Implementasi LightGBM dengan KNN *Imputation* untuk Deteksi Dini Risiko Kehamilan

Syahnur Alawiyah[1], Dian Yuliati[2*], Nurissaidah Ulinnuha[3]
Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Ampel Surabaya
Jl. Dr. Ir. H. Soekarno No. 682, Surabaya 60294, Jawa Timur, Indonesia
Email: [1]syahnuralawiyah01@gmail.com, [2]dian.yuliati@uinsa.ac.id, [3]nuris.ulinnuha@uinsa.ac.id

Abstract. *Pregnancy risks are a critical issue in maternal health that contributes to high rates of maternal and infant morbidity and mortality; therefore, accurate analytical methods are needed for early detection. This study aims to develop and evaluate a pregnancy risk classification model using K-Nearest Neighbor Imputation (KNNI) to handle missing values and LightGBM as the primary method. The model was optimized through parameter tuning and evaluated using Stratified K-Fold Cross-Validation (SKCV). The results show that the proposed model achieved an accuracy of 97.64%, demonstrating excellent performance in classifying pregnancy risk levels. Thus, the approach used has the potential to be developed as a decision support system in the field of maternal health.*
Keywords: *High-Risk Pregnancy, KNNI, LightGBM, SKCV.*

Abstrak. *Risiko kehamilan merupakan isu penting dalam kesehatan maternal yang berkontribusi pada tingginya angka kesakitan dan kematian ibu serta bayi, sehingga diperlukan metode analisis yang akurat untuk deteksi dini. Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi model klasifikasi tingkat risiko kehamilan dengan menggunakan K-Nearest Neighbor Imputation (KNNI) untuk menangani missing value dan LightGBM sebagai metode utama. Model dioptimalkan melalui uji parameter dan dievaluasi menggunakan Stratified K-Fold Cross-Validation (SKCV). Hasil penelitian menunjukkan bahwa model yang diusulkan mampu mencapai akurasi sebesar 97,64%, sehingga menunjukkan kinerja yang sangat baik dalam mengklasifikasikan tingkat risiko kehamilan. Dengan demikian, pendekatan yang digunakan memiliki potensi untuk dikembangkan sebagai sistem pendukung keputusan dalam bidang kesehatan maternal.*
Kata Kunci: *Kehamilan Berisiko Tinggi, KNNI, LightGBM, SKCV.*

1. Pendahuluan

Kesehatan kehamilan merupakan komponen penting dalam kesehatan maternal karena berpengaruh langsung terhadap keselamatan ibu dan bayi [1]. Selama masa kehamilan, tubuh wanita akan mengalami berbagai perubahan fisiologis dan metabolik yang signifikan, sehingga meningkatkan kerentanan terhadap berbagai komplikasi medis seperti hipertensi kehamilan, diabetes gestasional, anemia, dan infeksi [2]. Kondisi tersebut menyebabkan kehamilan memiliki tingkat risiko yang beragam, mulai dari risiko rendah hingga risiko tinggi. Apabila risiko tersebut tidak terdeteksi secara dini, maka dapat menimbulkan komplikasi serius yang berpotensi menyebabkan kelahiran prematur, keguguran, hingga kematian ibu dan bayi [2]. Oleh karena itu, deteksi dini terhadap risiko kehamilan menjadi langkah penting dalam upaya meningkatkan kualitas pelayanan kesehatan maternal.

Permasalahan kesehatan maternal masih menjadi tantangan besar di berbagai negara di dunia. WHO melaporkan bahwa lebih dari 260.000 perempuan meninggal selama dan setelah kehamilan serta persalinan pada tahun 2023, atau setara dengan satu kematian setiap dua menit [3]. Sebagian besar kematian tersebut disebabkan oleh kondisi yang sebenarnya dapat dicegah melalui deteksi dini dan penanganan yang tepat, seperti hipertensi kehamilan, perdarahan, dan infeksi. Selain itu, komplikasi selama kehamilan juga berdampak pada kesehatan bayi yang dilahirkan. UNICEF melaporkan bahwa pada tahun 2023 angka kematian neonatal secara global

mencapai sekitar 17 kematian per 1.000 kelahiran hidup [4]. Data tersebut menunjukkan bahwa upaya identifikasi dan pengelolaan risiko kehamilan masih perlu ditingkatkan untuk mengurangi angka kematian ibu dan bayi.

Kondisi serupa juga terjadi di Indonesia, di mana kesehatan maternal masih menjadi salah satu prioritas dalam pembangunan kesehatan nasional. Berdasarkan data Badan Pusat Statistik, Angka Kematian Ibu (AKI) di Indonesia mencapai 189 per 100.000 kelahiran hidup, yang masih belum memenuhi target *Sustainable Development Goals* (SDGs) sebesar 70 per 100.000 kelahiran hidup pada tahun 2030 [5]. Selain itu, Angka Kematian Bayi (AKB) di Indonesia juga masih relatif tinggi, yaitu sebesar 16,85 per 1.000 kelahiran hidup [6]. Kondisi ini mengindikasikan bahwa proses identifikasi risiko kehamilan belum sepenuhnya optimal sehingga diperlukan pendekatan yang lebih efektif dalam menganalisis faktor-faktor yang berkontribusi terhadap risiko kehamilan.

