JURNAL ELEKTRO & INFORMATIKA
SWADHARMA (JEIS)
VOLUME 05 NOMOR 01 JANUARI 2025
ISSN 2774 – 5775 | eISSN 2774-5767

# ANALISIS PERBANDINGAN ENSEMBLE MACHINE LEARNING DENGAN TEKNIK SMOTE UNTUK PREDIKSI DIABETES

**Nur Tri Ramadhanti Adiningrum[1)], Nisa Hanum Harani[2)]**
[1,2] Prodi Teknik Informatika, Sekolah Vokasi Universitas Logistik dan Bisnis Internasional

Correspondence author: N.T.R. Adiningrum, nurtrira06@gmail.com, Bandung, Indonesia

**Abstract**

High blood glucose levels characterize a chronic disease called diabetes. Patients with diabetes will eventually experience health problems. These cases show that early detection and better diagnosis are needed. Although several Machine Learning (ML) models have been widely used in diabetes diagnosis, the algorithm performance is still between 70 - 79%. This study evaluates the use of Ensemble Machine Learning to predict diabetes using the Pima Indian Diabetes dataset. The models compared are Support Vector Machine, Linear Regression, Naive Bayes, Random Forest, AdaBoost, K Nearest Neighbour, and Decision Tree. The dataset will also be balanced using the Synthetic Minority Over-sampling Technique (SMOTE) to reduce accuracy bias. Cross-Industry Standard Process For Data Mining (CRISP-DM) is the methodology used. The accuracy results show that Random Forest with Bagging and Hard-Voting produces the best accuracy of other models. Where Random Forest produces an accuracy of 81.16% and Hard-Voting also produces an accuracy of 81.16%.

**Keywords:** prediction, diabetes, ensemble machine learning, smote, crisp-dm

**Abstrak**

Penyakit kronis yang disebut diabetes ditandai dengan kadar glukosa darah yang tinggi. Pasien dengan diabetes pada akhirnya akan mengalami masalah kesehatan. Kasus-kasus ini menunjukkan bahwa deteksi dini dan diagnosis yang lebih baik diperlukan. Meskipun beberapa model *Machine Learning* (ML) telah banyak digunakan dalam diagnosis diabetes, kinerja algoritmanya masih antara 70 - 79%. Untuk memutuskan apakah seseorang menderita diabetes atau tidak, penelitian ini mengevaluasi penggunaan *Ensemble Machine Learning* untuk memprediksi diabetes menggunakan dataset Diabetes Pima Indian. Model yang dibandingkan adalah *Support Vector Machine*, *Linear Regression*, Naive Bayes, *Random Forest*, *Adaboost*, *K Nearest Neighbor*, dan *Decision Tree*. Untuk mengurangi bias akurasi, dataset juga akan diseimbangkan menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE). *Cross-Industry Standard Process For Data Mining* (CRISP-DM) adalah metodologi yang digunakan. Hasil akurasi menunjukkan bahwa *Random Forest* dengan *Bagging* dan *Hard-Voting* menghasilkan akurasi terbaik dari model lainnya. Dimana *Random Forest* menghasilkan akurasi sebesar 81,16% dan *Hard-Voting* juga menghasilkan akurasi sebesar 81,16%.

**Kata Kunci**: prediksi, diabetes, *ensemble machine learning*, smote, crisp-dm

## A. PENDAHULUAN

Diabetes melitus adalah penyakit metabolik kronis yang ditandai oleh peningkatan kadar glukosa darah yang berisiko menyebabkan berbagai komplikasi kesehatan, seperti kerusakan pada ginjal, mata, dan jantung (Tsybikova et al., 2024). Menurut World Health Organization (WHO), sekitar 422 juta orang di seluruh dunia menderita diabetes, dimana 1,5 juta kematian secara langsung dikaitkan dengan diabetes setiap tahunnya (Retta et al., 2023). Penyakit ini adalah salah satu masalah kesehatan yang paling cepat berkembang di dunia pada abad ke-2. Berdasar International Diabetes Federation (IDF), diabetes menginfeksi 463 juta orang secara global pada tahun 2019, yang berarti 1 dari 11 orang dewasa (usia 20-79 tahun) menderita diabetes. Selanjutnya, IDF memperkirakan diabetes akan meningkat menjadi 643 juta pada tahun 2030 dan 783 juta pada tahun 2045 (Suprayitna et al., 2023).

