Seminar Nasional Sains Data 2024 (SENADA 2024)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Penerapan Algoritma Stacking Ensemble Machine Learning Berbasis Pohon untuk Prediksi Penyakit Diabetes

Atina Nora Haya[1], Mega Yuliani Ramme[2]
[1, 2]*Program Studi Sains Data, Institut Teknologi Telkom Purwokerto*
[1]2211110043@ittelkom-pwt.ac.id,
Corresponding author: [2]2211110027@ittelkom-pwt.ac.id

Abstract: *Machine learning is commonly used to predict diabetes patient outcomes. However, ensemble techniques can further improve prediction model accuracy. Stacking, an ensemble technique, combines several machine-learning models. This study aims to apply a stacking ensemble to classify patient clinical data into diabetes or non-diabetes classes. The classifier algorithm is tree-based, including CatBoost, XGBoost, Extra Tree, Decision Tree, and Random Forest as the base learners and Light Gradient-Boosting Machine (LGBM) as the meta learner. The results of this research show that the prediction model for the base learners is accurate as follows: CatBoost reached 76.04%, XGBoost reached 75.71%, Random Forest reached 84.82%, Extra Tree reached 84.96%, Decision Tree reached 80.66%, and Light Gradient-Boosting Machine reached 72.87%. Meanwhile, the stacking ensemble model built based on the base learners with the LGBM meta learner achieved an accuracy of 87.47%. Therefore, the research indicates that the stacking model outperforms the individual models.*
Keywords: *diabetes, prediction, tree-based-machine-learning, stacking, ensemble*

Abstrak: Prediksi kelas pasien diabetes menggunakan *machine learning* sudah banyak dilakukan. Model prediksi menggunakan *machine learning* secara tunggal masih bisa dioptimalkan untuk mencapai akurasi lebih baik, biasanya menggunakan teknik *ensemble*. *Stacking* adalah salah satu teknik *ensemble* yang menggabungkan beberapa model pembelajaran mesin. Penelitian ini bertujuan menerapkan *stacking ensemble* untuk mengklasifikasi data klinis pasien kedalam kelas diabetes atau tidak diabetes. Algoritma *classifier* yang digunakan adalah berbasis pohon, antara lain *CatBoost, XGBoost, Extra Tree, Decision Tree, dan Random Forest* sebagai *base learner* dan *Light Gradient-Boosting Machine* (LGBM) sebagai *meta learner*. Hasil penerapan menunjukkan akurasi model prediksi untuk *base learner* yaitu *CatBoost* mencapai 76,04%, *XGBoost* mencapai 75,71%, *Random Forest* mencapai 84,82%, *Extra Tree* mencapai 84,96%, *Decision Tree* mencapai 80,66%, *Light Gradient-Boosting* mencapai 72,87%. Sedangkan model *stacking ensemble* yang dibangun berdasarkan *base learner* tersebut dengan *meta learner* LGBM mencapai akurasi 87,47%. Dengan demikian, penelitian menunjukkan bahwa model *stacking* lebih baik daripada model individu.
Kata Kunci: diabetes, prediksi, *machine learning* berbasis pohon, *stacking,* ensembel

I. PENDAHULUAN

Diabetes merupakan masalah kesehatan yang mendapat perhatian besar di seluruh dunia. Menurut data *CDC Health Surveillance*[1], jumlah penderita diabetes yang terus meningkat setiap tahunnya menunjukkan pentingnya upaya pencegahan dan pengobatan penyakit ini. Salah satu tantangan utama dalam diagnosa diabetes adalah lamanya waktu pengambilan keputusan. Dalam mengatasi masalah tersebut, penelitian terkait pengembangan model prediksi dengan algoritma pembelajaran mesin banyak dilakukan oleh para peneliti[2]. Namun, model prediksi tersebut tidak selalu mencapai akurasi yang tinggi. Dalam domain kesehatan, prediksi yang tepat menjadi satu hal yang harus dipenuhi karena hasil prediksi yang salah akan berakibat fatal pada pengambilan keputusan. Dalam banyak penelitian model prediksi berbasis pembelajaran mesin[3], model pembelajaran mesin tunggal seringkali perlu optimasi agar dapat menghasilkan akurasi yang tinggi. Dalam penelitian[4], teknik *stacking* dapat meningkatkan akurasi model prediksi. Cara kerja Algoritma ini menggabungkan beberapa model pembelajaran mesin tunggal sebagai *base learner* dan algoritma *machine learning* lainnya sebagai *meta learner*. Teknik *stacking* ini merupakan salah satu metode *ensemble.*

