JURNAL FOURIER | Oktober 2026, Vol. 15, No. 2, 92-102 ISSN 2252-763X
DOI: 10.14421/fourier.2026.152.92-102 E-ISSN 2541-5239

Implementasi Light Gradient Boosting Machine Untuk Klasifikasi Penyakit Diabetes Melitus Dengan Deteksi Outlier Robust Random Cut Forest

Siti Nurus Shofiyah[1], Putroue Keumala Intan[2]

[12]Program Studi Matematika, Fakultas Sains dan Teknologi, UIN Sunan Ampel Surabaya

Jalan Dr. Ir. H. Soekarno No. 682, Gunung Anyar, Kecamatan Gunung Anyar, Surabaya, Jawa Timur 60294, Indonesia

Korespondensi; Putroue Keumala Intan Email: putroue@uinsa.ac.id

Abstrak

Diabetes melitus merupakan penyakit tidak menular yang ditandai dengan tingginya kadar glukosa dalam darah dan menjadi salah satu masalah kesehatan utama di Indonesia. Tingginya jumlah penderita menuntut adanya metode deteksi dini yang akurat untuk mendukung pencegahan komplikasi serta membantu pengambilan keputusan di bidang kesehatan. Penelitian ini bertujuan mengevaluasi efektivitas model *Light Gradient Boosting Machine* dalam mengklasifikasikan penyakit diabetes melitus. Sebelum proses klasifikasi, dilakukan deteksi dan penanganan *outlier* menggunakan *Robust Random Cut Forest* dengan persentase penanganan sebesar 5%, serta penyeimbangan distribusi kelas menggunakan *Synthetic Minority Over-sampling Technique*. Evaluasi model dilakukan menggunakan *K*-Fold Cross Validation sebanyak lima lipatan. Selain itu, dilakukan pengujian beberapa kombinasi parameter, yaitu *learning rate* sebesar 0.05, 0.075, dan 0.1, *num leaves* sebanyak 2, 4, 6, 8, 10, 12, 14, dan 16, serta *min data in leaf* sebesar 10, 20, 40, dan 80 untuk memperoleh model terbaik. Hasil penelitian menunjukkan bahwa kombinasi metode Robust Random Cut Forest (RRCF), SMOTE, dan Light Gradient Boosting Machine (LightGBM) menghasilkan nilai akurasi sebesar 85,36%, sensitivitas sebesar 79,03%, dan spesifisitas sebesar 85,77%. Hasil tersebut menunjukkan bahwa kombinasi metode yang digunakan mampu melakukan klasifikasi diabetes melitus dengan kinerja yang cukup baik berdasarkan ketiga metrik evaluasi tersebut. Dengan demikian, penelitian ini menunjukkan penerapan RRCF untuk penanganan outlier, SMOTE untuk penyeimbangan kelas, dan LightGBM untuk klasifikasi diabetes melitus.

Kata Kunci: Diabetes Melitus, Klasifikasi, LightGBM, RRCF

Abstract

Abstract—Diabetes mellitus is a non-communicable disease characterized by elevated blood glucose levels and remains one of the major public health challenges in Indonesia. The increasing number of diabetes cases highlights the need for accurate early detection methods to prevent complications and support clinical decision-making. This study aims to evaluate the effectiveness of the Light Gradient Boosting Machine (LightGBM) model for diabetes mellitus classification. Prior to classification, outlier detection and treatment were performed using the Robust Random Cut Forest (RRCF) algorithm with an outlier handling proportion of 5%, while class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE). Model performance was evaluated using five-fold cross-validation. In addition, several combinations of LightGBM hyperparameters were examined, including learning rates of 0.05, 0.075, and 0.10, *num leaves* values of 2, 4, 6, 8, 10, 12, 14, and 16, and *min data in leaf* values of 10, 20, 40, and 80 to obtain the optimal model. The research results indicate that the combination of the Robust Random Cut Forest (RRCF) method, SMOTE, and Light Gradient Boosting Machine (LightGBM) yields an accuracy of 85.36%, a sensitivity of 79.03%, and a specificity of 85.77%. These results demonstrate that the combined methodology is capable of classifying diabetes mellitus with a commendable level of performance across these three evaluation metrics. Consequently, this study highlights the application of RRCF for handling outliers, SMOTE for class balancing, and LightGBM for the classification of diabetes mellitus.

Keywords Diabetes Mellitus, Classification, LightGBM, RRCF

© 2026 JURNAL FOURIER

Pendahuluan

Diabetes melitus merupakan salah satu penyakit tidak menular yang menjadi masalah kesehatan masyarakat di dunia. Penyakit ini ditandai dengan meningkatnya kadar glukosa dalam darah akibat gangguan produksi atau kerja insulin sehingga dapat menimbulkan berbagai komplikasi serius, seperti penyakit jantung, gagal ginjal, kebutaan, hingga amputasi apabila tidak ditangani secara tepat [1]. Menurut International Diabetes Federation (IDF), jumlah penderita diabetes terus meningkat setiap tahun. Pada tahun 2024 Indonesia menempati peringkat kelima dunia dengan sekitar 20,4 juta penderita, sehingga diperlukan upaya deteksi dini yang akurat untuk menekan angka komplikasi dan kematian akibat diabetes melitus [2].

