e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 8 No. 2, Desember, 2024, **Hal. 714-723**
DOI: 10.29408/edumatic.v8i2.27990

Analisis Performa Model Random Forest dan CatBoost dengan Teknik SMOTE dalam Prediksi Risiko Diabetes

Rony Irfannandhy [1,*], Lekso Budi Handoko [1], Noval Ariyanto [1]

[1] Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Indonesia
[*] Correspondence: 111202113825@mhs.dinus.ac.id

Received: 30 Oktober 2024 | Revised: 20 November 2024 | Accepted: 3 Desember 2024 | Published: 19 Desember 2024

Abstrak
Diabetes melitus (DM) prevalensinya meningkat global menjadi masalah kesehatan serius. Deteksi dini mengurangi komplikasi jangka panjang. Tujuan penelitian kami adalah mengevaluasi serta membandingkan efektivitas model *Random Forest* (RF) dan *CatBoost* dengan teknik *SMOTE* dalam memprediksi risiko DM berdasarkan data uji yang diolah untuk menghasilkan performa analisis perbandingan kedua model berupa *precission*, *recall*, *F1-Score* dan akurasi. Jenis penelitian kami adalah kuantitatif menggunakan metode yang mencakup *EDA*, transformasi, membagi data uji dan *training*, implementasi metode RF dan *CatBoost* dengan SMOTE serta evaluasi performa model. Dataset dari platform (Kaggle) mencakup 768 data kesehatan individu terdiri delapan variabel independen kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, Indeks Massa Tubuh (IMT), riwayat DM, usia serta satu variabel target (*outcome*) status DM. Teknik analisis SMOTE diterapkan untuk menyeimbangkan distribusi kelas serta meningkatkan representasi kelas minoritas model prediksi lebih akurat dan stabil. Hasil temuan model SMOTE-*RF* akurasinya 82% dan SMOTE *CatBoost* akurasinya 81%. Berdasarkan analisis *feature importances*, variabel utama memengaruhi prediksi risiko DM kedua model kadar Glukosa, IMT dan usia. Variabel glukosa adalah indikator risiko utama DM yang digunakan prediksi agar lebih efisien. Implikasi praktis peningkatan deteksi dini *machine learning* berpotensi mendukung pengambilan keputusan dokter lebih akurat untuk mencegah komplikasi lebih serius pada diabetes melitus.

Kata kunci: *catboost*; diabetes; model prediksi; pembelajaran mesin; *random forest*

Abstract
Diabetes mellitus (DM) is increasing in prevalence globally and is becoming a serious health problem. Early detection reduces long-term complications. The purpose of our research is to evaluate and compare the effectiveness of Random Forest (RF) and CatBoost models with SMOTE technique in predicting DM risk based on test data processed to produce comparative analysis performance of both models in the form of precission, recall, F1-Score and accuracy. Our research type is quantitative using methods that include EDA, transformation, dividing test and training data, implementation of RF and CatBoost methods with SMOTE and evaluation of model performance. The dataset from the platform (Kaggle) includes 768 individual health data consisting of eight independent variables of pregnancy, glucose, blood pressure, skin thickness, insulin, Body Mass Index (BMI), DM history, age as well as one target (outcome) variable of DM status. The SMOTE analysis technique was applied to balance the class distribution and improve the representation of the minority class, making the prediction model more accurate and stable. The findings of the SMOTE-RF model were 82% accuracy and SMOTE CatBoost 81% accuracy. Based on the feature importances analysis, the main variables affecting DM risk prediction of both models are glucose, BMI and age. Glucose variable is the main DM risk indicator used for prediction to be more efficient. The practical implication of improved machine learning early detection has the potential to support doctors' decision making more accurately to prevent more serious complications in diabetes mellitus.

Keywords: *catboost; diabetes; prediction model; machine learning; random forest*

PENDAHULUAN

Diabetes melitus (DM) merupakan salah satu masalah kesehatan global yang terus meningkat cepat menjadi penyebab utama komplikasi serius seperti penyakit kardiovaskular, gagal ginjal dan neuropati (Li et al., 2019). Menurut (Ardiansyah et al., 2021), DM yakni penyakit kronis metabolik dimana glukosa meningkat tinggi. Data terbaru dari International Diabetes Federation (IDF) pada tahun 2021, mengungkapkan lebih dari 537 juta orang dewasa terjangkit DM. Angka ini diperkirakan meningkat 643 juta pada tahun 2030 (Resti et al., 2021). Hingga saat ini DM masih dicatat menjadi penyebab kematian tertinggi di dunia peningkatan signifikan setiap tahun (Irwansyah et al., 2021). Peningkatan prevalensi DM menunjukkan pentingnya deteksi dini dan *intervensi* tepat waktu agar dampak penyakit terhadap kualitas hidup pasien dapat diminimalkan (Dennison et al., 2021).

Salah satu penyebab peningkatan signifikan prevalensi masalah kesehatan pada DM yakni kadar glukosa darah berlebih dalam tubuh. Pola makan tidak terkontrol tinggi glukosa beserta kurangnya aktivitas fisik seperti olahraga tidak teratur, teridentifikasi penyebab utama kondisi ini (Hidayat et al., 2020). Terobosan serius dibutuhkan untuk masalah tersebut dengan mengembangkan metode deteksi dini yang kompleks guna mengurangi indikator DM terhadap kualitas hidup pasien (Maulana et al, 2024). Seiring perkembangan teknologi, dalam penelitian (Taufik & Kurniawan, 2023) *machine learning* menjadi pendekatan terobosan industri melalui *study* data dengan sajian pola yang ada sehingga menjanjikan dalam prediksi risiko DM. Kemampuan algoritma pembelajaran mesin mengolah data berukuran besar serta kompleks memungkinkan analisis lebih akurat dibanding metode konvensional (Wang et al., 2020). Berbagai model algoritma, *Random Forest* (RF) dan *CatBoost* dipilih karena telah terbukti memberikan performa unggul prediksi risiko penyakit kronis DM (Sabili et al., 2024).

