UNIVERSITAS
MUHAMMADIYAH
SIDOARJO

UMSIDA
DARI SINI PENCERAHAN BERSEMI

Prediksi Angka Harapan Hidup Berdasarkan Faktor Sosioekonomi Dan Kesehatan Menggunakan Algoritma Random Forest Regressor

Oleh:

Hamzah Zufarul Furqon

Mochamad. Alfan Rosid

Progam Studi Informatika

Universitas Muhammadiyah Sidoarjo

Juli, 2024

Pendahuluan

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.

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.

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

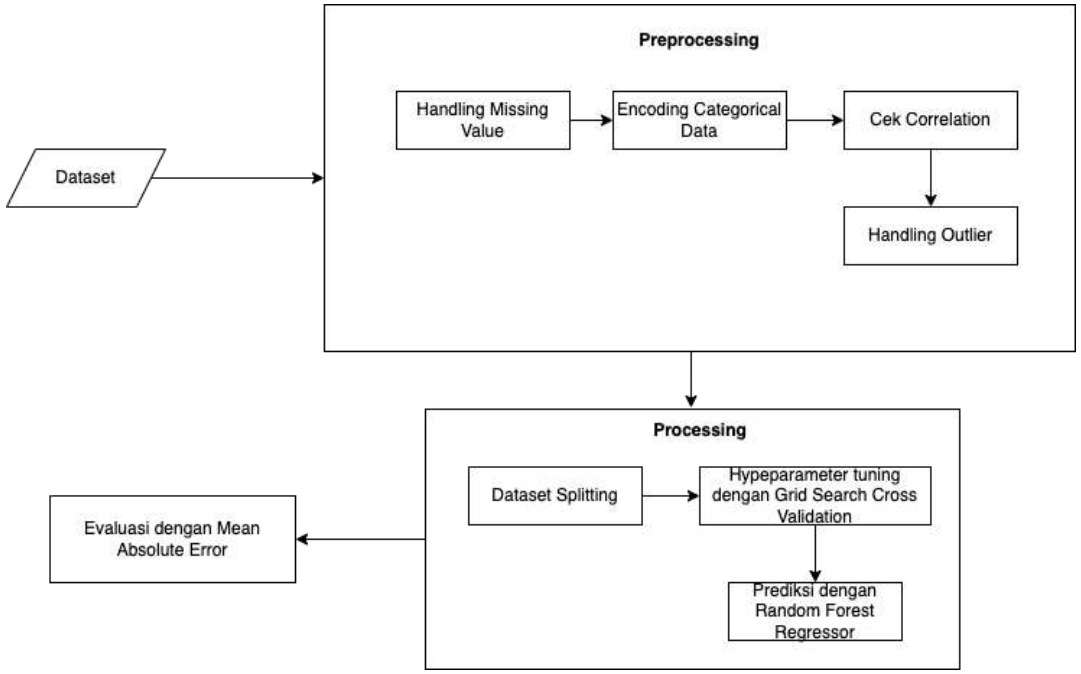
Pertanyaan Penelitian (Rumusan Masalah)

1. Bagaimana Random Forest Regressor dapat digunakan untuk menganalisis dan memprediksi pengaruh berbagai faktor sosioekonomi dan kesehatan terhadap angka harapan hidup di berbagai negara?

2. Apa saja faktor sosioekonomi dan kesehatan utama yang paling berpengaruh terhadap perbedaan angka harapan hidup antar negara?

3. Bagaimana efektivitas Random Forest Regressor dalam menganalisis pengaruh faktor sosioekonomi dan kesehatan terhadap angka harapan hidup, dengan penilaian efektivitas berdasarkan Confusion Matrix?

Metode



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 indikator tingkat harapan hidup suatu negara yang diambil dari tahun 2000 hingga 2015 di 193 negara di Asia.

Metode

PREPROCESSING

1. ***Handling Missing Value***

   Data masukan dilakukan handling missing value terlebih dahulu. Missing value dapat terjadi karena kesalahan penginputan data atau memang data tersebut tidak ada.

2. ***Encoding Categorical Data***

   Mengubah data yang bersifat kategorik menjadi numerik. Hal ini dikarenakan mesin hanya dapat membaca data berupa angka saja.

3. ***Cek Correlation***

   Koefisien korelasi data yang lemah akan dihapus sehingga menyisakan korelasi data cukup, kuat dan sempurna.

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

4. ***Handling Outlier***

   Keberadaan outlier dapat menjadi sumber ketidakakuratan dan interpretasi yang salah, serta memengaruhi performa model machine learning, Salah satu metode yang umum digunakan untuk mendeteksi dan mengelola outlier adalah dengan menggunakan Z-Score.

