ISSN (Online) 2747-0563
Volume 6 Tahun 2026

Evaluasi Kinerja Model CatBoost pada Penanganan Missing Value

Shalsa Adinda Dwiska Putri[1], Faisal Muttaqin[2], Fetty Tri Anggraeny[3]

[1,2,3] Informatika, UPN Veteran Jawa Timur
[2]faisalmuttaqin.if@upnjatim.ac.id
[3]fettyanggraeny.if@upnjatim.ac.id

*Corresponding author email: 22081010346@student.upnjatim.ac.id

Abstrak*— Hipertensi merupakan salah satu penyakit tidak menular dengan prevalensi yang terus meningkat dan sering tidak terdeteksi sejak dini, sehingga diperlukan model prediksi yang akurat untuk mendukung upaya deteksi dini. Kualitas data menjadi faktor penting dalam pembangunan model *machine learning*, terutama terkait keberadaan *missing value* yang dapat memengaruhi performa prediksi. Penelitian ini bertujuan untuk mengevaluasi pengaruh dua strategi penanganan *missing value*, yaitu imputasi dan penghapusan data (*drop missing value*), terhadap kinerja algoritma CatBoost dalam memprediksi risiko hipertensi. Dataset yang digunakan adalah *Hypertension Risk Prediction Dataset* yang terdiri dari 1.985 sampel dan 11 variabel yang mencakup karakteristik individu, gaya hidup, serta riwayat kesehatan. Pada strategi imputasi, nilai kosong pada variabel *Medication* diisi dengan kategori *Non-Medication*, sedangkan pada strategi kedua seluruh baris yang mengandung nilai kosong dihapus. Model dievaluasi menggunakan metrik *accuracy, precision, recall, F1-score,* dan *confusion matrix*. Hasil penelitian menunjukkan bahwa kedua pendekatan menghasilkan performa yang sangat tinggi dengan nilai *accuracy* di atas 99%. Strategi imputasi memperoleh *accuracy* sebesar 99,60%, *precision* 100%, *recall* 99,22%, dan *F1-score* 99,61%, sedangkan strategi penghapusan data memperoleh *accuracy* 99,41%, *precision* 99,52%, *recall* 99,35%, dan *F1-score* 99,44%. Analisis *confusion matrix* menunjukkan bahwa strategi imputasi menghasilkan jumlah *false negative* yang lebih rendah serta mampu mempertahankan jumlah data yang lebih besar. Oleh karena itu, strategi imputasi dinilai lebih efektif dalam mendukung prediksi risiko hipertensi menggunakan CatBoost.

Kata Kunci— **Hipertensi, *Machine Learning*, CatBoost, *Missing Value*, Prediksi Risiko**

I. PENDAHULUAN

Hipertensi merupakan masalah kesehatan global yang terus meningkat dan menjadi salah satu penyebab utama kematian dini di dunia. WHO melaporkan bahwa jumlah kasus hipertensi pada orang dewasa usia 30–79 tahun menunjukkan peningkatan yang signifikan, dari sekitar 650 juta kasus pada tahun 1990 menjadi sekitar 1,3 miliar pada tahun 2019, dan pada tahun 2024 jumlahnya diperkirakan mencapai 1,4 miliar atau sekitar 33% populasi dewasa secara global [1][2]. WHO juga memproyeksikan jumlah ini dapat mencapai 1,5 miliar pada tahun 2025, dengan dampak yang lebih besar pada negara dengan akses layanan kesehatan terbatas. Di kawasan Asia Tenggara, sekitar sepertiga populasi mengalami hipertensi, yang berkontribusi terhadap sekitar 1,5 juta kematian setiap tahun [3].

Di benua Asia, jumlah penderita hipertensi tercatat sebesar 38,4 juta pada tahun 2000 dan diperkirakan meningkat menjadi 67,4 juta pada tahun 2025. Di Asia Tenggara, prevalensi hipertensi menempati peringkat ketiga tertinggi, yaitu sekitar 25% dari total populasi. Di Indonesia, sekitar 17,21% penduduk menderita hipertensi, dengan sebagian besar kasus tidak terdeteksi [4]. Secara nasional, prevalensi hipertensi mencapai 25,8%, namun hanya sekitar 8% kasus yang tercatat oleh tenaga kesehatan, dan hanya 26,97% penderita yang menyadari kondisi mereka [5]. Hal ini menunjukkan bahwa hipertensi masih menjadi masalah kesehatan yang signifikan dan memerlukan perhatian serius, terutama dalam hal deteksi dini. Hipertensi adalah kondisi ketika tekanan darah mencapai atau melebihi 140 mmHg untuk sistolik atau 90 mmHg untuk diastolik. Penyakit ini merupakan faktor risiko utama penyakit kardiovaskular dan berbagai komplikasi kesehatan lainnya, seperti serangan jantung, gagal jantung, stroke, dan gagal ginjal kronis. Meskipun sering dikaitkan dengan kelompok usia lanjut, hipertensi juga dapat terjadi pada usia muda (20–40 tahun), dengan prevalensi sekitar 1 dari 8 orang dewasa [6].