Diabetes adalah kondisi kronis yang ditandai dengan peningkatan kadar glukosa darah. Diabetes menyebabkan kerusakan ginjal, mata, dan jantung yang progresif dari waktu ke waktu (Rastogi & Bansal, 2023). Diabetes dapat menyebabkan masalah kesehatan (Khanam & Foo, 2021). Dalam jangka waktu yang lama, hal ini meningkatkan kemungkinan bahwa pasien dengan diabetes akan mengalami masalah kesehatan lainnya (Makroum et al., 2022) termasuk kejadian kardio-serebrovaskular, penyakit ginjal, kerusakan mata, dan kerusakan sistem saraf (Nicolucci et al., 2022). Yang lebih mengkhawatirkan adalah efeknya pada kehamilan – sekitar 7% kehamilan dipengaruhi oleh diabetes setiap tahunnya (Olisah et al., 2022). Berdasarkan kasus tersebut, diperlukan deteksi dini, diagnosis yang lebih baik, dan pencegahan primer diabetes dan komplikasinya (Mistry et al., 2023).

Berdasarkan paparan sebelumnya, diperlukan suatu kebutuhan berupa teknik komputasi untuk membantu penerapan analisis dalam diagnosis diabetes (Carter et al., 2019). Hal ini karena kemajuan teknologi membuat pendekatan ini jauh lebih efektif seiring berjalannya waktu (Chang et al., 2022). Kemajuan teknologi, khususnya dalam bidang Machine Learning (ML), telah membuka peluang besar dalam pengembangan model prediksi penyakit seperti diabetes. Selain itu, pada tahap awal penyakit, indikator diabetes dapat lebih mudah diidentifikasi melalui teknologi daripada pemeriksaan manual (Chaki et al., 2022). Hingga saat ini, berbagai model *Machine Learning* (ML) telah menjadi penggunaan utama dalam mendiagnosis diabetes (Nicolucci et al., 2022).

Banyak penelitian telah melakukan perbandingan prediksi diagnosis diabetes menggunakan metode *Machine Learning* untuk menemukan hasil akurasi yang paling baik. Para peneliti telah bereksperimen dengan berbagai pendekatan ML untuk memprediksi penyakit sedini mungkin (Kibria et al., 2022). Pada penelitian (Khanam & Foo, 2021), prediksi diabetes dilakukan dengan algoritma Naive Bayes (NB), SVM, *Linear Regression*, *Adaboost*, *Random Forest*, *K Nearest Neighbor* (KNN), *Decision Tree* (DT). Ditemukan bahwa model dengan *Logistic Regression* (LR) dan *Support Vector Machine* (SVM) bekerja dengan baik pada prediksi diabetes dengan akurasi 77%–78%. Pada penelitian (Rajendra & Latifi, 2021), dirancang model prediksi apakah pasien menderita diabetes berdasarkan pengukuran diagnostik dan menjelajahi berbagai teknik untuk meningkatkan kinerja dan akurasi. Hasil penelitian menunjukan akurasi tertinggi diperoleh sekitar 77,83% untuk dataset Pima Indian Dataset Diabetes, setelah menggunakan teknik ansambel-*Max Voting*. Pada penelitian (Kumari et al., 2021), mengusulkan ansambel algoritma dengan model *Machine Learning* yaitu, *Random Forest, Logistic Regression*, dan Naïve Bayes dengan *soft voting classifier* untuk

JURNAL ELEKTRO & INFORMATIKA
SWADHARMA (JEIS)
VOLUME 05 NOMOR 01 JANUARI 2025
ISSN 2774 – 5775 | eISSN 2774-5767

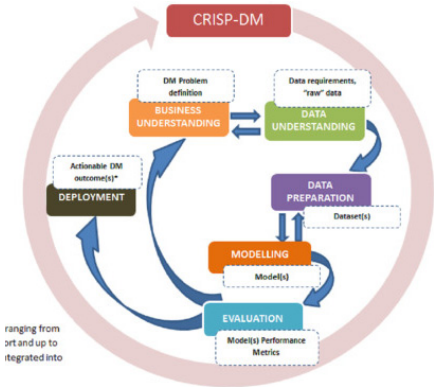
klasifikasi penyakit diabetes. Hasil menunjukan akurasi sebesar 79.08%. Penelitian (Joshi & Dhakal, 2021) menyajikan persamaan prediksi diabetes untuk memberikan pemahaman yang lebih baik tentang faktor risiko yang dapat membantu dalam mengklasifikasikan individu berisiko tinggi dengan algoritma *Logistic Regression* dengan hasil model yang dibuat memiliki akurasi sebesar 78,26%. Pada penelitian (Su et al., 2023), penggunaan *Linear Regression*, *Logistic Regression*, *Polynomial Regression* (PR), *Neural Network* (NN), *Support Vector Machines* (SVM), *Random Forest* (RF), and *XGboost* (XGB) dilakukan dalam pengklasifikasian prediksi risiko diabetes dengan membuat model adaptasi berdasarkan usia. Penelitian tersebut menghasilkan akurasi terbesar pada model *Logistic Regression* sebesar 78.80% dengan melakukan minoritas *over-sampling* dengan SMOTE (*Synthetic Minoritas Over-Sampling Technique*).

