INJOPHARS (Innovation Journal of Pharmaceutical Science)
Vol. xx, No. x, November 2026, pp. 49-57
DOI: 10.68300/injophars.v2i2.112
E-ISSN 3108-9542

*Research Article Open Access (CC–BY-SA)*

# Deteksi Dini Non-Persistensi Pengobatan di Apotek Menggunakan Risk-Calibrated and Explainable Ensemble Learning

# *Early Detection of Non-Persistence with Medication in Pharmacies Using Risk-Calibrated and Explainable Ensemble Learning*

**Wahyuni Azis[1]; Pertiwi Ishak[2]; Sustrin Abasa[3]; Firman Aziz[4,*]**
[1,3] *Branch Educatian and Training Center, Makassar 90155, Indonesia*
[2, 4] *Universitas Pancasakti, Makassar 90121, Indonesia*
[1] *wahyuni.betc@gmail.com*; [2] *pertiwi.ishak@unpacti.ac.id*; [3]*sustrin.abasa@unpacti.ac.id*; [4] *firman.aziz@unpacti.ac.id*
** Corresponding author*

**Abstrak**

Non-persistensi pengobatan merupakan tantangan dalam pelayanan farmasi karena dapat menghambat keberhasilan terapi. Penelitian ini mengevaluasi empat algoritma untuk klasifikasi status persistensi pengobatan pada dataset Data Fix.csv yang berisi 3.424 catatan awal dan 68 kolom. Setelah enam baris duplikat dihapus, tersisa 3.418 observasi. Model yang dibandingkan adalah Decision Tree, Random Forest, Extra Trees, dan Gradient Boosting. Pada validasi silang berstrata lima lipatan, Gradient Boosting memperoleh ROC-AUC rata-rata 0,851 dan F1-score 0,838. Pada data uji, model ini memperoleh accuracy 0,776, precision 0,794, recall 0,866, F1-score 0,828, dan ROC-AUC 0,830. Kalibrasi sigmoid memperbaiki Brier score secara kecil dari 0,15823 menjadi 0,15774, tetapi log loss sedikit memburuk dari 0,48920 menjadi 0,48945; recall turun dari 0,866 menjadi sekitar 0,850. Analisis SHAP menunjukkan kontribusi relatif terbesar dari indikator terapi obat jangka panjang, kunjungan untuk pemeriksaan, kunjungan umum, kunjungan imunisasi, dan kategori spesialisasi. Temuan ini menunjukkan bahwa kinerja klasifikasi dan kualitas probabilitas perlu dinilai secara terpisah. Karena waktu ketersediaan seluruh fitur belum dapat diverifikasi dan validasi eksternal belum dilakukan, hasil ini hanya mendukung klasifikasi pada dataset yang dianalisis dan belum membuktikan kesiapan deteksi dini klinis.

**Kata Kunci:** non-persistensi, ensemble learning, kalibrasi probabilitas, SHAP, pembelajaran mesin, pelayanan farmasi.

***Abstract***

*Medication non-persistence is a challenge in pharmacy services because it may compromise treatment outcomes. This study evaluates four algorithms for medication persistence status classification using Data Fix.csv, which initially contained 3,424 records and 68 columns. After six duplicate rows were removed, 3,418 observations remained. The compared models were Decision Tree, Random Forest, Extra Trees, and Gradient Boosting. In stratified five-fold cross-validation, Gradient Boosting achieved a mean ROC-AUC of 0.851 and F1-score of 0.838. On the held-out test set, it achieved an accuracy of 0.776, precision of 0.794, recall of 0.866, F1-score of 0.828, and ROC-AUC of 0.830. Sigmoid calibration slightly improved the Brier score from 0.15823 to 0.15774, while log loss slightly worsened from 0.48920 to 0.48945; recall decreased from 0.866 to approximately 0.850. SHAP analysis ranked indicators of long-term drug therapy, screening-related encounters, general examinations, immunization encounters, and specialty category among the largest contributors to model output. These findings show that classification discrimination and probability quality must be evaluated separately. Because the temporal availability of all features has not been verified and external validation has not been performed, the results support classification on the analyzed dataset but do not establish readiness for clinical early detection.*

***Keywords:*** *non-persistence; ensemble learning; probability calibration; SHAP; machine learning; pharmacy services.*

## Pendahuluan

Persistensi pengobatan (medication persistence) merupakan salah satu aspek penting dalam keberhasilan terapi, khususnya pada pasien yang memerlukan pengobatan dalam jangka panjang [1]. Persistensi mengacu pada lamanya pasien melanjutkan terapi sejak pengobatan dimulai hingga penghentian terapi, sedangkan kepatuhan pengobatan (medication adherence) berkaitan dengan kesesuaian perilaku penggunaan obat terhadap regimen yang

direkomendasikan [2]. Kedua konsep tersebut saling berkaitan, tetapi tidak dapat diperlakukan sebagai istilah yang sepenuhnya sama. Ketidakberlanjutan terapi dapat menghambat pencapaian tujuan pengobatan dan menyulitkan tenaga kesehatan dalam melakukan pemantauan terhadap keberlanjutan terapi pasien [3]. Oleh karena itu, identifikasi risiko non-persistensi menjadi salah satu aspek yang relevan dalam pengembangan sistem pendukung keputusan di bidang farmasi.