Istilah *the silent disease* sering digunakan untuk menggambarkan hipertensi, sebab sebagian besar penderita baru mengetahui kondisinya setelah menjalani pemeriksaan medis. Secara global, hipertensi termasuk dalam penyakit tidak menular utama yang meningkatkan beban penyakit, kecacatan, dan kematian dini [7]. Rendahnya kesadaran masyarakat terhadap risiko hipertensi menyebabkan banyak penderita tidak mendapatkan pengobatan yang memadai. Padahal, terdapat sejumlah faktor yang berperan dalam terjadinya hipertensi, seperti usia, jenis kelamin, nilai indeks massa tubuh, kebiasaan merokok, riwayat penyakit, pola makan, aktivitas fisik, dan kondisi kesehatan mental [8][9]. Oleh karena itu, deteksi dini dan prediksi risiko hipertensi menjadi sangat penting untuk mendukung upaya pencegahan dan pengelolaan penyakit secara efektif. Seiring dengan perkembangan teknologi, metode *machine learning* telah banyak digunakan dalam bidang kesehatan untuk membantu prediksi penyakit, termasuk hipertensi [10][11]. Namun, keberhasilan model prediksi sangat dipengaruhi oleh kualitas data yang digunakan. Dataset kesehatan umumnya memiliki karakteristik kompleks, termasuk ukuran yang besar, variasi fitur, serta keberadaan *missing value* atau data yang tidak lengkap. *Missing value*

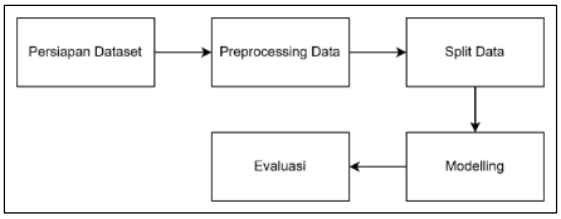
ISSN (Online) 2747-0563
Volume 6 Tahun 2026

merupakan salah satu tantangan utama dalam analisis data karena dapat menurunkan akurasi dan kinerja model prediksi [12][13]. Berbagai penelitian telah memanfaatkan algoritma *machine learning* untuk prediksi hipertensi, termasuk algoritma berbasis *gradient boosting* seperti CatBoost, XGBoost, dan LightGBM. Penelitian sebelumnya menunjukkan bahwa algoritma tersebut mampu menghasilkan performa prediksi yang baik melalui optimasi hyperparameter, penanganan ketidakseimbangan kelas, maupun penerapan metode ensemble untuk meningkatkan akurasi model. Selain itu, CatBoost diketahui memiliki kemampuan untuk *menangani missing value* secara internal sehingga sering digunakan pada berbagai kasus klasifikasi di bidang Kesehatan [14]. Meskipun demikian, sebagian besar penelitian terdahulu berfokus pada peningkatan performa model melalui pemilihan algoritma dan optimasi parameter [15], sementara kajian mengenai pengaruh strategi penanganan *missing value* terhadap kinerja model masih terbatas. Padahal, keberadaan *missing value* merupakan permasalahan yang umum ditemukan pada dataset kesehatan dan dapat memengaruhi kualitas model yang dihasilkan. Di sisi lain, kemampuan CatBoost dalam menangani *missing value* secara internal menimbulkan pertanyaan apakah proses imputasi masih diperlukan atau justru penghapusan data yang tidak lengkap dapat memberikan hasil yang lebih baik. Hingga saat ini, masih terbatas penelitian yang secara khusus membandingkan kedua pendekatan tersebut pada model CatBoost untuk prediksi hipertensi.

