Jurnal Ilmu Komputer dan Sistem Informasi (JIRSI) **e-ISSN 2830-3954**
Volume: 5, Nomor: 3, September 2026: 1124-1141 **p-ISSN 2830-6031**
https://jurnal.unity-academy.sch.id/index.php/jirsi/index

Klasifikasi Risiko *No-Show* Pasien Berdasarkan Karakteristik Penjadwalan dan Sosiodemografi Menggunakan Kombinasi *Instance Hardness Threshold* dan *Symbolic Regression*

Patient No-Show Risk Classification Based on Scheduling and Sociodemographic Characteristics Using a Combination of Instance Hardness Threshold and Symbolic Regression

Meidhika Nawa Sapta*[1], Arief Hermawan[2], Donny Avianto[3]
[1,2]Magister Teknologi Informasi, Fakultas Pascasarjana, Universitas Teknologi Yogyakarta
[3]Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Teknologi Yogyakarta
E-mail: [1]meidhikanawas@gmail.com, [2]ariefdb@uty.ac.id, [3]donny@uty.ac.id

Abstrak

Ketidakhadiran pasien pada janji medis terjadwal (no-show) menyebabkan inefisiensi sumber daya, gangguan kontinuitas perawatan, dan kerugian finansial dalam sistem pelayanan kesehatan. Dataset janji medis yang umumnya tidak seimbang (imbalanced) menyulitkan identifikasi kelas minoritas no-show secara akurat melalui pendekatan klasifikasi konvensional. Di sisi lain, sebagian besar penelitian terdahulu lebih berfokus pada upaya peningkatan akurasi prediksi melalui model black-box yang kompleks, sementara aspek interpretabilitas model terhadap pengambilan keputusan klinis masih kurang mendapatkan perhatian, sehingga model yang dihasilkan tidak mampu menghasilkan penjelasan yang dapat langsung ditindaklanjuti oleh tenaga kesehatan secara operasional. Penelitian ini mengusulkan pipeline klasifikasi dua tahap yang mengintegrasikan Instance Hardness Threshold (IHT) dan Symbolic Regression (SR) berbasis pemrograman genetika. Dataset yang digunakan adalah no-show Appointments Kaggle (N=110.527 rekaman, Vitória, Brasil) dengan 14 fitur sosiodemografi dan penjadwalan. IHT mengeliminasi instance mayoritas yang ambigu secara selektif, sementara SR menghasilkan formula matematika eksplisit yang dapat dibaca dan diterjemahkan langsung menjadi aturan prioritas intervensi tanpa memerlukan alat interpretabilitas tambahan. IHT mereduksi rasio ketidakseimbangan dari 3,95:1 menjadi 1,08:1 dengan menghapus 44.790 instance. SR mengidentifikasi waiting_days dan same_day_appt sebagai pasangan fitur terkritik, menghasilkan formula simbolik eksplisit dengan recall 92%, F1-Score 0,43, dan ROC-AUC 0,67 pada data uji, kompetitif terhadap Gradient Boosting terbaik (F1=0,43, ROC-AUC=0,73) dengan keunggulan interpretabilitas yang signifikan. Formula SR secara langsung dapat dioperasionalkan sebagai aturan prioritas pengiriman pengingat, pasien dengan waktu tunggu panjang yang bukan janji hari yang sama diprioritaskan untuk intervensi, sementara pasien janji hari yang sama menunjukkan risiko no-show mendekati nol. Penelitian lanjutan direkomendasikan mencakup eksplorasi multi-objective SR, penambahan fitur kontekstual, serta validasi pada dataset klinis Indonesia.

***Kata kunci**: no-show pasien, Instance Hardness Threshold, Symbolic Regression, interpretabilitas, machine learning kesehatan*

Abstract

Patient non-attendance at scheduled medical appointments (no-show) places significant burdens on healthcare systems, including resource inefficiency, disrupted care continuity, and

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

substantial financial losses. Medical appointment datasets are inherently imbalanced, making accurate identification of the no-show minority class challenging with conventional classification methods. Moreover, most prior studies have primarily focused on improving predictive accuracy through complex black-box models, while model interpretability for clinical decision-making has received comparatively little attention, leaving such models unable to produce explanations that healthcare personnel can directly act upon in operational settings. This study proposes a two-stage classification pipeline integrating Instance Hardness Threshold (IHT) and genetic programming-based Symbolic Regression (SR), applied to the no-show Appointments dataset from Kaggle (N=110,527 records, Vitória, Brazil) with 14 sociodemographic and scheduling features. IHT selectively eliminates ambiguous majority-class instances, while SR produces explicit mathematical formulas that can be read and translated directly into intervention priority rules without requiring additional interpretability tools. IHT reduced the class imbalance ratio from 3.95:1 to 1.08:1 by removing 44,790 instances. SR identified waiting_days and same_day_appt as the most critical feature pair, generating an explicit symbolic formula that achieved 92% recall, an F1-Score of 0.43, and a ROC-AUC of 0.67 on the test set, competitive against the best-performing Gradient Boosting baseline (F1=0.43, ROC-AUC=0.73) with markedly superior interpretability. The SR formula can be directly operationalized as a reminder dispatch priority rule, patients with longer waiting times who did not book same-day appointments are prioritized for intervention, while same-day appointment patients exhibit near-zero no-show risk. Future work is recommended to explore multi-objective SR, contextual feature integration, and validation on Indonesian clinical datasets.

***Keywords**: patient no-show, Instance Hardness Threshold, Symbolic Regression, interpretability, healthcare machine learning*

1. PENDAHULUAN

Sistem pelayanan kesehatan modern menghadapi tekanan ganda, meningkatnya permintaan layanan di satu sisi, dan keterbatasan kapasitas sumber daya di sisi lain. Dalam konteks manajemen penjadwalan medis, fenomena *no-show* yaitu ketidakhadiran pasien pada janji medis yang telah terjadwal tanpa pemberitahuan sebelumnya telah lama diidentifikasi sebagai salah satu sumber utama inefisiensi operasional fasilitas kesehatan [1]. Ketidakhadiran pasien secara signifikan mengganggu pemberian layanan kesehatan dan menyebabkan pemanfaatan sumber daya medis yang kurang optimal [2], [3], [4]. Selain berdampak pada efisiensi operasional, kemangkiran janji medis juga menyebabkan pasien kehilangan kesempatan untuk dirawat secara tepat waktu, yang pada akhirnya dapat membahayakan keselamatan serta memperburuk kondisi kesehatan mereka [1], [5], [6]. Secara global, tingkat *no-show* dilaporkan berkisar antara 12% hingga 30% bergantung pada jenis fasilitas, spesialisasi layanan, dan karakteristik demografi pasien [2]. Angka ini tidak hanya mewakili kerugian finansial langsung, tetapi juga berdampak pada antrian pasien lain yang membutuhkan layanan, serta menurunkan kualitas kontinuitas perawatan bagi pasien itu sendiri.

Dampak *no-show* bersifat multidimensi. Dari perspektif penyedia layanan, setiap kuota jadwal yang kosong akibat *no-show* menciptakan kapasitas terbuang yang tidak dapat dikompensasi secara *real-time*, mengakibatkan pengurangan *volume* layanan yang diberikan dan penurunan pendapatan operasional [7]. Sistem

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

kesehatan Amerika Serikat saja diperkirakan menanggung kerugian lebih dari USD 150 miliar per tahun akibat janji yang tidak ditepati [2]. Dari sisi pasien, *no-show* meningkatkan risiko perburukan kondisi medis akibat terputusnya rangkaian perawatan, serta memperpanjang waktu tunggu bagi pasien lain untuk mendapatkan kuota pengganti [1].

