

KOMPARASI ALGORITMA LIGHTGBM DAN CATBOOST UNTUK PREDIKSI DIABETES BERBASIS DATA KLINIS

LAPORAN SCIENTIST

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Informatika



Disusun oleh :

MUHAMMAD SIDIK LATUCONSINA

22.11.4676

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

KOMPARASI ALGORITMA LIGHTGBM DAN CATBOOST UNTUK PREDIKSI DIABETES BERBASIS DATA KLINIS

LAPORAN SCIENTIST

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Informatika



Disusun oleh :

MUHAMMAD SIDIK LATUCONSINA

22.11.4676

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

HALAMAN PERSETUJUAN

JALUR NON-REGULER

**KOMPARASI ALGORITMA LIGHTGBM DAN CATBOOST
UNTUK PREDIKSI DIABETES BERBASIS DATA KLINIS**

yang disusun dan diajukan oleh

Muhammad Sidik Latuconsina

22.11.4676

telah disetujui oleh Dosen Pembimbing

pada tanggal 27 Februari 2026

Dosen Pembimbing,



Majid Bahardi, S.Kom., M.Eng.

NIK. 190302393

HALAMAN PENGESAHAN

JALUR NON-REGULER

**KOMPARASI ALGORITMA LIGHTGBM DAN CATBOOST UNTUK
PREDIKSI DIABETES BERBASIS DATA KLINIS**

yang disusun dan diajukan oleh

Muhammad Sidik Latuconsina
22.11.4676

Telah dipertahankan di depan Dewan Penguji
pada tanggal 27 Februari 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Arifiyanto Hadinegoro, S.Kom., M.T.
NIK. 190302289

Dr. Hartatik, S.T., M.Cs.
NIK. 190302232

Majid Rahardi, S.Kom., M.Eng.
NIK. 190302393



Laporan ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 27 Februari 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusrini, M.Kom.

NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN KARYA

Yang bertandatangan di bawah ini,

Nama mahasiswa : Muhammad Sidik Latuconsina

NIM : 22.11.4676

Menyatakan bahwa Laporan dengan judul berikut:

Tuliskan Judul Karya

Dosen Pembimbing : Majid Rahardi, S.Kom., M.Eng.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan kegiatan SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidak-benaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 27 Februari 2026

