



PREDIKSI RISIKO KEHAMILAN BERDASARKAN VARIABEL KLINIS MENGGUNAKAN ALGORITMA CATBOOST

Nafisah Adawiyah¹, Farid Akbar Siregar²

Fakultas Ilmu Komputer Dan Teknologi Informasi, Jurusan Teknologi Informasi
Universitas Muhammadiyah Sumatera Utara

Email: nfsahadwyah@gmail.com

ABSTRAK

Risiko kehamilan merupakan kondisi yang perlu dideteksi sejak dini karena dapat berdampak pada kesehatan ibu dan janin. Tingginya angka komplikasi kehamilan menuntut adanya sistem yang mampu membantu tenaga kesehatan dalam melakukan identifikasi tingkat risiko kehamilan secara cepat dan akurat. Penelitian ini bertujuan untuk membangun model prediksi risiko kehamilan menggunakan algoritma Categorical Boosting (*CatBoost*) berdasarkan variabel klinis ibu hamil. Dataset yang digunakan diperoleh dari platform Kaggle dengan jumlah 1.014 data yang terdiri dari fitur Age, SystolicBP, DiastolicBP, BS (Blood Sugar), BodyTemp, dan HeartRate, serta target RiskLevel yang diklasifikasikan ke dalam tiga kategori yaitu High Risk, Mid Risk, dan Low Risk. Proses penelitian meliputi tahap preprocessing data, transformasi label menggunakan Label Encoding, pembagian data dengan rasio 80% data latih dan 20% data uji, pelatihan model *CatBoost*, serta evaluasi menggunakan confusion matrix, accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa model *CatBoost* mampu menghasilkan akurasi sebesar 81,28% dalam mengklasifikasikan tingkat risiko kehamilan. Nilai precision, recall, dan F1-score pada masing-masing kelas menunjukkan performa yang cukup baik, dengan kelas High Risk memperoleh precision 0,85, recall 0,87, dan F1-score 0,86. Hasil analisis feature importance menunjukkan bahwa fitur BS (Blood Sugar) memiliki pengaruh terbesar terhadap prediksi risiko kehamilan dengan nilai sebesar 32,58%, diikuti oleh Age sebesar 16,96% dan HeartRate sebesar 15,18%. Berdasarkan hasil tersebut, algoritma *CatBoost* terbukti mampu memberikan performa yang baik dalam memprediksi risiko kehamilan dan berpotensi digunakan sebagai pendukung pengambilan keputusan dalam bidang kesehatan.

Kata Kunci: Risiko Kehamilan, Machine Learning, *CatBoost*, Klasifikasi, Prediksi.

Article History

Received: September 2026

Reviewed: September 2026

Published: September 2026

Plagiarism Checker

No 237.u873247.fh8374

Copyright : Author

Publish by : Kohesi



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/)

I. PENDAHULUAN

Kehamilan merupakan suatu proses biologis yang berlangsung di dalam rahim seorang perempuan selama kurang lebih 40 minggu, dihitung sejak hari pertama haid terakhir. Proses ini diawali dengan terjadinya fertilisasi, kemudian dilanjutkan dengan nidasi atau implantasi embrio pada dinding rahim, serta berlanjut dengan tahap pertumbuhan dan perkembangan hingga janin siap untuk dilahirkan. Kehamilan dapat menjadi masalah atau komplikasi setiap saat. WHO (World Health Organization) memperkirakan bahwa sekitar 15% dari seluruh wanita yang hamil akan berkembang menjadi komplikasi yang berkaitan dengan kehamilannya serta dapat mengancam jiwanya (Rizky Yulia Efendi et al., 2022).

Risiko kehamilan adalah keadaan menyimpang dari normal, yang secara langsung menyebabkan kesakitan dan kematian ibu maupun bayi. Ibu hamil yang berisiko adalah ibu



hamil yang mempunyai faktor risiko dan risiko tinggi. Oleh karena itu, tenaga kesehatan harus dapat mendeteksi sedini mungkin terhadap tanda-tanda bahaya pada ibu hamil yang mungkin akan terjadi. Ibu hamil yang termasuk dalam kelompok risiko tinggi antara lain memiliki ciri-ciri riwayat kehamilan buruk, riwayat anemia, gangguan tekanan darah, riwayat penyakit kronis, serta faktor 4T (Terlalu muda <20 tahun, Terlalu tua >35 tahun, Terlalu dekat jarak kehamilan, dan Terlalu banyak anak).

