e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 8 No. 1, Juni, 2024, **Hal. 290-299**
DOI: 10.29408/edumatic.v8i1.25838

Peningkatan Performa Model Hard Voting Classifier dengan Teknik Oversampling ADASYN pada Penyakit Diabetes

Muhammad Ikhsan Anugrah [1,*] , Junta Zeniarja [1] , Dicky Setiawan [1]

[1] Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Indonesia
[*] Correspondence: 111202012708@mhs.dinus.ac.id

Received: 30 April 2024 | Revised: 3 Mei 2024 | Accepted: 20 Mei 2024 | Published: 20 Juni 2024

Abstrak
Diabetes menjadi penyakit kronis yang timbul dari kelebihan kadar gula yang ada di dalam tubuh dan kurangnya intensitas olahraga sehingga terjadi penumpukan di dalam darah. Indonesia menempati peringkat kelima sebagai negara dengan penyandang diabetes terbesar berdasarkan laporan dari International Diabetes Federation (IDF). Penyebabnya para penyandang diabetes belum menyadari bahwa mereka mengidap penyakit diabetes, sehingga perlu adanya deteksi dini dalam mengetahui hal ini. Tujuan penelitian ini melakukan peningkatan performa model *Hard Voting Classifier* yang menggabungkan algoritma *Decision Tree*, *Random Forest*, dan *XGBoost* dengan teknik *oversampling* ADASYN yang menangani ketidakseimbangan data pada prediksi penyakit diabetes. Penelitian ini menggunakan data informasi pasien dengan jumlah 1000 data dan 14 fitur dari laboratorium Rumah Sakit Medical City, Irak. Hasil dari penelitian ini terjadi peningkatan performa model prediksi dengan nilai akurasi 99,0%, presisi 99,1%, *recall* 9,.0%, dan *f1-score* 98.98% tanpa menggunakan ADASYN. Kemudian mendapatkan nilai akurasi 99,8%, presisi 99,8%, *recall* 99,8%, dan *f1-score* 99,8% dengan menggunakan ADASYN sebagai teknik *oversampling*. Hal ini menunjukkan bahwa ada kenaikan performa model *Hard Voting Classifier* sehingga menghasilkan prediksi penyakit diabetes dengan akurat yang mana tingkat kebenaran prediksi penyakit diabetes sangat baik.

Kata kunci: adasyn; *hard voting classifier;* penyakit diabetes

Abstract
Diabetes is a chronic disease that arises from excess sugar levels in the body and lack of exercise intensity resulting in a buildup in the blood. Indonesia ranks fifth as the country with the largest number of people with diabetes based on a report from the International Diabetes Federation (IDF). The reason is that people with diabetes do not realize that they have diabetes, so there is a need for early detection in knowing this. The purpose of this research is to improve the performance of the Hard Voting Classifier model combining the Decision Tree, Random Forest, and XGBoost algorithms with the ADASYN oversampling technique that handles data imbalance in diabetes prediction. This study uses patient information data with a total of 1000 data and 14 features from the Medical City Hospital laboratory, Iraq. The results of this study show an increase in the performance of the prediction model with an accuracy value of 99.0%, precision 99.1%, recall 99.0%, and f1-score 98.98% without using ADASYN. Then get an accuracy value of 99.8%, precision 99.8%, recall 99.8%, and f1-score 99.8% by using ADASYN as an oversampling technique. This shows that there is an increase in the performance of the Hard Voting Classifier model so that it produces accurate predictions of diabetes, where the correctness of diabetes prediction is very good.

Keywords: *adasyn; hard voting classifier; diabetes disease*

PENDAHULUAN

Diabetes merupakan kondisi kronis yang timbul akibat kekurangan hormon insulin yang merupakan produk dari pankreas (Nurani & Waluyo, 2022). Penyakit diabetes cenderung akan lebih cepat timbul ketika seseorang mengkonsumsi gula berlebihan dan kurang olahraga yang mengakibatkan penumpukan gula dalam darah (Gunawan et al., 2020) Hingga saat ini diabetes masih dicatat menjadi penyebab kematian yang tinggi di dunia. Setiap tahunnya mengalami peningkatan yang cukup signifikan (Irwansyah & Kasim, 2021). Menurut federasi diabetes Internasional International Diabetes Federation (IDF) (Yahyaoui et al., 2019), jumlah penderita diabetes akan terus bertambah setiap tahunnya hingga menyentuh angka 592 juta jiwa di tahun 2035. Diabetes juga dapat menyebabkan komplikasi serius terhadap penyakit lain seperti stroke, hiperosmolar non ketotik, dan penyakit kardiovaskular (Nurani & Waluyo, 2022). Indonesia berada di peringkat ke-5 sebagai negara dengan penderita penyakit diabetes terbanyak di dunia sebesar 28,6 juta jiwa berdasarkan hasil pengumpulan data dari IDF. IDF juga melaporkan bahwa 10,5% populasi orang dewasa dari umur 20 hingga 79 tahun menderita diabetes, namun setengahnya belum menyadari mereka mengidap penyakit diabetes. Dari permasalahan tersebut perlu adanya deteksi sejak dini agar dapat dilakukan penanganan dengan cepat terhadap pengidap diabetes (Gunawan et al., 2020).