Deteksi dini diabetes melitus umumnya dilakukan melalui pemeriksaan kadar glukosa darah menggunakan glukometer maupun pemeriksaan hemoglobin terglikasi (HbA1c) [3]. Meskipun kedua metode tersebut banyak digunakan, pemeriksaan glukometer memiliki keterbatasan akurasi yang dipengaruhi oleh teknik pengambilan sampel, kualitas alat, dan kondisi lingkungan, sedangkan pemeriksaan HbA1c memerlukan fasilitas laboratorium, waktu pemeriksaan yang lebih lama, serta tidak dapat diterapkan pada beberapa kondisi klinis tertentu [4]. Oleh karena itu, diperlukan metode pendukung yang mampu membantu proses identifikasi risiko diabetes secara cepat dan akurat berdasarkan data kesehatan pasien

Perkembangan *machine learning* telah memberikan kontribusi yang besar dalam bidang kesehatan, khususnya pada proses klasifikasi penyakit [5]. Berbagai algoritma telah dikembangkan untuk mempelajari pola dari data pasien sehingga mampu membantu tenaga medis dalam melakukan deteksi dini. Salah satu algoritma yang banyak digunakan adalah Light Gradient Boosting Machine (LightGBM) karena memiliki proses pelatihan yang cepat, mampu menangani data berdimensi besar, serta menghasilkan performa klasifikasi yang tinggi melalui pendekatan *gradient boosting decision tree* [6]. Keunggulan tersebut menjadikan LightGBM banyak dimanfaatkan dalam berbagai penelitian klasifikasi penyakit yang memerlukan tingkat akurasi tinggi dan efisiensi komputasi.

Beberapa penelitian sebelumnya menunjukkan bahwa metode LightGBM memiliki kinerja yang sangat baik dalam proses klasifikasi penyakit. Tawil dkk. membandingkan beberapa algoritma klasifikasi pada data kanker payudara dan menunjukkan bahwa LightGBM memperoleh akurasi tertinggi sebesar 98%, melampaui Support Vector Machine, Random Forest, Logistic Regression, K-Nearest Neighbor, dan Naïve Bayes [7]. Penelitian lain oleh Xiong dkk. mengombinasikan LightGBM dengan metode Isolation Forest untuk mendeteksi *outlier* pada klasifikasi *Sleep Apnea Syndrome* [8]. Hasil penelitian tersebut menunjukkan bahwa penanganan *outlier* mampu meningkatkan akurasi model dari 87,82% menjadi 92,62%, sehingga membuktikan bahwa kualitas data berpengaruh terhadap performa klasifikasi.

Penelitian terdahulu telah mengombinasikan *Light Gradient Boosting Machine* (LightGBM) dengan metode deteksi *outlier*. Metode Isolation Forest masih menjadi teknik yang paling banyak digunakan dalam proses penanganan *outlier*. Padahal, Robust Random Cut Forest (RRCF) merupakan metode yang dirancang khusus untuk mendeteksi *outlier* pada data multidimensi dan dilaporkan memiliki kemampuan yang lebih baik dalam mengidentifikasi anomali dibandingkan beberapa metode konvensional [9]. Selain itu, penelitian mengenai klasifikasi diabetes melitus yang mengombinasikan LightGBM dengan RRCF masih sangat terbatas, terutama pada data kesehatan yang memiliki ketidakseimbangan kelas.

Penelitian ini menerapkan *Robust Random Cut Forest* (RRCF) sebagai metode deteksi *outlier* sebelum proses pelatihan model *Light Gradient Boosting Machine* (LightGBM) untuk meningkatkan kualitas data yang digunakan dalam klasifikasi. Penanganan ketidakseimbangan kelas dilakukan menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) sehingga distribusi kelas menjadi

lebih seimbang. Kinerja model selanjutnya dievaluasi menggunakan *5-Fold Cross Validation* dengan pengujian berbagai kombinasi parameter LightGBM untuk memperoleh model terbaik..

Penelitian ini bertujuan mengevaluasi kinerja model LightGBM yang dioptimalkan menggunakan Robust Random Cut Forest dan SMOTE dalam mengklasifikasikan penyakit diabetes melitus. Kontribusi utama penelitian ini adalah mengintegrasikan deteksi *outlier*, penyeimbangan data, dan optimasi parameter dalam satu kerangka klasifikasi sehingga diharapkan mampu meningkatkan kemampuan model dalam mengenali pasien diabetes secara lebih akurat. Hasil penelitian ini diharapkan dapat menjadi alternatif metode pendukung deteksi dini diabetes melitus dan memberikan kontribusi bagi pengembangan sistem pendukung keputusan di bidang kesehatan.

