Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2023106480
Vol. 10, No. 2, April 2023, hlm. 457-466 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

STRATEGI PENANGANAN IMBALANCE CLASS PADA MODEL KLASIFIKASI PENERIMA KARTU INDONESIA PINTAR KULIAH BERBASIS NEURAL NETWORK MENGGUNAKAN KOMBINASI SMOTE DAN ENN

Zaqiatud Darojah*[1], Ronny Susetyoko[2], Nana Ramadijanti[3]

[1, 2, 3]Politeknik Elektronika Negeri Surabaya, Surabaya
Email: [1]zaqiah@pens.ac.id, [2] rony@pens.ac.id, [3]nana@pens.ac.id
*Penulis Korespondensi

(Naskah masuk: 09 Agustus 2022, diterima untuk diterbitkan: 12 April 2023)

Abstrak

Keterbatasan kuota penerima program Kartu Indonesia Pintar Kuliah (KIP Kuliah) dari pemerintah mengharuskan Perguruan Tinggi (PT) menyeleksi dengan cermat calon mahasiswa yang berhak menerima program tersebut. Pembentukan model klasifikasi penerima program KIP Kuliah merupakan salah satu cara yang dapat membantu PT dalam menyeleksi calon mahasiswa agar tepat sasaran berdasarkan data lampau. Penelitian ini bertujuan untuk membentuk model klasifikasi penerima KIP Kuliah menggunakan *Neural Network* (NN). Strategi *data processing level* digunakan untuk mengatasi ketidakseimbangan data atau *imbalance class* yang terjadi antara kelas penerima KIP Kuliah sebagai kelas minoritas dan kelas bukan penerima KIP Kuliah sebagai kelas mayoritas. Teknik yang digunakan pada penelitian ini adalah mengkombinaskan metode *oversampling Syntetic Minority Oversampling Technique* (SMOTE), metode *undersampling Edited Nearest Neighbor Rule* (ENN), dan metode *undersampling* dengan penghapusan langsung pada sampel terpilih. Skema penggabungan dilakukan dengan cara mengelompokkan terlebih dahulu kelas mayoritas menjadi beberapa sub kelas (*cluster*) menggunakan algoritma *k-means*. Metode SMOTE dan ENN diterapkan secara bersamaan menggunakan rasio sampling tertentu pada dataset yang berasal dari kelas minoritas dan sub kelas mayoritas yang merupakan tetangga terdekat kelas minoritas tersebut. Metode penghapusan sampel diterapkan pada sub kelas mayoritas yang memiliki jarak yang sangat signifikan dari kelas minoritas. Tujuan dari skema yang diajukan adalah untuk meminimalkan terjadinya pembangkitan *false sample* pada kelas minoritas dan penghapusan sampel informatif pada kelas mayoritas. Hasil simulasi menunjukkan bahwa kombinasi teknik *undersampling* dan *oversampling* dengan skema yang diusulkan mampu meningkatkan kinerja model klasifikasi NN secara signifikan. Model klasifikasi terbaik menghasilkan nilai *accuracy* sebesar 93.45%, TPR sebesar 90,00%, TNR sebesar 93.67%, *G-Mean* sebesar 91,51%, dan nMCC sebesar 81.25%.

Kata kunci: KIP Kuliah, *imbalance class*, *Neural Network*, *SMOTE, ENN*

NEURAL NETWORK USING HYBRID SAMPLING TECHNIQUE COMBINING SMOTE AND ENN FOR IMBALANCED KARTU INDONESIA PINTAR KULIAH RECIPIENTS DATA

Abstract

The limited quota for recipients of the Kartu Indonesia Pintar Kuliah (KIP Kuliah) program requires the university to select carefully the students who are entitled to receive the program. This study aims to build the classification model for KIP Kuliah recipients using Neural Network (NN) which can be utilized by universities in selecting prospective KIP Kuliah recipients students. To solve the imbalanced KIP Kuliah recipients data, we propose a hybrid sampling technique that combines the Synthetic Minority Over-Sampling Technique (SMOTE) and the Edited Nearest Neighbor (ENN) and also samples selected deletion method with a new scheme. Firstly, the majority class is clustered into several sub-classes using the k-means algorithm. The SMOTE and ENN methods are applied simultaneously on a dataset derived from a minority class and a majority sub-class that is the nearest neighbor of the minority class with a certain sampling ratio. Furthermore, the sample-selected deletion method is applied to the majority sub-classes that have a very significant distance from the minority class. Lastly, The resampling results of the proposed scheme are combined into one training dataset in ANN. The objective of the proposed scheme is to minimize the generation of 'false samples' in the minority class and the elimination of informative samples in the majority class. The results show that the proposed scheme can significantly improve

the performance of the NN classification model. The best classification model produces an accuracy value of 93.45%, TPR of 90.00%, TNR of 93.67%, G-Mean of 91.51%, and MCC of 81.25%.
Keywords: *KIP Kuliah*, *neural network*, *oversampling, cluster based undersampling.*

1. PENDAHULUAN

Kartu Indonesia Pintar Kuliah (KIP Kuliah) merupakan program bantuan pendidikan di Perguruan Tinggi (PT) yang diberikan oleh pemerintah kepada mahasiswa yang memiliki potensi akademik baik tetapi memiliki keterbatasan ekonomi. Penerima KIP Kuliah ditetapkan oleh Pusat Layanan Pembiayaan Pendidikan Kemndikbud (Puslapdik) atas usulan PT setiap tahunnya (Puslapdik, 2022). Permasalahan yang terjadi di lapangan adalah beberapa mahasiswa yang menerima KIP Kuliah ternyata mampu secara ekonomi. Sebaliknya, tidak sedikit mahasiswa yang tidak menerima KIP Kuliah melakukan permohonan keringanan UKT karena permasalahan ekonomi pada semester berjalan (Susetyoko, 2002. Pemerintah memberikan kebijakan adanya pembatalan dan pengalihan penerima KIP Kuliah bila kondisi ekonomi keluarganya meningkat atau tidak memenuhi standar minimum Indek Prestasi Komulaitif yang ditetapkan PT masing-masing (Yanuar, 2022). Berdasarkan hal tersebut dan adanya keterbatasan kuota penerima program KIP Kuliah, pembentukan model klasifikasi penerima KIP Kuliah akan sangat membantu PT dalam menyeleksi dan mengevaluasi kelayakan mahasiswa untuk menerima program tersebut. Penelitian ini bertujuan untuk membangun model klasifikasi penerima KIP Kuliah beradsarkan data lampau menggunakan metode *machine learning* NN. Model yang terbentuk dapat memprediksi apakah mahasiswa termasuk dalam kelas penerima KIP Kuliah, selanjutnya disebut dengan kelas KIP atau bukan penerima KIP Kuliah, selanjutnya disebut dengan kelas Non-KIP.