Perkembangan kecerdasan buatan (artificial intelligence, AI) dan pembelajaran mesin (machine learning, ML) membuka peluang untuk menganalisis pola yang terdapat dalam data kesehatan dan mendukung proses pengambilan keputusan berbasis data [4], [5]. Algoritma ML dapat mempelajari hubungan antara variabel demografis, karakteristik klinis, serta riwayat pelayanan kesehatan dengan suatu label target. Pada konteks persistensi pengobatan, pendekatan klasifikasi dapat digunakan untuk membedakan observasi yang dikategorikan sebagai persisten dan non-persisten berdasarkan karakteristik yang tersedia [6], [7]. Meskipun demikian, keberhasilan model tidak cukup dinilai berdasarkan ketepatan klasifikasi semata. Model yang digunakan untuk memperkirakan risiko juga perlu menghasilkan probabilitas yang dapat dipercaya dan memberikan penjelasan mengenai faktor yang berkontribusi terhadap hasil prediksi.

Penelitian terdahulu oleh Mardewi et al. telah mengevaluasi algoritma Decision Tree menggunakan kriteria pemisahan Gini, Entropy, dan Log Loss pada data resep yang terdiri atas 3.424 observasi dan 67 variabel. Penelitian tersebut melaporkan bahwa kriteria Entropy memperoleh akurasi sebesar 79,1%, diikuti Gini sebesar 78% dan Log Loss sebesar 77,9% [6]. Temuan tersebut memberikan dasar untuk pengembangan penelitian lanjutan mengenai klasifikasi persistensi pengobatan. Namun, penelitian tersebut berfokus pada perbandingan kriteria dalam algoritma Decision Tree. Dengan demikian, pengembangan evaluasi model ensemble, keandalan probabilitas risiko, dan interpretabilitas prediksi merupakan arah penelitian lanjutan yang relevan.

Salah satu pendekatan yang dapat dipertimbangkan adalah ensemble learning, yaitu metode yang menggabungkan beberapa model atau pohon keputusan untuk menghasilkan prediksi. Random Forest mengombinasikan sejumlah pohon keputusan melalui proses pengambilan sampel dan pemilihan fitur secara acak [8], sedangkan Gradient Boosting membangun model secara bertahap dengan memperbaiki kesalahan prediksi dari model sebelumnya [9]. Pendekatan tersebut memungkinkan peneliti mengevaluasi apakah penggabungan sejumlah model dapat memberikan kemampuan diskriminasi yang lebih baik dibandingkan satu pohon keputusan. Walaupun demikian, model ensemble tidak selalu menghasilkan kinerja yang lebih unggul pada semua dataset. Oleh sebab itu, perbandingan perlu dilakukan melalui protokol eksperimen yang konsisten, pemisahan data yang tepat, dan metrik evaluasi yang sesuai dengan karakteristik kelas target.

Selain kemampuan membedakan kelas, keandalan probabilitas prediksi merupakan aspek penting dalam pemodelan risiko. Sebuah model dapat memberikan probabilitas 0,80 untuk suatu observasi, tetapi probabilitas tersebut belum tentu mencerminkan frekuensi kejadian aktual yang sebanding. Kalibrasi probabilitas bertujuan menyelaraskan probabilitas prediksi dengan frekuensi empiris kejadian. Niculescu-Mizil dan Caruana membahas pentingnya kualitas estimasi probabilitas dalam berbagai metode pembelajaran mesin [10], sedangkan Guo et al. menunjukkan bahwa kalibrasi merupakan persoalan tersendiri yang perlu dievaluasi, bahkan pada model dengan kinerja klasifikasi yang baik [11]–[13]. Dalam penelitian ini, kalibrasi direncanakan melalui metode seperti sigmoid scaling, dengan evaluasi menggunakan Brier Score, log loss, dan kurva kalibrasi. Evaluasi tersebut diperlukan untuk mengetahui apakah probabilitas yang dihasilkan dapat digunakan sebagai estimasi risiko yang lebih dapat dipercaya, bukan hanya sebagai skor untuk menentukan kelas.

Aspek lain yang perlu diperhatikan adalah interpretabilitas model. Model dengan kinerja prediksi tinggi belum tentu mudah dipahami oleh pengguna, terutama ketika keputusan didasarkan pada banyak variabel. Explainable artificial intelligence (XAI) menyediakan pendekatan untuk memeriksa bagaimana fitur berkontribusi terhadap keluaran model. Salah satu metode yang dapat digunakan adalah SHAP (SHapley Additive exPlanations), yang mendistribusikan kontribusi fitur terhadap prediksi berdasarkan konsep nilai Shapley dari teori permainan kooperatif [14]. Dengan SHAP, penelitian dapat mengidentifikasi fitur yang paling berkontribusi terhadap prediksi non-persistensi dan menelaah penjelasan pada tingkat observasi individual. Namun, kontribusi fitur terhadap keluaran model tidak secara otomatis menunjukkan hubungan sebab-akibat dalam konteks klinis.

