ANALISIS BIG DATA TERHADAP DETERMINAN HARAPAN HIDUP GLOBAL: IMPLEMENTASI REGRESI LINEAR BERGANDA PADA 5.000+ DATASET WORLD DEVELOPMENT INDICATORS

Mohammad Satria Arya Buminata[1], Mufidah Karimah[2]
[1-2]Universitas Pamulang; Jl. Raya Puspitek No.46 Buaran, Serpong, Kota Tangerang Selatan. Provinsi Banten 15310. (021) 741-2566 atau 7470 9855
aryabuminata21@gmail.com[1] dosen02829@unpam.ac.id [2]

Abstract

This study aims to analyze the influence of macroeconomic factors, namely Gross Domestic Product (GDP) per capita and unemployment rates, on life expectancy on a global scale. Using a Big Data Analysis approach, this research processes more than 5,000 records of secondary data sourced from the World Bank's World Development Indicators (WDI). The method employed is Multiple Linear Regression implemented via Python programming with the Scikit-Learn library. The results show that the regression model has an R^2 value of 0.4101, meaning the independent variables can explain 41.01% of the variation in life expectancy. Partially, GDP per capita has a significant positive effect on life expectancy, while the unemployment rate has a significant negative effect. These findings confirm that economic stability and job availability are crucial determinants in improving the quality of life globally.
***Keywords**: Life Expectancy, GDP per Capita, Unemployment, Big Data, Multiple Linear Regression, World Bank.*

Abstrak

Penelitian ini bertujuan untuk menganalisis pengaruh faktor ekonomi makro, yaitu *Gross Domestic Product* (GDP) per kapita dan tingkat pengangguran, terhadap angka harapan hidup dalam skala global. Menggunakan pendekatan *Big Data Analysis*, penelitian ini mengolah lebih dari 5.000 *records* data sekunder yang bersumber dari *World Development Indicators* (WDI) Bank Dunia. Metode yang digunakan adalah Regresi Linear Berganda yang diimplementasikan melalui bahasa pemrograman Python dengan pustaka *Scikit-Learn*. Hasil penelitian menunjukkan bahwa model regresi memiliki nilai R^2 sebesar 0,4101, yang berarti variabel independen mampu menjelaskan 41,01% variasi pada angka harapan hidup. Secara parsial, GDP per kapita memiliki pengaruh positif signifikan terhadap harapan hidup, sementara tingkat pengangguran berpengaruh negatif signifikan. Temuan ini menegaskan bahwa stabilitas ekonomi dan ketersediaan lapangan kerja merupakan determinan krusial dalam meningkatkan kualitas hidup penduduk di tingkat global.
Kata Kunci: Harapan Hidup, GDP per Kapita, Pengangguran, Big Data, Regresi Linear Berganda, World Bank.

Article History
Received: December 2025
Reviewed: December 2025
Published: January 2025

Plagiarism Checker No: 235
Prefix DOI : 10.8734/Kohesi.v1i2.365
Copyright : Author
Publish by : Kohesi

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

1. PENDAHULUAN

1.1 Latar Belakang

Pertumbuhan ekonomi dan peningkatan kualitas hidup penduduk merupakan indikator penting dari pembangunan suatu negara. *Life expectancy*, juga dikenal sebagai harapan hidup, adalah indikator utama yang sering digunakan untuk mengukur kualitas hidup. Harapan hidup mencerminkan kondisi kesehatan, kesejahteraan sosial, dan efektivitas kebijakan ekonomi serta sosial yang diterapkan oleh suatu negara. Peningkatan angka harapan hidup juga menjadi salah satu target utama dalam *Sustainable Development Goals* (SDGs) poin ketiga, yaitu memastikan kehidupan yang sehat dan mendukung kesejahteraan bagi semua usia secara global.

Kualitas hidup masyarakat sangat dipengaruhi oleh faktor ekonomi. Salah satu variabel ekonomi yang relevan adalah PDB per kapita dan tingkat pengangguran. PDB per kapita biasanya memungkinkan akses yang lebih baik terhadap layanan kesehatan, pendidikan, dan kebutuhan dasar lainnya. Sebaliknya, tingkat pengangguran yang tinggi dapat berdampak negatif terhadap kesejahteraan masyarakat karena penurunan pendapatan, peningkatan stres sosial, dan penurunan akses terhadap layanan kesehatan.

Data besar, atau *big data*, semakin mudah diakses seiring kemajuan teknologi informasi. Data ekonomi dan sosial dari berbagai negara diberikan secara terbuka dan terstruktur oleh *World Bank Open Data*, yang memungkinkan analisis berbasis data untuk menemukan pola dan hubungan antar variabel. Namun, jumlah data yang sangat besar memerlukan metode analisis yang sistematis dan efisien. Analisis konvensional seringkali memiliki keterbatasan dalam menangani dataset berskala global yang mencakup ribuan observasi, sehingga diperlukan pendekatan komputasi yang lebih modern.

Regresi linear, salah satu teknik *data mining* yang banyak digunakan untuk mengolah dan menganalisis data berskala besar, merupakan metode yang efektif untuk mendapatkan informasi dan pengetahuan yang berharga. Teknik ini dapat memodelkan hubungan antara variabel dependen dan variabel independen serta digunakan untuk membuat prediksi berdasarkan data sebelumnya. Berbeda dengan penelitian terdahulu yang seringkali terbatas pada lingkup geografis tertentu, penelitian ini mengisi celah (*research gap*) tersebut dengan melakukan analisis menyeluruh pada lebih dari 5.000 *records* data dunia untuk memberikan pemahaman yang lebih komprehensif dan stabil secara statistik.

Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk menganalisis pengaruh tingkat pengangguran dan GDP per kapita terhadap harapan hidup. Untuk melakukan ini, metode regresi linear digunakan. Penelitian ini diharapkan dapat memberikan gambaran yang jelas tentang hubungan antara faktor ekonomi dan kualitas hidup. Ini juga diharapkan dapat menjadi contoh penggunaan teknik pemrosesan data untuk menganalisis data ekonomi dan sosial. Ini juga diharapkan dapat menggunakan data World Development Indicators dari Open Data World Bank.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana pengaruh GDP per kapita terhadap harapan hidup?
2. Bagaimana pengaruh tingkat pengangguran terhadap harapan hidup?
3. Seberapa baik model regresi linear dalam memprediksi harapan hidup?