Penelitian ini bertujuan untuk mengevaluasi pengaruh dua strategi penanganan *missing value*, yaitu imputasi dan penghapusan data yang mengandung nilai kosong, terhadap kinerja model CatBoost dalam prediksi risiko hipertensi. Evaluasi dilakukan menggunakan metrik *accuracy, precision, recall, F1-score, confusion matrix*, serta *5-Fold Cross Validation* untuk memperoleh gambaran performa model yang lebih komprehensif.

II. METODOLOGI

Penelitian ini dilakukan melalui beberapa tahapan pengolahan data yang tersusun secara sistematis untuk memastikan proses analisis berjalan secara terstruktur dan menghasilkan model prediksi yang baik. Tahapan penelitian meliputi persiapan data, *preprocessing*, pembagian data (*data splitting*), validasi model menggunakan *Cross Validation*, pemodelan (*modeling*), serta evaluasi kinerja model. Alur tahapan penelitian ditunjukkan pada Gbr. 1.



Gbr. 1 Alur penelitian

Berdasarkan *flowchart* pada Gbr. 1, penelitian diawali dengan pengumpulan dan persiapan dataset hipertensi yang kemudian dilanjutkan dengan tahap *preprocessing* untuk menangani *missing value*. Selanjutnya, dataset dibagi menjadi data pelatihan dan data pengujian menggunakan rasio 80:20. Untuk memperoleh estimasi performa yang lebih stabil dan mengurangi pengaruh pembagian data secara acak, dilakukan validasi menggunakan *5-Fold Cross Validation* pada data pelatihan. Setelah itu, model CatBoost dibangun menggunakan data yang telah diproses dan divalidasi. Tahap terakhir adalah evaluasi model menggunakan metrik *accuracy*, *precision*, *recall, F1-score*, serta *confusion matrix* untuk membandingkan pengaruh dua strategi penanganan *missing value* terhadap kinerja model prediksi hipertensi. Setiap tahapan dalam proses penelitian dijelaskan sebagai berikut.

1. Dataset

Sumber data dalam penelitian ini adalah *Hypertension Risk Prediction Dataset* yang diperoleh dari platform Kaggle. Dataset ini berisi informasi mengenai karakteristik individu, gaya hidup, serta riwayat kesehatan yang berpotensi memengaruhi risiko hipertensi. Secara keseluruhan, dataset terdiri dari 1.985 sampel dengan 11 variabel, yang mencakup 10 fitur prediktor dan 1 variabel target, yaitu Has_Hypertension.Fitur-fitur dalam dataset meliputi kombinasi data numerik dan kategorikal, seperti usia (Age), konsumsi garam (Salt_Intake), tingkat stres (Stress_Score), riwayat tekanan darah (BP_History), durasi tidur (Sleep_Duration), indeks massa tubuh (BMI), penggunaan obat (Medication), riwayat keluarga (Family_History), tingkat aktivitas fisik (Exercise_Level), serta status merokok (Smoking_Status). Variabel Has_Hypertension digunakan sebagai target untuk menentukan status hipertensi responden. Distribusi kelas pada variabel target relatif seimbang, dengan 1.032 data hipertensi dan 953 data non-hipertensi, sehingga dataset ini sesuai digunakan untuk pemodelan klasifikasi biner.

2. *Preprocessing Data*

Pada tahap *preprocessing* dilakukan penanganan terhadap nilai kosong (*missing value*) yang terdapat pada dataset. Dalam penelitian ini digunakan dua pendekatan untuk menangani nilai kosong pada variabel Medication. Pendekatan pertama dilakukan dengan mengonversi nilai kosong menjadi kategori Non-Medication, Hal ini didasarkan pada karakteristik data, di mana nilai kosong pada variabel Medication sebagian besar merepresentasikan nilai “None” yang menunjukkan bahwa responden tidak mengonsumsi obat terkait hipertensi. Oleh karena itu, dilakukan imputasi dengan mengganti nilai kosong menjadi kategori Non-Medication untuk mempertahankan informasi data. Pendekatan kedua dilakukan dengan menghapus baris data yang mengandung nilai kosong. Kedua pendekatan tersebut digunakan untuk melihat pengaruh metode penanganan *missing value* terhadap kinerja model yang dihasilkan. Seluruh proses ini dilakukan untuk memastikan kualitas data yang digunakan dalam pemodelan tetap optimal.

ISSN (Online) 2747-0563
Volume 6 Tahun 2026

3. Split Data

Datset kemudian dibagi menjadi dua bagian, yaitu data pelatihan dan pengujian. Pembagian data ini bertujuan untuk melatih model serta mengevaluasi kemampuan model dalam melakukan prediksi terhadap data yang belum pernah digunakan pada proses pelatihan.

