



Optimasi Algoritma Decision Tree dalam Memprediksi Penyakit Diabetes Menggunakan GridSearchCV

Edi Cahya Pratama

Program Studi Sistem Informasi, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia

E-Mail : edicahyapratama97@gmail.com

Article Info

Article history:

Received Dec 03, 2024

Revised Dec 17, 2024

Accepted Dec 25, 2024

Kata Kunci:

DecisionTree,
Prediksi Diabetes,
GridSearchCV,
SMOTE,
Pembelajaran Mesin

Keywords:

Decision Tree,
Diabetes Prediction,
GridSearchCV,
SMOTE,
Machine Learning

ABSTRAK

Penelitian ini bertujuan untuk mengoptimalkan algoritma Decision Tree dalam memprediksi penyakit diabetes melalui penerapan hyperparameter tuning menggunakan GridSearchCV serta penyeimbangan data dengan metode SMOTE (Synthetic Minority Over-sampling Technique). Dataset yang digunakan adalah Pima Indians Diabetes Dataset, yang terdiri dari delapan fitur medis dan satu label target. Model awal yang dibangun dengan parameter default menghasilkan akurasi sebesar 95,42%, namun memiliki kelemahan dalam mendeteksi pasien yang benar-benar menderita diabetes. Untuk mengatasi hal tersebut, dilakukan tuning terhadap parameter `max_depth`, `min_samples_leaf`, dan `criterion`, dan diperoleh konfigurasi terbaik: `max_depth = 5`, `min_samples_leaf = 2`, dan `criterion = 'entropy'`. Model hasil optimasi menunjukkan peningkatan performa signifikan, dengan akurasi meningkat menjadi 97,06%, serta nilai `precision`, `recall`, dan `F1-score` sebesar 0,97 untuk kelas positif. Hasil ini menunjukkan bahwa kombinasi antara optimasi parameter dan penyeimbangan data secara sistematis mampu meningkatkan akurasi dan keandalan model prediksi untuk diagnosis dini penyakit diabetes.

ABSTRACT

This research aims to optimize the Decision Tree algorithm for predicting diabetes by applying hyperparameter tuning using GridSearchCV and addressing data imbalance through the SMOTE (Synthetic Minority Over-sampling Technique) method. The dataset used is the Pima Indians Diabetes Dataset, consisting of eight medical features and one target label. The initial model built with default parameters achieved an accuracy of 95.42%, but struggled to detect actual diabetic patients. To address this, tuning was performed on `max_depth`, `min_samples_leaf`, and `criterion`, yielding the best configuration: `max_depth = 5`, `min_samples_leaf = 2`, and `criterion = 'entropy'`. The optimized model demonstrated significant performance improvement, increasing accuracy to 97.06%, with `precision`, `recall`, and `F1-score` reaching 0.97 for the positive class. These results confirm that systematic parameter optimization and class balancing can greatly enhance the accuracy and reliability of predictive models for early diabetes diagnosis.

This is an open access article under the [CC BY-NC](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Edi Cahya Pratama,

Program Studi Sistem Informasi, STIKOM Tunas Bangsa Pematangsiantar,

Jalan Jendral Sudirman Blok A, No. 1/2/3, Siantar Barat, Kota Pematangsiantar, Sumatera Utara, 21127, Indonesia

Email: edicahyapratama97@gmail.com

1. PENDAHULUAN

Penerapan teknologi kecerdasan buatan, khususnya algoritma pembelajaran mesin (*machine learning*), telah menjadi fokus utama dalam pengembangan sistem pendukung keputusan di bidang kesehatan (Zharif et al., 2025). Salah satu tantangan utama dalam dunia medis adalah mendeteksi penyakit kronis seperti diabetes secara dini dan akurat. Diabetes melitus merupakan penyakit metabolik kronis yang ditandai oleh tingginya kadar glukosa darah dan menjadi salah satu penyebab kematian tertinggi di dunia (Febrian et al.,

2024). Menurut WHO, deteksi dini dan penanganan yang tepat dapat menurunkan risiko komplikasi secara signifikan (WHO, 2021).

Algoritma *Decision Tree* merupakan salah satu metode klasifikasi yang populer karena kemampuannya dalam menghasilkan model yang mudah dipahami dan diinterpretasikan (Lile & Suharjo, 2024). Selain itu, algoritma ini juga tidak memerlukan asumsi distribusi data dan dapat menangani fitur numerik maupun kategorikal secara langsung (Amir et al., 2023). Penggunaan *Decision Tree* telah diterapkan dalam berbagai studi prediksi, seperti deteksi kanker payudara (Irfan et al., 2023), prediksi gagal jantung (Nugraha & Setiawan, 2022), serta klasifikasi risiko kehamilan (Ramadhan et al., 2023). Dalam konteks prediksi diabetes, algoritma ini menawarkan fleksibilitas dalam membangun model klasifikasi berdasarkan data klinis pasien.

Namun, kinerja dari *Decision Tree* sangat dipengaruhi oleh parameter-parameter tertentu, seperti kedalaman pohon (*max_depth*), jumlah minimum sampel untuk pemisahan (*min_samples_split*), dan kriteria pemisahan (*criterion*). Pengaturan parameter yang tidak optimal dapat menyebabkan model mengalami *overfitting* atau *underfitting*, yang berdampak pada menurunnya akurasi prediksi (Sari et al., 2022). Oleh karena itu, diperlukan pendekatan sistematis dalam memilih parameter terbaik guna meningkatkan performa model. Salah satu metode yang efektif untuk mengoptimalkan hyperparameter adalah *GridSearchCV* (Darmawan & Dianta, 2023). Metode ini melakukan eksplorasi menyeluruh terhadap kombinasi parameter dengan memanfaatkan validasi silang untuk menghindari bias dan *overfitting* (Yaqin et al., 2025). Beberapa penelitian menunjukkan bahwa penerapan *GridSearchCV* mampu meningkatkan kinerja model secara signifikan, baik pada algoritma *Random Forest* (Putra & Dewi, 2021), *Support Vector Machine* (Azzahra et al., 2023), maupun *Decision Tree* dalam konteks prediksi pendidikan dan keuangan (Wulandari & Hidayat, 2023).