Upaya prediksi *no-show* menggunakan *machine learning* telah berkembang pesat dalam satu dekade terakhir. Banyak penelitian telah mengeksplorasi penggunaan algoritma *machine learning* untuk memprediksi probabilitas ketidakhadiran pasien dengan memanfaatkan rekam medis historis dan data penjadwalan [8], [9], [10]. Pemanfaatan teknik *machine learning* telah terbukti mampu memberikan alat pendukung keputusan klinis yang lebih kuat dalam memandu manajemen sumber daya rumah sakit [7], [11], [12]. Berbagai algoritma telah diujicobakan secara ekstensif oleh banyak peneliti, mulai dari *Logistic Regression*, *Decision Tree*, *Random Forest*, *Gradient Boosting*, hingga jaringan saraf tiruan.

Namun demikian, terdapat dua kesenjangan penelitian yang signifikan yang belum teratasi secara komprehensif. Pertama, sebagian besar penelitian belum secara serius menangani masalah ketidakseimbangan kelas (*class imbalance*) yang inheren pada *dataset* janji medis. Dalam kumpulan data layanan kesehatan di dunia nyata, kejadian pasien yang tidak hadir seringkali sangat minoritas dibandingkan dengan yang hadir, sehingga menimbulkan bias klasifikasi pada model prediktif yang cenderung memprioritaskan kelas mayoritas [13], [14], [15]. Oleh karena itu, penerapan teknik penyeimbangan data (*data balancing*) sangat krusial untuk meningkatkan kepekaan model dalam mengidentifikasi pola kelas minoritas dengan akurat [16], [17]. Kelas *no-show* (minoritas) biasanya hanya mewakili 12 - 30% dari total data, dan metode *resampling* yang umum digunakan seperti SMOTE atau *Random Undersampling* memiliki keterbatasan. SMOTE berpotensi menciptakan sampel sintetis yang tidak realistis, sementara *Random Undersampling* membuang informasi berharga secara acak. *Instance Hardness Threshold* (IHT) menawarkan pendekatan yang lebih cerdas dengan menghapus hanya *instance* mayoritas yang ambigu secara statistik, tetapi penerapannya dalam konteks prediksi *no-show* masih sangat terbatas [18].

Kedua, sebagian besar model yang dihasilkan bersifat *black-box* (akurat secara statistik tetapi tidak dapat diinterpretasi secara klinis). Klinisi dan manajer fasilitas kesehatan membutuhkan penjelasan yang transparan mengapa seorang pasien teridentifikasi berisiko tinggi *no-show*, bukan sekadar nilai probabilitas [19]. Kebutuhan akan model prediktif yang dapat diinterpretasi sangat mendesak di sektor medis, di mana para tenaga kesehatan serta pemangku kepentingan perlu memahami dasar penalaran suatu prediksi sebelum merumuskan tindakan intervensi. *Symbolic Regression* (SR) muncul sebagai pendekatan yang mampu menjembatani akurasi prediktif dengan interpretabilitas intrinsik melalui pencarian formula matematika eksplisit, namun penggunaannya dalam klasifikasi risiko *no-show* masih sangat jarang dieksplorasi [20], [21].

Studi yang paling relevan dengan penelitian ini adalah yang pertama mengombinasikan IHT dan SR untuk prediksi *no-show* diuji pada dua *dataset* rumah sakit Brasil (data *CT scan* tidak terpublikasi, N=8.371, tingkat *no-show* 6,65%; dan

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

subset dataset Kaggle wilayah Jardim Camburi, N=7.717, tingkat *no-show* 19,03%) [7]. Melalui skema *double z-fold cross-validation* dengan 100 replikasi, studi tersebut membandingkan SR dan IHT terhadap sejumlah algoritma klasifikasi (KNN, SVM) dan teknik *resampling* lain (SMOTE, RUS, NearMiss), serta melaporkan sensitivitas di atas 0,94 pada kedua *dataset*. Namun, evaluasi pada studi tersebut berhenti pada level metrik performa dan frekuensi seleksi fitur, formula matematika eksplisit yang dihasilkan SR tidak ditampilkan maupun diinterpretasikan secara klinis, sehingga potensi operasionalisasi langsung formula tersebut sebagai aturan intervensi belum tereksplorasi. *Dataset* yang digunakan juga berskala relatif kecil (kurang dari 12.500 rekaman per *dataset*) dan belum dibandingkan terhadap model *black-box* modern seperti *Gradient Boosting*. Sebagai konteks tambahan, sebuah pendekatan *deep learning* interpretatif telah dikembangkan untuk memprediksi risiko *no-show* di rumah sakit pediatri dengan tingkat kemangkiran sebesar 20%, kondisi ketidakseimbangan kelas yang mirip dengan *dataset* dalam penelitian ini [1], sementara pengujian algoritma *Gradient Boosting* tercatat mencapai akurasi hingga 78% pada data rawat jalan di Malaysia, meskipun orientasi pengembangannya belum mempertimbangkan aspek interpretabilitas model [5]. Lebih lanjut, penerapan metode SR berbasis *feature engineering automation tool* juga telah dibuktikan mampu menghasilkan model klinis yang akurat sekaligus interpretatif dari data rekam medis elektronik berdimensi tinggi [19].

Penelitian ini termotivasi oleh kebutuhan untuk menghadirkan solusi prediksi *no-show* yang tidak hanya akurat, tetapi juga dapat dipertanggungjawabkan secara ilmiah dan diimplementasikan secara praktis oleh tenaga kesehatan. Kombinasi IHT dan SR dipilih karena keduanya saling melengkapi: IHT membersihkan ruang latih dari *noise* statistik yang dapat menyebabkan SR menemukan formula yang salah, sementara SR menghasilkan *output* yang dapat langsung diterjemahkan menjadi kebijakan operasional misalnya, aturan prioritisasi pengingat berbasis formula matematis.

Berbeda dengan penelitian sebelumnya yang menekankan perbandingan metrik performa lintas algoritma dan teknik *resampling* [7], penelitian ini bertujuan membangun *pipeline* klasifikasi risiko *no-show* berskala penuh yang mengintegrasikan IHT sebagai mekanisme *undersampling* berbasis kesulitan *instance* dan SR sebagai model klasifikasi interpretatif, dengan penekanan khusus pada ekstraksi dan interpretasi klinis formula matematika yang dihasilkan SR. Penelitian ini juga mengevaluasi efektivitas IHT dalam mereduksi ketidakseimbangan kelas pada *dataset* janji medis nyata, menganalisis formula matematika yang dihasilkan SR guna mengidentifikasi kombinasi fitur yang paling diskriminatif dalam memprediksi *no-show*, serta membandingkan performa SR dengan model *baseline* (*Gradient Boosting*) dari perspektif akurasi prediktif dan interpretabilitas. Pada akhirnya, penelitian ini diharapkan mampu memberikan rekomendasi berbasis bukti untuk strategi intervensi klinis yang dapat diterapkan dalam upaya mengurangi tingkat *no-show* secara operasional.

Secara eksplisit, kontribusi ilmiah penelitian ini dapat dirumuskan dalam tiga poin yang secara spesifik memperluas temuan penelitian sebelumnya [7]. Pertama, penelitian ini menguji kombinasi IHT dan SR pada skala *dataset* yang jauh lebih besar (110.527 rekaman, dibandingkan 7.717 rekaman pada *subset* Kaggle