Dalam penelitian ini, pembagian dataset dilakukan dengan rasio 80:20, yaitu 80% data digunakan sebagai data pelatihan dan 20% sisanya sebagai data pengujian. Data pelatihan dimanfaatkan untuk membangun model CatBoost, sedangkan data pengujian digunakan untuk menilai kinerja model yang telah dikembangkan. Pembagian data dilakukan menggunakan *stratified splitting* agar proporsi kelas pada data pelatihan dan pengujian tetap seimbang dan merepresentasikan distribusi kelas pada keseluruhan dataset.

4. Cross Validation

Untuk memperoleh estimasi performa model yang lebih stabil dan mengurangi pengaruh pembagian data secara acak, penelitian ini menerapkan *Stratified 5-Fold Cross Validation* pada data pelatihan. Pada metode ini, data pelatihan dibagi menjadi lima *fold* dengan ukuran yang relatif sama dan mempertahankan proporsi kelas pada setiap *fold* agar tetap seimbang. Selanjutnya, model dilatih menggunakan empat *fold* dan diuji pada satu *fold* yang tersisa secara bergantian hingga seluruh *fold* pernah digunakan sebagai data validasi. Nilai evaluasi dari setiap iterasi kemudian dirata-ratakan untuk memperoleh gambaran performa model yang lebih stabil sebelum dilakukan pengujian akhir menggunakan data *testing*.

5. Modeling

Pada tahap pemodelan, digunakan algoritma CatBoost untuk membangun model klasifikasi dalam memprediksi risiko hipertensi berdasarkan fitur-fitur yang tersedia pada dataset. CatBoost adalah algoritma *machine learning* berbasis *gradient boosting* yang memiliki keunggulan khusus dalam mengolah data kategorik secara efisien [13].

CatBoost memiliki mekanisme internal untuk menangani *missing value* pada fitur numerik tanpa memerlukan proses imputasi eksplisit. Dalam proses pembentukan pohon keputusan, CatBoost memperlakukan nilai hilang sebagai nilai khusus dan secara otomatis menentukan arah split optimal (*default direction*) untuk observasi yang memiliki *missing value* pada setiap *node*. Mekanisme ini terutama bekerja pada fitur numerik, sedangkan pada fitur kategorikal, penanganan nilai hilang tidak selalu dieksplorasi secara eksplisit dalam proses *encoding* [16]. Oleh karena itu, dalam penelitian ini dilakukan perbandingan antara strategi imputasi dan penghapusan data pada variabel Medication untuk mengevaluasi apakah perlakuan eksplisit terhadap *missing value* masih memberikan dampak terhadap kinerja model dalam kasus klasifikasi hipertensi.

Proses pemodelan dilakukan menggunakan data yang telah melalui tahap *preprocessing* dan pembagian data. Model CatBoost dilatih dengan memanfaatkan data pelatihan guna mengenali hubungan antara variabel prediktor dan variabel target, yaitu Has_Hypertension. Berikutnya, model yang sudah dilatih digunakan untuk memprediksi data pengujian sebagai bagian dari evaluasi kemampuan generalisasi model.

Pemodelan dilakukan pada dua skenario penanganan *missing value*, yaitu dengan pendekatan imputasi dan penghapusan baris data yang mengandung nilai kosong. Hasil dari kedua pendekatan tersebut kemudian dianalisis untuk melihat pengaruhnya terhadap kinerja model yang dihasilkan.

6. Evaluasi

Tahapan evaluasi bertujuan untuk mengukur performa model CatBoost dalam mengklasifikasikan status hipertensi. Penilaian dilakukan dengan memanfaatkan data pengujian yang tidak digunakan selama proses pelatihan, sehingga hasil evaluasi dapat mencerminkan kemampuan model dalam menghadapi data baru.

Pengukuran kinerja dilakukan menggunakan beberapa metrik klasifikasi, yaitu *accuracy, precision, recall*, dan *F1-score*. Nilai *accuracy* menunjukkan tingkat ketepatan prediksi secara keseluruhan. Sementara itu, *precision* merepresentasikan proporsi prediksi positif yang benar, dan *recall* menggambarkan kemampuan model dalam mendeteksi kasus yang benar-benar positif. Adapun *F1-score* merupakan ukuran gabungan antara *precision* dan *recall* yang mencerminkan keseimbangan kinerja model.