Ensemble learning, termasuk teknik seperti *bagging* dan *boosting* telah banyak digunakan dalam prediksi berbagai penyakit[5]. Penelitian[4] dalam ulasannya menyatakan bahwa meskipun teknik

seperti *bagging* dan *boosting* lebih sering digunakan, teknik *stacking* sering kali memberikan hasil yang paling akurat dalam memprediksi penyakit seperti diabetes, penyakit kulit, penyakit ginjal, penyakit hati, dan penyakit jantung. Lebih lanjut, penelitian[6] menunjukkan bahwa model *stacking* dapat meningkatkan akurasi prediksi penyakit jantung dengan menggabungkan beberapa algoritma dasar. Penelitian mereka menggarisbawahi pentingnya prediksi penyakit jantung yang akurat mengingat tingginya angka kematian akibat penyakit ini. Dengan menggunakan data kesehatan pasien, model *stacking* yang dikembangkan berhasil memberikan peningkatan kinerja yang signifikan dalam prediksi penyakit jantung.

Selain itu, penelitian[7] meninjau penggunaan algoritma klasifikasi berbasis jaringan saraf dalam sistem kesehatan untuk diagnosa penyakit kronis seperti diabetes, penyakit jantung, dan kanker. Studi mereka menunjukkan bahwa metode *ensemble* berbasis jaringan saraf dapat mencapai tingkat akurasi yang sangat tinggi, dengan beberapa metode mencapai hingga 100% akurasi. Penelitian ini menyoroti tantangan dan potensi peningkatan dalam penerapan *machine learning* untuk diagnosa penyakit kronis, menunjukkan bahwa pendekatan *ensemble* dapat lebih efektif dibandingkan dengan algoritma dasar tunggal.

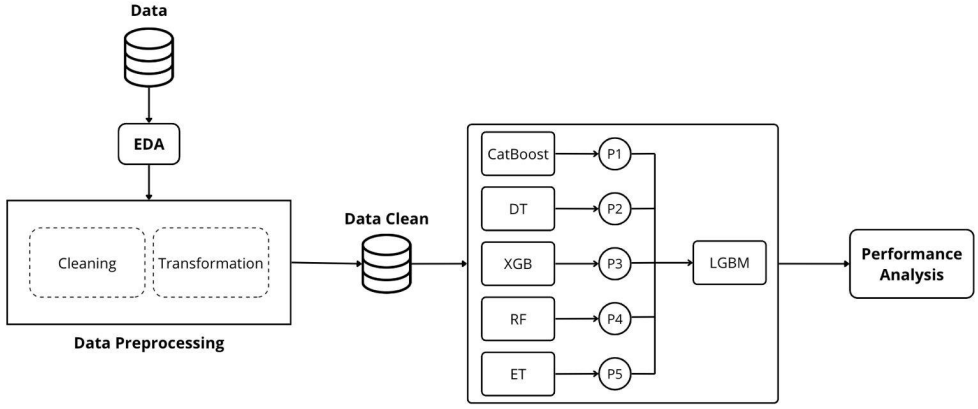
Dalam konteks prediksi penyakit diabetes, penggunaan algoritma *stacking ensemble classifier* diharapkan dapat memberikan hasil prediksi yang lebih akurat dan membantu dalam penentuan langkah-langkah pengobatan yang tepat. Dalam penelitian[6] menunjukkan bahwa penggunaan algoritma *stacking ensemble classifier* terbukti efektif dalam memprediksi berbagai jenis penyakit, termasuk diabetes. Hasil penelitian tersebut menunjukkan bahwa metode ini dapat meningkatkan akurasi prediksi hingga 10% dibandingkan dengan model tunggal. Penelitian ini dapat membantu meningkatkan akurasi prediksi yang dapat digunakan dalam praktik medis. Selain itu, dengan membandingkan hasil prediksi antara algoritma stacking dengan metode prediksi lainnya, manfaat praktis dari penelitian ini adalah dapat meningkatkan kualitas prediksi diabetes dan membantu mengambil keputusan medis yang lebih akurat.