Model prediksi diabetes perlu dikembangkan agar dapat memberikan dampak yang signifikan dalam mendeteksi penyakit diabetes sejak dini. Dengan kemajuan teknologi yang begitu pesat terutama di bidang kecerdasan buatan model prediksi dapat dikembangkan dengan bantuan pembelajaran komputer tentang *machine learning* (Manongga et al., 2022). *Machine learning* menjadi salah satu topik dari kecerdasan buatan yang mempelajari tentang data dalam statistika sehingga dapat membantu membuat berbagai model klasifikasi (Armansyah & Ramli, 2022; Hidayat et al., 2021; Prayoga et al., 2023; Wedashwara et al., 2022) dan prediksi (Amri et al., 2023; Armansyah & Ramli, 2022; Efriadi et al., 2023; Naufal & Kusuma, 2023). Dalam penelitian (Kurniawan et al., 2023) menyebutkan *machine learning* dapat memprediksi sebuah permasalahan dalam industri melalui pembelajaran data yang telah disajikan dan mengenai pola-pola yang ada. Dari berbagai pernyataan yang telah dikemukakan oleh berbagai penelitian, *machine learning* menjadi solusi yang baik untuk bidang kesehatan dalam memberikan model prediksi kepada pasien yang terindikasi mengalami diabetes.

Berbagai model telah dikembangkan dalam membuat prediksi penyakit diabetes, namun model hanya menggunakan algoritma tunggal. Dari berbagai percobaan yang dilakukan dengan menggunakan algoritma tunggal dalam memprediksi penyakit diabetes belum mendapatkan hasil yang terbaik. Hal ini disebabkan setiap algoritma bergantung pada karakteristik data. Dengan perbedaan karakteristik setiap data perlu adanya model yang dapat menyelesaikan permasalahan tersebut, sehingga model yang dibuat bisa membaca karakteristik data yang berbeda-beda. Upaya dalam mengatasi permasalahan tersebut dengan menggabungkan algoritma tunggal menjadi satu model yang baik dalam menangani perbedaan karakteristik data. seperti pada penelitian yang dilakukan oleh (Kibria et al., 2022) yang mencoba melakukan *ensemble learning* atau menggabungkan algoritma pohon keputusan dalam memprediksi penyakit diabetes, dari percobaan tersebut teknik *ensemble learning* mendapatkan nilai akurasi tinggi daripada menggunakan algoritma tunggal. Dengan menggabungkan beberapa algoritma, diharapkan dapat membantu dengan kelebihan masing-masing algoritma dan mengurangi kelemahan algoritma yang digabungkan(Masacgi & Rohman, 2023). *Hard Voting Classifier* adalah salah satu metode dari teknik *ensemble* yang menggabungkan beberapa algoritma dengan memilih label kelas mayoritas (Altaf et al., 2022).

Ketidakseimbangan data juga menjadi salah satu masalah dalam penerapan model prediksi. *Imbalance data* atau data tidak seimbang terjadi ketika terdapat kelas yang dominan dari beberapa kelas yang ada sebagai kelas mayoritas, sedangkan kelas minoritas menjadi kejadian yang jarang terjadi (Sinaga & Agustian, 2022). Teknik *oversampling* menjadi sebuah

teknik dalam menyelesaikan ketidakseimbangan data dalam *machine learning*. Teknik ini membuat kelas minoritas menyamakan distribusi kelas dengan kelas mayoritas. Teknik *oversampling* yang digunakan dalam penelitian ini adalah *Adaptive Synthetic Sampling* (ADASYN) yang mana teknik ini memanfaatkan bobot distribusi pada data kelas minoritas berdasarkan tingkat kesulitan pembelajaran data dan mode (Wicaksono et al., 2024).

Pada penelitian sebelumnya tentang prediksi *machine learning* dengan dilakukan menggunakan algoritma individu dalam memperdiksi diabetes menggunakan algoritma Decision Tree dan Naïve Bayes. Memberikan hasil akurasi 95,58% untuk algoritma Decision Tree dan 87,69% untuk algoritma Naïve Bayes (Permana & Patwari, 2021). Penelitian lain dilakukan oleh (Depari et al., 2022) tentang penerapan prediksi penyakit jantung menggunakan algortima *Decision Tree, Naive Bayes* dan *Random Forest*, dimana pada penelitian mereka meghasilkan *Decision Tree* akurasi sebesar 71%, *Naive Bayes* 72% dan *Random Forest* 75%. Penelitian lainya (Andryan & Fajri, 2022) mencoba mengkomparasi antara algoritma *Support Vector Machine* (SVM) dan *XGBoost* untuk diagnosis penyakit kangker payudara mengahsilkan nilai akurasi pada SVM sebesar 90,24% dan *XGBoost* sebesar 95,12%.

