JTERA (Jurnal Teknologi Rekayasa), Vol. 10, No. 2, Desember 2025, Hal. 59-68 p-ISSN 2548-737X
Terakreditasi "**Peringkat 3**" oleh Kemdiktisaintek, Nomor : 10/C/C3/DT.05.00/2025 e-ISSN 2548-8678
DOI: 10.31544/jtera.v10.i2.2025.59-68

Model Prediktif Indeks Kebahagiaan Berbasis *Gradient Boosting Regressor* dengan Optimalisasi Seleksi Fitur dan Implementasi Web

Dani Ferinan[1], Nisa Hanum Harani[2], Syafrial Fachri Pane[3]
[1,2,3]Program Studi D4 Teknik Informatika, Universitas Logistik dan Bisnis Internasional
Jl. Sariasih No.54, Sarijadi, Kec. Sukasari, Kota Bandung, Jawa Barat 40151, Indonesia
1214050@std.ulbi.ac.id

Abstrak

Penelitian ini menghadapi tantangan dalam memodelkan Indeks Kebahagiaan 2021 dari Badan Pusat Statistik (BPS) yang memiliki dimensi fitur sangat tinggi dan potensi redundansi, yang dapat menurunkan akurasi dan interpretabilitas model. Tujuan utama penelitian ini adalah untuk mengidentifikasi fitur-fitur paling berpengaruh dalam data tersebut untuk meningkatkan akurasi, efisiensi komputasi, dan transparansi model prediksi berbasis pohon keputusan. Metodologi mencakup pra-pemrosesan data dengan imputasi modus, transformasi Yeo-Johnson, dan *Robust Scaler*. Tiga algoritma regresi diuji: *Decision Tree*, *Random Forest*, dan *Gradient Boosting Regressor*, yang dioptimalkan menggunakan *Particle Swarm Optimization* (PSO). Model terbaik dievaluasi menggunakan metrik R^2, MSE, RMSE, dan MAE serta dianalisis lebih lanjut menggunakan SHAP untuk interpretasi. Hasil menunjukkan bahwa *Gradient Boosting Regressor* adalah model paling unggul dengan nilai R^2 sebesar 0,696 saat menggunakan 20 fitur terseleksi. Selain itu, sebagai bentuk implementasi praktis, model diimplementasikan ke dalam sebuah aplikasi *web* interaktif berbasis Flask yang memungkinkan pengguna memasukkan data melalui antarmuka kuisioner dan menerima prediksi indeks kebahagiaan secara *real-time*. Integrasi ini menjembatani hasil riset dengan pemanfaatan nyata oleh pengguna akhir.

Kata kunci: Indeks Kebahagiaan, *Machine Learning*, *Feature Selection*, Gradient Boosting, Aplikasi Web

Abstract

This study addresses the challenges of modeling Indonesia's 2021 Happiness Index from the Central Bureau of Statistics (BPS), which features high-dimensional data with potential redundancy, reducing model accuracy and interpretability. The primary objective is to identify the most influential features in the dataset to improve the performance and transparency of tree-based predictive models. The methodology involves preprocessing through mode imputation, Yeo-Johnson transformation, and Robust Scaler. Three regression models were compared—Decision Tree, Random Forest, and Gradient Boosting Regressor—with hyperparameter tuning using Particle Swarm Optimization (PSO). The best-performing model was evaluated using R^2, MSE, RMSE, and MAE, and further interpreted with SHAP values. The Gradient Boosting Regressor emerged as the top model, achieving an R^2 of 0.696 with 20 selected features. Furthermore, to bridge technical results with practical use, the model was integrated into a real-time interactive web application using the Flask framework. The application enables users to input questionnaire data and instantly receive happiness index predictions, providing an accessible interface for end-users and stakeholders.

Keywords: *Happiness Index, Machine Learning, Feature Selection, Gradient Boosting, Web Application*

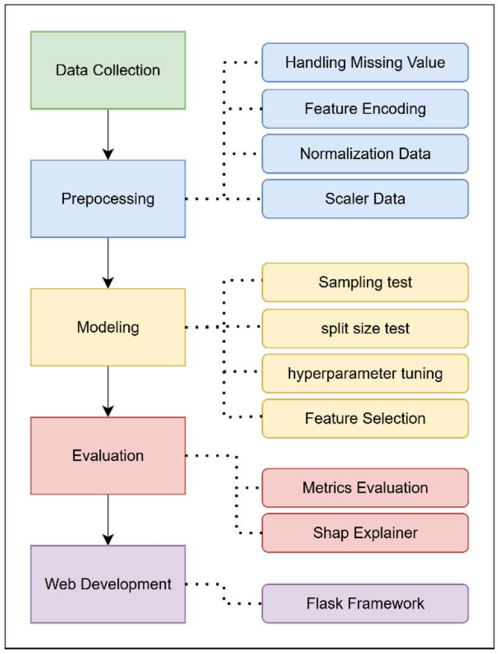
I. PENDAHULUAN

Perhatian terhadap kesejahteraan subjektif atau kebahagiaan masyarakat semakin meningkat telah mendorong pemanfaatannya sebagai indikator keberhasilan pembangunan suatu negara, melengkapi data objektif seperti Produk Domestik Bruto (PDB) dan Indeks Pembangunan Manusia

(IPM) [1]. Namun, pengukuran kebahagiaan melalui data survei, seperti data Indeks Kebahagiaan 2021 oleh Badan Pusat Statistik (BPS), menyajikan tantangan signifikan dalam pemodelan. Tantangan ini muncul karena survei tersebut dirancang sangat komprehensif dan mencakup puluhan sampai ratusan pertanyaan untuk menangkap berbagai aspek kehidupan [2]**.** Akibatnya, data yang dihasilkan memiliki karakteristik dimensi fitur yang sangat tinggi serta potensi adanya fitur yang tidak relevan atau redundan, yang dapat mengurangi akurasi dan interpretabilitas model *machine learning* [3]. Permasalahan ini menghambat kemampuan untuk mengidentifikasi faktor-faktor penentu kebahagiaan secara presisi, yang krusial untuk perancangan kebijakan publik yang efektif. Oleh karena itu, diperlukan pendekatan yang mampu menyederhanakan kompleksitas data tanpa mengorbankan informasi penting.

