Available online at www.elektro.itn.ac.id

Journal Of ALINIER Artificial Intelligence & Applications

Journal homepage: www.elektro.itn.ac.id

TEKNIK ELEKTRO ITN MALANG

Perbandingan Kinerja Algoritma XGBoost dan CatBoost dalam Klasifikasi Risiko Penyakit Diabetes

Setyobudi Utomo[a], Syauqita Sigit[b], Niken Sulistyowati[c], Dwi Arman Prasetya[d],Tresna Maulana Fahrudin[e]

[a,b,c,d,e] Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Indonesia
[e] Department of Information and Communication Systems, Okayama University, Okayama 700-8530, Japan
[a] 22083010112@student.upnjatim.ac.id
[b] 22083010083@student.upnjatim.ac.id
[c] 22083010091@student.upnjatim.ac.id
[d] arman.prasetya.sada@upnjatim.ac.id
[e] tresnamf@s.okayama-u.ac.jp

Kata Kunci :

klasifikasi diabetes, XGBoost, CatBoost

ABSTRAK

Diabetes merupakan penyakit kronis dengan prevalensi global yang terus meningkat, termasuk di Indonesia. Deteksi dini sangat penting untuk mencegah komplikasi serius yang terkait dengan kondisi ini. Seiring dengan kemajuan teknologi, pendekatan machine learning semakin banyak diterapkan dalam klasifikasi penyakit dan prediksi risiko. Penelitian ini bertujuan untuk membandingkan kinerja dua algoritma gradient boosting, yaitu XGBoost dan CatBoost, dalam mengklasifikasikan risiko diabetes berdasarkan data medis pasien. Dataset yang digunakan adalah Diabetes Dataset, yang terdiri dari delapan fitur medis dan satu label target. Penelitian ini mencakup proses pra-pemrosesan data, pelatihan model dengan pembagian data latih dan uji, serta evaluasi menggunakan metrik klasifikasi seperti akurasi, presisi, recall, dan F1-score. XGBoost dipilih karena efisiensinya dalam komputasi serta adanya regularisasi bawaan yang membantu mencegah overfitting pada data numerik berskala besar. Sementara itu, CatBoost digunakan karena kemampuannya dalam menangani fitur kategorikal secara langsung melalui teknik random permutation dan ordered boosting. Hasil dari penelitian ini diharapkan dapat berkontribusi pada pengembangan sistem prediksi diabetes yang lebih akurat dan efisien serta menjadi referensi dalam penerapan algoritma boosting di bidang medis.

Diabetes classification, XGBoost, CatBoost

Diabetes is a chronic disease with an increasing global prevalence, including in Indonesia. Early detection is crucial to prevent severe complications associated with this condition. With technological advancement, machine learning approaches are increasingly applied in disease classification and risk prediction. This study aims to compare the performance of two gradient boosting algorithms, XGBoost and CatBoost, in classifying diabetes risk based on patient

medical data. The dataset used is the Pima Indians Diabetes Dataset, which consists of eight medical features and one target label. The research involves data preprocessing, model training with training-testing split, and evaluation using classification metrics such as accuracy, precision, recall, and F1-score. XGBoost is chosen for its computational efficiency and built-in regularization, which helps prevent overfitting in large-scale numerical data. CatBoost is utilized for its strength in handling categorical features natively using random permutation and ordered boosting techniques. The results of this study are expected to contribute to the development of more accurate and efficient diabetes prediction systems and serve as a reference for applying boosting algorithms in the medical domain.

1. PENDAHULUAN

1.1. Latar Belakang

Diabetes adalah salah satu penyakit jangka panjang yang menjadi isu kesehatan di seluruh dunia. penyakit ini ditandai dengan hiperglikemia, atau kadar glukosa darah yang tinggi, akibat gangguan dalam sekresi insulin, kerja insulin, atau keduanya. [1]. Faktor utama penyebab penyakit ini adalah perubahan dalam pola makan dan cara hidup. Faktor tersebut akan mempengaruhi kondisi insulin, tekanan darah, dan kandungan kadar glukosa darah yang meningkat yaitu ≥ 200 mg/dL dan kadang-kadang disertai dengan penurunan berat badan yang tidak diketahui penyebabnya [2]. Hal tersebut merupakan faktor utama yang mempengaruhi manusia terjangkit penyakit diabetes.

Menurut Federasi Diabetes Internasional (IDF), diperkirakan ada 537 juta orang di seluruh dunia yang akan mengalami diabetes pada tahun 2021. Dengan 19,5 juta orang yang memiliki diabetes pada tahun tersebut, Indonesia akan berada di posisi keenam. Diabetes di Indonesia dianggap sebagai masalah kesehatan yang signifikan dan telah menjadi perhatian sejak awal tahun 1980-an. Negara ini memiliki tingkat prevalensi sebesar 6,2% dengan lebih dari 10 juta orang yang hidup dengan diabetes [3]. Jika tidak ditangani dengan baik, diabetes dapat menyebabkan berbagai komplikasi serius seperti penyakit jantung, stroke, gagal ginjal, gangguan penglihatan, hingga amputasi. Oleh karena itu, pendeteksian dini dan pemantauan kadar gula darah menjadi langkah yang sangat penting untuk mencegah komplikasi yang lebih lanjut.