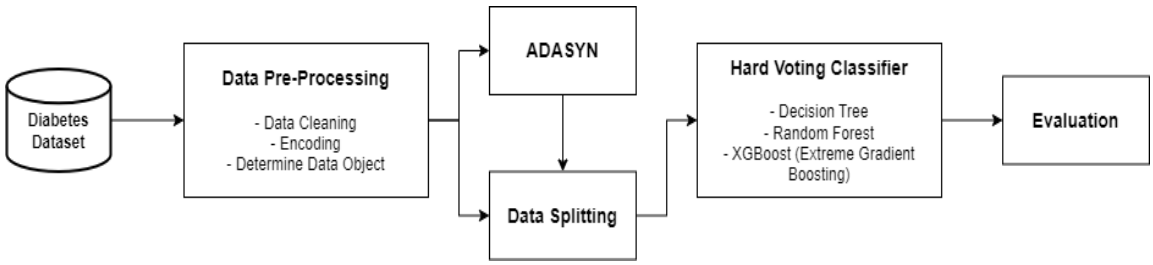
Pada penelitian (Sharma & Singhal, 2023) menerapkan algoritma *Xgboost Classifier* untuk memprediksi penyakit jantung. Pada penelitian tersebut membandingkan penggunaan teknik *oversampling* SMOTE dan ADAYSN dalam menangani kasus ketidakseimbanagan data. Hasil yang didapatkan adalah teknik*A daptive Synthetic Sampling* (ADASYN) mendapatkan nilai akurasi tertinggi menggunakan teknik *oversampling* sebesar 94,7%. ADASYN juga memberikan hasil yang tinggi terhadap permasalahan ketidakseimbangan data pada penelitian (Kaope & Pristyanto, 2023) yang dikombinasikan dengan algoritma *Random Forest, AdaBoost*, dan *K-Nearest Neighbor* menghasilkan nilai akurasi tertinggi pada algoritma Random Forest dari 86% menjadi 93%. Dari temuan tersebut ADASYN menjadi teknik *oversampling* yang cocok dikombinasikan dengan algoritma yang tergabung dalam pohon keputusan seperti *Random Forest* dan *XGBoost*.

Berbagai penelitian tersebut masih menggunakan algoritma individu yang masih tergolong rentan dalam mengatasi data dengan jumlah besar dan beragam jenis data. Sehingga perlu adanya model yang bisa menangani data yang besar dan beragam. Salah satu penlitian dari (Atif et al., 2022) membahas tentang meningkatkan model prediksi *machine learning* dengan metode *Hard Voting Classifier*, algoritma yang digunakan adalah *Logistic Regression*, Decision Tree, dan Support Vector Machine, dengan menggunakan dua dataset yang berbeda yaitu PIMA Indian Diabetes dengan hasil akurasi 94.24% dan Early Stage Diabetes Risk Prediction mendapatkan nilai akurasi 81,12%. Kemudian pada penelitian(Maniruzzaman et al., 2020) mencoba untuk mengkombinasikan berbagai model *machine learning*, model yang digunakan dalam penelitian tersebut adalah Logistic Regression, Odds Ratio, Naive Bayes, Decision Tree, Adaboost, dan Random Forest, dari hasil percobaan kombinasi dilakukan menghasilkan kombinasi yang terbaik yaitu kombinasi model Logistic Regression dan Random Forest dengan tingkat akurasi sebesar 94,25%. Penerapan penggabungan model dapat menutipi keruangan dari model individu yang belum memberikan performa terbaik untuk data yang beragam. Namun terdapat masalah terhadap data dengan jumlah besar yang mana sering terdapat ketidakseimbangan data dimana hal tersebut juga dapat mempengaruhi performa model.

Penelitian ini bertujuan melakukan peningkatan performa model *Hard Voting Classifier* dalam memprediksi penyakit diabetes dengan mengatasi permasalahan yang ada pada dataset dari penelitian ini dengan menggunakan teknik *oversampling* ADASYN. Dengan menerapkan penanganan ketidakseimbangan data terdapat peningkatan hasil akurasi, presisi, *recall*, dan *f1-score* model prediksi. Dengan adanya peningkatan nilai akurasi, presisi, *recall*, dan *f1-score* model *Hard Voting Classifier* dalam memprediksi penyakit diabetes dapat memberikan acuan terhadap pelaku di bidang kesehatan untuk mendeteksi penyakit diabetes sejak dini.