Data untuk membangun model klasifikasi penerima KIP Kuliah memiliki ketidakseimbangan kelas yang sangat signifikan. Proporsi antar kelas KIP sebagai kelas minoritas dengan kelas Non-KIP sebagai kelas mayoritas adalah 4.9% dengan 95.1%. Data yang memiliki proporsi jauh berbeda antar satu kelas data dengan kelas data lainnya, yang dikenal dengan istilah *imbalance class* atau *imbalance dataset*, dapat menyebabkan algoritma *supervised machine learning* mengalami penurunan kinerja klasifikasi dengan signifikan (Kumar et al., 2021). Algoritma *machine learning* yang tidak mempertimbangkan ketidakseimbangan kelas akan menyebabkan kelas minoritas sering disalah-klasifikasikan sebagai kelas mayoritas.

Penelitian terkait dengan *imbalance class* menjadi banyak perhatian oleh para peneliti sejak beberapa tahun terakhir dengan tujuan menghasilkan model klasifikasi yang mampu memprediksi dengan tepat terutama pada kelas minoritas. Strategi pendekatan dasar dalam menyelesaikan *imbalance class* pada *machine learning* terbagi menjadi dua, yaitu *data processing level* dan *learning algorithm level* (Kumar et al., 2021). Penelitian ini menggunakan strategi *data processing level* dalam menangani ketidakseimbangan kelas yang terjadi. Startegi *data level processing* atau disebut dengan *eksternal level* dilakukan dengan cara menyeimbangkan distribusi antara kelas minoritas dan kelas mayoritas melalui teknik *undersampling, oversampling*, atau kombinasi dari keduanya.

Teknik *undersampling* diterapkan pada kelas mayoritas yang dilakukan dengan cara mengeliminasi sampel secara acak dari kelas tersebut. Beberapa algoritma berbasis teknik *undesrampling* diantaranya adalah *Random Over Sampler* (ROS), *Condensed Nearest Neighbour* (CNN), *Edited Nearest Neighbor Rule* (ENN), *Neighbour Cleaning Rule* (NCL), *K-Nearest Neigbor* (K-NN), *Tomek Links*, dan *One-sided Selection* (OSS) (Devi, Debashree, Biswas, Saroj Kr., Purkayastha, 2020; Kumar et al., 2021). Teknik *undersampling* terbukti mampu meningkatkan akurasi kinerja klasifikasi dari kelas minoritas. Namun, penghapusan sampel pada kelas mayoritas secara acak dapat berpotensi menghilangkan beberapa sampel yang informatif pada kelas mayoritas. Oleh karena itu, penerapan teknik *undersampling* membutuhkan strategi peningkatan kinerja secara kontinu (Deng et al., 2021). Strategi yang paling sering digunakan adalah menggabungkan metode *undersampling* dengan metode *clustering* (Devi, Debashree, Biswas, Saroj Kr., Purkayastha, 2020). Tsai et al. (2019) mengkombinasikan metode *clustering analysis* dan *instance selection. Clustering analysis* mengelompokkan sampel kelas mayoritas yang memiliki kemiripan data ke dalam beberapa sub-kelas. Sedangkan, *instance selection* manyaring sampel data yang tidak representatif dari masing-masing sub-kelas untuk dihilangkan. Nugraha et al. (2020) menerapkan teknik *undersampling* dengan strategi *clustering*. Sampel pada setiap *cluster* akan dipilih secara acak dan digabungkan dengan sampel-sampel dari *clsuter* lainnya sedemikian hingga distribusi kelas menjadi setimbang. Strategi ini mampu meningkatkan nilai *sensitivity* dan AUC dengan sangat signifikan.

Teknik *oversampling* diterapkan pada kelas minoritas dengan cara mensintesis sampel baru atau menduplikasi sampel secara acak pada kelas tersebut. Metode berbasis teknik *oversampling* diantaranya adalah *Random Over Sampler* (ROS), SMOTE, dan algoritma pengembangan dari SMOTE. Permasalahan yang timbul dalam mensintesis sampel baru adalah terjadinya *overfitting*, *boundary samples*, *noise sample*, dan *overlapping sample.* Oleh karena itu, penerapan teknik *oversampling* juga

membutuhkan strategi peningkatan kinerja (Deng et al., 2021; Hassanat, Tarawneh, Abed, et al., 2022). Strategi yang diterapkan oleh Xu et al. (2021) adalah mengkombinasikan SMOTE dangan algoritma k-*means cluster-based filter startegy* yang dikenal dengan KNSMOTE. Kombinasi tersebut mampu meningkatkan kinerja klasifikasi algoritam RF pada data *medical* dengan nilai *sensitify* dan *specificity* masing-masing sebesar 99.84% dan 95.56%. Deng et al. (2021) menggunkan strategi *classification ranking* and *weight setting* dalam melakukan *oversampling* kelas. Strategi tersebut bertujuan untuk menghasilkan *oversampled sample* yang dapat mempertahankan karakteristik distribusi spasial dan fitur asli sampel asli sambil menyeimbangkan jumlah data antara beberapa kelas. Hasil studi *critical review* terbaru terhadap metode *oversampling* dilakukan oleh Hassanat et al. (2022) yang mengungkapkan bahwa penggunaan metode *oversampling* SMOTE memungkinkan memunculkan sampel-sampel sintesis yang tidak benar-benar mewakili kelas minoritas sehingga dapat mengakibatkan prediksi yang salah pada *real-life dataset*. Hassanat et al. (2022) melakukan validasi dari teknik *oversampling* dengan cara menutup atau menyembunyikan subset kelas mayoritas atau yang disebut dengan *hidden subset*. Teknik *oversampling* diterapkan pada kelas minoritas dan subset kelas mayoritas tersisa untuk meng*enerate* sample baru yang disebut dengan *synthetic subset*. Dataset baru terbentuk dari sampel-sampel kelas mayoritas asli digabung dengan sampel-sampel yang dihasilkan pada *sysnthetic subset* setelah dilakukan pengecekan apakah sampel-sampel tersebut benar-benar mewakili kelas minoritas berdasarkan derajat kemiripan menggunakan nilai *Hassanant distance*. Hasil validasi pada empat *real-world dataset* mengungkapkan bahwa semua metode *oversampling* menghasilkan terbentuknya *false sample,* sampel yang sebenarnya bukan milik kelas minoritas. Hal tersebut menyebabkan model klasifikasi berkerja dengan baik di laboratorium tetapi lebih mungkin gagal dalam praktiknya.