Seiring perkembangan teknologi informasi, metode *machine learning* semakin banyak dimanfaatkan dalam bidang kesehatan untuk menganalisis data medis dan mengidentifikasi pola yang kompleks [7]. Salah satu algoritma yang sering digunakan adalah *Light Gradient Boosting Machine* (LightGBM), yang dikenal karena memiliki kecepatan komputasi tinggi, efisiensi memori, dan mampu menangani data berdimensi besar dengan membangun pohon keputusan secara bertahap menggunakan metode *leaf-wise growth* yang memprioritaskan pemisahan pada daun dengan nilai *loss* terbesar [8]. Meskipun LightGBM sering digunakan pada *dataset* berskala besar, dalam konteks penelitian ini keunggulan tersebut tidak menjadi fokus utama, namun metode ini tetap dipilih karena kemampuannya dalam menghasilkan model klasifikasi yang akurat dan stabil. Penelitian oleh Noviandy menunjukkan bahwa metode LightGBM mampu menghasilkan performa klasifikasi yang baik dalam mendeteksi risiko kehamilan dengan tingkat akurasi sebesar 84.73% [9]. Selain itu, penelitian Kanber dalam mendiagnosis kanker payudara menunjukkan bahwa LightGBM menghasilkan akurasi sebesar 95.98% [10]. Temuan tersebut menunjukkan bahwa LightGBM memiliki potensi yang baik dalam berbagai permasalahan klasifikasi di bidang kesehatan.

Penerapan metode *machine learning* pada data kesehatan juga sering kali menghadapi permasalahan kualitas data, terutama adanya *missing value*. Data yang tidak lengkap dapat menurunkan performa model klasifikasi jika tidak ditangani dengan metode yang tepat [11]. Salah satu metode yang dapat digunakan untuk mengatasi permasalahan tersebut adalah *K-Nearest Neighbor Imputation* (KNNI), yaitu metode imputasi yang mengisi nilai hilang berdasarkan kemiripan dengan data tetangga terdekat sehingga pola data tetap terjaga [12]. Salah satu penelitian yang menerapkan KNNI dilakukan oleh Chen, yang menunjukkan bahwa penggunaan KNNI mampu meningkatkan performa model klasifikasi secara signifikan dibandingkan dengan metode penghapusan *missing value* dengan peningkatan akurasi dari 79.93% menjadi sebesar 99.41% [13].

Meskipun berbagai penelitian telah memanfaatkan metode *machine learning* dalam klasifikasi kesehatan, sebagian besar masih berfokus pada algoritma klasifikasi tanpa menangani secara optimal permasalahan kualitas data, khususnya *missing value*. Padahal, data yang tidak lengkap dapat menurunkan performa model. Selain itu, penelitian yang mengombinasikan teknik imputasi data dengan algoritma LightGBM dalam deteksi risiko kehamilan masih relatif terbatas, serta cenderung menggunakan teknik imputasi sederhana sehingga belum mampu mengoptimalkan keterkaitan antar data, sementara penelitian ini menawarkan integrasi KNNI dan LightGBM dalam satu kerangka kerja dengan evaluasi yang tidak hanya meningkatkan akurasi, tetapi juga kualitas data setelah imputasi. Dengan demikian, penelitian ini menerapkan metode KNNI untuk menangani *missing value* serta menggunakan algoritma LightGBM untuk deteksi dini risiko kehamilan. Kombinasi kedua metode ini diharapkan mampu meningkatkan kualitas data dan performa model klasifikasi sehingga dapat mendukung pengambilan keputusan klinis serta perumusan strategi kesehatan ibu hamil secara lebih tepat.

2. Tinjauan Pustaka

2.1. Risiko Kehamilan

Risiko kehamilan merupakan kondisi yang menunjukkan adanya kemungkinan terjadinya gangguan kesehatan selama masa kehamilan yang dapat memengaruhi keselamatan ibu maupun janin [14]. Risiko ini dipengaruhi oleh berbagai faktor fisiologis dan klinis seperti kadar gula darah, suhu tubuh, tekanan darah, denyut jantung, serta indeks massa tubuh ibu. Perubahan kondisi tersebut dapat menjadi indikator adanya komplikasi medis, misalnya hipertensi gestasional, diabetes gestasional, anemia, maupun kelahiran prematur [2]. Apabila kondisi tersebut tidak terdeteksi sejak awal, maka risiko komplikasi dapat meningkat dan berpotensi menimbulkan dampak serius bagi kesehatan ibu dan bayi. Oleh karena itu, identifikasi faktor risiko secara dini menjadi langkah penting dalam upaya pencegahan komplikasi selama kehamilan.

2.2. *Preprocessing*

Preprocessing merupakan tahap awal dalam analisis data yang bertujuan untuk memperbaiki dan meningkatkan kualitas *dataset* sebelum digunakan dalam proses pemodelan [15]. Tahapan ini mencakup kegiatan seperti penanganan nilai hilang (*missing value*), serta transformasi data agar memiliki skala yang seragam. Proses ini penting karena kualitas data yang kurang baik dapat memengaruhi akurasi model *machine learning* [16]. Dengan melakukan *preprocessing* secara sistematis, data yang digunakan dalam penelitian menjadi lebih konsisten dan mampu menggambarkan pola hubungan antarvariabel secara lebih jelas.