METODE

Penelitian ini menggunakan teknik eksperimen yang mana bertujuan untuk meningkatkan performa model prediksi penyakit diabetes dengan mengatasi ketidakseimbangan data pada dataset Diabetes Dataset. Model prediksi yang digunakan adalah *Hard Voting Classifier* yang merupakan model dengan metode *ensemble* atau penggabungan beberapa algoritma prediksi. Algoritma yang digunakan dalam metode penggabungan diantaranya adalah algoritma *Decision Tree, Random Forest*, dan *XGBoost*. Untuk mengatasi ketidakseimbangan data sehingga dapat meningkatkan performa model yang telah dibuat penelitian ini menggunakan teknik *oversampling* ADASYN. Setelah itu akan dilihat perbandingan performa model sebelum dan sesudah menggunakan teknik *oversampling* untuk melihat seberapa besar peningkatan performa model prediksi. Adapun langkah - langkah dalam penelitian ini terlihat pada gambar 1.



Gambar 1. Tahapan penelitian

Pada tahapan awal dilakukan pengumpulan data dengan menggunakan dataset yang relevan seperti Diabetes Dataset yang dihimpun dari laman Kaggle. Dataset tersebut dikumpulkan oleh rumah sakit Pusat Spesialis Endokrinologi dan Rumah Sakit Pendidikan Diabetes-Al-Kindy dari masyarakat Irak sebanyak 1000 data dan 14 fitur. Setelah melewati tahapan pengumpulan data perlu adanya proses *pre-processing* agar dapat memastikan dataset seragam dan mempermudah untuk proses selanjutnya (Fajri et al., 2022). Melalui tahap ini, dilakukan pembersihan data yang terdapat *missing value* dan duplikasi data. Tidak hanya itu penelitian ini melibatkan proses *encoding* untuk mengubah beberapa data yang bersifat kategorikal menjadi numerik. Kedua hal tersebut berperan penting agar hasil yang didapat melalui proses *modelling* menjadi lebih maksimal. Diakhiri dengan *determine data object* untuk menampilkan sebaran data dari 14 fitur. Setelah melalui tahap *pre-processing*, masuk ke tahap penyelesaian masalah di ketidakseimbangan data yang mana peran dari penyelesaian ini dapat meningkatkan performa model yang akan digunakan, teknik yang diterapkan dalam mengatasi ketidakseimbangan data yaitu *Adaptive Synthetic Sampling* (ADASYN), teknik ini menyeimbangkan data dari fitur class yang menjadi target dari model yang akan dibuat.

Sebelum masuk ke tahapan *modelling*, data akan dibagi menjadi dua yaitu *data test* sebesar 20% dan *data train* sebesar 80% kemudian akan diproses ke tahap *modelling* menggunakan teknik *ensemble* dengan metode *Hard Voting Classifier.*Algoritma yang digabungkan dalam penelitian ini diantaranya Decision Tree, Random Forest, XGBoost. Ketiga algoritma tersebut akan digunakan dalam pembuatan model prediksi penyakit diabetes. Tahapa terakhir merupakan evaluasi dalam penelitian ini, evaluasi ini bertujuan untuk mengetahui kinerja model prediksi dengan menggunakan *confusion matrix* yang akan membrikan nilai *accuracy, precision, recall, dan F1 – Score* (Setiawan et al., 2024).

HASIL DAN PEMBAHASAN

Hasil

Pada penelitian ini menggunakan dataset yang telah dihimpun dari Diabetes dataset merupakan dataset umum untuk digunakan dalam penelitian terkait diagnosis penyakit diabetes. Dengan 1000 rekaman pasien dan 14 fitur dataset tersebut memberikan variasi dari berbagai isi data pasien yang dapat diidentifikasi dan memungkinkan untuk diuji dengan model *Hard Voting Classifier*. Informasi lebih rinci terdapat pada tabel 1 dimana menujukkan 14 fitur yang akan digunakan dalam model prediksi penyakit diabetes. Kemudian pada tabel 2 menunjukkan pembagian fitur *class* yang menjadi target dari model prediksi dimana terbagi menjadi tiga class yaitu bukan penderita diabtes (0), diperdiksi penderita diabetes (1), dan pasien yang mengalami penyakit diabetes (2).