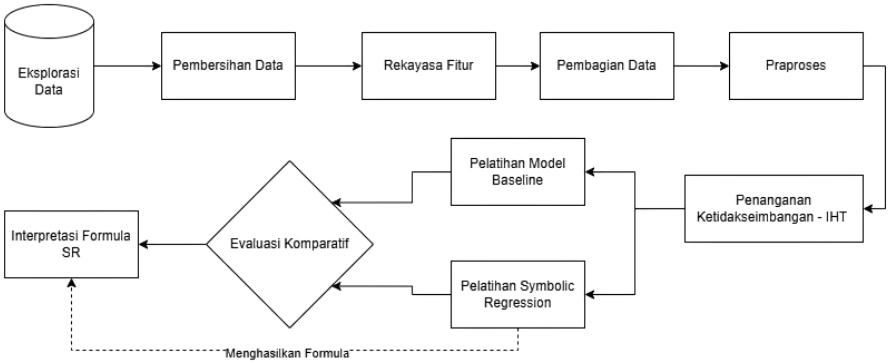
© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

yang digunakan pada studi rujukan), sekaligus secara eksplisit menampilkan dan menginterpretasikan formula matematika hasil SR secara klinis, alih-alih hanya melaporkan metrik performa dan frekuensi seleksi fitur sebagaimana dilakukan pada studi sebelumnya. Kedua, penelitian ini menawarkan kerangka evaluasi yang membandingkan model interpretatif (SR) dengan model *black-box* berperforma tinggi (*Gradient Boosting*), yang lebih merepresentasikan praktik pemodelan *machine learning* kontemporer dibandingkan KNN dan SVM yang menjadi pembanding pada penelitian sebelumnya. Ketiga, formula simbolik yang dihasilkan menyediakan artefak ilmiah konkret berupa aturan prioritas intervensi berbasis *waiting_days* dan *same_day_appt* yang dapat langsung diadopsi oleh fasilitas kesehatan tanpa memerlukan infrastruktur *machine learning* maupun alat interpretabilitas tambahan seperti SHAP bentuk operasionalisasi yang belum didemonstrasikan pada penelitian kombinasi IHT-SR sebelumnya.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan paradigma klasifikasi terbimbing (*supervised classification*). Kontribusi utama penelitian terletak pada penggabungan dua teknik yang saling melengkapi. IHT sebagai strategi perbaikan kualitas data latih, dan *Symbolic Regression* sebagai model klasifikasi yang mengutamakan interpretabilitas. Desain penelitian diarahkan bukan sekadar untuk membandingkan performa model secara numerik, melainkan untuk mendemonstrasikan bahwa model yang interpretatif dalam hal ini SR dapat memberikan nilai ilmiah dan klinis yang tidak dimiliki model *ensemble* berperforma tinggi.

Alur kerja penelitian dirancang secara bertahap dan sistematis mengikuti kaidah *machine learning pipeline* yang ketat guna menjamin validitas internal hasil eksperimen.



Gambar 1. Alur Kerja Penelitian Klasifikasi *No-Show* dengan IHT dan *Symbolic Regression*

Desain ini menempatkan SR sebagai metode utama yang diusulkan, sementara model *baseline* berfungsi sebagai tolok ukur perbandingan untuk mengkontekstualisasikan performa SR.

2.1. *Dataset*

Dataset yang digunakan dalam penelitian ini adalah *no-show Appointments* yang tersedia secara publik melalui *platform* Kaggle (*Medical Appointment No*

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

Show), mencakup 110.527 rekaman janji medis dari fasilitas kesehatan publik di kota Vitória, Brasil, yang dikumpulkan selama periode April - Juni 2016. *Dataset* ini telah digunakan secara luas dalam literatur *machine learning* bidang kesehatan dan menjadi *benchmark* yang representatif untuk penelitian prediksi *no*-show [7], [8], [22]. Distribusi kelas target menunjukkan 88.208 pasien hadir (79,8%) dan 22.319 pasien *no-show* (20,2%), menghasilkan rasio ketidakseimbangan sebesar 3,95:1.

Dataset ini mencakup 14 atribut asli, meliputi: *PatientId* dan *AppointmentID* (*identifier*), *Gender* (jenis kelamin), *ScheduledDay* (tanggal dan waktu penjadwalan), *AppointmentDay* (tanggal janji), *Age* (usia), *Neighbourhood* (lingkungan tempat tinggal, 81 kategori unik), serta atribut kondisi kesehatan biner: *Scholarship* (penerima bantuan sosial), *Hipertension*, *Diabetes*, *Alcoholism*, *Handcap*, *SMS_received*, dan variabel target *No-show*.

Tabel 1. Deskripsi Atribut *Dataset* Mentah

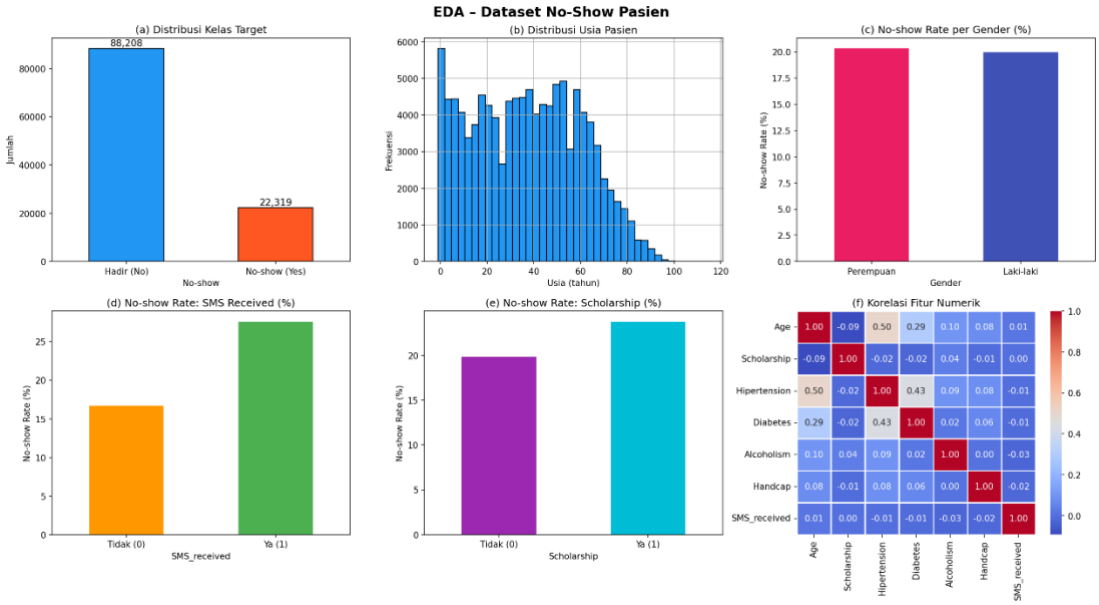
Nama Atribut	Tipe Data	Deskripsi / Keterangan
PatientId	Numerik	Nomor identifikasi unik untuk setiap pasien.
AppointmentID	Numerik	Nomor identifikasi unik untuk setiap pendaftaran janji temu.
Gender	Kategorikal	Jenis kelamin pasien (F = *Female*/Perempuan, M = *Male*/Laki-laki).
ScheduledDay	*Datetime*	Tanggal dan waktu ketika pasien mendaftarkan atau membuat janji temu.
AppointmentDay	*Datetime*	Tanggal hari pemeriksaan atau observasi medis pasien.
Age	Numerik	Usia pasien dalam hitungan tahun.
Neighbourhood	Kategorikal	Nama distrik atau lokasi tempat rumah sakit/klinik berada.
Scholarship	Biner	Indikator apakah pasien terdaftar dalam program bantuan sosial/beasiswa (0 = Tidak, 1 = Ya).
Hipertension	Biner	Indikator riwayat penyakit darah tinggi atau hipertensi (0 = Tidak, 1 = Ya).
Diabetes	Biner	Indikator riwayat penyakit diabetes (0 = Tidak, 1 = Ya).
Alcoholism	Biner	Indikator riwayat ketergantungan alkohol (0 = Tidak, 1 = Ya).
Handcap	Kategorikal	Tingkat disabilitas pasien (diukur dari rentang 0 hingga 4).
SMS_received	Biner	Indikator apakah pasien menerima pesan pengingat (SMS) untuk janji temunya (0 = Tidak, 1 = Ya).
No-show	Kategorikal	Status target kehadiran. "*No*" berarti pasien hadir, "*Yes*" berarti pasien batal hadir (*no-show*).