Penelitian seperti ini menunjukkan bahwa penggunaan dataset Diabetes Pima Indian untuk klasifikasi penyakit diabetes dengan algoritma *Machine Learning* telah diterapkan (Kumari et al., 2021), namun masih ada penelitian yang menghasilkan performa algoritma diantara 70-79% yang dianggap cukup namun belum optimal. Hasil tersebut menunjukan performa model masih dibawah 80% (cukup). Pendekatan *Ensemble Machine Learning* muncul sebagai solusi untuk meningkatkan akurasi prediksi dengan cara menggabungkan beberapa model ML untuk memberikan hasil prediksi yang lebih stabil dan akurat. Teknik *ensemble* pada *Machine Learning* berfungsi sebagai pendorong untuk meningkatkan performa suatu model *Machine Learning*. Sehingga *ensemble* cocok untuk meningkatkan akurasi prediksi (Tran & Kim, 2023). Di sisi lain, dataset yang tidak seimbang sering menjadi tantangan dalam penelitian ML karena dapat menyebabkan bias pada hasil prediksi. Oleh karena itu, *Synthetic Minority Over-sampling Technique* (SMOTE) digunakan dalam penelitian ini untuk menyeimbangkan dataset dan mengurangi bias akurasi (Joloudari et al., 2023).

Penelitian ini bertujuan untuk membandingkan performa model *Ensemble Machine Learning* dalam memprediksi diabetes menggunakan dataset Pima Indian Diabetes yang telah diseimbangkan dengan SMOTE. Berdasarkan hal tersebut penelitian ini berfokus pada perbandingan model *Ensemble Machine Learning* untuk prediksi diabetes menggunakan dataset Diabetes Pima Indian untuk mengidentifikasi apakah pasien tertentu menderita diabetes atau tidak (Rajendra & Latifi, 2021). Dataset juga akan dilakukan penyeimbangan data dengan SMOTE untuk mengurangi bias akurasi. Nantinya model tersebut dibandingkan untuk melihat model mana yang berforma dengan hasil terbaik.

## B. METODE PENELITIAN

Alur untuk melakukan pemecahan masalah pada penelitian dilakukan dengan metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM). CRISP-DM adalah metodologi yang paling umum digunakan untuk mengelola proyek *data mining* atau *data science*. CRISP-DM menyediakan kerangka kerja yang terstruktur untuk memahami dan menyelesaikan masalah bisnis melalui analisis data. CRISP-DM terdiri dari enam tahapan yang dieksekusi secara berulang seperti terlihat pada Gambar 1.




Gambar 1. CRISP-DM

Adapun tahapan tersebut dalam penelitian ini adalah sebagai berikut (Plotnikova et al., 2022):

1. *Business Understanding*

*Business Understanding* berfokus pada identifikasi tujuan bisnis dan persyaratan proyek. Dalam fase ini, tujuan penelitian dilakukan untuk membandingkan performa model *Machine Learning* yang dipadukan dengan *ensemble learning* dengan data yang diseimbangkan. Perbandingan dinilai berdasarkan nilai akurasi model prediksi.

2. *Data Understanding*

Setelah menetapkan tujuan bisnis, peneliti mulai mengumpulkan dan mengeksplorasi data yang tersedia. *Data Understanding* berfokus pada pengumpulan data dan eksplorasi data. Dalam penelitian ini, data yang digunakan adalah dataset Pima Indian Diabetes (PID) yang diambil dari repositori Kaggle. Dataset berisi informasi tentang 768 pasien dan sembilan atribut.

3. *Data Preparation*

Tahap ini melibatkan pembersihan dan transformasi data untuk mempersiapkannya dalam analisis selanjutnya. Langkah ini dapat mencakup penanganan data yang hilang, penghapusan *outlier*, normalisasi data, dan pemilihan fitur yang relevan untuk model.

