KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA CATEGORICAL BOOSTING DENGAN FAKTOR RISIKO DIABETES

Naufal Levi Sabili, Fajri Rakhmat Umbara, Melina
Teknik Informatika, Universitas Jenderal Achmad Yani
Jl. Terusan Jend. Sudirman, Kota Cimahi, Jawa Barat 40525 Indonesia
levisabili6@gmail.com

ABSTRAK

Indonesia menjadi salah satu negara yang memiliki jumlah penderita Diabetes Melitus tertinggi di dunia. Penyakit diabetes dapat menimbulkan komplikasi serius yang berpotensi membahayakan bagi penderitanya. Penelitian ini bertujuan untuk mengembangkan model prediksi yang akurat untuk mengklasifikasikan penyakit diabetes menggunakan algoritma *Categorical Boosting* (CatBoost) dengan mempertimbangkan berbagai risiko penyakit diabetes. CatBoost dikenal karena kemampuannya menangani data kategorikal dengan baik. Tahap awal dalam penelitian ini adalah pengolahan data atau *pre-processing*, yang meliputi pembersihan data untuk menangani masalah data yang tidak bersih, penanganan data dengan nilai ekstrem, dan memperbaiki tipe data yang tidak sesuai. Selanjutnya, dilakukan tahap pembuatan model prediksi menggunakan algoritma CatBoost yang merupakan metode *gradient boosting* yang efektif dalam pengambilan keputusan. Evaluasi model dilakukan menggunakan Confusion Matrix untuk menilai performa klasifikasi. Hasil penelitian menunjukkan akurasi yang cukup tinggi dalam klasifikasi pada penyakit diabetes yaitu sebesar 98,63% berdasarkan atribut yang digunakan pada data. Diharapkan, penelitian ini dapat memberikan kontribusi dalam memberikan pemahaman dan upaya pengelolaan risiko diabetes serta tingkat kematian yang disebabkan oleh penyakit tersebut.

***Kata kunci** : algoritma, CatBoost, diabetes, klasifikasi, prediksi*

1. PENDAHULUAN

Diabetes melitus (DM) merupakan penyakit autoimun yang disebabkan dari faktor keturunan, lingkungan, pola makan, dan faktor lainnya [1]. DM merupakan penyakit yang berkaitan dengan kesehatan pankreas, dimana ketidaknormalan dalam produksi hormon insulin oleh pankreas dapat mengakibatkan peningkatan kadar glukosa dalam darah. Kenaikan kadar glukosa dalam tubuh manusia dapat mengganggu fungsi organ-organ vital seperti ginjal, jantung, dan otak [2]. Pada tahun 2019, Organisasi Kesehatan Dunia (WHO) mencatat bahwa setidaknya terdapat 2 juta kasus kematian yang dapat dikaitkan dengan kondisi diabetes [3]. Menurut laporan resmi Kementerian Kesehatan RI tahun 2018 dalam Konferensi Suara Dunia Perangi Diabetes, Indonesia menempati peringkat keenam sebagai negara dengan jumlah penderita diabetes terbanyak di dunia. Data tersebut menunjukkan bahwa jumlah penderita diabetes di Indonesia pada rentang usia 20-79 tahun mencapai sekitar 10,3 juta orang [4].

Keterlambatan diagnosis penyakit diabetes merupakan salah satu faktor yang menyebabkan terjadinya peningkatan jumlah penderita diabetes, yang apabila tidak dikontrol dengan baik, dapat menimbulkan komplikasi penyakit yang membahayakan nyawa penderitanya [5]. Dalam upaya pencegahan dan menurunkan jumlah pasien penderita DM telah dilakukan berbagai cara tetapi masih belum menunjukkan hasil yang diharapkan sehingga perlu dilakukan penelitian-penelitian lanjut untuk mengembangkan sistem pendeteksi diabetes sebagai pencegahan dan pengobatan segera [6]. Penelitian terbaru di bidang bioinformatika menunjukkan bahwa deteksi dini penyakit diabetes melitus menggunakan mesin pembelajaran (ML) lebih baik dan efisien dibandingkan deteksi manual [7].

Algoritma CatBoost dan XGBoost merupakan pengembangan dari algoritma ML [8]. CatBoost, singkatan dari "*Categorical Boosting*," difokuskan pada fitur kategorik. Ini merupakan algoritma pohon keputusan peningkatan gradien yang menggunakan *decision trees* sebagai prediktor dasar [9]. *CatBoost* diperkenalkan oleh Prokhorenkova [10], dan Dorogush [11] untuk mencapai kerugian terendah dalam konteks fitur kategorikal dan dataset yang besar [12], untuk tingkat kehilangan informasi yang minimal [13]. Catboost memiliki salah satu ciri khas yaitu mekanisme *gradient boosting* yang dirancang khusus untuk menangani data yang heterogen, meningkatkan stabilitas, dan memberikan kemampuan prediksi yang optimal. Pendekatan ini menggunakan pengkodean yang efisien dengan tujuan mengurangi risiko *overfitting*. Konsep ini memiliki potensi untuk mempengaruhi tingkat akurasi model, dan oleh karena itu, Catboost diciptakan untuk mengatasi kelemahan tersebut dengan maksud meningkatkan tingkat akurasi secara signifikan [14].