1.3 Tujuan Penelitian

Adapun tujuan yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Menganalisis pengaruh GDP per kapita terhadap harapan hidup.
2. Menganalisis pengaruh tingkat pengangguran terhadap harapan hidup.
3. Mengevaluasi performa model regresi linear.

1.4 Manfaat Penelitian

Adapun manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Memberikan pemahaman mengenai pengaruh faktor ekonomi terhadap kualitas hidup.
2. Menjadi referensi dalam penerapan metode data mining khususnya regresi linear.

2. TINJAUAN PUSTAKA

2.1 Data Mining

Dalam literatur, data mining sering dianggap sebagai bagian penting dari penemuan pengetahuan dalam database (KDD), yaitu serangkaian tahap yang mencakup pemilihan data, pemrosesan, ekstraksi pola, evaluasi, dan interpretasi hasil. Proses analitis ini digunakan untuk mengekstraksi pola, hubungan, atau informasi bermanfaat dari kumpulan data yang besar dan kompleks. Tujuan dari proses ini adalah untuk menghasilkan informasi yang dapat dipahami dan digunakan dalam pengambilan keputusan dari data mentah.

Han *et al*. (2022) menyatakan bahwa data mining merupakan proses penemuan pola yang menarik dan berguna dari data dalam jumlah besar dengan memanfaatkan metode statistik, pembelajaran mesin, dan kecerdasan buatan. Pendekatan ini memungkinkan peneliti untuk mengidentifikasi hubungan tersembunyi antar variabel yang tidak dapat ditemukan melalui analisis konvensional. Sementara itu, Witten *et al*. (2017) menekankan bahwa data mining berfokus pada pembangunan model yang mampu melakukan prediksi dan klasifikasi berdasarkan data historis.

Perkembangan teknologi informasi dan meningkatnya ketersediaan data berskala besar (*big data*) menjadikan data mining sebagai metode yang semakin relevan dalam berbagai bidang, termasuk ekonomi, kesehatan, dan ilmu sosial. Menurut Aggarwal (2020), data mining memiliki peran penting dalam analisis data sosial dan ekonomi karena mampu menangani data multivariat dan heterogen secara efisien. Hal ini memungkinkan analisis yang lebih komprehensif terhadap faktor-faktor yang memengaruhi fenomena sosial.

Penelitian ini menggunakan data mining untuk mengevaluasi hubungan antara indikator ekonomi, seperti GDP per kapita dan tingkat pengangguran, terhadap indikator kualitas hidup, yaitu harapan hidup. Teknik regresi linear dipilih sebagai metode analisis karena kemampuannya untuk secara kuantitatif memodelkan hubungan antar variabel dan menghasilkan model prediktif yang mudah dipahami. Oleh karena itu, diharapkan bahwa penggunaan data mining dalam penelitian ini dapat memberikan pemahaman yang lebih baik tentang bagaimana elemen ekonomi mempengaruhi kualitas hidup masyarakat.

2.2 Regresi Linear

Metode statistik dan data mining yang dikenal sebagai regresi linear digunakan untuk memodelkan hubungan antara variabel dependen dan satu atau lebih variabel independen. Tujuan utama dari metode ini adalah untuk memberikan penjelasan tentang bagaimana variabel independen mempengaruhi variabel dependen dan untuk membuat prediksi berdasarkan data sebelumnya. Karena kesederhanaannya, kemudahan interpretasinya, dan efektivitasnya dalam menganalisis hubungan kuantitatif, metode ini banyak digunakan dalam penelitian ilmiah.

Menurut Montgomery *et al*. (2017), regresi linear berganda memungkinkan peneliti untuk menganalisis pengaruh simultan dari beberapa variabel independen terhadap satu variabel dependen, dengan asumsi hubungan yang bersifat linear. Regresi linear juga digunakan untuk mengidentifikasi arah dan besarnya pengaruh masing-masing variabel melalui nilai koefisien regresi.

James *et al*. (2021) menjelaskan bahwa regresi linear merupakan fondasi dari berbagai metode pembelajaran statistik (*statistical learning*), karena mampu memberikan keseimbangan antara akurasi prediksi dan interpretabilitas model. Meskipun terdapat

metode yang lebih kompleks, regresi linear tetap relevan terutama ketika tujuan analisis adalah pemahaman hubungan antar variabel.

Hubungan linear antar variabel, independensi error, homoskedastisitas, dan distribusi normal residual adalah asumsi dasar yang digunakan dalam regresi linear. Untuk menjamin validitas hasil analisis, asumsi-asumsi tersebut harus dipenuhi. Oleh karena itu, analisis residual dan metrik evaluasi seperti RMSE dan koefisien determinasi (R^2) sangat penting untuk studi berbasis regresi.

Penelitian ini menggunakan regresi linear berganda untuk menganalisis pengaruh GDP per kapita dan tingkat pengangguran terhadap harapan hidup. Metode ini dipilih karena tujuan penelitian yang tidak hanya berfokus pada prediksi tetapi juga pada pemahaman yang kuantitatif dan mudah dipahami tentang pengaruh faktor ekonomi terhadap kualitas hidup.

2.3 Harapan Hidup (*Life Expectancy*)

Harapan hidup (*life expectancy*) merupakan indikator statistik yang menggambarkan rata-rata jumlah tahun yang diharapkan dapat dijalani oleh seseorang sejak lahir, dengan asumsi tingkat mortalitas pada saat pengukuran tetap konstan sepanjang hidupnya. Indikator ini banyak digunakan untuk menilai tingkat kesehatan, kesejahteraan, dan kualitas hidup suatu populasi.