4. *Modeling*

Pada tahap ini, peneliti memilih teknik analisis dan algoritma yang sesuai, seperti *Machine Learning*, untuk membangun model prediksi. Data yang telah disiapkan kemudian digunakan untuk melatih model, dan dilakukan *tuning* parameter untuk meningkatkan performa. *Modelling* adalah proses yang berfokus pada membangun model setelah memilih metode dan teknik yang sesuai. Model yang digunakan untuk prediksi adalah *Naive Bayes* (NB), SVM, *Linear Regression* (LR), *Adaboost*, *Random Forest*, *K Nearest Neighbor* (KNN), dan *Decision Tree* (DT).

Pada pemodelan ML dengan *ensemble Bagging*, tujuh model digunakan secara masing-masing dengan *Bagging*. Hasil pemodelan ini menghasilkan nilai akurasi pada masing-masing model ML yang dibaurkan dengan *Bagging*.

5. *Evaluation*

Tahap evaluasi berkaitan dengan penilaian kualitas dan konfirmasi bahwa tujuan bisnis proyek terpenuhi. Evaluasi model yang dilakukan pada penelitian ini adalah menguji performa model *Ensemble Machine Learning* menggunakan nilai akurasi. Dalam pekerjaan ini, evaluasi model dilakukan untuk melihat seberapa baik kinerja model pembelajaran mesin *ensemble* menggunakan nilai akurasi.

6. *Deployment*

*Fase Deployment* adalah tahap merencanakan dan memantau hasil pengembangan. Pada fase ini, model akan dibandingkan berdasarkan nilai akurasi untuk melihat model dengan performa terbaik dalam melakukan prediksi diabetes menggunakan dataset Pima Indian Diabetes.

## C. HASIL DAN PEMBAHASAN

### Pemahaman Dataset Yang Digunakan

Pemahaman data ini berfokus pada eksplorasi data yang digunakan dalam penelitian. Dalam penelitian ini, data yang digunakan adalah dataset Pima Indian Diabetes (PID) yang diambil dari repositori Kaggle. Dataset berisi informasi tentang 768 pasien dan sembilan atribut. Tabel 2 menunjukan deskripsi atribut dataset ini.

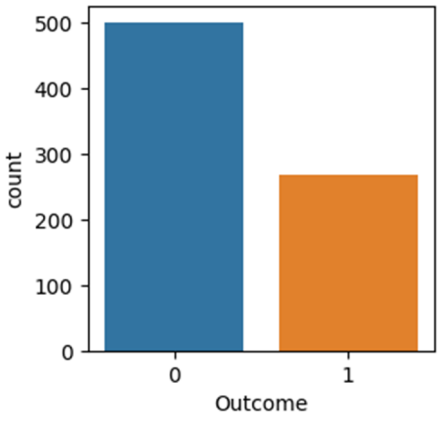
Tabel 1. Atribut Dataset

| Atribut | Keterangan | Tipe Data |
|---|---|---|
| Pregnancies | Berapa kali hamil. | Numerik |
| Glucose | Konsentrasi glukosa plasma 2 jam dalam tes toleransi glukosa oral. | Numerik |
| BloodPressure | Tekanan darah diastolik (mm Hg). | Numerik |
| SkinThickness | Ketebalan lipatan kulit trisep (mm). | Numerik |

JURNAL ELEKTRO & INFORMATIKA
SWADHARMA (JEIS)
VOLUME 05 NOMOR 01 JANUARI 2025
ISSN 2774 – 5775 | eISSN 2774-5767

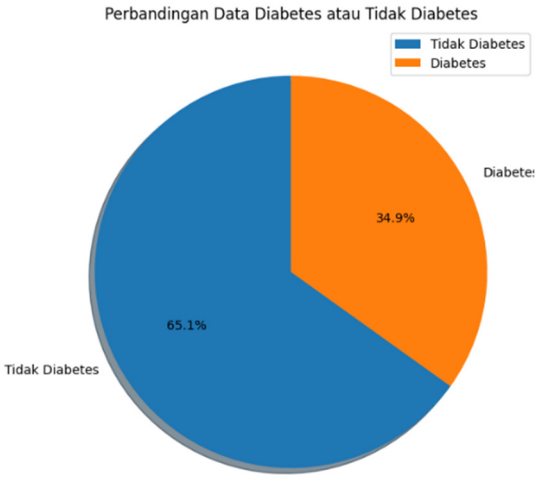
| Atribut | Keterangan | Tipe Data |
|---|---|---|
| Insulin | Insulin serum 2 jam (µlU/mL) | Numerik |
| BMI | Indeks massa tubuh (kg/m2$^2$). | Numerik |
| Age | Usia (tahun). | Numerik |
| DiabetesPedigreeFunction | Fungsi silsilah diabete | Numerik |
| Outcome | Hasil diagnosa diabetes (tes_positif: 1, tes_negatif: 0) | Numerik |

Pada Gambar 2, menunjukan hasil eksplorasi data, sebanyak 500 records memiliki status tidak diabetes dan 268 records memiliki status diabetes. Dalam hal ini, dataset tersebut memiliki data yang tidak seimbang.