Beberapa penelitian sebelumnya yang telah mengkaji tentang klasifikasi diabetes, yaitu penelitian dilakukan oleh Bhoi, S.K. tentang prediksi diabetes pada wanita keturunan *Indian Pima* yang diambil dari dataset *Pima Indians* Diabetes Database. Penelitian ini menggunakan klasifikasi biner untuk memprediksi apakah seorang wanita keturunan *Indian Pima* menderita diabetes atau tidak, hasilnya ditemukan bahwa algoritma *Logistic Regression* mencapai tingkat akurasi sebesar 76,80% [15]. Selanjutnya penelitian

oleh Agatsa dkk tentang klasifikasi pasien pengidap diabetes berdasarkan data yang diperoleh dari dataset *Pima Indians* Diabetes Databese, yang mencapai nilai akurasi tertinggi sebesar 77.92% dengan menggunakan algoritma *Support Vector Machine* (SVM) [16]. Penelitian yang dilakukan oleh Sisodia dkk yang membahas tentang deteksi dini diabetes dan upaya untuk memprediksi penyakit tersebut menggunakan algoritma klasifikasi, menggunakan dataset dari Pima Indians Diabetes yang mencapai tingkat akurasi sebesar 76.30% dengan menerapkan metode *Naive Bayes* [17]. Penelitian yang berjudul "Implementasi Catboost Menggunakan Hyper-Parameter Tuning Bayesian Search Untuk Memprediksi Penyakit Diabetes menggunakan CatBoost untuk memprediksi penyakit diabetes" dataset yang digunakan dalam penelitian ini adalah *Pima Indian Diabetes (PID*) dataset, dimana penelitian ini mendapatkan nilai AUC sebesar 0,868 dan presisi sebesar 62,5% dan menggunakan *CatBoost* dengan *hyper-parameter Tuning Bayesian Search* memperoleh hasil kinerja model dengan nilai AUC sebesar 0.901 dan akurasi sebesar 63,46% [14].

Berdasarkan latar belakang, penelitian ini bertujuan menggunakan algoritma CatBoost untuk melakukan klasifikasi terhadap penyakit diabetes serta mengevaluasi nilai akurasi yang dihasilkan dari algoritma Catboost dengan menggunakan 14 atribut Kategorik yaitu: ID, No_Pation, Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI, dan CLASS.

2. TINJAUAN PUSTAKA

2.1. Diabetes

Diabetes merupakan suatu kondisi yang mempengaruhi metabolisme tubuh, menunjukkan gejala hiperglikemia akibat gangguan dalam sekresi insulin, aksi insulin, atau keduanya. Diabetes melitus menjadi masalah kesehatan global karena tingkat morbiditas dan mortalitas yang tinggi yang terkait dengan penyakit ini [18].

Gaya hidup merupakan salah satu faktor penting dalam prevalensi diabetes mellitus (DM) di masyarakat. DM termasuk dalam kategori penyakit tidak menular yang jumlah kasusnya terus meningkat setiap tahun [19].

2.2. Klasifikasi

Proses klasifikasi merujuk pada aktivitas pengelompokan objek berdasarkan pola yang ada pada model objek klasifikasi atau memilah data berdasarkan kelompoknya [20]. Pembentukan model klasifikasi dilakukan dengan memanfaatkan data pelatihan yang ada, dan model tersebut selanjutnya digunakan untuk mengategorikan data yang baru muncul. Dalam konsepnya, klasifikasi dapat dinyatakan sebagai suatu tindakan yang melibatkan pelatihan atau pembelajaran fungsi tujuan yang mengaitkan setiap kumpulan atribut (karakteristik) dengan sejumlah label kelas yang tersedia [21].

2.3. Machine Learning

Machine Learning (ML) merupakan bidang ilmiah yang terkait dengan cara di mana mesin memperoleh pengetahuan dari pengalaman [22]. Fokus ML adalah mengembangkan sistem komputer yang dapat menyesuaikan diri dan memperoleh pengetahuan melalui pengalaman yang mereka dapatkan [23]. Teknologi ML memungkinkan deteksi pola yang tersembunyi dalam data dan penyesuaian algoritma untuk meningkatkan keakuratan hasil [24].

2.4. Data Mining

Data mining adalah proses menemukan pola-pola potensial yang bermanfaat dengan memanfaatkan kumpulan data yang sangat besar. Berbagai aspek dari kegiatan data mining meliputi klasifikasi data, integrasi data, transformasi data, diskritisasi data, serta evaluasi pola, dan lain-lain. Teknik penambangan data diterapkan untuk mengungkapkan hubungan tersembunyi dan tidak terduga antara data, dengan aplikasinya [25].