Seiring dengan perkembangan zaman, kemajuan teknologi telah memberikan dampak besar dalam berbagai aspek kehidupan, termasuk di bidang kesehatan. Teknologi medis terus mengalami perkembangan yang pesat, membawa inovasi-inovasi yang mengubah cara diagnosis, perawatan, dan pemantauan kesehatan dilakukan [4]. Machine learning telah menjadi alat yang potensial dalam dunia medis, khususnya dalam identifikasi penyakit diabetes dapat dilakukan dengan melakukan pengklasifikasian untuk membantu dalam penanggulangan penyakit diabetes, maka langkah untuk melakukan analisis dan mencari solusi yang nantinya dapat membantu untuk menurunkan angka tingkat diabetes yang semakin tinggi dengan dilakukannya klasifikasi.

Penelitian sebelumnya yang berjudul "Implementasi Data Mining dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest dan XGBoost" Dalam penelitian tersebut akurasi yang dihasilkan oleh penggunaan algoritma XGboost bernilai 76%. Sedangkan nilai akurasi yang dihasilkan dalam penggunaan algoritma Random Forest bernilai

74% [5]. Penelitian lain berjudul “Analisis Algoritma Gradient Boosting, AdaBoost, dan CatBoost dalam Klasifikasi Kualitas Air” juga membandingkan beberapa algoritma boosting. Dalam penelitian tersebut, algoritma Gradient Boosting menghasilkan akurasi sebesar 60%, AdaBoost sebesar 58%, dan CatBoost menunjukkan akurasi tertinggi sebesar 68% [6].

Dengan ini peneliti mengusulkan pendekatan lanjutan dengan membandingkan algoritma XGBoost dan CatBoost dalam klasifikasi penyakit diabetes. CatBoost adalah algoritma gradient boosting yang dikembangkan oleh Yandex dan dioptimalkan untuk menangani data kategorikal secara langsung. Dalam pengolahan data kategorikal, CatBoost melakukan eksekusi permutasi secara acak daripada mengubahnya menjadi biner, serta menghitung nilai rata-rata untuk setiap label [7]. CatBoost digunakan untuk membangun model klasifikasi yang mempelajari pola dalam data pasien dan memberikan prediksi apakah pasien tersebut berisiko menderita diabetes. Dengan pendekatan ini, diharapkan hasil penelitian dapat memberikan kontribusi nyata dalam pengembangan sistem prediksi penyakit berbasis *machine learning* yang lebih akurat dan efisien.

2. METODE

2.1 Data

Dataset yang digunakan dalam penelitian ini merupakan data mengenai kondisi pasien yang berkaitan dengan diabetes. Dataset ini berisi sejumlah variabel yang menggambarkan kondisi fisiologis atau biologis pasien serta status diabetes mereka. Tujuan dataset ini digunakan untuk membangun dan menguji model klasifikasi guna memprediksi kemungkinan seseorang menderita diabetes berdasarkan faktor-faktor fisiologis atau biologis dan medis. Pemrosesan data akan meliputi penanganan nilai nol, normalisasi, dan pemilihan fitur relevan.

Tabel 1. Data Atribute

No.	Nama Atribut	Deskripsi
1	Pregnancies	Jumlah kehamilan pasien
2	Glucose	Konsentrasi glukosa dalam plasma (mg/dL)
3	BloodPressure	Tekanan darah diastolik (mm Hg)
4	SkinThickness	Ketebalan lipatan kulit triceps (mm)
5	Insulin	Kadar insulin serum 2 jam (mu U/ml)
6	BMI	Indeks massa tubuh (berat badan dalam kg dibagi kuadrat tinggi badan m^2
7	DiabetesPedigreeFunction	Fungsi silsilah diabetes (indikator risiko genetic diabetes)
8	Age	Usia pasien (dalam tahun)
9	Outcome	Status diabetes (0 = tidak diabetes, 1 = diabetes)

Semua atribut dalam dataset ini bertipe numerik dan atribut outcome bertipe kategorikal biner.

2.2. XGBoost

Extreme Gradient Boosting (XGBoost) adalah algoritma berbasis gradient boosting decision tree (GBDT) yang dikembangkan untuk meningkatkan efisiensi komputasi dan akurasi prediksi[10]. Algoritma ini dikembangkan oleh Tianqi Chen dan Carlos Guestrin dan dirancang untuk mengatasi berbagai kelemahan dari metode boosting tradisional, seperti *overfitting* dengan kecepatan pemrosesan yang lambat pada dataset berskala besar[10]. XGBoost bekerja dengan menambahkan pohon keputusan secara iteratif, di mana setiap pohon baru bertugas meminimalkan kesalahan dari pohon-pohon sebelumnya menggunakan pendekatan gradient descent. Selain itu, XGBoost memiliki fungsi regularisasi (baik L1 maupun L2) yang membantu mengontrol kompleksitas model dan mencegah *overfitting*[10].