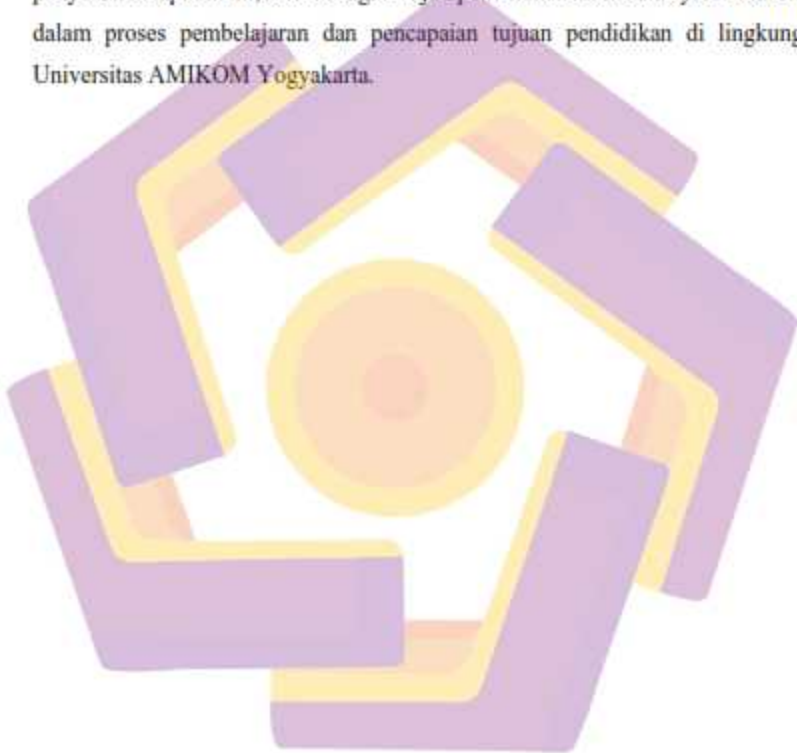
Yang Menyatakan,



Muhammad Sidik Latuconsina

HALAMAN PERSEMBAHAN

Laporan ini dipersembahkan kepada Universitas AMIKOM Yogyakarta sebagai bentuk kontribusi akademik dan tanggung jawab penulis dalam mendukung pengembangan ilmu pengetahuan, khususnya di bidang yang relevan dengan penyusunan laporan ini, serta sebagai wujud pemenuhan salah satu syarat akademik dalam proses pembelajaran dan pencapaian tujuan pendidikan di lingkungan Universitas AMIKOM Yogyakarta.



KATA PENGANTAR

Puji syukur kehadiran Allah SWT yang telah melimpahkan rahmat, hidayah, dan karunia-Nya, sehingga penulis dapat menyelesaikan penyusunan laporan naskah publikasi ini dengan baik. Laporan ini disusun sebagai salah satu syarat untuk mencapai gelar Sarjana Komputer pada Program Studi Informatika, Universitas Amikom Yogyakarta.

Penulis menyadari bahwa selesainya laporan penelitian ini tidak terlepas dari bantuan, semangat, dan doa dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati, penulis ingin menyampaikan rasa terima kasih dan penghargaan yang setinggi-tingginya kepada:

1. Teristimewa kepada kedua orang tua tercinta, Bapak Ichsan Latuconsina dan Ibu Erna Ery Tubaka, yang senantiasa memberikan kasih sayang tak terhingga, memanjatkan doa tulus di setiap sujudnya, serta memberikan dukungan moril maupun materil yang menjadi motivasi terbesar bagi penulis untuk menyelesaikan studi ini.
2. Bapak Majid Rahardi, S.Kom., M.Eng., selaku Dosen Pembimbing, yang telah dengan sabar meluangkan waktu, tenaga, dan pikiran untuk memberikan bimbingan, arahan, serta masukan yang sangat berharga sejak awal penelitian hingga terselesaikannya laporan ini.

Penulis menyadari bahwa laporan ini masih jauh dari kata sempurna karena keterbatasan pengetahuan dan pengalaman penulis. Oleh karena itu, kritik dan saran yang bersifat membangun sangat diharapkan demi penyempurnaan di masa mendatang. Semoga laporan penelitian ini dapat memberikan manfaat bagi pembaca serta berkontribusi terhadap pengembangan ilmu pengetahuan, khususnya di bidang *Machine Learning* dan informatika medis.