2.5. SMOTE

Synthetic *Minority Over-sampling Technique* (SMOTE) adalah sebuah metode inovatif yang digunakan untuk mengatasi masalah ketidakseimbangan kelas dalam dataset, yang sering kali menjadi tantangan dalam analisis data. Metode ini bekerja dengan cara menciptakan sampel sintetis tambahan untuk kelas minoritas melalui proses interpolasi antara contoh-contoh minoritas yang sudah ada. SMOTE tidak hanya menambah jumlah sampel dari kelas minoritas tetapi juga mempertahankan karakteristik data yang ada. Peningkatan representasi kelas minoritas dalam dataset ini membantu model klasifikasi dalam mempelajari fitur-fitur yang relevan, sehingga meningkatkan kemampuan model untuk mengenali dan mengidentifikasi pola yang berhubungan dengan kelas minoritas secara lebih akurat dan efisien [26].

2.6. Gradient Boosting

Teknik *Machine Learning* (ML) lain yang digunakan untuk regresi dan klasifikasi adalah *Gradient Boosting* (GB). Teknik ini melakukan prediksi dalam bentuk ansambel model prediksi lemah *Decision Tree* (DT). Konsep peningkatan pertama kali diajukan dalam referensi, yang menyatakan bahwa peningkatan dapat diinterpretasikan sebagai algoritma optimasi untuk fungsi biaya tertentu. Pada tahap selanjutnya, algoritma regresi GB diimplementasikan oleh dengan pendekatan yang berbeda dan penerapan praktis Sebagai hasilnya, perkembangan berbagai algoritma peningkatan teramati di berbagai bidang seperti statistik dan kecerdasan buatan (*Artificial Intelligence*) [28].

Metode Gradient Boosting menggunakan pengklasifikasi yang memiliki korelasi kuat dengan klasifikasi yang sesungguhnya. Model ini dikembangkan dengan menggabungkan beberapa

pengklasifikasi yang memiliki korelasi rendah dengan klasifikasi aktual untuk menghasilkan prediksi yang lebih akurat. Pendekatan ini diimplementasikan secara berulang [29].

2.7. CatBoost

Categorical Boosting (CatBoost) adalah algoritma penguatan bertingkat yang didasarkan pada metode penguatan gradien, yang menggunakan pohon keputusan untuk menyelesaikan tugas regresi. Dalam algoritma ini, setiap pohon bertanggung jawab untuk membagi ruang fitur dengan cara tertentu dan menghasilkan nilai keluaran yang sesuai [30].

CatBoost menangani data kategorikal secara otomatis dengan memanfaatkan metode statistik. Algoritma ini mampu mencegah overfitting dengan mengoptimalkan berbagai parameter input dan mengabaikan karakteristik kategori selama pemrosesan data [27]. Dalam pengolahan data kategorikal, CatBoost melakukan eksekusi permutasi secara acak daripada mengubahnya menjadi biner, serta menghitung nilai rata-rata untuk setiap label [28].

2.8. Transform Data

Transform Data adalah proses mengubah dan mengkonsolidasikan data ke dalam bentuk yang dapat digunakan untuk penambangan melalui operasi ringkasan atau agregasi dikenal sebagai transformasi data [29]. Proses pemrosesan data terdiri dari 7 (tujuh) tahapan, yang empat tahap pertama dikenal sebagai proses *preprocessing data*, yang terdiri dari pembersihan data, integrasi data, pemilihan data, dan transformasi data. Implementasi proses ini membutuhkan waktu sekitar 60% dari total proses. Metode transformasi data termasuk *smoothing*, *generalization*, *normalization*, *aggregation*, dan *attribute construction* [30]. Penelitian ini menggunakan teknik *binning* karena termasuk kedalam metode *smoothing* dalam melakukan transform data.

2.9. Confusion Matrix

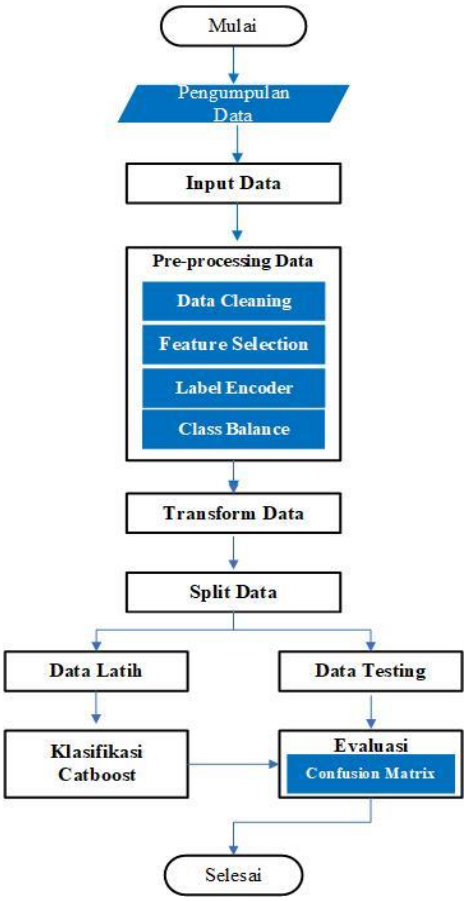
Tabel confusion matrix berfungsi untuk mengevaluasi performa model klasifikasi dengan membandingkan prediksi yang dihasilkan oleh model terhadap nilai sebenarnya dari data. *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) adalah empat komponen confusion matrix [31]. Confusion matrix dapat dilihat pada Tabel 1.