Meskipun demikian, penerapan *GridSearchCV* dalam optimasi *Decision Tree* untuk prediksi penyakit diabetes masih jarang dilakukan, terutama yang disertai penyeimbangan data menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) (Trianto et al., 2022). Padahal, data medis seperti diagnosis diabetes sering kali tidak seimbang, di mana jumlah pasien non-diabetes jauh lebih banyak dibandingkan pasien yang positif diabetes (Sidiq et al., 2025).

Dengan latar belakang tersebut, penelitian ini bertujuan untuk mengoptimalkan algoritma *Decision Tree* dalam memprediksi penyakit diabetes menggunakan *GridSearchCV*, serta meningkatkan akurasi model melalui teknik oversampling SMOTE. Penelitian ini diharapkan dapat memberikan kontribusi praktis dalam pengembangan sistem pendukung diagnosis dini, serta menjadi acuan ilmiah dalam penerapan pembelajaran mesin pada bidang medis.

2. METODE PENELITIAN

2.1. Dataset Penelitian

Penelitian ini menggunakan *Pima Indians Diabetes Dataset* yang tersedia secara publik melalui platform Kaggle (Aswad, 2022). Dataset ini berisi 768 data pasien wanita keturunan Pima Indian yang berusia di atas 21 tahun, dengan 8 atribut utama yang berkaitan dengan risiko diabetes, yaitu *Pregnancies* (jumlah kehamilan), *Glucose* (kadar glukosa plasma), *BloodPressure* (tekanan darah diastolik), *SkinThickness* (ketebalan lipatan kulit trisep), *Insulin* (kadar insulin dalam darah), *BMI* (indeks massa tubuh), *DiabetesPedigreeFunction* (riwayat genetik diabetes), *Age* (usia pasien). Atribut kesembilan adalah *Outcome* (0: tidak diabetes, 1: diabetes) sebagai label target. Seluruh fitur tersebut berperan penting dalam memprediksi kemungkinan seseorang mengidap diabetes. Data telah disusun dalam format terstruktur dan dimuat menggunakan perangkat lunak pengolah data untuk dilakukan analisis lebih lanjut. Gambar 2 menunjukkan lima baris pertama dari dataset yang digunakan.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Gambar 1. Beberapa Baris Dataset

2.2. Pra-pemrosesan Data

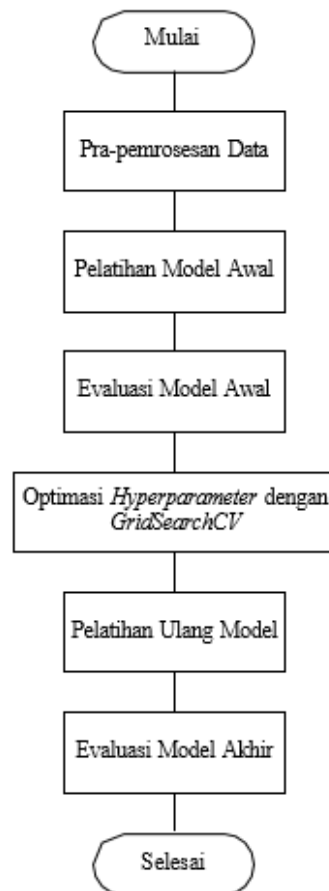
Pra-pemrosesan dilakukan untuk memastikan data dalam kondisi layak digunakan untuk pelatihan model. Langkah-langkah yang diterapkan meliputi imputasi nilai nol pada atribut, normalisasi data, pembagian data dan *oversampling*. Walaupun dataset tidak mengandung nilai kosong secara eksplisit, terdapat fitur-fitur medis seperti *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, dan *BMI* yang memiliki nilai nol (0). Secara medis, nilai nol pada fitur-fitur tersebut dianggap tidak mungkin atau tidak logis, karena tidak ada orang sehat dengan kadar glukosa atau tekanan darah nol. Untuk mengatasi hal ini, digunakan metode imputasi dengan median. Median dipilih karena lebih tahan terhadap outlier dibandingkan dengan rata-rata (Wulandari et al., 2023).

Agar perhitungan jarak dan pemisahan pohon tidak bias terhadap fitur dengan skala besar, dilakukan normalisasi menggunakan metode *Min-Max Scaling*. Skala setiap fitur dikonversi ke dalam rentang 0 hingga 1. Ini penting karena meskipun *Decision Tree* tidak terlalu sensitif terhadap skala, normalisasi membantu stabilitas saat menggunakan teknik *ensemble* lanjutan (Nugraha & Setiawan, 2022). Data dibagi menjadi dua bagian yaitu data latih (80%) dari total data dan data uji (20%) dari total data. Pembagian dilakukan secara *stratified*, yaitu menjaga proporsi kelas target (0 dan 1) tetap seimbang di kedua subset.