Variabel klinis merupakan kumpulan informasi medis yang diperoleh dari hasil pemeriksaan kesehatan pasien seperti usia (Age), tekanan darah sistolik (SystolicBP), tekanan darah diastolik (DiastolicBP), kadar gula darah (BS), suhu tubuh (BodyTemp), dan detak jantung (HeartRate). Parameter klinis tersebut sering digunakan oleh tenaga kesehatan sebagai dasar dalam menilai kondisi kesehatan ibu hamil. Perkembangan teknologi khususnya dalam bidang Machine Learning memberikan peluang baru dalam dunia medis untuk meningkatkan kualitas layanan kesehatan melalui sistem prediksi risiko otomatis.

Salah satu algoritma Machine Learning ensemble yang handal adalah Categorical Boosting (CatBoost). CatBoost merupakan algoritma gradient boosting berbasis pohon keputusan yang memiliki keunggulan dalam penanganan fitur kategorikal, efisiensi eksekusi, serta pencegahan overfitting melalui mekanisme Ordered Boosting. Oleh karena itu, penelitian ini bertujuan untuk mengimplementasikan dan menguji performa algoritma CatBoost dalam memprediksi tingkat risiko kehamilan berdasarkan variabel klinis ibu hamil.

II. LANDASAN TEORI DAN METODOLOGI

a. Algoritma boosting

Salah satu terobosan terbaru yang memanfaatkan teknologi adalah penerapan machine learning dibidang kesehatan. Diantara berbagai pendekatan machine learning, algoritma boosting merupakan salah satu metode yang baik untuk melakukan prediksi. Algoritma boosting merupakan metode yang menggabungkan beberapa model untuk meningkatkan performa dan akurasi.

b. Categorical Boosting

CatBoost adalah *Algoritma Boosting* open-source yang dikembangkan oleh tim Yandex, perusahaan teknologi terkemuka di Rusia. *CatBoost* merupakan singkatan dari "*Category Boosting*," yang menunjukkan keunggulan algoritma ini dalam menangani fitur kategorikal dalam data. *Algoritma CatBoost* dikembangkan untuk meningkatkan performa model Machine Learning dengan fokus pada kecepatan, akurasi, dan kemampuan penanganan fitur kategorikal. *CatBoost* atau *Categorical Boosting* adalah varian algoritma Gradient Boosting yang memanfaatkan struktur bilangan biner pada decision tree sebagai dasar dalam menghasilkan prediksi data (Jasman et al., 2022).

CatBoost membangun model secara bertahap dengan menambahkan pohon keputusan untuk memperbaiki error sebelumnya.

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

CatBoost meminimalkan fungsi loss menggunakan gradient descent. Untuk klasifikasi biner:

$$L(y, \hat{y}) = -\sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Gradient digunakan untuk mengetahui arah perbaikan error.



$$g_i = \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}$$

Keunggulan utama CatBoost adalah Ordered Boosting untuk mengurangi overfitting dan prediction shift.

Transformasi kategori:

$$\hat{x}_k = \frac{\sum_{j=1}^{p-1} -[x_j = x_p]y_i + aP}{\sum_{j=1}^{p-1} [x_j = x_p] + a}$$

Hasil akhir CatBoost adalah penjumlahan seluruh pohon keputusan.

$$\hat{y} = \sum_{n=1}^M \eta h_m(x)$$

c. Alur Penelitian

Penelitian ini menggunakan pendekatan kuantitatif. Tahapan penelitian meliputi :

1. Pengumpulan data :
Mengambil dataset ibu hamil berdasarkan variabel klinis dari platform kaggle sebanyak 1.014 baris data.
2. Preprocessing data :
Meliputi pemeriksaan missing value, normalisasi, dan label encoding pada variabel (high risk = 0, low risk = 1, dan mid risk = 2).
3. Pembagian data :
Dataset dibagi menjadi 80% data latih (811 sampel) dan 20% data uji (203 sampel).
4. Pelatihan model :
Melatih model catboost dengan konfigurasi hyperparameter optimal
5. Evaluasi performa :
Mengukur kinerja model berdasarkan *Confusion Matrix*, *Accuracy*, *Precision*, *Recall* Dan *F1-Score*.