Penelitian sebelumnya telah membuktikan keberhasilan penerapan *machine learning* untuk memprediksi kebahagiaan dan isu terkait. Sebagai contoh, *Gradient Boosting Classifier* (GBC) telah berhasil digunakan untuk memprediksi kesejahteraan subjektif [2], dan sebuah model *Gradient Boosting* juga telah diusulkan untuk mengoptimalkan prediksi *human well-being* (SWB) [4]. Di bidang terkait, LightGBM juga sukses diterapkan untuk memprediksi masalah kesehatan mental [5]. Tren penelitian terkini menunjukkan bahwa algoritma berbasis *ensemble*, khususnya varian dari gradient boosting, sangat efektif dalam menangani data sosial yang kompleks. Namun, fokus utama dari banyak penelitian ini lebih pada pencapaian akurasi prediktif, dengan eksplorasi yang lebih terbatas pada metode seleksi fitur sistematis dan interpretabilitas kontribusi setiap fitur secara mendalam. Sebagai tambahan, sebuah penelitian lain telah menerapkan kombinasi regresi *ensemble Random Forest Regression*, *XGBoost Regression*, dan *Decision Tree Regression* untuk memprediksi skor kebahagiaan berdasarkan *World Happiness Index* dan *Human Development Index* [1]. Penelitian tersebut juga melakukan analisis *feature importance* dan *permutation importance* guna mengidentifikasi fitur-fitur kunci seperti GNI per kapita yang memberikan kontribusi terbesar dalam model, sehingga menyediakan wawasan sistematis mengenai peran masing-masing variabel dalam prediksi kebahagiaan.

Untuk menjawab kesenjangan dalam pengembangan sistem prediksi kebahagiaan yang akurat, efisien, dan dapat diinterpretasikan, penelitian ini menawarkan empat kontribusi utama yang bersifat strategis sekaligus aplikatif. Pertama, dilakukan analisis komparatif secara sistematis antara model prediktif yang dilatih menggunakan seluruh fitur dan model yang hanya menggunakan subset fitur hasil seleksi. Tujuan dari pendekatan ini adalah untuk menunjukkan bahwa penyederhanaan model tidak hanya mampu mengurangi kompleksitas dan waktu komputasi, tetapi juga dapat meningkatkan akurasi serta memusatkan perhatian pada variabel-variabel yang paling signifikan secara substantif. Kedua, penelitian ini mengadopsi PSO (*Particle Swarm Optimizatio*) sebagai teknik optimasi hiperparameter yang efisien dan adaptif. PSO digunakan untuk menemukan konfigurasi parameter terbaik secara otomatis, sehingga menghindari proses tuning manual yang melelahkan dan sekaligus mendorong peningkatan kinerja model. Ketiga, aspek interpretabilitas model diperkuat melalui penerapan SHAP (*SHapley Additive exPlanations*), yang memberikan wawasan kuantitatif dan visual mengenai kontribusi serta interaksi masing-masing fitur terhadap *output* model. Ini memungkinkan transparansi tinggi dalam pengambilan keputusan berbasis model. Keempat, sebagai bentuk kontribusi praktis, seluruh sistem diimplementasikan ke dalam aplikasi *web* interaktif yang memungkinkan pengguna, baik pemangku kebijakan maupun masyarakat umum, untuk mengeksplorasi hasil prediksi dan skenario secara dinamis. Dengan mengintegrasikan akurasi prediksi, efisiensi komputasi, transparansi model, dan kemudahan akses, penelitian ini menghadirkan solusi *end-to-end* yang relevan dan siap diterapkan dalam konteks pengambilan keputusan berbasis data.



Gambar 1. Metodologi penelitian

II. METODE PENELITIAN

Alur metodologi penelitian yang dilakukan dalam penelitian ini disajikan secara ringkas pada Gambar 1. Tahapan penelitian ini dibagi menjadi lima fase utama yang berurutan, dimulai dari *Data Collection*, diikuti oleh *Preprocessing*, *Modeling*, *Evaluation*, dan diakhiri dengan *Web Development*.

A. *Data Collection*

Penelitian ini menggunakan *dataset* komprehensif yaitu Indeks Kebahagiaan 2021 yang di keluarkan oleh BPS [6], yang terdiri dari 74.684 responden dan 276 kolom. Kolom-kolom ini mencakup beragam indikator yang berkaitan dengan kebahagiaan. Setiap responden memberikan informasi mendalam mengenai berbagai dimensi kebahagiaan, yang diukur berdasarkan serangkaian indikator. Indikator-indikator ini dikelompokkan ke dalam beberapa kategori utama, yaitu kepuasan hidup personal, kepuasan hidup sosial, perasaan, dan makna hidup. Tabel 1 merinci dimensi-dimensi kebahagiaan dan indikator-indikator beserta bobot yang digunakan dalam penentuan indeks Kebahagiaan.

Indeks Kebahagiaan yang dihitung untuk setiap responden didasarkan pada dimensi-dimensi tersebut, yang mencakup berbagai aspek kehidupan. Setiap dimensi kemudian dibagi lagi menjadi indikator-indikator spesifik yang menggambarkan faktor-faktor kunci yang berkontribusi pada tingkat kebahagiaan individu.

Tabel 1. Indikator indeks kebahagiaan

Dimensi	Sub-Dimensi	Indikator	Bobot
Kepuasan Hidup (34,80)	Kepuasan Hidup Personal (50)	Pendidikan dan Keterampilan [R505]	18,34
		Pekerjaan/Usaha/Kegiatan Utama [R601B2/R605]	21,67
		Pendapatan Rumah Tangga [R610]	22,81
		Kesehatan [R708]	17,04
		Kondisi Rumah dan Fasilitas Rumah [R1211]	20,14
	Kepuasan Hidup Sosial (50)	Keharmonisan Keluarga [R802]	19,41
		Ketersediaan Waktu Luang [R903]	18,93
		Hubungan Sosial [R1005]	22,13
		Keadaan Lingkungan [R1102]	20,64
		Kondisi Keamanan [R1108]	18,89
Perasaan (31,18)	-	Perasaan Senang/Riang/Gembira [R1302]	25,86
		Perasaan Tidak Khawatir/Cemas [R1304]	36,90
		Perasaan Tidak Tertekan [R1306]	37,34
Makna Hidup (34,02)	-	Kemandirian [R1402]	16,56
		Penguasaan Lingkungan [R1404]	18,44
		Pengembangan Diri [R1406]	15,27
		Hubungan Positif dengan Orang Lain [R1408]	15,48
		Tujuan Hidup [R1410]	17,48
		Penerimaan Diri [R1412]	16,78