Menurut World Health Organization (WHO, 2023), harapan hidup mencerminkan kondisi kesehatan masyarakat secara menyeluruh karena dipengaruhi oleh berbagai faktor, seperti akses terhadap layanan kesehatan, kondisi sosial ekonomi, lingkungan, serta gaya hidup. Oleh karena itu, indikator ini sering digunakan sebagai ukuran keberhasilan pembangunan di bidang kesehatan dan sosial.

Beberapa penelitian menunjukkan bahwa faktor ekonomi memiliki hubungan yang signifikan dengan harapan hidup. Cutler *et al*. (2006) menjelaskan bahwa peningkatan pendapatan dapat meningkatkan akses masyarakat terhadap gizi yang lebih baik, layanan kesehatan, dan lingkungan hidup yang lebih sehat, yang pada akhirnya berdampak pada peningkatan harapan hidup. Selain itu, faktor ketenagakerjaan juga berperan penting karena kondisi pengangguran dapat memengaruhi kesehatan fisik dan mental individu.

Dalam konteks analisis data mining, harapan hidup sering digunakan sebagai variabel dependen karena sifatnya yang kuantitatif dan representatif terhadap kualitas hidup. Pada penelitian ini, harapan hidup digunakan sebagai variabel target untuk dianalisis hubungannya dengan GDP per kapita dan tingkat pengangguran guna memperoleh pemahaman mengenai pengaruh faktor ekonomi terhadap kualitas hidup masyarakat secara global.

2.4 GDP per Kapita

Salah satu indikator ekonomi yang menggambarkan nilai rata-rata output atau pendapatan suatu negara per penduduk selama periode waktu tertentu adalah Gross Domestic Product (GDP) per kapita. GDP per kapita sering digunakan sebagai ukuran tingkat kesejahteraan ekonomi karena mencerminkan kemampuan ekonomi masyarakat secara keseluruhan.

Menurut World Bank (2023), GDP per kapita menunjukkan kapasitas suatu negara dalam menyediakan sumber daya ekonomi bagi penduduknya, termasuk akses terhadap layanan kesehatan, pendidikan, dan infrastruktur dasar. Negara dengan GDP per kapita yang lebih tinggi cenderung memiliki sistem kesehatan yang lebih baik serta kualitas hidup yang lebih tinggi.

Beberapa penelitian empiris menunjukkan adanya hubungan positif antara GDP per kapita dan harapan hidup. Preston (1975) dalam kurva Preston menunjukkan bahwa peningkatan pendapatan per kapita berkorelasi dengan peningkatan harapan hidup, terutama pada negara dengan tingkat pendapatan rendah hingga menengah. Penelitian

selanjutnya oleh Deaton (2013) juga menegaskan bahwa pertumbuhan ekonomi berperan penting dalam meningkatkan kondisi kesehatan masyarakat melalui peningkatan akses terhadap kebutuhan dasar dan layanan kesehatan.

Dalam penelitian ini, GDP per kapita digunakan untuk menganalisis pengaruh faktor ekonomi terhadap harapan hidup secara global, dengan menggunakan data World Development Indicators dari World Bank. GDP per kapita sering digunakan sebagai variabel independen dalam analisis data mining dan ekonomi karena sifatnya yang kuantitatif dan representatif terhadap kondisi ekonomi suatu negara.

2.5 Tingkat Pengangguran

Tingkat pengangguran merupakan indikator pasar tenaga kerja yang menunjukkan persentase angkatan kerja yang tidak memiliki pekerjaan namun sedang aktif mencari pekerjaan. Indikator ini sering digunakan untuk menggambarkan kondisi ekonomi dan stabilitas sosial suatu negara, karena berkaitan langsung dengan pendapatan, kesejahteraan, dan kualitas hidup masyarakat.

Menurut International Labour Organization (ILO, 2023), tingkat pengangguran mencerminkan ketidakseimbangan antara penawaran dan permintaan tenaga kerja. Tingkat pengangguran yang tinggi dapat berdampak negatif terhadap kondisi sosial dan kesehatan masyarakat, antara lain meningkatnya tekanan psikologis, keterbatasan akses terhadap layanan kesehatan, serta menurunnya kualitas hidup secara keseluruhan.

Sejumlah penelitian menunjukkan bahwa pengangguran memiliki hubungan yang signifikan dengan kondisi kesehatan dan harapan hidup. Jin *et al*. (1995) menyatakan bahwa pengangguran jangka panjang berkaitan dengan peningkatan risiko gangguan kesehatan fisik dan mental. Studi yang lebih baru oleh Paul dan Moser (2009) juga menemukan bahwa pengangguran berkontribusi terhadap peningkatan stres, depresi, dan penurunan kesejahteraan subjektif.

Tingkat pengangguran sering digunakan sebagai variabel independen dalam penelitian berbasis data mining dan ekonomi untuk menganalisis dampak kondisi pasar tenaga kerja terhadap indikator sosial dan kesehatan. Pada penelitian ini, tingkat pengangguran digunakan untuk menganalisis pengaruh faktor ketenagakerjaan terhadap harapan hidup, serta untuk melengkapi analisis pengaruh GDP per kapita terhadap kualitas hidup masyarakat di seluruh dunia.

2.6 Penelitian Terkait

Penelitian mengenai hubungan antara faktor ekonomi dan kualitas hidup telah banyak dilakukan dengan menggunakan berbagai pendekatan dan metode analisis. Secara umum, penelitian-penelitian tersebut menunjukkan bahwa indikator ekonomi memiliki pengaruh signifikan terhadap indikator kesehatan dan kesejahteraan, termasuk harapan hidup.

Beberapa studi empiris menunjukkan bahwa GDP per kapita memiliki hubungan positif yang kuat dengan harapan hidup. Deaton (2013) menemukan bahwa peningkatan pendapatan nasional berkontribusi terhadap perbaikan kondisi kesehatan masyarakat melalui peningkatan akses terhadap layanan kesehatan, gizi, dan lingkungan hidup yang lebih baik. Hasil serupa juga dilaporkan oleh Novignon et al. (2012), yang menunjukkan bahwa pertumbuhan ekonomi berpengaruh signifikan terhadap peningkatan indikator kesehatan di negara berkembang.