Arsitektur XGBoost terdiri dari beberapa komponen utama:

1. **Fungsi Obektif dan Optimasi**

   Untuk klasifikasi biner, fungsi obektif biasanya adalah log loss:

$$\mathcal{L}(\emptyset) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega\left(f_k\right) \quad (1)$$

   Dengan:

   - $l\left(y_i, \hat{y}_i^{(t)}\right)$: Fungsi loss antara label aktual $y_i$dan prediksi $\hat{y}_i$.
   - $\Omega(f_k)$: Fungsi regularisasi dari pohon ke-k,yang didefinisikan sebagai:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (2)$$

Di mana:
T: jumlah daun dalam pohon
w_j:bobot prediksi pada daun ke-j

Untuk meminimalkan loss ini, CatBoost menggunakan gradient boosting seperti GBDT, namun dengan pendekatan ordered boosting yang membedakan aliran data test menjadi dua bagian secara acak untuk menghindari *prediction shift*.

2. Perhitungan Prediksi

Prediksi total pada iterasi ke-*t* dihitung dengan menjumlahkan kontribusi semua pohon sebelumnya:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (1)$$

Gradient dan hessian pada setiap iterasi dihitung sebagai:

$$g_i = \frac{\partial l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i}, \; h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^2} \quad (2)$$

Kemudian, gain dari suatu pemisahan *(split)* dalam pohon dihitung menggunakan rumus:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \quad (3)$$

Di mana:
G_L, H_L: jumlah gradient dan hessian untuk node kiri
G_R, H_R: jumlah gradient dan hessian untuk node kanan,
γ: penalty kompleksitas
Split terbaik dipilih berdasarkan gain tertinggi

2.3. CatBoost

CatBoost adalah gradient-boosting yang dirancang khusus untuk mengatasi tantangan dalam menangani data kategorikal dalam pembelajaran mesin. CatBoost, singkatan dari "Categorical Boosting," dirancang untuk menangani data kategorikal dengan lebih efisien tanpa memerlukan pra-pemrosesan yang ekstensif seperti one-hot encoding, yang sering digunakan dalam algoritma lain [8]. CatBoost termasuk dalam keluarga algoritma gradient boosting, tetapi dengan beberapa perbaikan khusus yang membuatnya lebih stabil, lebih cepat, dan mampu menangani fitur kategorikal secara native [9]. CatBoost bekerja dengan membangun model berbasis pohon keputusan (decision tree) secara bertahap, di mana setiap pohon baru berusaha memperbaiki kesalahan dari pohon-pohon sebelumnya. Algoritma ini menggunakan pendekatan boosting, di mana model secara bertahap dibangun dengan memperbaiki kesalahan dari model sebelumnya, sehingga meningkatkan akurasi secara progresif, khususnya dalam menghadapi data kategori dan menangani *overfitting*.

Arsitektur CatBoost terdiri dari beberapa komponen utama:

1. Fungsi Obektif dan Optimasi
   Untuk klasifikasi biner, fungsi obektif biasanya adalah log loss:

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

Untuk meminimalkan loss ini, CatBoost menggunakan *gradient boosting* seperti GBDT, namun dengan pendekatan ordered boosting yang membedakan aliran data test menjadi dua bagian secara acak untuk menghindari *prediction shift.*

2. Perhitungan dan Penanganan Data Kategorikal
   Berbeda dari algoritma lain yang melakukan encoding di awal, CatBoost mengubah data kategorikal secara internal menggunakan:

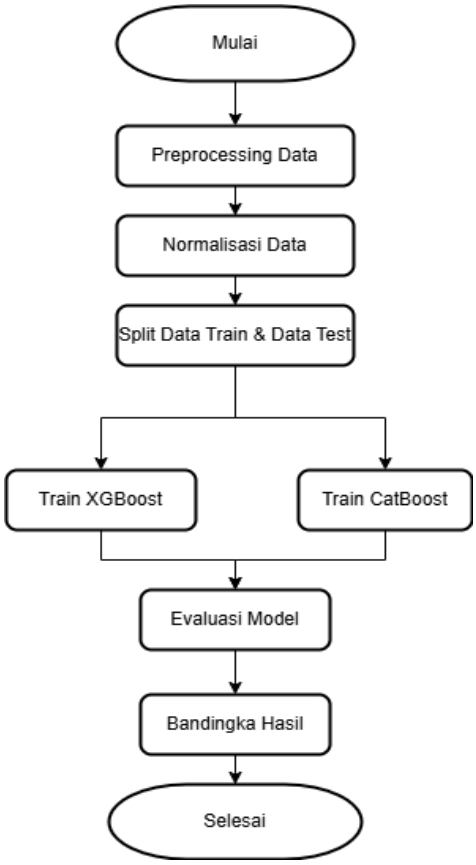
$$\text{MeanTarget}(x) = \frac{\sum_{i \in Perm(x)} y_i + a.p}{n_x + a} \quad (2)$$