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

Untuk memberikan gambaran karakteristik data yang lebih rinci, tiga aspek berikut perlu digarisbawahi. Pertama, terkait nilai hilang (*missing value*), pemeriksaan kelengkapan data menunjukkan bahwa seluruh 14 atribut asli tidak memiliki satu pun nilai kosong (0% *missing*), sehingga tidak diperlukan prosedur imputasi pada data mentah, imputasi median/modus pada tahap praproses (Subbab 2.6) semata-mata berfungsi sebagai mekanisme pengaman (*safeguard*) terhadap kemungkinan nilai hilang pada data baru saat model diterapkan secara operasional. Kedua, *dataset* dibagi menjadi tiga *subset* menggunakan *stratified split* 70:15:15, menghasilkan 76.920 rekaman data latih, 16.483 rekaman data validasi, dan 16.483 rekaman data uji, dengan proporsi kelas *no-show* yang konsisten sebesar ±20,1% pada setiap *subset* (data latih: 61.455 hadir dan 15.465 *no-show*; data validasi dan data uji: masing-masing sekitar 13.169 hadir dan 3.314 *no-show*) rincian lebih lanjut disajikan pada Subbab 2.5. Ketiga, penerapan IHT pada data latih mengubah rasio ketidakseimbangan kelas dari 3,97:1 (61.455 hadir: 15.465 *no-show*) menjadi 1,08:1 (16.665 hadir: 15.465 *no-show*) melalui eliminasi 44.790 *instance* mayoritas yang ambigu, sebagaimana dibahas secara rinci pada Subbab 2.7 dan Subbab 3.1.

2.2. Analisis Data Eksploratif

Tahap EDA dilakukan untuk memahami karakteristik distribusi data sebelum *preprocessing*. Temuan utama EDA meliputi: (1) tidak ditemukan nilai *missing* pada seluruh kolom, menunjukkan kelengkapan *dataset* yang baik; (2) ditemukan 1 baris dengan nilai usia negatif (-1 tahun) yang merupakan anomali data; (3) terdapat 635 baris duplikat; (4) kolom *Handcap* memiliki distribusi ordinal 0 - 4 dengan nilai di atas 1 yang sangat jarang (< 0,2% total data); dan (5) korelasi antar fitur numerik bersifat rendah hingga sedang, tidak ditemukan multikolinearitas kritis.



Gambar 2. Analisis Eksploratif *Dataset No-Show Appointments*

2.3. Pembersihan Data (*Data Cleaning*)

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

Proses pembersihan data dilakukan secara sistematis melalui beberapa tahap. Pertama, kolom *PatientId* dan *AppointmentID* dihapus karena bersifat *identifier* unik yang tidak memiliki nilai prediktif dan berpotensi menyebabkan *data memorization*. Kedua, 1 baris dengan usia negatif dan 635 baris duplikat dieliminasi. Ketiga, 5 baris dengan *AppointmentDay* sebelum *ScheduledDay* dihapus sebagai anomali logis (janji tidak dapat terjadwal di masa lalu). Keempat, kolom *Handcap* dibinarisasi (0 = tidak ada keterbatasan, 1 = ada keterbatasan) untuk menstabilkan representasi fitur. Kelima, *label* target di-*encode*: *No* → 0 (hadir), *Yes* → 1 (*no-show*). *Dataset* bersih menghasilkan 109.886 rekaman dengan distribusi {Hadir: 87.793, *No-show*: 22.093}.

2.4. Rekayasa Fitur (*Feature Engineering)*

Kolom *ScheduledDay* dan *AppointmentDay* dalam format *datetime* ISO 8601 mengandung informasi temporal yang kaya namun perlu diekstraksi secara eksplisit. Mengacu pada literatur prediksi *no-show* [1], [5], [7], penelitian ini membuat tujuh fitur baru dari kedua kolom tersebut. Fitur pertama adalah *waiting_days*, yaitu selisih hari antara tanggal penjadwalan dan tanggal janji (rentang: 0 - 179 hari) yang dihitung dari komponen tanggal saja, bukan selisih *datetime* mentah guna menghindari nilai negatif akibat perbedaan waktu aktual. Dari kolom *ScheduledDay* diekstraksi tiga fitur tambahan: *scheduled_hour* (jam penjadwalan, skala 0 - 23), *scheduled_dayofweek* (hari dalam minggu saat penjadwalan, 0=Senin hingga 6=Minggu), dan *scheduled_month* (bulan penjadwalan, 1 - 12). Secara paralel, kolom *AppointmentDay* menghasilkan dua fitur: *appointment_dayofweek* dan *appointment_month*, yang merepresentasikan hari dan bulan saat janji berlangsung. Fitur ketujuh adalah *same_day_appt*, indikator biner bernilai 1 apabila *waiting_days* = 0 dan 0 apabila sebaliknya; sebanyak 38.477 rekaman (35,0% dari total) teridentifikasi sebagai janji hari yang sama. Setelah kolom *datetime* asli dihapus, *dataset* akhir memiliki 16 fitur prediktif dan satu variabel target dengan dimensi (109.886 × 17).

2.5. Pembagian Data *(Data Splitting)*

Untuk mencegah *data leakage*, pembagian data dilakukan sebelum tahap *preprocessing* apapun menggunakan strategi *stratified split* 70/15/15: 70% untuk data latih (76.920 rekaman), 15% untuk data validasi (16.483 rekaman), dan 15% untuk data uji (16.483 rekaman). Stratifikasi memastikan rasio *no-show* konsisten di ketiga *subset* (*no-show rate* ≈ 0,201 di seluruh *split*).

2.6. *Pipeline Preprocessing*

Pipeline preprocessing dibangun menggunakan *ColumnTransformer* dari *Scikit-learn*, dengan tiga jalur transformasi yang berbeda berdasarkan tipe fitur. Prinsip utama *preprocessor* di-*fit* hanya pada data latih, data validasi dan uji hanya melalui tahap *transform*. Tiga jalur transformasi adalah: (1) Fitur Numerik (7 kolom: *Age*, *waiting_days*, dan 5 fitur temporal): *SimpleImputer*(median) → *StandardScaler*; (2) Fitur Kategorikal (2 kolom: *Gender* dan *Neighbourhood*): *SimpleImputer*(modus) → *OneHotEncoder* (*handle_unknown*= '*ignore*', *drop*= '*if_binary*'), menghasilkan 82

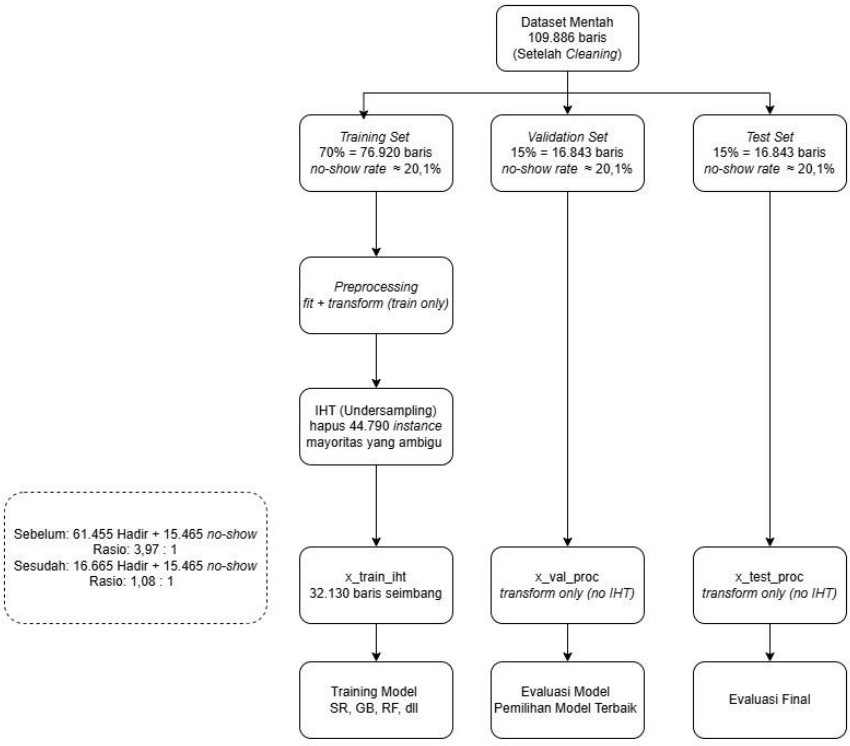
© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

kolom OHE; (3) Fitur Biner (7 kolom): hanya *SimpleImputer*(modus). Total fitur setelah *preprocessing*: 96 fitur.