Dalam penghitungan Indeks Kebahagiaan, bobot masing-masing indikator sangat penting untuk menentukan kontribusinya terhadap total indeks. Bobot ini, yang disajikan pada Tabel 1, diperoleh dari buku Indeks Kebahagiaan 2021 yang dirilis oleh Badan Pusat Statistik (BPS). Formula perhitungan total Indeks Kebahagiaan yang dirilis oleh BPS pada tahun 2021 dijadikan acuan utama untuk menghitung Indeks Kebahagiaan pada setiap responden. Setiap bobot indikator menunjukkan seberapa besar faktor tersebut memengaruhi kebahagiaan individu. Rumus perhitungan dimensi kepuasan hidup penyusun kebahagiaan dan indeks kebahagiaan dapat dilihat pada persamaan (1) – (6).

$$index\_khp = \frac{\sum w_i * x_i}{\sum w_i} \quad (1)$$

$$indeks\_khs = \frac{\sum w_i * x_i}{\sum w_i} \quad (2)$$

Persamaan (1) dan (2) dapat di jelaskan sebagai berikut:

indeks_khp: Skor sub-dimensi Kepuasan Hidup Personal.

indeks_khs: Skor sub-dimensi Kepuasan Hidup Sosial.

x_i: Merupakan skor indikator ke-i.

w_i: Merupakan bobot dari indikator ke-i.

$$indeks\_kh = \frac{w_1 * indeks\_khp + w_2 * indeks\_khs}{w_1 + w_2} \quad (3)$$

Persamaan (3) dapat di jelaskan sebagai berikut:

indeks_kh: Skor dimensi kepuasan hidup.
w_1: Bobot sub-imensi Kepuasan Hidup Personal.
w_2: Bobot sub-imensi Kepuasan Hidup Sosial.

$$indeks\_p = \frac{\sum w_i * x_i}{\sum w_i} \quad (4)$$

$$indeks\_m = \frac{\sum w_i * x_i}{\sum w_i} \quad (5)$$

Persamaan (4) dan (5) dapat di jelaskan sebagai berikut:
indeks_p: Skor sub-dimensi Perasaan.
indeks_m: Skor sub-dimensi Makna Hidup.
x_i: Merupakan skor indikator ke-i.
w_i: Merupakan bobot dari indikator ke-i.

$$indeks\_k = \frac{w_1 * indeks\_kh + w_2 * indeks\_p + w_3 * indeks\_m}{w_1 + w_2 + w_3} \quad (6)$$

Persamaan (6) dapat dijelaskan sebagai berikut:
*indeks*_k: Skor indeks kebahagiaan.
indeks_kh: Skor dimensi kepuasan hidup.
indeks_p: Skor dimensi perasaan.
indeks_m: Skor dimensi makna hidup.
w_1: Bobot dimensi Kepuasan Hidup.
w_2: Bobot dimensi Perasaan.
w_3: Bobot dimensi Makna Hidup.

Dataset ini mengandung *missing value* yang disebabkan oleh *skip logic* dalam survei. *Skip logic* adalah mekanisme yang memungkinkan pertanyaan tertentu dilewati berdasarkan jawaban responden sebelumnya, sehingga tidak semua responden mengisi setiap kolom data [7]. Oleh karena itu, proses pengolahan *missing value* harus dilakukan dengan cermat untuk memastikan integritas data. Beberapa metode umum yang dapat digunakan meliputi imputasi nilai yang hilang, penghapusan baris atau kolom yang mengandung *missing value*, atau menggunakan model prediksi untuk memperkirakan nilai yang tidak ada.

Secara keseluruhan, *dataset* ini menyediakan informasi yang komprehensif mengenai faktor-faktor yang mempengaruhi kebahagiaan individu. Dengan menganalisis data dan menghitung Indeks Kebahagiaan berdasarkan indikator-indikator yang relevan, dapat diperoleh wawasan yang lebih mendalam mengenai aspek-aspek kehidupan yang paling mempengaruhi kebahagiaan masyarakat.

B. *Prepocessing*

Pra-pemrosesan adalah tahapan penting dalam penelitian yang bertujuan untuk menyiapkan data [8]. Pra-pemrosesan dilakukan secara sistematis untuk menjamin integritas dan kesiapan data sebelum pemodelan [9]. Proses ini diawali dengan penanganan nilai hilang, yang mayoritas disebabkan oleh *skip logic* survei [10], menggunakan imputasi modus untuk variabel kategorikal. Selanjutnya, variabel non-numerik ditransformasi menjadi format numerik melalui standardisasi variabel biner dan *one-hot encoding* untuk variabel kategori nominal. Berdasarkan hasil Uji Shapiro-Wilk, data terbukti tidak terdistribusi secara normal ($p < 0.001$), sehingga diperlukan transformasi [11]. Oleh karena itu, diterapkan transformasi Yeo-Johnson [12] pada fitur numerik dan transformasi logaritmik (log1p) pada variabel target. Sebagai tahap akhir, seluruh fitur numerik diskalakan menggunakan *Robust Scaler* [13] untuk meminimalkan pengaruh nilai ekstrem (*outlier*), sehingga menghasilkan *dataset* yang bersih dan optimal untuk analisis.

C. *Modeling*

Tahap pemodelan dirancang secara komprehensif untuk optimasi dan evaluasi model melalui serangkaian pengujian sistematis. Fokus utama mencakup empat area: penentuan ukuran sampel, evaluasi rasio pembagian data, optimasi *hyperparameter*, dan analisis kepentingan fitur. Setiap pengujian bertujuan untuk menyempurnakan model agar mencapai performa prediktif yang maksimal dan dapat diandalkan.