Dengan:

- y_i: label target
- n_x: jumlah observasi sebelumnya dengan kategori x
- a: parameter smoothing
- p: rata-rata label secara global
- Perm(x): urutan permutasi acak dari data

Perhitungan ini dilakukan berdasarkan permutasi acak dari data, bukan semua data sekaligus, untuk menghindari informasi target bocor saat testing

2.4. Flowchart



Gambar 1. Flowhchart

Gambar di atas menunjukkan alur proses penelitian klasifikasi risiko diabetes menggunakan algoritma XGBoost dan CatBoost. Adapun tahapan-tahapan yang dilakukan dijelaskan sebagai berikut:

a. Mulai
   Tahap awal dimulai dengan perumusan masalah dan penentuan tujuan penelitian, yaitu membandingkan performa dua algoritma boosting dalam klasifikasi risiko diabetes
b. Ambil Dataset
   Data yang digunakan merupakan dataset diabetes yang bersumber dari platform *Kaggle*. Dataset ini berisi sejumlah variabel klinis pasien yang menjadi dasar dalam proses prediksi
c. Preprocessing Data
   Tahap ini bertujuan untuk membersihkan dan mempersiapkan data agar layak digunakan dalam pelatihan model. Beberapa langkah di antaranya adalah penghapusan fitur yang tidak relevan, penanganan nilai kosong atau tidak valid, serta transformasi awal data
d. Normalisasi Data
   Untuk menyetarakan skala antar fitur numerik, dilakukan normalisasi data menggunakan metode *Min-Max Scaling*. Hal ini penting agar setiap fitur memiliki kontribusi yang seimbang dalam proses pembelajaran model
e. Split Data Train dan Data Test
   Dataset dibagi menjadi dua subset, yaitu data latih dan data uji, guna untuk menghindari *overfitting* dan memungkinkan evaluasi model terhadao data yang belum pernah dilihat sebelumnya
f. Latih Model XGBoost dan CatBoost
   Kedua model pembelajaran dilatih secara terpisah:
   - XGBoost: dipilih karena kecepatan dan efisiensinya dalam menangani data numerik besar, serta memiliki mekanisme regularisasi
   - CatBoost: digunakan karena kemampuannya menangani data kategorikal secara langsung tanpa encoding tambahan, serta tahan terhadap *overfitting*
g. Bandingkan hasil
   Setelah kedua model dilatih, dilakukan evaluasi terhadap masing-masing model menggunakan metrik klasifikasi seperti akurasi, precision, recall, dan F1-score. Perbandingan ini bertujuan untuk mengetahui algoritma mana yang lebih unggul dalam konteks dataset yang digunakan
h. Selesai
   Tahap akhir ini adalah penyimpulan hasil, dengan mempertimbangkan kelebihan dan kekurangan masing-masing model. Hasil penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem deteksi penyakit diabetes berbasis *machine learning.*

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan *Diabetes Dataset* yang tersedia secara publik dan dapat diakses melalui platform *online* Kaggle. Dataset ini merupakan kumpulan data medis yang sering digunakan dalam penelitian klasifikasi dan deteksi penyakit diabetes. Data ini berisi informasi rekam medis pasien yang telah dilabeli berdasarkan kondisi diabetes mereka, sehingga sangat cocok untuk diterapkan dalam model pembelajaran mesin yang memerlukan data terstruktur dan beranotasi. Dengan menggunakan dataset ini, penelitian bertujuan untuk mengembangkan model prediktif yang akurat dan dapat diandalkan dalam konteks deteksi dini penyakit diabetes.

Tabel 2. Data Diabetes

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcame
2	138	62	35	0	33.6	0.127	47	1
0	84	82	31	125	38.2	0.233	23	0
0	145	0	0	0	44.2	0.630	31	1
0	135	68	42	250	42.3	0.365	24	1
1	139	62	41	480	40.7	0.536	21	0
0	173	78	32	265	46.5	1.159	58	0
4	99	72	17	0	25.6	0.294	28	0
8	194	80	0	0	26.1	0.551	67	0
2	83	65	28	66	36.8	0.629	24	0
2	89	90	30	0	33.5	0.292	42	0

Pada Tabel 2. ditampilkan sebagian data atribut yang digunakan dalam penelitian untuk mendeteksi diabetes menggunakan model pembelajaran mesin. Atribut-atribut tersebut meliputi jumlah kehamilan (*Pregnancies*), kadar glukosa (*Glucose*), tekanan darah (*Blood Pressure*), ketebalan kulit (*Skin Thickness*), kadar insulin (*Insulin*), indeks massa tubuh (*BMI*), fungsi silsilah diabetes (*Diabetes Pedigree Function*), usia (*Age*), serta hasil diagnosis (*Outcome*). Seluruh atribut ini merepresentasikan faktor-faktor medis yang relevan dalam mendeteksi penyakit diabetes. Data tersebut akan dianalisis menggunakan algoritma *Extreme Gradient Boosting* (XGBoost) dan *Categorical Boosting* (CatBoost), yang merupakan algoritma *ensemble learning* berbasis *gradient boosting* yang terkenal dengan akurasi tinggi dan kemampuan menangani data numerik maupun kategorikal secara efisien. Model ini diharapkan mampu memberikan performa yang optimal dalam mengklasifikasikan individu sebagai penderita diabetes atau tidak berdasarkan atribut yang tersedia.