Penggabungan teknik *undersampling* dan teknik *oversampling* sebagai strategi dalam mengatasi ketidakseimbangan kelas juga menjadi pilihan para peneliti. Bach et al. (2017) mengkombinasikan beberapa metode *oversampling* dan *undersampling* pada data *osteoporosis*. Kombinasi algoritma yang paling optimal dipilih berdasarkan kinerja klasifikasi algoritma Random Foreset (RF). Hasil simulasi menunjukkan bahwa kombinasi SMOTE dan ENN meningkatkan kinerja algoritma klasifiaksi RF secara signifikan. Xu et al. (2020) mengkombinasikan ENN dan *Missclassification-Oriented* SMOTE (M-SMOTE) dalam menangani ketidakseimbangan kelas pada data medis berdasarkan algoritma RF. Algoritma M-SMOTE merupakan pengembangan dari SMOTE untuk menghindari terjadinya *overfitting* dengan cara menggantikan tingkat ketidakseimbangan sampel dengan tingkat terjadinya *misclassification* yang diperoleh dari RF. Koziarski (2021) mengajukan algoritma *Combined Synthetic Oversampling and Undersampling Technique* (CSMOUTE) yang mengintegrasikasn SMOTE dengan *Synthetic Majority Undersampling Technique* (SMUTE). Algoritma CSMOUTE mampu menghasilkan kinerja yang signifikan ketika digabungkan dengan algoritma MLP dan SVM.

Hasil investigasi yang dilakukan menujukkan bahwa penggunaan metode-metode dalam menyeimbangkan kelas tetap memerlukan strategi untuk mengkompensasi kelemahan dari masing-masing metode. Penelitian ini mengkombinasikan teknik *oversampling* pada kelas minoritas dan teknik *undersampling* pada kelas mayoritas melalui skema berbasis jarak antar kelas minoritas dengan sub-sub kelas mayoritas hasil *clustering* menggunakan algoritma *k-means*. Perbedaan strategi dengan penelitian-penelitian sebelumnya adalah tidak setiap sub kelas hasil *clusterring* akan dikenai teknik *undersampling* atau teknik *oversampling*. Teknik *resampling* hanya dikenakan pada sub kelas tertentu dengan tujuan untuk meminimalkan terjadinya penghapusan sampel-sampel informatif yang dapat berpengaruh terhadap kinerja model klasifikasi. Strategi yang digunakan tersebut menyebabkan rasio antara kelas mayoritas dan kelas minoritas tidak berada tepat di titik setimbang. Penelitian ini menyimulasikan beberapa skema dataset untuk melihat pengaruh dari teknik yang diajukan terhadap model klasifikasi yang dihasilkan. Pemilihan metode klasifikasi NN pada penelitian ini berdasarkan hasil investigasi yang menunjukkan bahwa kinerja klasifikasi NN lebih baik dibandingkan dengan algoritma klasifikasi lainnya (Saritas & Yasar, 2019).

2. METODE PENELITIAN

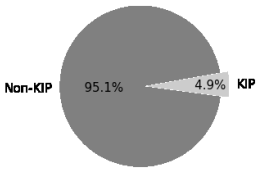
2.1 Sumber Data

Data yang digunakan pada pembentukan model klasifikasi penerima KIP Kuliah adalah data pendaftar yang diterima melalui jalur Seleksi Bersama Masuk Politeknik Negeri di Politeknik Elektronika Negeri Surabaya (SBMPN-PENS) Tahun 2019-2021. Data asli memiliki 31 fitur yang masing-masing berjumlah 873 data. Dari 31 fitur dipilih sebanyak tujuh fitur (X_1 sampai X_7) sebagai data input pada NN yang ditunjukkan pada Tabel 1. Sedangkan targetnya (Y) adalah status pendaftar apakah termasuk dalam kelas KIP atau Non-KIP.

Tabel 1. Data Peneltian

Variabel	Keterangan	Jenis data
X_1	Penghasilan orang tua	Numerik
X_2	Jumlah rumah	Kategorik
X_3	Status rumah	Kategorik
X_4	Jumlah Motor	Kategorik
X_5	Jumlah Mobil	Kategorik
X_6	Jumlah Anak	Kategorik
X_7	Daya Listrik	Kategorik
Y	Status Pendaftar	Boolean

Pembersihan data dilakukan pada data asli dalam bentuk menghapus data yang tidak lengkap dan mendeteksi data yang tidak akurat serta menghilangkannya dari dataset. Jumlah data setelah dilakukan pembersihan menjadi 837, dimana jumlah mahasiswa yang masuk dalam kelas Non-KIP sebanyak 796 pendaftar dan kelas KIP sebanyak 41 pendaftar atau sebanyak 4.9%. Gambar 1 merupakan penyajian data dalam bentuk *pie chart* yang menunjukkan terjadi ketidakseimbangan data antara kelas KIP dan kelas Non-KIP.