Penelitian ini membandingkan tiga model regresi. *Decision Tree Regressor* bekerja dengan membuat serangkaian aturan keputusan sederhana untuk memprediksi nilai target [14]. *Random Forest Regressor* meningkatkan pendekatan ini dengan membangun banyak pohon keputusan dari sampel data acak. Prediksi akhir ditentukan oleh suara mayoritas (*voting*) dari seluruh pohon tersebut [15]. Sementara itu, *Gradient Boosting Regressor* membangun pohon secara berurutan, di mana setiap pohon baru secara spesifik belajar untuk memperbaiki kesalahan dari pohon sebelumnya, sering kali menghasilkan performa prediktif yang superior [16].

Eksperimen dimulai dengan menguji berbagai ukuran sampel (dari 300 hingga seluruh data) dan rasio test split (0.1, 0.2, 0.3, dan 0.4) untuk menemukan konfigurasi data yang paling optimal. Selanjutnya, *hyperparameter* model disetel secara cermat menggunakan algoritma *Particle Swarm Optimization* (PSO), *hyperparameter tuning* sangat penting untuk meningkatkan kinerja dan akurasi model [17]. Tahap terakhir adalah analisis kepentingan fitur, di mana model terbaik dievaluasi dengan jumlah fitur yang bervariasi (20, 50, 100, dan seluruhnya) untuk mengidentifikasi variabel paling signifikan. Penting dicatat, dalam analisis ini, fitur-fitur hasil *one-hot encoding* dari pertanyaan yang sama dikelompokkan berdasarkan prefiks untuk menilai dampak variabel asli secara utuh dan koheren.

D. ***Evaluation***

Bagian evaluasi dalam penelitian ini dilakukan untuk mengukur kinerja model regresi yang telah dikembangkan dalam memprediksi indeks kebahagiaan. Metrik evaluasi yang digunakan adalah R^2 (Koefisien Determinasi), MSE (*Mean Squared Error*), RMSE (*Root Mean Squared Error*), dan MAE (*Mean Absolute Error*) [18].

Kinerja prediktif model diukur menggunakan beberapa metrik standar. Koefisien Determinasi (R^2) digunakan untuk mengukur proporsi varians yang dapat dijelaskan oleh model, dengan nilai terbaik 1, dan terburuk 0 [19] persamaan R2 dapat dilihat pada persamaan (7), dengan yi adalah nilai aktual, $\hat{y}_i$ adalah nilai prediksi dan adalah $\bar{y}$ nilai rata-rata dari seluruh data aktual.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

Kesalahan prediksi dievaluasi lebih lanjut menggunakan tiga metrik. *Mean Squared Error* (MSE) menghitung rata-rata kuadrat selisih dan sensitif terhadap *outlier* [20]. Rumus MSE dapat dilihat pada persamaan (8), dengan *n* adalah jumlah data, y_i adalah nilai aktual dan $\hat{y}_i$ adalah nilai prediksi.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (8)$$

Untuk mempermudah interpretasi, digunakan *Root Mean Squared Error* (RMSE), yang memiliki satuan sama dengan variabel target [21]. Rumus RMSE dapat dilihat pada persamaan (9), dengan *n* adalah jumlah data, y_i adalah nilai aktual dan $\hat{y}_i$ adalah nilai prediksi.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

Terakhir, *Mean Absolute Error* (MAE), yang lebih robust terhadap *outlier*, mengukur rata-rata selisih absolut [22]. Rumus MAE dapat dilihat pada persamaan (10), dengan *n* adalah jumlah data, y_i adalah jumlah nilai aktual $\hat{y}_i$ adalah nilai prediksi

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (10)$$

Untuk semua metrik berbasis kesalahan (MSE, RMSE, MAE), nilai yang lebih kecil menunjukkan performa yang lebih baik.

Kedua, untuk untuk membuat model prediksi yang paling akurat menjadi lebih transparan. Dengan menghitung kontribusi setiap variabel, SHAP (*SHapley Additive exPlanations*) berhasil mengidentifikasi faktor-faktor sosioekonomi yang paling berpengaruh terhadap skor kebahagiaan dan menjelaskan bagaimana pengaruh tersebut bekerja, baik secara umum maupun pada kasus individual

E. ***Web Development***

Pengembangan antarmuka *web* bertujuan untuk mengimplementasikan model terbaik ke dalam sebuah aplikasi yang fungsional dan mudah diakses. Arsitektur aplikasi ini menggunakan Flask, sebuah *micro-framework* Python, sebagai *backend* untuk menangani logika dan pemrosesan. Model yang telah terlatih dan dioptimalkan dengan fitur-fitur terseleksi diintegrasikan ke dalam *backend* melalui proses serialisasi, memungkinkannya dimuat saat aplikasi berjalan.

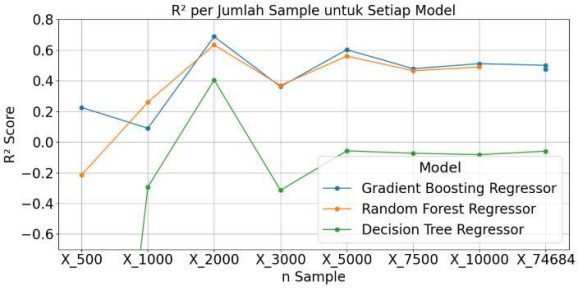
III. HASIL DAN PEMBAHASAN

A. ***Data Extraction***

Indeks Kebahagiaan tidak langsung tersedia dalam *dataset*. Indeks ini dihitung secara spesifik berdasarkan nilai-nilai indikator dan bobot yang telah ditetapkan oleh BPS, mengacu pada formula perhitungan total Indeks Kebahagiaan yang dirilis pada tahun 2021. Untuk setiap responden, Indeks Kebahagiaan dihitung dengan mengalikan nilai pada setiap indikator dengan bobot yang sesuai, lalu menjumlahkan hasilnya.

Berdasarkan perhitungan ini, hasil Indeks Kebahagiaan akan memberikan gambaran mendalam mengenai tingkat kebahagiaan individu yang tergambar dari berbagai dimensi kepuasan hidup personal, kepuasan hidup sosial, perasaan, dan makna hidup. Analisis selanjutnya akan membahas distribusi Indeks Kebahagiaan di antara responden serta faktor-faktor dominan yang memengaruhi tingkat kebahagiaan berdasarkan bobot indikator dan formulasi yang telah ditetapkan BPS. Pembahasan juga akan mencakup dampak dari penanganan *missing value* terhadap integritas dan interpretasi hasil Indeks Kebahagiaan.

