MIND (Multimedia Artificial Intelligent Networking Database) Journal

MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 10 | No. 2 | Halaman 221 - 234
DOI : https://doi.org/10.26760/mindjournal.v10i2.221-234 Desember 2025

Prediksi Penyakit Diabetes menggunakan Teknik Imputasi *Missforest* dan Klasifikasi *LightGBM*

ALDOVA FERDIANSYAH, FAJRI RAKHMAT UMBARA, FATAN KASYIDI

Universitas Jenderal Achmad Yani, Cimahi, Indonesia
Email : aldovaferdiansyah12@gmail.com

Received 14 Agustus 2025 **|** *Revised* 25 September 2025 **|** *Accepted* 12 Oktober 2025

ABSTRAK

Diabetes adalah salah satu penyakit kronis dengan grafik prevalensinya meningkat secara global. Penyakit ini disebabkan oleh gangguan metabolisme tubuh yang memengaruhi kadar gula darah, dan jika tidak ditangani sejak dini dapat menimbulkan komplikasi serius seperti stroke, gagal ginjal, kebutaan, hingga kematian. Penelitian ini mengembangkan model prediksi risiko diabetes berbasis klasifikasi biner menggunakan algoritma LightGBM yang dikombinasikan dengan teknik imputasi Missforest untuk menangani data yang hilang. Dataset yang digunakan berasal dari Pima Indian, tersedia secara publik di Kaggle. Tahapan pre-processing mencakup imputasi data hilang, penanganan outlier dengan Isolution Forest, pembagian data menjadi 80:20. Evaluasi model menunjukkan hasil akurasi sebesar 91,84% dan ROC AUC 0.9614. BMI menjadi faktor paling berpengaruh dalam prediksi yang diikuti oleh DiabetesPedigreeFunction dan Glucose.

***Kata kunci**: diabetes melitus, data mining, klasifikasi, LightGBM, missforest*

ABSTRACT

Diabetes mellitus is one of the most common chronic diseases, with a globally increasing prevalence. It is caused by metabolic disorders that affect blood glucose levels and, if not treated early, can lead to serious complications such as stroke, kidney failure, blindness, and even death. This research develops a diabetes risk prediction model based on binary classification using the LightGBM algorithm combined with the Missforest imputation technique to handle missing data. The dataset used is the publicly available Pima Indian dataset from Kaggle. The pre-processing stages include missing value imputation, outlier handling using Isolution Forest, an 80:20 data split. Model evaluation shows an accuracy of 91.84% and a ROC AUC 0.9614. BMI was found to be the most influential factor in the prediction, followed by DiabetesPedigreeFunction and Glucose.

***Keywords**: diabetes mellitus, data mining, classification, LightGBM, missforest*

1. PENDAHULUAN

Diabetes mellitus merupakan penyakit metabolik yang ditandai dengan meningkatnya kadar gula darah akibat gangguan hormon insulin dalam menjaga kestabilan glukosa tubuh **(*American Diabetes Association*, 2017)**. Penyakit ini menyerang anak-anak hingga orang dewasa, dan beresiko menimbulkan komplikasi yang serius seperti stroke, amputasi, gagal ginjal, kebutaan, bahkan kematian **(Tomic, Shaw, and Magliano, 2022)**. *Internasional Diabetes Federation* (IDF) mencatat bahwa pada tahun 2021 terdapat kurang lebih 537 juga penderita diabetes berusia 20-79 tahun di dunia atau 10,5% dari populasi global kelompok usia tersebut, dan hampir separuhnya tidak menyadari bahwa mereka mengidap penyakit diabetes. Diperkirakan angka ini akan terus meningkat menjadi 637 pada tahun 2030 dan tahun 2045 mencapai 786. Di kawasan Asia Tenggara sendiri kasus kematian yang diakibatkan penyakit diabetes ini mencapai 747.000 jiwa pada tahun 2021, dengan estimasi peningkatan jumlah penderita menjadi 152 juta jiwa pada tahun 2045, atau naik sebesar 68% dari tahun-tahun sebelumnya.

Berbagai penelitian sebelumnya telah mencoba memprediksi penyakit diabetes dengan menggunakan pendekatan klasifikasi berbasis *data mining*. Salah satunya studi menggunakan metode *Support Vector Machine (SVM)* dan *Naïve Bayes (NB)* dengan memperoleh akurasi masing-masing sebesar 78,04% dan 76,98% **(Maulidah, dkk, 2021)**. Studi lain menggunakan kombinasi *SVM* dengan *kernel Radial Basis Function* dan metode seleksi atribut *forward selection*, menghasilkan akurasi sebesar 91,2% **(Hovi, Id Hadiana, and Rakhmat Umbara, 2022)**. Penelitian menggunakan *algoritma C4.5* juga menunjukkan hasil yang cukup baik dengan akurasi mencapai 90.00% **(Ucha Putri, dkk, 2021)**. Sementara itu, pada penelitian lainnya menunjukkan bahwa metode *LightGBM* memiliki performa tinggi dibandingkan dengan metode klasifikasi lainnya seperti *SVM, KNN, NB, XGBoost, Random Forest*, dan *Bagging*, dengan akurasi mencapai 98,1% **(Rufo, dkk, 2021)**. Namun, sebagian besar penelitian tersebut belum memanfaatkan teknik imputasi canggih untuk menangani data hilang secara optimal, yang padahal hal tersebut sangat penting dalam meningkatkan kualitas pembelajaran model.

Masalah umum dalam penggunaan dataset adalah masih ditemukannya beberapa nilai kosong (*missing values*) pada sebagian atribut di dalam dataset. Teknik imputasi *Missforest* merupakan salah satu metode berbasis *Random Forest* yang terbukti efektif dalam mengisi data hilang secara iteratif, baik untuk data numerik maupun kategorikal, namun teknik imputasi *Missforest* ini masih belum banyak diterapkan dalam penelitian yang serupa. Maka dari itu, pada penelitian ini bertujuan mengembangkan model prediksi penyakit diabetes dengan menggabungkan metode klasifikasi *LightGBM* serta teknik imputasi *Missforest* dalam penanganan data yang hilang dalam dataset, serta penelitian ini memiliki tujuan lain yaitu diharapkan dapat meningkatkan akurasi pada penelitian-penelitian sebelumnya, khususnya pada penelitian **(Maulidah, dkk, 2021)**.