Yogyakarta, 06 Februari 2026

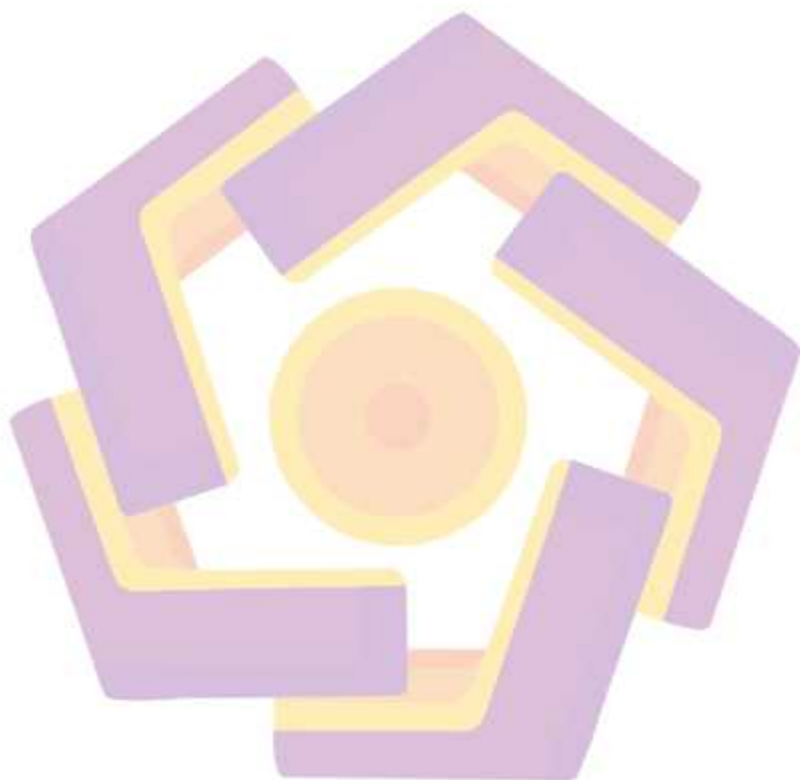
Penulis

DAFTAR ISI

Halaman Judul.....	i
Halaman Persetujuan.....	ii
Halaman Pengesahan	iii
Halaman Pernyataan Keaslian Karya.....	iv
Halaman Persembahan	v
Kata Pengantar	vi
Daftar Isi.....	vii
Daftar Tabel	x
Daftar Gambar	xi
Daftar Lampiran.....	xii
Daftar Lambang dan Singkatan.....	xiii
Daftar Istilah.....	xiv
Intisari	xv
<i>Abstract</i>	xvi
Bab I Pendahuluan	1
1.1. Gambaran Umum	1
1.2. Rumusan Masalah	3
1.3. Batasan Masalah	3
1.4. Tujuan.....	4
Bab II TINJAUAN PUSTAKA DAN DASAR TEORI	5
2.1. Studi Literatur.....	5
2.2. Landasan Teori	7
2.2.1 Diabetes Melitus.....	7
2.2.2 Data Mining dan <i>Machine Learning</i>	8
2.2.3 Pra-pemrosesan Data (<i>Data Preprocessing</i>)	9
2.2.4 Normalisasi Data	9
2.2. Imbalanced Data dan <i>SMOTE</i>	10
2.2.4 Algoritma Baseline.....	10
2.2.5 Algoritma <i>Gradient boosting</i>	11
2.2.6 Explainable AI (XAI) dan SHAP	11
2.2.9 Evaluasi dan Validasi Model	12

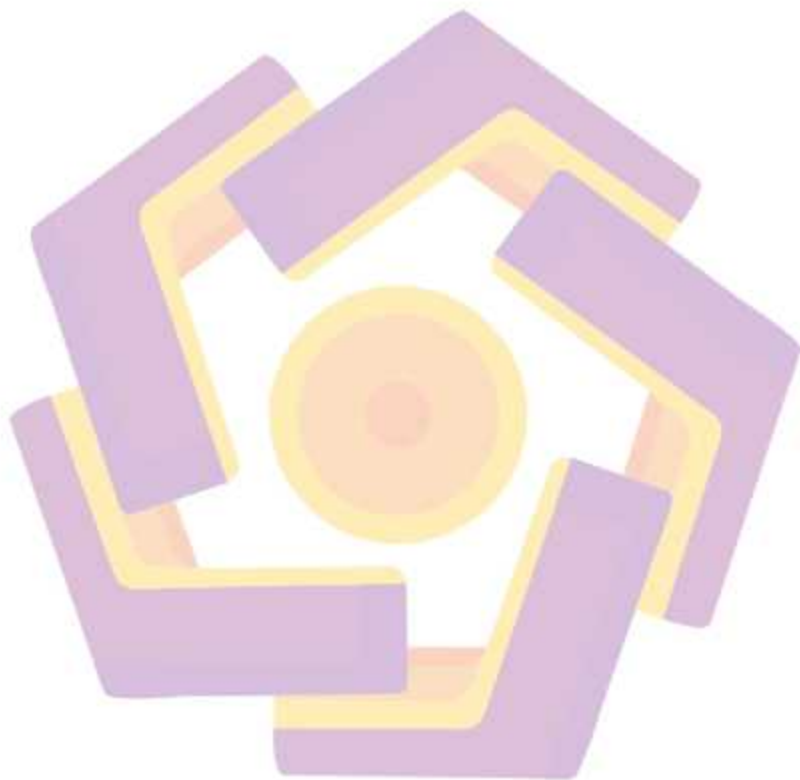
BAB III Metode Penelitian	13
3.1. Alur Penelitian	13
3.2. Data Penelitian	13
3.2.1. Analisis Eksplorasi Data (EDA)	14
3.3. Pra-pemrosesan Data.....	16
3.3.1. Pembersihan Data.....	16
3.3.2. Feature Encoding	16
3.3.3. Feature Scaling.....	16
3.4. Implementasi <i>SMOTE</i>	17
3.6. Evaluasi Model.....	21
3.7. Validasi Model (<i>K-Fold Cross-Validation</i>).....	22
BAB IV PEMBAHASAN.....	23
4.1. Gambaran Umum Hasil Eksperimen	23
4.2. Evaluasi Performa Model.....	23
4.2.1. Perbandingan Metrik Statistik.....	23
4.2.2. Analisis Confusion Matrix	25
4.2.3. Analisis Kurva ROC dan AUC	27
4.3. Pembahasan.....	28
4.3.1. Analisis Komparatif Algoritma: <i>LightGBM</i> vs <i>CatBoost</i>	28
4.3.2. Urgensi Sensitivitas.....	29
4.4. Interpretasi Model (<i>Explainable AI</i>).....	29
4.4.1. Analisis Fitur Dominan.....	29
4.4.2. Analisis Arah Pengaruh Fitur.....	31
BAB V Kesimpulan	32
5.1. Kesimpulan.....	32
5.2. Saran.....	33
Referensi	34
Curriculum Vitae.....	37
Lampiran dan Bukti Pendukung.....	39
Muhammad Sidik Latuconsina ^{1*} , Majid Rahardi ^{2*}	46
I. Pendahuluan	46
II. Metode.....	48