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan pendekatan Risk-Calibrated and Explainable Ensemble Learning untuk klasifikasi non-persistensi pengobatan menggunakan data Data Fix.csv. Penelitian dirancang untuk membandingkan Decision Tree sebagai model dasar dengan beberapa model ensemble, termasuk

Random Forest, Extra Trees, dan Gradient Boosting. Kinerja model dievaluasi melalui accuracy, precision, recall, F1-score, ROC-AUC, dan PR-AUC. Selanjutnya, kalibrasi probabilitas dievaluasi menggunakan Brier Score, log loss, dan kurva kalibrasi, sedangkan interpretabilitas model dianalisis menggunakan SHAP. Rancangan ini memungkinkan evaluasi yang lebih komprehensif terhadap kemampuan diskriminasi, keandalan probabilitas, dan penjelasan prediksi.

Kontribusi yang diharapkan dari penelitian ini adalah tiga hal. Pertama, menyediakan perbandingan empiris antara model Decision Tree dan model ensemble pada dataset dan protokol evaluasi yang sama. Kedua, mengevaluasi kalibrasi probabilitas sebagai dimensi yang berbeda dari kinerja klasifikasi. Ketiga, menggunakan SHAP untuk mengidentifikasi fitur yang berkontribusi terhadap prediksi non-persistensi. Berbeda dari penelitian sebelumnya yang membandingkan kriteria pemisahan Decision Tree, penelitian ini mengarahkan evaluasi pada kombinasi kinerja ensemble, kalibrasi risiko, dan interpretabilitas model.

Dengan demikian, penelitian ini bertujuan mengevaluasi kemampuan model ensemble dalam mengklasifikasikan non-persistensi pengobatan, menilai keandalan probabilitas prediksi yang dihasilkan, dan mengidentifikasi fitur yang berkontribusi terhadap hasil prediksi. Istilah deteksi dini dalam judul penelitian perlu ditafsirkan berdasarkan ketersediaan fitur sebelum kejadian non-persistensi diketahui. Apabila urutan waktu tersebut tidak dapat dibuktikan dari dokumentasi dataset, hasil penelitian akan dibatasi sebagai klasifikasi non-persistensi dan tidak diklaim sebagai prediksi prospektif atau bukti klinis.

## Metode

### *A. Desain Penelitian dan Sumber Data*

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen berbasis machine learning untuk mengklasifikasikan status persistensi pengobatan menjadi dua kelas, yaitu Persistent dan Non-Persistent. Fokus penelitian adalah membangun model klasifikasi berbasis ensemble learning yang tidak hanya menghasilkan prediksi kelas, tetapi juga menyediakan estimasi probabilitas risiko yang terkalibrasi serta penjelasan mengenai kontribusi fitur terhadap prediksi.

Dataset yang digunakan adalah Data Fix.csv, yang berisi catatan terkait persistensi pengobatan beserta atribut pasien, terapi, dan karakteristik klinis atau administratif yang tersedia dalam data. Berdasarkan pemeriksaan awal, dataset terdiri atas 3.424 baris dan 68 kolom, termasuk variabel target Persistency_Flag. Distribusi awal kelas target terdiri atas 2.135 catatan Non-Persistent dan 1.289 catatan Persistent. Dataset tersebut digunakan sebagai satu-satunya sumber data empiris dalam penelitian ini.

Sebelum pemodelan, dilakukan pemeriksaan struktur data, tipe variabel, nilai kosong, duplikasi, distribusi kelas, dan kesesuaian setiap atribut terhadap tujuan prediksi. Pemeriksaan waktu ketersediaan fitur juga menjadi bagian penting dari desain penelitian. Prediktor yang baru diketahui setelah periode pengobatan berlangsung atau setelah status persistensi terbentuk tidak boleh digunakan untuk mendukung klaim deteksi dini karena dapat menimbulkan target leakage. Oleh sebab itu, klaim deteksi dini perlu dibatasi pada fitur yang secara realistis telah tersedia pada saat prediksi dilakukan.

### *B. Pra-pemrosesan Data dan Seleksi Fitur*

Tahap pra-pemrosesan dilakukan untuk meningkatkan kualitas data dan mengurangi potensi bias dalam proses pembelajaran. Pertama, data diperiksa untuk menemukan baris duplikat yang identik. Pemeriksaan awal menemukan enam baris duplikat; penghapusan duplikasi dilakukan sebelum pembentukan data latih dan data uji agar catatan yang sama tidak tersebar pada kedua bagian data. Selanjutnya, distribusi kelas target diperiksa untuk menentukan strategi pembagian data yang sesuai.

Kedua, variabel target Persistency_Flag ditetapkan sebagai label klasifikasi, sedangkan atribut lainnya diperlakukan sebagai kandidat prediktor. Fitur yang berpotensi mengungkapkan status atau perubahan kondisi selama terapi, seperti Risk_Segment_During_Rx, Tscore_Bucket_During_Rx, Change_T_Score, Change_Risk_Segment, dan Adherent_Flag, perlu dievaluasi berdasarkan definisi serta waktu pencatatannya. Fitur-fitur tersebut dikeluarkan apabila terbukti mengandung informasi yang hanya tersedia setelah atau bersamaan dengan pembentukan label target.