Penelitian ini dibatasi pada penggunaan satu teknik imputasi saja yaitu teknik imputasi *Missforest*, tanpa terdapatnya perbandingan dengan teknik imputasi lainnya, serta hanya menggunakan satu model klasifikasi yaitu model klasifikasi *LightGBM*, jadi tidak terdapat perbandingan model klasifikasi yang digunakan yaitu *LightGBM* dengan model klasifikasi lainnya seperti *KNN, XGBoost, Random Forest, Support Vector Machine, Naïve Bayes* dan lainnya. Model yang dibangun diharapkan dapat memberikan prediksi yang lebih akurat, cepat, serta dapat diandalkan untuk membantu deteksi dini risiko seseorang terpapar atau terkena penyakit diabetes mellitus.

Pada penelitian ini istilah prediksi digunakan dalam konteks *machine learning*, yaitu membangun model untuk memperkirakan kelas (label) suatu data baru. Dengan demikian, prediksi yang dimaksud bukanlah untuk meramalkan kejadian di masa depan, melainkan proses klasifikasi biner untuk menentukan apakah seseorang berisiko diabetes atau tidak.

1.1. *Data Mining*

Data mining adalah proses penggalian suatu pengetahuan atau sebuah informasi penting yang bernilai dari kumpulan data dalam basis data yang berukuran besar. Teknik ini bertujuan untuk menambang data secara sistematis guna memperoleh informasi yang relevan melalui proses penyaringan yang lebih terstruktur dan akurat, sehingga memungkinkan analisis data yang lebih efektif **(Candra Permana and Dewi Patwari, 2021; Hovi, Id Hadiana, and Rakhmat Umbara, 2022)**. Sejak kemunculannya pada awalan tahun 1990-an, *data mining* ini telah menjadi pendekatan yang efektif untuk menganalisis pola hubungan antar data serta mengelompokkan objek-objek ke dalam *cluster* tertentu, yang dimana objek dalam satu *cluster* memiliki tingkat kemiripan yang lebih tinggi dibandingkan dengan objek dalam *cluster* lainnya **(Ucha Putri, dkk, 2021)**. Beberapa teknik utama dalam *data mining* yang umum digunakan meliputi aturan asosiasi, klasifikasi, pengelompokkan *(clustering)*, prediksi dan model sekuensial **(Derisma, 2020)**.

1.2. Teknik Imputasi *Missforest*

Data hilang dapat diartikan sebagai ketidaksesuaian antara data yang seharusnya dikumpulkan dengan data yang berhasil diperoleh selama proses pengumpulan data berlangsung. Kondisi ini merujuk pada situasi dimana hasil pengamatan tidak tersedia untuk atau lebih variabel yang dibutuhkan dalam analisis. Data hilang sering kali disebabkan oleh berbagai faktor, seperti kegagalan dalam mengamati sejumlah individu atau entitas tertentu, kesalahan saat memasukkan data, atau kekeliruan dalam proses pengeditan data. Pada dataset yang digunakan dalam penelitian ini *https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database*, keberadaan data yang hilang seraca signifikan dapat menyebabkan bias dalam proses pengolahan data. Maka dari itu, diperlukan metode imputasi yang tepat untuk mengatasi permasalahan tersebut **(Alfebi and Anasanti, 2023; Bemi and Nooraeni, 2019)**.

Metode imputasi *Missforest* memanfaatkan algoritma *Random Forest* untuk menangani imputasi data *fenotipik*. Prosesnya melibatkan pelatihan algoritma *Random Forest* pada data yang tersedia untuk setiap variabelnya. Metode ini menggunakan pendekatan *iteratif* guna memperkirakan nilai-nilai yang hilang hingga memenuhi persyaratan penghentian. Selain itu, *Missforest* dapat dijalankan secara paralel untuk mempercepat proses komputasi dan mengevaluasi kesalahan imputasi menggunakan metode *Out-of-Bag (OOB)* baik pada data kontinu maupun kategorikal **(Novianto and Anasanti, 2023)**.

1.3. Klasifikasi *LightGBM*

LightGBM merupakan algoritma berbasis *Gradient Boosting Decision Tree (GBDT)* yang dirancang untuk memberikan efisiensi tinggi dalam memproses data berskala besar. Algoritma ini dikembangkan oleh *Microsoft Research Asia* untuk meningkatkan kecepatan pelatihan, efisiensi komputasi, serta penggunaan memori yang lebih rendah dibandingkan metode *GBDT* lainnya. Selain itu, *LightGBM* mendukung pembelajaran paralel, penggunaan *GPU*, serta mampu menangani berbagai tugas pembelajaran mesin seperti klasifikasi, regresi, dan prediksi. Dengan kemampuan klasifikasinya yang unggul, algoritma ini telah banyak digunakan dalam analisis klinis untuk mendiagnosis penyakit dan memprediksi hasil kesehatan, serta memberikan kontribusi signifikan dalam pencegahan dan pengendalian penyakit seperti diabetes **(Hou, dkk, 2020; Septiana Rizky, Haiban Hirzi, and Hidayaturrohman, 2022; Wardhana, dkk, 2022)**.

Model *LightGBM* secara otomatis menangani perhitungan iterasi serta pembaruan yang terkait dengan Persamaan (1) – (2) :

$$\hat{y}^{(t)} = \hat{y}\,(\mathrm{t}-1) + \mathrm{fi}(\mathrm{xi}) \quad (1)$$

Pada Persamaan (1), *LightGBM* diterapkan sebagai algoritma *Gradient Boosting* yang berfokus pada pengoptimalan fungsi kerugian *(loss function)* secara iteratif dengan menambahkan pohon keputusan baru pada setiap iterasi **(Septiana Rizky, Haiban Hirzi, and Hidayaturrohman, 2022)**.

$$\Omega(\mathrm{f}) = \gamma\mathrm{T} + \frac{\lambda}{2}||\omega||^2 \quad (2)$$

Pada Persamaan (2), algoritma *LightGBM* mengontrol kompleksitas model dan mencegah kemungkinan terjadinya *overfitting* melalui penalti terhadap pohon keputusan yang terbentuk **(Septiana Rizky, Haiban Hirzi, and Hidayaturrohman, 2022)**.

1.4. *K-Fold Cross Validation*

K-fold cross validation merupakan salah satu metode evaluasi model yang bertujuan untuk mengukur kemampuan generalisasi model secara menyeluruh dan objektif. Teknik ini bekerja dengan membagi data pelatihan menjadi k bagian *(fold),* kemudian model akan dilatih sebanyak k kali dengan menggunakan k-1 bagian sebagain data latih dan satu bagian lainnya sebagai data validasi yang dilakukan secara bergantian. Seluruh hasil evaluasi dari setiap iterasi kemudian dirata-ratakan untuk menghasilkan metrik performa akhir. Metode ini sangat berguna untuk mencegah terjadinya *overfitting* serta dapat memberikan evaluasi yang stabil, terutama saat dataset yang digunakan terbatas. Hal ini menunjukkan bahwa *k-fold cross validation* mampu untuk memberikan penilaian yang adil terhadap semua data karena setiap *fold* akan berperan sebagai data uji tepat satu kali. Maka dari itu, metode ini menjadi salah satu pendekatan yang direkomendasikan dalam evaluasi kinerja model klasifikasi **(Fuadah, dkk, 2022; Tuntun, Kusrini, and Kusnawi, 2022)**.