III. HASIL DAN PEMBAHASAN

a. Pengujian Model Prediksi

Ada juga hasil pengujian prediksi model pada data uji sebagai berikut :

Tabel 1 Hasil Pengujian Prediksi

No	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	Hasil Prediksi
1	25	130	80	15	98	86	High Risk
2	35	140	90	13	98	70	High Risk
3	29	90	70	7.8	100	80	High Risk
4	30	140	85	7	98	70	High Risk
5	35	120	60	6.1	98	76	Mid Risk
6	23	130	70	7.0	98	78	Mid Risk



7	32	120	90	6.9	98	70	Mid Risk
8	22	110	75	6.0	98	72	Low Risk
9	25	120	80	6.2	98	74	Low Risk
10	19	120	70	4.8	98.4	70	Low Risk

Berdasarkan hasil tersebut, model mampu mengklasifikasikan setiap data ke dalam kategori risiko kehamilan (*High Risk*, *Mid Risk*, dan *Low Risk*) sesuai dengan pola yang telah dipelajari dari data training.

b. Pengujian hyperparameter catboost

Tabel 2 Pengujian Hyperparameter Catboost

Pengujian ke-	Max Iterations	Learning Rate	Akurasi Akhir
1	100	0.01	84.43 %
2	200	0.02	84.31%
3	300	0.03	84.39%
4	400	0.04	84.47%
5	500	0.01	87.02%
6	500	0.03	87.75%
7	600	0.03	87.60%
8	800	0.01	87.60%
9	1000	0.02	89.50%
10	1000	0.03	89.35%

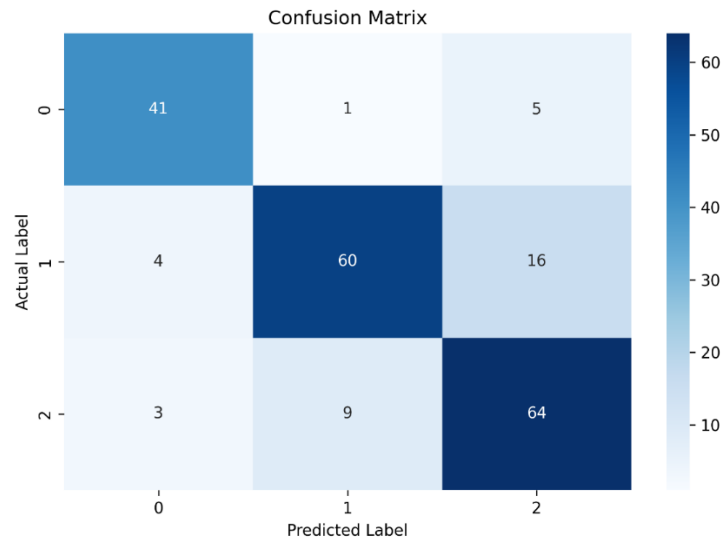
Berdasarkan tabel 4, konfigurasi optimal dicapai pada pengujian ke-9 dengan 1000 literasi dan learning rate 0,02 yang menghasilkan akurasi pelatihan mencapai 89,50%

c. Evaluasi model pada data uji (testing set)

Tabel 3 Evaluasi model

Kelas	Precision	Recall	F1-score	Support	Akurasi perkelas
High Risk	0.85	0.87	0.86	47	87.23%
Low Risk	0.86	0.75	0.80	80	75.00%
Mid Risk	0.75	0.84	0.80	76	84.21%
Accuracy			0,81	203	

d. Hasil Gambar Confusion Matrix



Gambar 1 Confusion Matrix

Berdasarkan hasil pengujian 203 data uji adalah sebagai berikut :