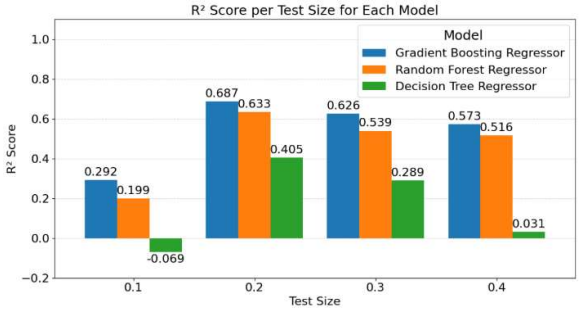
B. ***Modeling***



Gambar 2. Uji Ukuran Sampel

Gambar 2 menyajikan hasil uji kinerja tiga model regresi yaitu *Gradient Boosting Regressor*, *Random Forest Regressor*, dan *Decision Tree Regressor* pada berbagai ukuran sampling *dataset*. Observasi menunjukkan bahwa *Gradient Boosting Regressor* secara konsisten menghasilkan nilai R2 tertinggi pada berbagai skala data, mencapai puncaknya di

0,687 pada *dataset* X_2000, meskipun terdapat sedikit penurunan pada keseluruhan *dataset* (X_74684), sementara *Decision Tree Regressor* secara persisten menunjukkan R2 negatif. Dalam hal efisiensi komputasi, *Random Forest Regressor* memerlukan waktu yang signifikan lebih lama, terutama pada *dataset* besar, dibandingkan dengan *Gradient Boosting Regressor* yang menawarkan keseimbangan optimal antara akurasi dan kecepatan, serta *Decision Tree Regressor* yang tercepat namun dengan akurasi terburuk. Hasil ini mengindikasikan bahwa model *ensemble* berbasis *boosting* secara signifikan lebih unggul dalam memodelkan hubungan data dan mencapai kinerja prediksi superior.



Gambar 3. Uji Ukuran Pembagian Dataset

Analisis kinerja model regresi pada berbagai ukuran *test-set* menunjukkan variabilitas yang signifikan dalam kapabilitas prediksi. Seperti diilustrasikan pada Gambar 3, *Gradient Boosting Regressor* secara konsisten menghasilkan nilai R2 tertinggi di seluruh rentang *test-size* yang diuji, mencapai puncak 0,687 pada *test-size* 0,2. *Random Forest Regressor* menunjukkan kinerja yang solid sebagai model *ensemble* kedua terbaik, mencapai R2 0,633 pada *test-size* yang sama. Sebaliknya, *Decision Tree Regressor* menunjukkan kinerja paling tidak stabil dan seringkali buruk, bahkan menghasilkan R2 negatif (-0,069) pada *test-size* 0,1, menandakan model tersebut lebih buruk daripada prediksi rata-rata data. Temuan ini menegaskan keunggulan metode *ensemble* dalam meningkatkan stabilitas dan akurasi prediksi dibandingkan pohon keputusan tunggal, serta menyoroti pentingnya pemilihan ukuran test set yang optimal untuk evaluasi model yang *robust*.

Tabel 2. Evaluasi Hasil Hyper Paramater Tunning PSO

No	Model	R2	MSE	RMSE	MAE
1	*Decision Tree Regressor*	0,239	0,022	0,147	0,093
2	*Random Forest Regressor*	0,637	0,010	0,102	0,071
3	*Gradient Boosting Regressor*	0,693	0,009	0,093	0,068

Setiap model dievaluasi menggunakan metrik R2, MSE, RMSE, dan MAE, Selanjutnya dioptimalkan menggunaan *Particle Swarm Optimization* (PSO) untuk penentuan parameter optimal dengan data. Hasil evaluasi yang disajikan pada Tabel 2 menunjukkan bahwa *Gradient Boosting Regressor* memberikan kinerja terbaik dengan nilai R2 tertinggi sebesar 0,693, serta nilai MSE dan RMSE terendah, masing-masing 0,009 dan 0,093. *Random Forest Regressor* menempati posisi kedua dengan R2 sebesar 0,634, sementara *Decision Tree Regressor* memiliki performa paling rendah dengan R2 hanya 0,239. Perbandingan ini mengindikasikan bahwa *ensemble learning model*, khususnya *boosting*, secara signifikan lebih unggul dalam memprediksi variabel target dibandingkan model pohon tunggal, menunjukkan efektivitas strategi optimasi yang diterapkan dalam meningkatkan akurasi prediksi model.

C. ***Feature Selection***

Berdasarkan hasil optimasi yang menunjukkan bahwa *Gradient Boosting Regressor* adalah model dengan kinerja terbaik, proses seleksi fitur ini difokuskan pada analisis kontribusi variabel dalam model tersebut untuk menentukan peringkat kepentingannya.

Tabel 3. *Feeature Importance.*

No	Feature	Pertanyaan	Importance
1	R1202	Luas lantai bangunan tempat tinggal	0.127618
2	R1002E	Kepercayaan terhadap aparat pemerintahan kabupaten/kota	0.103786
3	R1002B	Selalu menyempatkan hadir di rumah duka saat ada kejadian kematian di lingkungan tempat tinggal	0.071125
4	R1003G	Tidak keberatan dan bersedia meluangkan waktu jika terpilih menjadi responden dalam sebuah survei yang diselenggarakan pemerintah	0.067268
5	R1003E	Menghormati dan menaati keputusan hasil musyawarah warga, meskipun hal tersebut bertentangan dengan kehendak dan pendapat	0.056854

Tabel 3 menyajikan lima fitur teratas dengan tingkat *importance* tertinggi, yang merupakan hasil dari model terbaik. *Feature importance* mengindikasikan seberapa besar kontribusi masing-masing fitur terhadap kinerja prediksi model. Fitur dengan *importance* tertinggi adalah 'Luas lantai bangunan tempat tinggal' (R1202) dengan nilai 0,127618, diikuti oleh 'Kepercayaan terhadap aparat pemerintahan kabupaten/kota' (R1002E) sebesar 0,103786.