1.5. Evaluasi *Confusion Matrix*

Confusion Matrix merupakan metode untuk keperluan evaluasi yang umum digunakan dalam mengukur kinerja model klasifikasi, terutama pada kasus dengan dua atau lebih kelas keluaran. *Matrix* ini berbentuk tabel yang terdiri dari empat kombinasi nilai yang mencerminkan hasil prediksi model terhadap data aktual. Empat komponen utama dalam *Confusion Matrix* adalah sebagai berikut *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN). True Negative (TN)* mengacu pada data negatif yang berhasil diklasifikasikan dengan benar oleh model, sedangkan *False Positive (FP)* adalah data negatif yang secara keliru diklasifikasikan sebagai positif. Selanjutnya, *True Positive (TP)* menunjukkan data positif yang diklasifikasikan dengan benar, sementara *False Negative (FN*) merupakan data positif yang keliru diklasifikasikan sebagai negatif kebalikan dari True Positive **(Hovi, Id Hadiana, and Rakhmat Umbara, 2022)**.

Setelah mendapatkan data hasil proses klasifikasi dari *Confusion Matrix*, maka data akan dikalkulasikan dalam perhitungan tiap nilainya dengan *accuracy, precision, recall, dan F-1 Score.*

a. Perhitungan *Accuracy*

Accuracy ini menggambarkan seberapa akuratnya model yang akan digunakan dalam proses klasifikasi. Perhitungan *accuracy* dapat dilihat pada Persamaan (3).

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \text{ X } 100\% \quad (3)$$

b. Perhitungan *Precision*

Precision ini memperlihatkan nilai akurasi antara sebuah data yang diminta dengan hasil prediksi yang dihasilkan model. Perhitungan *precision* dapat dilihat pada Persamaan (4).

$$\text{Precision} = \frac{TP}{(TP+FP)} \text{ X } 100\% \quad (4)$$

c. Perhitungan *Recall*

Recall ini memperlihatkan keberhasilan model dalam menentukan sebuah informasi. Perhitungan *recall* dapat dilihat pada Persamaan (5).

$$\text{Recall} = \frac{TP}{(TP+FN)} \text{ X } 100\% \quad (5)$$

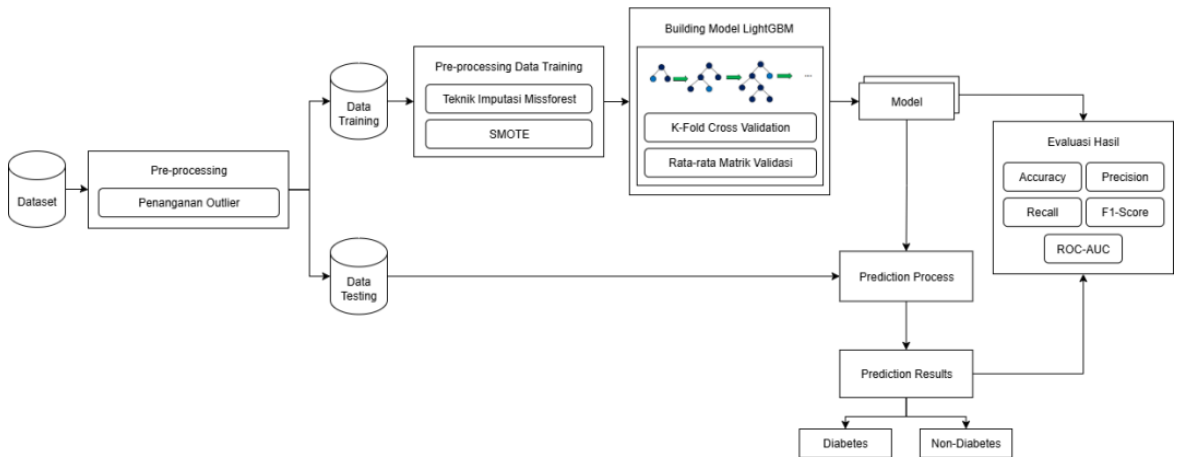
d. Perhitungan *F-1 Score*

F-1 Score ini memperlihatkan sebuah perbandingan nilai rata-rata precision dan recall yang dibobotkan. Perhitungan *F-1 Score* dapat dilihat pada Persamaan (6).

$$\text{F-1 Score} = \frac{(2\ x\ precision\ x\ recall)}{(precision+recall)} \text{ X } 100\% \quad (6)$$

2. METODOLOGI

Penelitian yang dilakukan yaitu untuk memprediksi penyakit diabetes serta meningkatkan akurasi dari penelitian sebelumnya, pada penelitian yang serupa menggunakan metode yang berbeda dengan menghasilkan tingkat akurasi untuk *SVM* sebesar 78% dan *NB* sebesar 76%. Dataset ini akan diperiksa terlebih dahulu apakah terdapat *missing value* atau tidak, apabila terdeteksi terdapatnya *missing value* maka dataset akan dilakukan tahapan teknik imputasi menggunakan *Missforest.* Setelah itu file model yang dihasilkan dari *building model LightGBM* akan diproses untuk menghasilkan suatu bagan yang menunjukkan atribut mana yang paling mempengaruhi prediksi diabetes. Model prediksi yang dibangun dalam penelitian ini menggunakan algoritma *LightGBM* dengan pendekatan klasifikasi biner, yang artinya keluaran dari model ini bukan berupa nilai kontinu, tetapi kategori "diabetes" dan "non-diabetes", seperti metode penelitian pada Gambar 1.



Gambar 1. Metode Penelitian

2.1. *Pre-Processing*
Pada tahapan awal adalah mengidentifikasi terdapatnya data ekstrem atau *outlier* di dalam dataset, yang dimana penanganan *outlier* yang dilakukan menggunakan algoritma *Isolution Forest* penanganan ini berfungsi untuk menghindari pengaruh nilai ekstrem dalam dataset terhadap hasil prediksi diabetes serta untuk meningkatkan akurasi model.