Landasan Teori

Diabetes Melitus

Diabetes melitus adalah sekelompok penyakit metabolik yang ditandai dengan peningkatan kadar gula darah (hiperglikemia) yang disebabkan oleh kelainan sekresi dan/atau aktivitas insulin [10]Insulin merupakan hormon yang diproduksi oleh sel-sel beta Langerhans kelenjar pankreas yang memiliki fungsi sebagai pengatur kadar glukosa dalam darah [11]. Kementrian Kesehatan mendefinisikan penyakit diabetes melitus sebagai suatu penyakit atau gangguan metabolisme kronis dengan multi etiologi yang ditandai dengan tingginya kadar gula darah disertai dengan gangguan metabolisme karbohidrat, lipid, dan protein sebagai akibat insufisiensi fungsi insulin [12]. Insufisiensi fungsi insulin merupakan sel- sel tubuh terhadap insulin kurang responsif dan produksi insulin oleh sel- sel beta Langerhans kelenjar pankreas mengalami defisiensi yang menghambat penyerapan glukosa oleh sel tubuh sehingga merubah sel- sel beta Langerhans menjadi sel- sel progenitor endokrin yang tidak berfungsi [13].

Robust Random Cut Forest

Robust random cut forest (RRCF) ialah metode ensemble untuk mendeteksi outlier dalam data. Robust Random Cut Forest dikembangkan untuk mengurangi kompleksitas memori dan meningkatkan efisiensi pada dataset berdimensi tinggi atau pada model OOD (out of distribution)[14]. Secara spesifik Robust Random Cut Forest memberikan kinerja yang baik pada data berdimensi tinggi dan menangani duplikat dengan baik yang dapat menutupi keberadaan outlier. Robust Random Cut Forest menunjukkan kinerja yang lebih baik daripada Isolation Forest, karena pemotongan dimensi ditentukan berdasarkan rentang geometrik data, sedangkan algoritma Isolation Forest memilih pemotongan dimensi secara acak[15].

Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE merupakan penanganan data tidak seimbang yang mana data minoritas akan ditambahkan data baru (data sintesis) dengan melakukan interpolasi pada jarak[16]. Algoritma ini mencari tetangga terdekat (K-Nearest Neighbors) dan menyisipkan titik data baru di antara data yang telah ada. Akan ditunjukkan persamaan SMOTE untuk mencari data baru (data sintesis) pada persamaan berikut

$$x_{\text{baru}} = x_0 + \delta(x_{nn} - x_0) \quad (1)$$

Dimana,

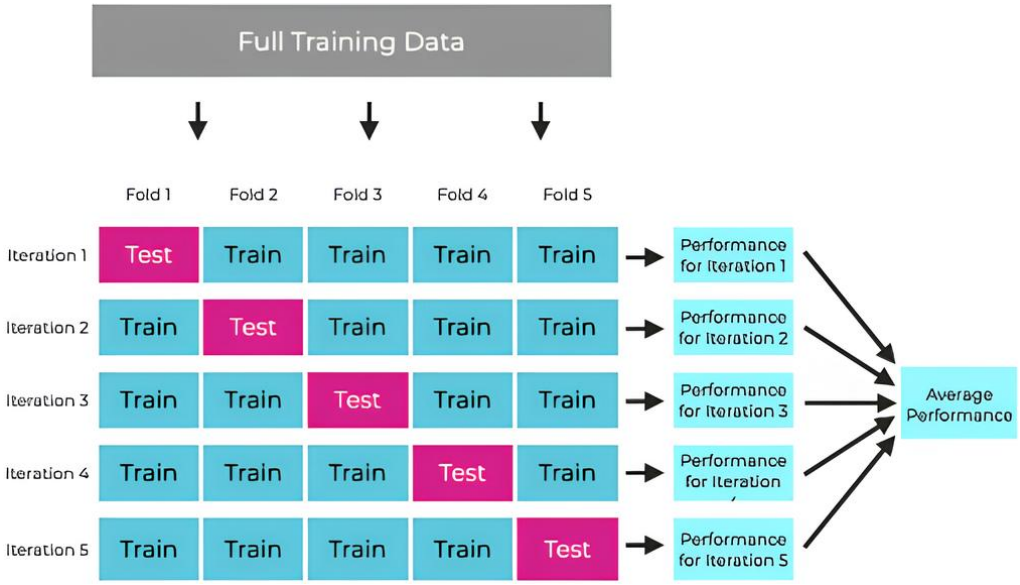
x_0 : Data minoritas yang dipilih,

δ : Bilangan acak antara 0 dan 1

x_{nn} : Tetangga terdekat dari x_0

K-Fold Cross Validation

K-Fold cross validation adalah teknik evaluasi yang paling banyak digunakan untuk memperkirakan kinerja generalisasi model dengan membagi data menjadi k subset (*fold*) dan melakukan pelatihan atau validasi berulang [17]. Pembagian data menggunakan K-Fold Cross Validation dapat dilihat pada Gambar 1 [18].