Ketiga, atribut kategorikal dan numerik diproses sesuai kebutuhan algoritma. Variabel kategorikal dikodekan ke dalam representasi numerik yang sesuai, sedangkan variabel numerik dipertahankan atau dinormalisasi apabila diperlukan oleh model. Penanganan nilai kosong dilakukan berdasarkan hasil audit data, bukan dengan mengasumsikan bahwa semua variabel selalu lengkap. Seluruh proses yang mempelajari parameter dari data—misalnya imputasi, pengodean berbasis kategori, dan seleksi fitur—harus dipasang pada data latih saja melalui pipeline untuk mencegah kebocoran informasi dari data uji.

Setelah penghapusan duplikasi dan penyaringan fitur yang berisiko mengalami kebocoran, jumlah akhir sampel dan prediktor harus dicatat kembali dalam laporan eksperimen. Jumlah 57 prediktor yang dibahas pada rancangan awal hanya dapat dipertahankan apabila hasil pemeriksaan akhir benar-benar mengonfirmasi jumlah tersebut.

### *C. Pengembangan Model dan Validasi*

Penelitian ini membandingkan empat algoritma klasifikasi, yaitu Decision Tree, Random Forest, Extra Trees, dan Gradient Boosting. Pemilihan algoritma tersebut bertujuan untuk mengevaluasi perbedaan kinerja antara model berbasis pohon keputusan tunggal dan metode ensemble yang menggabungkan beberapa pohon keputusan. Random Forest membangun sejumlah pohon menggunakan sampel dan subset fitur yang bervariasi [8], sedangkan Gradient Boosting membentuk model secara bertahap dengan mengoptimalkan fungsi kerugian [9], [15].

Untuk Random Forest, probabilitas suatu sampel $x$ termasuk kelas Non-Persistent dapat direpresentasikan sebagai rata-rata probabilitas dari seluruh pohon:

$$[\hat{p}(y = 1 \mid x) = \frac{1}{T}\sum_{t=1}^{T} \quad \widehat{p_t}(y = 1 \mid x)] \qquad (1)$$

dengan *T* menyatakan jumlah pohon dan $\widehat{p_t}$ merupakan probabilitas kelas yang dihasilkan oleh pohon ke-*t*. Pada penelitian ini, $y = 1$ ditetapkan sebagai kelas Non-Persistent. Representasi tersebut menjelaskan prinsip agregasi probabilitas pada Random Forest; mekanisme agregasi dan optimasi pada model lain mengikuti karakteristik masing-masing algoritma.

Dataset dibagi menjadi data latih sebesar 80% dan data uji sebesar 20% menggunakan stratified split agar proporsi kelas target relatif terjaga pada kedua bagian. Random seed ditetapkan sebesar 42 untuk mendukung reprodusibilitas. Selanjutnya, validasi silang berstrata lima lipatan (stratified five-fold cross-validation) dilakukan pada data pengembangan untuk menilai kestabilan model dan membantu pemilihan konfigurasi hiperparameter. Data uji disimpan untuk evaluasi akhir dan tidak digunakan dalam pemilihan model maupun penyetelan hiperparameter.

Hiperparameter yang perlu ditentukan dan dilaporkan meliputi jumlah pohon, kedalaman maksimum pohon, jumlah minimum sampel pada pemisahan simpul, jumlah minimum sampel pada daun, serta parameter khusus untuk setiap algoritma. Pada Gradient Boosting, misalnya, parameter yang relevan mencakup learning rate, jumlah estimator, dan kedalaman atau kompleksitas setiap pohon. Nilai akhir parameter harus berasal dari konfigurasi yang benar-benar dijalankan dan tidak ditentukan hanya berdasarkan perkiraan.

### *D. Kalibrasi Probabilitas dan Evaluasi Kinerja*

Evaluasi model dilakukan pada data uji dengan memperhatikan kinerja klasifikasi dan kualitas estimasi probabilitas. Akurasi, presisi, recall, dan F1-score digunakan untuk mengukur kemampuan model dalam membedakan kelas target. Karena fokus penelitian adalah mendeteksi kemungkinan Non-Persistent, kelas tersebut diperlakukan sebagai kelas positif.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

$$F1 = 2\frac{Precision \times Recall}{Precision+Recall} \qquad (5)$$

dengan *TP* adalah jumlah kasus Non-Persistent yang teridentifikasi dengan benar, *FP* adalah kasus Persistent yang salah diprediksi sebagai Non-Persistent, dan *FN* adalah kasus Non-Persistent yang tidak terdeteksi. Recall menjadi metrik penting karena menunjukkan proporsi kasus non-persistensi yang berhasil dikenali oleh model. Namun, peningkatan recall perlu dinilai bersama presisi agar tidak menghasilkan terlalu banyak peringatan yang keliru.

Selain metrik klasifikasi, penelitian ini menggunakan Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) dan Precision–Recall Area Under the Curve (PR-AUC) untuk mengevaluasi kemampuan pemeringkatan risiko pada berbagai ambang keputusan. PR-AUC juga memberikan informasi yang berguna ketika distribusi kelas tidak seimbang.

Tahap berikutnya adalah kalibrasi probabilitas. Model dengan kemampuan klasifikasi yang baik belum tentu menghasilkan probabilitas risiko yang sesuai dengan frekuensi kejadian sebenarnya [5], [6]. Oleh karena itu, probabilitas dari model terpilih dikalibrasi menggunakan metode sigmoid calibration apabila didukung oleh hasil validasi. Secara umum, transformasi sigmoid dapat ditulis sebagai:

$$[\widehat{p_{cal}}(y = 1 \mid x) = \frac{1}{1+\exp(Af(x)+B)}] \quad (6)$$

dengan $f(x)$ merupakan skor model sebelum kalibrasi, sedangkan *A* dan *B* adalah parameter yang dipelajari dari data kalibrasi. Tanda dan nilai parameter tersebut ditentukan melalui proses pelatihan kalibrator.