Melalui penggabungan berbagai metode *ensemble*, khususnya teknik *stacking*, penelitian ini bertujuan untuk mengeksplorasi dan meningkatkan akurasi prediksi penyakit. Pendekatan ini tidak hanya memberikan hasil yang lebih akurat tetapi juga membuka peluang untuk pengembangan sistem prediksi penyakit yang lebih andal dan efisien dalam praktik klinis.

II. METODE

Penelitian ini menerapkan pendekatan *machine learning* dengan algoritma *Stacking Ensemble Classifier* untuk mengklasifikasi data survei kesehatan CDC *(Centers for Disease Control)*[8] sebagai data yang masuk untuk klasifikasi terindikasinya pengidap diabetes dan tidak terindikasi diabetes. Tahapan dalam membangun model prediksi ditunjukkan Gambar 1.

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 1. Tahapan Membangun Model Klasifikasi

2.1 Akuisis Data

Dataset dari *Centers for Disease Control and Prevention* (CDC) yang terdiri dari 253.680 observasi dengan 22 variabel mencakup berbagai aspek kesehatan, demografi, dan perilaku kesehatan. Dataset ini, yang bisa berasal dari survei seperti *Behavioral Risk Factor Surveillance System* (BFSS), bertujuan untuk memantau prevalensi penyakit dan faktor kesehatan diantara penduduk AS. Variabel prediktor dalam dataset ini mencakup jenis kelamin, usia, indeks massa tubuh (BMI), status merokok, konsumsi alkohol, keterlibatan dalam aktivitas fisik, status riwayat penyakit jantung, riwayat stroke, kesehatan mental dan fisik. Variabel target dalam analisis ini adalah status diabetes dengan dua kategori, yaitu diabetes dan non-diabetes. Data ini sangat penting untuk analisis epidemiologis, penelitian kesehatan masyarakat, dan evaluasi kebijakan kesehatan yang bertujuan mengidentifikasi dan mengelola faktor risiko diabetes.

2.2. *Exploratory Data Analysis (EDA)*

Pada tahap ini merupakan langkah awal yang penting sebelum melakukan analisis statistik lebih lanjut atau pemodelan prediktif. Tujuannya adalah untuk memperoleh pemahaman menyeluruh tentang fitur, struktur, dan komponen utama dari dataset. Tahap ini meliputi memeriksa distribusi data, mengidentifikasi pencilan *(outliers)*, dan menyelidiki faktor-faktor yang terkait dengan kasus diabetes berdasarkan tertinggi usia, jenis kelamin, dan BMI. Melalui EDA, peneliti menjelajahi dimensi dataset, memuat dan memeriksa baris-baris awal untuk memahami formatnya, serta memeriksa nilai missing value setiap variabel. Ringkasan statistik seperti mean, median dan deviasi standar memberikan wawasan tentang variabel numerik, sementara visualisasi seperti histogram dan box plot mengungkapkan distribusinya dan menyoroti *outlier* dengan memeriksa faktor demografis seperti usia dan jenis kelamin, bersama dengan indikator kesehatan seperti BMI, EDA memfasilitasi identifikasi pola dan tren yang dapat mempengaruhi prevalensi diabetes. Proses ini tidak hanya mempersiapkan data untuk analisis lebih lanjut, tetapi juga memastikan bahwa upaya pemodelan yang dilakukan berdasarkan pada pemahaman yang komprehensif tentang karakteristik tersebut.

2.3 *Preprocessing*

Pada proses pembersihan dan *transformasi* data menjadi krusial sebelum melanjutkan ke tahap analisis identifikasi, pemahaman, dan penanganan terhadap data yang kotor, tidak lengkap, atau tidak sesuai. Langkah-langkah ini mencakup penghapusan nilai yang *outlier* yang signifikan, serta penanganan duplikat data untuk memastikan keakuratan dan konsistensi dataset. *Transformasi* data dilakukan untuk mengubah format atau struktur data mentah menjadi format yang lebih sesuai dan

konsisten untuk analisis lebih lanjut. Contoh *transformasi* meliputi normalisasi data numerik, pengkodean ulang data kategorikal, atau ekstraksi fitur dari data mentah untuk memperkaya informasi yang relevan.