Gambar 2. Jumlah Data Diabetes dan Tidak Diabetes

Dalam *piechart* pada Gambar 3 juga menunjukan bahwa data tidak seimbang antara jumlah yang terkena diabetes dengan yang tidak terkena diabetes. Hal ini juga mendukung bahwa perbandingan persen antara label Tidak Diabetes dan label Diabetes memiliki dominasi terkuat pada label Tidak Diabetes.



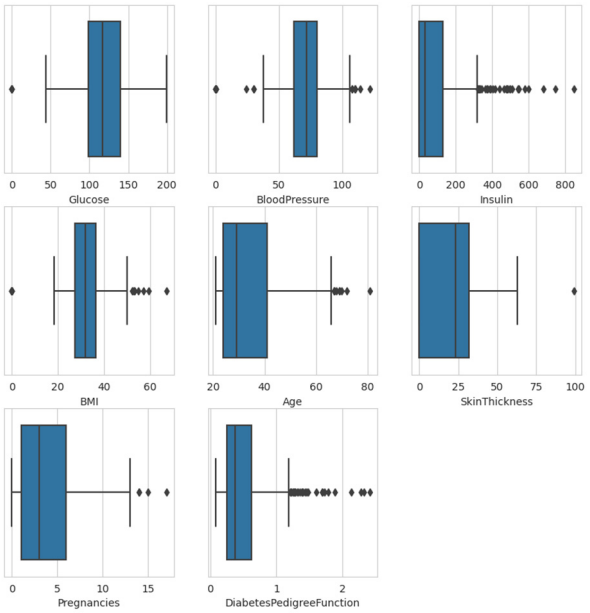

Gambar 3. Perbandingan Data Diabetes dan Tidak Diabetes

***Data Preparation***

*Data Preparation* adalah proses pembersihan dan transformasi data mentah sebelum diproses dan dianalisis. Hal yang dilakukan pada fase ini adalah mengatasi *outlier*, melakukan pemilihan fitur, dan melakukan penyeimbangan data dengan Teknik SMOTE.

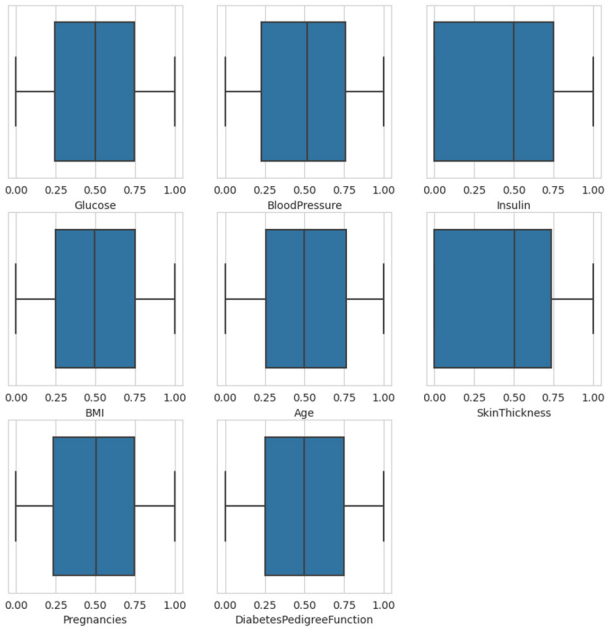
1. Identifikasi Dan Mengatasi Outlier

Pada Gambar 4 ditemukan nilai *outlier* pada data.




Gambar 4. *Outlier* Data

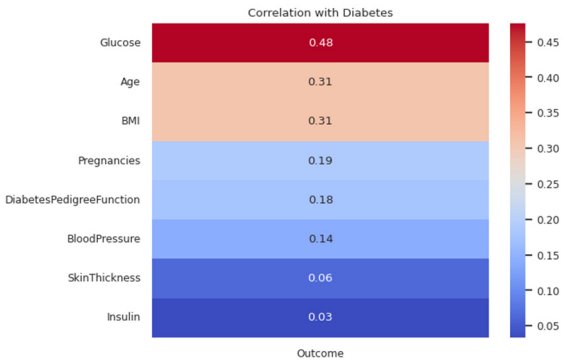
*Outlier* yang ada diatasi dengan Teknik *Quantile Transformer*. Metode ini mengubah fitur untuk mengikuti distribusi seragam atau normal. Sehingga *outlier* tergantikan oleh nilai distribusi data. Gambar 5 menunjukan hasil dari proses mengatasi outlier pada data.