2.2. *Splitting* Data
Setelah proses *pre-processing* selesai, yaitu mencakup penanganan data hilang dan penanganan *outlier*, dataset disimpan dan telah siap untuk digunakan dalam proses pelatihan model *LightGBM*. Sebelum dilatih, data perlu dibagi menjadi dua bagian terlebih dahulu, yaitu 80% sebagai data pelatihan dan 20% sebagai data pengujian dari total keseluruhan 768 data.

Pembagian data dengan menggunakan rasio 80:20 ini bukan dilakukan secara acak melainkan didasarkan pada hasil penelitian yang dilakukan oleh **(Gholamy, Kreinovich, and Kosheleva, 2018)**. Penelitian tersebut menyatakan bahwa pembagian 70:30 atau 80:20 merupakan pilihan yang ideal karena mampu memberikan estimasi kinerja model yang akurat dan stabil. Rasio ini juga membantu mengurangi risiko *overfitting*, sehingga hasil prediksi model menjadi lebih andal dan seimbang.

2.3. *Pre-Processing Data Training*
Setelah proses *splitting* dataset menjadi data pelatihan dan penguji berhasil dilakukan, maka langkah berikutnya adalah mengecek keberadaan data yang hilang dalam data pelatihan serta melakukan penyeimbangan data. Hal ini penting untuk mencegah terjadinya *data leakage*, yaitu kebocoran informasi dari data pengujian ke dalam proses pelatihan model. Menurut **(Demircioğlu, 2024)**.

Meskipun seluruh kolom dalam data pelatihan terlihat telah terisi, namun beberapa atribut seperti *Glucose, BloodPressure, SkinThickness, Insulin*, dan *BMI* tidak seharusnya memiliki nilai 0. Oleh karena itu, nilai pada atribut-atribut tersebut harus diubah menjadi bentuk *NaN* atau *Null* agar dapat diidentifikasi sebagai data hilang. Proses ini memungkinkan sistem untuk dapat menghitung jumlah data yang hilang secara akurat sebelum dilakukannya tahapan imputasi data. Setelah data hilang dapat teridentifikasi, maka proses dilanjukan dengan imputasi menggunakan metode *Missforest*, yang memanfaatkan algoritma *Random Forest* untuk memperkirakan nilai yang hilang secara iteratif. Setelah melalui tahap pengisian data hilang menggunakan imputasi *Missforest* selanjutnya adalah memeriksa apakah data pelatihan memiliki distribusi kelas yang seimbang. Jika ditemukan ketidakseimbangan antara kelas mayoritas dan minoritas maka diperlukan penanganan agar saat model menggunakan data untuk pelatihan tidak akan menimbulkan bias atau model lebih memprioritaskan kelas yang lebih unggul dan mengabaikan kelas yang minoritas.

2.4. *Building Model LightGBM*
Selanjutnya, model dilatih menggunakan algoritma *LightGBM* yang membangun prediksi secara bertahap melalui pendekatan *boosting*, dimana setiap iterasi akan memperbaiki kesalahan dari iterasi sebelumnya yang dimana hal ini dilakukan oleh rumus nomor 1. Untuk menghindari terjadinya *overfitting*, digunakan mekanisme penalti kompleksitas model yang dilakukan oleh rumus nomor 2. Selain itu, digunakan teknik *K-fold Cross Validation* guna memastikan model mampu melakukan generalisasi dengan baik. Teknik ini membagi data pelatihan menjadi beberapa bagian dan menjalankan pelatihan serta validasi secara bergantian untuk memperoleh rata-rata performa model yang lebih stabil dan akurat.

2.5. *Prediction Process*

Pada tahapan ini adalah memproses hasil pembelajaran data yang dihasilkan oleh *building model LightGBM* untuk menghasilkan suatu prediksi, yang nantinya model tersebut akan digabungkan dengan data *testing* yang dihasilkan pada tahapan *splitting* dataset dalam bentuk *dataframe*, yang nantinya *dataframe* hasil penggabungan data *testing* dan data hasil pembelajaran model *(model final)* akan digunakan untuk menampilkan suatu prediksi atribut mana yang paling mempengaruhi terhadap prediksi diabetes. Output dari prediksi akan berbentuk bagan dengan menunjukkan atribut mana yang paling mempengaruhi terhadap prediksi diabetes.

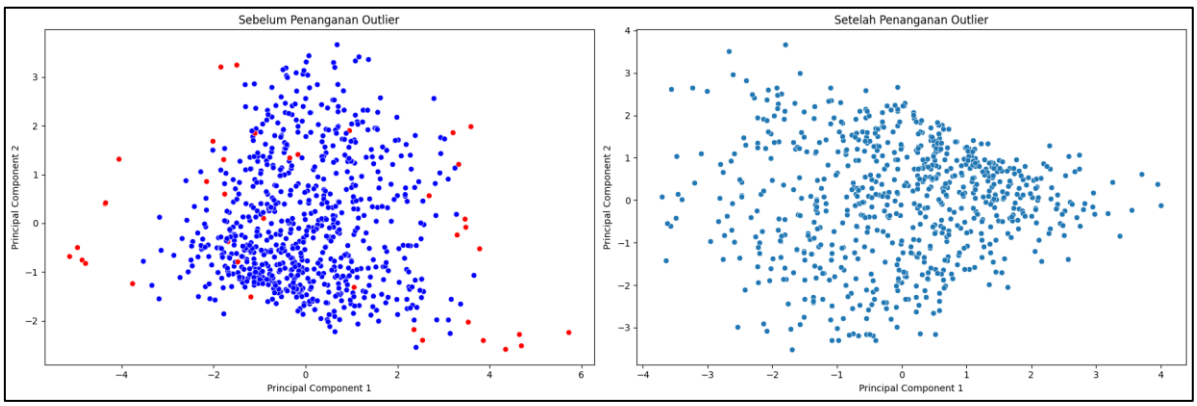
2.6. Evaluasi Hasil

Setelah hasil prediksi diperoleh, evaluasi model *LightGBM* dilakukan menggunakan *Confusion Matrix* yang nantinya akan menghasilkan empat komponen utama yaitu *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN).* Berdasarkan nilai-nilai perhitungan ini, kinerja model dapat dinilai menggunakan perhitungan *matrix*. *Accuracy* menunjukkan tingkat keseluruhan prediksi yang benar, *precision* mengukur ketepatan prediksi positif, *recall* menilai kemampuan model dalam mendeteksi semua kasus positif, dan *F-1 Score* bertujuan untuk memberikan keseimbangan antara *precision* dan *recall*, terutama penting saat data tidak seimbang. Evaluasi ini penting untuk dapat memastikan apakah model dapat memprediksi penyakit diabetes secara akurat dan andal atau tidak.