Kualitas kalibrasi dievaluasi menggunakan Brier score dan log loss. Untuk *N* sampel dengan label biner $y_i$ dan probabilitas prediksi $\widehat{p_i}$, Brier score dirumuskan sebagai:

$$[BS = \frac{1}{N}\sum_{i=1}^{N}(\widehat{p_i} - y_i)^2] \quad (7)$$

Nilai Brier score yang lebih rendah menunjukkan probabilitas prediksi yang lebih baik menurut ukuran tersebut. Log loss digunakan sebagai ukuran pelengkap yang memberikan penalti besar pada prediksi probabilitas yang sangat keliru. Hasil sebelum dan sesudah kalibrasi dibandingkan untuk mengetahui apakah kalibrasi benar-benar meningkatkan kualitas probabilitas. Kalibrasi tidak dianggap berhasil hanya karena diterapkan; peningkatannya harus didukung oleh hasil evaluasi pada data yang tidak digunakan untuk melatih kalibrator.

### *E. Interpretabilitas Model dan Analisis Risiko*

Untuk meningkatkan keterjelasan prediksi, penelitian ini menggunakan SHapley Additive exPlanations (SHAP) guna mengestimasi kontribusi fitur terhadap keluaran model [7]. SHAP memberikan penjelasan pada tingkat global maupun individual. Analisis global digunakan untuk mengidentifikasi fitur yang paling berkontribusi terhadap prediksi model secara keseluruhan, sedangkan analisis individual digunakan untuk menjelaskan faktor yang meningkatkan atau menurunkan estimasi risiko pada satu catatan.

Penting untuk ditegaskan bahwa besarnya nilai SHAP menunjukkan kontribusi suatu fitur terhadap prediksi model, bukan bukti bahwa fitur tersebut menyebabkan non-persistensi. Interpretasi juga perlu mempertimbangkan hubungan antarf fitur, kemungkinan korelasi, serta konteks klinis dan administratif dari setiap variabel.

Selanjutnya, probabilitas yang telah dikalibrasi dapat digunakan sebagai dasar pengelompokan risiko untuk mendukung penentuan prioritas tindak lanjut. Namun, ambang keputusan atau kategori risiko harus ditentukan berdasarkan validasi, tujuan penggunaan, dan konsekuensi kesalahan prediksi; ambang tersebut tidak boleh dipilih semata-mata untuk memaksimalkan kinerja pada data uji. Apabila belum tersedia ambang yang tervalidasi secara klinis, keluaran penelitian sebaiknya diperlakukan sebagai estimasi risiko untuk mendukung peninjauan oleh tenaga terkait, bukan sebagai keputusan otomatis mengenai pasien.

Secara keseluruhan, metode ini menggabungkan perbandingan algoritma ensemble, evaluasi klasifikasi, pemeriksaan kalibrasi probabilitas, dan interpretabilitas berbasis SHAP. Kombinasi tersebut ditujukan untuk menghasilkan model yang tidak hanya mampu mengidentifikasi kelas Non-Persistent, tetapi juga memberikan estimasi risiko yang lebih dapat dievaluasi dan penjelasan prediksi yang lebih transparan. Validitas klaim deteksi dini tetap bergantung pada verifikasi bahwa seluruh fitur yang digunakan tersedia pada saat prediksi dilakukan.

## Hasil dan Diskusi

### *A. Pemeriksaan dan persiapan data*

Berkas Data Fix.csv memuat 3.424 baris dan 68 kolom. Ditemukan 6 baris duplikat identik, sehingga jumlah observasi setelah penghapusan duplikasi menjadi 3.418. Target penelitian adalah Persistency_Flag, dengan kelas Non-Persistent sebagai kelas positif.

Sepuluh fitur yang menunjukkan indikasi informasi selama terapi, perubahan kondisi, atau status kepatuhan dikeluarkan dari eksperimen awal. Dengan demikian, model menggunakan 57 fitur prediktor yang tersisa. Langkah ini mengurangi risiko kebocoran informasi, tetapi belum membuktikan bahwa seluruh fitur yang tersisa tersedia sebelum non-persistensi terjadi.

Data dibagi menjadi 80% data pengembangan dan 20% data uji dengan stratifikasi kelas. Validasi silang stratified 5-fold dilakukan pada data pengembangan. Empat model yang dibandingkan adalah Decision Tree, Random Forest, Extra Trees, dan Gradient Boosting.

### *B. Hasil validasi silang*

Tabel berikut menampilkan rata-rata dan simpangan baku ROC-AUC dari validasi silang 5-fold.