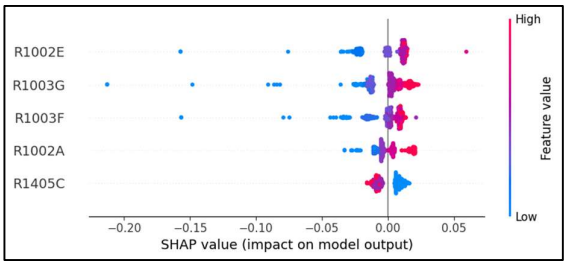
Urutan ini menunjukkan bahwa aspek fisik bangunan tempat tinggal dan dimensi kepercayaan sosial-pemerintahan merupakan faktor-faktor prediktif paling dominan yang diekstrak oleh model, menegaskan relevansi karakteristik demografi dan sosial dalam memengaruhi variabel target.

Tabel 4. Evaluasi model terbaik berdasarkan jumlah fitur yang dipakai.

No	Jumlah Pertanyaan	R2	RMSE	Time
1	20	0,696	0,093	0,16s
2	50	0,694	0,093	0,30s
3	100	0,683	0,095	0,53s
4	Semua Pertanyaan	0,693	0,093	0,92s

Analisis yang disajikan pada Tabel 4 mengevaluasi kinerja model *Gradient Boosting Regressor*, yang diidentifikasi sebagai model terbaik sebelumnya, pada jumlah pertanyaan yang bervariasi (20, 50, 100, dan Semua pertanyaan) dengan jumlah sampel terbaik (2000) dan setelah optimasi menggunakan *Particle Swarm Optimization* (PSO). Hasil menunjukkan bahwa model mencapai R2 tertinggi sebesar 0,696 dengan RMSE 0,093 saat menggunakan 20 pertanyaan, dengan waktu komputasi tercepat 0,16 detik.

Meskipun penambahan jumlah pertanyaan hingga 100 dan "Semua" meningkatkan waktu komputasi secara signifikan (menjadi 0,53 detik dan 0,92 detik masing-masing), nilai R2 justru sedikit menurun menjadi 0,683 dan 0,693, diikuti dengan peningkatan RMSE. Ini mengindikasikan adanya titik optimum dalam jumlah fitur (pertanyaan) yang digunakan, di mana penambahan fitur yang tidak relevan atau berlebihan dapat memperkenalkan noise atau kompleksitas yang tidak perlu, yang pada akhirnya dapat sedikit mengurangi kemampuan generalisasi model, meskipun waktu komputasi meningkat seiring dengan peningkatan dimensi data.



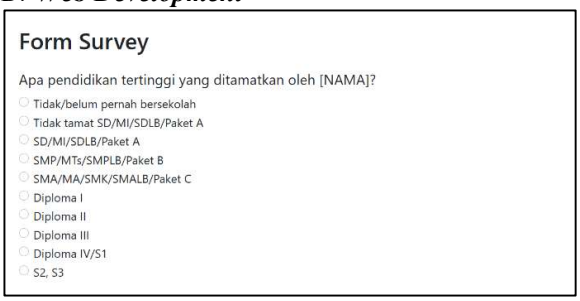
Gambar 4. Shap Explanation

Untuk memahami kontribusi setiap fitur pada model dengan performa terbaik, dilakukan analisis SHAP (*SHapley Additive exPlanations*). Analisis ini difokuskan pada 20 fitur pilihan yang menghasilkan performa model optimal (sebagaimana ditunjukkan pada Tabel 4), dengan visualisasi lima fitur paling berpengaruh disajikan pada Gambar 4.

Penting untuk dicatat bahwa analisis SHAP mengukur dampak setiap fitur pada skala logaritmik, sesuai dengan transformasi log(1+indeks_kebahagiaan) yang diterapkan pada variabel target. Meskipun demikian, karena transformasi ini bersifat monotonik, arah pengaruh (positif atau negatif) dan peringkat kepentingan fitur yang ditunjukkan oleh SHAP tetap valid untuk diinterpretasikan.

Dengan pemahaman tersebut, hasil analisis mengungkapkan bahwa modal sosial dan kepercayaan merupakan prediktor paling dominan dalam model indeks kebahagiaan. Fitur-fitur seperti kepercayaan terhadap pemerintah daerah (R1002E), kesediaan berpartisipasi dalam aktivitas sipil (R1003G, R1003F), dan keyakinan pada dukungan komunitas (R1002A) secara konsisten menunjukkan hubungan positif yang signifikan. Nilai fitur yang lebih tinggi pada variabel-variabel ini berkorelasi dengan peningkatan *output* model (prediksi kebahagiaan yang lebih tinggi), dengan kepercayaan terhadap pemerintah daerah (R1002E) menunjukkan magnitudo dampak terbesar. Temuan ini menegaskan bahwa rasa kecukupan berfungsi sebagai salah satu pendorong positif kebahagiaan dalam model yang diuji.

D. *Web Development*



Gambar 5. Implementasi Pada *Website*

Sebagai implementasi praktis dari hasil penelitian, model *Gradient Boosting Regressor* yang terpilih diintegrasikan ke dalam sebuah aplikasi *web* interaktif. Aplikasi ini memungkinkan pengguna memasukkan data melalui antarmuka kuesioner seperti ditunjukkan pada Gambar 5 dan secara langsung menerima hasil prediksi skor Indeks Kebahagiaan

E. ***Diskusi***

Hasil pengujian menunjukkan adanya perbedaan kinerja yang signifikan antar model, baik dari sisi jumlah sampel maupun proporsi pembagian data. Model *ensemble* seperti *Gradient Boosting Regressor* dan *Random Forest Regressor* secara konsisten memberikan hasil prediktif yang lebih baik, sementara *Decision Tree Regressor* cenderung menghasilkan nilai R^2 negatif, menandakan ketidakmampuannya dalam menangkap pola data secara efektif. Oleh karena itu, analisis selanjutnya difokuskan pada *Gradient Boosting Regressor* yang menunjukkan kinerja terbaik dengan R^2 sebesar 0,696. Meskipun demikian, model ini masih belum mampu menjelaskan sekitar 30,4% varians dalam Indeks Kebahagiaan, mengindikasikan bahwa masih terdapat faktor lain yang belum terakomodasi dalam model dan mencerminkan kompleksitas multidimensional dari konsep kebahagiaan itu sendiri.