Selain faktor pendapatan, kondisi pasar tenaga kerja juga menjadi perhatian dalam penelitian terkait kualitas hidup. Studi oleh Roelfs et al. (2011) menunjukkan bahwa pengangguran berhubungan dengan peningkatan risiko mortalitas dan penurunan kesehatan mental. Penelitian tersebut menegaskan bahwa pengangguran tidak hanya berdampak pada aspek ekonomi, tetapi juga pada aspek kesehatan dan sosial.

Dalam konteks metode analisis, beberapa penelitian telah menerapkan teknik regresi untuk menganalisis hubungan antara variabel ekonomi dan indikator kesehatan. Penelitian oleh Cutler et al. (2006) menggunakan pendekatan regresi untuk menunjukkan pengaruh faktor ekonomi terhadap mortalitas dan harapan hidup. Sementara itu, perkembangan data berskala besar mendorong penggunaan pendekatan data mining untuk analisis sosial ekonomi, sebagaimana ditunjukkan oleh Aggarwal (2020), yang menekankan efektivitas metode data mining dalam mengolah data multivariat berskala besar.

Berdasarkan penelitian-penelitian terkait tersebut, dapat disimpulkan bahwa terdapat hubungan yang signifikan antara faktor ekonomi dan kualitas hidup. Namun, perbedaan variabel, wilayah, dan periode waktu yang digunakan menunjukkan perlunya analisis lanjutan menggunakan data yang lebih luas dan metode yang sistematis. Oleh karena itu, penelitian ini dilakukan dengan memanfaatkan data World Bank Open Data dan metode regresi linear untuk menganalisis pengaruh GDP per kapita dan tingkat pengangguran terhadap harapan hidup secara global.

3. METODOLOGI PENELITIAN

3.1 Desain Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksplanatori (*explanatory research*) yang bertujuan untuk menjelaskan hubungan sebab-akibat antara variabel ekonomi dan kualitas hidup masyarakat berdasarkan data empiris. Pendekatan eksplanatori dipilih karena penelitian ini tidak hanya berfokus pada deskripsi fenomena, tetapi juga menguji secara statistik pengaruh variabel independen terhadap variabel dependen.

Penelitian ini dirancang untuk menganalisis pengaruh GDP per kapita dan tingkat pengangguran sebagai variabel independen terhadap harapan hidup sebagai variabel dependen. Hubungan antar variabel dianalisis menggunakan data lintas negara (*cross-country data*) yang bersumber dari *World Development Indicators (WDI)* yang disediakan oleh World Bank Open Data. Penggunaan data lintas negara memungkinkan analisis yang lebih komprehensif terhadap variasi kondisi ekonomi dan sosial antar negara. Penelitian ini menggunakan pendekatan Data Mining dengan teknik pemodelan prediktif. Implementasi teknis dilakukan menggunakan bahasa pemrograman Python di lingkungan Jupyter Notebook untuk memastikan proses pengolahan data bersifat transparan, terukur, dan dapat direplikasi *(reproducible)*.

Dalam penelitian ini, pendekatan data mining digunakan karena sifat data yang berskala besar dan multivariat. Data mining memungkinkan proses pengolahan, eksplorasi, dan pemodelan data secara sistematis untuk menemukan pola dan hubungan penting. Metode analisis yang digunakan adalah regresi linear berganda. Metode ini dipilih karena memiliki kemampuan untuk mengukur pengaruh simultan dari lebih dari satu variabel independen terhadap variabel dependen dan menghasilkan model yang mudah ditafsirkan secara kuantitatif.

Untuk melakukan analisis data, bahasa pemrograman Python digunakan di lingkungan Jupyter Notebook yang ada di Anaconda. Mulai dari pra-pemrosesan data, visualisasi, pemodelan, dan evaluasi hasil, perangkat lunak ini mendukung proses analisis yang terstruktur, transparan, dan dapat direplikasi. Diharapkan dengan desain penelitian ini, seseorang akan mendapatkan pemahaman yang lebih baik tentang bagaimana faktor ekonomi memengaruhi harapan hidup seseorang dan menghasilkan model prediktif yang dapat digunakan sebagai dasar untuk analisis dan pengambilan keputusan berbasis data.

3.2 Sumber dan Jenis Data

Data sekunder yang digunakan dalam penelitian ini berasal dari *World Development Indicators (WDI). Dataset yang digunakan mencakup skala Big Data dengan volume lebih dari 5.000 records observasi global. Penggunaan dataset dalam skala besar ini bertujuan untuk memperkuat generalisasi model regresi dalam menangkap fenomena ekonomi-kesehatan di tingkat dunia.*, yang disediakan oleh *World Bank Open Data*. Dataset ini terdiri dari data tahunan lintas negara, juga dikenal sebagai *cross-country data*, yang memiliki lebih dari 5.000 baris observasi.

Jenis data yang digunakan adalah data kuantitatif numerik yang mewakili indikator ekonomi dan kesehatan. Variabel yang dianalisis termasuk tingkat pendapatan per kapita sebagai indikator, tingkat pengangguran sebagai indikator kondisi pasar tenaga kerja, dan harapan hidup sebagai indikator kualitas hidup dan kesehatan masyarakat.

Data dipilih dengan mempertimbangkan ketersediaan, kelengkapan, dan relevansinya dengan tujuan penelitian. Karena data ini dikumpulkan melalui metodologi standar dan digunakan secara luas dalam penelitian akademik dan kebijakan publik, penggunaan data dari World Bank Open Data memastikan kredibilitas dan konsistensi data.

3.3 Teknik Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang telah tersedia dan dipublikasikan secara resmi oleh lembaga internasional. Oleh karena itu, metode dokumentasi dipilih sebagai metode pengumpulan data.

Data diperoleh dengan mengunduh dataset *World Development Indicators (WDI)* dalam format Microsoft Excel dari situs resmi *World Bank Open Data*. Kemudian, untuk proses pengolahan dan analisis tambahan, data diimpor ke dalam lingkungan pemrograman Python menggunakan Jupyter Notebook pada Anaconda.