3. HASIL DAN PEMBAHASAN

3.1. *Pre-processing*

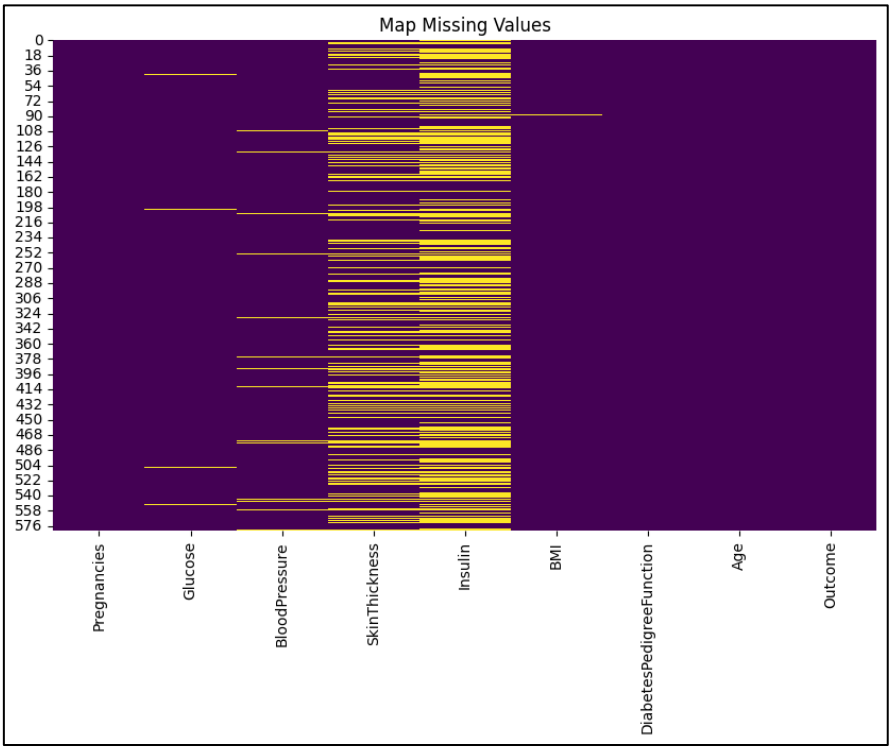
Pada tahapan awal adalah mendeteksi terdapatnya *outlier* atau data ekstrem di dalam dataset yang dimana data *outlier* harus ditangani terlebih dahulu sebelum masuk dalam tahapan pembelajaran *machine learning*. Pada Gambar 2 menunjukkan perbandingan distribusi data sebelum dan sesudah penanganan *outlier* yang menggunakan metode deteksi dan pembersihan data yaitu *Isolution Forest*. Pada grafik sebelah kiri, terdapat beberapa titik berwarna merah yang menandakan keberadaan *outlier* yang menyebar jauh dari pusat data utama. Keberadaan *outlier* ini dapat memengaruhi kualitas model nantinya, karena berpotensi menimbulkan bias pada proses pembelajaran. Setelah dilakukan penanganan, seperti terlihat pada grafik sebelah kanan, data menjadi lebih terpusat dan *homogeny* tanpa adanya titik ekstrem. Hal ini menunjukkan bahwa distribusi data lebih representative, sehingga diharapkan dapat meningkatkan kinerja model dan mengurangi resiko terjadinya *overfitting*.



Gambar 2. Penanganan *Outlier*

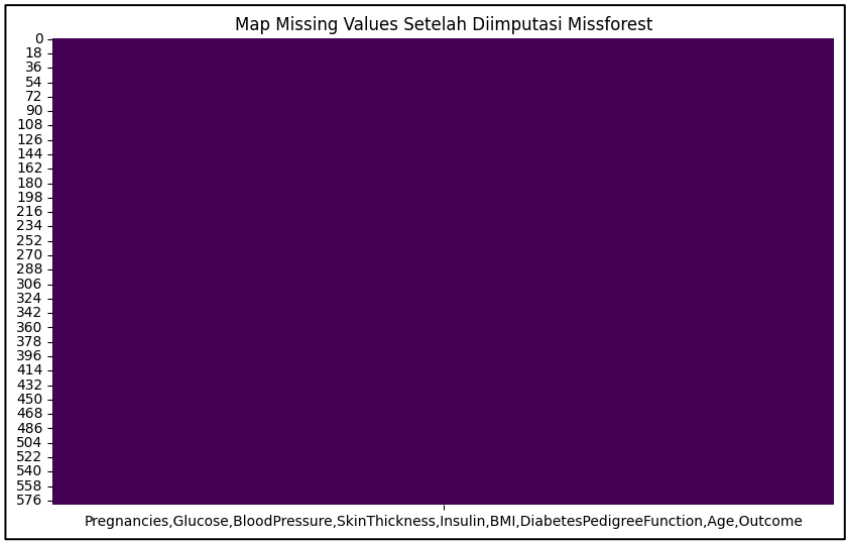
3.2 Penanganan *Missing Value*

Data *training* terlebih dahulu dilakukan pengecekan terhadap keberadaan *missing value*. Meskipun awalnya terlihat tidak terdapat nilai yang hilang, ditemukan bahwa terdapat beberapa atribut dalam dataset yang memiliki nilai 0, secara logika seharusnya tidak mungkin bernilai 0 seperti atribut *glucose* dan *blood pressure*. Oleh karena itu, nilai 0 pada atribut-atribut tersebut diubah menjadi *NaN* agar dapat diidentifikasi sebagai *missing value.*



Gambar 3. Sebelum Penanganan Imputasi Data

Visualisasi pada Gambar 3 menunjukkan terdapatnya *missing value* yang ditandai dengan garis-garis berwarna kuning. Untuk menangani hal tersebut digunakan metode imputasi *Missforest* dengan berbasis algoritma *Random Forest*.