Dalam menginterpretasikan faktor penentu kebahagiaan, terdapat perbedaan penekanan antara hasil *feature importance* (Tabel 3) dan analisis SHAP (Gambar 4), yang wajar mengingat keduanya mengukur "kepentingan" fitur secara berbeda. *Feature importance* dari *Gradient Boosting Regressor* menyoroti 'Luas lantai bangunan tempat tinggal' (R1202) sebagai fitur utama berdasarkan kontribusi rata-rata terhadap penurunan *error* model secara global, sedangkan analisis SHAP lebih menonjolkan variabel kepercayaan dan modal sosial seperti 'Kepercayaan kepada Pemerintah Daerah' (R1002E), karena mengukur dampak aditif setiap fitur terhadap *output* prediksi individual. SHAP unggul dalam memberikan interpretasi lokal pada setiap prediksi, sedangkan *feature importance* bersifat agregat. Meskipun keduanya tumpang tindih, perbedaan ini mengingatkan bahwa hasil model bersifat asosiatif, bukan kausal. Selain itu, karena target variabel ditransformasi menggunakan logaritma (np.log1p), nilai SHAP merefleksikan dampak fitur pada skala logaritmik, sehingga meskipun arah pengaruh tetap valid, besaran dampaknya tidak dapat diartikan secara langsung pada skala asli.

Tabel 5. Perbendingan dengan penelitian terkait

Author	Dataset	Model	Preprocessing				Feature Selection	Evaluasi Metriks	PSO	SHAP	Akurasi	Implementasi Aplikasi
			Missing value	Encode Data	Normalization	Scaler						
[1]	HDI	Random Forest Regression	✓	✗	✗	✓	✗	✓	✗	✗	0,937	✗
[2]	Survei *Online*	Gradient Boosting Classifier	✗	✓	✗	✓	✓	✓	✗	✗	0,90	✗
[5]	Survei Kesehatan	LightGBM	✓	✓	✓	✓	✗	✓	✗	✗	0,986	✗
[23]	Data Survei	XGBoost	✓	✓	✗	✗	✓	✓	✗	✗	0,927	✗
[4]	Data Survei	Gradient Boosting	✓	✓	✗	✗	✓	✓	✗	✓	0,318	✗
Metode yang di usulkan	Data Survei	*Gradient Boosting Regressor*	✓	✓	✓	✓	✓	✓	✓	✓	0,696	✓

Berdasarkan Tabel 5, metode yang diusulkan dalam penelitian ini menunjukkan keunggulan komparatif dengan mengadopsi praktik terbaik dari studi-studi sebelumnya, khususnya dalam konteks prediksi berbasis data survei subjektif. Sejalan dengan pendekatan yang digunakan oleh [2], [23]dan [4], penelitian ini menerapkan model *Gradient Boosting Regressor* yang telah terbukti efektif dalam menangani data nonlinier dan kompleks. Namun, berbeda dari [1] dan [5] yang tidak menyertakan tahapan seleksi fitur secara eksplisit, penelitian ini mengintegrasikan proses seleksi fitur guna meningkatkan efisiensi komputasi dan mempertajam fokus model pada variabel-variabel yang paling berkontribusi terhadap indeks kebahagiaan.

Keunggulan penelitian ini tidak hanya terletak pada tahap pra-pemrosesan data yang robus meliputi penanganan nilai hilang, *encoding* kategorikal, normalisasi, dan *scaling* tetapi juga pada integrasi sinergis antara teknik interpretabilitas dan optimasi model. Serupa dengan penelitian [4], kami

menerapkan SHAP untuk mendapatkan wawasan kuantitatif yang mendalam terhadap pengaruh setiap fitur. Namun, keunikan studi ini terletak pada penggabungan analisis granular SHAP dengan algoritma optimasi *Particle Swarm Optimization* (PSO) untuk *tuning hyperparameter*. Pendekatan terpadu ini memperkuat validasi model secara adaptif dan objektif, yang membedakannya dari studi sebelumnya seperti [1] dan [2] yang umumnya terbatas pada peringkat fitur tanpa analisis mendalam maupun optimasi sistematis. Penggunaan metrik evaluasi komprehensif (R^2, MAE, RMSE) semakin menegaskan validitas hasil yang dicapai.

Keunggulan lain yang membedakan penelitian ini secara signifikan adalah implementasi praktis dari model ke dalam sebuah aplikasi *web* berbasis Flask. Integrasi ini memungkinkan visualisasi prediksi indeks kebahagiaan secara interaktif dan dinamis, menjembatani hasil analisis ilmiah dengan pengguna akhir seperti pemangku kebijakan dan masyarakat umum. Aspek ini belum banyak dieksplorasi dalam studi sebelumnya, yang umumnya berhenti pada tahap eksperimen akademik tanpa realisasi aplikatif. Dengan demikian, pendekatan yang diajukan tidak hanya unggul dari sisi teknis dan metodologis, tetapi juga memiliki nilai aplikatif yang tinggi sebagai alat bantu pengambilan keputusan yang berbasis data.

IV. KESIMPULAN

Penelitian ini berhasil mengembangkan model *Gradient Boosting Regressor* (GBR) yang dioptimasi menggunakan *Particle Swarm Optimization* (PSO) dan seleksi fitur untuk memprediksi Indeks Kebahagiaan 2021. Model mencapai nilai R^2 0,696 dengan 20 fitur terpilih, dan analisis SHAP mengonfirmasi modal sosial serta kepercayaan pemerintah sebagai prediktor kunci. Meskipun sistem telah diimplementasikan dalam aplikasi web Flask fungsional, performa model (R^2) masih belum melampaui beberapa studi terdahulu ($R^2 > 0,9$). Keterbatasan utama lainnya terletak pada cakupan data yang terbatas pada satu survei dan ketiadaan validasi lintas domain. Oleh karena itu, penelitian mendatang direkomendasikan untuk mengintegrasikan model *ensemble* lanjutan seperti stacking, memperluas himpunan data dengan fitur representatif, serta menguji generalisasi model pada domain berbeda guna meningkatkan akurasi dan adaptabilitas sistem.