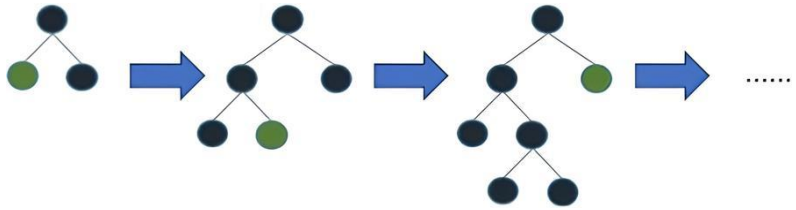


Gambar 1. *K-Fold Cross Validation*

Gambar 1. menunjukkan pembagian data menggunakan K-Fold Cross Validation dengan bagian pertama berwarna ungu digunakan sebagai data uji, sedangkan bagian data lainnya digunakan untuk data latih yang ditunjukan pada warna biru. Langkah tersebut akan diulang hingga fold terakhir yang telah ditentukan.

Light Gradient Boosting Machine

LightGBM merupakan salah satu algoritma *boosting* yang populer dan juga merupakan pesaing algoritma *boosting* lain seperti XGBoost dan algoritma boosting lainnya. LightGBM adalah implementasi modern dari gradient boosting *decision trees* yang dirancang untuk efisiensi memori untuk membuat prediksi yang lebih cepat. LightGBM memiliki beberapa kelebihan diantaranya efisisensi yang lebih tinggi, penggunaan memori yang lebih rendah, akurasi yang lebih baik, dan dapat menangani data yang besar. Ada beberapa parameter dalam LightGBM *seperti learning rate, max depth, n estimators, num leaves, min child samples*. LightGBM membagi pohon secara *leaf-wise* (berurutan daun) seperti yang ditunjukkan pada Gambar 2[19].



Gambar 2. Pohon *Life-Wise* LightGBM

Gambar 2 menunjukkan LightGBM menggunakan fungsi *life-wise tree growth* atau pertumbuhan pohon berdassarkan daunnya. Pohon dapat menjadi tidak seimbang dan tumbuh hanya dalam satu arah yang terlihat pada pohon ketiga pada gambar. Pohon seperti itu rentan terhadap *overfitting*. Akan tetapi dapat diatasi parameter *max depth* untuk membatasi kedalaman pohon atau mengontrol kedalaman maksimum pohon. Tujuan dari algoritma LightGBM adalah untuk meminimalkan fungsi kerugian, yang biasanya didefinisikan sebagai jumlah kerugian untuk setiap contoh dalam dataset. Fungsi kerugian dapat ditulis seperti pada persamaan berikut[20]:

$$L(y_i, \hat{p}_i) = -[y_i \ln(\hat{p}_i) + (1 - y_i) \ln(1 - \hat{p}_i)] \quad (2)$$

Dimana,

y_i : Label aktual dari observasi ke-i.

$\hat{p}_i$: Probabilitis hasil prediksi baru

Confussion Matrix

Confusion matrix adalah metode dasar untuk mengevaluasi kinerja klasifikasi dengan menyajikan jumlah TP (True Positive) data aktual positif diprediksi sebagai positif, TN (True Negative) yaitu data aktual negatif diprediksi negatif, FP (False Positive) yaitu data aktual negatif diprediksi positif, dan FN (False Negative) yaitu data aktual positif diprediksi negative [21]. Perbandingan antara label aktual dan klasifikasi yang diprediksi oleh model dapat dilihat dari Tabel 1 berikut:

Tabel 1. *Confussion Matrix*

		Prediksi	
		Sehat	DM
Aktual	Sehat	TN	FP
	DM	FN	TP

Perhitungan hasil evaluasi pada penelitian ini dapat dihitung menggunakan persamaan berikut:

Sensitivitas atau recall mengukur seberapa baik model dalam menemukan semua data yang benar benar positif. Nilai recall dapat dihitung menggunakan persaman berikut:

$$\text{Sensitivitas} = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

Spesifisitas digunakan untuk mengetahui seberapa efektif pengklasifikasian data. Nilai spesifitas dapat dihitung menggunakan persamaan berikut:

$$\text{Spesifitas} = \frac{TN}{FP+TN} \times 100\% \quad (4)$$

Akurasi adalah standar dasar untuk menentukan seberapa baik suatu model dapat menghasilkan prediksi dengan membandingkan jumalah data yang memiliki kasifikasi benar dan jumlah seluruh data. Nilai akurasi dapat dihitung menggunakan persamaan berikut:

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (5)$$

Metode

Sumber Data

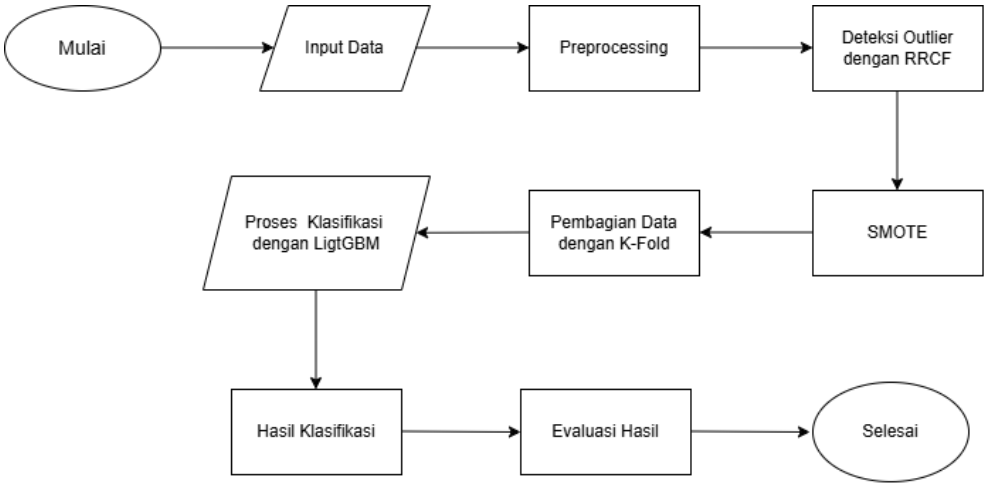
Sumber data pada penelitian ini diambil dari dari situs *Mendeley Data*. Data ini merupakan data kuantitatif yang menggunakan data sekunder dari data asli yang mana pengambilan data dilakukan oleh petugas Kesehatan Masyarakat dari Yayasan Palli Karma-Sahayak Negara Bangladesh. Data pada penelitian ini berjumlah 5.288 data dengan 342 data penderita Diabetes Melitus dan 4.946 data sehat. Tabel 1 merupakan variabel indikator diabetes melitus:

Tabel 2. Indikator Diabetes Melitus

Variabel	Indikator
X_1	Usia
X_2	Jenis Kelamin
X_3	Frekuensi Denyut Jantung
X_4	Tekanan Darah Sistolik
X_5	Tekanan Darah Diastolik
X_6	Glukosa
X_7	Tinggi Badan
X_8	Berat Badan
X_9	Indeks Massa Tubuh
X_{10}	Diabetes Keluarga
X_{11}	Hipertensi
X_{12}	Hipertensi Keluarga
X_{13}	Penyakit Kardiovaskular
X_{14}	Stroke
Y	Diabetes

Langkah – langkah

Diagram alir yang digunakan dalam penelitian ini ditunjukkan pada Gambar 3.



Gambar 3. Diagram Alir Penelitian

Berdasarkan diagram alir pada Gambar 3, alur penelitian dapat dijabarkan sebagai berikut:

a. Langkah pertama yaitu menginput data yang mana Y merupakan variabel dependen dengan dua kelas dan X merupakan variabel independen dengan faktor -faktor yang memengaruhi diabetes melitus.
b. Langkah ke dua pada penelitian ini yaitu melakakun *preprocessing* data agar data siap pakai dengan menghapus data kosong dan *encoding* data dari bentuk string ke dalam bentuk integer.
c. Langkah ketiga dilakukan deteksi *outlier* menggunakan metode Robust Random Cut Forest untuk pengendali data-data yang memiliki nilai yang jauh dari sebagian besar dataset.
d. Langkah keempat yaitu dilakukan penanganan ketidakseimbangan data menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE) dengan membuat data sintesis baru pada kelas minoritas.
e. Langkah kelima dilakukan pembagian data dengan mengaplikasikan teknik *K-Fold Cross Validation* yang diimplementasikan dalam pembagian data pada dataset Diabetes Melitus menggunakan nilai k = 5.
f. Langkah keenam yaitu proses klasifikasi menggunkan metode LightGBM untuk memeroleh probabilitas baru.
g. Langkah ketujuh dilakukan evaluasi model menggunakan Confussion Matrix.

Hasil dan Pembahasan

Langkah pertama pada penelitian ini yaitu *preprocessing* meliputi mengubah data string menjadi data numerik menggunakan *encoding*. Hal ini bertujuan untuk menjaga konsistensi representasi data, sehingga setiap kategori memiliki nilai numerik yang tetap. Tahap selanjutnya, dilakukan penanganan data *outlier* menggunkan metode *Robust Random Cut Forest* yang mana pada penelitian ini akan dihapus data *outlier* yang berpotensi mengganggu pembentukan pola klasifikasi. Pada penelitian ini digunakan penanganan outlier sebesar 5% yang mana 5% dari data yang memiliki skor tertinggi dianggap sebagai outlier yang dapat dilihat pada Tabel

Tabel 3. Hasil Data Setelah Dilakukan Penanganan Outlier 5%

Keterangan	Data Asli		Setelah Penanganan Outlier 5%	
	Sehat	Diabetes Melitus	Sehat	Diabetes Melitus