Distribusi kelas pada dataset ini tidak seimbang, dengan jumlah pasien non-diabetes lebih banyak dibandingkan pasien diabetes. Untuk menanganinya, digunakan teknik SMOTE (*Synthetic Minority Oversampling Technique*), yaitu teknik *oversampling* yang menghasilkan data sintesis berdasarkan tetangga terdekat dari kelas minoritas (Nguyen et al., 2023).

2.3. Flowchart Penelitian

Untuk memastikan bahwa setiap langkah ditangani secara sistematis dan untuk meningkatkan kinerja model, penelitian ini mengikuti alur kerja yang terstruktur. Gambar 1 menunjukkan proses yang digunakan dalam penelitian ini.



Gambar 2. Flowchart Penelitian

Arsitektur penelitian ini disusun secara sistematis agar setiap tahap dalam proses pengembangan model prediksi dapat berjalan secara terarah dan terukur. Penelitian diawali dengan tahap pengumpulan data,

yaitu mengambil dataset *Pima Indians Diabetes* dari sumber terbuka di platform *Kaggle*. Setelah data diperoleh, dilakukan pra-pemrosesan data yang mencakup pemeriksaan nilai tidak valid, imputasi data nol menggunakan median, normalisasi fitur numerik menggunakan teknik *Min-Max Scaling*, dan pembagian data menjadi data latih dan data uji dengan rasio 80:20 secara stratifikasi untuk menjaga keseimbangan distribusi kelas. Selanjutnya, diterapkan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) untuk menangani ketidakseimbangan kelas, khususnya karena jumlah pasien dengan label positif diabetes lebih sedikit dibandingkan label negatif.

Setelah data siap, tahap berikutnya adalah pelatihan model awal menggunakan algoritma *Decision Tree* dengan parameter default. Tujuan dari pelatihan awal ini adalah untuk memperoleh baseline performa model sebelum dilakukan optimasi. Evaluasi terhadap model awal dilakukan dengan menghitung metrik-metrik evaluasi klasifikasi seperti akurasi, precision, recall, F1-score, serta menyusun *confusion matrix* guna melihat distribusi kesalahan prediksi antar kelas.

Langkah selanjutnya adalah optimasi model menggunakan metode *GridSearchCV*, yaitu teknik tuning hiperparameter yang mengevaluasi kombinasi parameter terbaik melalui validasi silang (*cross-validation*). Parameter yang dioptimasi meliputi kedalaman pohon (*max_depth*), jumlah minimum sampel daun (*min_samples_leaf*), dan kriteria pemisahan (*criterion*). Setelah diperoleh kombinasi parameter terbaik, model dilatih ulang menggunakan konfigurasi tersebut dengan data latih yang telah diseimbangkan.

Tahap akhir dari arsitektur ini adalah melakukan evaluasi model akhir untuk melihat sejauh mana peningkatan performa model setelah dilakukan tuning. Hasil evaluasi model akhir kemudian dibandingkan dengan performa model awal untuk menilai efektivitas dari proses optimasi yang dilakukan. Seluruh rangkaian ini dirancang untuk menghasilkan model prediksi penyakit diabetes yang optimal, akurat, dan dapat diandalkan dalam konteks diagnosis dini di bidang kesehatan.

2.4. Algoritma *Decision Tree*

Algoritma *Decision Tree* merupakan metode klasifikasi berbasis struktur pohon di mana keputusan dibuat dari akar ke daun berdasarkan nilai fitur (Setio et al., 2020). Kelebihannya antara lain adalah mudah diinterpretasikan, tidak memerlukan normalisasi skala data (meskipun tetap dilakukan di sini untuk alasan stabilitas model), dapat menangani fitur kategorikal dan numerik (Al & Subhiyacto, 2025).

Namun, kekurangan dari algoritma ini adalah rentan terhadap *overfitting*, terutama jika tidak diatur parameter seperti kedalaman pohon (*max_depth*) atau ukuran minimum daun (*min_samples_leaf*). Oleh karena itu, diperlukan proses *tuning parameter* (Sari et al., 2022).

2.5. Tuning Hiperparameter dengan *GridSearchCV*

Untuk mengoptimalkan performa model, dilakukan *tuning hyperparameter* dengan menggunakan metode *GridSearchCV*, yang menguji semua kombinasi parameter dan mengevaluasinya menggunakan teknik *cross-validation* (CV). Parameter yang diuji *Max_depth* : [3, 5, 7, None] – mengontrol kedalaman maksimal dari pohon, *Min_samples_leaf* : [1, 2, 4] – jumlah minimum sampel pada node daun, dan *Criterion* : ['gini', 'entropy'] – fungsi yang digunakan untuk mengukur kualitas split. Teknik *5-fold cross-validation* digunakan untuk memvalidasi performa setiap kombinasi parameter secara adil (Geddfdforge & Sumathi, 2020).

2.6. Evaluasi Model

Evaluasi model dalam penelitian ini dilakukan untuk mengukur kinerja algoritma *Decision Tree* dalam memprediksi apakah seorang pasien menderita diabetes atau tidak berdasarkan delapan fitur medis yang tersedia. Proses evaluasi menggunakan beberapa metrik klasifikasi yang umum digunakan dalam pembelajaran mesin. Metrik pertama adalah akurasi, yaitu persentase jumlah prediksi yang benar terhadap keseluruhan data uji. Selanjutnya, digunakan precision, yaitu proporsi prediksi positif (penderita diabetes) yang benar-benar tepat, serta recall, yang mengukur seberapa baik model dalam menangkap semua kasus positif sebenarnya. Kedua metrik ini kemudian dirangkum ke dalam satu metrik gabungan bernama *F1-Score*, yaitu rata-rata harmonik antara precision dan recall yang memberikan gambaran kinerja model dalam situasi distribusi data yang tidak seimbang. Selain itu, evaluasi juga dilengkapi dengan *confusion matrix*, yaitu sebuah tabel dua dimensi yang menunjukkan distribusi hasil prediksi model dibandingkan dengan label aktual. Matriks ini memvisualisasikan jumlah kasus benar dan salah dalam kategori positif dan negatif.