1. High risk (47 sampel) : 41 diprediksi benar, 1 diprediksi low risk, 5 diprediksi mid risk
2. Low risk (80 sampel) : 60 diprediksi benar, 4 diprediksi high risk, 16 prediksi mid risk
3. Mid risk (76 sampel) : 64 diprediksi benar, 3 diprediksi high risk, 9 diprediksi low

e. Fitur Yang Mempengaruhi

Adapun tabel fitur yang mempengaruhi prediksi resiko kehamilan berdasarkan presentase dengan menggunakan CatBoost.

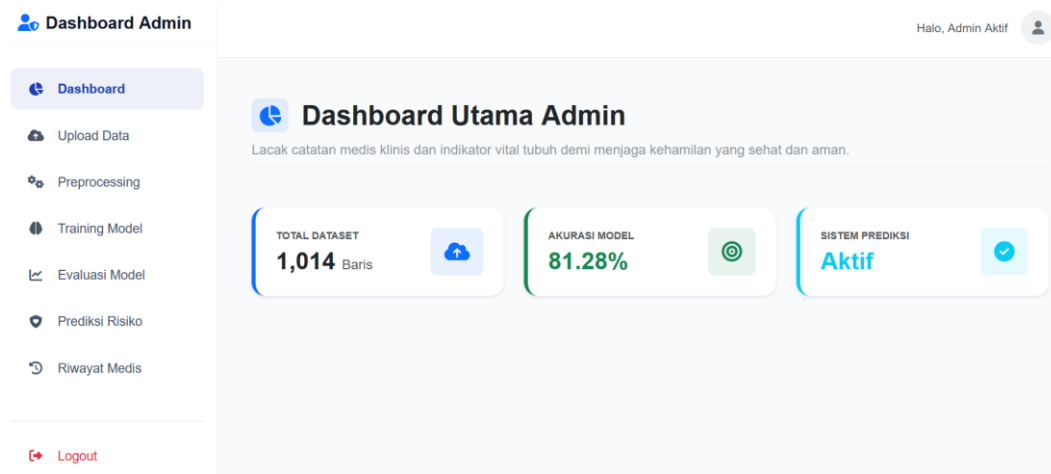
Tabel 4 Fitur yang Mempengaruhi

Fitur	Tingkat Pengaruh %
BS	32.58%
Age	16.96%
HeartRate	15.18%
SystolicBP	13.57%
DiastolicBP	12.37%
BodyTemp	9.31%

Proses klasifikasi risiko kehamilan oleh model CatBoost. Fitur BS (Blood Sugar) merupakan variabel yang paling berpengaruh dengan nilai 32.58%, yang menunjukkan bahwa kadar gula darah memiliki peran dominan dalam menentukan tingkat risiko kehamilan.

f. Implementasi berbasis web

1. Dashboard admin



Gambar 2 Dashboard Admin

Halaman Dashboard Utama Admin ini berfungsi sebagai pusat kendali sistem yang menyajikan ringkasan informasi penting secara real-time. Pada bagian kiri, terdapat menu navigasi lengkap mulai dari proses unggah data, preprocessing, pelatihan dan evaluasi model, hingga prediksi risiko dan riwayat medis serta tombol logout di bagian bawah. Sementara itu, area utama menampilkan tiga kartu indikator performa sistem, yaitu total data yang terkumpul sebanyak 1.014 baris, tingkat akurasi model sebesar 81.28%, dan status sistem prediksi yang sedang aktif berjalan.