Metode

PROCESSING

1. ***Dataset Splitting***

Terdapat dua bagian data antara lain data latih dan data uji dengan prosentase masing-masing sebesar 20% untuk data uji dan 80% untuk data latih. Nantinya data uji akan dipakai untuk pengujian model yang dihasilkan sedangkan data latih akan dipakai untuk proses pembelajaran mesin.

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

Proses menemukan kombinasi terbaik dari hyperparameter untuk model machine learning guna mencapai kinerja optimal.

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

.

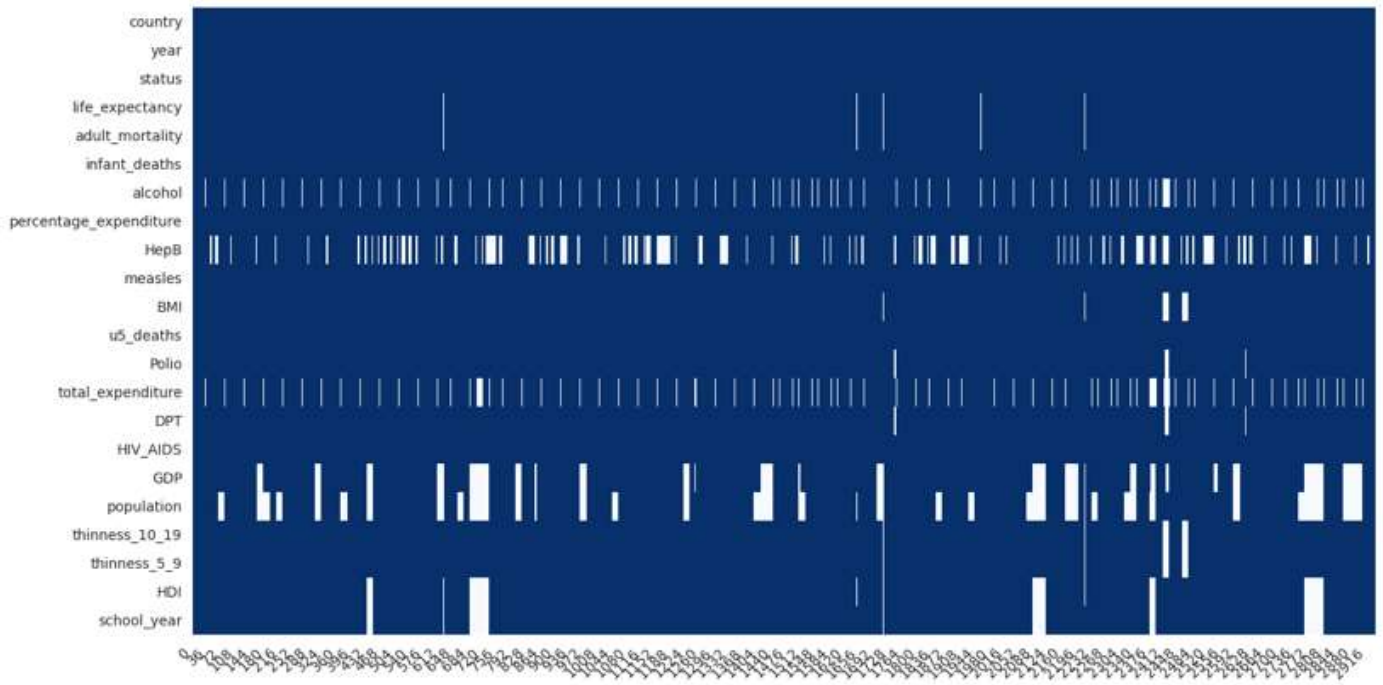
3. ***Pemodelan dengan Random Forest Regressor***

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 .

Hasil

PREPROCESSING

Tahapan ini menghasilkan data yang telah dibersihkan dan diubah. Visualisasi dari kelengkapan data ditunjukkan pada Gambar, dimana area berwarna putih menunjukkan keberadaan nilai yang hilang atau missing value pada kolom tertentu. Dari Gambar tersebut, dapat dilihat bahwa beberapa kolom memiliki missing value dalam jumlah yang beragam.