2.2.1. *K-Nearest Neighbor Imputation* (KNNI)

KNNI merupakan metode yang digunakan untuk menangani *missing value* dengan memanfaatkan kemiripan antar data dalam *dataset* [12]. Nilai yang hilang digantikan menggunakan nilai dari K tetangga terdekat yang memiliki karakteristik paling mirip dengan observasi yang mengalami kehilangan data [17]. Kemiripan antar data umumnya dihitung menggunakan jarak *Euclidean* dengan terlebih dahulu menentukan jumlah tetangga terdekat (K). Jarak antar observasi dapat dihitung menggunakan Persamaan 1:

$$d_{(x_u,x_v)} = \sqrt{\sum_{l}^{m}(x_{ul} - x_{vl})^2} \tag{1}$$

dengan,

$d_{(x_u,x_v)}$ = jarak antara observasi target (x_u) dan observasi pembanding (x_v),
m = jumlah variabel yang digunakan dalam perhitungan,
x_{ul} = nilai data pada variabel ke-l dari observasi target x_{ul},
x_{vl} = nilai data pada variabel ke-l dari observasi pembanding x_{vl},
l = 1,2,…,m.

Setelah jarak dihitung, dipilih sejumlah tetangga terdekat sebanyak (K). Selanjutnya, nilai imputasi dihitung menggunakan rata-rata nilai dari tetangga tersebut, yang dirumuskan dalam Persamaan 2:

$$\bar{x}_l = \frac{1}{K}\sum_{k=1}^{K} x_k \tag{2}$$

di mana, $\bar{x}_l$ adalah nilai imputasi untuk fitur ke-l dan x_k adalah nilai observasi data yang tidak hilang.

2.2.2. Normalisasi

Normalisasi merupakan proses transformasi data untuk menyamakan perbedaan skala antar variabel dapat memengaruhi kontribusi setiap fitur dalam proses pembelajaran model, sehingga diperlukan penyesuaian agar semua fitur berperan secara seimbang [18]. Salah satu metode yang sering digunakan adalah *Min-Max Normalization*, yaitu teknik yang mentransformasikan nilai data ke dalam rentang tertentu, biasanya antara 0 hingga 1 [19]. Persamaan normalisasi dapat dituliskan sebagai Persamaan 3:

$$x_b = \frac{x_s - x_{min}}{x_{max} - x_{min}} \tag{3}$$

di mana, x_s adalah data setelah penangan *missing value*, x_{min} merupakan data terkecil dari semua data, sedangkan x_{max} merupakan data terbesarnya

2.3. *Stratified K-Fold Cross-Validation* (SKCV)

SKCV merupakan pengembangan dari metode *K-Fold Cross-Validation* yang dirancang untuk permasalahan klasifikasi dengan ketidakseimbangan distribusi kelas. Metode ini bekerja dengan membagi data ke dalam beberapa *fold* sambil mempertahankan proporsi tiap kelas pada setiap *fold* agar menyerupai distribusi keseluruhan *dataset* [20]. Dengan demikian, setiap data digunakan secara bergantian sebagai data *training* dan *testing*, sehingga mengurangi bias terhadap kelas mayoritas serta menghasilkan evaluasi model yang lebih stabil, objektif, dan reliabel.

2.4. *Light Gradient Boosting Machine* (LightGBM)

LightGBM adalah algoritma *gradient boosting* berbasis pohon keputusan yang dikembangkan oleh Microsoft dan dikenal memiliki kecepatan pelatihan tinggi serta efisiensi memori pada *dataset* berdimensi besar [8]. LightGBM menggunakan strategi *leaf-wise growth*, yaitu membagi *node* berdasarkan daun yang memberikan penurunan *loss* terbesar sehingga menghasilkan model yang lebih optimal [21]. Pada klasifikasi biner, LightGBM menggunakan fungsi *binary log loss* untuk mengukur kesalahan prediksi (lihat Persamaan 4):

$$L(Y, \hat{P}) = -\left[Y\, log(\hat{P}) + (1 - Y) log(1 - \hat{P})\right] \tag{4}$$

dengan Y sebagai label aktual dan $\hat{P}$ sebagai probabilitas prediksi model. Pada tahap awal, model melakukan inisialisasi probabilitas berdasarkan proporsi kelas target (N_c) pada data yang dibagi dengan total jumlah data (n) (lihat Persamaan 5).

$$\hat{P} = \frac{N_c}{n} \tag{5}$$

Selanjutnya dihitung nilai residual *(R)* dan *hessian (H)*, serta *information gain* untuk menentukan pembelahan *node* (lihat Persamaan 6-8).

$$R = \hat{P} - Y \tag{6}$$

$$H = \hat{P} \times \left(1 - \hat{P}\right) \tag{7}$$

$$Gain = \left[\frac{R_{kiri}^2}{H_{kiri} + \lambda} + \frac{R_{kanan}^2}{H_{kanan} + \lambda} - \frac{(R_{kiri} + R_{kanan})^2}{H_{kiri} + H_{kanan} + \lambda}\right] - \gamma \tag{8}$$

di mana, parameter λ berfungsi sebagai regulasi untuk menghindari *overfitting*, dan γ (*minimum split gain*) digunakan untuk mengontrol besar perubahan bobot model.

Selanjutnya, pembelahan node dilakukan berdasarkan nilai *gain* tertinggi, dan jika nilai $Gain < 0$ atau jumlah data tidak memenuhi batas minimum, maka node tersebut ditetapkan sebagai *leaf node* dan dilanjutkan menghitung daun *(d)* (lihat Persamaan 9):

$$d = -\frac{\sum_{m=1}^{g} R_m}{\sum_{m=1}^{g} H_m + \lambda} \tag{9}$$

Selanjutnya dihitung nilai logit *(b)* (lihat Persamaan 10) yang kemudian diakhiri dengan probabilitas prediksi diperoleh menggunakan fungsi sigmoid *($\hat{P}$)* (lihat Persamaan 11):

$$b = \sum_{t=1}^{S} d_t \times \kappa \tag{10}$$

$$\hat{P} = \frac{1}{1 + e^{-b}} \tag{11}$$

di mana, κ adalah *learning rate* dan probabilitas yang diperoleh digunakan untuk menentukan kelas prediksi berdasarkan ambang batas 0.5.