Gambar 5. Hasil Mengatasi *Outlier*

2. Pemilihan Fitur

Metode korelasi Pearson adalah metode yang popular untuk menemukan fitur yang paling relevan (Khanam & Foo, 2021). Koefisien korelasi dihitung dalam metode ini, yang berkorelasi dengan atribut Output. Gambar 6 menunjukan korelasi seluruh variabel prediktor dengan variabel *Outcome*.

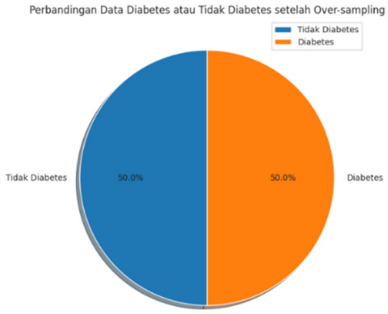



Gambar 6. Korelasi Atribut Dengan Variabel *Outcome*

Nilai 0,19 digunakan sebagai batasan untuk atribut yang relevan. Karenanya fitur DiabetesPedigreeFunction, BloodPresure, SkinThickness, dan Insulin dihapus. Glucose, Age, BMI, Insulin, dan Pregnancies adalah lima atribut pilihan yang paling relevan.

3. Penyeimbangan Data

Data yang tidak seimbang berdasarkan Gambar 3, perlu dilakukan penyeimbangan agar nilai akurasi tidak bias atau tidak hanya menguntungkan kelas mayoritas atau data terbanyak (Raghuwanshi & Shukla, 2020). Untuk mengatasi masalah ketidakseimbangan, digunakan Teknik SMOTE. Gambar 7 menunjukan hasil perbandingan data diabetes dan tidak diabetes setelah dilakukan proses SMOTE.




Gambar 7. Perbandingan Data Diabetes Dan Tidak Diabetes Setelah SMOTE

4. Pemisahan Data dan Pemodelan

a. Pemisahan *Data Train* dan *Data Test*

Sebelum dilakukan pemodelan, data dipisah menjadi *Train* dan *Test* menggunakan metode *train/test split*. Pemisahan masing-masing dibagi menjadi 80% untuk data training dan 20% untuk data testing.

b. Implementasi Model *Ensemble Machine Learning*

Data dilatih menggunakan model ML yang dipadukan dengan *ensemble learning*. Metode *ensemble* yang digunakan adalah *Hard Voting Classifier* dan *Bagging*.

Pada penerapan model *Ensemble Machine Learning* digunakan dua *Ensemble Machine Learning*, yaitu *Hard-Voting Classifier* dan *Bagging*. Pada penggunaan

JURNAL ELEKTRO & INFORMATIKA
SWADHARMA (JEIS)
VOLUME 05 NOMOR 01 JANUARI 2025
ISSN 2774 – 5775 | eISSN 2774-5767

model *Ensemble Machine Learning Hard-Voting*, keseluruhan model digunakan sebagai vote dalam pemungutan suara. Seluruh model akan divoting dan menghasilkan hasil prediksi voting yang akurat. Sedangkan pada penggunaan model *Machine Learning ensemble bagging*, seluruh model akan dilatih satu per satu dengan ensemble bagging. Hasil yang diperoleh dari model ini adalah nilai akurasi prediksi untuk masing-masing model.

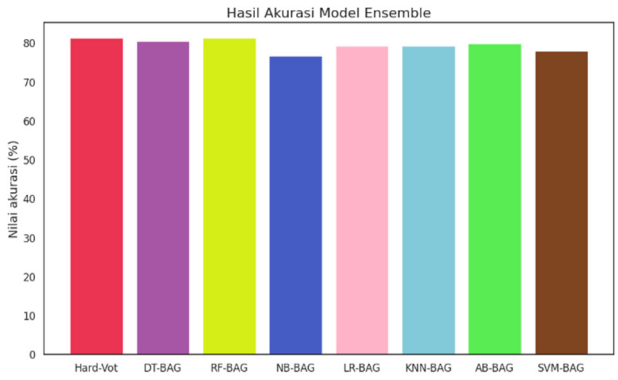
Akurasi output model digunakan untuk menguji *outcome* prediksi diabetes menggunakan *Ensemble Machine Learning*. Hasil akurasi prediksi diabetes menggunakan model *Ensemble Machine Learning* disajikan pada Tabel 2.