Gambar 1. Persentase Pendaftar Berdasarkan Kelas

2.2 Metode Penanganan *Imbalance Class*

Prinsip kerja dalam menangani *imbalance class* adalah menyeimbangakan jumlah data antara kelas mayoritas dan kelas minoritas sehingga menghasilkan proporsi data yang seimbang atau tidak berbeda secara signifikan. Metode *oversampling* kelas minoritas yang digunakan pada penilitian ini adalah metode SMOTE, sedangkan metode *undersampling* kelas mayoritas yang digunakan adalah metode ENN.

a) Metode SMOTE

SMOTE merupakan teknik pengembangan metode *oversampling* dari kelas minoritas dan merupakan salah satu algoritma yang paling banyak digunakan oleh para peneliti (Fernández et al., 2018; Kumar et al., 2021). SMOTE bekerja dengan cara meng*enerate* sampel baru dari kelas minoritas dengan interpolasi. SMOTE menggunakan *K-Nearest Neigbors* (KNN) dalam menemukan sampel baru (sampel sintesis) untuk setiap data di kelas minoritas. Satu sampel sintesis diciptakan melalui beberapa tahapan, yaitu: algoritma KNN menyeleksi tetangga terdekat (*nearest neighbor*) sebanyak k dari sebuah sampel di kelas minoritas (x_i), satu sampel dari k tetangga terdekat dipilih secara acak ($\hat{x}_i$), kemudian sampel sintesis dig*enerate* dengan atribut random diantara x_i dan $\hat{x}_i$ menggunakan Persamaan (1). Proses ini diulang sebanyak n kali untuk setiap sampel x_i dari sampel set kelas minoritas, dimana $n = b/100$, dan b adalah persentase *oversampling* yang diperlukan untuk menyeimbangkan dataset.

$$x_{sintesis} = x_i + (\hat{x}_i - x_i)\delta \quad (1)$$

dimana δ adalah bilangan random antar 0 dan 1.

b) Metode ENN

Metode ENN merupakan salah satu teknik pengembangan metode *undersampling* kelas mayoritas yang dilakukan dengan cara menghapus sampel pada kelas mayoritas yang dinilai memiliki perbedaan dari kelas mayoritas berdasarkan aturan tetangga terdekat (*the nearest-neighbor rule*). Aturan ini menyatakan bahwa sampel tetangga terdekat dari setiap sampel mayoritas ditemukan berdasarkan jarak antara dua sampel dan identifikasi apakah sampel mayoritas merupkan *noise sampel* atau bukan dengan menilai konsistensi label tetangga terdekatnya (Xu et al., 2020).

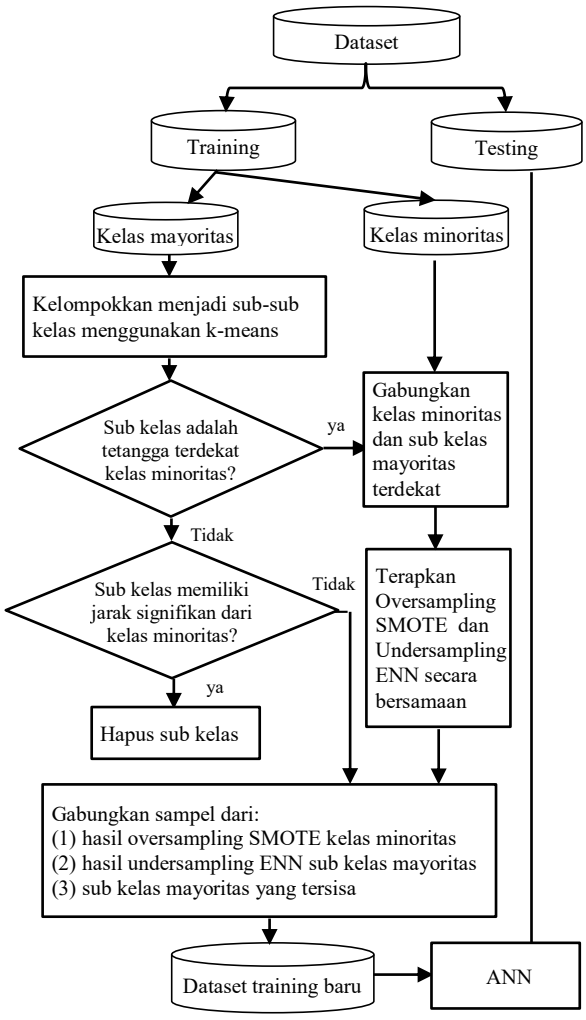
Tetangga terdekat sebanyak k *(k-nearest neighbor*) dari sampel x_i secara matematis dinyatakan:

$$\mathrm{KNN}(x_i, k) = \{y_i \in X | dist(x_j, x_i) \leq dist(x_i', x_i)\} \quad (2)$$

dimana x_i' adalah sampel tetangga terdekat ke- k dari x_i pada dataset X dan $dist$ adalah jarak antar sampel x_i dan tetangga terdekatnya yang umumnya menggunakan rumus *euclidean distance.*

2.3 Teknik Kombinasi *Undersampling* dan *Oversampling* yang Diusulkan

Teknik kombinasi *undersampling* dan *oversampling* untuk menyeimbangkan kelas pada penelitian ini dilakukan melalui serangkaian prosedur yang ditunjukkan pada Gambar 2.



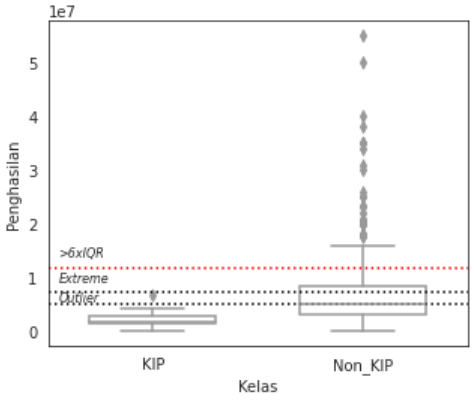
Gambar 2. Prosedur Kombinasi *Undersampling* dan Cluster-Based *Oversampling*

Langkah pertama adalah membagi dataset menjadi dua bagian, yaitu data training dan data testing. Data training akan dikenakan teknik *resampling* dalam rangka untuk membangun model klasifikasi NN, sedangkan data testing digunakan untuk menguji model klasifikasi NN yang didapatkan dari proses training. Data training dibagi berdasarkan status penerima KIP Kuliah, yaitu kelas KIP (kelas minoritas) dan kelas Non-KIP (kelas mayoritas).