3.1. Data Preprocessing

Pada tahap selanjutnya, dilakukan proses preprocessing data untuk memastikan data siap digunakan oleh model. Tahapan ini dimulai dengan data cleansing untuk memeriksa dan menangani nilai yang hilang (*missing values*) agar tidak memengaruhi kinerja model. Langkah ini penting agar data bersih dan model dapat bekerja secara optimal. Data yang sudah dibersihkan dapat dilihat pada tabel 3.

Tabel 3. Data Cleansing

No.	Glucose	Blood Pressure	BMI	Age	Outcame
0	138.0	62.0000	33.6	47	1
1	84.0	82.0000	38.2	23	0
2	145.0	69.1455	44.2	31	1
3	135.0	68.0000	42.3	24	1
4	139.0	62.0000	40.7	21	0
...	...	...	...	...	...
1995	75.0	64.0000	29.7	33	0
1996	179.0	72.0000	32.7	36	1
1997	85.0	78.0000	31.2	42	0
1998	129.0	110.0000	67.1	26	1
1999	81.0	72.0000	30.1	25	0

Tabel 3 menunjukkan data sebelum dilakukan proses normalisasi. Beberapa fitur seperti *Glucose*, *BloodPressure*, *BMI*, dan *Age* memiliki rentang nilai yang berbeda-beda, yang dapat mempengaruhi kinerja model *machine learning* jika tidak distandarisasi. Misalnya, nilai *Glucose* berkisar antara 75 hingga 179, sedangkan nilai *Age* berkisar dari 21 hingga 47 pada data yang ditampilkan.

3.2. Normalisasi Data

Untuk mengatasi ketimpangan skala antar fitur tersebut, dilakukan proses normalisasi data. Hasil dari proses ini ditampilkan pada Tabel 4.

Tabel 4. Normalisasi Data

No.	Glucose	Blood Pressure	BMI	Age	Outcame
0	0.606452	0.387755	0.246795	0.433333	1
1	0.258065	0.591837	0.320513	0.033333	0
2	0.651613	0.460668	0.416667	0.166667	1
3	0.587097	0.448980	0.386218	0.050000	1
4	0.612903	0.387755	0.360577	0.000000	0
...	...	...	...	...	...
1995	0.200000	0.408163	0.184295	0.200000	0
1996	0.870968	0.489796	0.232372	0.250000	1
1997	0.264516	0.551020	0.208333	0.350000	0
1998	0.548387	0.877551	0.783654	0.083333	1
1999	0.238710	0.489796	0.190705	0.066667	0

Dalam tabel tersebut, setiap fitur telah ditransformasikan ke dalam rentang nilai antara 0 dan 1 menggunakan metode *Min-Max Scaling*. Tujuan normalisasi ini adalah agar semua fitur memiliki kontribusi yang seimbang saat digunakan dalam pelatihan model, sehingga dapat meningkatkan akurasi dan stabilitas model prediksi.

3.3. Data Training

Dari hasil pembagian data, terdapat 1.600 data yang digunakan untuk proses pelatihan (training) dan 400 data lainnya digunakan untuk pengujian (testing). Dapat dilihat dari tabel 5.

Tabel 5. Data Training

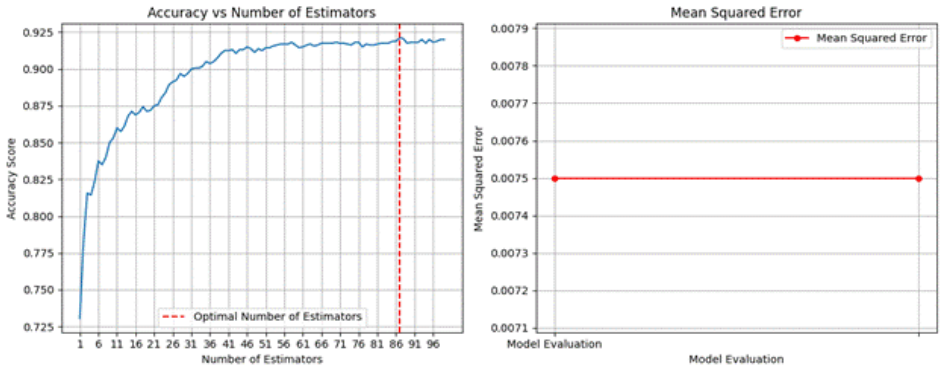
Data Training	1.600
Data Testing	400

Pembagian ini dilakukan agar model dapat belajar terlebih dahulu dari data training, lalu diuji kemampuannya menggunakan data testing. Dengan cara ini, kita bisa mengetahui seberapa baik model memprediksi data yang belum pernah dilihat sebelumnya, sehingga hasilnya bisa lebih akurat dan tidak hanya bagus pada data latih saja.