Selain itu, analisis juga dilengkapi dengan *confusion matrix* untuk memberikan gambaran jumlah prediksi yang tepat maupun keliru pada masing-masing kelas.

III. *HASIL DAN PEMBAHASAN*

Pada subbab ini, akan disajikan hasil analisis data, pemodelan, dan evaluasi kinerja model CatBoost dalam memprediksi risiko hipertensi. Hasil penelitian dibagi menjadi beberapa bagian, dimulai dengan analisis kondisi dataset dan penanganan *missing value*, dilanjutkan dengan evaluasi performa model menggunakan metrik klasifikasi, serta interpretasi *confusion matrix* untuk menilai kemampuan model dalam mengenali sampel positif dan negatif. Pembahasan difokuskan pada perbandingan dua strategi penanganan *missing value*, yaitu imputasi dan penghapusan baris yang mengandung nilai kosong, untuk menilai pengaruhnya terhadap kinerja model dan keakuratan prediksi.

1. Analisis Data dan *Missing Value*

Sebelum pemodelan, dilakukan analisis data untuk meninjau kondisi dataset secara menyeluruh, termasuk struktur data, distribusi kelas, serta keberadaan *missing value* yang dapat memengaruhi proses pembelajaran model. Dataset yang digunakan terdiri dari 1.985 sampel dengan 11 variabel, yang mencakup 10 fitur prediktor dan 1 variabel target (Has_Hypertension). Pada tahap ini, dilakukan eksplorasi awal

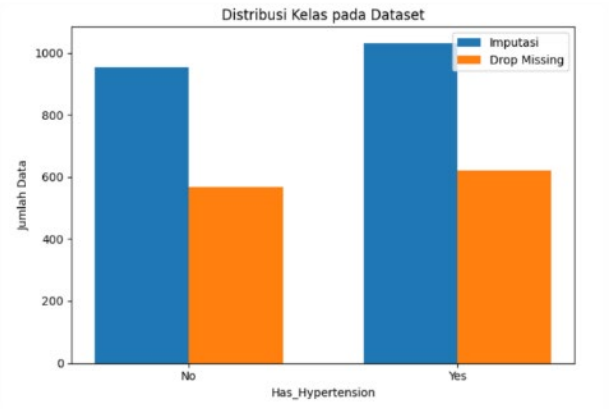
ISSN (Online) 2747-0563
Volume 6 Tahun 2026

untuk memahami karakteristik setiap variabel serta memastikan tidak terdapat anomali data yang dapat mengganggu proses analisis lebih lanjut. Hasil pengecekan menunjukkan bahwa kolom Medication memiliki 799 nilai kosong, sedangkan variabel lainnya lengkap tanpa *missing value* yang signifikan. Rincian jumlah *missing value* pada setiap variabel disajikan pada Gbr. 2.

Fitur	Jumlah Missing
Medication	799

Gbr. 2 Jumlah *missing value*

Dua pendekatan penanganan *missing value* diterapkan, yaitu imputasi dengan mengisi nilai kosong menjadi Non-Medication dan metode *drop* dengan menghapus baris yang memiliki nilai kosong. Setelah *preprocessing*, ukuran dataset menjadi berbeda, di mana dataset hasil imputasi tetap berjumlah 1.985 sampel, sedangkan dataset dengan metode drop berkurang menjadi 1.186 sampel.
Perbandingan distribusi kelas pada kedua dataset ditunjukkan pada Gbr. 3, terlihat bahwa jumlah sampel pada metode imputasi lebih besar dibandingkan metode *drop*, baik pada kelas hipertensi (*Yes*) maupun non-hipertensi (*No*).



Gbr. 3 Distribusi target

2. Evaluasi Model CatBoost

Evaluasi model dilakukan menggunakan dua pendekatan, yaitu *5-Fold Cross Validation* pada data pelatihan untuk mengukur kestabilan model, serta evaluasi pada data pengujian (*test set*) untuk menilai kemampuan generalisasi model terhadap data baru yang belum pernah dilihat sebelumnya.
Hasil *5-Fold Cross Validation* menunjukkan bahwa model CatBoost memiliki performa yang stabil pada kedua strategi penanganan *missing value*. Pada strategi imputasi diperoleh rata-rata *accuracy* sebesar 0.9960, *precision* sebesar 1.0000, *recall* sebesar 0.9922, dan *F1-score* sebesar 0.9961. Sementara itu, pada strategi penghapusan *missing value* diperoleh rata-rata *accuracy* sebesar 0.9941, *precision* sebesar 0.9952, *recall* sebesar 0.9935, dan *F1-score* sebesar 0.9944. Hasil ini menunjukkan performa yang konsisten antara *fold*, sehingga model memiliki stabilitas yang baik dalam proses pembelajaran. Evaluasi utama dalam penelitian ini menggunakan data pengujian (*test set*). Hasil evaluasi ditunjukkan pada Tabel 1.