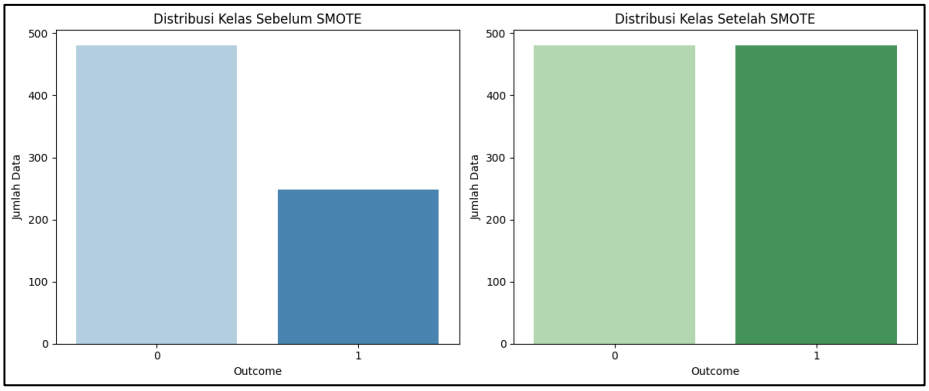


Gambar 4. Setelah Penanganan Imputasi Data

Setelah proses imputasi selesai, visualisasi pada Gambar 4 menunjukkan bahwa tidak terdapat lagi garis-garis kuning, yang dimana hal tersebut menandakan bahwa dataset telah bersih dari *missing value,* dan secara tidak langsung menandakan bahwa data telah siap untuk masuk ke dalam tahapan selanjutnya yaitu penyeimbangan data tidak seimbang.

3.3. Penanganan Data Tidak Seimbang

Penanganan data tidak seimbang hanya dilakukan pada data *training* saja dimana hal tersebut dilakukan untuk menghindari terjadinya *data leakage*. Data *training* ditemukan memiliki distribusi kelas yang tidak seimbang, dimana kelas 0 (non-diabetes) jauh lebih dominan dibandingkan dengan kelas 1 (diabetes). Ketimpangan ini berpotensi menimbulkan bias dalam proses pelatihan model, karena algoritma cenderung mempelajari pola dari kelas mayoritas saja.



Gambar 5. Visualisasi Sebelum dan Sesudah SMOTE

Untuk mengatasi hal tersebut, digunakan metode *Synthetic Minority Over-sampling Technique (SMOTE)* yang menghasilkan sistetik untuk kelas minoritas tanpa melakukan duplikasi langsung. Pada Gambar 5 memperlihatkan visualisasi distribusi label sebelum (kiri) dan sesudah (kanan) penerapan *SMOTE*. Terlihat bahwa setelah penyeimbangan, jumlah data kelas 0 dan 1 menjadi seimbang. Dengan distribusi data yang lebih proporsional, model diharapkan dapat melakukan klasifikasi dengan lebih adil dan meningkatkan performa, khususnya dalam hal *recall* dan *f-1 score* untuk kelas minoritas.

3.4. Implementasi Model *LightGBM*

Model *LightGBM* berhasil diterapkan pada data pelatihan *(training)* yang dimana data *training* ini telah melalui berbagai macam tahapan seperti proses *pre-processing* dan penyeimbangan kelas menggunakan *smote*, implementasi model *LightGBM* sendiri menerapkan beberapa parameter utama seperti *boosting_type='gbdt', objective='binary',* serta *regularisasi reg_alpha=1.5* dan *reg_lambda=1.5* yang digunakan untuk mengurangi risiko terkenanya *overfitting*. Evaluasi performa model sendiri menggunakan metode *k-fold cross validation* dengan menggunakan lipatan atau nilai k sebanyak k=5.

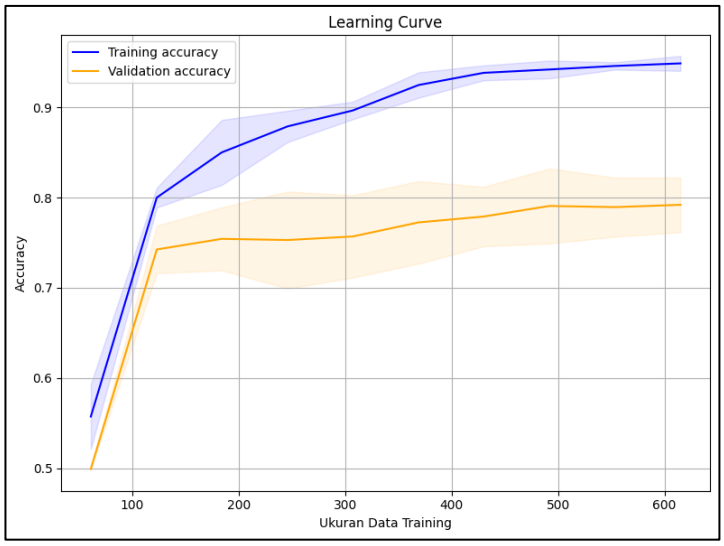
Hasil evaluasi model dapat dilihat pada Tabel 1, yang menunjukkan bahwa model mampu memberikan performa yang konsisten disetiap foldnya, dengan rata-rata nilai *accuracy* sebesar 79.20%, *precision* sebesar 77.67%, *recall* sebesar 81.81%, dan *f-1 score* sebesar 79.62%. Nilai *accuracy* yang tergolong tinggi menunjukkan kemampuan model dalam mengklasifikasikan data secara umum dengan sangat baik. Sementara itu, nilai *precision* dan *recall* yang tinggi menandakan bahwa model tidak hanya akurat dalam memprediksi kelas positif (diabetes), tetapi juga mampu mendeteksi sebagian besar kasus positif secara efektif.

Rata-rata *f-1 score* yang mencapai 79.62% juga menunjukkan keseimbangan yang optimal antara *precision* dan *recall.* Konsistensi performa ini mencerminkan bahwa model memiliki kemampuan generalisasi yang baik serta menandakan tidak terjadinya *overfitting*, sehingga layak digunakan pada tahapan pengujian dengan data testing yang belum pernah dilihat sebelumnya pada tahapan pembelajaran model.

Tabel 1. Hasil *K-Fold Cross Validation*

Fold	***Accuracy***	***Precision***	***Recall***	***F-1 Score***
1	0.811688	0.779070	0.870130	0.822086
2	0.824675	0.812500	0.844156	0.828025
3	0.753247	0.740741	0.779221	0.759494
4	0.727273	0.739726	0.701299	0.720000
5	0.843137	0.811765	0.896104	0.851852
Rata-rata	0.792004	0.776760	0.818182	0.796291