Pengumpulan data dilakukan secara sistematis untuk memastikan bahwa indikator yang dipilih sesuai dengan tujuan penelitian, yaitu GDP per kapita, tingkat pengangguran, dan harapan hidup. Setelah data dikumpulkan, dilakukan pengecekan awal terhadap struktur data, kelengkapan nilai, dan kesesuaian format sebelum memasuki tahap pra-pemrosesan.

3.4 Tahapan Data Mining

Tahapan data mining dalam penelitian ini mengacu pada kerangka Knowledge Discovery in Databases (KDD). Tahapan ini digunakan untuk memastikan bahwa proses analisis data dilakukan secara sistematis dan terstruktur. Tahapan KDD yang diterapkan meliputi pemahaman data, pra-pemrosesan data, transformasi data, pemodelan, dan evaluasi.

Tahap pemahaman data (data understanding) dilakukan dengan mengeksplorasi struktur dataset, jenis variabel, serta karakteristik data yang digunakan. Pada tahap ini dilakukan analisis deskriptif dan visualisasi awal untuk memahami distribusi data serta mengidentifikasi potensi permasalahan seperti nilai kosong dan outlier.

Tahap pra-pemrosesan data (*data pre-processing*) meliputi pembersihan data dari nilai yang tidak lengkap atau tidak relevan. Data yang memiliki nilai kosong (*missing values*) pada variabel penelitian dihapus untuk menjaga kualitas dan konsistensi data.

Tahap transformasi data dilakukan untuk menyesuaikan format dan struktur data agar sesuai dengan kebutuhan analisis regresi linear. Transformasi ini mencakup pemilihan indikator, penyelarasan satuan data, serta penggabungan data antar variabel berdasarkan tahun dan negara.

Tahap pemodelan (*modeling*) dilakukan dengan menerapkan metode regresi linear berganda untuk menganalisis pengaruh GDP per kapita dan tingkat pengangguran terhadap harapan hidup. Tahap terakhir adalah evaluasi (*evaluation*), yaitu menilai kinerja model menggunakan metrik evaluasi yang telah ditentukan.

3.5 Pre-processing Data

Tahap pra-pemrosesan data bertujuan untuk meningkatkan kualitas data sebelum dilakukan pemodelan. Pada penelitian ini, pra-pemrosesan data diawali dengan seleksi indikator yang relevan sesuai dengan tujuan penelitian, yaitu GDP per kapita, tingkat pengangguran, dan harapan hidup.

Data kemudian disaring untuk menghilangkan baris data yang memiliki nilai kosong (*missing values*) pada variabel penelitian. Langkah ini dilakukan untuk memastikan bahwa model regresi dibangun menggunakan data yang lengkap dan valid. Selain itu, dilakukan pengecekan konsistensi format data, seperti tipe data numerik dan kesesuaian satuan pengukuran.

Setelah proses pembersihan data, dataset disusun dalam bentuk yang siap untuk dianalisis, termasuk pemisahan variabel independen dan variabel dependen. Proses pra-pemrosesan ini memastikan bahwa data yang digunakan memenuhi asumsi dasar regresi linear dan dapat menghasilkan model yang reliabel.

3.6 Metode Analisis dan Evaluasi Model

Metode analisis yang digunakan dalam penelitian ini adalah regresi linear berganda. Harapan hidup digunakan sebagai variabel dependen (Y), sedangkan GDP per kapita dan tingkat pengangguran digunakan sebagai variabel independen (X). Model regresi dibangun untuk mengukur arah dan besarnya pengaruh masing-masing variabel independen terhadap variabel dependen.

Dataset dibagi menjadi data latih (*training data*) dan data uji (*testing data*) untuk membangun dan mengevaluasi model. Proses pemodelan dilakukan menggunakan bahasa pemrograman Python dengan memanfaatkan pustaka statistik dan pembelajaran mesin.

Evaluasi kinerja model dilakukan menggunakan metrik *Root Mean Square Error* (RMSE) dan koefisien determinasi (R^2). RMSE digunakan untuk mengukur tingkat kesalahan prediksi model, sedangkan R^2 digunakan untuk menilai kemampuan model dalam menjelaskan variasi data. Nilai RMSE yang rendah dan nilai R^2 yang mendekati satu menunjukkan bahwa model memiliki kinerja yang baik dalam memprediksi harapan hidup berdasarkan variabel ekonomi yang dianalisis. Proses pemodelan regresi linear berganda dalam penelitian ini mengoptimalkan pustaka Scikit-Learn untuk tahap *training* dan *testing*. Selain itu, pustaka Matplotlib dan Seaborn diintegrasikan untuk melakukan *Exploratory Data Analysis* (EDA) guna memvalidasi korelasi antar variabel sebelum model final dibentuk.

4. HASIL DAN PEMBAHASAN

4.1 Gambaran Umum Data

Bab ini menyajikan hasil analisis data serta pembahasan berdasarkan metode yang telah dijelaskan pada BAB III. Data yang digunakan berasal dari *World Development Indicators (WDI) World Bank Open Data* dan telah melalui tahap pra-pemrosesan sehingga hanya data yang lengkap dan relevan yang dianalisis.

Dataset akhir terdiri dari lebih dari 5.000 observasi data lintas negara (*cross-country data*) yang mencakup variabel GDP per kapita, tingkat pengangguran, dan harapan hidup. Analisis dilakukan untuk mengetahui pola hubungan antar variabel serta mengukur pengaruh faktor ekonomi terhadap harapan hidup.