3.4. Pemodelan XGboost

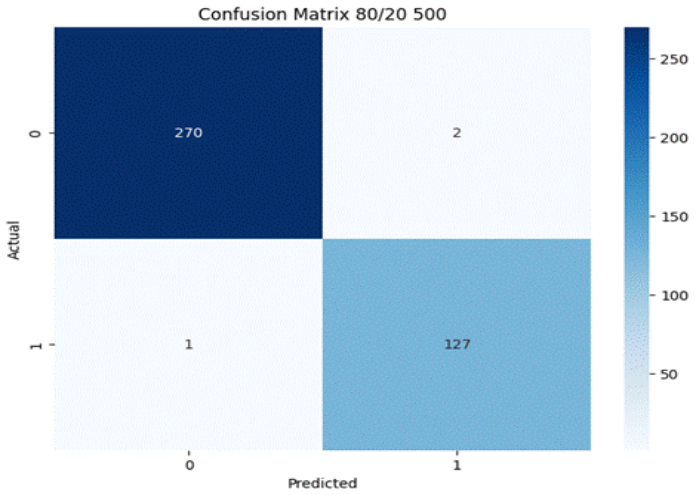
Pada penelitian ini, digunakan algoritma XGBoost Classifier sebagai model pembelajaran mesin untuk menyelesaikan permasalahan klasifikasi. Model diinisialisasi dengan parameter `n_estimators=100`, `learning_rate=0.01`, dan `max_depth=3`, yang kemudian divalidasi menggunakan teknik *k-fold cross-validation* sebanyak 3 lipatan dengan metode *StratifiedKFold* untuk menjaga proporsi distribusi kelas pada setiap lipatan data. Hasil *cross-validation* menunjukkan nilai akurasi sebesar 0.7752, 0.7505, dan 0.7729, dengan rata-rata akurasi sebesar 76.62%, yang menggambarkan stabilitas model pada data pelatihan.

Selanjutnya, dilakukan proses tuning terhadap jumlah estimators untuk menentukan parameter optimal dalam meningkatkan performa model. Berdasarkan grafik hubungan antara jumlah estimators dan akurasi, terlihat bahwa akurasi model meningkat secara signifikan pada jumlah estimators rendah, kemudian mencapai titik stabil mendekati 92% pada kisaran 80 estimators. Optimalisasi model menunjukkan bahwa jumlah estimators terbaik adalah 86, ditandai dengan nilai *Mean Squared Error* (MSE) sebesar 0.0075 dan akurasi prediksi pada data uji sebesar 92.13%, yang menunjukkan bahwa model mampu melakukan generalisasi dengan sangat baik terhadap data yang belum pernah dilihat sebelumnya.



Gambar 2. Akurasi dan MSE Model XGboost

Evaluasi akhir dilakukan menggunakan *confusion matrix* yang memberikan gambaran rinci mengenai performa klasifikasi model. Dari total 400 data uji, model berhasil mengklasifikasikan 270 data kelas 0 dan 127 data kelas 1 dengan benar, sementara hanya terjadi 3 kesalahan klasifikasi.



Gambar 3. Confusion Matrix Model XGboost

Hasil matrik evaluasi menunjukkan bahwa model memiliki akurasi sebesar 99%, *precision* sebesar 1.00 untuk kelas 0 dan 0.98 untuk kelas 1, serta *recall* sebesar 0.99 untuk kedua kelas. Selain itu, nilai *macro average* dan *weighted average* untuk precision, recall, dan F1-score adalah 0.99, menandakan bahwa model tidak hanya akurat secara keseluruhan, tetapi juga adil dan seimbang dalam memprediksi masing-masing kelas.

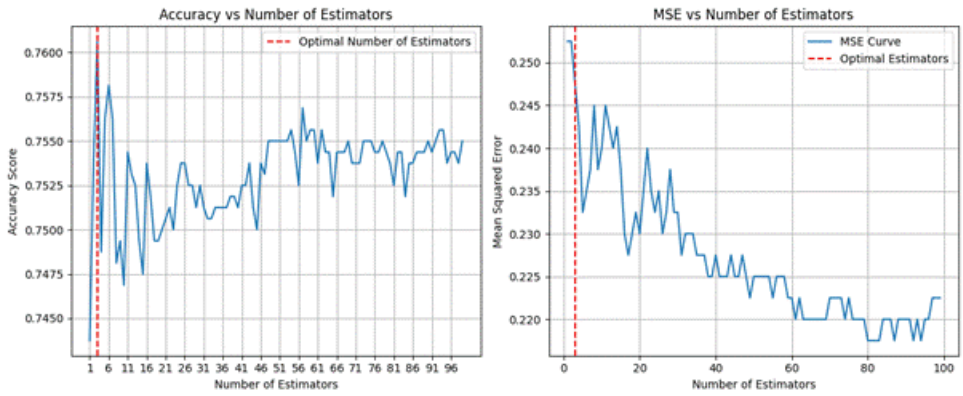
Tabel 6. Evaluasi Confusion Matrix Model XGboosr

Class	Precision	Recall	F1-Score	Support
0	1.00	0.99	0.99	272
1	0.98	0.99	0.99	128
Accuracy			**0.99**	**400**
Macro Avg	0.99	0.99	0.99	400
Weighted Avg	0.99	0.99	0.99	400

Secara keseluruhan, hasil ini menunjukkan bahwa algoritma XGBoost sangat efektif dalam menangani permasalahan klasifikasi pada dataset yang digunakan.