2.7. *Instance Hardness Threshold (IHT)*

IHT merupakan teknik *undersampling* yang pertama kali diperkenalkan guna mengukur tingkat "kesulitan" setiap *instance* berdasarkan probabilitas kesalahan klasifikasi melalui *cross-validation* internal [18]. Berbeda dengan *Random Undersampling* yang membuang sampel mayoritas secara acak, IHT secara selektif menghapus *instance* mayoritas yang sulit diklasifikasikan (*hard instances*) yaitu *instance* yang berada di area *overlap* antar kelas. Melalui penyaringan *instance* yang tumpang tindih ini, IHT memungkinkan sistem analitik untuk mengesampingkan sampel-sampel yang mengganggu tanpa mereduksi representasi inti dari data yang sedang dipelajari [15], [18], [23].

Dalam penelitian ini, IHT dikonfigurasi dengan *Random Forest* (100 pohon) sebagai estimator internal untuk mengukur *instance hardness*, strategi *undersampling* pada kelas mayoritas (hadir), dan *5-fold cross-validation*. IHT hanya diterapkan pada data latih (*x_train_proc, y_train*) untuk menghindari kontaminasi data validasi dan uji.



Gambar 3. Penerapan IHT pada *Dataset No-Show Appointment*

2.8. Model *Symbolic Regression* (SR)

SR berbasis pemrograman genetika merupakan pendekatan *evolutionary computation* yang mencari ekspresi matematika simbolik optimal secara otomatis dari data. Dengan memanfaatkan metode optimasi evolusioner, SR mampu memformulasikan model matematis nonlinear yang tidak hanya transparan tetapi juga seringkali kompetitif dalam mengoptimalkan performa tanpa membebani

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

sistem dengan kompleksitas berlebih [21]. Berbeda dengan model *black-box*, SR menghasilkan formula eksplisit yang dapat dibaca, diverifikasi, dan diterjemahkan secara langsung oleh praktisi klinis [19].
Implementasi menggunakan *library gplearn* (*SymbolicClassifier*) dengan konfigurasi: populasi 2.000 individu, 30 generasi evolusi, *tournament selection* berukuran 20, fungsi matematika yang diizinkan mencakup {*add, sub, mul, div, sqrt, log, abs, neg, max, min*}, metrik optimisasi *log loss*, koefisien parsimoni 0,005 (penalti kompleksitas formula), serta probabilitas operator genetika: *crossover* 70%, *subtree mutation* 10%, *hoist mutation* 5%, *point mutation* 10%. Model dilatih pada data *x_train_iht* (hasil IHT) untuk mencegah SR mempelajari pola dari *instance* yang ambigu.

2.9. Model *Baseline* Pembanding

Sebagai pembanding, lima model *baseline* juga dilatih pada data yang sama (*x_train_iht*): *Logistic Regression* (*class_weight= 'balanced'*, C=1.0), *Random Forest* (300 pohon, *class_weight= 'balanced'*), *Gradient Boosting* (200 estimator, *learning_rate*=0,05, *max_depth*=5), *XGBoost* (300 estimator, *scale_pos_weight* proporsional), dan *LightGBM* (300 estimator, *num_leaves*=63, *class_weight= 'balanced'*). *Gradient Boosting* dipilih sebagai pembanding utama karena menunjukkan performa terbaik pada data validasi.

2.10. Metrik Evaluasi

Mengingat ketidakseimbangan kelas inheren, metrik evaluasi utama yang digunakan adalah *F1-Score* (*harmonic mean* antara *precision* dan *recall*), *Recall* (sensitivitas terhadap kelas *no-show*), *ROC-AUC* (kemampuan diskriminatif keseluruhan), dan *Precision*. Akurasi dilaporkan namun bukan metrik utama karena rentan terhadap dominasi kelas mayoritas. Evaluasi dilakukan pada tiga tahap, data validasi (pemilihan model), data uji (evaluasi final), dan analisis *confusion matrix* serta kurva ROC.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Penerapan IHT

Penerapan IHT secara dramatis mengubah distribusi data latih. Dari total 76.920 sampel awal (61.455 hadir + 15.465 *no-show*), IHT menghapus 44.790 instance mayoritas yang teridentifikasi sebagai *hard instances* yaitu sampel kelas hadir yang berada di wilayah *overlap* dengan kelas *no-show* dan konsisten salah diklasifikasi oleh *Random Forest* internal dalam proses 5-fold CV. Hasilnya adalah *dataset* latih berisi 32.130 sampel (16.665 hadir + 15.465 *no-show*) dengan rasio 1,08:1 mendekati seimbang sempurna, tanpa menambah sampel sintetis dan tanpa membuang informasi minoritas. Ini menunjukkan bahwa hampir 73% sampel kelas mayoritas memiliki karakteristik yang ambigu dan berpotensi menyesatkan model klasifikasi.

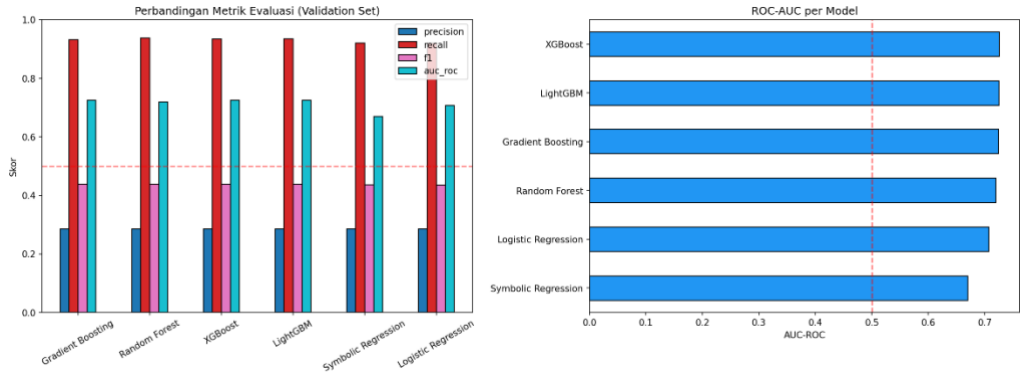
3.2. Performa Model pada Data Validasi

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

Seluruh model dilatih menggunakan data *post*-IHT dan dievaluasi pada data validasi yang tidak tersentuh proses IHT. Tabel 2 merangkum performa komparatif seluruh model.

Tabel 2. Perbandingan Metrik Evaluasi pada Data Validasi

Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***ROC-AUC***
Symbolic Regression	0,5204	0,2853	0,9203	0,4356	0,6701
Gradient Boosting	0,5175	0,2856	0,9327	0,4374	0,7247
XGBoost	0,5165	0,2854	0,9342	0,4372	0,7256
LightGBM	0,5154	0,2851	0,9357	0,4371	0,7250
Random Forest	0,5143	0,2850	0,9384	0,4373	0,7196
Logistic Regression	0,5183	0,2845	0,9212	0,4347	0,7073



Gambar 4. Perbandingan Metrik Evaluasi Seluruh Model pada Data Validasi

Beberapa temuan penting dapat ditarik dari Tabel 2. Pertama, seluruh model menunjukkan recall tinggi (> 0,92) untuk kelas *no-show*, mengindikasikan bahwa kombinasi IHT + data *post-resampling* efektif dalam menggeser sensitivitas model terhadap kelas minoritas. Kedua, perbedaan *F1-Score* antar model sangat kecil (0,4347–0,4374), menunjukkan bahwa setelah IHT diterapkan, kemampuan pemisahan kelas menjadi relatif setara di semua model, tanpa ada model yang mendominasi secara signifikan. Ketiga, meski SR memiliki ROC-*AUC* lebih rendah (0,6701) dibandingkan model *ensemble* (0,70–0,73), ini menandakan adanya *trade-off* interpretabilitas, diskriminasi yang dapat dibenarkan dalam konteks klinis. Keempat, *Gradient Boosting* menempati posisi terbaik dalam *F1-Score* (0,4374) di antara model *baseline*, menjadikannya referensi komparatif utama.