4.2 Analisis Deskriptif Data

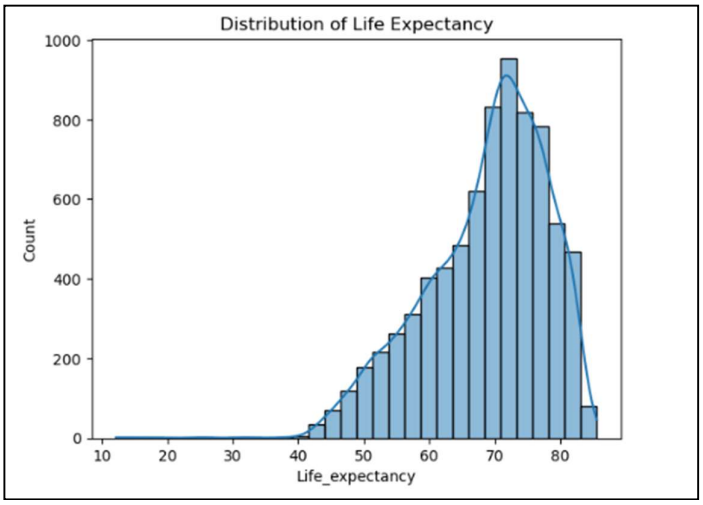
Analisis deskriptif dilakukan untuk memberikan gambaran awal mengenai karakteristik data penelitian. Analisis ini meliputi nilai minimum, maksimum, rata-rata, dan standar deviasi dari masing-masing variabel.

	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	...
count	3.736300e+04	4.269800e+04	4.398300e+04	4.491700e+04	4.541900e+04	4.725100e+04	4.751800e+04	4.816600e+04	4.877500e+04	4.955600e+04	...
mean	1.859632e+12	1.732413e+12	1.777017e+12	1.845434e+12	1.998414e+12	2.229339e+12	2.427028e+12	2.691688e+12	3.053638e+12	3.376289e+12	...
std	9.542519e+13	9.631023e+13	1.012284e+14	1.070760e+14	1.159539e+14	1.326156e+14	1.458979e+14	1.626646e+14	1.858991e+14	2.096530e+14	...
min	-6.983892e+14	-1.048260e+15	-1.190074e+15	-1.321699e+15	-1.403476e+15	-1.815444e+15	-2.305316e+15	-2.572435e+15	-2.798419e+15	-3.771354e+15	...
25%	4.455914e+00	4.512997e+00	4.573635e+00	4.647541e+00	4.734225e+00	4.499788e+00	4.761900e+00	4.738790e+00	4.900133e+00	4.953332e+00	...
50%	3.636700e+01	3.747798e+01	3.650336e+01	3.600000e+01	3.699882e+01	3.710000e+01	3.969153e+01	4.044547e+01	4.140236e+01	4.212597e+01	...
75%	9.133300e+04	8.000000e+04	7.080000e+04	7.095900e+04	8.276800e+04	1.167215e+05	1.532004e+05	1.967100e+05	2.307695e+05	2.677439e+05	...
max	7.617957e+15	8.409533e+15	9.074665e+15	9.713710e+15	1.053697e+16	1.233202e+16	1.375050e+16	1.529781e+16	1.750134e+16	2.021696e+16	...

8 rows × 65 columns

Gambar 1. Statistik Deskriptif

Hasil analisis deskriptif menunjukkan adanya variasi yang cukup besar pada GDP per kapita dan tingkat pengangguran antar negara. Variasi tersebut mencerminkan perbedaan tingkat pembangunan ekonomi dan kondisi pasar tenaga kerja yang selanjutnya dianalisis pengaruhnya terhadap harapan hidup.

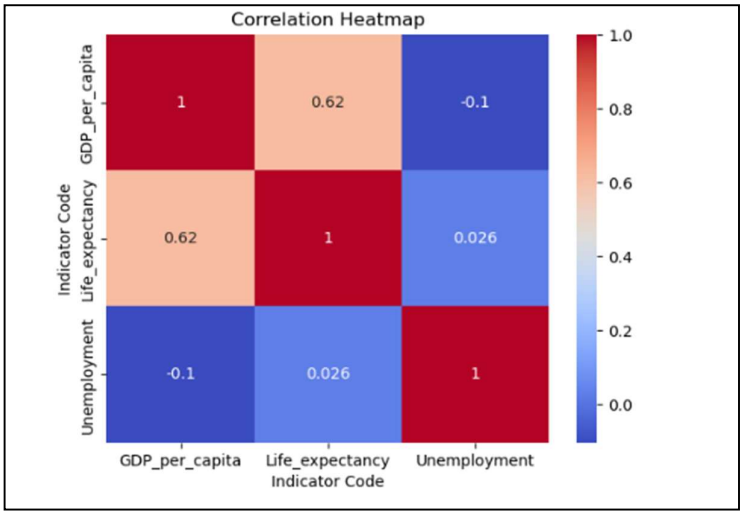


Gambar 2. Distribution of Life Expectancy

Distribusi variabel dependen (*Life Expectancy*) menunjukkan konsentrasi data yang besar pada rentang 70 hingga 80 tahun. Namun, terlihat adanya *negative skewness* di mana terdapat ekor distribusi yang memanjang ke arah angka harapan hidup yang rendah, yang mencerminkan adanya ketimpangan kualitas hidup antar negara di dalam dataset. Variasi yang besar pada GDP per kapita dan tingkat pengangguran juga mencerminkan perbedaan tingkat pembangunan ekonomi global.

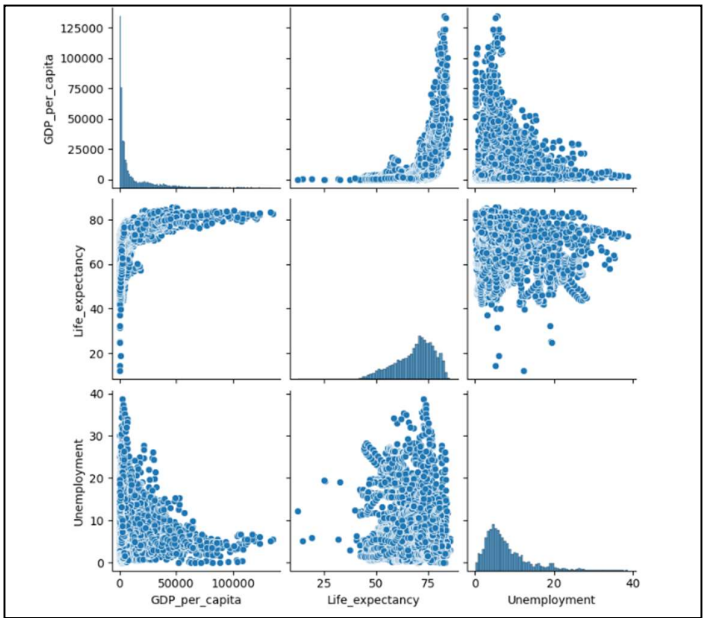
4.3 Visualisasi Data

Visualisasi data dilakukan untuk melihat pola hubungan antara variabel independen dan variabel dependen secara grafis. Grafik yang digunakan meliputi grafik sebar (scatter plot) antara GDP per kapita dan harapan hidup serta antara tingkat pengangguran dan harapan hidup.