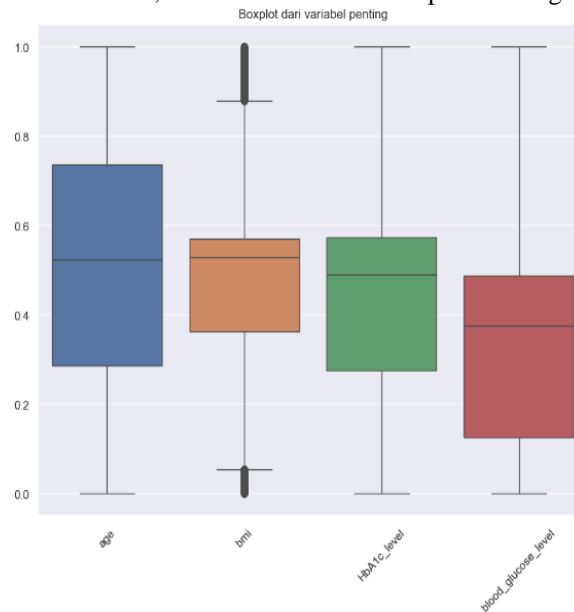
Evaluasi dilakukan sebanyak dua kali. Pertama, model dievaluasi dalam bentuk awalnya, yaitu menggunakan parameter default dari *Decision Tree Classifier*, untuk mendapatkan gambaran performa dasar tanpa optimasi. Kemudian, setelah dilakukan tuning hiperparameter menggunakan metode *GridSearchCV* dan model dilatih ulang dengan parameter terbaik, dilakukan evaluasi kedua. Tujuan dari evaluasi ini adalah untuk mengetahui sejauh mana peningkatan kinerja model setelah dilakukan proses optimasi parameter.

Dengan membandingkan hasil evaluasi sebelum dan sesudah tuning, efektivitas proses optimasi dapat diukur secara kuantitatif. Evaluasi ini sangat penting untuk memastikan bahwa model yang dibangun tidak hanya memiliki akurasi tinggi, tetapi juga mampu mendeteksi kasus positif dengan tepat, yang sangat krusial dalam konteks diagnosis penyakit seperti diabetes.

3. HASIL DAN PEMBAHASAN

3.1. Analisis Awal Data

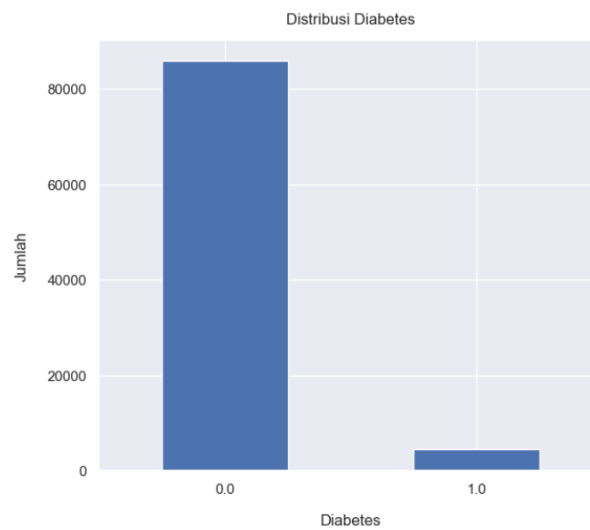
Tahapan awal dalam proses analisis dimulai dengan mengevaluasi karakteristik distribusi dari fitur-fitur penting dalam dataset yang berkaitan erat dengan diagnosis diabetes. Variabel yang dianalisis secara visual meliputi *age* (usia pasien), *BMI* (indeks massa tubuh), *HbA1c level* (kadar hemoglobin terglikasi), dan *bloodglucose level* (kadar glukosa darah). Visualisasi dalam bentuk boxplot pada gambar 3 digunakan untuk memahami rentang nilai, persebaran data, serta keberadaan outlier pada masing-masing fitur tersebut.



Gambar 2. Visualisasi *Outliers*

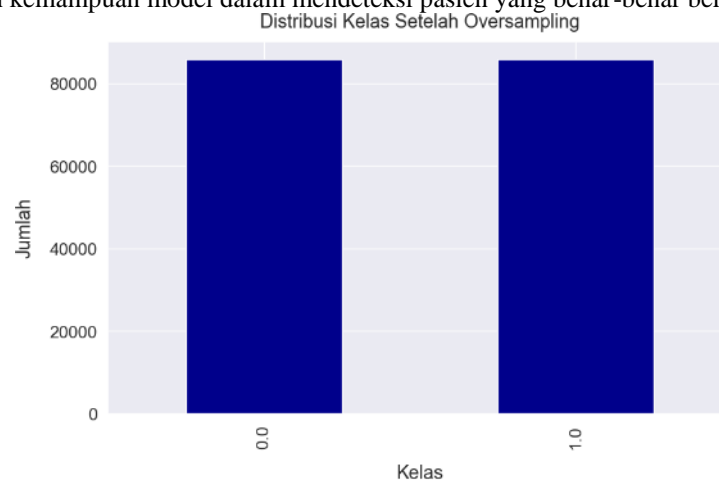
Berdasarkan boxplot tersebut, dapat diamati bahwa sebagian besar fitur memiliki distribusi data yang relatif simetris dan menyebar dalam rentang normal. Namun, pada fitur BMI, terdapat sejumlah *outlier* yang cukup mencolok. *Outlier* ini ditunjukkan oleh titik-titik data yang berada jauh di atas atau di bawah batas kuartil atas dan bawah. Keberadaan outlier ini bisa disebabkan oleh kondisi medis ekstrem atau pencatatan data yang tidak akurat. Meskipun demikian, karena data telah melalui tahap *normalisasi* menggunakan metode *Min-Max Scaling*, nilai-nilai ekstrem tersebut tidak memberikan dampak signifikan dalam proses pelatihan model. Normalisasi ini membuat seluruh fitur berada pada skala yang sama, sehingga mencegah satu fitur mendominasi perhitungan pada algoritma klasifikasi.