Tabel 1. Confusion Matrix

	Positive	*Negative*
Positive	*True Positive*	*False Positive*
Negative	*False Negative*	*True Negative*

3. METODE PENELITIAN

Metode yang digunakan pada penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Metode Penelitian

3.1 Data Collection

Pada penelitian ini, data yang digunakan adalah data pasien dari laboratorium Rumah Sakit *Medical City* (Pusat Spesialisasi Endokrinologi dan Rumah Sakit Pendidikan Diabetes-Al-Kindy). Berkas pasien diambil dan data diekstrak dan dimasukkan ke dalam database untuk membangun kumpulan data diabetes. Data tersebut memiliki 1000 baris data dengan 14 atribut yang terdiri dari informasi medis dan analisis laboratorium.

3.2 Preprocessing

Tahap ini melibatkan beberapa proses penting untuk mempersiapkan data yang nantinya digunakan dalam proses analisi dan pemodelan. Proses yang dilakukan pada penelitian ini termasuk *Data Cleaning, Feature Selection, Label Encoder dan Class Balancing*.

3.3 Transformasi Data

Transformasi data menggunakan teknik binning untuk mengubah fitur numerik menjadi kategori. Proses ini dimulai dengan mengidentifikasi fitur berisi nilai kontinu yang perlu diubah, seperti usia, kolesterol, dan BMI. Kemudian, distribusi data dianalisis menggunakan histogram atau grafik lainnya untuk menentukan interval bin yang tepat. Interval ini ditetapkan berdasarkan pengetahuan domain atau distribusi data, misalnya usia dibagi menjadi 'Remaja',

'Dewasa', dan 'Lansia'. Selanjutnya, nilai kontinu diubah menjadi kategori sesuai interval yang ditentukan menggunakan fungsi tertentu. Terakhir, hasil binning diperiksa untuk memastikan setiap nilai kontinu dikonversi dengan benar dan distribusi data masuk akal.

3.4 Klasifikasi CatBoost

Pada tahapan ini, data yang telah melewati tahapan pra-proses sebelumnya akan di-split menjadi dua subset, yaitu data latih dan data uji. Data latih digunakan untuk melatih model CatBoost, sedangkan data uji digunakan untuk mengevaluasi kinerja model. Persentase pembagian data adalah 80% untuk data latih dan 20% untuk data uji.

Model CatBoostClassifier dilatih menggunakan parameter-parameter yang telah ditentukan, yaitu learning_rate=1, depth=16, dan iterations=50. Proses pelatihan dilakukan dengan menggunakan data latih dan fitur kategorikal yang telah ditentukan. CatBoost memanfaatkan teknik boosting gradient untuk membangun model secara iteratif dengan menambahkan pohon keputusan baru untuk mengurangi kesalahan yang dibuat oleh model sebelumnya.

3.5 Evaluasi Model

Evaluasi dilakukan menggunakan data uji. Prediksi dilakukan pada data uji dan hasil prediksi dibandingkan dengan label sebenarnya untuk menghitung metrik evaluasi seperti akurasi, precision, recall, dan F1-score. Selain itu, matriks kebingungan (confusion matrix) juga digunakan untuk mengevaluasi kinerja model.

4. HASIL DAN PEMBAHASAN

Penelitian ini melakukan klasifikasi penyakit diabetes menggunakan metode CatBoost. Tujuan penelitian ini dilakukan untuk menghasilkan nilai akurasi pada metode klasifikasi yang dipakai untuk mengklasifikasi penyakit diabetes.

4.1. Data Collection

Tabel 2 menunjukan contoh dari dataset yang digunakan pada penelitian ini.

Tabel 2. Dataset Diabetes

No	ID	No_Pation	Gender	...	CLASS
1	502	17975	F	…	N
2	735	34221	M	…	N
3	420	47975	F	…	N
4	680	87656	F	…	N
5	504	34223	M	…	N
…	…	…	…	…	…
995	198	454316	M	…	Y
996	199	454316	M	…	Y
997	200	454317	M	…	Y
998	671	876534	M	…	Y
999	669	87654	M	…	Y
1000	99	24004	M	…	Y

Pada Tabel 3 menunjukan atribut yang ada pada dalam dataset yang digunakan dala penelitian ini.

Tabel 3. Atribut Data

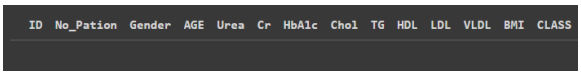
No	Atribut Data
1	ID
2	No_Pation
3	Gender
4	Age
5	Urea
6	Cr
7	HBA1c
8	Chol
9	TG
10	HDL
11	LDL
12	VLDL
13	BMI
14	Class

Berdasarkan data yang diperoleh, diketahui bahwa data penyakit diabetes memiliki 14 atribut. Dari 14 atribut tersebut, 2 di antaranya bertipe kategorikal, sementara 12 atribut lainnya merupakan tipe numerik.