3.5. Pemodelan Catboost

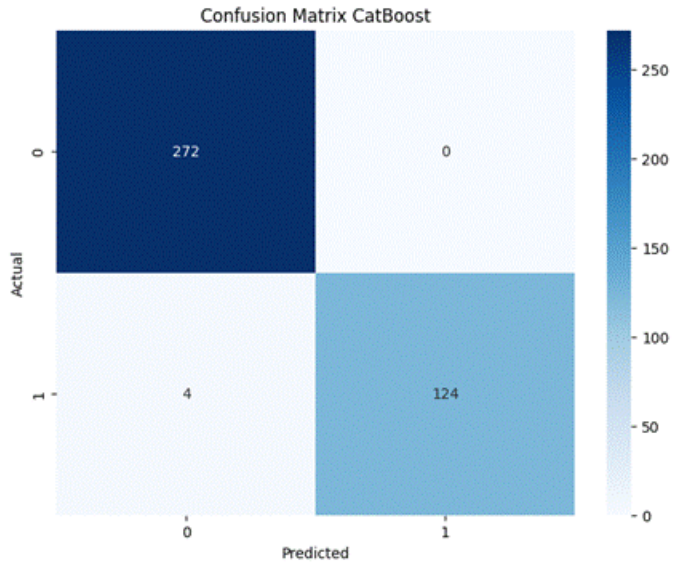
Dalam penelitian ini, algoritma CatBoost Classifier digunakan untuk melakukan klasifikasi pada dataset yang tersedia. CatBoost merupakan algoritma gradient boosting yang efektif menangani data kategorikal dan mampu memberikan performa yang tinggi tanpa banyak prapemrosesan. Proses pelatihan model melibatkan pencarian jumlah estimators yang optimal berdasarkan grafik akurasi terhadap jumlah estimators dan *Mean Squared Error* (MSE). Dari grafik tersebut, terlihat bahwa estimasi optimal tercapai pada titik awal (sekitar estimators ke-6), dengan nilai akurasi pada data uji sebesar 76.06% dan nilai MSE sebesar 0.2475.



Gambar 4. Akurasi dan MSE Model Catboost

Meskipun nilai akurasi pada data uji lebih rendah dibandingkan model sebelumnya (XGBoost), model CatBoost menunjukkan ketahanan dan kestabilan yang cukup baik terhadap fluktuasi jumlah estimators, sebagaimana ditunjukkan oleh pola MSE yang konsisten menurun.

Evaluasi lanjutan dilakukan menggunakan *confusion matrix*, yang menunjukkan performa klasifikasi model secara rinci. Dari total 400 data uji, model berhasil memprediksi 272 data kelas 0 dan 124 data kelas 1 dengan benar, serta hanya melakukan 4 kesalahan klasifikasi pada kelas 1.



Gambar 5. Confusion Matrix Model Catboost

Nilai akurasi model mencapai 99%, dengan nilai precision, recall, dan f1-score masing-masing sebesar 0.99 untuk kelas 0 dan 0.98 untuk kelas 1. Nilai *macro average* dan *weighted average* untuk seluruh metrik evaluasi juga menunjukkan skor tinggi, yaitu 0.99, yang menandakan bahwa model memiliki kemampuan klasifikasi yang sangat baik secara menyeluruh dan seimbang antar kelas.

Tabel 7. Evaluasi Confusion Matrix Model Catboost

Class	Precision	Recall	F1-Score	Support
0	0.99	1.00	0.99	272
1	1.00	0.97	0.98	128
Accuracy			**0.99**	**400**
Macro Avg	0.99	0.98	0.99	400
Weighted Avg	0.99	0.99	0.99	400

Hasil ini mengindikasikan bahwa meskipun akurasi pada data uji tidak setinggi model XGBoost, CatBoost tetap memberikan performa klasifikasi yang sangat akurat, terutama dalam konteks data yang mungkin mengandung fitur kategorikal atau membutuhkan pendekatan boosting yang lebih robust terhadap noise dan ketidakseimbangan data.

4. KESIMPULAN

Penelitian ini berhasil membandingkan performa dua algoritma gradient boosting, yaitu XGBoost dan CatBoost, dalam melakukan klasifikasi risiko diabetes berdasarkan data medis dari *Diabetes Dataset*. Proses penelitian mencakup tahapan penting seperti *preprocessing data*, normalisasi menggunakan *Min-Max Scaling*, pembagian data menjadi *training* dan *testing*, hingga pelatihan dan evaluasi model dengan metrik klasifikasi.