Langkah berikutnya adalah menganalisis distribusi kelas target (*label*), yaitu apakah pasien dalam dataset menderita diabetes (*label 1*) atau tidak (*label 0*). Hasil visualisasi distribusi label target dalam bentuk diagram batang pada gambar 4 menunjukkan bahwa dataset mengalami ketidakseimbangan kelas (*class imbalance*) yang cukup besar. Terlihat bahwa jumlah pasien dengan *label 0* (*negatif diabetes*) jauh lebih banyak dibandingkan pasien dengan *label 1* (*positif diabetes*). Ketimpangan ini berbahaya dalam konteks pemodelan pembelajaran mesin karena model cenderung akan lebih akurat dalam mengenali kelas mayoritas, namun gagal mengidentifikasi kasus pada kelas minoritas, dalam hal ini pasien yang benar-benar menderita diabetes. Hal tersebut dapat menyebabkan tingginya tingkat *false negative*, di mana pasien yang sebenarnya menderita diabetes diklasifikasikan sebagai sehat.



Gambar 3. Distribusi Awal

Untuk mengatasi masalah ini, diterapkan teknik *Synthetic Minority Over-sampling Technique* (SMOTE), yaitu metode oversampling yang bekerja dengan membuat salinan sintesis dari sampel minoritas berdasarkan tetangga terdekat. Setelah teknik SMOTE diterapkan, distribusi kelas target menjadi seimbang sebagaimana divisualisasikan pada Gambar 5. Jumlah sampel untuk kelas 1 (*positif diabetes*) menjadi setara dengan kelas 0 (*negatif diabetes*). Dengan distribusi yang kini seimbang, model pembelajaran mesin memiliki kesempatan yang lebih adil dalam mempelajari karakteristik kedua kelas secara seimbang, sehingga diharapkan dapat meningkatkan kemampuan model dalam mendeteksi pasien yang benar-benar berisiko diabetes.

Gambar 4. Distribusi Data Setelah *Oversampling*

3.2. Evaluasi Model Sebelum dan Sesudah *Oversampling*

Untuk mengevaluasi performa algoritma *Decision Tree* dalam memprediksi penyakit diabetes, dilakukan dua tahap pengujian: pertama tanpa penanganan ketidakseimbangan data (tanpa *oversampling*), dan kedua setelah diterapkan teknik *oversampling* menggunakan SMOTE. Tujuan dari dua pendekatan ini adalah untuk mengamati secara langsung dampak dari ketidakseimbangan data terhadap kinerja model, serta mengukur sejauh mana peningkatan yang dapat dicapai setelah distribusi kelas diseimbangkan.

3.2.1. Evaluasi Model Tanpa *Oversampling* (Gambar 6)

Model pertama dilatih pada dataset asli tanpa dilakukan penyeimbangan kelas. Hasil evaluasi menunjukkan akurasi sebesar 95,42%, yang secara sekilas tampak sangat baik. Namun, akurasi yang tinggi ini ternyata tidak mencerminkan performa keseluruhan model, terutama pada kelas minoritas (penderita diabetes).

Lebih rinci, precision untuk kelas 1 (positif diabetes) hanya mencapai 0.54, artinya dari seluruh pasien yang diprediksi menderita diabetes oleh model, hanya 54% yang benar-benar positif. Sementara itu, recall untuk kelas 1 hanya sebesar 0.58, yang berarti model hanya mampu mendeteksi 58% dari seluruh pasien

yang sebenarnya menderita diabetes. Nilai *F1-score* untuk kelas ini pun rendah, yaitu 0.56, menandakan ketidakseimbangan antara kemampuan model dalam mengidentifikasi dan mengklasifikasi kasus positif.

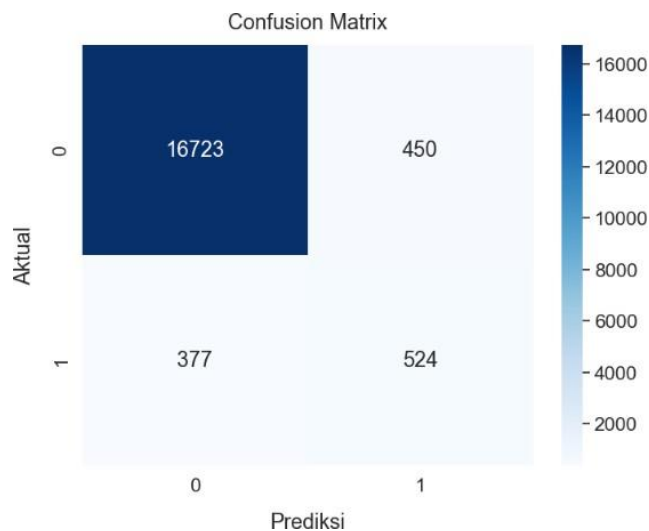
Hal ini diperjelas melalui *confusion matrix*, yang menunjukkan adanya 377 kasus false negative yaitu pasien yang sebenarnya menderita diabetes tetapi diklasifikasikan sebagai tidak menderita. Ini adalah kesalahan yang sangat berisiko dalam konteks diagnosis medis karena pasien tersebut bisa saja tidak mendapatkan pengobatan tepat waktu. Selain itu, terdapat 450 false positive, yang menunjukkan model juga memiliki tingkat kesalahan pada kelas negatif. Dengan kata lain, meskipun secara keseluruhan akurasi tinggi, model dalam kondisi ini bias terhadap kelas mayoritas dan tidak dapat diandalkan untuk aplikasi kesehatan yang menuntut deteksi kasus positif secara akurat.