4.2. Preprocessing

Langkah berikutnya adalah melakukan preprocessing data. Tahap ini mencakup beberapa proses utama seperti *data cleaning, feature selection, label encoder* dan *class balancing*.

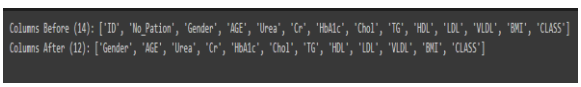
Langkah pertama dalam tahap ini adalah membersihkan data, di mana data yang diperoleh akan melalui proses pembersihan untuk mengatasi informasi yang tidak lengkap, seperti data curah hujan yang tidak lengkap. Tujuan dari pembersihan ini adalah untuk mempermudah proses analisis dan pengolahan data lebih lanjut, yang bisa dilihat pada Gambar 2.



Gambar 2. Data Cleaning

Berdasarkan Gambar 2 dapat dilihat bahwa dalam data penyakit diabetes tidak terdapat nilai yang hilang.

Pada Gambar 3 dapat disimpulkan bahwa atribut yang akan digunakan sebagai fitur adalah *Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI.* Sedangkan untuk target atau kelas yang akan digunakan adalah atribut *Class.*



Gambar 3. Feature Selection

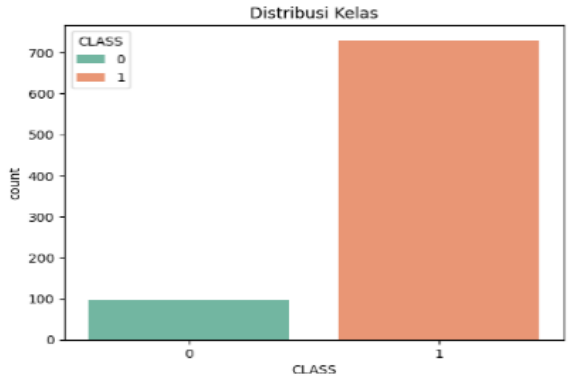
Tahap selanjutnya adalah transformasi fitur yang awalnya berbentuk kategorikal diubah menjadi numerik menggunakan Label Encoder bisa dilihat pada Gambar 4. Langkah ini diperlukan agar data

dapat digunakan untuk keperluan oversampling dengan metode *Synthetic Minority Over-sampling Technique* (SMOTE) dan untuk memastikan kompatibilitas dengan algoritma *Machine Learning* yang digunakan.

```
Before: ['F', 'M']
Before: ['N', 'Y']
After: [0, 1]
After: [0, 1]
```

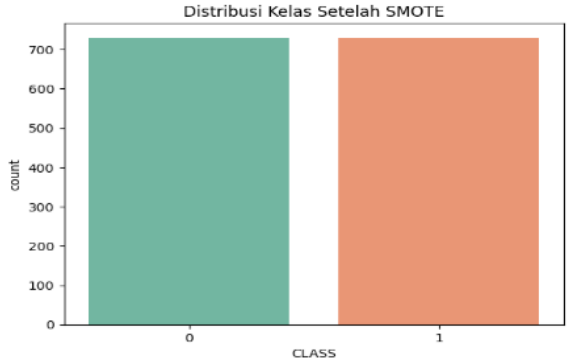
Gambar 4. Label Encoder

Tahap selanjutnya adalah menyeimbangkan kelas dalam data yang diperoleh, di mana terdapat ketidakseimbangan kelas seperti yang terlihat pada Gambar 5.



Gambar 5. Sebelum SMOTE

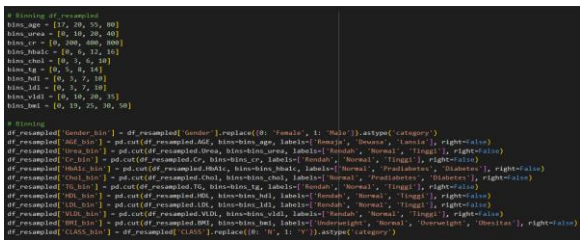
Pada Gambar 5 diperlihatkan dua kelas utama, yaitu "0" dan "1", yang menunjukkan ketidakseimbangan yang cukup signifikan. Untuk mengatasi ketidakseimbangan ini, dilakukan *oversampling* menggunakan SMOTE pada kelas minoritas agar data yang sebelumnya tidak seimbang seperti yang terlihat pada Gambar 6, menjadi seimbang,



Gambar 6. Sesudah SMOTE

4.3. Transform Data

Setelah pemrosesan terhadap data telah selesai, langkah selanjutnya adalah mentransformasi data dengan teknik binning untuk mengubah fitur numerik menjadi kategorik. Proses binning dimulai dengan mengidentifikasi fitur ber-nilai kontinu yang perlu diubah menjadi kategori, seperti usia, tingkat kolesterol, atau BMI. Selanjutnya, distribusi data dianalisis menggunakan histogram atau grafik distribusi lainnya untuk menentukan interval bin yang sesuai, berdasarkan rentang dan frekuensi nilai. Interval bin ditentukan berdasarkan pengetahuan domain atau hasil analisis, misalnya, usia dapat dikategorikan menjadi 'Remaja', 'Dewasa', dan 'Lansia'. Nilai kontinu kemudian diubah menjadi kategori sesuai dengan interval bin yang ditetapkan, menggunakan fungsi khusus dalam bahasa pemrograman. Terakhir, hasil binning diperiksa untuk memastikan bahwa konversi nilai kontinu ke kategori sudah benar dan distribusi data dalam kategori tersebut masuk akal. Hasil binning bisa dilihat pada Gambar 7.