2.5. Confusion Matrix

Confusion Matrix adalah teknik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi dengan membandingkan hasil prediksi dengan label sebenarnya [22]. Informasi tersebut disajikan dalam Tabel 1.

Tabel 1. Confusion Matrix

Kelas Sebenarnya	Kelas Prediksi	
	Low	***High***
Low	TN	FP
High	FN	TP

di mana,
TN = data risiko *Low* yang diprediksi *Low*.
FP = data risiko *Low* yang diprediksi *High*.
FN = data risiko *High* yang diprediksi *Low*.
TP = data risiko *High* yang diprediksi *High*.

Berdasarkan nilai tersebut, beberapa metrik evaluasi dapat dihitung, antara lain akurasi yang mengukur proporsi prediksi yang benar terhadap seluruh data (lihat Persamaan 12), sensitivitas (*recall*) yang menunjukkan kemampuan model dalam mengidentifikasi kelas risiko *High* (lihat Persamaan 13), serta spesifisitas yang mengukur kemampuan model dalam mengidentifikasi kelas risiko *Low* (lihat Persamaan 14). Selain itu, digunakan pula presisi. untuk mengukur tingkat ketepatan model dalam memprediksi kelas risiko *High*, yaitu seberapa banyak prediksi positif yang benar-benar sesuai (lihat Persamaan 15). Untuk melengkapi evaluasi, digunakan *F1-score* yang merupakan rata-rata harmonik antara presisi dan sensitivitas, sehingga memberikan gambaran keseimbangan kinerja model, terutama pada kondisi data yang tidak seimbang (lihat Persamaan 16). Selain itu, digunakan pula ROC-AUC (*Receiver Operating Characteristic – Area Under Curve*) (lihat Persamaan 17) yang mengukur kemampuan model dalam membedakan kelas secara keseluruhan berdasarkan hubungan antara sensitivitas dan *False Positive Rate* (FPR), tanpa bergantung pada threshold tertentu [23].

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{12}$$

$$Sensitivitas = \frac{TP}{TP + FN} \times 100\% \tag{13}$$

$$Spesifisitas = \frac{TN}{TN + FP} \times 100\% \tag{14}$$

$$Presisi = \frac{TP}{TP + FP} \times 100\% \tag{15}$$

$$F1-score = 2 \times \frac{Precision \times Sensitivitas}{Precision + Sensitivitas} \times 100\% \quad (16)$$

$$ROC\text{-}AUC = \sum_{i=1}^{n-1} \frac{Sensitivitas_{\ i+1} + Sensitivitas}{2} \times (FPR_{\ i+1} - FPR) \quad (17)$$

dengan $FPR = \frac{FP}{FP+TN}$

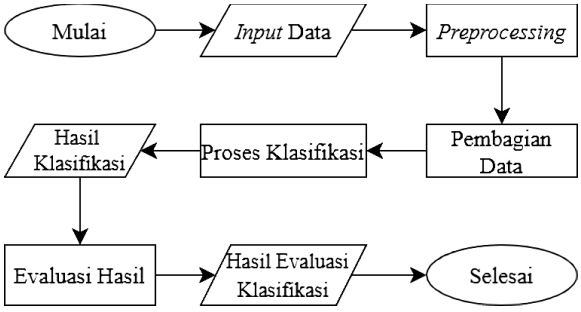
3. Metodologi Penelitian

Dataset yang digunakan dalam penelitian ini berupa data sekunder yang diperoleh dari platform Mendeley Data [24]. Data berisi informasi kesehatan ibu hamil berupa variabel klinis, fisiologis, dan riwayat kesehatan yang terdiri dari 1187 data pasien dengan 11 variabel prediktor dan 1 variabel respons. Variabel respons menunjukkan status risiko kehamilan yang diklasifikasikan menjadi dua kategori, yaitu risiko rendah (*low*) dan risiko tinggi (*high*). Dari keseluruhan data, sebanyak 713 data termasuk risiko rendah dan 474 data termasuk risiko tinggi. *Dataset* juga mengandung 24 *missing value* yang perlu ditangani sebelum proses analisis. Variabel prediktor yang digunakan dalam penelitian ini ditunjukkan pada Tabel 2.

Tabel 2. Variabel Prediktor Penelitian

No	Variabel	Deskripsi	Range
1	*Age* (X_1)	Usia pasien (tahun)	10–65
2	*Systolic* BP (X_2)	Tekanan darah sistolik (mmHg)	70–200
3	*Diastolic* (X_3)	Tekanan darah diastolik (mmHg)	40–140
4	BS (X_4)	Kadar gula darah (mg/dL)	54–342
5	*Body Temp* (X_5)	Suhu tubuh (°C)	36.1–39.4
6	BMI (X_6)	*Body Mass Index* (kg/m²)	15–37
7	*Previous Complications* (X_7)	Riwayat komplikasi kehamilan	0 = Tidak, 1 = Ya
8	*Preexisting Diabetes* (X_8)	Riwayat diabetes sebelum kehamilan	0 = Tidak, 1 = Ya
9	*Gestational Diabetes* (X_9)	Diabetes selama kehamilan	0 = Tidak, 1 = Ya
10	*Mental Health* (X_{10})	Masalah kesehatan mental	0 = Tidak, 1 = Ya
11	*Heart Rate* (X_{11})	Denyut jantung (bpm)	58–92