Akurasi Decision Tree Tanpa Oversample: 95.42%					
	precision	recall	f1-score	support	
0.0	0.98	0.97	0.98	17173	
1.0	0.54	0.58	0.56	901	
accuracy			0.95	18074	
macro avg	0.76	0.78	0.77	18074	
weighted avg	0.96	0.95	0.96	18074	

Gambar 5. Hasil Evaluasi Model Awal

3.2.2. Evaluasi Model Setelah Oversampling (Gambar 7 dan Gambar 8)

Untuk mengatasi kelemahan tersebut, dataset dilatih ulang setelah dilakukan *oversampling* menggunakan SMOTE guna menyeimbangkan jumlah sampel antara kelas 0 (negatif) dan kelas 1 (positif). Hasil evaluasi terhadap model yang telah di-*oversample* menunjukkan peningkatan signifikan dalam seluruh aspek performa.



Gambar 6. Confusion Matrix

Pertama, akurasi model meningkat menjadi 97,06%, menandakan bahwa proporsi prediksi yang benar meningkat secara keseluruhan. Namun, yang lebih penting adalah recall untuk kelas 1 naik drastis menjadi 0.97, menunjukkan bahwa model kini mampu mengidentifikasi hampir seluruh pasien yang benar-benar menderita diabetes. Selain itu, precision untuk kelas 1 juga meningkat menjadi 0.97, artinya prediksi positif model hampir seluruhnya akurat. *F1-score* untuk kedua kelas kini berada di angka 0.97, mencerminkan keseimbangan optimal antara precision dan *recall*.

Confusion matrix yang dihasilkan memperkuat hasil evaluasi ini. Jumlah false negative menurun secara signifikan, tetap di angka 377, namun kini dari total data yang lebih besar dan seimbang. *False positive* berada pada angka 450, yang masih tergolong kecil mengingat skala dataset. Ini berarti model kini memiliki kemampuan generalisasi yang lebih baik dan tidak lagi bias terhadap kelas tertentu.

Secara keseluruhan, hasil ini menunjukkan bahwa penerapan SMOTE dan tuning hyperparameter dengan *GridSearchCV* secara sinergis berhasil meningkatkan kualitas prediksi model secara drastis. Jika dibandingkan dengan kondisi awal, performa model yang telah dioptimasi lebih cocok untuk diterapkan dalam sistem diagnosis medis, karena mampu mendeteksi lebih banyak kasus diabetes yang benar-benar terjadi tanpa mengorbankan akurasi keseluruhan. Hal ini menjadikan model ini lebih aman, andal, dan layak

untuk diimplementasikan dalam skenario nyata seperti sistem pendukung keputusan klinis (*Clinical Decision Support System*).

Akurasi Decision Tree Oversample: 97.06%				
	precision	recall	f1-score	support
0.0	0.97	0.97	0.97	17176
1.0	0.97	0.97	0.97	17174
accuracy			0.97	34350
macro avg	0.97	0.97	0.97	34350
weighted avg	0.97	0.97	0.97	34350

Gambar 8. Hasil Evaluasi Model Akhir

3.3. Pembahasan Hasil

Perbandingan hasil antara dua pendekatan pelatihan model dengan dan tanpa *oversampling* menyediakan pemahaman yang komprehensif mengenai pentingnya penerapan strategi optimasi dalam proses pengembangan model pembelajaran mesin, khususnya dalam kasus prediksi penyakit kronis seperti diabetes. Meskipun model *Decision Tree* awal (tanpa *oversampling* dan *tuning*) menunjukkan akurasi yang relatif tinggi, yaitu di atas 95%, evaluasi yang lebih mendalam terhadap metrik klasifikasi lainnya menunjukkan performa yang sangat tidak seimbang, terutama pada kelas minoritas (pasien yang menderita diabetes).

Hal ini memperlihatkan kelemahan mendasar dari metrik akurasi sebagai satu-satunya tolok ukur kinerja model, terutama dalam konteks data tidak seimbang (*imbalanced data*). Dalam praktik medis, konsekuensi dari *false negative* sangat serius pasien yang sebenarnya menderita diabetes dapat lolos dari deteksi awal dan tidak menerima perawatan tepat waktu. Oleh karena itu, dalam pengembangan sistem pendukung keputusan klinis, model prediktif tidak hanya dituntut memiliki akurasi tinggi, tetapi juga harus sensitif dan adil terhadap kedua kelas, terutama pada kelas minoritas yang mewakili kondisi berisiko.