Gambar 7. Binning

4.4. Split Data

Setelah selesai melakukan transformasi data, tahapan selanjutnya adalah pembagian data. Pada tahapan ini, data yang telah melewati tahapan pra-proses sebelumnya akan di-split menjadi dua subset, yaitu data latih dan data uji. Data latih digunakan untuk melatih model CatBoost, sedangkan data uji digunakan untuk mengevaluasi kinerja model. Persentase pembagian data adalah 80% untuk data latih dan 20% untuk data uji.

Fitur-fitur yang bersifat kategorikal dalam dataset didefinisikan untuk membantu CatBoost dalam mengidentifikasi dan menangani variabel kategorikal secara otomatis. Dalam penelitian ini, fitur kategorikal adalah Gender_bin, AGE_bin, Urea_bin, Cr_bin, HbA1c_bin, Chol_bin, TG_bin, HDL_bin, LDL_bin, VLDL_bin, dan BMI_bin.

4.5. Klasifikasi CatBoost

Tahap berikutnya adalah proses klasifikasi, di mana penelitian ini melakukan prediksi menggunakan metode CatBoost. Proses dimulai dengan membangun model CatBoost, yang memerlukan pengaturan beberapa parameter yang telah ditentukan melalui berbagai eksperimen untuk menghasilkan prediksi yang optimal. Pada tahap ini, model akan menggunakan parameter yang telah disesuaikan. Beberapa parameter yang digunakan dalam model CatBoost ini dapat dilihat pada Tabel 4.

Tabel 4. Parameter CatBoost

Parameter	Kegunaan
iteration	Batas tertinggi dari jumlah pohon yang bisa dibuat saat menyelesaikan permasalahan dalam pembelajaran mesin..
learning_rate	Menentukan seberapa cepat model belajar dari data
depth	Parameter untuk menentukan kedalaman maksimum dari setiap pohon.

4.6. Evaluasi Model

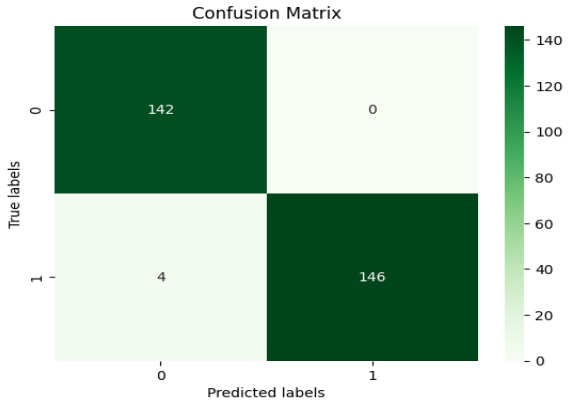
Setelah menyelesaikan tahapan sebelumnya, dilakukan serangkaian percobaan untuk memperoleh hasil prediksi terbaik berdasarkan data penyakit diabetes melalui penyetelan parameter. Eksperimen penyetelan parameter ini bertujuan untuk meningkatkan kinerja model CatBoost, dan hasil dari eksperimen tersebut dapat dilihat pada Tabel 5.

Tabel 5. Eksperimen Parameter

Parameter	Percobaan 1	Percobaan 2	Percobaan 3
iteration	50	100	500
learningrate	1	0.01	0.1
depth	16	8	10
Akurasi	98.6%	95.5%	97.9%

Dari Tabel 5 dapat disimpulkan bahwa akurasi yang dihasilkan dari setiap eksperimen tidak menunjukkan perbedaan yang signifikan, dan eksperimen-eksperimen tersebut telah didasarkan pada penelitian sebelumnya yang melakukan pendekatan serupa.

Pada eksperimen kedua, fokus utama adalah penerapan model CatBoost dengan mengevaluasi akurasi yang diperoleh. Untuk menguji performa model, penelitian ini menggunakan evaluasi dengan confusion matrix. Perhitungan melalui confusion matrix digunakan untuk menentukan persentase data yang benar sesuai kenyataan dibandingkan dengan data yang tidak sesuai atau tidak ada. Hasil dari confusion matrix dapat dilihat pada Gambar 8.



Gambar 8. Confusion Matrix

Pada Gambar 8, terdapat beberapa nilai yang digunakan untuk menghitung akurasi model. Dari hasil confusion matrix tersebut, didapatkan nilai TP sebesar 142, FP sebesar 4, FN sebesar 146, dan TN sebesar 0, yang kemudian digunakan untuk menghitung nilai akurasi, presisi, recall, dan f1-score, dengan hasil yang disajikan pada Tabel 6.