**Tabel 1.** Rata-rata dan simpangan baku ROC-AUC

| *Model* | *ROC-AUC, rata-rata ± SD* | *F1-score, rata-rata ± SD* |
|---|---|---|
| Decision Tree | 0,750 ± 0,008 | 0,746 ± 0,009 |

| | | |
|---|---|---|
| Random Forest | 0,848 ± 0,013 | 0,822 ± 0,008 |
| Extra Trees | 0,845 ± 0,012 | 0,820 ± 0,007 |
| Gradient Boosting | 0,851 ± 0,015 | 0,838 ± 0,006 |

Gradient Boosting memperoleh rata-rata ROC-AUC tertinggi sebesar 0,851 dan F1-score tertinggi sebesar 0,838. Random Forest memberikan ROC-AUC rata-rata sebesar 0,848, sehingga selisih antara kedua model relatif kecil.

Hasil ini menunjukkan bahwa model ensemble memiliki kemampuan diskriminasi yang lebih baik daripada Decision Tree pada eksperimen ini. Namun, perbedaan rata-rata tersebut belum membuktikan keunggulan yang signifikan secara statistik. Simpangan baku juga tidak menggantikan pengujian statistik berpasangan.

### *C. Evaluasi pada data uji*

Model yang telah dilatih kemudian dievaluasi pada data uji yang tidak digunakan untuk pelatihan maupun validasi silang.

**Tabel 2.** Kinerja model

| *Model* | *Accuracy* | *Precision* | *Recall* | *F1-score* | *ROC-AUC* |
|---|---|---|---|---|---|
| Decision Tree | 0,708 | 0,787 | 0,728 | 0,756 | 0,758 |
| Random Forest | 0,768 | 0,820 | 0,803 | 0,811 | 0,826 |
| Extra Trees | 0,763 | 0,813 | 0,805 | 0,809 | 0,822 |
| Gradient Boosting | 0,776 | 0,794 | 0,866 | 0,828 | 0,830 |

Gradient Boosting menghasilkan accuracy 77,63%, recall 86,62%, dan F1-score 82,83%. Model berhasil mengidentifikasi 369 dari 426 observasi non-persisten pada data uji. Sebanyak 57 observasi non-persisten salah diklasifikasikan sebagai persisten.

Berdasarkan angka yang dilaporkan untuk Gradient Boosting, confusion matrix pada data uji adalah TP=369, FN=57, FP=96, dan TN=162. Dengan demikian, 57 observasi Non-Persistent terlewat, sedangkan 96 observasi Persistent berpotensi memicu peninjauan tambahan. Interpretasi ini bergantung pada ambang klasifikasi yang digunakan dalam eksperimen; ambang tersebut harus dikonfirmasi dari kode asli sebelum laporan final.

Hasil ini mengindikasikan bahwa Gradient Boosting cukup menjanjikan ketika tujuan evaluasi menekankan identifikasi kelas non-persisten. Namun, model juga menghasilkan 96 prediksi non-persisten yang sebenarnya termasuk kelas persisten. Dalam konteks farmasi, kondisi ini dapat menambah kebutuhan pemeriksaan lanjutan. Karena itu, ambang keputusan perlu dievaluasi berdasarkan konsekuensi kesalahan, bukan hanya berdasarkan ambang standar 0,5.

### *D. Evaluasi kalibrasi risiko*

Kalibrasi sigmoid diterapkan pada Gradient Boosting menggunakan validasi silang pada data pengembangan. Hasilnya dibandingkan dengan probabilitas model sebelum kalibrasi.

**Tabel 3.** Perbandingan Kalibrasi

| *Metrik* | *Sebelum kalibrasi* | *Sesudah kalibrasi* |
|---|---|---|
| Accuracy | 0,776 | 0,772 |
| Recall | 0,866 | 0,850 |
| F1-score | 0,828 | 0,823 |
| ROC-AUC | 0,830 | 0,830 |
| PR-AUC | 0,872 | 0,872 |
| Brier Score | 0,15823 | 0,15774 |
| Log loss | 0,48920 | 0,48945 |

Kalibrasi sigmoid sedikit memperbaiki Brier Score, dari 0,15823 menjadi 0,15774, tetapi log loss sedikit memburuk. Recall juga menurun dari 86,62% menjadi 84,98%. Perubahan ini menunjukkan bahwa kalibrasi tidak meningkatkan seluruh aspek kinerja secara bersamaan.

Oleh karena itu, hasil validasi belum mendukung klaim bahwa kalibrasi sigmoid secara konsisten meningkatkan kualitas probabilitas risiko. Analisis calibration curve dan perbandingan metode kalibrasi lain masih diperlukan sebelum model dapat disebut terkalibrasi dengan baik.

### *E. Analisis SHAP*

Analisis SHAP dilakukan pada model Gradient Boosting menggunakan sampel 150 observasi dari data uji. Nilai absolut SHAP dirata-ratakan dan diagregasikan berdasarkan fitur asli untuk melihat kontribusi relatif fitur terhadap keluaran model. Lima fitur teratas berdasarkan rata-rata absolut SHAP adalah:

**Tabel 4.** Kontribusi relatif fitur

| *Peringkat* | *Fitur* | *Rata-rata absolut SHAP* |
|---|---|---|
| 1 | Comorb_Long_Term_Current_Drug_Therapy | 0,3571 |
| 2 | Comorb_Encounter_For_Screening_For_Malignant_Neoplasms | 0,3457 |
| 3 | Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx | 0,2809 |
| 4 | Comorb_Encounter_For_Immunization | 0,2342 |
| 5 | Ntm_Speciality_Bucket | 0,1575 |

Hasil tersebut menunjukkan bahwa indikator terkait terapi obat jangka panjang, kunjungan medis, dan kategori spesialis merupakan fitur yang relatif berpengaruh terhadap keluaran model. Namun, nilai absolut SHAP hanya mengukur besarnya kontribusi, bukan arah pengaruh. Untuk menyimpulkan apakah suatu nilai fitur mendorong prediksi ke arah non-persistensi atau persistensi, diperlukan analisis arah SHAP dan pemeriksaan distribusi fitur.