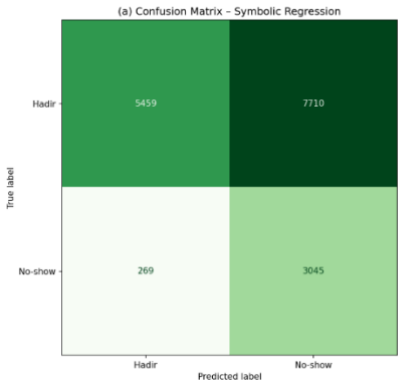
© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

3.3. Evaluasi Final pada Data Uji

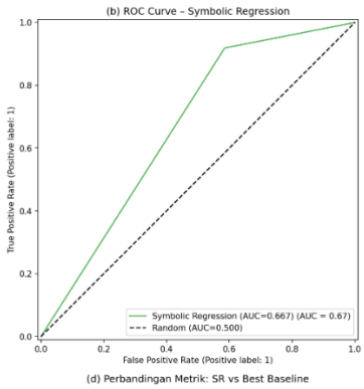
Evaluasi final dilakukan pada data uji yang sepenuhnya tersegregasi dari proses latih dan validasi. Tabel 3 menyajikan perbandingan detail antara SR dan *Gradient Boosting* sebagai model *baseline* terbaik.

Tabel 3. Perbandingan Performa SR dan *Gradient Boosting* pada Data Uji

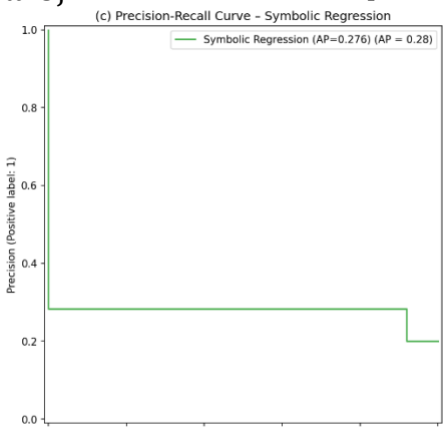
Metrik	***Symbolic Regression***	***Gradient Boosting***
Accuracy	0,52	0,51
Precision (no-show)	0,28	0,28
Recall (no-show)	0,92	0,93
F1-Score (no-show)	0,43	0,43
ROC-AUC	0,6667	0,7250
Interpretabilitas	Tinggi (formula)	Rendah *(black-box)*



Gambar 5. *Confusion Matrix Symbolic Regression* pada Data Uji



Gambar 6. Kurva *ROC Symbolic Regression* pada Data Uji (*AUC* = 0,667)



Gambar 7. Kurva *Precision-Recall Symbolic Regression* pada Data Uji

Dari *confusion matrix* SR (Gambar 5), dapat dilihat bahwa model berhasil mengidentifikasi 3.045 dari 3.314 pasien *no-show* yang sebenarnya (*True Positive* tinggi), dengan biaya 7.710 pasien hadir yang salah diprediksi sebagai *no-show*

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

(*False Positive*). Pola ini mencerminkan preferensi model terhadap *recall* yang tinggi sebuah perilaku yang dikehendaki dalam konteks klinis, di mana lebih baik mengirimkan pengingat kepada pasien yang sebenarnya akan hadir (*False Positive* yang ringan dan tidak berbahaya) daripada melewatkan pasien yang benar-benar akan absen (*False Negative* yang berpotensi merugikan).

3.4. Analisis Formula *Symbolic Regression*

Salah satu kontribusi paling penting dari pendekatan *symbolic regression* (SR) adalah diperolehnya formula matematika eksplisit melalui 30 generasi evolusi dengan populasi sebanyak 2.000 individu. Formula akhir memiliki kompleksitas 15 *node* dan *raw fitness* (*log loss*) sebesar 0,1786, yaitu:

$$f(x) = \frac{4.\ waiting_days}{same_day_appt}$$

di mana fungsi pembagian terproteksi didefinisikan sebagai:

$$div_protected(a,b) = \begin{cases} a/b, & \text{jika } |b| > 0.001 \\ 1, & \text{jika } |b| \leq 0.001 \end{cases}$$

Dalam ruang fitur hasil praproses yang terdiri atas 96 fitur, *waiting_days* berada pada indeks X1 sebagai fitur numerik kedua setelah *Age*, sedangkan *same_day_appt* berada pada indeks X95 sebagai fitur biner terakhir. Dalam notasi *gplearn*, formula tersebut pada awalnya dituliskan sebagai:

$$add\Big(add\Big(add\big(div(X1,X95), div(X1,X95)\big), div(X1,X95)\Big), div(X1,X95)\Big)$$

yang secara aljabar ekuivalen dengan:

$$4\ x\ div(X_{1,}X_{95})$$

3.5. Interpretasi Perilaku Formula

Formula SR menghasilkan dua skenario skor yang berbeda berdasarkan nilai *same_day_appt*:

Skenario A, janji hari yang sama (*same_day_appt* = 1, *waiting_days* = 0):

$$f(x) = 4 \times (0/1) = 0.$$

Skor mendekati nol mengindikasikan risiko *no-show* yang sangat rendah. Pasien yang mendaftar dan langsung mendatangi janji pada hari yang sama memiliki keterikatan tinggi dan hambatan logistik minimal.

Skenario B, janji terjadwal (*same_day_appt* = 0, *waiting_days* > 0):

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

$$f(x) = 4 \times div_protected(waiting_days, 0) = 4 \times 1 = 4$$

Skor konstanta tinggi mengindikasikan risiko *no-show* yang *elevated* untuk seluruh pasien dengan janji terjadwal ke depan. Meskipun *waiting_days* aktual tidak muncul secara eksplisit dalam Skenario B akibat sifat pembagian terproteksi, keberadaannya dalam formula mengkonfirmasi perannya sebagai fitur diskriminatif utama yang diidentifikasi oleh evolusi genetika.

3.6. Signifikansi Klinis Formula

Formula SR memberikan tiga *insight* klinis yang dapat ditindaklanjuti secara langsung. Pertama, *same_day_appt* adalah prediktor terkuat: pasien yang membuat dan mengonfirmasi janji pada hari yang sama memiliki motivasi intrinsik dan mobilitas akses yang memadai mereka tidak memerlukan intervensi tambahan. Kedua, *waiting_days* merupakan prediktor kedua yang dikonfirmasi oleh SR sebagai variabel kritis, penelitian terdahulu secara konsisten menunjukkan bahwa semakin panjang jeda antara waktu penjadwalan dan waktu janji, semakin tinggi probabilitas *no-show*. Ketiga, ketiadaan fitur demografi (usia, gender, kondisi medis) dalam formula SR mengisyaratkan bahwa karakteristik *scheduling behavior* lebih prediktif daripada profil demografi dalam konteks *dataset* ini sebuah temuan yang konsisten dengan argumen penelitian terdahulu bahwa fitur penjadwalan adalah prediktor dominan.

3.7. Diskusi: *Trade-off* Interpretabilitas vs. Diskriminasi

Perbedaan *ROC-AUC* antara SR (0,667) dan *Gradient Boosting* (0,725) sebesar 0,058 merepresentasikan *trade-off* yang menjadi inti dari paradigma *interpretable machine learning* [24]. Dari perspektif klinis, pemilihan model bukan semata-mata fungsi dari *ROC-AUC* tertinggi, melainkan dari relevansi praktisnya. SR unggul dalam tiga dimensi kritis: (1) Transparansi formula yang dihasilkan dapat dibaca, dipahami, dan diaudit tanpa memerlukan *tools* interpretabilitas tambahan seperti SHAP, (2) *Deployability* formula dapat diimplementasikan langsung dalam sistem informasi manajemen rumah sakit tanpa memerlukan infrastruktur *machine learning* yang kompleks, (3) Regulasi pedoman FDA dan badan regulasi kesehatan internasional semakin mensyaratkan transparansi model dalam aplikasi klinis [20].