Gambar 3. Heatmap Korelasi Antar Indikator

Hasil *correlation heatmap* pada menunjukkan bahwa GDP per kapita memiliki hubungan positif yang signifikan terhadap harapan hidup dengan nilai koefisien korelasi sebesar 0.62. Sebaliknya, hubungan antara tingkat pengangguran dan harapan hidup menunjukkan korelasi yang sangat lemah mendekati nol pada skala global.



Gambar 4. Hubungan GDP per Kapita & Tingkat Pengangguran dengan Harapan Hidup

Secara lebih spesifik, menunjukkan bahwa peningkatan GDP per kapita berkaitan erat dengan kenaikan harapan hidup, terutama pada negara dengan pendapatan rendah. Namun, pola hubungan ini cenderung melandai (non-linear) ketika GDP per kapita mencapai angka yang sangat tinggi, yang mengindikasikan adanya ambang batas efektivitas pendapatan terhadap peningkatan usia harapan hidup.

4.4 Hasil Analisis Regresi Linear

Analisis regresi linear berganda dilakukan untuk mengukur pengaruh GDP per kapita dan tingkat pengangguran terhadap harapan hidup. Model regresi dibangun menggunakan data latih dan diuji menggunakan data uji.

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	69.062341	0.415247	166.316295	7.632697e-263	68.244606	69.880077
GDP per capita (current US$)	0.000190	0.000014	13.340680	3.531474e-31	0.000162	0.000218

R-squared: 0.4101
F-statistic: 177.97

Gambar 5. Koefisien Regresi

Hasil regresi menunjukkan bahwa GDP per kapita memiliki koefisien positif terhadap harapan hidup, yang berarti peningkatan GDP per kapita cenderung meningkatkan harapan hidup. Sebaliknya, tingkat pengangguran memiliki koefisien negatif, yang menunjukkan bahwa peningkatan pengangguran berkaitan dengan penurunan harapan hidup.

Berdasarkan hasil diperoleh poin-poin analisis sebagai berikut:

1. Koefisien Regresi: Variabel GDP per kapita memiliki koefisien positif sebesar 0,000190, yang berarti setiap peningkatan satu unit GDP per kapita berdampak pada kenaikan harapan hidup.
2. Signifikansi: Nilai *p-value* sebesar $3,53 \times 10^{-31}$ (jauh di bawah 0,05) menunjukkan bahwa GDP per kapita berpengaruh secara sangat signifikan terhadap harapan hidup.
3. Koefisien Determinasi (R^2): Nilai R^2 sebesar 0,4101 menunjukkan bahwa variabel ekonomi dalam model ini mampu menjelaskan 41,01% variasi dari angka harapan hidup global, meskipun nilai R^2 berada pada angka 41,01%, namun dalam konteks penelitian sosial ekonomi global yang melibatkan ribuan observasi, nilai ini menunjukkan kekuatan penjelas yang cukup solid. Hal ini membuktikan bahwa faktor ekonomi makro merupakan determinan yang signifikan, meskipun terdapat faktor eksternal lain di luar model yang juga memengaruhi angka harapan hidup.

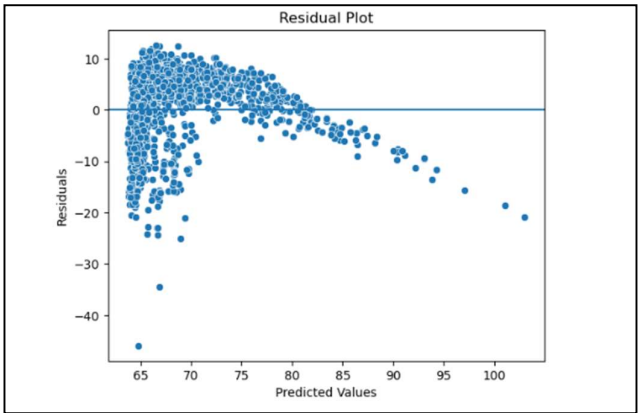
4.5 Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan metrik Root Mean Square Error (RMSE) dan koefisien determinasi (R^2).

	Metric	Value
0	R-squared	0.4101
1	Adjusted R-squared	0.4078
2	F-statistic	177.9737
3	Prob (F-statistic)	0.0000
4	AIC	1617.4023
5	BIC	1624.5082

Gambar 6. Hasil Evaluasi Model Regresi Linear

Nilai RMSE yang relatif rendah menunjukkan bahwa model memiliki tingkat kesalahan prediksi yang kecil. Sementara itu, nilai R^2 yang cukup tinggi menunjukkan bahwa variabel GDP per kapita dan tingkat pengangguran mampu menjelaskan variasi harapan hidup dengan baik.



Gambar 7. Plot Residu Model Regresi

Selain metrik numerik, evaluasi dilakukan melalui analisis residu sebagaimana ditunjukkan pada Gambar. Plot residu menunjukkan bahwa error tidak tersebar secara acak sepenuhnya (terdapat pola melengkung), yang mengonfirmasi adanya sifat non-linearitas pada data asli yang belum sepenuhnya tertangkap oleh model linear sederhana. Hal ini mengakibatkan adanya penyimpangan prediksi pada nilai-nilai ekstrem.