III.	Result And Discussion.....	51
IV.	Kesimpulan.....	54



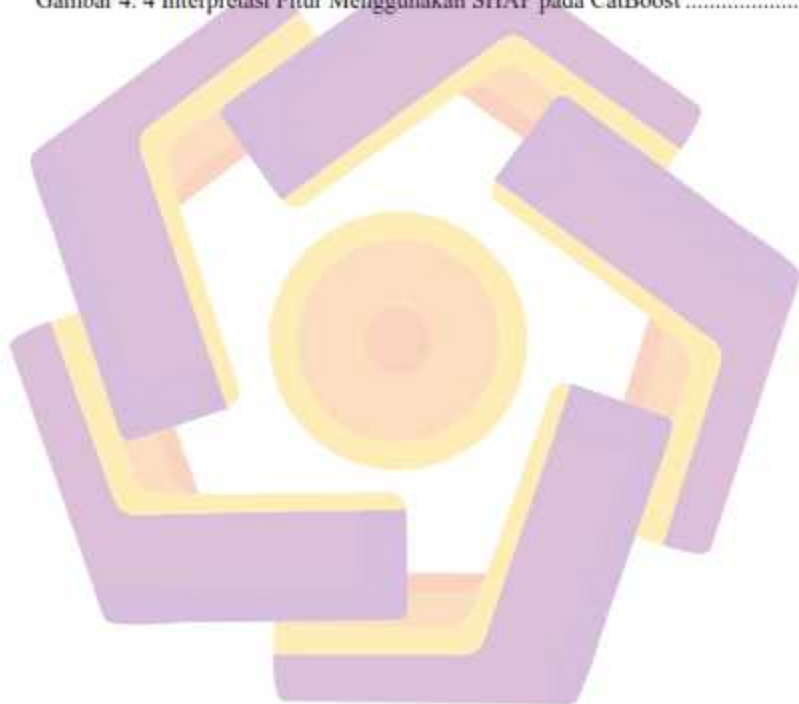
DAFTAR TABEL

Tabel 3. 1 Karakteristik Dataset.....	14
Tabel 3. 2 Ringkasan Pra-pemrosesan	16
Tabel 4. 1 Perbandingan Performa Algoritma Klasifikasi	24