IV. KESIMPULAN DAN SARAN

a. Kesimpulan

1. Algoritma catboost berhadil diimplementasikan untuk memprediksi tingkat risiko kehamilan berdasarkan 6 variabel klinis (age, systolicBP, DistolicBP, BS, BodyTemp, dan HeartRate) pada 1.014 dataset ibu hamil
2. Model catboost menghasilkan performa klasifikasi yang baik dengan akurasi 81.28% pada data uji. Kelas High Eisk memperoleh Precision 0.85, recall 0.87, dan F1-Score 0.86.
3. Variabel klinis yang paling berpengaruh terhadap prediksi adalah blood sugar sebesar (32.58%), diikuti oleh Usia (16.96%) dan detak jantung (15.18%).
4. Eksperimen hyperparameter diperoleh pada pengujian ke-9 dengan nilai Max Iterations sebesar 1000 dan Learning Rate sebesar 0,02. Kombinasi parameter tersebut menghasilkan akurasi tertinggi, yaitu 89,50% memberikan hasil yang paling optimal

b. Saran

1. Penelitian selanjutnya disarankan untuk menambah variabel klinis lain sepperti memperluas ruang lingkup penelitian dengan menambahkan faktor-faktor lain yang berkaitan dengan risiko kehamilan.
2. Mengembangkan dan membandingkan model dengan algoritma esamble lainnya seperti XGBoost, LightGBM, atau Random Forest.
3. Memperluas jangkauan uji coba klinis secara langsung pada fasilitas pelayanan seperti dipuskesmas/klinik.

**DAFTAR PUSTAKA**

- A.Siroj, R., Afgani, W., Fatimah, Septaria, D., Zahira, G., & Salsabila. (2024). METODE PENELITIAN KUANTITATIF PENDEKATAN ILMIAH UNTUK ANALISIS DATA. *Jurnal Review Pendidikan dan Pengajaran*, 11279-11289.
- Ali Hafid, M., & Ade Kurniawan, M. (2024). Deteksi Akun Kaggle Bot Menggunakan Linear Regression. *Journal homepage: Journal of Electrical Engineering and Computer (JEECOM)*, 449-459.
- Amaliah, S., Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, 121-127.
- Anita, N., & Diana, S. A. (2025). Deteksi Dini Faktor Resiko Kehamilan dan Peningkatan Pengetahuan Kader Tentang Screening Serta Pemeriksaan Kesehatan pada Ibu Hamil di Desa Mosso. *Kolaborasi: Jurnal Pengabdian Masyarakat*, 246-256.
- Arif Abdul Syukur, M., & Chamidy, T. (2025). Prediksi Kualitas Udara Menggunakan Metode CatBoost. *JISKA (Jurnal Informatika Sunan Kalijaga)*, 249-258.
- Dimas Syandika, N., & Yustanti, W. (2023). Deteksi Anomali Terhadap Pembatalan Transaksi Pada Platform Tiktok Shop dengan Algoritma Categorical Boosting (Catboost). *Journal of Informatics and Computer Science*, 1377-1385.
- Enggar Saputra, G., Hanindia Prami Swari, M., & Lina Nurlaili, A. (2025). Implementasi Algoritma XGBoost, CatBoost, dan LGBM untuk Klasifikasi Pencemaran Udara. *JIIP (Jurnal Ilmiah Ilmu Pendidikan)*, 14135-14138.
- Jasman, T. Z., Fadhlullah, M. A., Pratama, A. L., & Rismayani, R. (2022). Analisis Algoritma Gradient Boosting, AdaBoost dan Catboost dalam Klasifikasi Kualitas Air. *Jurnal Teknik Informatika dan Sistem Informasi*, 392-401.
- Kusuma, E. J., Nurmandhani, R., Aryani, L., Pantiawati, I., & Shidik, G. F. (2025). Optimasi Model Extreme Gradient Boosting Dalam Upaya Penentuan Tingkat Risiko Pada Ibu Hamil Berbasis Bayesian Optimization (BOXGB). *Jurnal Teknologi Informasi dan Ilmu Komputer*, 111-120.
- Mawardi, A. B., Pradini, R. S., & Haris, M. S. (2025). KOMPARASI ALGORITMA BOOSTING UNTUK PREDIKSI GANGGUAN TIDUR. *JITET (Jurnal Informatika dan Teknik Elektro Terapan)*, 1377-1385.
- Meihartati, T., Norhapifah, H., Abiyoga, A., Wiyata, S., & Samarinda, H. (2020). Tanda-Tanda Bahaya dan Resiko Pada Resiko Kehamilan. *STIKES Wiyata Husada Samarinda*.