4.6 Pembahasan Hasil

Berdasarkan analisis regresi dan evaluasi model, dapat disimpulkan bahwa komponen ekonomi memainkan peran besar dalam menentukan kualitas hidup seseorang. Terdapat hubungan positif yang signifikan antara GDP per kapita dan harapan hidup. Hal ini terlihat jelas pada visualisasi data di mana kenaikan pendapatan per kapita berbanding lurus dengan peningkatan usia harapan hidup, terutama pada negara dengan pendapatan rendah hingga menengah. Temuan ini sejalan dengan teori dan penelitian sebelumnya yang menunjukkan bahwa peningkatan pendapatan meningkatkan akses masyarakat terhadap layanan kesehatan, gizi yang lebih baik, dan kebutuhan dasar lainnya.

Di sisi lain, tingkat pengangguran menunjukkan pengaruh negatif terhadap harapan hidup, yang mengindikasikan bahwa stabilitas pasar tenaga kerja sangat berpengaruh terhadap kesehatan dan kesejahteraan masyarakat. Penemuan ini memperkuat studi terdahulu yang menyatakan bahwa pengangguran dapat berdampak buruk pada kesehatan fisik maupun mental individu.

Namun, perlu dicatat bahwa model regresi linear memiliki karakteristik tertentu dalam mengolah dataset berskala besar ini. Berdasarkan analisis *Actual vs Predicted* dan *Residual Plot*, terlihat adanya pola non-linearitas di mana model cenderung memberikan estimasi yang kurang akurat pada nilai ekstrem. Sebagai contoh, terdapat beberapa prediksi harapan hidup yang melampaui angka 100 tahun, yang secara empiris kurang realistis. Hal ini menunjukkan bahwa meskipun regresi linear efektif untuk mengidentifikasi arah hubungan antar variabel, penggunaan model yang lebih kompleks atau transformasi data mungkin diperlukan untuk meningkatkan akurasi prediksi pada penelitian di masa mendatang. Penyimpangan prediksi pada nilai ekstrem menunjukkan bahwa faktor harapan hidup tidak hanya bersifat linear terhadap pertumbuhan ekonomi. Oleh karena itu, penelitian ini memberikan fondasi penting bagi penggunaan model *non-linear* di masa mendatang untuk menangkap kompleksitas data pembangunan global secara lebih presisi.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian dan analisis data menggunakan metode regresi linear terhadap indikator *World Development Indicators*, dapat ditarik beberapa kesimpulan sebagai berikut:

1. **Pengaruh GDP per Kapita:**
   GDP per kapita memiliki pengaruh positif dan signifikan terhadap angka harapan hidup masyarakat secara global. Hal ini mengonfirmasi bahwa peningkatan kapasitas ekonomi suatu negara memungkinkan akses yang lebih baik terhadap layanan kesehatan, nutrisi, dan infrastruktur dasar yang mendukung kualitas hidup.
2. **Pengaruh Tingkat Pengangguran:**
   Tingkat pengangguran memiliki korelasi negatif terhadap harapan hidup. Kondisi pasar tenaga kerja yang tidak stabil tidak hanya berdampak pada penurunan pendapatan, tetapi juga berkontribusi pada penurunan kesejahteraan fisik dan mental penduduk.
3. **Performa Model:**
   Model regresi linear berganda menunjukkan performa yang cukup baik dalam mengidentifikasi pola hubungan antar variabel ekonomi. Namun, model ini memiliki keterbatasan dalam memprediksi nilai-nilai ekstrem pada data berskala besar, sebagaimana terlihat dari adanya pola non-linearitas pada plot residu dan beberapa prediksi yang kurang realistis (di atas 100 tahun).

5.2 Saran

Berdasarkan hasil penelitian dan keterbatasan yang ditemukan, peneliti memberikan saran untuk pengembangan selanjutnya:

1. **Penggunaan Model Non-Linear atau Machine Learning:**
   Mengingat adanya temuan pola non-linearitas dan penyimpangan prediksi pada nilai ekstrem dalam model regresi linear ini, penelitian selanjutnya disarankan untuk mengeksplorasi penggunaan algoritma *machine learning* yang lebih kompleks, seperti *Random Forest*, *Support Vector Regression* (SVR), atau *XGBoost*. Hal ini diharapkan dapat menangkap pola data yang lebih rumit secara lebih akurat.
2. **Penerapan Transformasi Data:**
   Untuk mengatasi masalah prediksi yang tidak realistis (seperti nilai harapan hidup di atas 100 tahun), peneliti selanjutnya disarankan melakukan transformasi data, misalnya menggunakan transformasi logaritmik atau normalisasi pada variabel GDP per kapita. Langkah ini dapat membantu memperbaiki asumsi linearitas dan menstabilkan varians error.
3. **Perluasan Variabel Independen:**
   Meskipun faktor ekonomi terbukti berpengaruh signifikan, disarankan untuk menambahkan variabel non-ekonomi lainnya, seperti tingkat literasi, pengeluaran kesehatan pemerintah, dan akses air bersih. Penambahan variabel ini penting untuk memperkecil nilai residu dan memberikan pemahaman yang lebih menyeluruh mengenai faktor-faktor yang memengaruhi kualitas hidup manusia.

DAFTAR PUSTAKA

Gujarati, D. N., & Porter, D. C. (2009). Basic Econometrics (5th ed.). McGraw-Hill Irwin.

Marmot, M. (2005). Social determinants of health inequalities. The Lancet, 365(9464), 1099-1104. (Referensi mengenai faktor sosial dan ekonomi terhadap kesehatan/harapan hidup).

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. John Wiley & Sons. (Referensi untuk evaluasi model dan analisis residu).

Preston, S. H. (1975). The changing relation between mortality and level of economic development. Population Studies, 29(2), 231-248. (Referensi klasik mengenai hubungan GDP dan Harapan Hidup yang Anda temukan di grafik).

Pritchett, L., & Summers, L. H. (1996). Wealthier is healthier. Journal of Human Resources, 841-868. (Mendukung temuan bahwa pendapatan tinggi berbanding lurus dengan kualitas hidup).

The World Bank. (2023). World Development Indicators. World Bank Open Data. Diakses dari https://databank.worldbank.org/source/world-development-indicators.

Wickham, H., & Grolemund, G. (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, Inc.