

Penerapan Metode *Random Over-Under Sampling* Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes

Eko Saputro¹, Didi Rosiyadi²

Universitas Bina Sarana Informatika¹, Universitas Nusa Mandiri²
eko.eto@bsi.ac.id¹, didi.rosiyadi@gmail.com²

Abstrak - Penyakit diabetes merupakan salah satu penyakit yang mematikan dan jumlah penderita setiap tahunnya meningkat. Upaya pencegahan dan pengendalian diabetes ini sebaiknya dilakukan melalui edukasi deteksi dini sebagai identifikasi awal individu. Jumlah data penderita diabetes melitus yang banyak dan perlu dilakukan seleksi fitur-fitur pada dataset. Penggunaan teknik machine learning dapat memberikan kemudahan dalam melakukan pemodelan tetapi juga terdapat beberapa permasalahan. Penggunaan algoritma yang tidak sesuai akan menurunkan tingkat akurasi dari klasifikasi. Permasalahan yang lain yaitu apabila dataset yang digunakan merupakan dataset dengan distribusi kelas yang tidak seimbang. Untuk mengatasi permasalahan tersebut dengan menerapkan teknik pendekatan level data dengan menerapkan metode resampling serta membandingkan beberapa metode algoritma seperti Algoritma C4.5, Naive Bayes, K-Nearest Neighbour, Support Vector Machine, Neural Network dan Random Forest. Hasil kinerja yang diperoleh menunjukkan bahwa model pengklasifikasi Random Over-Under Sampling Random Forest memiliki nilai accuracy yang lebih tinggi dibandingkan dengan beberapa model lainnya dengan nilai accuracy sebesar 0,9808 atau 98,08% yang dan nilai AUC sebesar 0.9809 atau 98,09%. Pada pengujian dataset asli juga menghasilkan akurasi yang tinggi dengan nilai akurasi yaitu 0,9923 atau 99,23% dan nilai AUC 0,9919. Maka dapat disimpulkan bahwa Algoritma tersebut memiliki performa terbaik dan dapat digunakan untuk menyelesaikan masalah pada klasifikasi penentuan penyakit diabetes.

Kata Kunci: Klasifikasi diabetes, Random Over-Under Sampling, Random Forest

Abstract - Diabetes is one of the deadly diseases. The number of sufferers is increasing every year. Efforts to prevent and control diabetes should be carried out through early detection as an individual early. The amount of data for people with diabetes mellitus is large and it is necessary to select features in the dataset. The use of machine learning techniques can provide convenience in modeling but there are also some problems. Inappropriate use will reduce the accuracy of the classification. Another problem is if the data set used is a data set with an unbalanced class distribution. To overcome this problem by applying a data level approach by applying the resampling method and comparing several algorithm methods such as the C4.5 Algorithm, Naive Bayes, K-Nearest Neighbor, Support Vector Machine, Neural Network and Random Forest. The results obtained indicate that the classifier of the Random Over-Under Sampling Random Forest model has a higher accuracy value compared to several other models with an accuracy value of 0.9808 or 98.08% and an AUC value of 0.9809 or 98.09%. In testing the original dataset, the quality of accuracy is also high with an accuracy value of 0.9923 or 99.23% and an AUC value of 0.9919. So it can be said that the algorithm has the best performance and can be used to solve problems in various diabetes diseases.

Keywords: Diabetes classification, Random Over-Under Sampling, Random For

I. PENDAHULUAN

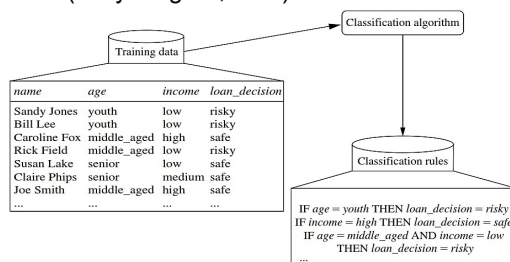
Penyakit diabetes merupakan salah satu penyakit menyebabkan kematian prematur di seluruh dunia. Penyakit seperti kebutaan, jantung dan gagal ginjal penyebab utamanya yaitu diabetes. organisasi Internasional Diabetes Federation (IDF) memperkirakan sedikitnya terdapat 463 juta orang pada usia 20 sampai 79 tahun di dunia menderita diabetes pada tahun 2019. Prevalensi diabetes meningkat seiring penambahan umur penduduk menjadi 19,9% atau 111,2 juta orang pada umur 65 sampai 79 tahun. Angka diprediksi

meningkat terus sehingga mencapai 578 juta di tahun 2030(KEMENKES, 2020).