Skema *clustering* pada penelitian ini hanya dikenakan pada kelas mayoritas saja. Kelas Non-KIP sebagai kelas mayoritas dibagi menjadi berberapa sub-kelas menggunakan algoritma *k-means*. Sub kelas Non-KIP dengan jarak terdekat dengan kelas KIP dipilih untuk kemudian digabungkan dengan kelas KIP. Dataset hasil penggabungan kedua kelas tersebut diterapkan teknik *oversampling* SMOTE pada kelas KIP dan *undersampling* ENN pada kelas Non-KIP terpilih. Pada teknik *oversampling* SMOTE, penggunaan dataset yang berasal dari kelas KIP dan sub kelas Non-KIP terdekat merupakan bentuk strategi penutupan sub kelas Non-KIP lainnya yang dalam studi Hassanat et al. (2022) diistilahkan *hidden subset.* Langkah tersebut merupakan teknik untuk menghindari *false sample* yang mungkin terjadi seperti yang diungkapkan oleh Hassanat et al. (2022). Penerapan teknik *undersampling* ENN yang hanya diterapkan pada kelas Non-KIP terpilih tersebut juga dapat menghindari penghapusan sampel secara acak yang dapat berpotensi menghilangkan sampel-sampel informatif pada kelas mayoritas secara keseluruhan.

Teknik *undersampling* juga dikenakan pada sub kelas Non-KIP yang memiliki jarak signifikan dari kelas minoritas menggunakan metode penghapusan sampel. Rumusan untuk menentukan jarak signifikan pada penelitian ini menggunakan rumusan dari box plot, yaitu ketika nilai minimal dari sub kelas Non-KIP berada di area *extreme* dari kelas KIP, yaitu lebih dari $3 \times Inter\ Quartil\ Range\ (IQR)$ dari box plot kelas KIP. Pada penelitian ini batas area yang ditetapkan untuk dapat diterapkan metode penghapusan langsung adalah lebih $6 \times IQR$ dari kelas KIP. Metode penghapusan sampel secara langsung didasarkan pada lebarnya jarak yang sangat ekstrim antar kedua kelas tersebut. Hal ini menunjukkan bahwa kedua kelas tersebut merupakan dua kelas yang berbeda. Oleh karena itu, sampel-sampel pada sub kelas mayoritas tersebut dapat langsung dilakukan penghapusan secara langsung. Hasil resampling dari semua metode dijadikan satu kesatuan dataset training untuk pembentukan model klasifikasi NN. Penerapan teknik SMOTE dan ENN pada penelitian ini menggunakan bantuan *toolboox imbalanced-learn* (Lemaitre et al., 2017).

Penerapan strategi yang diusulkan berdasarkan pada karakteristik data yang digunakan. Karakteristik data pada penelitian ini mengacu pada data yang bertipe numerik, yaitu variabel penghasilan. Gambar 3 merupakan visualisasi box plot dari variabel penghasilan.



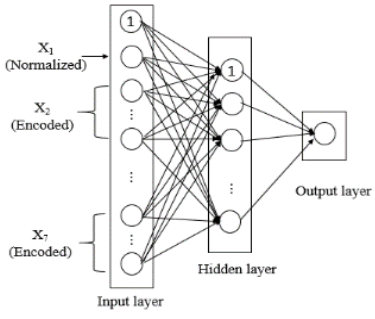
Gambar 3. Box Plot Variabel Penghasilan

Kelas Non-KIP memiliki cukup banyak pencilan data dengan jarak signifikan dari pusat data (data *outlier* dan data *extreme*). Di sisi lain, Kelas KIP memiliki banyak irisan data dengan kelas Non-KIP pada area *lower whisker* hingga *median*. Beberapa data *outlier* di kelas KIP mencapai area di *upper whisker* kelas Non-KIP. Oleh karena itu, kelas Non-KIP dikelompokkan menjadi beberapa sub kelas menggunakan algoritma *k-means clustering* dan dilihat jarak *cluster center* antar masing-masing sub kelas Non-KIP dengan kelas KIP. *Decision boundary* antara kelas KIP dan kelas Non-KIP terjadi pada area dimana antara kedua kelas memiliki kedekatan jarak, sehingga teknik *undersampling* SMOTE dan *oversampling* ENN hanya dikenakan pada kedua kelas tesebut. Metode penghapusan sampel secara langsung diterapkan pada sub kelas Non-KIP yang memiliki jarak ekstrim dari kelas KIP untuk lebih menyeimbangkan kelas.

Kelas KIP juga memiliki beberapa data di area *outlier* yang dapat berpengaruh terhadap pembentukan model NN. Namun demikian, variabel input pada NN tidak hanya variabel penghasilan, ada enam variabel lain yang juga memiliki pengaruh terhadap pembentukan model NN.

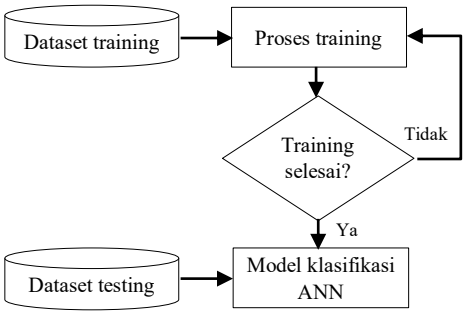
2.4 Pembentukan Model Klasifikasi NN

Pembentukan model klasifikasi menggunakan NN dibangun melalui proses training dengan menetapkan arsitektur jaringan terlebih dahulu. Arsitektur jaringan terdiri tiga *layer*, yaitu *input layer*, satu *hidden layer*, dan *output layer*. Arsitektur NN pada penelitian disajikan pada Gambar 4.