Tabel 6. Matrix Evaluasi

Matriks Evaluasi	Skor
Accuracy	98.63%
Precision	1.0%
Recall	97.33%
F1-Score	98.64%

1) Akurasi dihitung menggunakan persamaan (1).

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Hasil:

$$Akurasi = \frac{142 + 146}{142 + 0 + 4 + 146} = 0.9863$$

2) Precision dihitung mengunakan persamaan (2).

$$Presisi = \frac{TP}{TP + FP} \quad (2)$$

Hasil:

$$Presisi = \frac{146}{146 + 0} = \frac{146}{146} = 1.0$$

3) Recall dihitung menggunakan persamaan (3).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Hasil:

$$Recall = \frac{146}{146 + 4} = \frac{146}{150} = 0.9733$$

4) F1 Score, yang merupakan harmonic mean dari precision dan recall, dihitung menggunakan persamaan (4).

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (4)$$

Hasil:

$$F1 - Score = 2 \times \frac{1.0 \times 0.9733}{1.0 + 0.9733} = 2 \times \frac{0.9733}{1.9733} = 0.9864$$

■

5. KESIMPULAN DAN SARAN

Penelitian ini mengembangkan model prediksi untuk klasifikasi penyakit diabetes menggunakan algoritma CatBoost, dengan dataset 1460 sampel, yang dibagi dari 80% untuk pelatihan dan 20% untuk pengujian. Model yang dikembangkan menunjukkan performa yang sangat baik dalam mengklasifikasikan penyakit diabetes. Selain itu, hasil evaluasi menggunakan Confusion Matrix menunjukkan bahwa model yang dikembangkan mencapai akurasi sebesar 98,63%, presisi sebesar 100%, recall sebesar 97,33%, dan F1-Score sebesar 98,64%. Oleh karena itu, dapat disimpulkan bahwa algoritma CatBoost terbukti efektif dalam menangani data kategorikal tanpa perlu

melakukan encoding, dengan kinerja yang sangat baik dalam prediksi diabetes. Tingginya akurasi, presisi, recall, dan F1-Score juga menunjukkan bahwa model yang dikembangkan dalam penelitian ini dapat digunakan dalam sistem pendukung keputusan medis dalam mengidentifikasi pasien berisiko diabetes lebih awal. Untuk penelitian mendatang, disarankan agar menggunakan dataset yang lebih besar dan beragam serta mencoba algoritma alternatif seperti XGBoost atau teknik deep learning. Pengoptimalan hyperparameter, penambahan fitur relevan seperti data genetik dan riwayat keluarga, serta penerapan teknik validasi kompleks seperti cross-validation dapat meningkatkan validitas dan pemahaman model. Selain itu, perbandingan dengan metode pre-processing lainnya, implementasi dalam konteks nyata, analisis interpretabilitas model, dan pemanfaatan data temporal untuk memantau perubahan penyakit seiring waktu juga layak dipertimbangkan.

DAFTAR PUSTAKA

[1] S. S. Bhat, M. Banu, G. A. Ansari, and V. Selvam, "A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms," *Healthcare Analytics*, vol. 4, p. 100273, 2023, doi: https://doi.org/10.1016/j.health.2023.100273.

[2] M. F. Fahrul and W. Hadikurniawati, "Mohammad Faisal Fahrul Klasifikasi Diabetes Pada Wanita Klasifikasi Diabetes Pada Wanita Menggunakan Metode Naive Bayes Classifier."

[3] M. Rifqi Maulana, M. Faizal Kurniawan, and M. Adib Al Karomi, "BULLETIN OF COMPUTER SCIENCE RESEARCH Komparasi Algoritma Data Mining untuk Klasifikasi Penyakit Diabetes," *Media Online*, vol. 3, no. 5, pp. 343–350, 2023, doi: 10.47065/bulletincsr.v3i5.280.

[4] Gde Agung Brahmana Suryanegara, Adiwijaya, and Mahendra Dwifebri Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 114–122, Feb. 2021, doi: 10.29207/resti.v5i1.2880.

[5] S. Ucha Putri, E. Irawan, F. Rizky, S. Tunas Bangsa, P. A. -Indonesia Jln Sudirman Blok No, and S. Utara, "Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5," 2021.

[6] R. Rizki, R. Athallah, I. Cholissodin, and P. P. Adikara, "Prediksi Potensi Pengidap Penyakit Diabetes berdasarkan Faktor Risiko Menggunakan Algoritme Kernel K-Nearest Neighbor," 2022. [Online]. Available: http://j-ptiik.ub.ac.id

[7] N. Permatasari, S. Asy Syahidah, A. Leofiro Irfiansyah, and M. G. Al-Haqqoni, "PREDICTING DIABETES MELLITUS USING CATBOOST CLASSIFIER AND SHAPLEY ADDITIVE EXPLANATION (SHAP) APPROACH," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 16, no. 2, pp. 615–624, Jun. 2022, doi: 10.30598/barekengvol16iss2pp615-624.