Di sisi lain, penurunan *ROC-AUC* pada SR menunjukkan bahwa model kehilangan sebagian kemampuan diskriminatif yang dimiliki oleh *ensemble* terutama dalam kasus-kasus *borderline* di mana kombinasi fitur kompleks diperlukan untuk membedakan hadir vs. *no-show*. Dalam penelitian ini, *F1-Score* yang setara (0,43) mengindikasikan bahwa pada ambang keputusan standar (0,5), SR dan *Gradient Boosting* menghasilkan akurasi diagnosis yang praktis identik. Ini mendukung argumentasi bahwa SR merupakan pilihan yang tepat ketika interpretabilitas diprioritaskan.

KESIMPULAN

Penelitian ini berhasil membangun dan memvalidasi sebuah *pipeline* klasifikasi risiko *no-show* pasien yang mengintegrasikan *Instance Hardness*

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

Threshold (IHT) dan *Symbolic Regression* (SR) menggunakan *dataset* janji medis publik berskala besar. Hasil eksperimen membuktikan bahwa IHT sangat efektif bekerja sebagai teknik *undersampling* cerdas yang mampu menyeimbangkan rasio kelas secara masif melalui eliminasi *instance* mayoritas yang ambigu, tanpa harus merekayasa data sintetis ataupun membuang informasi berharga pada kelas minoritas secara acak.

Melalui komputasi evolusioner, model SR berhasil mengekstraksi formula matematika eksplisit yang secara cermat mengidentifikasi fitur *waiting_days* dan *same_day_appt* sebagai pasangan prediktor paling kritis. Formula matematis tersebut memiliki nilai aplikatif tinggi karena dapat langsung ditranslasikan menjadi aturan kebijakan operasional intervensi klinis di fasilitas kesehatan tanpa membutuhkan instrumen interpretabilitas tambahan seperti SHAP.

Dari aspek performa prediktif, model SR menunjukkan nilai kapabilitas dan sensitivitas kemangkiran yang sangat kompetitif dan setara secara praktis dengan model ansambel tingkat tinggi (*Gradient Boosting*), dengan *trade-off* penurunan nilai *ROC-AUC* yang sangat minimal demi mendapatkan keunggulan transparansi model yang absolut sebuah perbandingan langsung terhadap model *black-box* modern yang belum dilakukan pada penelitian kombinasi IHT-SR sebelumnya. Dominasi fitur temporal ini mengonfirmasi bahwa karakteristik *scheduling behavior* jauh lebih diskriminatif dalam mendeteksi risiko *no-show* pasien dibandingkan dengan profil sosiodemografi maupun riwayat kondisi medis internal. Selain itu, validitas metodologi dan keandalan evaluasi dalam penelitian ini sepenuhnya terjamin melalui penerapan arsitektur *pipeline* yang ketat guna mencegah terjadinya kebocoran data (*data leakage*).

Berdasarkan temuan serta keterbatasan yang ditemukan dalam penelitian ini, beberapa arah pengembangan direkomendasikan untuk penelitian lanjutan. Dari sisi metodologi, proses SR dalam penelitian ini hanya mengoptimalkan satu metrik tunggal yaitu *log loss*, yang berpotensi menghasilkan formula yang bias terhadap satu jenis kesalahan klasifikasi. Oleh karena itu, penerapan *multi-objective* SR yang mengoptimalkan akurasi, interpretabilitas, dan panjang formula secara simultan berpeluang menghasilkan solusi Pareto yang lebih seimbang. Selain itu, skema validasi dalam penelitian ini menggunakan *stratified split* tunggal (70:15:15), Skema validasi tunggal ini berpotensi belum sepenuhnya menangkap variabilitas performa model akibat komposisi acak data latih, validasi, dan uji, sehingga penerapan skema *cross-validation* berulang pada penelitian mendatang direkomendasikan untuk memperkuat generalisabilitas dan stabilitas hasil. Sejalan dengan itu, konfigurasi arsitektur evolusi SR dalam penelitian ini masih bersifat *general-purpose*, sehingga pencarian hiperparameter secara sistematis mencakup ukuran populasi, jumlah generasi, koefisien parsimoni, dan variasi *function set* menggunakan metode *Bayesian optimization* berpotensi meningkatkan daya diskriminatif model SR hingga mendekati performa model ansambel *black-box*. Sebagai alternatif metodologis, eksplorasi menggunakan *library* SR modern berbasis arsitektur nonevolusioner seperti *PySR, FEAT*, atau *SymScore* juga dapat dipertimbangkan, mengingat pendekatan tersebut berpotensi menemukan struktur formula yang lebih kompleks dan akurat namun tetap mempertahankan sifat interpretabilitas intrinsiknya.

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

Dari sisi data dan generalisabilitas, penelitian mendatang dapat memperkaya daya diskriminasi model SR melalui penambahan variabel kontekstual eksternal. Integrasi data cuaca lokal, jarak geografis tempat tinggal pasien ke fasilitas kesehatan, riwayat *no-show* pada janji medis sebelumnya, serta data sosioekonomi yang lebih granular berpotensi meningkatkan akurasi prediksi secara signifikan. Di samping itu, *dataset* yang digunakan dalam penelitian ini berasal dari konteks pelayanan kesehatan Brasil, sehingga validasi *pipeline* kombinasi IHT dan SR pada data janji medis riil di berbagai fasilitas kesehatan di Indonesia sangat direkomendasikan guna menguji kelayakan transfer metodologi, mengingat adanya perbedaan karakteristik demografi, pola penjadwalan, serta sistem penjaminan asuransi kesehatan antar kedua negara.

Dari perspektif implementasi jangka pendek, karakteristik formula SR yang sederhana, deterministik, dan transparan menjadikannya sangat ideal untuk diintegrasikan sebagai modul *Clinical Decision Support* (CDS) dalam Sistem Informasi Manajemen Rumah Sakit (SIMRS). Penelitian implementasi lanjutan diperlukan untuk mengevaluasi dampak nyata penerapan formula ini terhadap penurunan angka *no-show* pasien dan peningkatan efisiensi operasional secara empiris di lingkungan klinis yang sesungguhnya.

Secara lebih spesifik pada konteks Indonesia, formula SR yang dihasilkan berpotensi diadopsi sebagai modul pendukung keputusan sederhana dalam Sistem Informasi Manajemen Rumah Sakit (SIMRS) di rumah sakit maupun puskesmas, mengingat fasilitas kesehatan di Indonesia pada umumnya belum memiliki infrastruktur *machine learning* yang memadai, namun tetap dapat mengimplementasikan aturan berbasis formula matematis sederhana (empat kali *waiting_days* dibagi *same_day_appt*) melalui *query* basis data maupun *spreadsheet* tanpa investasi teknologi tambahan yang signifikan. Sebagai ilustrasi manfaat operasional, dengan mengacu pada tingkat *no-show* rumah sakit yang secara umum dilaporkan berkisar antara 12-30%, penerapan strategi prioritisasi pengingat berbasis formula SR pada segmen pasien berisiko tinggi (janji terjadwal dengan *waiting_days* panjang) berpotensi meningkatkan efektivitas intervensi pengingat dibandingkan pengiriman pengingat secara seragam (*blanket reminder*) ke seluruh pasien, sehingga berpeluang menghemat sumber daya administratif rumah sakit. Dengan sensitivitas (*recall*) formula sebesar 92% dalam mengidentifikasi pasien berisiko *no-show* pada data uji, penerapan strategi intervensi bertarget hanya pada pasien yang teridentifikasi berisiko berpotensi menekan proporsi slot jadwal yang terbuang akibat ketidakhadiran secara lebih efisien dibandingkan strategi pengingat seragam yang selama ini umum diterapkan. Meskipun demikian, estimasi kuantitatif besaran penurunan angka *no-show* secara nyata memerlukan validasi lapangan pada data rumah sakit di Indonesia, mengingat adanya perbedaan karakteristik demografi, pola penjadwalan, dan sistem penjaminan kesehatan dengan *dataset* Brasil yang digunakan dalam penelitian ini.