Tabel 1. Diabetes dataset

Fitur	Keterangan
id	Id data pasien
no pation	Nomor setiap pasien
gender	Jenis kelamin pasien
age	Umur pasien
urea	Kadar urea pada pasien
cr	Rasio kreatinin
hba1c	Rata - rata glukosa
chol	Tingkat kolesterol dalam darah
tg	Lemak dalam darah
hdl	Jumlah kolesterol baik dalam darah
ldl	Jumlah kolesterol jahat dalam darah
vldl	Kadar lipoprotein berdensitas sangat rendah
bmi	Berat badan pasien
class	Indikator diabetes

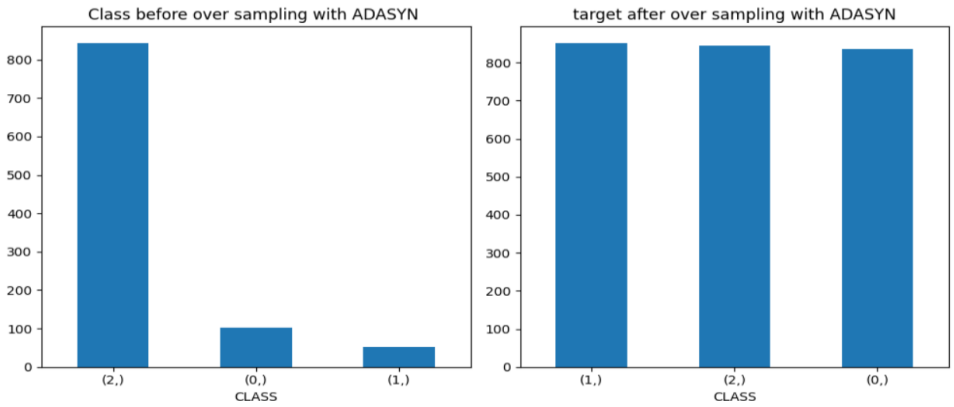
Hasil dari eksperimen yang telah dilakukan penelitian ini dimana pada tahapan *pre-processing* menghasilkan data yang digunakan dalam penelitian ini tidak terdapat *missing value* dan *duplicate value*, yang mana tidak terdapat nilai kosong dan tidak memiliki dua nilai yang sama. Setelah menyelesaikan tahapan *data cleaning* pada *pre-processing* dilanjut dengan proses *encoding* untuk menyesuaikan beberapa tipe data yang ada. Terdapat dua fitur yang bertipe data *object* sehingga tidak dapat memberikan hasil dalam eksperimen model kemudian, dilakukan perubahan tipe data pada kedua fitur tersebut. Langkah terakhir adalah *determine data object* untuk mengidentifikasi setiap fitur data serta memilih fitur yang akan digunakan ke proses selanjutnya. Kemudian proses penyeimbangan data pada tahapan *oversampling* menggunakan metode ADASYN menghasilkan seperti pada gambar 2

Tabel 2. Pembagian jumlah fitur *class*

Class	Jumlah	Keterangan
0	103	Bukan Penderita Diabetes
1	53	Diprediksi Penderita Diabetes
2	844	Penderita Diabetes

.

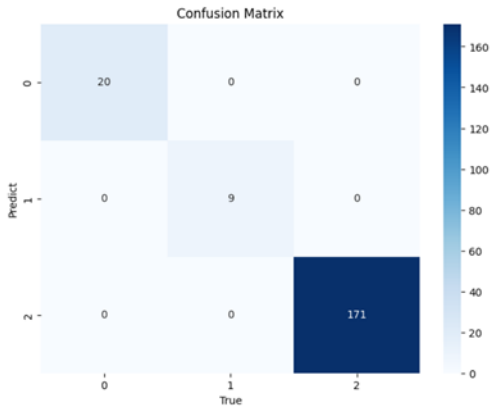
Gambar 2 menunjukkan sebaran data dari fitur class yang menjadi target sebelum dilakukannya penyeimbangan data dengan teknik *oversampling* ADASYN dan sesudah menggunakan teknik *oversampling* ADASYN menunjukkan hasil setiap *class* yang sebelumnya memiliki data yang berbeda menjadi memiliki data yang seimbang dengan jumlah masing-masing *class* sebanyak 836 untuk *class* 0,852 untuk *class* 1, dan 844 untuk *class* 2 karena sangat sedikit selisih dari masing - masing *class*. Hasil ini menunjukkan penggunaan teknik *oversampling* ADASYN memberikan peran signifikan untuk mengatasi ketidakseimbangan data apalagi untuk jumlah data pada class minoritas. Dan berikut adalah tabel yang menunjukkan hasil akurasi, presisi, *recall*, dan *f1-score* dari eksperimen yang dilakukan yaitu model tanpa teknik *oversampling* dan model menggunakan teknik *oversampling*



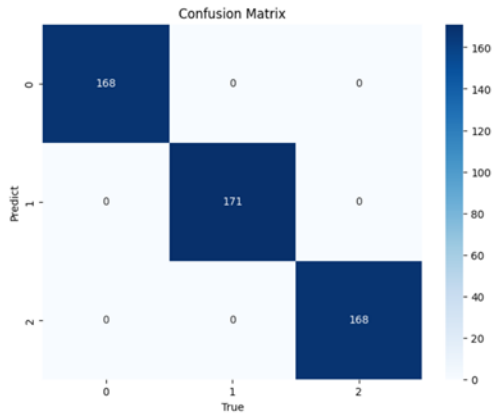
Gambar 2. Distribusi *class* sebelum dan sesudah *oversampling*