Gambar 4. Arsitektur NN

Input layer berisikan data input yang berupa tujuh varibael input yang telah dinormalisasi. Normalisasi untuk variabel input dengan tipe data berbentuk numerik menggunakan fungsi *Z-score Normalization*. Fungsi ini akan menskalakan data input ke dalam distribusi nilai dengan $\mu = 0$ dan $\sigma = 1$. Normalisasi untuk tipe data berbetuk kategorik menggunakan metode One-hot-encoding. Metode One-hot encoding akan menciptkan vektor biner baru sebagai reperesentasi data kategorik yang bernilai integer 0 dan 1. Hasil investigasi oleh (Hancock & Khoshgoftaar, 2020), metode One-hot-Encoding untuk varibel kategorik pada NN cukup baik dalam merancang NN. Jumlah neuron pada *hidden layer* disimulasikan di awal dan ditetapkan untuk seluruh model yang dibangun. Hal ini bertujuan untuk melihat pengaruh dari teknik *imbalance class* yang digunakan terhadap kinerja model yang dibangun tanpa melakukan perubahan pada arsitektur jaringan yang ditetapkan. *Output layer* berupa satu neuron dengan dua label nilai, yaitu nilai nol untuk kelas Non-KIP dan nilai satu untuk kelas KIP.



Gambar 5. Proses Pembentukan Model NN

Proses pembentukan model klasifikasi NN secara umum disajikan pada Gambar 5. Pada fase training, NN akan melakukan pelatihan jaringan dari dataset training yang telah dinormalisasi. Jaringan akan belajar terhadap pasangan data input-terget dari dataset training sedemikian hingga meminimkan fungsi *loss* dengan tujuan output dari jaringan nilainya mendekati atau sama dengan target data training. Proses training akan berhenti jika jumlah *epoch* telah mencapai dari yang ditentukan atau ketika metrik evaluasi kinerja model sudah mencapai nilai yang dinginkan (Farizawani et al., 2020). Hasil dari proses training ini adalah model klasifikasi terbaik selama proses simulasi dilakukan. Dataset testing digunakan untuk menguji model klasifikasi yang sudah terbentuk dan dihitung metriks pengukuran kinerja untuk kemudian dilakukan analisa kinerja dari setiap teknik yang digunakan.

Proses pembangunan model pada penelitian ini menggunkan Tensorflow yang merupakan *open machine learning platform* yang dikembangkan oleh Google. Algoritma pelatihan menggunakan *backpropagation* dengan fungsi aktivasi RELU, *sigmoid function* sebagai *optimizer*, dan nilai MSE sebagai fungsi *loss*.

2.5 Pengukuran Kinerja Klasifikasi

Nilai *accuracy* merupakan ukuran kinerja yang paling umum digunakan pada *machine learning*. Namun demikian, untuk kasus *imbalances class*, nilai *accuracy* tidak tepat untuk mengekspresikan kemampuan kinerja model klasifikasi karena nilai *accuracy* yang tinggi tidak mencerminkan kapasitas prediksi untuk kelas minoritas.

Luque et al. (2019) melakukan identifikasi terhadap beberapa metriks kinerja klasifikasi pada *imbalance class* berdasarkan *confusion matrix* (CM). Hasil simulasi menunjukkan bahwa parameter *Gometric Mean* (G-Mean) merupakan metrik kinerja *unbias* terbaik dan nilai *Matthews Correlation Coeffecient* (MCC) adalah pilihan terbaik jika klasifikasi error juga menjadi satu perimbangan. G-Mean merepresentasikan ukuran keseimbangan antara kinerja klasifikasi di kelas mayoritas dan kelas minoritas, sedangkan MCC merepresentasikan koefesien korelasi antara data aktual dengan data prediksi. MCC akan menghasilkan skor tinggi hanya jika prediksi memperoleh hasil yang baik di semua empat kategori CM. MCC memiliki rentang nilai $[-1, 1]$, sehingga pada beberapa kasus digunakan rumusan *normalized* MCC (nMCC) yang memiliki rentang $[0, 1]$. MCC mengghasilkan skor yang lebih informatif dan terpercaya dalam mengevaluasi klasifikasi biner dibandingkan dengan nilai *accuracy* dan F1 score (Chicco & Jurman, 2020). Penelitian ini juga menggunakan parameter yang umum digunakan untuk kinerja klasifikasi, yaitu *Sensitifity* atau *True Positif Rate* (TPR) yang mendeskripsikan tingkat prediksi kelas KIP yang benar dan *Specificity* atau *False Positif Rate* (FPR) yang mendeskripsikan tingkat prediksi kelas Non-KIP yang benar.

Tabel 2. Formasi Confusion Matrix

		Kelas Prediksi	
		KIP	**Non-KIP**
Kelas Aktual	**KIP**	**True Positive (TP)** (Terdeteksi benar kelas KIP)	**False Negatif (FN)** (Terdeteksi salah kelas Non-KIP)
	Non-KIP	**False Positive (FP)** (Terdeteksi salah kelas KIP)	**True Negative (TN)** (Terdeteksi benar kelas Non-KIP)

Tabel 2 merupakan formasi CM pada penelitian ini. Kelas minoritas (kelas KIP) memiliki label kelas positif, sedangkan kelas mayoritas (kelas Non-KIP) memiliki label negatif. Ada empat parameter pada CM, yaitu *True Positif* (TP) yang bermakna bahwa model mendeteksi kelas KIP dengan benar, *False Negatif* (FN) yang bermakna bahwa model mendeteksi salah sebagai kelas negatif yang seharusnya kelas positif, *True Negatif* (TN) yang bermakna bahwa model mendeteksi dengan benar kelas negatif, dan *False Positif* (FP) yang bermakna bahwa model mendeteksi salah sebagi kelas positif yang seharusnya kelas negatif. Formula matematis

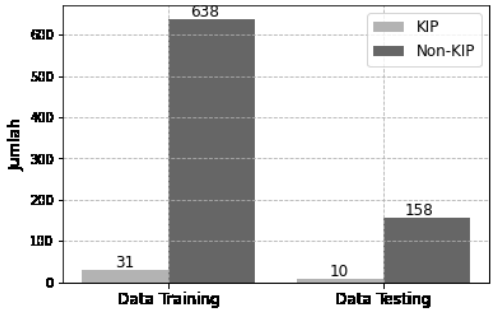
dari metriks ukuran kinerja yang digunakan pada penelitian ini disajikan pada Tabel 3 (Chicco & Jurman, 2020).