Hasil evaluasi menunjukkan bahwa kedua algoritma memiliki performa yang sangat baik, dengan nilai akurasi masing-masing sebesar 99% pada data pengujian. Algoritma XGBoost menunjukkan performa terbaik pada estimators ke-86, menghasilkan akurasi 92.13% dengan nilai MSE 0.0075 pada saat *tuning*, serta *precision, recall, dan f1-score* yang hampir sempurna. Sedangkan CatBoost, meskipun akurasi awalnya sedikit lebih rendah pada saat *tuning* (76.06% dengan MSE 0.2475), tetap memberikan hasil akhir yang seimbang dan kuat, terutama dalam menangani fitur kategorikal tanpa banyak preprocessing tambahan.

Dari *confusion matrix*, kedua model berhasil memprediksi sebagian besar data dengan benar. XGBoost menghasilkan *precision* 1.00 untuk kelas 0 dan 0.98 untuk kelas 1, dengan *recall* 0.99 untuk kedua kelas. Sementara itu, CatBoost menunjukkan *precision* 0.99 untuk kelas 0 dan 1.00 untuk kelas 1, dengan *recall* 0.99 dan 0.97.

Secara keseluruhan, kedua algoritma mampu melakukan klasifikasi dengan sangat baik dan akurat. Namun, jika mempertimbangkan kestabilan prediksi dan efisiensi pada data numerik, XGBoost lebih unggul dalam konteks dataset ini. Di sisi lain, CatBoost tetap menjadi alternatif kuat, khususnya jika menghadapi data yang mengandung banyak fitur kategorikal. Hasil penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem prediksi penyakit diabetes berbasis machine learning yang lebih akurat dan efisien.

Ucapan Terima Kasih

Kami, penulis, menyampaikan terima kasih kepada Universitas Pembangunan Nasional "Veteran" Jawa Timur, khususnya Program Studi Sains Data, atas dukungan fasilitas dan kesempatan yang diberikan sehingga penelitian ini dapat terlaksana dengan baik. Kami juga mengapresiasi semangat kerja sama tim selama proses penyusunan jurnal ini. Ucapan terima kasih turut kami sampaikan kepada semua pihak yang telah membantu, baik secara langsung maupun tidak langsung, dalam penyelesaian artikel ini.

Daftar Pustaka

Depari, Sevilla. (2024). Diabetes: Pengenalan Dasar tentang Penyakit Kronis untuk Meningkatkan Kesadaran. 10.13140/RG.2.2.27175.97440.

Sasmita, Anggi Marta Dwi. "Faktor-faktor yang mempengaruhi tingkat kepatuhan berobat pasien diabetes melitus." Jurnal medika hutama 2.04 Juli (2021): 1105-1111.

Ligita, Titan, et al. "How people living with diabetes in Indonesia learn about their disease: A grounded theory study." PloS one 14.2 (2019): e0212019.

Yunus, Haikal. Buatan Kecerdasan Kesehatan & Perawatan. (2023). Masa Depan Teknologi Medis: Inovasi Terkini dalam Perawatan Kesehatan Pendahuluan.

Salsabil, Muhammad, Nuril Lutvi Azizah, and Ade Eviyanti. "Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost." Jurnal Ilmiah Komputasi 23.1 (2024): 51-58.

Jasman, Taufik Zulhaq, et al. "Analisis Algoritma Gradient Boosting, Adaboost dan Catboost dalam Klasifikasi Kualitas Air." Jurnal Teknik Informatika dan Sistem Informasi 8.2 (2022): 392-402.

Kumar, P. Suresh, et al. "CatBoost ensemble approach for diabetes risk prediction at early stages." 2021 1st Odisha international conference on electrical power engineering, communication and computing technology (ODICON). IEEE, 2021.

Istianto, Andrian Febriansyah, Asep Id Hadiana, and Fajri Rakhmat Umbara. "Prediksi curah hujan menggunakan metode categorical boosting (Catboost)." JATI (Jurnal Mahasiswa Teknik Informatika) 7.4 (2023): 2930-2937.

Ptr, Agus Fahmi Limas, Muhammad Mizan Siregar, and Irwan Daniel. "Analysis of Gradient Boosting, XGBoost, and CatBoost on Mobile Phone Classification." J. Comput. Networks, Archit. High Perform. Comput 6.2 (2024): 661-670.

L. V. Fausett, Fundamentals of Neural Networks: Architectures, Algorithms and Application, Pearson Education India, 2006.

C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," arXiv preprint arXiv:1811.03378, 2018.

L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, pp. 6638–6648, 2018.

A. Muhaimin, D. D. Prastyo and H. Horng-Shing Lu, "Forecasting with Recurrent Neural Network in Intermittent Demand Data," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 802-809, doi: 10.1109/Confluence51648.2021.9376880.