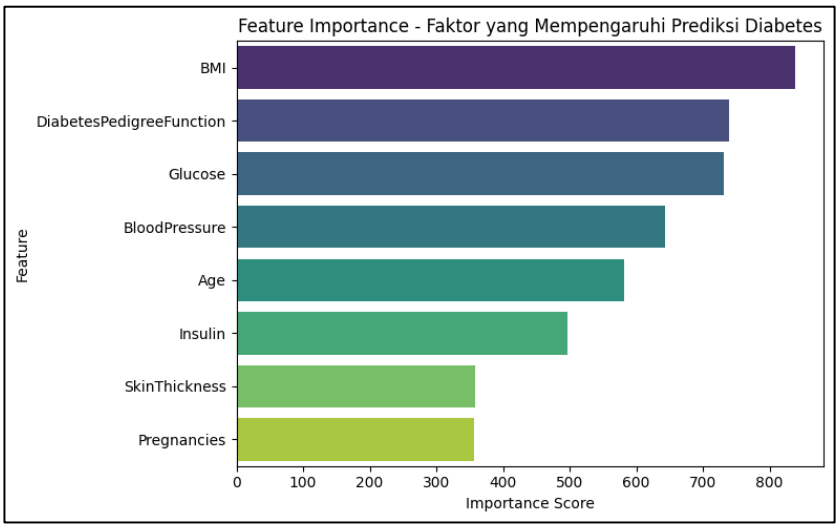
3.5. *Learning Curve* Model



Gambar 6. *Learning Curve Model*

Model *LightGBM* dievaluasi terhadap risiko *overfitting* dan *underfitting* dengan menggunakan grafik *learning curve* yang dapat dilihat pada Gambar 6. Grafik tersebut menunjukkan bahwa akurasi pada data *training* meningkat secara konsisten dari sekitar 55% hingga mencapai 95%, sementara akurasi validasi juga meningkat dari kisaran 50% hingga sekitar kisaran 80%. Meskipun terdapat selisih antara akurasi training dan testing, keduanya cenderung meningkat seiring bertambahnya jumlah data yang digunakan. Hal ini menandakan bahwa model mampu belajar dengan baik dari data tanpa mengalami *overfitting* yang berlebihan. Temuan ini sejalan dengan **(Halabaku and Bytyçi, 2024)**, yang menjelaskan bahwa overfitting umumnya terjadi ketika selisih akurasi training dan validasi sangat besar, seperti pada kasus *decision tree* yang mencapai nilai 99.84% untuk akurasi training sedangkan 77.30% untuk akurasi validasi. Maka dari itu, berdasarkan hasil *learning curve* pada Gambar 6, dapat disimpulkan bahwa model memiliki kemampuan generalisasi yang baik dan siap untuk digunakan pada tahapan selanjutnya yaitu prediksi.

3.6. *Prediction Result*

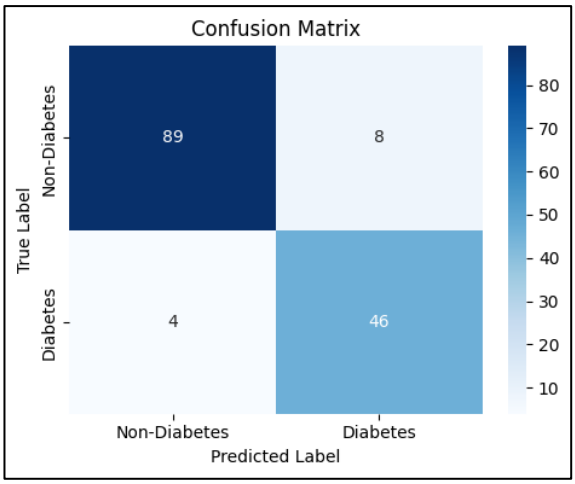


Gambar 7. Hasil Prediksi

Hasil prediksi model *LightGBM* terhadap data pengujian menunjukkan bahwa atribut *BMI* memiliki pengaruh paling dominan dalam menentukan klasifikasi diabetes, yang diikuti oleh *DiabetesPedigreeFunction* dan *Glucose*, hal tersebut berdasarkan pada Gambar 7 yang menunjukkan atribut dalam dataset yang paling mempengaruhi terhadap prediksi diabetes. Atribut lain seperti *BloodPressure, Age, Insulin, SkinThickness,* dan *Pregnancies* juga memberikan kontribusi meskipun tergolong rendah. Identifikasi atribut paling mempengaruhi oleh model ini tidak hanya membantu meningkatkan akurasi prediksi saja, tetapi juga dapat memberikan pemahaman medis yang lebih mendalam terhadap faktor risiko diabetes dalam konteks data.

3.7. Evaluasi Model

Evaluasi akhir terhadap model *LightGBM* menggunakan 147 data uji dengan menunjukkan performa yang sangat baik, dengan akurasi mencapai 91.84% dan nilai *ROC AUC* sebesar 0.9614, yang dimana hal tersebut menggambarkan kemampuan model tergolong tinggi dalam membedakan antara kelas diabetes dan non-diabetes.



Gambar 8. Hasil *Confusion Matrix*

Berdasarkan visualisasi *Confusion Matrix* pada Gambar 8, menunjukkan bahwa model berhasil mengklasifikasikan 89 dari 97 data non-diabetes dan 46 dari 50 data diabetes dengan benar, dengan menunjukkan tingkat kesalahan yang sangat rendah. Hasil ini mengkonfirmasi bahwa model yang telah dibangun memiliki kemampuan generalisasi yang kuat dan layak digunakan untuk prediksi diabetes pada data baru (data uji).

Tabel 2. Hasil Evaluasi Model (*Classification Report*)

Label	***Precision***	***Recall***	***F1-Score***	***Support***
0 (Non-Diabetes)	0.96	0.92	0.94	97
1 (Diabetes)	0.85	0.92	0.88	50
Accuracy			0.92	147
Macro Avg	0.90	0.92	0.91	147
Weighted Avg	0.92	0.92	0.92	147

Berdasarkan hasil evaluasi pada Tabel 2, model *LightGBM* menunjukkan performa klasifikasi yang sangat baik dengan diterapkannya teknik *imputasi Missforest*. *Precision* mencapai 0.96 untuk kelas non-diabetes dan 0.85 untuk kelas diabetes, dengan *recall* sebesar 0.92 pada kedua kelas, *f1-score* juga menghasilkan nilai yang tinggi yitu 0.94 untuk kelas non-diabetes dan 0.88 untuk kelas diabetes. *Accuracy* ini menunjukkan bahwa penggabungan *Missforest* dan *LightGBM* dalam penelitian dapat memberikan dampak positif terhadap performa model, terutama dalam meningkatkan kualitas data yang digunakan sehingga dapat menghasilkan prediksi yang lebih akurat dan seimbang antar kelas.