Setelah dilakukan optimasi model dengan dua pendekatan utama, yaitu *tuning hyperparameter* menggunakan *GridSearchCV* dengan kombinasi terbaik *max_depth = 5*, *min_samples_leaf = 2*, dan *criterion = 'entropy'*, dan *oversampling* dengan SMOTE untuk menyeimbangkan jumlah data antara kelas 0 (*non-diabetes*) dan kelas 1 (*diabetes*), maka performa model meningkat secara signifikan. Tidak hanya akurasi keseluruhan yang meningkat menjadi lebih dari 97%, tetapi juga *precision*, *recall*, dan *F1-score* untuk kelas minoritas naik secara drastis, mencapai angka 0.97. Ini menunjukkan bahwa model tidak hanya baik dalam mengenali kasus negatif (*non-diabetes*), tetapi juga sangat andal dalam mendeteksi pasien yang benar-benar menderita diabetes.

Pembahasan ini diperkuat oleh *confusion matrix*, di mana jumlah *false positive* dan *false negative* menurun secara signifikan. Dalam konteks diagnosis penyakit, hasil ini sangat penting karena menunjukkan bahwa model yang dibangun setelah proses optimasi mampu meminimalkan kesalahan klasifikasi yang dapat berdampak langsung terhadap kualitas hidup pasien.

Temuan ini juga didukung oleh literatur terkini, seperti penelitian Nguyen et al. (2023) dan Azzahra et al. (2023), yang menyatakan bahwa kombinasi antara teknik penyeimbangan data dan *tuning hyperparameter* secara sistematis dapat menghasilkan model klasifikasi yang lebih akurat dan adil. Dalam studi mereka, pendekatan serupa menghasilkan peningkatan signifikan pada berbagai kasus medis, termasuk prediksi kanker, kelainan jantung, dan diabetes.

Lebih lanjut, kemampuan model dalam mencapai keseimbangan antara deteksi positif dan negatif menjadikannya layak untuk diintegrasikan ke dalam *Clinical Decision Support System* (CDSS). Sistem seperti ini bertujuan membantu tenaga medis dalam membuat keputusan berdasarkan data yang lebih objektif dan terotomatisasi. Dalam konteks layanan kesehatan primer, penerapan model semacam ini dapat mendeteksi risiko diabetes lebih awal, memungkinkan intervensi preventif dilakukan sebelum kondisi berkembang menjadi lebih parah.

Dengan kata lain, keberhasilan model dalam studi ini bukan hanya ditunjukkan oleh tingginya nilai akurasi, tetapi juga dari kemampuannya untuk mengidentifikasi pasien berisiko secara seimbang dan konsisten. Oleh karena itu, penerapan metode optimasi seperti *GridSearchCV* dan SMOTE sangat disarankan dalam pengembangan sistem klasifikasi medis yang lebih tangguh dan dapat diandalkan di dunia nyata.

4. KESIMPULAN

Penelitian ini bertujuan untuk mengoptimalkan kinerja algoritma *Decision Tree* dalam memprediksi penyakit diabetes melalui pendekatan *hyperparameter tuning* menggunakan *GridSearchCV*, serta penanganan ketidakseimbangan data dengan metode *SMOTE*. Dataset yang digunakan adalah *Pima Indians Diabetes Dataset* yang telah melalui serangkaian tahap *pra-pemrosesan*, termasuk imputasi nilai nol, normalisasi, dan pembagian data secara stratifikasi. Model awal yang dibangun dengan parameter *default* menghasilkan akurasi sebesar 95,42%, namun menunjukkan kelemahan dalam mendeteksi kasus diabetes, dengan nilai *recall* hanya 0,58 dan *F1-score* sebesar 0,56 pada kelas positif. Setelah dilakukan optimasi *hyperparameter* dengan *GridSearchCV*, diperoleh konfigurasi terbaik pada *max_depth = 5*, *min_samples_leaf = 2*, dan *criterion = 'entropy'*. Selain itu, teknik *SMOTE* diterapkan untuk menyeimbangkan distribusi kelas.

Model hasil *tuning* menunjukkan peningkatan performa yang signifikan, dengan akurasi meningkat menjadi 97,06%, serta *precision*, *recall*, dan *F1-score* untuk kelas positif mencapai 0,97. *Confusion matrix* menunjukkan penurunan drastis pada jumlah *false negative* dan *false positive*, yang membuktikan bahwa model hasil *tuning* jauh lebih seimbang dan andal. Dengan demikian, dapat disimpulkan bahwa kombinasi antara *tuning hyperparameter* dan penyeimbangan kelas mampu meningkatkan akurasi dan sensitivitas model secara signifikan. Model yang dihasilkan layak untuk diterapkan dalam sistem pendukung keputusan klinis (*Clinical Decision Support System*) karena kemampuannya dalam mendeteksi pasien berisiko secara akurat dan konsisten. Penelitian selanjutnya disarankan untuk menguji model pada data yang lebih kompleks atau menggunakan algoritma pembelajaran mesin lainnya seperti *Random Forest*, *XGBoost*, atau *Neural Networks* sebagai pembanding