DAFTAR PUSTAKA

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

[1] D. Liu *et al.*, "Machine Learning Approaches to Predicting No-Shows in Pediatric Medical Appointment," *NPJ Digit. Med.*, vol. 5, no. 1, pp. 1–11, Dec. 2022, doi: 10.1038/s41746-022-00594-w.

[2] Q. A. Hathaway, N. Yanamala, T. Narumanchi, and J. Narumanchi, "Prediction of Pediatric No-Show Rates Before and During the COVID-19 Pandemic Utilizing Machine Learning," *medRxiv*, pp. 1–2, Nov. 2024, doi: 10.1101/2024.11.23.24317836.

[3] D. Valero-Bover *et al.*, "Reducing Non-Attendance in Outpatient Appointments: Predictive Model Development, Validation, and Clinical Assessment," *BMC Health Serv. Res.*, vol. 22, no. 451, pp. 1–9, Dec. 2022, doi: 10.1186/s12913-022-07865-y.

[4] A. Russotto *et al.*, "Understanding No-Show Patterns in Healthcare: A Retrospective Study from Northern Italy," *Healthcare (Switzerland)*, vol. 13, no. 15, pp. 1–10, Jul. 2025, doi: 10.3390/HEALTHCARE13151869.

[5] A. F. A. Hamdan and A. A. Bakar, "Machine Learning Predictions on Outpatient No-Show Appointments in a Malaysia Major Tertiary Hospital," *Malaysian Journal of Medical Sciences*, vol. 30, no. 5, pp. 169–180, Oct. 2023, doi: 10.21315/mjms2023.30.5.14.

[6] A. M. Husein and M. K. Harahap, "Teknik Data Science di Bidang Kesehatan: Menganalisis Faktor penting Ketidakhadiran Janji Medis Pasien," *Data Sciences Indonesia (DSI)*, vol. 1, no. 2, pp. 50–61, Jan. 2022, doi: 10.47709/dsi.v1i2.1314.

[7] C. Deina, F. S. Fogliatto, G. J. C. da Silveira, and M. J. Anzanello, "Decision Analysis Framework for Predicting No-Shows to Appointments Using Machine Learning Algorithms," *BMC Health Serv. Res.*, vol. 24, no. 37, pp. 1–17, Dec. 2024, doi: 10.1186/s12913-023-10418-6.

[8] L. H. A. Salazar, V. R. Q. Leithardt, W. D. Parreira, A. M. da Rocha Fernandes, J. L. V. Barbosa, and S. D. Correia, "Application of Machine Learning Techniques to Predict A Patient's No-Show in The Healthcare Sector," *Future Internet*, vol. 14, no. 3, pp. 1–21, Dec. 2022, doi: 10.3390/fi14010003.

[9] J. Joseph, S. Senith, A. Alfred Kirubaraj, and S. R. Jino Ramson, "Machine Learning for Prediction of Clinical Appointment No-Shows," *International Journal of Mathematical, Engineering and Management Sciences*, vol. 7, no. 4, pp. 558–574, Aug. 2022, doi: 10.33889/IJMEMS.2022.7.4.036.

[10] G. Fan, Z. Deng, Q. Ye, and B. Wang, "Machine Learning Based Prediction Models for Patients No-Show in Online Outpatient Appointments," *Data Science and Management*, vol. 2, pp. 45–52, Jun. 2021, doi: 10.1016/j.dsm.2021.06.002.

[11] A. Alshammari, R. Almalki, and R. Alshammari, "Developing a Predictive Model of Predicting Appointment No-Show by Using Machine Learning Algorithms," *Journal of Advances in Information Technology*, vol. 12, no. 3, Aug. 2021.

[12] K. M. Toffaha, M. C. E. Simsekler, M. A. Omar, and I. ElKebbi, "Predicting Patient No-Shows Using Machine Learning: A Comprehensive Review and Future Research Agenda," *Intell. Based. Med.*, vol. 11, pp. 2666–5212, Feb. 2025, doi: 10.1016/j.ibmed.2025.100229.

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

[13] E. F. Agyemang *et al.*, "Addressing Class Imbalance Problem in Health Data Classification: Practical Application From an Oversampling Viewpoint," *Applied Computational Intelligence and Soft Computing*, vol. 2025, no. 1, pp. 1–0, 2025, doi: 10.1155/acis/1013769.

[14] S. Al-Ahmari and F. Nadeem, "Improving Surgical Site Infection Prediction Using Machine Learning: Addressing Challenges of Highly Imbalanced Data," *Diagnostics*, vol. 15, no. 4, pp. 1–34, Feb. 2025, doi: 10.3390/diagnostics15040501.

[15] M. Yadav Mamilla, R. Al-Haddad, and S. Chowdhury, "Resampling Imbalanced Healthcare Data for Predictive Modelling," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 16, no. 2, p. 2025, 2025, [Online]. Available: www.ijacsa.thesai.org. [Accessed: Apr. 23, 2026]

[16] A. Malik and C. Tepe, "Improving Early Diabetes Diagnosis, A Machine Learning Approach for Class Imbalance Mitigation," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 11, no. 3, pp. 861–874, Jun. 2025.

[17] J. Dunstan *et al.*, "Predicting No-Show Appointments in A Pediatric Hospital in Chile Using Machine Learning," *Health Care Manag. Sci.*, vol. 26, no. 2, pp. 313–329, Jan. 2023, doi: 10.1007/s10729-022-09626-z.

[18] M. R. Smith, T. Martinez, and C. Giraud-Carrier, "An instance Level Analysis of Data Complexity," *Mach Learn*, vol. 95, no. 2, pp. 225–256, 2014, doi: 10.1007/s10994-013-5422-z.

[19] W. G. La Cava *et al.*, "A Flexible Symbolic Regression Method for Constructing Interpretable Clinical Prediction Models," *NPJ Digit. Med.*, vol. 6, no. 107, pp. 1–14, Dec. 2023, doi: 10.1038/s41746-023-00833-8.

[20] F. O. de Franca *et al.*, "Interpretable Symbolic Regression for Data Science: Analysis of the 2022 Competition," pp. 1–13, Jul. 2023, [Online]. Available: http://arxiv.org/abs/2304.01117. [Accessed: Jun. 05, 2026]

[21] D. Ferrari, V. Guidetti, Y. Wang, and V. Curcin, "Multi-objective Symbolic Regression to Generate Data-driven, Non-fixed Structure and Intelligible Mortality Predictors using EHR: Binary Classification Methodology and Comparison with State-of-the-art," Germany, 2022. [Online]. Available: https://github.com/davideferrari92/multiobjective_symbolic_regression. [Accessed: Jun. 05, 2026]

[22] L. H. A. Salazar, W. D. Parreira, A. M. da R. Fernandes, and V. R. Q. Leithardt, "No-Show in Medical Appointments with Machine Learning Techniques: A Systematic Literature Review," *Information (Switzerland)*, vol. 13, no. 11, pp. 1–13, Nov. 2022, doi: 10.3390/info13110507.

[23] A. Chatzimparmpas, F. V. Paulovich, and A. Kerren, "HardVis: Visual Analytics to Handle Instance Hardness Using Undersampling and Oversampling Techniques," *Computer Graphics Forum*, vol. 42, no. 1, pp. 135–154, Feb. 2023, doi: 10.1111/cgf.14726.

[24] Y. Alabdulkarim, M. Almukaynizi, A. Alameer, B. Makanati, R. Althumairy, and A. Almaslukh, "Predicting No-Shows for Dental Appointments," *PeerJ Comput. Sci.*, vol. 8, pp. 1–24, Nov. 2022, doi: 10.7717/PEERJ-CS.1147.

© 2026 The Author(s). Published by UNITY ACADEMY . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)