Jumlah data penderita diabetes melitus yang banyak dan perlu dilakukan seleksi fitur-fitur pada dataset maka teknik klasifikasi secara manual sudah tidak lagi efektif digunakan karena membutuhkan waktu yang cukup lama dan tingkat akurasi data yang baik. Secara medis sendiri masih mengalami kesulitan dan bahkan mengalami reduksi data saat mendiagnosis terhadap penyakit Diabetes Melitus. Data medis yang memiliki sejumlah fitur yang tidak relevan, dan redundant dapat

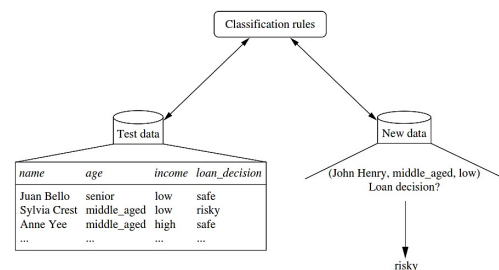
memberikan pengaruh terhadap kualitas dari diagnosis penyakit(Nurahman and Prihandoko, 2019).

Klasifikasi merupakan proses menemukan model atau fungsi yang menggambarkan dan membedakan kelas atau konsep data. Model ini digunakan untuk memprediksi label kelas objek yang label kelasnya tidak diketahui, model yang diturunkan berdasarkan analisis dari set data pelatihan yaitu objek data yang label kelasnya diketahui (Wanto, 2020). Penggunaan teknik machine learning selain dapat memberikan kemudahan dalam melakukan pemodelan juga terdapat beberapa permasalahan yang akan terjadi, seperti ditunjukkan pada Gambar 1. Penggunaan algoritma yang tidak sesuai akan menurunkan tingkat akurasi dari klasifikasi. Banyak penelitian tentang diabetes yang sudah menggunakan berbagai macam algoritma seperti Algoritma C4.5(Fatmawati, 2016), Naive Bayes(Fatmawati, 2016), K-Nearest Neighbour(Aulia, 2015), Support Vector Machine(Aulia, 2015), Neural Network(Purnama, 2019) maupun Random Forest(Suryanegara,2021).



Sumber : J. Han et al , 2012

Gambar 1. Tahap Learning



Sumber : J. Han et al , 2012

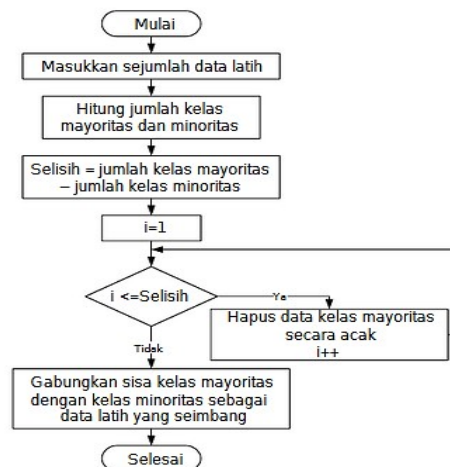
Gambar 2. Klasifikasi

Salah satu permasalahan yang dihadapi pada klasifikasi dengan menggunakan machine learning yaitu apabila dataset yang digunakan merupakan dataset dengan distribusi kelas yang tidak seimbang. Oleh karena itu masalah ketidakseimbangan kelas menjadi sangat penting. Ketidakseimbangan kelas terjadi pada saat kelas minoritas jauh lebih kecil atau lebih

sedikit dari kelas mayoritas(F. Ren at al, 2016). Pembuatan model yang menggunakan data tidak seimbang akan menghasilkan akurasi prediksi minoritas yang rendah. Informasi yang kaya dari kelas mayoritas mendominasi kelas minoritas sehingga menyebabkan batas-batas keputusan yang bias dalam sistem klasifikasi(C. Jian at al, 2016).