Tahapan penelitian secara umum ditunjukkan pada Gambar berikut:



Gambar 1. Diagram Alir Penelitian

Berdasarkan diagram alir tersebut, penelitian dimulai dengan *input* data yang diperoleh dari Mendeley Data. Tahap *preprocessing* meliputi penanganan *missing value* menggunakan metode *K-Nearest Neighbor Imputation* (KNNI) dengan nilai $K = 5$. Pemilihan nilai $K = 5$ didasarkan pada pertimbangan keseimbangan antara sensitivitas terhadap *noise* dan kemampuan mempertahankan pola lokal data, sehingga nilai ini dianggap mampu menghasilkan imputasi yang stabil tanpa menghilangkan karakteristik data [25]. Selanjutnya, dilakukan normalisasi data numerik ke dalam rentang 0 hingga 1 untuk menyamakan skala antarvariabel dan meningkatkan kestabilan proses pembelajaran model. Data kemudian dibagi menggunakan

metode *Stratified K-Fold Cross-Validation* (SKCV) dengan $k = 10$, yang dipilih karena mampu memberikan estimasi performa yang lebih stabil serta menjaga proporsi distribusi kelas pada setiap *fold* [26].

Proses klasifikasi dilakukan menggunakan metode LightGBM dengan uji coba beberapa kombinasi parameter, yaitu *n_estimators*, *learning_rate*, *max_depth*, dan *min_child_samples*. Nilai parameter yang digunakan dalam proses pengujian ditunjukkan pada Tabel 3, di mana nilai parameter ditentukan berdasarkan literatur dan dokumentasi LightGBM serta mempertimbangkan keseimbangan antara kompleksitas model dan kemampuan generalisasi. Proses *hyperparameter tuning* dilakukan secara terpisah pada setiap *fold*, di mana kombinasi parameter hanya diterapkan pada data *training* dalam masing-masing *fold*, sedangkan data pada *fold* tersebut digunakan secara eksklusif untuk evaluasi model guna menghindari kebocoran data (*data leakage*) antara data *training* dan data *testing*. Selanjutnya, kinerja model dievaluasi menggunakan *confusion matrix* dengan menghitung akurasi, sensitivitas, spesifisitas, presisi, *F1-score, dan* ROC-AUC.

Tabel 3. Parameter Pengujian LightGBM

Parameter	*Value*	Fungsi	Sumber
n_estimators	[100, 250, 500]	Menentukan jumlah pohon keputusan yang digunakan dalam proses *boosting*.	[27]
learning_rate	[0.01, 0.05, 0.1]	Mengatur besarnya kontribusi setiap pohon dalam memperbarui model pada setiap iterasi.	[28]
max_depth	[5, 10, 15]	Menentukan kedalaman maksimum pohon keputusan untuk mengontrol kompleksitas model.	[29]
min_child_samples	[100, 250, 500]	Menentukan jumlah minimum sampel pada sebuah *leaf node* untuk mencegah *overfitting*.	[30]

4. Hasil dan Diskusi

Berdasarkan Gambar 1, tahap setelah *input* data dengan sampel data yang digunakan dalam penelitian ini disajikan pada Tabel 4 dilakukan tahap *preprocessing* yang meliputi penanganan *missing value* dan normalisasi. Terdapat 24 *missing value* yang tersebar dalam 21 baris atau sebanyak 0.17% dari keseluruhan data yang kemudian ditangani dengan metode KNNI sehingga didapatkan data yang lebih lengkap seperti pada Tabel 5. Namun, karena jumlah *missing value* sangat kecil (0.17%), pengaruh metode imputasi terhadap hasil analisis kemungkinan tidak terlalu besar. Selanjutnya, pada variabel numerik dinormalisasi menggunakan Min-Max untuk memastikan distribusi data yang lebih seragam, seperti yang ditunjukkan pada Tabel 6.

Tabel 4. Sampel Data Penelitian

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	Y
1	22	90	60	162	37.8	18	1	1	0	1	80	*High*
2	22	110	70	128	36.7	20.4	0	0	0	0	74	*Low*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
15	22	100	60	128	36.7	*NaN*	0	0	0	0	74	*Low*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1186	23	130	100	92	36.7	27	0	0	1	1	60	*High*
1187	26	120	90	121	36.7	23.9	0	0	1	0	58	*High*

Tabel 5. Data Penelitian Setelah Penanganan *Missing Value*

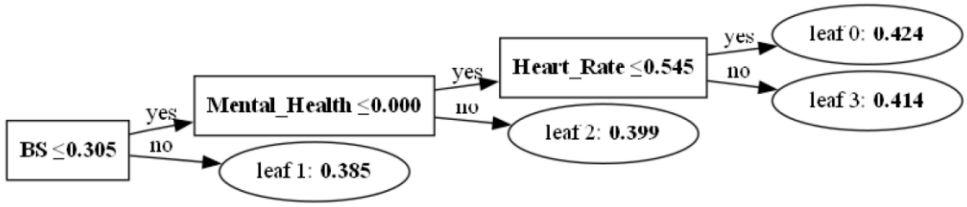
No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	Y
1	22	90	60	162	37.8	18	1	1	0	1	80	*High*
2	22	110	70	128	36.7	20.4	0	0	0	0	74	*Low*

⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
15	22	100	60	128	36.7	22.88	0	0	0	0	74	*Low*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1186	23	130	100	92	36.7	27	0	0	1	1	60	*High*
1187	26	120	90	121	36.7	23.9	0	0	1	0	58	*High*

Tabel 6. Data Penelitian Setelah Normalisasi

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	Y
1	0.22	0.15	0.2	0.38	0.52	0.14	1	1	0	1	0.65	*High*
2	0.22	0.31	0.3	0.26	0.18	0.25	0	0	0	0	0.47	*Low*
3	0.31	0.31	0.3	0.28	0.18	0.36	1	0	0	0	0.41	*Low*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1185	0.2	0.62	0.7	0.12	0.18	0.3	0	0	1	1	0.24	*High*
1186	0.24	0.46	0.6	0.13	0.18	0.55	0	0	1	1	0.06	*High*
1187	0.29	0.38	0.5	0.23	0.18	0.4	0	0	1	0	0	*High*

Sebelum melakukan klasifikasi data dibagi menjadi data *training* dan *testing* menggunakan SKCV dengan $k = 10$ untuk memperoleh evaluasi model yang lebih stabil. Pada tahap pelatihan model dilakukan menggunakan algoritma LightGBM membangun serangkaian pohon keputusan secara bertahap, di mana setiap pohon berfungsi untuk memperbaiki kesalahan dari pohon sebelumnya. Sebagai ilustrasi, ditampilkan contoh pohon pertama yang terbentuk pada model LightGBM.



Gambar 2. Pohon Keputusan LightGBM

Berdasarkan Gambar 2, model membagi data mulai dari variabel $X_4 \leq 0.305$, kemudian dilanjutkan oleh $X_{10} \leq 0$, dan $X_{10} \leq 0.545$ hingga mencapai *leaf node* sebagai hasil prediksi. Hal ini mengindikasikan bahwa model memanfaatkan beberapa variabel secara bertingkat dalam proses klasifikasi. Setelah itu, dalam menentukan hasil evaluasi terbaik dilakukan uji coba parameter pada Tabel 3 untuk menentukan kombinasi parameter yang menghasilkan performa optimal. Kombinasi parameter tersebut disajikan di bawah ini pada Tabel 7. Model dengan parameter terbaik tersebut kemudian digunakan untuk melakukan klasifikasi, dan kinerjanya dievaluasi menggunakan metrik seperti akurasi, sensitivitas, dan spesifisitas yang disajikan dalam Tabel 8.

Tabel 7. Kombinasi Parameter Terbaik

Metode	*n_estimators*	*learning_rate*	*max_depth*	*min_child_samples*
LightGBM+Hapus *Missing Value*	100	0.1	10	100
LightGBM+KNNI	100	0.1	5	100

Tabel 8. Hasil Evaluasi Parameter Terbaik

Metode	Akurasi	Sensitivitas	Spesifisitas	Presisi	*F1-score*	*ROC-AUC*
LightGBM+ Hapus *Missing Value*	97.42%	96.31%	98.16%	97.37%	96.65%	0.99

LightGBM+KNNI	97.64%	96.81%	98.18%	97.37%	96.95%	0.99

Berdasarkan Tabel 8, hasil evaluasi yang didapatkan menunjukkan bahwa proses imputasi menggunakan metode KNNI mampu membantu mengatasi permasalahan *missing value* sehingga kualitas data menjadi lebih baik dan berdampak pada peningkatan performa model dalam melakukan klasifikasi risiko kehamilan. Peningkatan ini terjadi karena KNNI mengisi nilai yang hilang berdasarkan kedekatan antardata sehingga hubungan antarvariabel tetap terjaga, yang memungkinkan model menghasilkan prediksi lebih akurat serta berpotensi mendukung deteksi dini secara klinis. Meskipun demikian, peningkatan performa yang diperoleh relatif kecil, yaitu sebesar 0.22% (dari 97.42% menjadi 97.64%), sehingga belum menunjukkan peningkatan yang signifikan secara praktis dan berpotensi dipengaruhi oleh variasi data (*noise*), mengingat proporsi *missing value* yang sangat rendah. Dengan demikian, pada kondisi *missing value* yang sangat rendah, metode sederhana seperti penghapusan data juga dapat memberikan hasil yang sebanding. Selanjutnya, untuk memperkuat hasil penelitian ini, dilakukan perbandingan dengan beberapa penelitian terdahulu yang menggunakan metode berbeda.

Penelitian ini menunjukkan performa yang baik dalam klasifikasi risiko kehamilan. Berdasarkan penelitian oleh Arif yang menggunakan *dataset* sama, metode XGBoost memperoleh akurasi sebesar 96,36%. Oleh karena itu, perbandingan dalam penelitian ini difokuskan pada studi yang menggunakan *dataset* yang sama agar interpretasi hasil lebih valid dan tidak menimbulkan bias. Adapun perbandingan hasil disajikan pada Tabel 9.