Tabel 3. Hasil akurasi, pesisi, *recall*, dan *f1-score*

Teknik	Akurasi	Presisi	Recall	F1-Score
Hard Voting Classifier	99,0%	99,1%	99,0 %	98,98%
Hard Voting Classifier +ADASYN	99,8%	99,8%	99,8%	99,8%



Gambar 3. *Confusion Matrix Hard Voting Classifier No* ADASYN

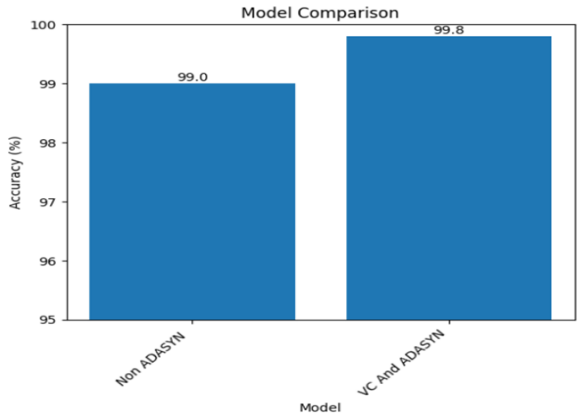


Gambar 4. *Confusion Matrix Hard Voting Classifier* + ADASYN

Pada tabel 3 menunjukkan hasil dari matriks evaluasi yang digunakan untuk mengevaluasi performa model yang telah dibuat. Dapat dilihat terjadi kenaikan nilai akurasi dari 99,0% menjadi 99,8%, presisi dari 99,1% menjadi 99,8%, recall dari 99,0% menjadi

99,8%, dan f1-score dari 98,98% menjadi 99,8% setelah dilakukan teknik *oversampling* ADASYN. Kemudian hasil dari *confusion matrix* dapat dilihat pada gambar 3 dan 4.

Pada gambar 3 dan 4 penggunaan teknik *oversampling* ADASYN sebagai metode dalam mengatasi masalah ketidakseimbangan data berhasil memprediksi dengan tepat pada *class 0* sebesar 168 kasus, *class 1* sebesar 171 kasus, dan *class 2* sebesar 168 kasus. Untuk memperjelas perbandingan yang dialami oleh model prediksi sebelum dan setelah menggunakan *oversampling* pada gambar 5 dan terjadi peningkatan akurasi dari 99,0% menjadi 99,8%.



Gambar 5. Hasil perbandingan model

Pembahasan

Penelitian ini menghasilkan peningkatan model prediksi dari penyakit diabetes yang menggunakan Diabetes Dataset dimana hal ini dilakukan untuk mencoba membantu pengembangan dari pembelajaran *machine learning* yang digunakan dalam bidang kesehatan. Tentunya dari peningkatan performa model yang dilakukan dapat memberikan penguatan kepada pihak yang memerlukan performa yang lebih baik dari prediksi yang telah dibuat.

Berdasarkan hasil eksperimen pertama hanya menggunakan model prediksi *Hard Voting Classifier* yang merupakan penggabungan algoritma dalam memprediksi penyakit diabetes mendapatkan hasil akurasi 99,0%, presisi 99,1%, recall 99,0%, dan f1-score 98,98% dalam memprediksi penyakit diabetes. Hasil tersebut didapatkan karena model prediksi Hard Voting Classifier dapat mengurangi bias yang terjadi pada penggunaan model tunggal dalam memprediksi penyakit diabetes (Atif et al., 2022).

Hasil temuan pada penelitian yang dilakukan oleh Altaf et al. (2022) juga menggunakan teknik *Hard Voting Classifier* guna mengidentifikasi penyakit dari berbagai dataset, dimana menghasilkan akurasi yang tinggi terhadap identifikasi penyakit. Namun pada penelitian tersebut masih terbatas pada percobaan penyakit di luar Indian Liver Patient Dataset (ILPD) dan dataset PIMA Indians, sehingga belum maksimal untuk penyakit diabetes. Tidak hanya itu penelitian tersebut juga masih belum menangani kasus ketidakseimbangan data yang berpengaruh untuk performa model.

Sementara pada percobaan kedua pada penelitian ini memasukkan teknik *oversampling* ADASYN untuk menangani ketidakseimbangan data dan menghasilkan peningkatan performa yang mendapatkan nilai akurasi 99,8%, presisi 99,8%, *recall* 99,8%, dan *f1-score* 99,8%. Hasil tersebut sejalan dengan efektifnya penggunaan teknik *oversampling* ADASYN dalam menangani kasus ketidakseimbangan data yang dapat mempengaruhi performa model sehingga belum maksimal. Performa tidak maksimal tersebut disebabkan oleh model kesulitan dalam mempelajari pola penting di kelas minoritas karena keterbatasan contoh dari ketidakseimbangan data.