RF sebagai *ensemble learning* sangat diandalkan beroperasi dengan menggabungkan banyak pohon keputusan untuk peningkatan akurasi prediksi lebih akurat dengan bergantung nilai vektor acak disampling independen dengan distribusi pohon agar semua sama (Sari et al., 2020). RF juga mengatasi masalah *overfitting* dengan mengurangi variasi dan meningkatkan generalisasi model yang terlalu kompleks (Tarigan & Dahlan, 2024). RF menangani dataset berdimensi tinggi yang memiliki karakteristik heterogen dan mampu menentukan fitur relevan dengan risiko DM untuk membantu keputusan klinis dokter (Zulfiansyah et al., 2023). Adapun algoritma *CatBoost* dipilih karena rancangannya dalam menangani data fitur kategorikal menggunakan pendekatan *ordered boosting* yang inovatif. Pendekatan ini mampu mengurangi kesalahan prediksi, menghasilkan model lebih stabil dan akurat serta dataset yang kompleks dan tidak seimbang (Syandika & Yustanti, 2023). *CatBoost* selain andal memprediksi, efisien memproses pelatihan model tanpa mengorbankan akurasi (Zhong et al., 2023). Kedua algoritma RF dan *CatBoost* meskipun sangat menjanjikan, memiliki tantangan menangani dataset medis yang sering kali tidak seimbang. Ketidakseimbangan terjadi pada kelas minoritas seperti pasien dengan risiko tinggi DM, dapat menyebabkan bias pada hasil prediksi. Teknik praproses optimalisasi kinerja algoritma model RF serta *CatBoost* menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan guna menyeimbangkan distribusi kelas (Siboro et al., 2024). SMOTE meningkatkan representasi kelas minoritas model prediksi lebih akurat dan stabil. Evaluasi kinerja model menggunakan teknik validasi silang dan analisis *hyperparameter tuning* untuk memastikan reliabilitas hasil (Palupi et al., 2023).

Penelitian sebelumnya telah menerapkan RF untuk memprediksi DM dengan akurasi sebesar 74% dan *XGBoost* 76%, menggunakan dataset berjumlah 768 entri dengan sembilan indikator yang diperoleh dari Kaggle.com (Salsabil et al., 2024). Penelitian lain oleh (Andi et al., 2023) menganalisis penyakit DM dengan menggunakan dataset berjumlah 768 entri dari

Kaggle.com, yang dibagi menjadi 70% data pelatihan dan 30% data uji. Temuan mereka telah menerapkan algoritma *machine learning* guna membandingkan performa beberapa pendekatan penggalian data (*data mining), yang* menghasilkan akurasi model *Decision Tree* sebesar 71,86%, RF 75,03%, *CatBoost* 76,19% dan *RF-Grid Search* 77,06%. Penelitian lainnya oleh (Nainggolan & Sinaga, 2023) membandingkan algoritma RF dan *Gradient Boosting* untuk diagnosis DM menggunakan dataset berjumlah 768 entri dengan sembilan indikator yang diperoleh dari Kaggle.com. Studi ini menghasilkan nilai perbandingan akurasi RF sebesar 79% dan *gradient boosting* 81%. Namun, dari berbagai hasil penelitian masih memiliki beberapa keterbatasan. Pertama, metode yang digunakan tidak menangani masalah ketidakseimbangan data dalam dataset, di mana kelas 1 minoritas pasien dengan DM cenderung kurang terwakili sehingga dapat menyebabkan bias pada model prediksi, terutama dalam mendeteksi kasus positif dengan risiko tinggi. Kedua, penelitian tersebut tidak memanfaatkan teknik optimasi seperti SMOTE, yang terbukti efektif meningkatkan representasi kelas minoritas dan stabilitas prediksi. Ketiga, tidak dilakukan analisis secara mendalam terhadap variabel-variabel utama *(feature importance)* yang memengaruhi risiko DM, sehingga membatasi wawasan terkait faktor-faktor signifikan yang dapat digunakan dalam deteksi dini DM.

Oleh karena itu, diperlukan pengembangan model lebih baik lagi guna meningkatkan performa model akurasi prediksi DM dengan mengintegrasikan *SMOTE* dalam menangani ketidakseimbangan data yang sejalan dengan penelitian oleh Syukron et al. (2020) membahas perbandingan performa model RF dengan meningkatkan akurasi sebesar 74,79% dengan SMOTE-RF 80,97% dan *XGBoost* sebesar 74,68% dengan SMOTE-*XGBoost* 78,63%. Penelitian tersebut menggunakan dataset dari *UCI Machine Learning Repository* mencakup 1385 data. Selain itu, penelitian lain oleh (Oktaviani et al., 2024) membandingkan RF dengan *oversampling* SMOTE-RF menggunakan data kuesioner mahasisswa tingkat akhir, dengan akurasi RF 54% menjadi 71% menguatkan bahwa setelah menerapkan SMOTE ampuh meningkatkan signifikan dalam performa akurasi prediksi. Di sisi lain, belum adanya analisis mendalam terhadap variabel utama *feature importance* sangat diperlukan untuk identifikasi faktor yang berkontribusi signifikan terhadap risiko DM yang membedakan dari penelitian sebelumnya. Namun, beberapa penelitian sebelumnya masih memerlukan metode yang lebih komprehensif dalam meningkatkan akurasi prediksi.