Tabel 9. Perbandingan Hasil Penelitian

Peneliti	Metode	Hasil
Arif* [31]	XGBoost	96.36%
Penelitian ini	LightGBM+KNNI	97.64%

Penelitian ini juga masih memiliki beberapa keterbatasan yang perlu diperhatikan. Keterbatasan jumlah dan cakupan *dataset* dapat memengaruhi kemampuan generalisasi model ketika diterapkan pada populasi yang lebih luas, sementara potensi bias data, seperti ketidakseimbangan distribusi kelas juga dapat memengaruhi hasil prediksi. Selain itu, metode yang digunakan memiliki keterbatasan, di mana KNNI sensitif terhadap pemilihan parameter *K* dan LightGBM berpotensi mengalami *overfitting* jika parameter model tidak diatur secara optimal. Oleh karena itu, penelitian selanjutnya disarankan menggunakan *dataset* yang lebih besar dan beragam serta melakukan eksplorasi parameter secara lebih optimal.

5. Kesimpulan dan Saran

Hasil penelitian ini menunjukkan bahwa penerapan metode KNNI dengan (K=5) efektif dalam menangani *missing value* dengan menghasilkan nilai imputasi yang tetap berada dalam rentang data, sehingga kualitas data menjadi lebih baik untuk proses klasifikasi. Hal tersebut dibuktikan dengan kombinasi parameter terbaik pada *n_estimators* = 100, *learning_rate* = 0.1, *max_depth* = 5, dan *min_child_samples* = 100 yang menghasilkan akurasi sebesar 97.64%, sensitivitas 96.81%, spesifisitas 98.18%, presisi 97.37%, *F1–score* 96.95%, dan ROC–AUC 0.99. Hasil ini menunjukkan bahwa kombinasi metode KNNI dan LightGBM dapat menjadi pendekatan yang efektif dalam mendeteksi risiko kehamilan secara lebih akurat. Pada penelitian selanjutnya disarankan untuk mengeksplorasi metode imputasi lain seperti *Multiple Imputation* atau *MissForest*, serta membandingkannya dengan KNNI untuk memperoleh hasil yang lebih komprehensif. Selain itu, penerapan algoritma klasifikasi lain seperti *Random Forest*, XGBoost, atau metode *deep learning* juga dapat dipertimbangkan untuk meningkatkan performa klasifikasi dan memperkaya pengembangan model dalam deteksi risiko kehamilan.

Referensi

[1] A. Raza, H. U. R. Siddiqui, K. Munir, M. Almutairi, F. Rustam, and I. Ashraf, "Ensemble Learning-Based Feature Engineering to Analyze Maternal Health During Pregnancy and Health Risk Prediction," *PLoS One*, vol. 17, no. 11, Nov. 2022, doi: 10.1371/journal.pone.0276525.

[2] J. Dol *et al.*, "Timing of Neonatal Mortality and Severe Morbidity During the Postnatal Period: A Systematic Review," *JBI Evid. Synth.*, vol. 21, no. 1, pp. 98–199, Jan. 2023, doi: 10.11124/JBIES-21-00479.

[3] WHO, "Maternal Mortality." Accessed: Oct. 01, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality

[4] UNICEF, "Neonatal Mortality." Accessed: Oct. 05, 2025. [Online]. Available: https://data.unicef.org/topic/child-survival/neonatal-mortality/

[5] Kementerian Kesehatan RI, *Laporan Kinerja Kementrian Kesehatan 2024*. Accessed: Oct. 25, 2025. [Online]. Available: https://repository.badankebijakan.kemkes.go.id/id/eprint/5852/1/LKj%20Kemenkes%202024.pdf

[6] BKKBN, "Laporan Kependudukan Indonesia 2024." Accessed: Oct. 05, 2025. [Online]. Available: https://siperindu.online/2023/pb/unduh_file/Laporan%20Kependudukan%20Indonesia%20-%20IND.pdf

[7] F. Chen, L. Yu, J. Mao, Q. Yang, D. Wang, and C. Yu, "A Novel Data-Characteristic-Driven Modeling Approach for Imputing Missing Value in Industrial Statistics: A Case Study of China Electricity Statistics," *Appl. Energy*, vol. 373, p. 123854, Nov. 2024, doi: 10.1016/j.apenergy.2024.123854.

[8] C. Lokker *et al.*, "Boosting Efficiency in a Clinical Literature Surveillance System with LightGBM," *PLOS Digit Health*, vol. 3, no. 9, p. e0000299, Sep. 2024, doi: 10.1371/journal.pdig.0000299.

[9] T. R. Noviandy, S. I. Nainggolan, R. Raihan, I. Firmansyah, and R. Idroes, "Maternal Health Risk Detection Using Light Gradient Boosting Machine Approach," *Infolitika J. Data Sci*, vol. 1, no. 2, pp. 48–55, Dec. 2023, doi: 10.60084/ijds.v1i2.123.

[10] B. M. Kanber, A. Al Smadi, N. F. Noaman, B. Liu, S. Gou, and M. K. Alsmadi, "LightGBM: A Leading Force in Breast Cancer Diagnosis Through Machine Learning and Image Processing," *IEEE Access*, vol. 12, pp. 39811–39832, 2024, doi: 10.1109/ACCESS.2024.3375755.

[11] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A Survey on Missing Data in Machine Learning," *J. Big Data*, vol. 8, no. 140, pp. 1–37, Oct. 2021, doi: 10.1186/s40537-021-00516-9.

[12] W. Sudrajat and I. Cholid, "K-Nearest Neighbor (K-NN) untuk Penanganan Missing Value pada Data UMKM," *JRSIT*, vol. 1, no. 2, pp. 54–63, Nov. 2023, doi: 10.59407/jrsit.v1i2.77.

[13] X. Chen *et al.*, "Cervical Cancer Detection Using K Nearest Neighbor Imputer and Stacked Ensemble Learning Model," *Digit. Health*, vol. 9, p. 20552076231203800, Jan. 2023, doi: 10.1177/20552076231203802.