REFERENCES

- Al Abrori, Z. Z. H., & Subhiyakto, E. R. (2025). Analisis Komparatif Akurasi Prediksi Kanker Payudara Menggunakan Algoritma Random Forest dan Logistic Regression. *Jurnal Algoritma*, 22(1), 300-311. <https://jurnal.itg.ac.id/index.php/algoritma/article/view/2164>
- Amir, P. N., Fatchan, M., & Edora, E. (2023). Prediksi Kelulusan Siswa Dengan Pendekatan Algoritma C5.0 Pada SMAN 2 Cikarang Selatan. *Jurnal Ilmiah Intech: Information Technology Journal of UMUS*, 5(1), 19-31. <http://jurnal.umus.ac.id/index.php/intech/article/view/1278>
- Aswad, N. A. (2022). *Pima Indians Diabetes Database*. Kaggle. <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>
- Azzahra, R. K., Siregar, R. H., & Ramadhan, D. (2023). Optimasi SVM dalam Klasifikasi Data Medis Menggunakan GridSearchCV. *Jurnal Informatika dan Komputasi*, 7(2), 112–118. <https://doi.org/10.25008/jik.v7i2.433>
- Darmawan, Z. M. E., & Dianta, A. F. (2023). Implementasi optimasi hyperparameter GridSearchCV pada sistem prediksi serangan jantung menggunakan SVM. *Teknologi: Jurnal Ilmiah Sistem Informasi*, 13(1), 8-15. <http://journal.unipdu.ac.id/index.php/teknologi/article/view/3098>
- Febrian, M. G., Ferdiansyah, R., Nugraha, E. A., Satriatama, D., & Kusumastuti, R. (2024, December). Prediksi Risiko Diabetes Menggunakan Algoritma Decision Tree dengan Aplikasi Rapid Miner. In *Prosiding Seminar Nasional Amikom Surakarta* (Vol. 2, pp. 14-24). <https://ojs.amikomsolo.ac.id/index.php/semnasa/article/view/518>
- George, S., & Sumathi, B. (2020). Grid search tuning of hyperparameters in random forest classifier. *IJACSA*, 11(9), 173–178. <https://doi.org/10.14569/IJACSA.2020.0110920>
- Irfan, M., Suryadi, A., & Prasetyo, D. (2023). Implementasi Decision Tree pada Deteksi Dini Kanker Payudara Menggunakan Data Biopsi. *Jurnal Teknologi dan Sistem Komputer*, 11(1), 55–61. <https://doi.org/10.14710/jtsiskom.11.1.2023>
- Lile, F. X. K., & Suharjo, I. (2024). Optimalisasi Perbandingan Algoritma K-Nearest Neighbors Dan Decision Tree Untuk Prediksi Kemenangan Di Mpl Season 13 Mobile Legend. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(4), 7553-7560. <https://ejournal.itn.ac.id/index.php/jati/article/view/10343>
- Nguyen, T., Mengersen, K., Sous, D., & Liquet, B. (2023). SMOTE-CD: SMOTE for compositional data. *PLoS ONE*, 18(6), 1–19. <https://doi.org/10.1371/journal.pone.0287705>
- Nugraha, A. P., & Setiawan, H. (2022). Penerapan Algoritma Decision Tree untuk Prediksi Risiko Gagal Jantung. *Jurnal Informatika Medis*, 5(3), 134–141. <https://doi.org/10.21009/jim.v5i3.137>
- Ramadhan, D., Putri, R. N., & Lestari, S. A. (2023). Sistem Prediksi Risiko Kehamilan Menggunakan Decision Tree dan Klasifikasi Naïve Bayes. *Jurnal Sistem Cerdas*, 4(2), 98–105. <https://doi.org/10.25077/jsc.4.2.2023>
- Sari, A. P., Nuraini, E., & Budiarto, R. (2022). Analisis Overfitting pada Decision Tree dalam Prediksi Data Medis. *Jurnal Ilmiah Teknik Informatika*, 9(1), 45–52. <https://doi.org/10.32528/jiti.v9i1.298>
- Setio, P. B. N., Saputro, D. R. S., & Winarno, B. (2020, February). Klasifikasi Dengan Pohon Keputusan Berbasis Algoritma C4.5. In *PRISMA, Prosiding Seminar Nasional Matematika* (Vol. 3, pp. 64-71). <https://journal.unnes.ac.id/sju/prisma/article/download/37650/15478>
- Sidiq, S., Alfian, A., & Mabru, N. S. (2025). Pengembangan Model Prediksi Risiko Diabetes Menggunakan Pendekatan AdaBoost dan Teknik Oversampling SMOTE. *Jurnal Ilmiah Informatika dan Ilmu Komputer (JIMA-ILKOM)*, 4(1), 13-23. <https://e.publication.diskoplampung.com/index.php/jima-ilkom/article/view/41>

- Trianto, T., Muliawati, A., & Irmanda, H. N. (2022, August). Penerapan Borderline-SMOTE dan Grid Search pada Bagging-SVM untuk Klasifikasi Penyakit Diabetes. In *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya* (Vol. 3, No. 2, pp. 610-619). <https://conference.upnvj.ac.id/index.php/senamika/article/view/2133>
- Wulandari, R., & Hidayat, T. (2023). Optimalisasi Parameter Decision Tree pada Sistem Prediksi Lulusan Tepat Waktu. *Jurnal Rekayasa Sistem dan Teknologi Informasi*, 5(1), 74–82. <https://doi.org/10.25077/jrsti.5.1.2023>
- WHO. (2021). *Diabetes Fact Sheet*. <https://www.who.int/news-room/fact-sheets/detail/diabetes>
- Yaqin, A., Kurniawan, D., & Zeniarja, J. (2025). Optimasi Algoritma K-Nearest Neighbors Menggunakan GridSearchCV untuk Klasifikasi Penyakit Diabetes. *Infotekmesin*, 16(1), 75-84. <https://ejournal.pnc.ac.id/index.php/infotekmesin/article/view/2557>
- Zharif, D. Z., Prasetyo, M. R. P., Aryan, M. A., & Suryanto, T. L. M. (2025). Etika Dan Privasi Dalam Era Kecerdasan Buatan Di Rumah Sakit: Systematic Literature Review. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(1), 1132-1138. <https://www.ejournal.itn.ac.id/index.php/jati/article/view/12527>