Mengatasi ketidakseimbangan kelas, beberapa metode dapat dibagi menjadi tiga kategori. Pertama yaitu dengan teknik tingkat data yang berusaha menyeimbangkan distribusi data dengan metode over-sampling dan under-sampling. Kedua yaitu pendekatan tingkat algoritma yaitu memodifikasi metode yang ada untuk memperhitungkan arti dari kelas minor atau dengan mengembangkan algoritma baru. Ketiga yaitu dengan mengkombinasikan pendekatan algoritma dan pendekatan level data(J. Xiao et al, 2012).

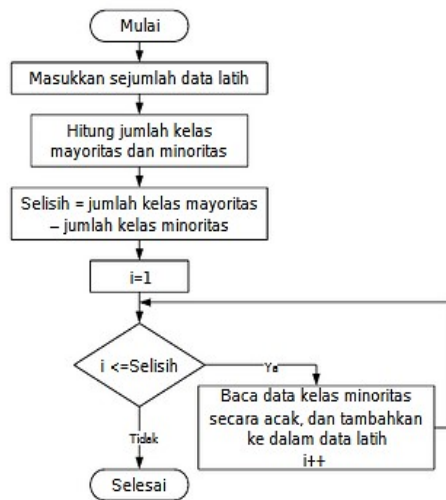
Oversampling bertujuan untuk meningkatkan sampel kelas minoritas sampai sama dengan kelas mayoritas lain dengan menduplikasi secara acak sampel kelas minoritas(H. He et al, 2018). Sedangkan *random undersampling* yaitu menghasilkan sub sampel acak dari *instance* kelas mayoritas(K. Rajesh and R. Dhuli, 2018). *Undersampling* merupakan metode *sampling* secara acak memilih sampel di kelas mayoritas dan menambahkannya ke kelas minoritas, membentuk sebuah *dataset* pelatihan baru(H. He et al, 2018) seperti yang ditunjukkan pada Gambar 3.



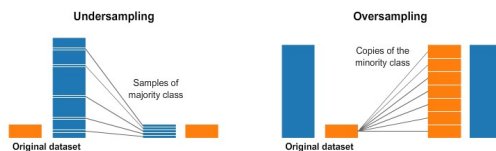
Sumber: A. Saifudin and R. S. Wahono, 2015

Gambar 3. Flowchart Algoritma Random Under sampling

Oversampling : Dalam metode ini, contoh dari kelas minoritas dipilih secara acak dan diduplikasi. *Instance* yang dihasilkan hanya meningkatkan besarnya jumlah kelas minoritas dengan hanya mereplikasi informasi yang sama (K. Agrawal et al, 2017). Salah satu teknik oversampling yang sering digunakan dalam mengatasi ketidakseimbangan kelas yaitu **SMOTE (Synthetic Minority Over-sampling Technique)**.



Sumber: A. Saifudin and R. S. Wahono, 2015
Gambar 4. Flowchart Algoritma Random Oversampling



Sumber: W. Badr, 2021
Gambar 5. Undersampling dan oversampling

II. METODOLOGI PENELITIAN

Penelitian eksperimen ini bertujuan untuk melakukan komparasi dan evaluasi model *algoritma* klasifikasi dan metode *resampling* untuk mengetahui algoritma yang memiliki nilai akurasi lebih tinggi pada klasifikasi tidak seimbang penentuan penyakit diabetes. Metode pengumpulan data yang dilakukan pada penelitian ini adalah menggunakan data yang diperoleh dari *University of California Irvine machine learning data repository* yang dapat diperoleh melalui alamat web <https://archive.ics.uci.edu/ml/machine-learning-databases/00529/>

Tahapan penelitian yang dilakukan mulai dari persiapan data, pemilihan model usulan algoritma yang digunakan, evaluasi dan validasi

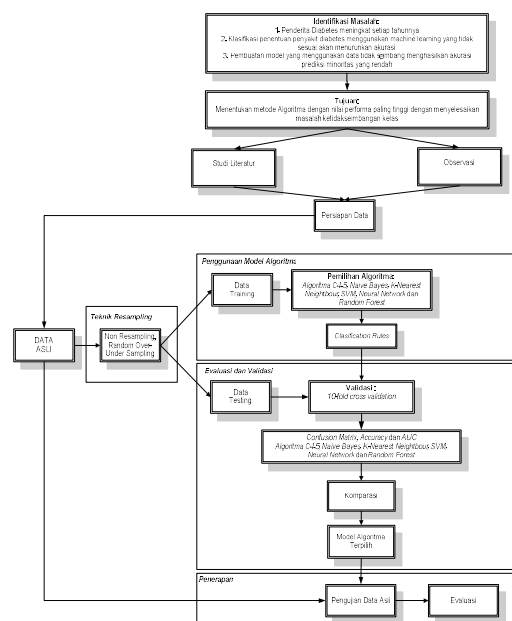
model sampai dengan penerapan model algoritma terpilih. Tahapan awal penelitian dimulai dari pengumpulan dataset yang akan digunakan eksperimen. Karena dataset yang didapat tidak terdapat missing value, maka tidak dilakukan proses pengolahan awal. Atribut yang digunakan adalah sebagai berikut:

Tabel 1. Atribut

| Atribut | Keterangan |
|--------------------|--|
| Age | Usia |
| Gender | Jenis kelamin |
| Polyuria | Sering buang air kecil |
| Polydipsia | Haus berlebihan |
| Sudden weight loss | Penurunan berat badan secara tiba-tiba |
| Weakness | Kelemahan |
| Polyphagia | Rasa lapar berlebihan |
| Genital thrush | Sariawan kelamin |
| Visual blurring | Pengaburan visual |
| Itching | Gatal |
| Irritability | Sifat lekas marah |
| Delayed healing | Penyembuhan tertunda |
| Partial paresis | Lumpuh sebagian |
| Muscle stiffness | Kekakuan otot |
| Alopecia | Rambut rontok |
| Obesity | Kegemukan |
| Class | Kelas |

Sumber: *University of California Irvine machine learning data repository*

Pada penelitian ini model yang diusulkan untuk mengatasi masalah ketidakseimbangan kelas yaitu dengan menerapkan kombinasi teknik pendekatan level data dengan metode resampling Random Over-Under Sampling (ROUS) maupun non resampling dan menggunakan pendekatan algoritma dengan 6 model algoritma yaitu Algoritma C4.5, Naive Bayes, K-Nearest Neighbour, SVM, Neural Network dan Random Forest seperti pada Gambar 6.



Sumber : (Saputro & Rosiyadi, 2021)
Gambar 6. Metode Usulan

Proses *resampling* dilakukan dengan aplikasi WEKA 3.9.1. dengan memilih filter=*Resample-B1.0-81-Z100.0* untuk proses *Random Over-Under Sampling* dan untuk tanpa resample cukup di kosongkan. Mengukur kinerja algoritma pengklasifikasi dilakukan dengan menggunakan *confusion matrix* yang diperoleh dari proses validasi menggunakan *k-fold cross validation* dengan metode evaluasi standarnya yaitu *10-fold cross validation* yang merupakan pilihan terbaik untuk mendapatkan hasil validasi yang akurat. Pengukuran dilakukan dengan menggunakan aplikasi WEKA 3.9.1. dengan memilih jenis *classifier* dan *test option cross validation fold 10*.

Hasil yang diperoleh dari pengukuran dihitung nilai *Accuracy* dan juga nilai *Area Under curve (AUC)*. Kemudian dibandingkan antara algoritma Algoritma C4.5, *Naive Bayes*, *K-Nearest Neighbour*, *SVM*, *Neural Network* dan *Random Forest* dengan kombinasi metode *resampling* maupun tanpa metode *resampling*. Dari hasil pengukuran akan mendapatkan algoritma terbaik yang akan digunakan sebagai penentuan penyakit diabetes. Penerapan model algoritma terpilih di uji menggunakan tingkat akurasi menggunakan data asli. Hasil pengujian dapat dilihat berdasarkan nilai *Accuracy*, *F-measure* dan juga nilai *Area Under curve (AUC)*. Tujuan pengujian ini adalah untuk mengetahui seberapa akurat model algoritma yang sudah terpilih.

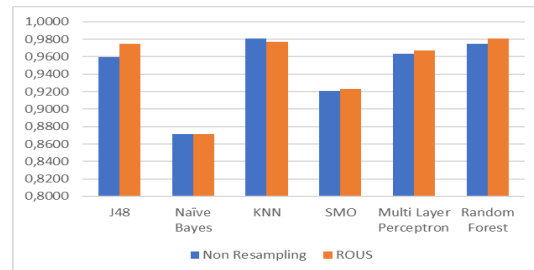
III. HASIL DAN PEMBAHASAN

Pada proses *resampling Random Over-Under sampling (ROUS)* proses yang dilakukan pada aplikasi Weka dilakukan penyeimbangan data pada masing-masing kelas dengan cara mengatur nilai *biasTouniformclass* menjadi 1,0. Sehingga jumlah dataset pada masing-masing kelas menjadi 260 *instance*. Sedangkan penerapan metode tanpa *resampling* yaitu dengan menggunakan data sebenarnya atau menggunakan data asli dimana total dari seluruh data adalah 520 *instance* yang terbagi menjadi 320 data *Positive* dan 200 data *Negative*.