Hasil dari kedua percobaan tersebut dengan mengatasi ketidakseimbangan data yang ada dengan teknik *oversampling* ADASYN terhadap model prediksi penyakit diabetes *Hard Voting*

Classifier dapat memberikan peningkatan nilai akurasi, presisi, *recall*, dan *f1-score* yang mana bisa dilihat pada tabel 3 dan gambar 5. Hal ini tentunya dapat meyakinkan kepada pihak yang akan menggunakan perdiksi ini untuk mendeteksi sejak dini pasien yang berersiko mengalami penyakit diabetes melalui model prediksi karena memiliki model prediksi yang akurat.

SIMPULAN

Berdasarkan hasil penelitian, implementasi model *Hard Voting Classifier* tanpa menggunakan teknik *oversampling* ADASYN memperoleh nilai akurasi 99,0%, presisi 99,1%, recall 99,0%, dan *f1-score* 98,98% sedangkan dengan menerapkan teknik *oversampling* untuk mengatasi ketidakseimbanagn data menghasilka nilai akurasi 99,8%, presisi 99,8%, *recall* 99,8%, dan *f1-score* 99,8%. Dari hasil tersebut mengatasi masalah ketidakseimbangan data dapat memberikan hasil yang maksimal pada performa model. Berkaitan dengan meningkatnya nilai akurasi pada model prediksi penyakit diabtes dapat meberikan hasil yang akurat dalam melakukan prediksi penyakit diabetes sehingga dapat dilakukan pencegahan sejak dini untuk penyakit diabetes.

REFERENSI

Altaf, I., Butt, M. A., & Zaman, M. (2022). Hard Voting Meta Classifier for Disease Diagnosis using Mean Decrease in Impurity for Tree Models. *Review of Computer Engineering Research*, *9*(2), 71–82. https://doi.org/10.18488/76.v9i2.3037

Amri, Z., Kusrini, K., & Kusnawi, K. (2023). Prediksi Tingkat Kelulusan Mahasiswa menggunakan Algoritma Naïve Bayes, Decision Tree, ANN, KNN, dan SVM. *Edumatic: Jurnal Pendidikan Informatika*, *7*(2), 187-196. https://doi.org/10.29408/edumatic.v7i2.18620

Andryan, M. R., & Fajri, M. (2022). Komparasi Kinerja Algoritma Xgboost Dan Algoritma Support Vector Machine (Svm) Untuk Diagnosa Penyakit Kanker Payudara. *JIKO (Jurnal Informatika dan Komputer)*, *6*(1), 1–5. https://doi.org/10.26798/jiko.v6i1.500

Armansyah, A., & Ramli, R. K. (2022). Model Prediksi Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes. *Edumatic: Jurnal Pendidikan Informatika*, *6*(1), 1-10. https://doi.org/10.29408/edumatic.v6i1.4789

Atif, M., Anwer, F., & Talib, F. (2022). An Ensemble Learning Approach for Effective Prediction of Diabetes Mellitus Using Hard Voting Classifier. *Indian Journal Of Science And Technology*, *15*(39), 1978–1986. https://doi.org/10.17485/IJST/v15i39.1520

Depari, D. H., Widiastiwi, Y., & Santoni, M. M. (2022). Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung. *Informatik : Jurnal Ilmu Komputer*, *18*(3), 239–248. https://doi.org/10.52958/iftk.v18i3.4694

Efriadi, D., Rahmaddeni, R., Agustin, A., & Junadhi, J. (2022). Prediksi Penambahan Piutang Iuran Jaminan Sosial Ketenagakerjaan menggunakan Algoritma K-Nearest Neighbor. *Edumatic: Jurnal Pendidikan Informatika*, *6*(1), 49-57. https://doi.org/10.29408/edumatic.v6i1.5255

Fajri, F., Tholib, A., & Yuliana, W. (2022). Application of Machine Learning Algorithm for Determining Elective Courses in Informatics Study Program. *Jurnal Teknik Informatika dan Sistem Informasi*, *8*(3), 485–496. https://doi.org/10.28932/jutisi.v8i3.3990

Gunawan, M. I., Sugiarto, D., & Mardianto, I. (2020). Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression. *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, *6*(3), 280–284. https://doi.org/10.26418/jp.v6i3.40718

Hidayat, W., Utami, E., Iskandar, A. F., Hartanto, A. D., & Prasetio, A. B. (2021). Perbandingan Performansi Model pada Algoritma K-NN terhadap Klasifikasi Berita

Fakta Hoaks Tentang Covid-19. *Edumatic: Jurnal Pendidikan Informatika*, *5*(2), 167-176. https://doi.org/10.29408/edumatic.v5i2.3664