Proses ini penting untuk memastikan bahwa data yang digunakan dalam analisis atau pemodelan memiliki kualitas yang tinggi dan siap digunakan. Dengan membersihkan data dari anomali dan melakukan transformasi yang diperlukan, peneliti dapat meningkatkan keandalan hasil analisis dan meminimalkan potensi bias akibat data yang tidak bersih. Hal ini juga mendukung validitas interpretasi dan keputusan yang diambil dari hasil analisis data[9].

2.3.1 *Data Cleaning*

Sebelum dataset dapat digunakan untuk analisis lebih lanjut, proses *preprocessing* data sangat penting dilakukan. Ini melibatkan serangkaian langkah untuk memeriksa serta membersihkan data *outlier* dan mengidentifikasi serta menghapus data duplikat. Sebagai contoh, pada dataset awal yang terdiri dari 253.680 data setelah melalui *preprocessing*, jumlah data bersih menurun menjadi 229.474. reduksi ini terutama disebabkan oleh penghapusan data duplikat yang ditemukan selama proses pembersihan. Proses preprocessing data yang teliti seperti ini penting untuk meningkatkan akurasi dan keandalan dataset sebelum dilakukan analisis statistik atau pemodelan. Dengan membersihkan data dari anomali dan menangani duplikasi, peneliti dapat memastikan bahwa hasil analisis yang dihasilkan adalah refleksi yang lebih akurat dari fenomena yang dipelajari.

2.3.2 *Data Transformation*

Transformasi data digunakan untuk mengolah data agar memiliki distribusi atau skala yang lebih konsisten dan sesuai analisis statistik atau pemodelan. Keberadan *outlier* dalam data dapat menyebabkan ketidakseragaman dalam skala atau distribusi data, yang dapat mempengaruhi kinerja modek atau hasil analisis yang dihasilkan. Salah satu teknik *transformasi* yang umum digunakan adalah *Standar Scaler*. *Standar Scaler* adalah teknik yang mengubah data sehingga memiliki mean nol dan deviasi standar satu. Ini dilakukan dengan menggunakan rumus *transformasi* berikut[10]:

$$z = \frac{x - \mu}{\sigma}$$

Keterangan:
z : nilai baru yang sudah distandarisasi
x: nilai asli dari variabel
μ: rata-rata dari variabel tersebut
σ: standar deviasi dari variabel tersebut

Proses *Standar Scaler* ini membantu dalam menghilangkan dampak skala yang berbeda antar variabel, sehingga mempermudah interpretasi hasil dan meningkatkan performa algoritma pembelajaran mesin yang sensitif terhadap skala.

2.3.3 *Oversampling*

Data yang digunakan dalam penelitian ini mempunyai distribusi tidak normal, yaitu ketidakseimbangan kelas sasaran. Untuk mengatasi permasalahan tersebut digunakan *oversampling* dengan cara menambah jumlah sampel kelas mayoritas. Teknik ini membantu model belajar lebih baik dari kelas minoritas dan meningkatkan performa model[11].

2.3.4 *Data Splitting*

Pembagian dataset menjadi dua bagian terpisah adalah praktik umum dalam pembangunan model machine learning untuk memastikan bahwa model yang dikembangkan dapat digeneralisasi dengan baik ke data baru yang belum pernah dilihat sebelumnya. Proses ini mengurangi risiko *overfitting*, di mana model terlalu "menghafal" data pelatihan dan tidak mampu menggeneralisasi dengan baik pada data baru. Dalam pembagian ini, 80% dari dataset digunakan untuk proses pelatihan (*training set*). Data pada *training set* digunakan untuk melatih model untuk mengenali pola dan karakteristik yang relevan dari data, sehingga model dapat belajar untuk membuat prediksi atau mengklasifikasi data dengan akurat. Sementara itu, 20% sisanya dari dataset digunakan untuk proses pengujian (*test set*). *Test set* digunakan untuk menguji kinerja model yang telah dilatih dengan data yang tidak pernah dilihat sebelumnya. Pengujian dilakukan untuk mengukur seberapa baik model dapat melakukan prediksi atau klasifikasi pada data baru yang belum terlihat, dan untuk mengevaluasi seberapa baik model dapat digeneralisasi. Pemisahan dataset menjadi training set dan test set memungkinkan peneliti atau praktisi *machine learning* untuk mendapatkan perkiraan yang lebih objektif tentang kinerja model mereka didunia nyata. Hal ini penting dalam memastikan bahwa model yang dikembangkan dapat diandalkan dan berguna untuk aplikasi praktis[12].