1. Evaluasi Kinerja

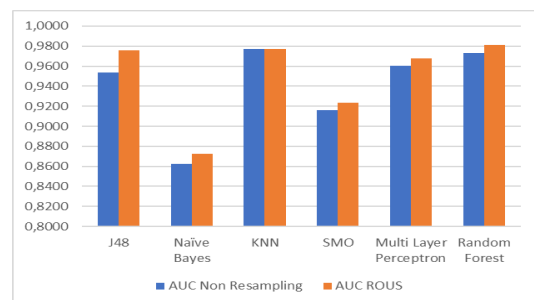
Berdasarkan pengukuran kinerja yang sudah dilakukan diperoleh kemampuan model *classifier* dalam mengklasifikasikan data secara umum (*accuracy*) perbandingan kinerja beberapa metode dapat dilihat melalui table 2. Pada tabel 2 menunjukkan bahwa dengan penerapan metode *Random Over-Under Sampling* pada dataset sebelum dilakukan pemodelan rata-rata dapat menghasilkan kinerja yang lebih baik dibandingkan dengan tanpa menggunakan metode *resampling*, kecuali metode KNN yang mengalami penurunan *accuracy* sebesar 0,39% dan

penurunan AUC sebesar 0.02%. Hasil kinerja yang diperoleh menunjukkan bahwa model pengklasifikasi *Random Over-Under Sampling (ROUS)+Random Forest* memiliki nilai *accuracy* yang lebih baik dibandingkan dengan beberapa model lainnya dengan nilai *accuracy* sebesar 0,9808 atau 98,08% yang mengalami kenaikan 0,59% dan nilai AUC sebesar 0.9809 atau 98,09% yang mengalami kenaikan 0,79%. Grafik perbandingan kinerja 6 model pengklasifikasi disajikan pada gambar 7 dan 8



Sumber: (Saputro & Rosiyadi, 2021)

Gambar 7. Perbandingan accuracy

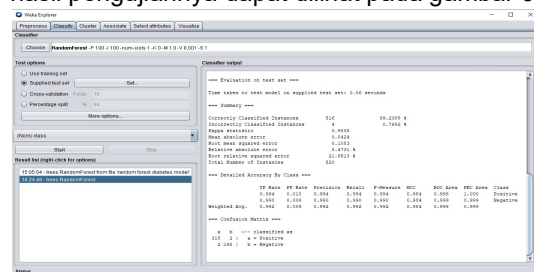


Sumber: (Saputro & Rosiyadi, 2021)

Gambar 8. Perbandingan AUC

2. Penerapan Model Algoritma Terpilih

Model Algoritma yang terpilih akan diuji menggunakan *dataset* asli. Hasil pengujian penerapan model yang terpilih yaitu model *Random Over-Under Sampling Random Forest*. Pengujian ini dilakukan untuk membuktikan bahwa model yang diusulkan apakah memiliki kinerja yang konsisten dalam mengukur nilai akurasi baik pada saat *training* maupun *testing*. Pengujian data asli dimasukkan di dalam *Supplied Test Set* pada aplikasi WEKA. Dari hasil pengujiannya dapat dilihat pada gambar 9.



Sumber: (Saputro & Rosiyadi, 2021)

Gambar 9. Pengujian data asli dengan *Supplied Test Set Random forest*

Tabel 2. Perbandingan Kinerja Model

| No | algoritma | Accuracy | | | AUC | | |
|----|------------------------|----------------|---------------|----------|----------------|---------------|----------|
| | | Non resampling | ROUS | Kenaikan | Non resampling | ROUS | Kenaikan |
| 1 | J48 | 0.9596 | 0.9750 | 1.60% | 0.9540 | 0.9759 | 2.29% |
| 2 | Naïve Bayes | 0.8712 | 0.8712 | 0.00% | 0.8622 | 0.8724 | 1.18% |
| 3 | KNN | 0.9808 | 0.9769 | -0.39% | 0.9774 | 0.9772 | -0.02% |
| 4 | SMO | 0.9212 | 0.9231 | 0.21% | 0.9158 | 0.9235 | 0.83% |
| 5 | Multi Layer Perceptron | 0.9635 | 0.9673 | 0.40% | 0.9603 | 0.9676 | 0.77% |
| 6 | Random Forest | 0.9750 | 0.9808 | 0.59% | 0.9732 | 0.9809 | 0.79% |