[8] A. Ilmiah Aplikasi Teknologi, Y. Purbolingga, D. Marta Putria, A. Rahmawatia, and B. Wajhi Akramunnas, "JURNAL APTEK Perbandingan Algoritma CatBoost dan XGBoost dalam Klasifikasi Penyakit Jantung," vol. 15, no. 2, pp. 126–133, 2023, [Online]. Available: http://journal.upp.ac.id/index.php/aptek

[9] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost Classifier with other Machine Learning Methods," 2020. [Online]. Available: www.ijacsa.thesai.org

[10] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.09516

[11] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.11363

[12] H. Wang and L. Cheng, "CatBoost model with synthetic features in application to loan risk assessment of small businesses," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.07954

[13] S. Ben Jabeur, C. Gharib, S. Mefteh-Wali, and W. Ben Arfi, "CatBoost model and artificial intelligence techniques for corporate failure prediction," *Technol Forecast Soc Change*, vol. 166, May 2021, doi: 10.1016/j.techfore.2021.120658.

[14] A. Darmawan *et al.*, "Implementasi Catboost Menggunakan Hyper-Parameter Tuning Bayesian Search Untuk Memprediksi Penyakit Diabetes."

[15] S. Kumar Bhoi *et al.*, "Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach," 2021.

[16] D. A. Agatsa, R. Rismala, and U. N. Wisesty, "Klasifikasi Pasien Pengidap Diabetes menggunakan Metode Support Vector Machine."

[17] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 1578–1585. doi: 10.1016/j.procs.2018.05.122.

[18] F. Eryuda and T. U. Soleha, "Artocarpus camansi) dalam Menurunkan Kadar Glukosa Darah pada Penderita Diabetes Melitus MAJORITY I Volume 5 I Nomor 4 I Oktober," 2016.

[19] M. M. Keperawatan, "Faktor-Faktor Gaya Hidup Yang Mempengaruhi Terjadinya Diabetes Melitus Lifestyle Factors That Influence The

Occurrence Of Diabetes Mellitus Alfi Rahim," *JONS : Journal Of Nursing*, vol. 1, no. 2, pp. 9–12, 2024, [Online]. Available: www.journal.medicpondasi.com/index.php/nursing/index

[20] R. Sukmawardani *et al.*, "IMPLEMENTASI ARSITEKTUR RESNET152 UNTUK KLASIFIKASI UANG KERTAS RUPIAH DENGAN METODE TRANSFER LEARNING," 2024.

[21] J. Homepage, B. Delvika, S. Nurhidayarnis, P. D. Rinada, N. Abror, and A. Hidayat, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Comparison of Classification Between Naive Bayes and K-Nearest Neighbor on Diabetes Risk in Pregnant Women Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes Pada Ibu Hamil," vol. 2, pp. 68–75, 2022.

[22] M. Melina, Sukono, H. Napitupulu, and N. Mohamed, "Modeling of Machine Learning-Based Extreme Value Theory in Stock Investment Risk Prediction: A Systematic Literature Review," *Big Data*, vol. 0, no. 0, p. null, doi: 10.1089/big.2023.0004.

[23] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," 2017, *Elsevier B.V.* doi: 10.1016/j.csbj.2016.12.005.

[24] A. Yosipof, R. C. Guedes, and A. T. García-Sosa, "Data mining and machine learning models for predicting drug likeness and their disease or organ category," *Front Chem*, vol. 6, no. MAY, May 2018, doi: 10.3389/fchem.2018.00162.

[25] D. Papakyriakou and I. S. Barbounakis, "Data Mining Methods: A Review," 2022.

[26] B. Nemade, V. Bharadi, S. S. Alegavi, and B. Marakarkandy, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons." [Online]. Available: www.ijisae.org

[27] M. Saber *et al.*, "Examining LightGBM and CatBoost Models for Wadi Flash Flood Susceptibility Prediction," *Geocarto Int*, vol. 37, Nov. 2021, doi: 10.1080/10106049.2021.1974959.

[28] P. S. Kumar, K. Anisha Kumari, S. Mohapatra, B. Naik, J. Nayak, and M. Mishra, "CatBoost ensemble approach for diabetes risk prediction at early stages," in *1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology, ODICON 2021*, Institute of Electrical and Electronics Engineers Inc., Jan. 2021. doi: 10.1109/ODICON50556.2021.9428943.

[29] J. Han, M. Kamber, and J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)," 2011.

[30] H. Junaedi *et al.*, "Prosiding Konferensi Nasional 'Inovasi dalam Desain dan Teknologi'-IDeaTech 2011 DATA TRANSFORMATION PADA DATA MINING".

[31] G. S. P. Ghantasala, N. V. Kumari, and R. Patan, "Chapter 9 - Cancer prediction and diagnosis hinged on HCML in IOMT environment," in *Machine Learning and the Internet of Medical Things in Healthcare*, K. K. Singh, M. Elhoseny, A. Singh, and A. A. Elngar, Eds., Academic Press, 2021, pp. 179–207. doi: https://doi.org/10.1016/B978-0-12-821229-5.00004-5