2.4 *Modeling*

Setelah data diproses, langkah berikutnya adalah menjalankan model klasifikasi menggunakan metode klasifikasi menggunakan metode pembelajaran ansambel. Data diklasifikasikan menggunakan algoritma seperti *CatBoost, XGBoost, Extra Tree, Decision Tree*, dan *Random Forest*. Teknik pembelajaran ansambel dengan pendekatan stacking juga diterapkan, dimana hasil klasifikasi dari setiap *base classifier* digabungkan dan digunakan sebagai input untuk *meta classifier*, yaitu *Light Gradient-Boosting Machine* (LGBM). Pendekatan ini bertujuan meningkatkan akurasi dan robustness model klasifikasi[13]. *CatBoost* efektif menangani data kategorik tanpa pra-pemrosesam tambahan, menggunakan pohon keputusan ansambel yang memperbaiki kesalahan model sebelumnya, keunggulan utamanya adalah penanganan variabel kategori dan penggunaan teknik pembobotan sampel, serta metode *ordered boosting* yang mencegah *overfitting*[14]. Sementara itu, LGBM dikenal karena efisiensi dan kecepatannya dalam tugas klasifikasi, mengurangi kompleksitas komputasi dan penggunaan memori, sehingga cocok untuk dataset besar[15].

Selanjutnya, *XGBoost* menggunakan *gradient boosting* untuk memperbaiki prediksi model sebelumnya. Algoritma ini dikenal karena efisiensi waktu pelatihan dan penggunaan memori, seta kemampuannya menangani data dengan fitur yang hilang dan melakukan regularisasi untuk mencegah *overfitting*[16]. *Random Forest* membangun banyak pohon keputusan dari subset acak data pelatihan dan fitur, menggabungkan prediksi dari setiap pohon untuk meningkatkan akurasi dan mencegah *overfitting*[17]. Di sisi lain, *Decision Tree* menggunakan struktur pohon untuk membuat keputusan berdasarkan input data[18]. Model ini menawarkan interpretabilitas tinggi, meskipun rentan terhadap *overfitting* jika terlalu kompleks. Sebagai variasi, *Extra Tree* meningkatkan randomisasi dalam pembuatan pohon keputusan, mengurangi korelasi antar pohon, dan membantu meminimalkan kesalahan keseluruhan[19].

Metode *stacking ensemble* menggabungkan beberapa model dasar untuk memprediksi hasil, mengurangi kesalahan prediksi dan *overfitting*. *Stacking* terdiri dari dua level: *base learners* sebagai level-0 dan *meta learner* sebagai level-1. *Base learners* menggunakan model yang berbeda untuk belajar dari dataset, dan output dari masing-masing model dikumpulkan untuk membuat dataset baru yang kemudian digunakan oleh *meta learner* untuk memberikan hasil akhir[20].

2.5 *Performance Analysis*

Confusion matrix adalah salah satu metode untuk menghitung akurasi model dengan memberikan informasi tentang perbandingan hasil klasifikasi yang dibuat oleh sistem dengan klasifikasi yang sebenarnya *(ground truth). Confusion matrix* berbentuk tabel matriks yang menggambarkan performa model klasifikasi berdasarkan sekumpulan data uji yang nilai sebenarnya telah diketahui. C*onfusion matrix* dapat mengevaluasi seberapa baik model dalam mengklasifikasikan data ke dalam kategori yang benar, serta mengidentifikasi kesalahan klasifikasi yang terjadi. Tabel 1 adalah *confusion matrix* yang dapat digunakan untuk memahami dan menganalisis performa model klasifikasi secara lebih mendalam[21].