Sumber : (Saputro & Rosiyadi, 2021)

Tabel 3. Confusion matrix Supplied Test Set Random Forest

| | Actual Positive | Actual Negative |
|-----------------|-----------------|-----------------|
| Predic Positive | 318 | 2 |
| Predic Negative | 2 | 198 |

Sumber: (Saputro & Rosiyadi, 2021)

Berdasarkan Tabel 3 tersebut didapatkan 318 *instance* prediksi *Positive* terklasifikasi benar sebagai kelas *Positive*, sedangkan 2 *instance* prediksi *Positive* ternyata *Negative*. Kemudian 198 *instance* diprediksi *Negative* terklasifikasi benar sebagai kelas *Negative*, sedangkan 2 *instance* prediksi *Negative* ternyata *positive*. Sehingga untuk proses perhitungan nilai akurasi sebagai berikut:

$$\begin{aligned}
 FPRate &= \frac{FP}{FP+TN} = \frac{2}{2+198} = 0.0100 \\
 TPRate/Recall &= \frac{TP}{TP+FN} = \frac{318}{318+2} = 0.9938 \\
 Precision &= \frac{TP}{TP+FP} = \frac{318}{318+2} = 0.9938 \\
 F-Measure &= \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} = \frac{2 \cdot 0.9938 \cdot 0.9938}{0.9938 + 0.9938} = 0.9938 \\
 Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} = \frac{318+198}{318+198+2+2} = 0.9923 \\
 AUC &= \frac{1+TPRate-FPRate}{2} = \frac{1+0.9938-0.0100}{2} = 0.9919
 \end{aligned}$$

Berdasarkan hasil pengujian model klasifikasi pada dataset asli yang telah dilakukan menunjukkan bahwa metode *Random Over-Under Sampling Random Forest* memiliki kinerja hasil pengujian yang baik. Dimana pada metode *Random Over-Under Sampling Random Forest* menghasilkan nilai akurasi yaitu 0,9923 atau 99,23% dan nilai AUC yaitu 0,9919.

IV. KESIMPULAN

Berdasarkan hasil eksperimen yang telah dilakukan pada *dataset prediksi penyakit diabetes* menunjukkan dengan penerapan metode *Random Over-Under Sampling* pada dataset sebelum dilakukan pemodelan, sebagian besar algoritma dapat menghasilkan kinerja yang lebih baik dibandingkan dengan tanpa menggunakan metode *resampling*, kecuali metode KNN yang mengalami

penurunan *accuracy* sebesar 0,39% dan penurunan AUC sebesar 0,02%. Maka disimpulkan bahwa *Random Over-Under Sampling* dapat meningkatkan performa akurasi secara efektif pada klasifikasi tidak seimbang untuk prediksi penyakit diabetes.

Hasil kinerja yang diperoleh menunjukkan bahwa model pengklasifikasi *Random Over-Under Sampling Random Forest* memiliki nilai *accuracy* yang lebih tinggi dibandingkan dengan beberapa model lainnya dengan nilai *accuracy* sebesar 0,9808 atau 98,08% yang dan nilai AUC sebesar 0.9809 atau 98,09%. Maka dapat disimpulkan bahwa Algoritma yang memiliki performa terbaik adalah *Random Over-Under Sampling Random Forest*. Pengujian *dataset* asli menggunakan algoritma terpilih menghasilkan akurasi yang tinggi dengan nilai akurasi yaitu 0,9923 atau 99,23% dan nilai AUC yaitu 0,9919. Maka dapat disimpulkan bahwa Algoritma tersebut dapat digunakan untuk menyelesaikan masalah pada klasifikasi penentuan penyakit diabetes.