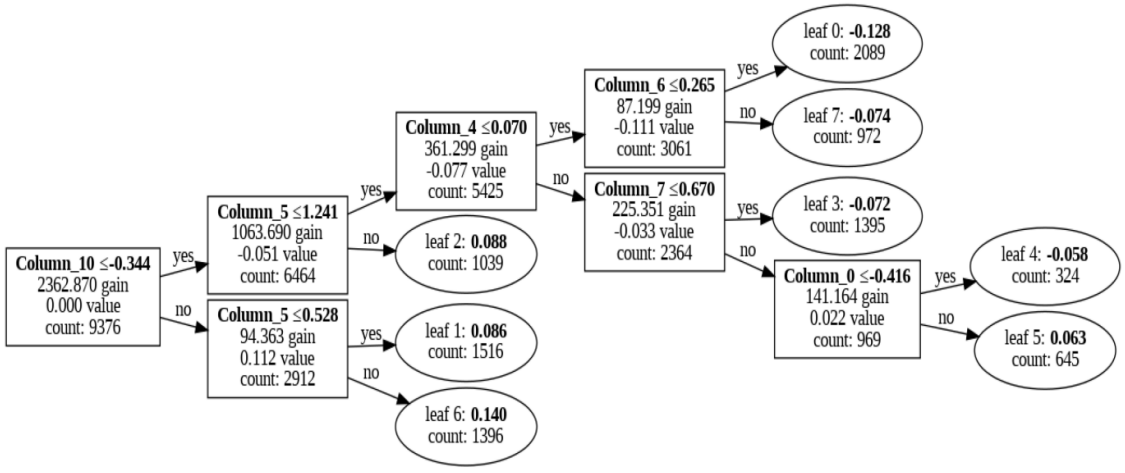
Jumlah Data	4946	342	4708	315
Total	5288		5023	

Tabel 3 menunjukkan data asli sebanyak 5.288 data, yang terdiri atas 4.946 data kategori sehat dan 342 data kategori diabetes melitus. Setelah dilakukan penanganan *outlier* menggunakan metode *Robust Random Cut Forest* dengan persentase penghapusan sebesar 5%, jumlah data berkurang menjadi 5.023 data, yang terdiri atas 4.708 data sehat dan 315 data diabetes melitus. Data setelah penanaganan *outlier* kemudian akan dilakukan penanganan data tidak seimbang menggunkaan metode SMOTE yang mana data minoritas akan ditambah dengan data sintesis sehingga jumlah data sehat dan diabetes melitus menjadi sama. Tahap selanjutnya yakni membagi data menggunakan metode K-Fold Cross Validation dengan nilai k=5 seperti pada Tabel berikut:

Tabel 4. Pembagian Data Menggunakan *K-Fold Cross Validation*

Fold ke-	Data Latih		Data Uji	
	Tidak Mengalami DM	Mengalami DM	Tidak Mengalami DM	Mengalami DM
1	3766	3766	942	63
2	3766	3766	942	63
3	3767	3767	941	63
4	3766	3767	941	63
5	3766	3766	942	63

Berdasarkan tabel tersebut, jumlah data pada setiap fold relatif konsisten, baik pada kelas tidak mengalami diabetes melitus maupun mengalami diabetes melitus. Konsistensi distribusi data ini menunjukkan bahwa proses pembagian data dilakukan secara merata sehingga karakteristik setiap fold tetap terjaga dan dapat merepresentasikan keseluruhan dataset. Kondisi tersebut penting untuk meminimalkan bias selama proses pelatihan dan pengujian model, karena setiap iterasi menggunakan distribusi data yang hampir sama. Kemudian akan dilakukan proses klasifiaksi menggunakan metode LightGBM.



Gambar 3. Pohon Klasifikasi LightGBM

Gambar 3 menunjukkan pohon keputusan yang dibentuk oleh algoritma LightGBM dalam proses klasifikasi diabetes melitus. Proses dimulai dengan memilih variabel yang paling mampu membedakan data, kemudian data dipisahkan ke dalam beberapa percabangan sesuai dengan nilai ambang (*threshold*) yang telah ditentukan. Proses pemisahan ini berlangsung secara berulang hingga setiap data mencapai bagian akhir pohon (*leaf*). Pada setiap titik percabangan ditampilkan nilai *count*, yaitu jumlah data yang berada pada bagian tersebut, serta *split gain*, yaitu ukuran yang menunjukkan seberapa besar peningkatan kualitas pemisahan data setelah dilakukan percabangan. Bagian akhir pohon (*leaf*) merupakan hasil akhir dari proses pengambilan keputusan. Nilai value pada setiap *leaf* menunjukkan kecenderungan prediksi yang dihasilkan oleh model terhadap data yang berada pada bagian tersebut. Semakin sesuai pola data dengan aturan yang terbentuk pada pohon, semakin baik model dalam membedakan antara kelas diabetes melitus dan non-diabetes. Proses kemudian akan dilanjutkan dengan evaluasi hasil menggunakan *confussion matrix*.

Berdasarkan perhitungan evaluasi hasil menggunakan persamaan (4), (5), dan (6) yang dilakukan pada ke lima fold, didapatkan nilai ke lima *fold* yang dapat dilihat pada table berikut.

Tabel 5. Nilai Evaluasi Tiap *Fold*

Fold	Akurasi	Sensitivitas	Spesifitas
1	84.98%	75.81%	85.58%
2	85.27%	69.3%	86.32%
3	84.58%	61.29%	86.11%
4	82.47%	72.58%	83.12%
5	85.36%	79.03%	85.77%

Berdasarkan table di atas, hasil terbaik diperoleh pada *fold* kelima dengan sensitivitas sebesar 79,03%, akurasi 85,36%, dan spesifisitas 85,77%. Penilaian kinerja model diprioritaskan pada sensitivitas karena menunjukkan kemampuan model dalam mengenali pasien yang benar-benar menderita diabetes melitus. Hal ini penting untuk meminimalkan kesalahan klasifikasi pasien diabetes sebagai pasien sehat (*false negative*). Oleh karena itu, pemilihan model didasarkan pada urutan sensitivitas, akurasi, kemudian spesifisitas.