Tabel 2. Hasil Akurasi Yang Dihasilkan

| Model | Akurasi |
|---|---|
| Hard-Voting Classifier | 81.16% |
| Decision Tree - Bagging | 80.51% |
| Random Forest - Bagging | 81.16% |
| Naïve Bayes - Bagging | 76.62% |
| Logistic Regression - Bagging | 79.22% |
| KNN - Bagging | 79.22% |
| AdaBoost - Bagging | 79.88% |
| SVM - Bagging | 77.92% |

Berdasarkan Tabel 3, dapat dilihat bahwa akurasi dari seluruh model *ensemble* menghasilkan akurasi diatas 75%. Selain itu kedua model berupa Hard-Voting dan Random Forest dengan Bagging menunjukan hasil akurasi yang paling baik yaitu masing-masing sebesar 81.16%.

Kinerja seluruh model dengan *ensemble* diplot melalui grafik pada Gambar 8.




Gambar 8. Grafik Akurasi *Ensemble Machine Learning*

Berdasarkan Gambar 8, visualisasi bar chart dari hasil perbandingan *Ensemble Machine Learning* untuk memprediksi penyakit diabetes menunjukan bahwa model ensemble Hard Voting dan Random Forest dengan Bagging menunjukan hasil akurasi paling baik yaitu masing-masing sebesar 81.16%. Sedangkan model *ensemble* Naïve Bayes dengan Hard Voting menunjukan hasil akurasi paling rendah diantara model ensemble lainnya yaitu sebesar 76.62%.

Pada hasil penelitian ini terdapat tiga model *Ensemble Machine Learning* yang memiliki performa akurasi diatas 80% (>80%) yaitu Hard Voting, Random Forest dengan Bagging, serta Decision Tree dengan Bagging. Hasil akurasi ini memperlihatkan bahwa performansi model prediksi mengalami peningkatan pada akurasi. Hasil akurasi juga meningkat akibat penggunaan teknik SMOTE untuk menyeimbangkan data dan dapat mengurangi bias akurasi. Hal ini dapat dilihat beberapa model memiliki hasil akurasi yang meningkat. Seperti pada penelitian oleh (Khanam & Foo, 2021), dimana prediksi diabetes menggunakan Random Forest menghasilkan akurasi yaitu sebesar 77.14%. Sedangkan pada penelitian ini penggunaan Random Forest yang dibaurkan dengan *ensemble* Bagging menghasilkan akurasi sebesar 81.16%. Hal tersebut menunjukan bahwa penggunaan teknik *ensemble learning* dan teknik penyeimbangan data SMOTE dapat

mempengaruhi hasil performa dari pemodelan prediksi suatu model.

## D. PENUTUP

Deteksi diabetes adalah salah satu tantangan dalam bidang Kesehatan. Dalam penelitian ini, dilakukan perbandingan model ensemble learning dengan teknik SMOTE untuk prediksi diabetes. Digunakan tujuh algoritma ML yang di-ensemble-kan berupa DT, KNN, RF, NB, AB, LR, SVM pada dataset PID. Seluruh model menghasilkan akurasi lebih dari 75%. Hard-Voting dan Random Forest dengan Bagging memberikan hasil akurasi paling baik yaitu masing-masing 81.16%. Akurasi ditemukan untuk Hard-Voting, dan DT, RF, LR, KNN, dan AB yang seluruh modelnya dipadukan ensemble bagging lebih baik daripada akurasi penelitian menggunakjan LR dan SVM ~77-78% (Khanam & Foo, 2021), Max-Voting~78% (Rajendra & Latifi, 2021), Soft-Voting~79.08% (Kumari et al., 2021), LR~78,26%% (Joshi & Dhakal, 2021), dan LR~78.80% (Su et al., 2023).

Pada penelitian ini, penggunaan model *Ensemble Machine Learning* hanya untuk *Hard-Voting dan Bagging*. Selain itu penggunaan teknik SMOTE untuk penyeimbangan data juga diterapkan. Dengan demikian, tantangan penelitian untuk perjalanan selanjutnya adalah menggunakan peningkatan akurasi dengan model pembelajaran mesin ansambel lainnya (seperti *Boosting Ensemble*) dan menggunakan teknik penyeimbangan yang lainnya (seperti SMOTE borderline atau ADASYN). Penelitian seperti ini dapat dilakukan untuk melihat performansi lain yang dapat mempengaruhi akurasi model prediksi pada penyakit diabetes.