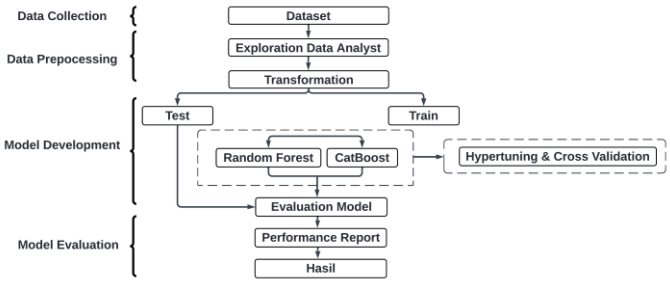
Penelitian ini bertujuan mengatasi keterbatasan tersebut dengan mengevaluasi serta membandingkan efektivitas model RF dan *CatBoost* dengan teknik *SMOTE* dalam memprediksi risiko DM berdasarkan data uji yang diolah untuk menghasilkan performa analisis perbandingan kedua model berupa *precission*, *recall*, *F1-Score* dan akurasi, serta mendapatkan variabel indikator risiko utama yang harus dihindari agar masyarakat tidak terjangkit DM. Temuan penelitian ini diharapkan berkontribusi nyata dalam meningkatkan performa model prediksi DM dengan akurasi yang lebih tinggi dibandingkan hasil dari penelitian sebelumnya, serta hasilnya dapat mendukung pengambilan keputusan klinis dokter secara akurat dan efisien untuk mencegah komplikasi yang lebih serius.

METODE

Penelitian ini bertujuan memprediksi risiko diabetes menggunakan algoritma model *Random Forest* (RF) dan *CatBoost* dengan teknik *oversampling* SMOTE untuk meningkatkan hasil performa akurasi. Adapun tahapan metode penelitian ini dikelompokkan empat langkah utama pada gambar 1. diantaranya *data collection*, *data prepocessing*, *model development* dan *model evaluation*. Pada tahap awal *data collection*, proses pengumpulan data mencakup eksplorasi mendalam guna memahami peran variabel guna memprediksi risiko DM. Dataset dari platform Kaggle.com dengan 768 sampel individu dan delapan variabel independen serta satu variabel target. Variabel independen mencakup jumlah kehamilan, glukosa, tekanan darah, ketebalan kulit, kadar insulin, IMT, riwayat DM keluarga dan usia. Variabel target

menunjukkan status DM, dengan kategori 0 (tidak diabetes) dan 1 (diabetes). Variabel kadar glukosa dan IMT terbukti pengaruh signifikan terhadap hasil prediksi. sedangkan variabel lainnya berfungsi sebagai faktor tambahan dalam model.

Langkah *data preppocessing* tahapan *Exploration Data Analyst* (EDA) melibatkan analisis statistik untuk identifikasi karakteristik variabel guna mengatasi ketidakseimbangan standarisasi data numerik keseimbangan kelas menggunakan SMOTE. Eksplorasi lanjut dilakukan pengecekan *missing value* atau *nan* untuk memastikan tidak ada sama sekali *missing* data yang di teliti dilanjutkan dengan melihat distribusi statistik data beserta analisis *outliers* untuk memeriksa distribusi hubungan antar variabel guna identifikasi faktor paling pengaruh terhadap prediksi. Hubungan kuat glukosa dan IMT di status DM memberikan dasar pemilihan fitur model. Selama proses ini, tren dan anomali dianalisis memastikan kualitas data optimal sebelum pemodelan. Transformasi data meningkatkan kestabilan akurasi model. Variabel target disesuaikan guna mengurangi variasi. Variabel numerik lainnya ditransformasi berdasarkan distribusi masing-masing. Pendekatan ini dirancang guna optimalisasi performa model secara keseluruhan.



Gambar 1. Tahapan penelitian

Setelah transformasi, data dibagi menjadi data pelatihan dan data uji yang selanjutnya diimplementasikan dengan algoritma RF dan *CatBoost*. RF merupakan algoritma *ensemble* yang menggabungkan banyak pohon keputusan untuk meningkatkan akurasi prediksi dan mengurangi risiko *overfitting*. Keunggulan utama RF mampu menangkap hubungan *non-linear* antar fitur, tahan terhadap data yang tidak seimbang dan efektif menangani *outliers*. RF dapat mengelola data berdimensi tinggi dengan baik sehingga relevan memprediksi pengambilan keputusan khususnya DM. Sedangkan algoritma *CatBoost* adalah algoritma *boosting* berbasis pohon keputusan dengan pendekatan *ordered boosting* yang mampu mengurangi dampak akumulasi kesalahan prediksi guna meningkatkan kestabilan dan keakuratan model. *CatBoost* efisien mengelola data heterogen tanpa memerlukan *tuning* parameter kompleks, menjadikannya sangat andal dalam analisis kesehatan.