Hasil ini juga tidak menunjukkan hubungan sebab-akibat. Fitur penting menurut model belum tentu merupakan faktor penyebab non-persistensi dalam praktik klinis.

## Kesimpulan

Penelitian ini mengusulkan pendekatan Risk-Calibrated and Explainable Ensemble Learning untuk mendukung deteksi dini non-persistensi pengobatan di apotek. Pendekatan yang dirancang mengintegrasikan perbandingan beberapa algoritma klasifikasi, yaitu Decision Tree, Random Forest, Extra Trees, dan Gradient Boosting, dengan evaluasi kualitas probabilitas melalui kalibrasi serta interpretasi prediksi menggunakan SHAP. Integrasi tersebut diharapkan dapat menghasilkan model yang tidak hanya mampu mengidentifikasi pasien berisiko mengalami non-persistensi, tetapi juga memberikan estimasi risiko yang lebih dapat dievaluasi dan penjelasan mengenai faktor yang memengaruhi prediksi. Evaluasi model dilakukan melalui metrik klasifikasi, meliputi akurasi, presisi, recall, F1-score, ROC-AUC, dan PR-AUC, serta metrik kalibrasi berupa Brier score dan log loss. Penggunaan metrik yang beragam diperlukan karena kemampuan membedakan kelas dan ketepatan estimasi probabilitas merupakan dua aspek yang berbeda. Sementara itu, SHAP digunakan untuk menjelaskan kontribusi fitur terhadap prediksi model pada tingkat global maupun individual. Di antara empat model yang diuji pada protokol evaluasi ini, Gradient Boosting menunjukkan hasil terbaik berdasarkan F1-score dan ROC-AUC validasi silang, dengan nilai masing-masing 0,838 dan 0,851. Pada data uji, model memperoleh accuracy 0,776, precision 0,794, recall 0,866, F1-score 0,828, dan ROC-AUC 0,830. Kalibrasi sigmoid hanya memperbaiki Brier score secara kecil, sementara log loss sedikit memburuk; karena itu, probabilitas belum dapat dinyatakan terkalibrasi dengan baik. SHAP mengidentifikasi indikator terapi jangka panjang, beberapa jenis kunjungan medis, dan kategori spesialisasi sebagai fitur yang berkontribusi terhadap prediksi, tetapi hasil ini tidak menunjukkan hubungan sebab-akibat. Keterbatasan utama penelitian adalah belum terverifikasinya waktu ketersediaan seluruh fitur, belum adanya validasi eksternal, dan belum dievaluasinya manfaat klinis. Dengan demikian, model belum dapat diklaim siap untuk deteksi dini atau penggunaan klinis otomatis; penelitian lanjutan perlu memverifikasi urutan waktu fitur, menguji kalibrasi lebih lanjut, dan mengevaluasi model pada data eksternal.

## Deklarasi Kepentingan

Penulis menyatakan bahwa tidak terdapat konflik kepentingan yang dapat memengaruhi perancangan penelitian, pengolahan data, pengembangan model, analisis hasil, maupun penyusunan artikel ini. Penelitian ini tidak memiliki kepentingan finansial, hubungan pribadi, atau afiliasi lain yang dapat memengaruhi objektivitas dan integritas hasil penelitian.

## Ucapan Terima Kasih

Penulis menyampaikan terima kasih kepada pihak-pihak yang telah mendukung pelaksanaan penelitian ini, khususnya dalam penyediaan dan pengelolaan data yang digunakan untuk analisis persistensi pengobatan. Penulis juga mengapresiasi dukungan yang diberikan selama proses pengolahan data, pengembangan model *machine learning*, dan penyusunan artikel ilmiah ini.

## Ketersediaan Data

Dataset yang digunakan dalam penelitian ini berasal dari berkas Data Fix.csv, yang memuat catatan terkait persistensi pengobatan dengan variabel target Persistency_Flag. Dataset tersebut digunakan untuk mengembangkan dan mengevaluasi model klasifikasi persistensi pengobatan menggunakan algoritma *Decision Tree*, *Random Forest*, *Extra Trees*, dan *Gradient Boosting*, serta untuk melakukan evaluasi kalibrasi probabilitas dan interpretasi model.

Ketersediaan dataset untuk publik bergantung pada sumber data dan ketentuan penggunaan yang berlaku. Apabila dataset merupakan data internal atau dibatasi oleh kebijakan pemilik data, dataset tidak dapat dibagikan secara bebas dan aksesnya harus memperoleh izin dari pihak yang berwenang. Penelitian ini menggunakan data untuk kepentingan analisis, dengan tetap memperhatikan kerahasiaan informasi dan ketentuan perlindungan data yang relevan.