UCAPAN TERIMA KASIH

Ucapan terima kasih yang tulus peneliti sampaikan kepada Program Studi Teknik Informatika, Universitas Logistik dan Bisnis Internasional, atas segala dukungan, arahan, dan fasilitas yang tak henti-hentinya peneliti terima dalam riset ini. Dedikasi terhadap keunggulan akademis dan inovasi yang ditunjukkan oleh program studi menjadi inspirasi besar bagi peneliti.

REFERENSI

[1] A. Jannani, N. Sael, and F. Benabbou, "Machine learning for the analysis of quality of life using the World Happiness Index and Human Development Indicators," *Mathematical Modeling and Computing*, vol. 10, no. 2, pp. 534–546, 2023, doi: 10.23939/mmc2023.02.534.

[2] N. Zhang *et al.*, "Prediction of adolescent subjective well-being: A machine learning approach," *Gen Psychiatr*, vol. 32, no. 5, Sep. 2019, doi: 10.1136/gpsych-2019-100096.

[3] M. A. Rohmaniar, R. Habibi, and S. F. Pane, "Pengaruh Metode Seleksi Fitur terhadap Akurasi Model SVM dalam Klasifikasi Customer Churn pada Perusahaan Telekomunikasi," *(IJAI) Indonesian Journal of Applied Informatics*, vol. 09, no. 01, pp. 94–103, 2024.

[4] E. Oparina *et al.*, "Machine learning in the prediction of human wellbeing," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-024-84137-1.

[5] A. Baba and K. Bunji, "Prediction of Mental Health Problem Using Annual Student Health Survey: Machine Learning Approach," *JMIR Ment Health*, vol. 10, 2023, doi: 10.2196/42420.

[6] U. Suchaini, W. P. S. Nugraha, Dwipayana I Kadek Dede, and S. A. Lestari, *Indeks Kebahagiaan 2021*. Badan Pusat Statistik RI, 2021.

[7] J. J. Palamar and A. Le, "Underreporting of drug use on a survey of electronic dance music party attendees," *Addiction Research and Theory*, vol. 28, no. 4, pp. 321–327, Jul. 2020, doi: 10.1080/16066359.2019.1653860.

[8] F. Abdullah, S. Fachri Pane, and R. Habibi, "Deteksi Emosi Pada Teks Berbahasa Indonesia Menggunakan Pendekatan Ensemble," *Jurnal Teknologi Terapan)* |, vol. 10, no. 2, 2024.

[9] N. H. Harani and C. Prianto, "Sentiment Analysis of Student Emotion During Online Learning Using Recurrent Neural Networks (RNN)," *International Journal of*

Information System & Technology Akreditasi, vol. 5, no. 3, pp. 299–307, 2021.

[10] N. Kalpourtzi, J. R. Carpenter, and G. Touloumi, "Handling Missing Values in Surveys With Complex Study Design: A Simulation Study," *J Surv Stat Methodol*, vol. 12, no. 1, pp. 105–129, Feb. 2024, doi: 10.1093/jssam/smac039.

[11] D. K. Lee, "Data transformation: A focus on the interpretation," *Korean J Anesthesiol*, vol. 73, no. 6, pp. 503–508, Dec. 2020, doi: 10.4097/kja.20137.

[12] J. Raymaekers and P. J. Rousseeuw, "Transforming variables to central normality," *Mach Learn*, vol. 113, no. 8, pp. 4953–4975, Aug. 2024, doi: 10.1007/s10994-021-05960-5.

[13] L. T. Quang, B. H. Baek, W. Yoon, S. K. Kim, and I. Park, "Comparison of Normalization Techniques for Radiomics Features From Magnetic Resonance Imaging in Predicting Histologic Grade of Meningiomas," *Investig Magn Reson Imaging*, vol. 28, no. 2, pp. 61–67, Jun. 2024, doi: 10.13104/imri.2024.0010.

[14] N. H. Harani and C. Prianto, "Penerapan algoritma Adaboost guna menentukan pola masuknya calon mahasiswa," *TRANSFORMTIKA*, vol. 18, no. 1, pp. 123–132, 2020.

[15] F. Özen, "Random forest regression for prediction of Covid-19 daily cases and deaths in Turkey," *Heliyon*, vol. 10, no. 4, Feb. 2024, doi: 10.1016/j.heliyon.2024.e25746.

[16] U. Singh, M. Rizwan, M. Alaraj, and I. Alsaidan, "A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments," *Energies (Basel)*, vol. 14, no. 16, Aug. 2021, doi: 10.3390/en14165196.

[17] B. Ramadhan and S. F. Pane, "Pengaruh Hyperparameter Tuning untuk Efektivitas pada Pendekatan Hybrid dalam Mendiagnosis Stres dan Depresi : Tinjauan Studi Literatur," *Jurnal Tekno Insentif*, vol. 18, no. 2, pp. 104–118, Dec. 2024, doi: 10.36787/jti.v18i2.1516.

[18] B. Zuhri and N. H. Harani, "Studi Literatur: Optimasi Algoritma Machine Learning Untuk Prediksi Penerimaan Mahasiswa Pascasarjana," *International Journal of Information System & Technology*, vol. 03, no. 05, pp. 299–307, 2019, [Online]. Available: https://ejurnalunsam.id/index.php/jicom/

[19] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.

[20] S. Rezaei Melal, M. Aminian, and S. M. Shekarian, "A machine learning method based on stacking heterogeneous ensemble learning for prediction of indoor humidity of greenhouse," *J Agric Food Res*, vol. 16, Jun. 2024, doi: 10.1016/j.jafr.2024.101107.

[21] A. Ahmad *et al.*, "Prediction of compressive strength of fly ash based concrete using individual and ensemble algorithm," *Materials*, vol. 14, no. 4, pp. 1–21, Feb. 2021, doi: 10.3390/ma14040794.

[22] P. Panicheva, L. Mararitsa, S. Sorokin, O. Koltsova, and P. Rosso, "Predicting subjective well-being in a high-risk sample of Russian mental health app users," *EPJ Data Sci*, vol. 11, no. 1, Dec. 2022, doi: 10.1140/epjds/s13688-022-00333-x.

[23] L. Zhang, "Subjective Well-Being Prediction Using Data Mining Techniques: Evidence from Chinese General Social Survey," *Applied and Computational Mathematics*, vol. 7, no. 4, p. 197, 2018, doi: 10.11648/j.acm.20180704.13.