Pada tahapan evaluasi, dilakukan *hyperparameter tuning* dengan langkah evaluasi kinerja model RF dan *CatBoost* menggunakan metrik akurasi, *precision*, *recall* dan *f1-score* untuk mengukur performa prediksi guna memberikan gambaran menyeluruh kekuatan model dalam mendeteksi pola keunggulan masing-masing model. Hasilnya diharapkan mendukung upaya deteksi dini dan manajemen risiko DM lebih akurat dan efisien dengan perbandingan variabel indikator tersignifikan dalam memberikan dampak buruk DM yang membantu dokter mengambil keputusan secara klinis.

HASIL DAN PEMBAHASAN

Hasil

Berdasarkan tahapan metode penelitian, pengujian algoritma RF dan *CatBoost* dilakukan guna memprediksi performa akurasi DM berdasarkan dataset dari Kaggle.com untuk di*preprocessing*. Analisis melalui *feature importance* dilakukan supaya kontribusi setiap fitur

mendapatkan variabel indikator utama yang dampak buruk pada DM, serta dampak penerapan SMOTE berhasil mengatasi ketidakseimbangan data sehingga meningkatkan performa hasil akurasi prediksi penyakit DM. Setelah praproses, dataset dibagi dua dengan rasio 80:20 yang menghasilkan 614 data latih dan 154 uji. Tujuannya memastikan model terlatih cukup data sekaligus menyediakan data independen yang menguji performa model secara objektif.

Metodologi EDA diterapkan karena berpengaruh mencakup analisis statistik deskriptif variabel utama dataset DM seperti glukosa, tekanan darah dan IMT dalam menentukan hasil. Matriks korelasi guna memahami keterkaitan variabel tersebut dengan risiko DM. Analisis statistik deskriptif memberikan pemahaman awal yang penting tentang karakteristik variabel dataset DM. Hasil pada tabel 1 menunjukkan rangkuman statistik deskriptif variabel utama meliputi nilai minimum, rata-rata (*mean*) dan maksimum setiap variabel. Hasil analisis statistik deskriptif variabel yang berhubungan menunjukkan bahwa Glukosa memiliki *mean* 120,9 dan nilai tertinggi mencapai 199, sehingga variabel glukosa paling dominan korelasi tertinggi terhadap *outcome* DM. Hal tersebut mengindikasikan bahwa glukosa faktor utama risiko DM. IMT memiliki *mean* sebesar 32,0, artinya responden memiliki obesitas atau tingkat risiko tinggi DM tipe 2. Riwayat DM dan Usia korelasinya signifikan dengan usia kisaran hingga 81 tahun menggambarkan risiko peningkatan bertambahnya usia. Rentang nilai insulin yang lebar (15 hingga 846) menunjukkan variasi signifikan responden menandakan resistensi insulin kelompok tertentu.

Tabel 1. Hasil statistik deskriptif

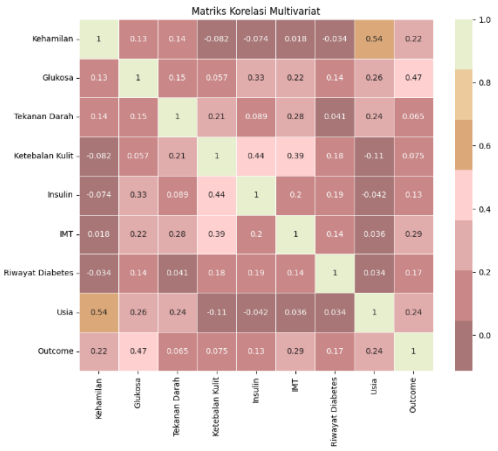
Variabel	*Min*	*Mean*	*Max*	Deskripsi
Kehamilan	0	3,8	17	Jumlah kehamilan risiko gestasional.
Glukosa	44	120,9	199	Kadar glukosa darah indikator diabetes.
Tekanan Darah	24	69,1	122	Tekanan darah gangguan metabolik.
Ketebalan Kulit	7	20,5	99	Ketebalan lemak risiko diabetes tipe 2.
Insulin	15	79,8	846	Kadar insulin indikasi resistensi insulin.
IMT	18,2	32,0	67,1	Indeks Massa Tubuh (IMT) risiko obesitas.
Riwayat Diabetes	0,08	0,47	2,42	Riwayat keluarga diabetes faktor genetik.
Usia	21	33,2	81	Usia lanjut penurunan metabolisme.

Setelah pemeriksaan nilai *outlier* variabel insulin dan ketebalan kulit, data diproses mengganti nilai ekstrem dengan nilai *median* agar stabilitas model tetap terjaga. *Outlier* insulin penting, mengingat rentang yang luas (15-846) untuk memastikan variabel tetap relevan dalam analisis risiko. Glukosa dan IMT merupakan variabel distandarisasi guna meningkatkan konsistensi hasil. Data yang hilang diatasi menggunakan imputasi *mean* variabel rendah risiko terhadap bias seperti tekanan darah dan ketebalan kulit. Transformasi diharapkan memperbaiki kualitas data untuk menghasilkan prediksi lebih akurat. Berdasarkan hasil Kehamilan menjadi indikator paling rendah dengan *mean* jumlah kehamilan gestasional 3,8 dan tertinggi 17.

Hasil analisis korelasi antara variabel kesehatan utama dan risiko DM menunjukkan hasil korelasi berkisar -1 hingga 1, nilai mendekati 1 menunjukkan hubungan positif kuat dan nilai mendekati -1 hubungan negatif yang signifikan. Korelasi mendekati 0 mengindikasikan tidak ada hubungan *linier* signifikan. Temuan ini mengidentifikasi variabel utama yang korelasi signifikan terhadap risiko DM serta memberikan kontribusi meningkatkan akurasi model prediksi.