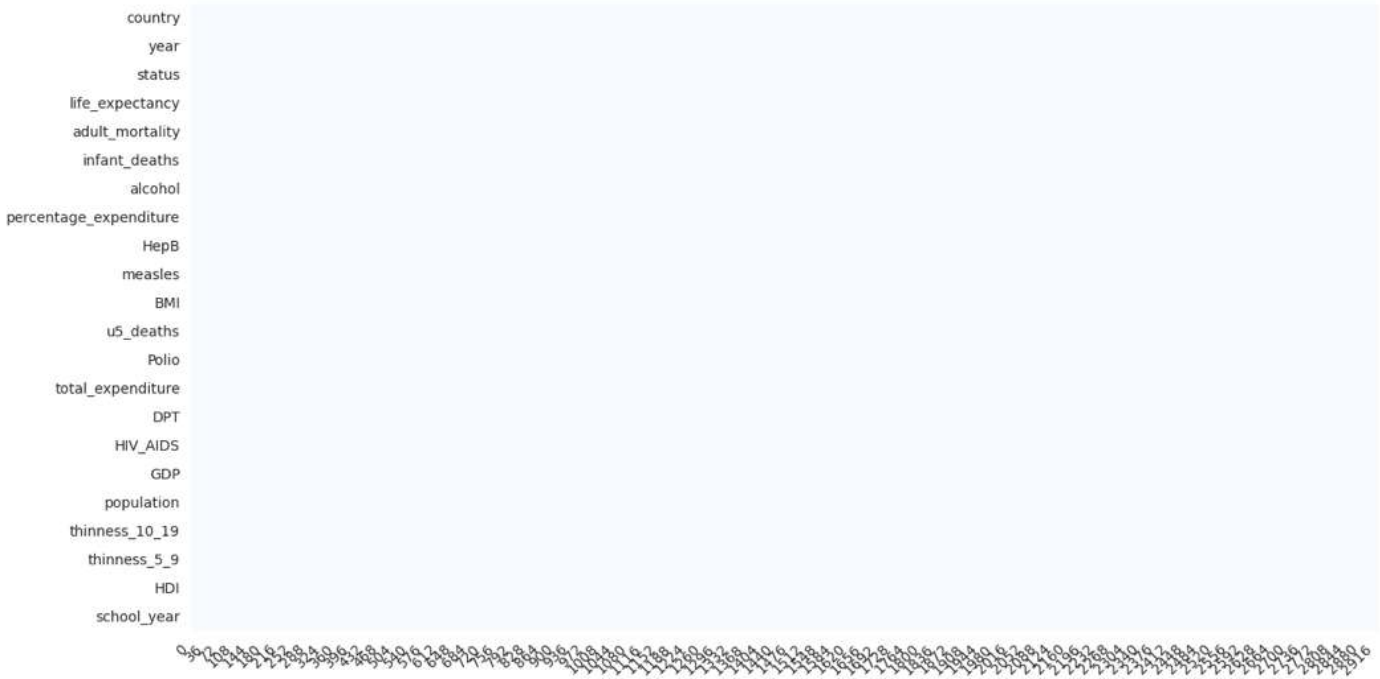


Sebelum Imputasi Missing Value

Hasil

PREPROCESSING

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

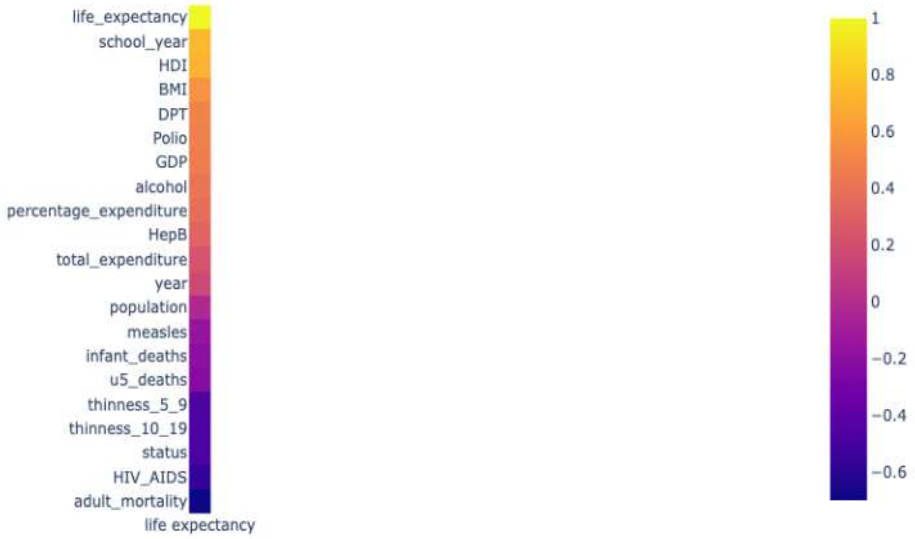


Setelah Imputasi Missing Value

Hasil

PREPROCESSING

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. Seperti yang diuraikan sebelumnya, untuk memudahkan dalam menentukan korelasi antara fitur dengan target, peneliti memanfaatkan fungsi `corr` dari pandas dan hasilnya divisualisasikan pada Gambar