Adapun uji coba pada penelitian ini dilakukan dengan menerapkan Pengujian pada algoritma LightGBM melalui proses *hyperparameter tuning*. Pengujian ini dilakukan untuk mendapatkan nilai akurasi, sensitivitas, dan spesifitas terbaik pada model klasifikasi. Parameter yang diuji meliputi *learning rate* sebesar 0.075, 0.05, dan 0.1 untuk mengatur 60 kecepatan pelatihan model dalam proses iterasi. Selanjutnya, parameter minimum jumlah data pada daun (*minimum data in leaf*) diuji menggunakan nilai 10, 20, 40, dan 80 untuk melihat pengaruh struktur pohon keputusan terhadap hasil klasifikasi. Kemudian, parameter *num levaes* nilai 2- 16 dengan interval 2 untuk mengatur jumlah daun pada pohon keputusan yang dibentuk oleh algoritma LightGBM.

Tabel 6. Perbandingan Parameter Terbaik pada Setiap *Learning Rate*

Learning Rate	*Num Leaves*	*Min Data in Leaf*	Akurasi (%)	Sensitivitas (%)	Spesivitas (%)
0.075	2	10	84.77	69.10	85.83
0.05	4	80	84.39	72.26	85.21
0.1	2	10	85.13	69.84	86.15

Berdasarkan Tabel di atas, parameter terbaik pada learning rate 0.05 diperoleh pada num leaves 4 dan min data in leaf 80 dengan sensitivitas sebesar 72,26%, akurasi 84,39%, dan spesifisitas 85,21%. Karena penelitian ini berfokus pada deteksi diabetes melitus, sensitivitas menjadi metrik utama yang dipertimbangkan. Sensitivitas yang tinggi menunjukkan kemampuan model yang lebih baik dalam mengidentifikasi pasien yang benar-benar menderita diabetes. Oleh karena itu, kombinasi parameter *learning rate* 0.05, *num leaves* 4, dan *min data in leaf* 80 dipilih sebagai model yang paling optimal meskipun memiliki akurasi dan spesifisitas yang sedikit lebih rendah dibandingkan beberapa kombinasi parameter lainnya. Dalam konteks diagnosis diabetes, kesalahan dalam mengidentifikasi pasien yang menderita diabetes (*false negative*) lebih berisiko dibandingkan kesalahan dalam mengidentifikasi pasien yang tidak menderita diabetes (*false positive*). Kombinasi parameter tersebut akan digunakan ke dalam proses klasifikasi menggunakan metode LightGBM.

Kesimpulan

Berdasarkan hasil uji coba parameter, diperoleh model LightGBM terbaik dengan kombinasi parameter *learning rate* sebesar 0,05, *num leaves* sebesar 4, dan *min data in leaf* sebesar 80. Kombinasi parameter tersebut mampu menghasilkan performa klasifikasi yang optimal dalam mengidentifikasi penyakit diabetes melitus. Hasil ini menunjukkan bahwa uji coba parameter berperan penting dalam meningkatkan kemampuan model untuk mempelajari pola data dan menghasilkan prediksi yang lebih baik.

Pada penelitian ini diketahui dari data yang digunakan, terdapat data yang memiliki nilai *outlier* sehingga pada penelitian ini dilakukan penanganan data *outlier* menggunakan metode Robust Random Cut Forest. Hasil dari penanganan data *outlier* tersebut menunjukkan pada nilai skor bahwa semakin besar nilainya maka data tersebut termasuk dalam data *outlier*. Berdasarkan hasil penelitian yang telah dilakukan, penggunaan metode RRCF dan SMOTE diketahui nilai sensitivitas sebesar 79.03%, nilai akurasi sebesar 85.36%, dan nilai spesifitas sebesar 85.77%. Hasil tersebut menunjukkan bahwa model mampu mengidentifikasi sebagian besar pasien yang terindikasi diabetes melitus, meskipun masih terdapat sejumlah pasien diabetes yang belum berhasil diklasifikasikan dengan tepat. Nilai sensitivitas sebesar 79,03% menunjukkan bahwa model mampu mengenali sekitar 79 dari setiap 100 pasien yang benar-benar mengalami diabetes melitus. Sementara itu, nilai spesifisitas sebesar 85,77% menunjukkan kemampuan model dalam mengenali individu yang tidak mengalami diabetes melitus. Dengan demikian, kombinasi RRCF dan SMOTE memberikan hasil klasifikasi yang cukup baik pada data penelitian, tetapi kinerjanya masih perlu ditingkatkan, khususnya dalam mengurangi kesalahan klasifikasi pada pasien diabetes.

Referensi

[1] S. Supriyo, M. Y. Ibrahim, P. N. D. Santoso, and S. U. Dwiningsih, "Diabetes Self Management Education (Dsme) Upaya Preventif Untuk Remaja Terhindar Diabetes Melitus," *J. Lintas Keperawatan*, vol. 5, no. 1, pp. 355–361, 2024, doi: 10.31983/jlk.v5i1.11403.