Hasil pada gambar 2 menampilkan Matriks Korelasi Multivariat, yang menggambarkan hubungan variabel dataset. Glukosa menunjukkan korelasi tertinggi (r = 0,47) dengan risiko DM meningkatkan glukosa secara signifikan berrisiko DM. Glukosa penting dalam model prediktif dalam membantu deteksi dini individu dengan glukosa tinggi. Temuan ini sejalan dengan literatur medis, sehingga glukosa tinggi sering dikaitkan resistensi insulin dalam mekanisme utama perkembangan

DM. IMT memiliki korelasi moderat (r = 0,29) dengan risiko DM, mencerminkan obesitas atau IMT tinggi turut berkontribusi resistensi insulin dan risiko DM. Usia menunjukkan korelasi sedang (r = 0,24), mengindikasikan risiko DM meningkat seiring bertambahnya usia akibat penurunan metabolisme.



Gambar 2. Matriks korelasi multivariat

Sebaliknya, variabel tekanan darah (r = 0,065) dan ketebalan kulit (r = 0,075) korelasinya rendah risiko DM. Meskipun kurang signifikan model *linier*, variabel penting model *non-linear* seperti *CatBoost* yang menangkap interaksi variabel kompleks. Tekanan darah tinggi bisa menjadi indikator gangguan metabolik. Insulin bukan faktor utama rendah korelasi model *linier* tetap dipertimbangkan model *non-linear* karena insulin ekstrem mengindikasikan resistensi insulin atau disfungsi pankreas sehingga keduanya penting dalam risiko DM. Variabel utama Glukosa, IMT dan usia penting dalam kedua model guna meningkatkan akurasi prediksi risiko DM. Glukosa berkontribusi penting pada akurasi dan keandalan deteksi dini. Korelasi tinggi seperti glukosa dan IMT, indikator awal deteksi dini yang memungkinkan tenaga medis memfokuskan individu dengan risiko lebih tinggi. Identifikasi variabel utama model prediktif deteksinya lebih akurat dan efisien mendukung intervensi preventif sebelum komplikasi DM berkembang sehingga mencegah dan manajemen DM dalam konteks klinis.

Tabel 2. Metrik evaluasi untuk model *random forest*

Metrik	Kelas 0 (Non-Diabetes)	Kelas 1 (Diabetes)
Precision	0,88	0,78
Recall	0,74	0,90
F1-Score	0,80	0,83
Akurasi	0,82	0,82

RF dan *CatBoost* diterapkan menggunakan dataset kesehatan yang diproses, RF memberikan stabilitas gabungan beberapa pohon keputusan. *CatBoost* unggul menangani data tidak seimbang dan fitur kategorikal. Keduanya dioptimalkan melalui *hyperparameter tuning* guna mencapai performa model. Analisis *feature importance* RF dan *CatBoost* dalam variabel glukosa sebagai prediktor utama risiko DM. IMT dan usia memiliki peran signifikan menunjukkan faktor metabolik dan usia yang penting meningkatkan akurasi deteksi risiko DM. Tekanan darah dan ketebalan kulit berpengaruh lebih rendah dan tetap signifikan, terutama model *non-linear* seperti *CatBoost* mampu menangkap interaksi antar variabel kompleks. Grafik *feature importance* menggambarkan kontribusi relatif tiap fitur terhadap *output* model, menyoroti pentingnya variabel glukosa, IMT dan usia memprediksi risiko DM. RF sebagai model *baseline* unggul melakukan *ensemble* beberapa *decision trees* untuk

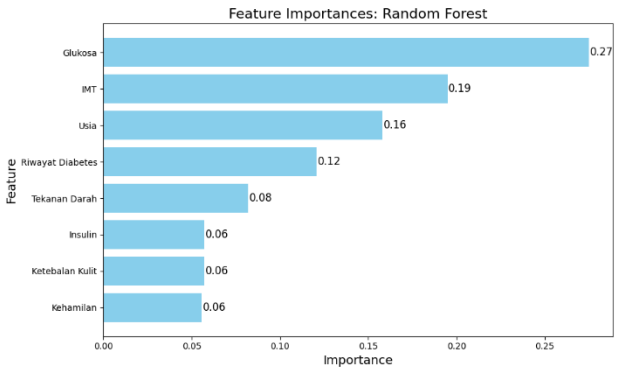
menangkap interaksi kompleks antar fitur. *Hyperparameter tuning* melalui proses *5-fold cross-validation*, yang mencakup eksplorasi 216 kombinasi parameter dengan total 1080 *fitting* guna meningkatkan hasil performa. Parameter optimal yang diperoleh adalah *max_depth* sebesar 20, *n_estimators* sebesar 50, *max_features* = 'sqrt', *min_samples_leaf* = 2, dan *min_samples_split* = 2.