[14] K. R. Nanda, "Maternal Age and Risk of Pregnancy Complications: A Qualitative Study," *Advances in Healthcare Research*, vol. 3, no. 2, pp. 132–147, May 2025, doi: 10.60079/ahr.v3i2.488.

[15] I. M. Hamdani, Nurhidayat, A. Karman, N. F. A. H, and A. H. Julyaningsih, "Edukasi dan Pelatihan Data Science dan Data Preprocessing," *Intisari*, vol. 2, no. 1, pp. 19–26, Jun. 2024, doi: 10.58227/intisari.v2i1.125.

[16] D. Liang, X. Jin, Y. Yuan, and R. Zou, "Performance Analysis of Machine Learning Methods," *J. Phys. Conf. Ser.*, vol. 2428, no. 1, p. 012039, 2023, doi: 10.1088/1742-6596/2428/1/012039.

[17] A. Fadlil, Herman, and M. D. Praseptian, "K Nearest Neighbor Imputation Performance on Missing Value Data Graduate User Satisfaction," *Jurnal RESTI*, vol. 6, no. 4, pp. 570–576, Aug. 2022, doi: 10.29207/RESTI.V6I4.4173.

[18] W. Wenny, "Normalisasi Data Kependudukan Dengan Model Min Max Dan Algoritma K-Means Untuk Pengelompokkan Tingkat Ekonomi Masyarakat," *Bios*, vol. 2, no. 2, pp. 53–63, Apr. 2024, doi: 10.62866/bios.v2i2.141.

[19] A. A. G. A. Pranandita and I. M. Widiartha, "Optimasi Metode Support Vector Machine (SVM) Mengunakan Particle Swarm Optimization pada Permasalahan Klasifikasi Diabetes," *Jnatia*, vol. 3, no. 4, pp. 879–888, Aug. 2025, doi: 10.24843/JNATIA.2025.V03.I04.P18.

[20] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042333.

[21] D. Wijayanto and B. P. Hartato, "Analisis Perbandingan Performa Algoritma XGBoost dan LightGBM pada Klasifikasi Kanker Payudara," *The Indonesian Journal of Computer Science*, vol. 13, no. 2, Apr. 2024, doi: 10.33022/ijcs.v13i2.3901.

[22] R. R. Adhitya, W. Witanti, and R. Yuniarti, "Perbandingan Metode CART Dan Naïve Bayes Untuk Klasifikasi Customer Churn," *INFOTECH journal*, vol. 9, no. 2, pp. 307–318, Jul. 2023, doi: 10.31949/infotech.v9i2.5641.

[23] B. S. W. Poetro, S. Mulyono, and V. A. Pramesti, "Prediksi Penyakit Batu Ginjal dengan Menerapkan Convolutional Neural Network," *Jurnal Buana Informatika*, vol. 15, no. 2, pp. 153–162, Oct. 2024, [Online]. Available: https://ojs.uajy.ac.id/index.php/jbi/article/view/9838

[24] M. U. Mojumdar *et al.*, "Maternal Health Risk Assessment Dataset," Mendeley Data. Accessed: Oct. 14, 2025. [Online]. Available: https://data.mendeley.com/datasets/p5w98dvbbk/1

[25] K. Muludi, R. Setianingsih, R. Sholehurrohman, and A. Junaidi, "Exploiting Nearest Neighbor Data and Fuzzy Membership Function to Address Missing Values in Classification," *PeerJ Comput. Sci.*, vol. 10, p. e1968, Mar. 2024, doi: 10.7717/peerj-cs.1968.

[26] S. Widodo, H. Brawijaya, and S. Samudi, "Stratified K-Fold Cross Validation Optimization on Machine Learning for Prediction," *SinkrOn*, vol. 7, no. 4, pp. 2407–2414, Oct. 2022, doi: 10.33395/sinkron.v7i4.11792.

[27] L. Deng, K. Lu, and H. Hu, "An Interpretable LightGBM Model for Predicting Coronary Heart Disease: Enhancing Clinical Decision-Making with Machine learning," *PLoS One*, vol. 20, no. 9 September, Sep. 2025, doi: 10.1371/journal.pone.0330377.

[28] J. Park and E. Hwang, "A Two-Stage Multistep-Ahead Electricity Load Forecasting Scheme Based on LightGBM and Attention-BilSTM," *Sensors*, vol. 21, no. 22, Nov. 2021, doi: 10.3390/s21227697.

[29] E. Ramadanti, D. A. Dinathi, C. Sri, K. Aditya, and R. Chandranegara, "Diabetes Disease Detection Classification Using Light Gradient Boosting (LightGBM) With Hyperparameter Tuning," *SinkrOn*, vol. 8, no. 2, pp. 956–963, Mar. 2024, doi: 10.33395/sinkron.v8i2.13530.

[30] K. T. Nguyen, T. N. Tran, and H. T. Nguyen, "Research on the Influence of Hyperparameters on the LightGBM Model in Load Forecasting," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 5, pp. 17005–17010, Oct. 2024, doi: 10.48084/etasr.8266.

[31] M. Arif, "Explainable AI in Maternal Health: Utilizing XGBoost and SHAP Values for Enhanced Risk Prediction and Interpretation," *Int. J. Emerg. Multidiscip.: Comput. Sci. Artif. Intell.*, vol. 4, no. 1, p. 16, Apr. 2025, doi: 10.54938/ijemdcsai.2025.04.1.419.