[2] IDF, "Diabetes in Indonesia," *International Diabetes Federation*, 2024. https://idf.org/our-network/regions-and-members/western-pacific/members/indonesia/ (accessed Sep. 01, 2025).

[3] H. Yun, J. won Park, and J. K. Kim, "A Comparative Evaluation of HbA1c Measurement Methods and Their Implications for Diabetes Management," *Diagnostics*, vol. 13, no. 22, 2023, doi: 10.3390/diagnostics13223449.

[4] N. Cassim, L. M. Coetzee, and D. K. Glencross, "Categorising specimen referral delays for CD4 testing: How inter-laboratory distances and travel times impact turn-around time across a national laboratory service in South Africa," *Afr. J. Lab. Med.*, vol. 9, no. 1, pp. 1–7, 2021, doi: 10.4102/AJLM.V9I1.1120.

[5] E. Alpaydin, "Introduction of Machine Learning," in *The MIT Press*, 3rd ed.The MIT Press, 2014.

[6] A. Ma and F. Fauzi, "Perbandingan Klasifikasi Random Forest , Support Vector Machines , dan LGBM Pada Klasifikasi Kualitas

Udara di Jakarta Comparison of Random Forest , Support Vector Machines , and LGBM Classification for Air," *J. Sist. dan Teknol. Inf. Indones.*, vol. 9, no. 2, pp. 99–108, 2024, [Online]. Available: https://ejurnal.unmuhjember.ac.id/index.php/JUSTINDO/article/view/1912

[7] A. Al Tawil, L. Almazaydeh, B. Alqudah, A. Z. Abualkishik, and A. A. Alwan, "Predictive modeling for breast cancer based on machine learning algorithms and features selection methods," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 2, pp. 1937–1947, 2024, doi: 10.11591/ijece.v14i2.pp1937-1947.

[8] X. Xiong *et al.*, "Application of LightGBM hybrid model based on TPE algorithm optimization in sleep apnea detection," *Front. Neurosci.*, vol. 18, no. February, pp. 1–17, 2024, doi: 10.3389/fnins.2024.1324933.

[9] S. Vashisth and A. Goyal, "Dynamic Anomaly Detection Using Robust Random Cut Forests in Resource-Constrained IoT Environments," *Inform.*, vol. 48, no. 23, pp. 107–120, 2024, doi: 10.31449/inf.v48i23.6862.

[10] I. Fadillah, Nurahmi, and L. B. Kurniawan, "Correlation between Serum Visfatin Level with Glycemic Control in Type 2 Diabetes Mellitus Subjects," *Indones. J. Clin. Pathol. Med. Lab.*, vol. 31, no. 3, pp. 299–303, 2025, doi: 10.24293/ijcpml.v31i3.2304.

[11] B. De la Cruz Concepción, Y. A. F. Cortez, M. I. B. Bonilla, J. M. M. Bello, and M. E. Rojo, "Insulin: A connection between pancreatic β cells and the hypothalamus," *World J. Diabetes*, vol. 14, no. 2, pp. 76–91, 2023, doi: 10.4239/wjd.v14.i2.76.

[12] K. Kesehatan, "Kariadi Podcast Episode 35- Diabetes Melitus," 2024. https://kms.kemkes.go.id/pengetahuan/detail/66b2f0c6108c3c91f387d845 (accessed Oct. 19, 2025).

[13] S. P. Collins *et al.*, "BUKU SAKU DIABETES MELITUS UNTUK AWAM," in *UNS Press*, 2021, pp. 167–186.

[14] S. Guha, N. Mishra, and G. Roy, "Robust Random Cut Forest Based Anomaly Detection On Streams," vol. 48, 2016.

[15] H. Vardhan and J. Sztipanovits, "Reduced Robust Random Cut Forest for Out-of-Distribution Detection in Machine Learning Models," *Int. J. Mach. Learn. Comput.*, vol. 12, no. 5, 2022, doi: 10.18178/ijmlc.2022.12.5.1104.

[16] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," vol. 16, pp. 321–357, 2002.

[17] Hartono, "Modul Digital Machine Learning,"

[18] V. W. Lumumba, D. Kiprotich, N. G. Makena, and M. D. Kavita, "Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models," *Am. J. Theor. Appl. Stat.*, vol. 13, no. 5, pp. 127–137, 2024.

[19] G. Ke *et al.*, "LightGBM : A Highly Efficient Gradient Boosting Decision Tree," no. Nips, pp. 1–9, 2017.

[20] Y. Wang and T. Wang, "applied sciences Application of Improved LightGBM Model in Blood Glucose Prediction," *J. Appl. Sci.*, 2020.

[21] E. Ramadanti, D. A. Dinathi, C. Sri, K. Aditya, and D. R. Chandranegara, "Diabetes Disease Detection Classification Using Light Gradient Boosting (LightGBM) With Hyperparameter Tuning," vol. 8, no. 2, pp. 956–963, 2024.