Hasil evaluasi Tabel 2. menunjukkan bahwa model RF akurasinya sebesar 82%, *precision* 0,88 untuk kelas *non*-diabetes dan 0,78 untuk kelas DM, *recall* 0,90 pada kelas positif (diabetes) menunjukkan kapabilitas model lebih unggul dari *CatBoost* dalam mendeteksi risiko DM serta *f1-score* sebesar 0,83 pada kelas positif menegaskan keseimbangan baik antara *precision* dan *recall*, sehingga menjadikannya andal memprediksi DM. Percobaan tambahan variasi parameter model RF, *max_depth* dan *n_estimators*, serta *learning_rate* model *CatBoost*, memberikan hasil optimal dengan *max_depth* diatur ke 25 dan *n_estimators* ke 100. RF mencapai akurasi 82%, sedikit lebih tinggi dibanding konfigurasi sebelumnya. *CatBoost learning_rate* lebih rendah, yaitu 0,005 meningkatkan akurasi hingga 81% meski membutuhkan waktu komputasi lebih lama. Hasil menunjukkan *tuning* lebih lanjut terhadap parameter penting karena berperan meningkatkan performa model dan mengurangi risiko *overfitting* atau *underfitting*. Eksperimen ini penting dilakukannya *hyperparameter tuning* dalam mencapai performa optimal.

Tabel 3. Metrik evaluasi untuk model *catboost*

Metrik	Kelas 0 (Non-Diabetes)	Kelas 1 (Diabetes)
Precision	0,87	0,77
Recall	0,73	0,89
F1-Score	0,79	0,83
Akurasi	0,81	0,81

Sebagai pembanding, *CatBoost* dipilih karena mampu menangani fitur kategorikal langsung tanpa perlu *preprocessing* tambahan dan kekuatannya menangkap pola interaksi kompleks melalui *boosting*. Dalam mengoptimalkan performa, dilakukan *hyperparameter tuning* dengan *3-fold cross-validation* dengan mengevaluasi 81 kombinasi parameter dan menghasilkan total 243 fits. Parameter terbaik yang diperoleh meliputi *depth* sebesar 8, *iterations* sebanyak 300, *learning_rate* = 0.01, dan *l2_leaf_reg* = 3. Hasil evaluasi Tabel 3. menunjukkan bahwa *CatBoost* akurasinya 81%, *precision* 0,87 kelas *non*-diabetes dan 0,77 kelas DM, *recall* 0,89 termasuk tinggi pada kelas positif yang andal mengidentifikasi kasus DM tingkat *false negative* rendah dan *f1-score* sebesar 0,83 pada kelas positif.



Gambar 3. *Feature importances model random forest*

Sementara hasil prediksi gambar 3 menunjukkan bahwa variabel Glukosa berpengaruh tinggi terjadinya DM. Hasil Glukosa diambil dari *Feature Importances* model RF. Selain Glukosa pengaruh yang paling tinggi, variabel berisiko selanjutnya IMT, Usia, Riwayat DM, Tekanan Darah, Insulin Ketebalan Kulit dan yang paling rendah variabel Kehamilan, sehingga berdasarkan hasil uji variabel Glukosa sangat dihindari untuk mencegah terjangkit DM.

Pembahasan

Penelitian ini menghasilkan akurasi signifikan model dalam memprediksi akurasi DM menggunakan dataset dari *platform* Kaggle.com. Penyeimbangan kelas dengan teknik SMOTE diterapkan karena dataset memiliki distribusi kelas tidak seimbang, berdasarkan jumlah sampel kategori kelas 0 (tidak diabetes) jauh lebih banyak dibandingkan kategori kelas 1 (diabetes). Kontribusi di bidang medis dalam *machine learning* diharapkan meningkatkan hasil performa yang utama, presisi, *recall* dan *f1-Score* memberikan prediksi lebih baik dari yang telah dibuat.

Setelah dilakukan *hyperparameter tuning* dan validasi silang *5-fold* RF akurasinya 82%, *precision* 0,88 kelas *non*-diabetes dan 0,78 kelas DM serta *recall* 0,90 kelas positif mampu memodelkan deteksi risiko DM. *CatBoost* dengan akurasi 81% kompetitif melalui nilai *recall* yang bersaing pada kelas positif. Kemampuan model mengidentifikasi sebagian besar sampel kategori DM penting mencegah *false negatives*. Mekanisme *ordered boosting CatBoost* memastikan prediksi dataset stabil dengan distribusi kelas tidak seimbang yang meningkatkan keandalan model efisiensi dalam menangani variabel numerik dan kategoris sehingga menjadi alternatif andal mendukung deteksi dini DM secara lebih akurat dan efisien.

Hasil performa SMOTE-RF dan SMOTE-*CatBoost* lebih unggul dengan akurasi RF sebesar 82% dan *CatBoost* 81%. Berbeda dengan penelitian yang dilakukan (Salsabil et al., 2024) membuktikan performa RF dengan akurasi 74% dan *XGBoost* 76%. Penelitian lain dari (Andi et al., 2023) menganalisis DM dengan dataset berjumlah 768 *entri* dari Kaggle.com dibagi 70% data pelatihan dan 30% uji menghasilkan nilai akurasi penelitian *Decision Tree* sebesar 71,86%, RF 75,03%, *CatBoost* 76,19% dan *RF-Grid Search* 77,06%. Penelitian lain oleh (Nainggolan & Sinaga, 2023) membandingkan RF dan *Gradient Boosting* menghasilkan akurasi RF sebesar 79% dan *Gradient Boosting* sebesar 81%. Berbagai penelitian sebelumnya, RF unggul tetapi akurasi masih kurang meskipun teknik lain seperti *Grid Search* diterapkan.