Tabel 1. Confusion Matrix

Confusion Matrix		**Kelas Prediktif**	
		Positif	Negatif
Kelas Aktual	Positif	TP	FN
	Negatif	FP	TN

Kinerja algoritma tersebut dapat dihitung seperti penjelasan sebagai berikut:
Accuracy merupakan persentase dari jumlah prediksi yang benar dari seluruh jumlah prediksi yang dilakukan oleh *classifier,* dihitung dengan rumus (1).

$$\text{accuracy} = \frac{TP+TN}{TP+FP+TN+FN} * 100\% \qquad (1)$$

Recall adalah persentase dari prediksi True Positive, dibandingkan dengan keseluruhan data positif, dihitung dengan rumus (2).

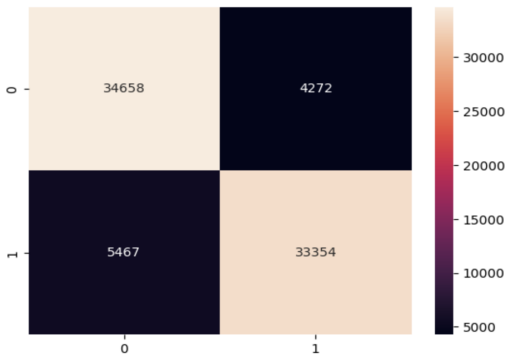
$$\text{recall} = \frac{TP}{TP+FN} * 100\% \qquad (2)$$

Precision adalah ukuran persentase dari prediksi True Positive dibandingkan keseluruhan hasil yang diprediksi sebagai positive.

$$\text{precision} = \frac{TP}{TP+FP} * 100\% \qquad (3)$$

III. HASIL DAN PEMBAHASAN

Dalam mengukur performance hasil pemodelan, pada penelitian ini menggunakan *confusion matrix, accuracy, recall, precision,* dan *F1-score*. Model *stacking* yang yang dibangun memiliki matrik konfusi pada Gambar 2.



Gambar 2. *Confusion Matrix Stacking*

Berdasarkan Gambar 2 di atas, diketahui bahwa sebanyak 34.658 data tidak diabetes dan 33.354 data diabetes berhasil diagnosa dengan benar oleh model. Namun, terdapat 4.272 data tidak diabetes dan 5.467 data diabetes tidak berhasil terdiagnosa dengan benar oleh model. Perbandingan performa model base dan *stacking* ditunjukkan Tabel 2.

Tabel 2. Hasil Perhitungan *Performa Model Base Learner dan Stacking*

Algoritma	Akurasi	Presisi	Recal	Skor F1
Catboost	76,04%	74,18%	79,79%	76,88%
Decision Tree	80,66%	77,79%	85,74%	81,57%
XGBoost	75,71%	73,85%	79,50%	76,57%
Random Forest	84,82%	82,01%	86,98%	84,37%
Extra Tree	84,96%	83,12%	87,68%	85,34%
LGBM	72,87%	72,92%	79,00%	75,84%
Stacking	**87,47%**	**88,65%**	**85,92%**	**87,26%**

Berdasarkan Tabel 2 hasil perhitungan *confusion matrix* di atas, diketahui bahwa nilai akurasi dari model *stacking* paling baik dibandingkan menggunakan model tunggal lainnya. Dengan demikian, dapat disimpulkan bahwa model *stacking* efektif untuk meningkatkan akurasi model prediksi.

IV. KESIMPULAN

Penelitian menggunakan kombinasi *single classifier* dengan teknik *ensemble stacking* menggunakan *CatBoost, XGBoost, Extra Tree, Decision Tree* dan *Random Forest* berfungsi sebagai *base learner* dan *Light Gradient-Boosting Machine* berfungsi sebagai *meta learner*. Nilai akurasi digunakan untuk mengukur kinerja hasil pengujian dari model yang digunakan. Menurut evaluasi model, nilai akurasi *stacking ensemble* paling efektif dalam penelitian prediksi penyakit diabetes dengan skor 87,47%, *CatBoost* 76,04%, *XGBoost* 75,71%, *Extra Tree* 84,96%, *Decision Tree* 80,66%, *Random Forest* 84,82%, dan LGBM 72,87%. Oleh karena itu, model *stacking ensemble* terbukti efektif dalam memprediksi penyakit diabetes.