Tabel 3. Metriks Pengukuran Kinerja

Metrics	Formula
Sensitivity/ True Positive Rate (TPR)	$\frac{TP}{TP+FN}$
Specificity/ TNegativeatif Rate (TNR)	$\frac{TN}{TN+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
G-Mean	$\sqrt{Recall \times Specifity}$
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}$
nMCC	$\frac{1+MCC}{2}$

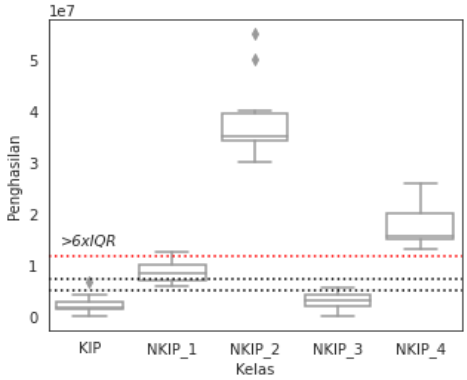
3. HASIL DAN PEMBAHASAN

Langkah pertama pada pembahasan ini adalah membagi dataset menjadi dua bagian, yaitu dataset training dan dataset testing. Proporsi dari pembagian data pada penelitian ini adalah 80% atau sebesar 669 sampel data digunakan untuk data training dan 20% atau sebesar 168 sampel data digunakan untuk testing. Pembagian data dilakukan secara acak dimana pada data training dan data testing sama-sama memiliki komposisi kelas mayoritas dan kelas minoritas yang tidak jauh berbeda. Gambar 6 merupakan komposisi KIP dan kelas Non-KIP pada data training dan data testing. Data training memiliki komposisi kelas KIP sebanyak 31 data atau 4.63% dari total data training, sedangkan data testing memiliki komposisi kelas KIP sebanyak 10 data atau 5.95% dari total data testing.



Gambar 6. Komposisi Kelas pada Data Training dan Data Testing

Tahap berikutnya adalah melakukan *clustering* menggunakan algoritma *k-means* pada data yang bertipe numerik, yaitu variabel penghasilan orang tua (X_1). Jumlah sub kelas (*cluster*) disimulasikan ke dalam beberapa nilai. Pada bagian ini divisualisasikan hasil pengelompokkan untuk empat sub kelas. Gambar 7 merupakan box plot variabel penghasilan pada kelas KIP dan empat sub kelas Non-KIP, yaitu sub kelas NKIP_1, NKIP_2, NKIP_3, dan NKIP_4. Gambar 7 memperlihatkan bahwa sub kelas NKIP_2 dan sub kelas NKIP_4 memiliki jarak yang signifikan jauh dengan *cluster center* kelas KIP. Nilai terkecil dari sub kelas NKIP_2 dan sub kelas NKIP_4 masing-masing adalah Rp.30.000.000,00 dan Rp.13.000.000,00 dimana kedua nilai tersebut lebih dari 6 kali dari nilai IQR kelas KIP, yaitu Rp.8.850.000,00. Oleh karena itu, kedua kelas tersebut dihapus secara langsung dari dataset. Sub kelas terdekat dengan kelas KIP adalah sub kelas NKIP_3.



Gambar 7. Box plot Variabel Penghasilan Pada Kelas KIP dan Sub-Sub Kelas Non-KIP Hasil *Clusterring*

Penggabungan teknik *undersampling* ENN dan teknik *oversampling* SMOTE dilakukan melalui beberapa skema dengan tujuan untuk melihat tingkat keberhasilan dari strategi penggabungan yang diusulkan. Tabel 4 merupakan dataset training dengan empat jenis dataset yang disimulasikan, yaitu: (1) Dataset_1, dataset asli tanpa melakukan *clusterring* pada kelas Non-KIP; (2) Dataset_2, dataset dengan *clusterring* kelas Non-KIP sebanyak 2 sub kelas; (3) Dataset_3, dataset dengan *clustering* kelas Non-KIP sebanyak 3 sub kelas; (4) Dataset_4, dataset dengan *clustering* kelas Non-KIP sebanyak 4 sub kelas.

Dataset_1 menyimulasikan tiga skema yang berbeda dengan perbedaan pada rasio sampling yang yang digunakan. Dataset_2 menyimulasikan 2 skema rasio yang berbeda dan 1 skema penghapusan langsung pada sub kelas Non-KIP yang berada di area *extreme*. Dataset_3 sampai dengan Dataset_4 menyimulasikan tiga skema berbeda dimana jumlah sub kelas yang dilakukan penghapusan sampel secara langsung dilakukan secara bertahap. Penghapusan sampel secara langsung dilakukan mulai dari kelas Non-KIP yang memiliki jarak terjauh dari *center* kelas KIP. Hal ini dilakukan untuk melihat pengaruh penghapusan langsung terhadap data-data yang berada di atas area *extreme*. Rasio sampling yang digunakan pada Tabel 4 merupkan hasil dari beberapa simulasi nilai dan diambil pada rasio sampling yang menghasilkan model klasifikasi dengan nilai metriks ukuran kinerja terbaik.