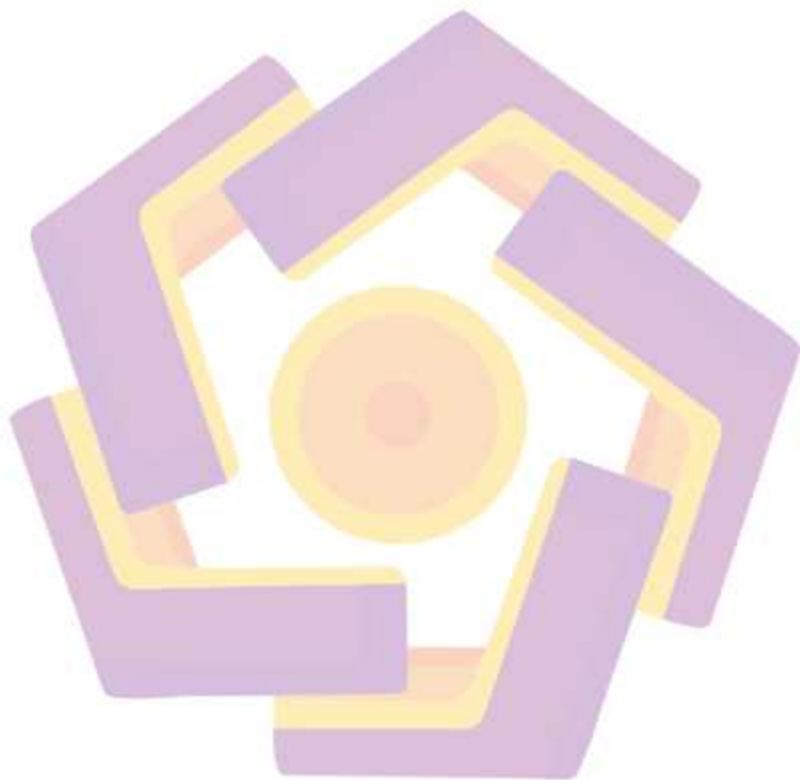
DAFTAR GAMBAR

Gambar 3. 1 Alur Penelitian.....	13
Gambar 3. 2 Grafik Bar Distribusi Kelas Target	15
Gambar 3. 3 Heatmap Correlation Matrix	15
Gambar 3. 4 Bar Chart Sebelum SMOTE.....	18
Gambar 3. 5 Bar Chart Sesudah SMOTE	18
Gambar 4. 1 Grafik Perbandingan Akurasi Antar Model	25
Gambar 4. 2 Komporasi Confusion Matrix.....	26
Gambar 4. 3 Kurva ROC dan Nilai AUC	27
Gambar 4. 4 Interpretasi Fitur Menggunakan SHAP pada CatBoost	30