Irwansyah, I., & Kasim, I. S. (2021). Indentifikasi Keterkaitan Lifestyle Dengan Risiko Diabetes Melitus. *Jurnal Ilmiah Kesehatan Sandi Husada*, *10*(1), 62–69. https://doi.org/10.35816/jiskh.v10i1.511

Kaope, C., & Pristyanto, Y. (2023). The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *22*(2), 227–238. https://doi.org/10.29408/edumatic.v5i2.3664

Kibria, H. B., Nahiduzzaman, M., Goni, Md. 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

Kurniawan, R., Wintoro, P. B., Mulyani, Y., & Komarudin, M. (2023). Implementasi Arsitektur Xception Pada Model Machine Learning Klasifikasi Sampah Anorganik. *Jurnal Informatika dan Teknik Elektro Terapan*, *11*(2), 233–236. https://doi.org/10.23960/jitet.v11i2.3034

Maniruzzaman, Md., Rahman, Md. J., Ahammed, B., & Abedin, Md. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems*, *8*(1), 7–24. https://doi.org/10.1007/s13755-019-0095-z

Manongga, D., Rahardja, U., Sembiring, I., Lutfiani, N., & Yadila, A. B. (2022). Dampak Kecerdasan Buatan Bagi Pendidikan. *ADI Bisnis Digital Interdisiplin Jurnal*, *3*(2), 41–55. https://doi.org/10.34306/abdi.v3i2.792

Masacgi, G. N., & Rohman, M. S. (2023). Optimasi Model Algoritma Klasifikasi menggunakan Metode Bagging pada Stunting Balita. *Edumatic: Jurnal Pendidikan Informatika*, *7*(2), 455–464. https://doi.org/10.29408/edumatic.v7i2.23812

Naufal, M. F., & Kusuma, S. F. (2023). Analisis Perbandingan Algoritma Machine Learning dan Deep Learning untuk Klasifikasi Citra Sistem Isyarat Bahasa Indonesia (SIBI). *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *10*(4), 873–882. https://doi.org/10.25126/jtiik.20241046823

Nurani, R. D., & Waluyo, A. (2022). Edukasi Senam Kaki Diabetes Dalam Pencegahan Komplikasi Penderita Diabetes Mellitus. *Jurnal Batikmu*, *2*(1), 86–89. https://doi.org/10.48144/batikmu.v2i1.1180

Permana, B. C., & Patwari, I. D. (2021). Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naïve Bayes Untuk Prediksi Penyakit Diabetes. *Infotek : Jurnal Informatika dan Teknologi*, *4*(1), 63–69. https://doi.org/10.29408/jit.v4i1.2994

Prayoga, P. R., Purnawansyah, P., Hasanuddin, T., & Darwis, H. (2023). Klasifikasi Daun Herbal Menggunakan K-Nearest Neighbor dan Support Vector Machine dengan Fitur Fourier Descriptor. *Edumatic: Jurnal Pendidikan Informatika*, *7*(1), 160-168. https://doi.org/10.29408/edumatic.v7i1.17521

Setiawan, D., Nugraha, A., & Luthfiarta, A. (2024). Komparasi Teknik Feature Selection Dalam Klasifikasi Serangan IoT Menggunakan Algoritma Decision Tree. *Jurnal Media Informatika Budidarma*, *8*(1), 83–93.

Sharma, S., & Singhal, A. (2023, November). A Novel Heart Disease Prediction System Using XGBoost Classifier Coupled With ADASYN SMOTE. In *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 76-81). IEEE. https://doi.org/10.1109/ICCCIS60361.2023.10425095

Sinaga, H. H., & Agustian, S. (2022). Pebandingan Metode Decision Tree dan XGBoost untuk Klasifikasi Sentimen Vaksin Covid-19 di Twitter. *Jurnal Nasional Teknologi dan Sistem Informasi*, *8*(3), 107–114. https://doi.org/10.25077/TEKNOSI.v8i3.2022.107-114

Wedashwara, W., Hidayat, A., Irmawati, B., & Zubaidi, A. (2022). Klasifikasi Teks menggunakan Genetic Programming dengan Implementasi Web Scraping dan Map Reduce. *Edumatic: Jurnal Pendidikan Informatika*, *6*(1), 58-67. https://doi.org/10.29408/edumatic.v6i1.5274

Wicaksono, D. F., Basuki, R. S., & Setiawan, D. (2024). Peningkatan Performa Model Machine Learning XGBoost Classifier melalui Teknik Oversampling dalam Prediksi Penyakit AIDS. *Jurnal Media Informatika Budidarma*, *8*(2), 736–747.

Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019). A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. *International Informatics and Software Engineering Conference (UBMYK)*, 1–4. IEEE. https://doi.org/10.1109/UBMYK48245.2019.8965556