TABEL I
EVALUASI MODEL

Metode	Accuracy	Precision	Recall	F1-Score
Imputasi	0.995970	1.000000	0.992238	0.996083
Hapus Data	0.994096	0.995199	0.993548	0.994351

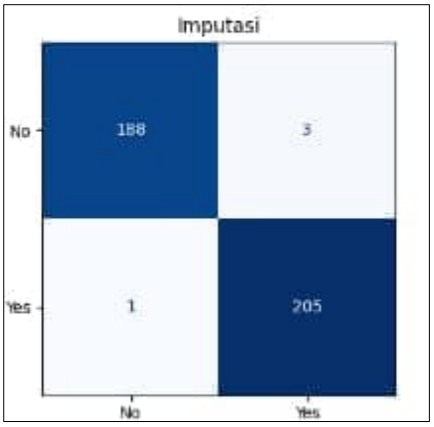
Berdasarkan hasil evaluasi pada *test set*, kedua strategi penanganan *missing value* menghasilkan performa yang sangat tinggi. Strategi imputasi menunjukkan performa yang sedikit lebih unggul dibandingkan strategi penghapusan data, terutama pada metrik *accuracy, precision*, dan *F1-score*. Pada metrik *precision*, strategi imputasi mencapai nilai sempurna yaitu 1.000000, yang menunjukkan tidak adanya *false positive* pada kelas positif. Sementara itu, strategi penghapusan data memperoleh *precision* sebesar 0.995199. Pada metrik *recall*, strategi penghapusan data sedikit lebih tinggi yaitu 0.993548 dibandingkan 0.992238 pada strategi imputasi, namun perbedaannya sangat kecil. Secara keseluruhan, strategi imputasi memberikan performa yang lebih baik karena mampu mempertahankan seluruh data pelatihan sehingga model dapat mempelajari pola data secara lebih lengkap dibandingkan strategi penghapusan data.

Strategi imputasi dapat memberikan kinerja yang lebih baik dibandingkan penghapusan data karena mampu mempertahankan seluruh informasi yang terdapat pada dataset. Penghapusan data menyebabkan berkurangnya jumlah sampel pelatihan, yang berpotensi mengurangi representasi distribusi data, terutama pada dataset dengan ukuran terbatas. Sebaliknya, imputasi memungkinkan seluruh sampel tetap digunakan dalam proses pelatihan sehingga distribusi data lebih terjaga dan model memiliki kemampuan generalisasi yang lebih stabil. Dalam konteks ini, meskipun imputasi sederhana yang digunakan dapat menimbulkan potensi bias, dampaknya lebih kecil dibandingkan kehilangan informasi akibat penghapusan data secara langsung.

3. Confusion Matrix

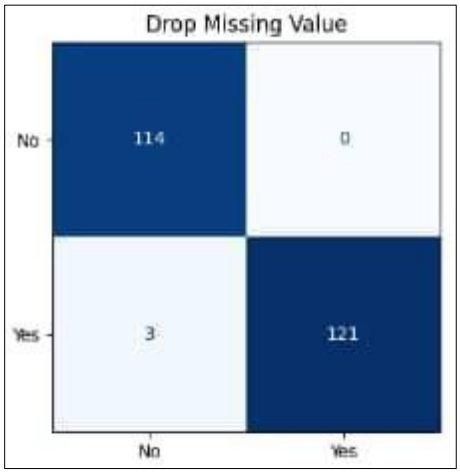
Analisis *confusion matrix* pada Gbr. 4 menunjukkan bahwa model dengan strategi imputasi menghasilkan 205 prediksi benar (*true positive*) dan hanya 1 prediksi salah (*false negative*) pada kelas positif (*Yes*). Pada kelas negatif (*No*), model menghasilkan 188 prediksi benar (*true negative*) dan 3 prediksi salah (*false positive*). Hasil ini menunjukkan bahwa model mampu mengenali hampir seluruh sampel positif maupun negatif dengan tingkat kesalahan yang sangat rendah. Jumlah *false negative* yang lebih sedikit dibandingkan model *drop missing value* menunjukkan bahwa strategi imputasi lebih efektif dalam mendeteksi kasus hipertensi sehingga berpotensi mengurangi risiko kasus yang tidak teridentifikasi.

