah Data**

**B. *Processing* dan Evaluasi**

Pada bab sebelumnya telah dijelaskan bahwa peneliti menggunakan *grid search cross validation* untuk mencari *hyperparameter* terbaik. Adapun hasil yang didapatkan adalah sebagai berikut.

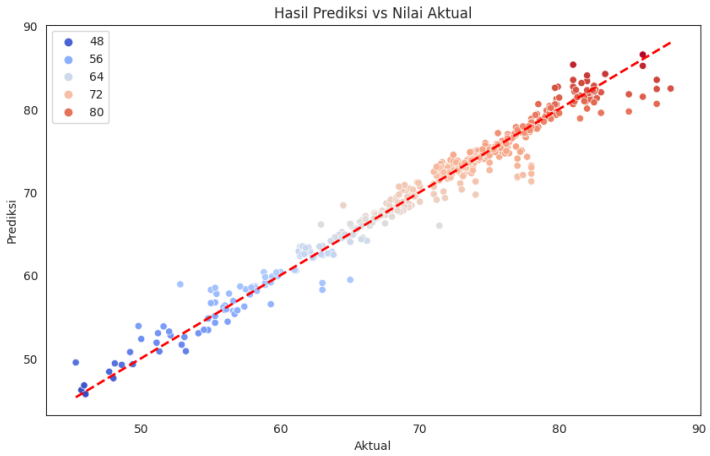
**Table 4 Hasil Hyperparameter Tuning**

| Hyperparameter | Value |
|---|---|
| Eta | 0.3 |
| Learning_rate | 0.1 |
| Max_depth | 12 |
| Min_child_weight | 8 |
| Reg_alpha | 1 |
| Reg_lambda | 0.1 |

Dari parameter terbaik yang dihasilkan oleh grid search cross validation yang terdapat pada tabel 4, selanjutnya dihasilkan skor train sebesar 99,6% dan skor test sebesar 96,8%. Selain itu, untuk skor *mean absolute error* yang dihasilkan adalah 0.97, jika skor MAE semakin mendekati 0 maka model yang telah dibangun semakin bagus.

**C. Uji Coba Model**

Gambar 10 merupakan visualisasi dari uji coba model yang telah dibangun. Garis lurus putus-putus (---) menunjukkan usia aktual sedangkan scatter plot menunjukkan prediksi usia. Dari gambar tersebut, model yang telah dibangun berhasil melakukan prediksi dengan baik dimana persebaran scatter plot mendekati garis.




**Gambar 10 Visualisasi Uji Coba Model**

## VII. SIMPULAN

Model yang dikembangkan pada penelitian ini berhasil memprediksi angka harapan hidup dengan tingkat akurasi mencapai 96,8%. Selain itu, hasil analisis menunjukkan bahwa nilai Mean Absolute Error (MAE) sebesar 0,97 menggambarkan rata-rata kesalahan prediksi model terhadap nilai sebenarnya. Temuan ini menunjukkan potensi model XGBoost Regressor sebagai alat prediksi yang efektif untuk mengestimasi angka harapan hidup di negara-negara wilayah Asia, dengan implikasi signifikan dalam pengambilan keputusan berbasis data dan perencanaan kebijakan kesejahteraan masyarakat.

## UCAPAN TERIMA KASIH

Ucapan terima kasih kami sampaikan kepada semua pihak yang telah memberikan dukungan berharga dalam pelaksanaan penelitian ini. Peneliti juga ingin mengakui peran UCI Machine Learning Repository atas penyediaan data yang digunakan dalam penelitian ini. Dukungan dari semua pihak ini telah memberikan sumbangan penting dalam upaya peneliti untuk mewujudkan penelitian ini.

## REFERENSI

[1] 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.

[2] *Cost and affordability of healthy diets across and within countries*. 2020. doi: 10.4060/cb2431en.

[3] 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.

[4] 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.

[5] C. Wang, C. Deng, and S. Wang, "Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost," *Pattern Recognit Lett*, vol. 136, 2020, doi: 10.1016/j.patrec.2020.05.035.

[6] S. P. Sinaga, A. Wanto, and S. Solikhun, "Implementasi Jaringan Syaraf Tiruan Resilient Backpropagation dalam Memprediksi Angka Harapan Hidup Masyarakat Sumatera Utara," *Jurnal Infomedia*, vol. 4, no. 2, 2020, doi: 10.30811/jim.v4i2.1573.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.*