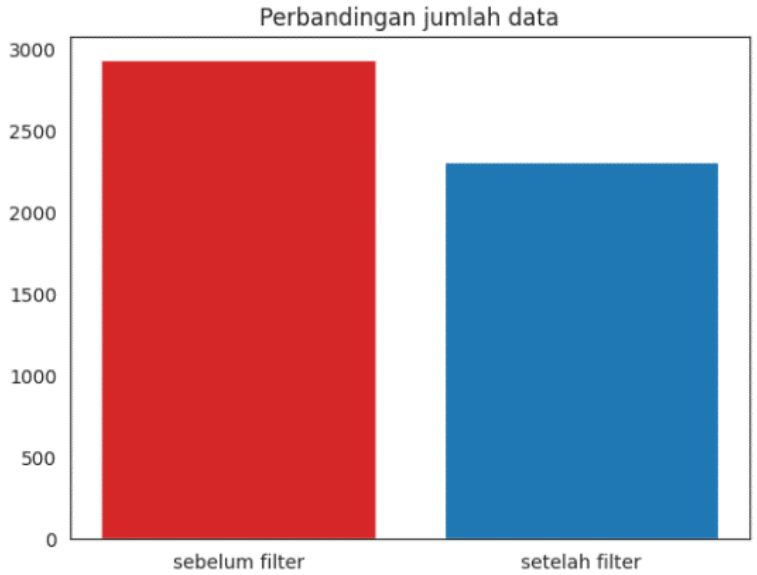
Visualisasi Korelasi

Hasil

PREPROCESSING

Berdasarkan Gambar yang ada pada slide sebelumnya, 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.



Perbandingan Jumlah Data

Hasil

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.

Hasil

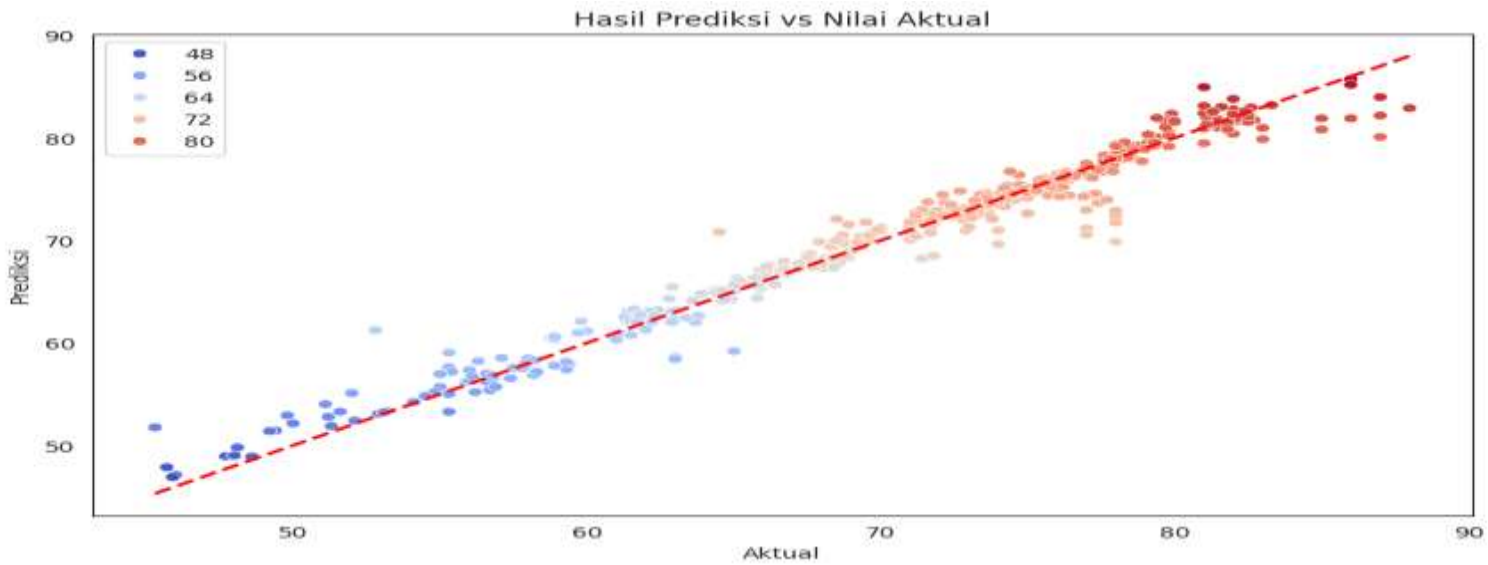
UJI COBA MODEL

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

Sampel Hasil Pengujian

Dari Sampel Hasil Pengujian 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.



Visualisasi Uji Coba Model

Pembahasan

Grafik yang disajikan pada Visualisasi Uji Coba Model 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.

Temuan Penting Penelitian

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.

Manfaat Penelitian

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.

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.

Referensi

[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.

Referensi

[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.


UMSIDA
DARI SINI PENCERAHAN BERSEMI