Temuan ini konsisten dengan penelitian yang dilakukan oleh (Syukron et al., 2020) yang menerapkan teknik SMOTE pada model RF dan *XGBoost*. Namun temuan mereka, nilai akurasinya masih minim, meskipun SMOTE berhasil meningkatkan akurasi dibanding model tanpa SMOTE. Performa akurasi yang diperoleh masih tergolong moderat, yaitu *SMOTE-RF* sebesar 80,97% dan SMOTE-*XGBoost* 78,63%. Selain itu, penelitian mereka belum mencakup SMOTE untuk diimplementasi ke dalam model *CatBoost*, yang secara teoretis memiliki potensi performa lebih baik karena kemampuannya menangani fitur kategori dan interaksi variabel yang kompleks. Implementasi SMOTE sejalan dengan penelitian Oktaviani et al. (2024), dimana hasil perbandingan model sebelum dan sesudah menggunakan teknik SMOTE akurasinya meningkat signifikan sehingga teknik ini efektif dalam menyeimbangkan distribusi kelas minoritas. Hasil penelitian menunjukkan bahwa kombinasi SMOTE dengan akurasi RF sebesar 82% dan *CatBoost* 81% memberikan performa lebih baik dibandingkan dengan temuan sebelumnya. Peningkatan dipengaruhi dengan SMOTE dalam mengatasi ketidakseimbangan data yang menciptakan sampel sintetis kelas minoritas, sehingga model lebih baik mendeteksi kasus positif DM. Analisis *feature importance* menunjukkan gambaran mendalam variabel utama glukosa, IMT, dan usia yang signifikan memengaruhi risiko DM.

Implementasi SMOTE dengan RF lebih unggul dibandingkan *CatBoost* karena RF pendekatannya sebagai algoritma *ensemble* yang menggabungkan banyak pohon keputusan untuk meningkatkan generalisasi dan akurasi prediksi yang tahan menghadapi masalah *overfitting*, serta andal menangani dataset berdimensi tinggi dengan fitur heterogen dan efektif menangkap interaksi *non-linear* variabel tanpa parameter *tuning* yang banyak. Sementara itu, *CatBoost* membutuhkan waktu komputasi lebih lama guna mencapai performa optimal karena proses *boosting* yang berurutan. Kombinasi SMOTE dengan model RF dan *CatBoost* dapat meningkatkan performa lebih signifikan, sehingga menjadi solusi pada bidang medis dalam memprediksi pengambilan keputusan dokter mendeteksi DM.

SIMPULAN

Temuan kombinasi teknik SMOTE dengan model *random forest* lebih optimal dan *CatBoost* menjadi pesaing kompetitif yang efektif memprediksi risiko DM dengan masing-masing akurasinya 82% dan 81% mengindikasikan performa akurasi lebih tinggi dibandingkan dengan temuan sebelumnya tanpa SMOTE. Temuan ini berhasil diimplementasikan dengan analisis mendalam *feature importance* guna mengidentifikasi variabel indikator utama Glukosa potensi signifikan terjangkit penyakit DM disusul IMT dan Usia. Kombinasi ini menjadi solusi keterbatasan temuan sebelumnya yang kurang optimal akurasinya dalam memprediksi DM, sehingga lebih efektif memprediksi performa akurasi. Hasil penelitian diharapkan membantu dokter melakukan keputusan klinis DM, sehingga mampu mencegah komplikasi lebih serius dan mengurangi biaya pengobatan pasien dikarenakan penanganan yang tepat. Penggunaan dataset ini relatif kecil, sehingga perlu penelitian lanjut dengan dataset relatif besar dan penambahan variabel beragam untuk meningkatkan keandalan model dalam penerapan klinis.

REFERENSI

Andi, A., Thamrin, T., Susanto, A., Wijaya, E., & Djohan, D. (2023). Analysis of the random forest and grid search algorithms in early detection of diabetes mellitus disease. *Jurnal Mantik*, *7*(2), 1117-1124.

Ardiansyah, M., Sunyoto, A., & Luthfi, E. T. (2021). Analisis Perbandingan Akurasi Algoritma Naïve Bayes Dan C4.5 untuk Klasifikasi Diabetes. *Edumatic: Jurnal Pendidikan Informatika*, *5*(2), 147–156. https://doi.org/10.29408/edumatic.v5i2.3424

Dennison, R. A., Chen, E. S., Green, M. E., Legard, C., Kotecha, D., Farmer, G., Sharp, S. J., Ward, R. J., Usher-Smith, J. A., & Griffin, S. J. (2021). The absolute and relative risk of type 2 diabetes after gestational diabetes: A systematic review and meta-analysis of 129 studies. *Diabetes Research and Clinical Practice*, *171*, 108625. https://doi.org/10.1016/j.diabres.2020.108625

Hidayat, T., Anelia, S. S., Pratiwi, R. I., Salsabila, N., & Prasvita, D. S. (2020). Perbandingan Akurasi Klasifikasi Penyakit Diabetes Menggunakan Algoritma Adaboost- Random Forest Dan Adaboost- Decision Tree Dengan Imputasi Median Dan Knn. *Seminar Nasional Mahasiswa Ilmu Komputer Dan Aplikasinya (SENAMIKA)*, *2*(1), 616–623.