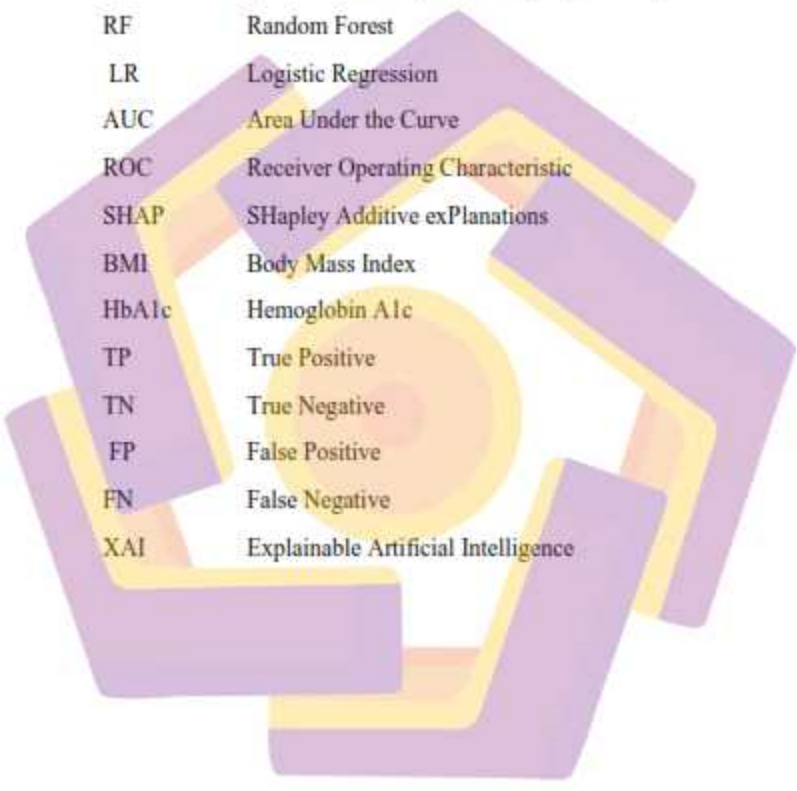


DAFTAR LAMPIRAN

Lampiran 1 Letter of Acceptance (LOA)	39
Lampiran 2 Lembar Review	39
Lampiran 3 Bukti Terbit/Terindex	45
Lampiran 4 Bukti Pembayaran.....	45
Lampiran 5 Naskah Jurnal	46

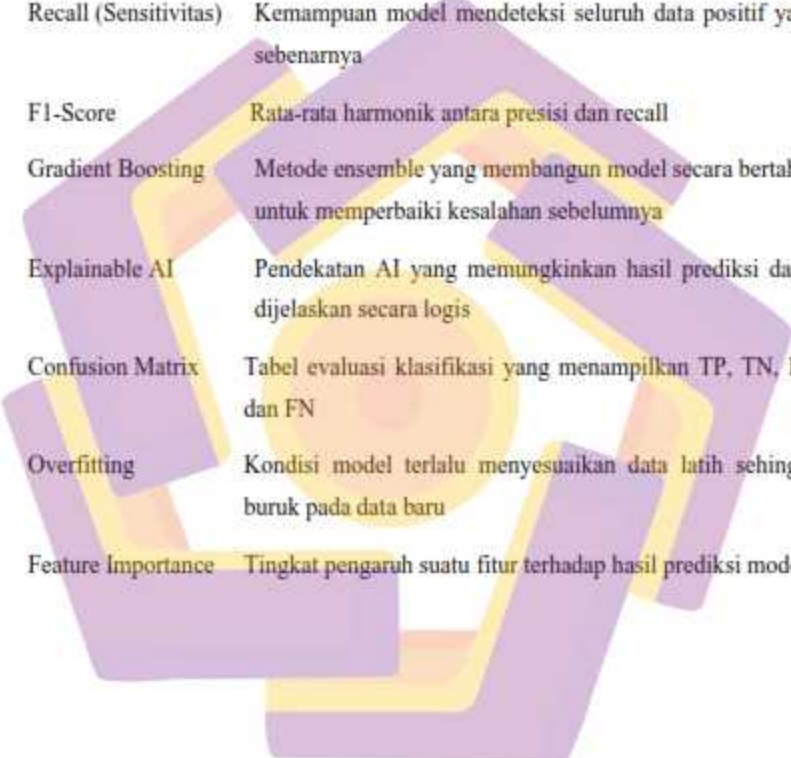


DAFTAR LAMBANG DAN SINGKATAN



DM	Diabetes Melitus
ML	Machine Learning
SMOTE	Synthetic Minority Over-sampling Technique
RF	Random Forest
LR	Logistic Regression
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
BMI	Body Mass Index
HbA1c	Hemoglobin A1c
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
XAI	Explainable Artificial Intelligence

DAFTAR ISTILAH



Imbalanced Data	Kondisi distribusi kelas yang tidak seimbang antara kelas mayoritas dan minoritas
Recall (Sensitivitas)	Kemampuan model mendeteksi seluruh data positif yang sebenarnya
F1-Score	Rata-rata harmonik antara presisi dan recall
Gradient Boosting	Metode ensemble yang membangun model secara bertahap untuk memperbaiki kesalahan sebelumnya
Explainable AI	Pendekatan AI yang memungkinkan hasil prediksi dapat dijelaskan secara logis
Confusion Matrix	Tabel evaluasi klasifikasi yang menampilkan TP, TN, FP, dan FN
Overfitting	Kondisi model terlalu menyesuaikan data latih sehingga buruk pada data baru
Feature Importance	Tingkat pengaruh suatu fitur terhadap hasil prediksi model