REFERENSI

1. "National Diabetes Statistics Report," U.S. Centers for Disease Control and Prevention.
2. N. R. Dzakiyullah, M. A. Burhanuddin, R. R. R. Ikram, K. A. Ghani, dan W. Setyonugroho, "Machine learning methods for diabetes prediction," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, hlm. 2199–2205, Okt 2019, doi: 10.35940/ijitee.L2973.1081219.
3. M. R. Pahlawan, "Penggunaan Explainable Machine Learning untuk Prediksi Pasien Diabetes," 2024. doi: https://doi.org/10.24089/j.sisfo.2024.05.002.
4. P. Mahajan, S. Uddin, F. Hajati, dan M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," *Healthcare (Switzerland)*, vol. 11, no. 12. MDPI, 1 Juni 2023. doi: 10.3390/healthcare11121808.
5. P. Y. Taser, "Application of Bagging and Boosting Approaches Using Decision Tree-Based Algorithms in Diabetes Risk Prediction," dalam *International Management Information Systems Conference*, MDPI AG, Mar 2021, hlm. 1–9. doi: 10.3390/proceedings2021074006.
6. S. Verma, R. Dhir, dan M. KUMAR, "Heart Disease Prediction Using Stacking Ensemble Model Based on Machine Learning Approach," 2023, hlm. 335–347. doi: 10.1007/978-981-99-3432-4_26.
7. J. Abdollahi, B. Nouri-Moghaddam, dan M. Ghazanfari, "Deep Neural Network Based Ensemble learning Algorithms for the healthcare system (diagnosis of chronic diseases)," 2021.
8. "CDC Diabetes Health Indicators," UC Irvine Machine Learning Repository.
9. E. Retnoningsih dan R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised dan Unsupervised Learning Menggunakan Python," *BINA INSANI ICT JOURNAL*, vol. 7, no. 2, hlm. 156–165, 2020, [Daring]. Tersedia pada: https://www.python.org/

10. A. Kaur dan M. Sarmadi, "Comparative Analysis of Data Preprocessing Methods, Feature Selection Techniques and Machine Learning Models for Improved Classification and Regression Performance on Imbalanced Genetic Data," Feb 2024, [Daring]. Tersedia pada: http://arxiv.org/abs/2402.14980
11. N. Cahyana, S. Khomsah, dan A. S. Wibowo, "Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting," 2019.
12. J. Brownlee, "Train-Test Split for Evaluating Machine Learning Algorithms," *Python Machine Learning*, Agu 2020.
13. Y. Zhang, J. Liu, dan W. Shen, "A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications," *Applied Sciences (Switzerland)*, vol. 12, no. 17. MDPI, 1 September 2022. doi: 10.3390/app12178654.
14. A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, dan G. A. Saheed, "Comparison of the CatBoost Classifier with other Machine Learning Methods," 2020. [Daring]. Tersedia pada: www.ijacsa.thesai.org
15. B. Shamreen Ahamed dan M. Sumeet Arya, "Prediction of Type-2 Diabetes using the LGBM Classifier Methods and Techniques," 2021.
16. I. Muslim Karo Karo, "Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan," 2020.
17. V. Ignatenko, A. Surkov, dan S. Koltcov, "Random forests with parametric entropybased information gains for classification and regression problems," *PeerJ Comput Sci*, vol. 10, 2024, doi: 10.7717/peerj-cs.1775.
18. B. Charbuty dan A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, hlm. 20–28, Mar 2021, doi: 10.38094/jastt20165.
19. A. García-Domínguez *dkk.*, "Diabetes Detection Models in Mexican Patients by Combining Machine Learning Algorithms and Feature Selection Techniques for Clinical and Paraclinical Attributes: A Comparative Evaluation," *J Diabetes Res*, vol. 2023, 2023, doi: 10.1155/2023/9713905.
20. S. K. Kalagotla, S. V Gangashetty, dan K. Giridhar, "A novel stacking technique for prediction of diabetes," *Comput Biol Med*, vol. 135, hlm. 104554, 2021, doi: https://doi.org/10.1016/j.compbiomed.2021.104554.
21. W. I. Rahayu, C. Prianto, dan E. A. Novia, "Perbandingan Algoritma K-Means dan Naive Bayes untuk Memprediksi Prioritas Pembayaran Tagihan Rumah Sakit berdasarkan Tingkat Kepentingan pada PT. Pertamina (Persero)," *Jurnal Teknik Informatika*, vol. 13, no. 2, hlm. 1–8, 2021.