ISSN (Online) 2747-0563
Volume 6 Tahun 2026



Gbr. 4 *Confusion matrix* imputasi

Sementara itu, pada Gbr. 5, model dengan penghapusan *missing value* menghasilkan 121 prediksi benar (*true positive*) dan 3 prediksi salah (false negative) pada kelas positif (*Yes*). Pada kelas negatif (*No*), seluruh sampel berhasil diklasifikasikan dengan benar sebanyak 114 data tanpa adanya prediksi salah (*false positive*). Hasil ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengidentifikasi kelas negatif karena tidak menghasilkan *false positive*. Namun, masih terdapat 3 kasus hipertensi yang tidak terdeteksi (*false negative*), sehingga kemampuan model dalam mengenali seluruh kasus positif sedikit lebih rendah dibandingkan model imputasi. Meskipun demikian, jumlah kesalahan yang dihasilkan tetap sangat kecil sehingga performa model secara keseluruhan masih tergolong sangat baik.



Gbr. 5 *Confusion matrix drop*

Secara keseluruhan, kedua model menunjukkan performa klasifikasi yang sangat baik dengan jumlah kesalahan yang relatif kecil. Model imputasi menghasilkan jumlah *false negative* yang lebih rendah dibandingkan model *drop missing value,* sehingga lebih efektif dalam mendeteksi kasus hipertensi. Sebaliknya, *model drop missing value* tidak menghasilkan *false positive*, namun masih memiliki jumlah *false negative* yang lebih tinggi. Hasil ini menunjukkan bahwa strategi imputasi memberikan keseimbangan yang lebih baik dalam mendeteksi kasus positif dan mempertahankan performa model secara keseluruhan.

4. Keterbatasan Masalah

Penelitian ini memiliki beberapa keterbatasan. Pertama, evaluasi model hanya dilakukan pada satu dataset, yaitu *Hypertension Risk Prediction Dataset*, sehingga hasil yang diperoleh belum tentu dapat digeneralisasikan pada dataset kesehatan lain dengan karakteristik yang berbeda. Kedua, penelitian ini hanya membandingkan dua pendekatan penanganan *missing value*, yaitu imputasi kategori dan penghapusan data yang mengandung nilai kosong. Metode penanganan *missing value* lainnya belum dievaluasi sehingga efektivitasnya relatif terhadap CatBoost masih perlu diteliti lebih lanjut.

Ketiga, penelitian hanya menggunakan algoritma CatBoost sebagai model klasifikasi. Selain itu, *missing value* pada dataset hanya ditemukan pada variabel Medication, sehingga hasil penelitian ini masih terbatas pada karakteristik data tersebut.

IV. KESIMPULAN

Penelitian ini bertujuan untuk mengevaluasi pengaruh strategi penanganan *missing value* terhadap kinerja algoritma CatBoost dalam prediksi risiko hipertensi. Hasil penelitian menunjukkan bahwa kedua strategi, yaitu imputasi dan penghapusan data yang mengandung *missing value*, mampu menghasilkan performa klasifikasi yang sangat baik dengan nilai evaluasi yang tinggi pada seluruh metrik yang digunakan, yaitu *accuracy, precision, recall*, dan *F1-score*.

Namun, strategi imputasi memberikan hasil yang lebih optimal dibandingkan strategi penghapusan data. Hal ini ditunjukkan oleh nilai *accuracy, precision*, dan *F1-score* yang sedikit lebih tinggi serta kemampuan mempertahankan seluruh jumlah data pelatihan, sehingga model dapat mempelajari pola data secara lebih lengkap dan representatif.

Analisis *confusion matrix* juga menunjukkan bahwa strategi imputasi menghasilkan jumlah *false negative* yang lebih rendah dibandingkan strategi penghapusan data. Hal ini menunjukkan bahwa strategi imputasi lebih efektif dalam mendeteksi kasus hipertensi yang berisiko tidak teridentifikasi, yang dalam konteks kesehatan merupakan aspek penting untuk mendukung deteksi dini.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa strategi imputasi lebih direkomendasikan dalam penanganan *missing value* pada dataset hipertensi berbasis CatBoost karena mampu memberikan keseimbangan yang lebih baik antara performa klasifikasi, stabilitas model, dan kemampuan deteksi kasus positif. Hal ini juga mengindikasikan bahwa penanganan *missing value* tetap memiliki pengaruh signifikan terhadap performa model meskipun algoritma CatBoost memiliki kemampuan internal dalam menangani data hilang.