## E. DAFTAR PUSTAKA

Carter, J. A., Long, C. S., Smith, B. P., Smith, T. L., & Donati, G. L. (2019). Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes. *Expert Systems with Applications*, *115*, 245–255. https://doi.org/10.1016/j.eswa.2018.08.002

Chaki, J., Ganesh, S. T., Cidham, S. ., & Theertan, S. A. (2022). Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University - Computer and Information Sciences*, *34*(6), 3204–3225. https://doi.org/10.1016/j.jksuci.2020.06.013

Chang, V., Ganatra, M. A., Hall, K., Golightly, L., & Xu, Q. A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics*, *2*, 100118. https://doi.org/10.1016/j.health.2022.100118

Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., & Hussain, S. (2023). Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences*, *13*(6), 4006. https://doi.org/10.3390/app13064006

Joshi, R. D., & Dhakal, C. K. (2021). Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. *International Journal of Environmental Research and Public Health*, *18*(14), 7346. https://doi.org/10.3390/ijerph18147346

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*(4), 432–439. https://doi.org/10.1016/j.icte.2021.02.004

Kibria, H. B., Nahiduzzaman, M., Goni, M. O. F., Ahsan, M., & Haider, J. (2022). An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI. *Sensors*, *22*(19), 7268. https://doi.org/10.3390/s22197268

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *4*, 40–46. https://doi.org/10.1016/j.ijcce.2021.01.001

Makroum, M. A., Adda, M., Bouzouane, A., & Ibrahim, H. (2022). Machine Learning and Smart Devices for Diabetes Management: Systematic Review. *Sensors*, *22*(5), 1843. https://doi.org/10.3390/s22051843

Mistry, S., Riches, N. O., Gouripeddi, R., & Facelli, J. C. (2023). Environmental exposures in machine learning and data mining approaches to diabetes etiology: A scoping review. *Artificial Intelligence in Medicine*, *135*, 102461. https://doi.org/10.1016/j.artmed.2022.102461

Nicolucci, A., Romeo, L., Bernardini, M., Vespasiani, M., Rossi, M. C., Petrelli, M., Ceriello, A., Bartolo, P. Di, Frontoni, E., & Vespasiani, G. (2022). Prediction of complications of type 2 Diabetes: A Machine learning approach. *Diabetes Research and Clinical Practice*, *190*, 110013. https://doi.org/10.1016/j.diabres.2022.110013

Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, *220*, 106773. https://doi.org/10.1016/j.cmpb.2022.106773

Plotnikova, V., Dumas, M., & Milani, F. P. (2022). Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements. *Data & Knowledge Engineering*, *139*, 102013. https://doi.org/10.1016/j.datak.2022.102013

Raghuwanshi, B. S., & Shukla, S. (2020). SMOTE based class-specific extreme learning machine for imbalanced learning. *Knowledge-Based Systems*, *187*, 104814. https://doi.org/10.1016/j.knosys.2019.06.022

Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. *Computer Methods and Programs in Biomedicine Update*, *1*, 100032. https://doi.org/10.1016/j.cmpbup.2021.100032

Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. *Measurement: Sensors*, *25*, 100605. https://doi.org/10.1016/j.measen.2022.100605

Retta, E., Kusumajaya, H., & Arjuna, A. (2023). Faktor – faktor yang Berhubungan dengan Pemilihan Pengobatan Herbal pada Pasien Diabetes Mellitus. *Jurnal Penelitian Perawat Profesional*, *5*(4), 1541–1552. https://doi.org/10.37287/jppp.v5i4.1891

Su, Y., Huang, C., Yin, W., Lyu, X., Ma, L., & Tao, Z. (2023). Diabetes Mellitus risk prediction using age adaptation models. *Biomedical Signal Processing and Control*, *80*(2), 104381. https://doi.org/10.1016/j.bspc.2022.104381

Suprayitna, M., Hajri, Z., Fatmawati, B. R., Prihatin, K., & Nadrati, B. (2023).

Deteksi Dini Diabetes Mellitus (DM) Melalui "Mawas DM." *BERNAS: Jurnal Pengabdian Kepada Masyarakat*, *4*(3), 2291–2296. https://doi.org/10.31949/jb.v4i3.5655

Tran, V.-L., & Kim, J.-K. (2023). Ensemble machine learning-based models for estimating the transfer length of strands in PSC beams. *Expert Systems with Applications*, *221*, 119768. https://doi.org/10.1016/j.eswa.2023.119768

Tsybikova, E. B., Kotlovsky, M. Y., & Kaigorodova, T. V. (2024). Diabetes Mellitus and Its Complications: Current State. Analytical Review. *Social Aspects of Population Health*, *70*(3), 13. https://doi.org/10.21045/2071-5021-2024-70-3-13