Irwansyah, I., Kasim, I. S., & Bohari, B. (2021). The relationship between lifestyle with the risk of diabetes mellitus in staff and lecturers of universitas megarezky. *Open Access Macedonian Journal of Medical Sciences*, *9*, 198–202. https://doi.org/10.3889/oamjms.2021.5681

Li, J., Cao, Y., Liu, W., Wang, Q., Qian, Y., & Lu, P. (2019). Correlations among Diabetic Microvascular Complications: A Systematic Review and Meta-analysis. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-40049-z

Maulana, M. R., Sucipto, A., & Mulyo, H. (2020). Optimisasi Parameter Support Vector Machine dengan Particle SWARM Optimization untuk Peningkatan Klasifikasi Diabetes. *Journal Informatika Teknologi dan Sains (JINTEKS)*, *7*(2), 802–812. https://doi.org/10.51401/jinteks.v6i4.4784

Nainggolan, S. P., & Sinaga, A. (2023). Comparative Analysis of Accuracy of Random Forest and Gradient Boosting Classifier Algorithm for Diabetes Classification. *Sebatik*, *27*(1), 97–102. https://doi.org/10.46984/sebatik.v27i1.2157

Oktaviani, V., Rosmawarni, N., & Muslim, M. P. (2024). Perbandingan Kinerja Random Forest Dan Smote Random Forest Dalam Mendeteksi Dan Mengukur Tingkat Stres Pada Mahasiswa Tingkat Akhir. *Informatik : Jurnal Ilmu Komputer*, *20*(1), 43–49. https://doi.org/10.52958/iftk.v20i1.9158

Palupi, L., Ihsanto, E., & Nugroho, F. (2023). Analisis Validasi dan Evaluasi Model Deteksi Objek Varian Jahe Menggunakan Algoritma Yolov5. *Journal of Information System*

Research (JOSH), *5*(1), 234–241. https://doi.org/10.47065/josh.v5i1.4380

Resti, Y., Kresnawati, E. S., Dewi, N. R., Zayanti, D. A., & Eliyati, N. (2021). Diagnosis of diabetes mellitus in women of reproductive age using the prediction methods of naive bayes, discriminant analysis, and logistic regression. *Science and Technology Indonesia*, *6*(2), 96–104. https://doi.org/10.26554/sti.2021.6.2.96-104

Sabili, N. L., Umbara, F. R., & Melina. (2024). Klasifikasi Penyakit Diabetes Menggunakan Algoritma Categorical Boosting dengan Faktor Risiko Diabetes. *Jurnal Mahasiswa Teknik Informatika (JATI)*, *8*(6), 11391–11398. https://doi.org/10.36040/jati.v8i6.11447

Salsabil, M., Azizah, N. L,, & Eviyanti, A. (2024). Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost. *Jurnal Ilmiah Komputasi*, *23*(1), 51–58. https://doi.org/10.32409/jikstik.23.1.3507

Sari, V. R., Firdausi, F., & Azhar, Y. (2020). Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes. *Edumatic: Jurnal Pendidikan Informatika*, *4*(2), 1–9. https://doi.org/10.29408/edumatic.v4i2.2202

Siboro, O., Banjarnahor, Y. P., Gultom, A., Siagian, N. A., & Silitonga, P. D. P. (2024). Penanganan Data Ketidakseimbangan dalam Pendekatan SMOTE Guna Meningkatkan akurasi. *Seminar Nasional Inovasi Sains Teknologi Informasi Komputer (SNISTIK)*, *1*(2), 473–478.

Syandika, N. D., & Yustanti, W. (2023). Deteksi Anomali Terhadap Pembatalan Transaksi Pada Platform Tiktok Shop dengan Algoritma Categorical Boosting (Catboost). *Journal of Informatics and Computer Science (JINACS)*, *5*(02), 149–156. https://doi.org/10.26740/jinacs.v5n02.p149-156

Syukron, M., Santoso, R., & Widiharih, T. (2020). Perbandingan Metode Smote Random Forest Dan Smote Xgboost Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data. *Jurnal Gaussian*, *9*(3), 227–236. https://doi.org/10.14710/j.gauss.v9i3.28915

Tarigan, L. R. A., & Dahlan. (2024). Optimalisasi Fitur dengan Forward Selection pada Estimasi Tingkat Penyakit Paru-Paru Menggunakan Algoritma Klasifikasi Random Forest. *Jurnal Mahasiswa Teknik Informatika (JATI)*, *8*(5), 10341–10348. https://doi.org/10.36040/jati.v8i5.11064

Taufik, I., & Kurniawan, A. A. (2023). Peran Artificial Intelligencedalam Inovasi Digital Marketing. *Seminar Nasional Ilmu, Manajemen, Ekonomi, Keuangan Dan Bisnis (SNIMEKB)*, *2*(1), 29–40. https://doi.org/10.55927/snimekb.v2i1.4602

Wang, L., Wang, X., Chen, A., Jin, X., & Che, H. (2020). Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model. *Healthcare (Switzerland)*, *8*(3), 1–11. https://doi.org/10.3390/healthcare8030247

Zhong, W., Zhang, D., Sun, Y., & Wang, Q. (2023). A CatBoost-Based Model for the Intensity Detection of Tropical Cyclones over the Western North Pacific Based on Satellite Cloud Images. *Remote Sensing*, *15*(14). https://doi.org/10.3390/rs15143510

Zulfiansyah, A. D. K., Kusuma, H., & Attamimi, M. (2023). Rancang Bangun Sistem Pendeteksi Keaslian Uang Kertas Rupiah Menggunakan Sinar UV dengan Metode Machine Learning. *Jurnal Teknik ITS*, *12*(2). https://doi.org/10.12962/j23373539.v12i2.118320