V. REFERENSI

- Agrawal, K., Baweja, Y., Dwivedi, D., Agrawal, S., & Chaturvedi, P. (2017). A Comparison of Class Imbalance Techniques for Real-World Landslide Predictions. *International Conference on Machine learning and Data Science*. IEEE. doi:10.1109/MLDS.2017.21
- Aulia, S., Hadiyoso, S., & Ramadhan, D. N. (2015). Analisis Perbandingan KNN dengan SVM untuk Klasifikasi Penyakit Diabetes Retinopati berdasarkan Citra Eksudat dan Mikroaneurisma. *Jurnal ELKOMIKA*, 3(1), 75-90.
- Badr, W. (n.d.). *towardsdatascience.com*. Retrieved Juli 25, 2021, from <https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb>
- Fatmawati, F. (2016). Perbandingan Algoritma Klasifikasi Data Mining model C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Diabetes. *Jurnal Techno Nusa Mandiri*, 8(1). Retrieved from

- <http://ejournal.nusamandiri.ac.id/index.php/techno/article/view/217/193>
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concept And Techniques*. California: Morgan Kaufmann.
- He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: adaption of different imbalance ratios. *Expert Systems With Applications*. doi:10.1016/j.eswa.2018.01.012
- Ichsan. (2013, November). Sistem Pendukung Keputusan Pemilihan Penerima Beasiswa Mahasiswa Kurang Mampu Pada STMIK BUDIDARMA Medan Menerapkan Metode Profile Matching. *Kursor*, 5(1), 2. Retrieved April 14, 2016, from <http://pelita-informatika.com/berkas/jurnal/1.%20TM%20Syahru.pdf>
- Jian, C., Gao, J., & Ao, Y. (2016). A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble. *Neurocomputing*. doi:<http://dx.doi.org/10.1016/j.neucom.2016.02.006>.
- Nurahman, N., & Prihandoko, P. (n.d.). Perbandingan Hasil Analisis Teknik Data Mining Metode Decision Tree, Naive Bayes, SMO an Part Untuk Mendiagnosa Penyakit Diabetes Mellitus. *INFORM*. Retrieved from <https://ejournal.unitomo.ac.id/index.php/inform/article/view/1403>
- Purnama, J. J., Rahayu, S., Nurdiani, S., Haryanti, T., & Mayangky, N. A. (2019). Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Diabetes. *Indonesian Journal on Computer and Information Technology*. Retrieved from <http://ejournal.bsi.ac.id/ejurnal/index.php/ijcit>
- Rajesh, K., & Dhuli, R. (2018). Classification Of Imbalanced ECG beats using re-sampling techniques And AdaBoost ensemble classifier. *Biomedical Signal Processing and Control*, 41, 242-254.
- Ren, F., Cao, P., Li, W., Zhao, D., & Zaiane, O. (2016). Ensemble Based Adaptive over-sampling method for imbalanced data Learning aided detection of microaneurysm. *Computerized Medical Imaging and Graphics*. doi:<http://dx.doi.org/10.1016/j.compmmedimag.2016.07.011>.
- Saifudin, A., & Wahono, R. S. (2015). Pendekatan Level Data untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software. *Journal of Software Engineering*, 1(2), 76–85.
- Saputro, E., & Rosiyadi, D. (2021). Penerapan Metode Random Over-Under Sampling Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes.
- Suryanegara, G. A., Adiwijaya, & Purbolaksono, M. D. (2021). Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi. *Jurnal RESTI(Rekayasa Sistem dan Teknologi Informasi)*, 5(1), 114 - 122.
- Wanto, A., Siregar, N. M., Windarto, A. P., Hartama, D., & Ginantra, N. L. (2020). *Data Mining : Algoritma dan Implementasi*. (2020). . Yayasan Kita Menulis.
- Xiao, J., Xie, L., He, C., & Jiang, X. (2012). Dynamic classifier ensemble model for customer classification with imbalanced class distribution. *Expert Systems with Applications*, 39, 3668-3675.
- N. Nurahman and P. Prihandoko, " Perbandingan Hasil Analisis Teknik Data Mining Metode Decision Tree, Naive Bayes, SMO an Part Untuk Mendiagnosa Penyakit Diabetes Mellitus," *INFORM*.
- S. Aulia, S. Hadiyoso and D. N. Ramadhan, "Analisis Perbandingan KNN dengan SVM untuk Klasifikasi Penyakit Diabetes Retinopati berdasarkan Citra Eksudat dan Mikroaneurisma," *Jurnal ELKOMIKA*, vol. 3, no. 1, pp. 75-90, 2015.