INTISARI

“Diabetes Melitus” merupakan tantangan kesehatan global yang memerlukan deteksi dini yang akurat guna mencegah komplikasi fatal. Namun, data klinis seringkali menunjukkan distribusi kelas yang tidak seimbang, sehingga menghambat model prediksi standar dalam mendeteksi pasien positif secara efektif. Penelitian ini bertujuan untuk membandingkan performa dua algoritma *Gradient boosting* modern, yaitu “*LightGBM*” dan “*CatBoost*”, dalam memprediksi risiko diabetes. Algoritma *Random Forest* dan *Logistic Regression* disertakan sebagai model *baseline* untuk mengukur efektivitas. Untuk menangani ketidakseimbangan data, teknik *Synthetic Minority Over-sampling Technique (SMOTE)* diterapkan pada tahap pra-pemrosesan data latih. *Dataset* bersumber dari repositori publik *Kaggle (Diabetes Prediction Dataset)*, yang terdiri dari 100.000 rekam medis pasien dengan atribut klinis seperti usia, indeks massa tubuh (IMT), dan kadar HbA1c. Evaluasi performa dilakukan menggunakan metrik Akurasi, Presisi, *Recall*, *F1-Score*, dan *Area Under the Curve (AUC)*. Hasil eksperimen menunjukkan kompetisi yang ketat, di mana “*LightGBM*” mencapai akurasi tertinggi sebesar 97,16%. Namun, “*CatBoost*” menunjukkan sensitivitas (*Recall*) yang unggul sebesar 69,71% dan *F1-Score* tertinggi sebesar 80,48%. Hal ini menjadikan “*CatBoost*” sebagai model yang paling andal dalam meminimalkan *False Negative* dibandingkan dengan “*LightGBM*” dan *Random Forest*, sementara *Logistic Regression* menunjukkan performa terendah. Selain itu, analisis interpretabilitas menggunakan “*SHAP*” (*SHapley Additive exPlanations*) mengungkapkan bahwa kadar HbA1c dan glukosa darah merupakan fitur yang paling dominan dalam deteksi, yang memvalidasi keselarasan model dengan diagnosis klinis. Penelitian ini menyimpulkan bahwa algoritma “*CatBoost*” yang dikombinasikan dengan “*SMOTE*” menawarkan prediksi diabetes yang lebih sensitif, transparan, dan efisien untuk *skrining* medis.

Kata kunci: Diabetes Melitus, *SHAP*, *SMOTE*, *LightGBM*, *CatBoost*.

ABSTRACT

"Diabetes Mellitus" is a global health challenge that requires accurate early detection to prevent fatal complications. However, clinical data often show an unbalanced distribution of classes, thus hampering standard prediction models in effectively detecting positive patients. This study aims to compare the performance of two modern Gradient boosting algorithms, namely "LightGBM" and "CatBoost", in predicting diabetes risk. The Random Forest and Logistic Regression algorithms are included as a baseline model to measure effectiveness. To deal with data imbalances, the Synthetic Minority Over-sampling Technique (SMOTE) technique is applied at the pre-processing stage of the training data. The Dataset is sourced from the public repository Kaggle (Diabetes Prediction Dataset), which consists of 100,000 medical records of patients with clinical attributes such as age, body mass index (BMI), and HbA1c levels. Performance evaluation was conducted using the Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC) metrics. The results of the experiment showed a fierce competition, where "LightGBM" achieved the highest accuracy of 97.16%. However, "CatBoost" showed superior sensitivity (Recall) of 69.71% and the highest F1-Score of 80.48%. This makes "CatBoost" the most reliable model for minimizing False Negatives compared to "LightGBM" and Random Forest, while Logistic Regression shows the lowest performance. In addition, interpretability analysis using "SHAP" (SHapley Additive exPlanations) revealed that HbA1c and blood glucose levels were the most dominant features in detection, which validated the model's alignment with clinical diagnosis. The study concluded that the "CatBoost" algorithm combined with "SMOTE" offers more sensitive, transparent, and efficient diabetes prediction for medical screening.

Keyword: Diabetes Mellitus, SHAP, SMOTE, LightGBM, CatBoost.