ISSN (Online) 2747-0563
Volume 6 Tahun 2026

REFERENSI

[1] "Global report on hypertension The race against a silent killer."
[2] World Health Organization, "Hypertension," World Health Organization. Accessed: Dec. 16, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension
[3] A. A. Lukito, "PANDUAN PROMOTIF DAN PREVENTIF HIPERTENSI 2023 Editor."
[4] U. Qalsum and W. Abidin, "Klasifikasi Penyakit Hipertensi Menggunakan Metode K-Means Clustering."
[5] L. ' Lu' Anjeli and F. Rizki, "ANALISIS PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN PERBANDINGAN ALGORITMA MACHINE LEARNING ANALYSIS OF HEART DISEASE PREDICTION USING A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS," vol. 4, no. 2, 2025, [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-
[6] T. Pustaka, H. Usia, M. Rahmawati, and R. P. Kasih, "Galenical is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License," 2023.
[7] C. Casmuti and A. I. Fibriana, "Kejadian Hipertensi di Wilayah Kerja Puskesmas Kedungmundu Kota Semarang," *HIGEIA (*Journal *of Public Health Research and Development)*, vol. 7, no. 1, pp. 123–134, Jan. 2023, doi: 10.15294/higeia.v7i1.64213.
[8] H. Muftisany, T. F. Efendi, and N. A. Rozaq Rais, "Perbandingan Kinerja Algoritma Random Forest, AdaBoost, dan Gradient Boosting dalam Memprediksi Risiko Penyakit Hipertensi," *Faktor Exacta*, vol. 18, no. 2, p. 161, Oct. 2025, doi: 10.30998/faktorexacta.v18i2.28959.
[9] F. V. Ongkosianbhadra and C. C. Lestari, "Pengembangan Model Prediksi Risiko Hipertensi Menggunakan Algoritma Gradient Boosting Decision Tree Yang Dioptimalkan," *Jurnal Informatika dan Sistem Informasi*, vol. 9, no. 2, pp. 90–99, Dec. 2023, doi: 10.37715/juisi.v9i2.4403.
[10] R. Maulana Yusup and E. Rijanto, "Analisis Komparatif Model Pembelajaran Mesin Untuk Memprediksi Hipertensi Ke Dalam Empat Kelas Berdasarkan JNC 8".
[11] M. A. Pradana, R. I. Maulana, R. S. Putra, S. Subairi, and F. T. Anggraeny, "Klasifikasi Penyakit Tanaman Tomat Menggunakan Metode Convolutional Neural Network (CNN) VGG16," *KERNEL: Jurnal Riset Inovasi Bidang Informatika dan Pendidikan Informatika*, vol. 4, no. 2, pp. 61–69, Dec. 2023, doi: 10.31284/j.kernel.2023.v4i2.6829.
[12] F. T. Anggraeny, I. Y. Purbasari, M. Syahrul Munir, F. Muttaqin, E. Prakarsa Mandyarta, and A. Akbar, "Analysis of Simple Data Imputation in Disease Dataset," 2018.
[13] M. L. Pratama, Y. V. Via, and E. P. Mandyartha, "ANALISIS PERFORMANSI NAIVE BAYES DAN RANDOM FOREST TERHADAP SENTIMEN KENAIKAN HARGA BBM DI INDONESIA 1."
[14] A. S. Alfath, A. K. Wardhana, and Rumini, "Hypertension Risk Prediction Using Stacking Ensemble of CatBoost, XGBoost, and LightGBM: A Machine Learning Approach," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 6, pp. 3146–3156, Dec. 2025.
[15] E. Crossesa and A. Sofro, "Application of XGBoost and CatBoost Algorithms for Elderly Hypertension Classification on IFLS 5 Data," *Leibniz: Jurnal Matematika*, vol. 6, no. 1, pp. 1–14, Jan. 2026.
[16] CatBoost Team, "Missing values processing," CatBoost Documentation. [Online]. Available: https://catboost.ai/en/docs/concepts/algorithm-missing-values-processing.html.