4. KESIMPULAN

Penelitian ini berhasil membangun model prediksi berbasis klasifikasi untuk risiko terpapar penyakit diabetes serta berhasil meningkatkan akurasi prediksi penyakit diabetes dibandingkan penelitian sebelumnya dengan membangun model berbasis data suku *Pima Indian*. Permasalahan data seperti nilai nol tidak valid, outlier, dan ketidakseimbagan kelas telah berhasil diatasi dengan menggunakan *Missforest*, *Isolution Forest* dan *SMOTE*. Model *LightGBM* yang digunakan mencapai akurasi rata-rata 79,20% pada validasi *K-Fold* dan 91,84% pada data *testing*, dengan nilai *ROC AUC* sebesar 0.9614. Dari hasil visualisasi dan analisis lebih lanjut, atribut *BMI* dalam dataset diketahui memiliki pengaruh paling besar dalam proses klasifikasi terhadap resiko seseorang terpapar penyakit diabetes, yang diikuti oleh *DiabetesPedigreeFunction* dan *Glucose*. Dengan adanya pendekatan ini diharapkan dapat mendukung diagnosis awal diabetes secara lebih akurat dan efisien ke depannya.

Pada penelitian selanjutnya diharapkan untuk membandingkan berbagai metode imputasi seperti *KNN, MICE,* dan imputasi statistik lainnya guna menemukan pendekatan terbaik sesuai dengan karakteristik data yang digunakan. Selain itu, penggunaan algoritma klasifikasi lain seperti *Random Forest, XGBoost,* atau model *deep learning* lainnya juga dapat diuji sebagai perbandingan untuk meningkatkan akurasi dan keandalan model prediksi penyakit diabetes.

DAFTAR RUJUKAN

Alfebi, Fadlan Hamid, and Mila Desi Anasanti. 2023. "Improving Cardiovascular Disease Prediction by Integrating Imputation, Imbalance Resampling, and Feature Selection

Techniques into Machine Learning Model." *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)* 17(1): 55. doi:10.22146/ijccs.80214.

Bemi, Windri Wucika, and Rani Nooraeni. 2019. "Dampak Redenominasi Terhadap Inflasi Indonesia: Penanganan Missing Menggunakan Metode Case Deletion, Pmm, Rf Dan Bayesian." *Indonesian Journal of Statistics and Its Applications* 3(3): 272–86. doi:10.29244/ijsa.v3i3.360.

Candra Permana, Baiq Andriska, and Intan Komala Dewi Patwari. 2021. "Komparasi Metode Klasifikasi Data Mining Decision Tree Dan Naïve Bayes Untuk Prediksi Penyakit Diabetes." *Infotek : Jurnal Informatika dan Teknologi* 4(1): 63–69. doi:10.29408/jit.v4i1.2994.

Demircioğlu, Aydin. 2024. "Applying Oversampling before Cross-Validation Will Lead to High Bias in Radiomics." *Scientific Reports* 14(1): 1–11. doi:10.1038/s41598-024-62585-z.

Derisma, D. 2020. "Perbandingan Kinerja Algoritma Untuk Prediksi Penyakit Jantung Dengan Teknik Data Mining." *Journal of Applied Informatics and Computing* 4(1): 84–88. doi:10.30871/jaic.v4i1.2152.

Fuadah, Yunendah Nur, Ibnu Dawan Ubaidullah, Nur Ibrahim, Fauzi Frahma Taliningsing, Nidaan Khofiya Sy, and Muhammad Adnan Pramuditho. 2022. "Optimasi Convolutional Neural Network Dan K-Fold Cross Validation Pada Sistem Klasifikasi Glaukoma." *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika* 10(3): 728. doi:10.26760/elkomika.v10i3.728.

Gholamy, Afshin, Vladik Kreinovich, and Olga Kosheleva. 2018. "Why 70 / 30 or 80 / 20 Relation Between Training and Testing Sets : A Pedagogical." *Departmental Technical Reports (CS)* 1209: 1–6. https://scholarworks.utep.edu/cs_techrep.

Halabaku, Erblin, and Eliot Bytyçi. 2024. "Overfitting in Machine Learning : A Comparative Analysis of Decision Trees and Random Forests." doi:10.32604/iasc.2024.059429.

Hou, Fan, Zhi Xiang Cheng, Luo Yao Kang, and Wen Zheng. 2020. "Prediction of Gestational Diabetes Based on LightGBM." *ACM International Conference Proceeding Series*: 161–65. doi:10.1145/3433996.3434025.

Hovi, Hovi Sohibul Wafa, Asep Id Hadiana, and Fajri Rakhmat Umbara. 2022. "Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM)." *Informatics and Digital Expert (INDEX)* 4(1): 40–45. doi:10.36423/index.v4i1.895.

Maulidah, Nurlaelatul, Riki Supriyadi, Dwi Yuni Utami, Fuad Nur Hasan, Ahmad Fauzi, and Ade Christian. 2021. "Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine Dan Naive Bayes." *Indonesian Journal on Software Engineering (IJSE)*

7(1): 63–68. doi:10.31294/ijse.v7i1.10279.

Novianto, Anton, and Mila Desi Anasanti. 2023. "Autism Spectrum Disorder (ASD) Identification Using Feature-Based Machine Learning Classification Model." *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)* 17(3): 259. doi:10.22146/ijccs.83585.

Rufo, Derara Duba, Taye Girma Debelee, Achim Ibenthal, and Worku Gachena Negera. 2021. "Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (Lightgbm)." *Diagnostics* 11(9): 1–14. doi:10.3390/diagnostics11091714.

Septiana Rizky, Putri, Ristu Haiban Hirzi, and Umam Hidayaturrohman. 2022. "Perbandingan Metode LightGBM Dan XGBoost Dalam Menangani Data Dengan Kelas Tidak Seimbang." *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika* 15(2): 228–36. doi:10.36456/jstat.vol15.no2.a5548.

Tomic, Dunya, Jonathan E. Shaw, and Dianna J. Magliano. 2022. "The Burden and Risks of Emerging Complications of Diabetes Mellitus." *Nature Reviews Endocrinology* 18(9): 525–39. doi:10.1038/s41574-022-00690-7.

Tuntun, Ritham, Kusrini Kusrini, and Kusnawi Kusnawi. 2022. "Analisis Perbandingan Kinerja Algoritma Klasifikasi Dengan Menggunakan Metode K-Fold Cross Validation." *Jurnal Media Informatika Budidarma* 6(4): 2111. doi:10.30865/mib.v6i4.4681.

Ucha Putri, Sanni, Eka Irawan, Fitri Rizky, Stikom Tunas Bangsa, Pematangsiantar A - Indonesia Jln Sudirman Blok No, and Sumatera Utara. 2021. "Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5." *Januari* 2(1): 39–46.

Wardhana, Indrawata, Musi Ariawijaya, Vandri Ahmad Isnaini, and Rahmi Putri Wirman. 2022. "Gradient Boosting Machine, Random Forest Dan Light GBM Untuk Klasifikasi Kacang Kering." *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)* 6(1): 92–99. doi:10.29207/resti.v6i1.3682.