## Penggunaan Ai Dan Deklarasi Penggunaan Ai Generatif

Dalam penyusunan artikel ini, penulis menggunakan alat kecerdasan buatan generatif sebagai alat bantu untuk mendukung penyuntingan bahasa, perbaikan struktur penulisan, penyusunan teks akademik, dan pengembangan redaksi artikel. Penggunaan alat tersebut dilakukan di bawah pengawasan dan kendali penulis.

Keputusan terkait desain penelitian, pemilihan dan prapemrosesan data, seleksi fitur, penerapan algoritma *machine learning*, evaluasi model, kalibrasi probabilitas, interpretasi hasil, dan penarikan kesimpulan menjadi tanggung jawab penulis. Penulis bertanggung jawab untuk memeriksa dan memverifikasi keakuratan metodologi, hasil eksperimen, referensi, serta seluruh isi akhir artikel. Penggunaan AI generatif tidak menggantikan kewajiban penulis untuk memastikan validitas ilmiah dan integritas penelitian.

## Daftar Pustaka

[1] F. Ayuningtyas, D. Setiawan, and G. F. Galistiani, "Narrative Review: The Influence of Single-Pill Combination (SPC) Antihypertensive Therapy on Medication Adherence in Patients with Hypertension," *BENZENA Pharm. Sci. J.*, vol. 5, no. 01, pp. 135–146, Jul. 2026, doi: 10.31941/BENZENA.V5I01.7982.

[2] "Model Kepatuhan Minum Obat Pada Pasien Hipertensi - Ns. Ahmad Abdul Ghofar Abdulloh, S.Kep., M.Kep. - Google Books." https://books.google.co.id/books?hl=en&lr=&id=RQrxEQAAQBAJ&oi=fnd&pg=PA1&dq=Persistensi+mengacu+pada+lamanya+pasien+melanjutkan+terapi+sejak+pengobatan+dimulai+hingga+penghentian+terapi,+sedangkan+kepatuhan+pengobatan+(medication+adherence)+berkaitan+dengan+kesesuaian+perilaku+penggunaan+obat+terhadap+regimen+yang+direkomendasikan&ots=iBo6pHLpTy&sig=S_7N38T3t-Q5KPV_jdABRFsflbE&redir_esc=y#v=onepage&q&f=false (accessed Oct. 10, 2026).

[3] J. Farmasi *et al.*, "Evaluasi Implementasi Pharmaceutical Care terhadap Kepatuhan Terapi Pasien Penyakit Kronis melalui Pendekatan Farmasi Klinis Terintegrasi," *J. Farm. Klin. dan Kesehat. Masy.*, vol. 1, no. 1, pp. 11–21, Jun. 2026, Accessed: Oct. 10, 2026. [Online]. Available: https://journal.unukaltim.ac.id/index.php/JFKKM/article/view/39.

[4] L. Banh and G. Strobel, "Generative artificial intelligence," *Electron. Mark. 2023 331*, vol. 33, no. 1, pp. 63-, Dec. 2023, doi: 10.1007/S12525-023-00680-1.

[5] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: 10.1109/ACCESS.2020.2988510.

[6] M. Mardewi, F. Aziz, S. Usman, and R. F. Syam, "Detection of Persistent vs. Non-Persistent Drugs in Pharmacy Using Decision Tree Classification Based on Gini, Entropy, and Log Loss Criteria," *Ilk. J. Ilm.*, vol. 17, no. 2, pp. 186–195, Aug. 2025, doi: 10.33096/ilkom.v17i2.2585.186-195.

[7] S. Abasa, F. Aziz, P. Ishak, and J. J. And, "Detection of Persistent vs Non-Persistent Medications in Pharmacy Using Artificial Intelligence: Development of Intelligent Algorithms for Pharmaceutical Product," *journal.unpacti.ac.id*, vol. 6, no. 1, pp. 2723–1240, 2025, doi: 10.61628/jsce.v6i1.1618.

[8] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.

[9] G. Biau, B. Cadre, and L. Rouvière, "Accelerated gradient boosting," *Mach. Learn.*, vol. 108, no. 6, pp. 971–992, Jun. 2019, doi: 10.1007/S10994-019-05787-1.

[10] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," *ICML 2005 - Proc. 22nd Int. Conf. Mach. Learn.*, pp. 625–632, 2005, doi: 10.1145/1102351.1102430.

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks." PMLR, pp. 1321–1330, Jul. 17, 2017, Accessed: Jul. 02, 2026. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html.

[12] M. Minderer *et al.*, "Revisiting the Calibration of Modern Neural Networks," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 15682–15694, Dec. 2021, Accessed: Oct. 10, 2026. [Online]. Available: https://github.com/google-research/robustness_metrics/tree/master/.

[13] C. Wang, "Calibration in Deep Learning: A Survey of the State-of-the-Art," Aug. 2023, Accessed: Oct. 10, 2026. [Online]. Available: https://arxiv.org/pdf/2308.01222.

[14] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, Accessed: Feb. 05, 2026. [Online]. Available: https://github.com/slundberg/shap.

[15] J. F.-A. of statistics and undefined 2001, "Greedy function approximation: a gradient boosting machine," *JSTORJH FriedmanAnnals Stat. 2001•JSTOR*, vol. 29, no. 5, pp. 1189–1232, 2001, Accessed: Dec. 20, 2023. [Online]. Available: https://www.jstor.org/stable/2699986.