Tabel 4. Pembentukan Dataset Training

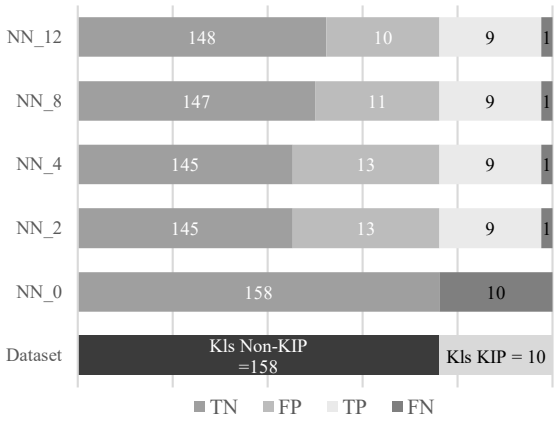
Kelas/Sub Kelas	Jumlah Sampel	Jumlah Sampel Setelah Resampling		
		Skema 1	Skema 2	Skema 3
Dataset_1 (tanpa clusterring)				
NKIP	638	495	513	504
KIP	31	453	167	129
Total	669	948	680	633
Rasio NKIP dan KIP		0.52:0.48	0.75:0.25	0.80:0.20
Dataset_2 (jumlah cluster = 2)				
NKIP_1	575	449	445	445
NKIP_2**	63	63	63	0
KIP	31	126	168	168
Total	669	581	676	613
Rasio NKIP dan KIP		0.80:0.20	0.75:0.25	0.73:0.27
Dataset_3 (jumlah cluster = 3)				
NKIP_1***	31	31	0	0
NKIP_2*	416	297	297	297
NKIP_3**	191	191	191	0
KIP	31	99	99	99
Total	669	618	587	396
Rasio NKIP dan KIP		0.86:0.14	0.86:0.14	0.85:0.15
Dataset_4 (jumlah cluster = 4)				
NKIP_1*	376	254	254	254
NKIP_2***	10	10	0	0
NKIP_3	199	199	199	199
NKIP_4***	53	53	53	0
KIP	31	83	83	83
Total	669	599	589	536
Rasio NKIP dan KIP		0.86:0.14	0.86:0.14	0.85:0.15

* Sub kelas NKIP terdekat dengan kelas KIP;
** Sub kelas NKIP yang berada di area *extreme* kelas KIP
***Sub kelas NKIP yang berada di area *>6xIQR* kelas KIP

Pembentukan model klasifikasi NN melalui proses training dilakukan pada seluruh skema dataset training pada Tabel 4. Empat dataset masing-masing memiliki tiga skema yang berbeda, sehingga model klasifikasi yang dihasilkan dari proses training ada sebanyak 12 model. Pada pembahasan ini juga akan ditunjukkan model klasifikasi NN yang terbentuk dari dataset training tanpa melakukan teknik *imbalance class*. Dataset training pada proses ini dinamai Dataset_0. Tabel 5 merupakan penamaan model klasifikasi yang dihasilkan dari proses training pada seluruh dataset yang terbentuk. Total terdapat 13 model klasifikasi yang dihasilkan.

Tabel 5. Hasil Pengukuran Kinerja

Dataset	Skema	Nama Model Klasifikasi
Dataset_0	-	NN_0
Dataset_1	Skema 1	NN_1
	Skema 2	NN_2
	Skema 3	NN_3
Dataset_2	Skema 1	NN_4
	Skema 2	NN_5
	Skema 3	NN_6
Dataset_3	Skema 1	NN_7
	Skema 2	NN_8
	Skema 3	NN_9
Dataset_3	Skema 1	NN_10
	Skema 2	NN_11
	Skema 3	NN_12



Gambar 8. Hasil Klasifikasi Model Terbaik Pada Setiap Formula Dataset

Model klasifikasi yang terbentuk kemudian diujikan pada data testing. Jumlah sampel pada data testing ada 168 sampel, dimana 158 sampel merupakan kelas Non-KIP dan 10 sampel merupakan kelas KIP. Gambar 8 merupakan hasil klasifikasi model dengan tingkat klasifikasi terbaik pada setiap formula dataset. Hasil simulasi menunjukkan bahwa Model NN_0 mampu mengklasifikasikan seluruh sampel pada kelas Non KIP dengan benar (TN = 158), namun seluruh sampel pada kelas KIP tidak mampu dikenali atau terklasifikasikan sebagai kelas Non KIP (TP=0). Model NN_0 terbentuk dari dataset yang tidak memperhitungkan *imbalance class*, akibatnya seluruh sampel pada kelas minoritas akan dikenali sebagai kelas mayoritas (FP=10). Penerapan teknik *imbalance class* yang digunakan pada penelitian ini mampu menghasilkan model yang dapat mengenali kelas KIP dengan jumlah yang sama. Dari 10 sampel pada kelas KIP hanya satu yang tidak mampu dikenali (FN=1 atau TP=9). Sedangkan pada kelas non KIP, model NN memiliki perbedaan jumlah sampel yang dikenali secara benar. NN_2 dan NN_4 mampu menghasilkan jumlah TN yang sama, yaitu 145 sampel. NN_8 memiliki jumlah TN sebesar 147 sampel dan NN_12 memiliki jumlah TN terbanyak, yaitu 148 sampel.

Hasil kinerja klasifikasi model secara lebih detail dapat dilihat dari nilai metriks ukuran kinerja yang digunakan pada penelitian ini. Tabel 6 merupakan hasil pengukuran kinerja berdasarkan nilai CM dari seluruh model NN yang dihasilkan. Model NN_0 memiliki nilai *accuray* tertinggi, yaitu 94.05%, namun demikian nilai TPR dan G-Mean adalah 0%. Hasil ini menunjukkan bahwa nilai *accuray* belum cukup mewakili untuk menunjukkan hasil kinerja klasifikasi secara keseluruhan dari suatu model. Model NN_2 memiliki hasil ukuran kinerja terbaik dibandingkan model lainnya pada formula Dataset_2 dengan nilai accuracy 91.67%, TPR sebesar 90%, TNR sebesar 90.88%, G-Mean sebeasar 90.88%, dan nMCC sebesar 78.86%. Hasil ini menunjukkan bahwa teknik kombinasi SMOTE dan ENN mampu meningkatkan nilai TPR, G-Mean, dan

nMCC dengan sangat signifikan. Hal yang menjadi perhatian ketika mengkombinasikan teknik *undersampling* dan *oversampling* adalah rasio sampling yang digunakan. Rasio sampling antara kelas mayoritas dan kelas minoritas yang menghasilkan kinerja model terbaik tidak berada tepat di titik setimbang atau sekitarnya. Formula Dataset_2 dengan rasio sampling 0.75: 0.25 mampu menghasilkan model dengan ukuran kinerja terbaik. Berdasarkan hasil tersebut, untuk rasio sampling pada dataset lainnya akan digunakan rasio yang lebih besar pada kelas mayoritas.

Tabel 6. Hasil Pengukuran Kinerja

Model	Metriks Kinerja (%)				
	Accuracy	TPR	TNR	G – Mean	nMCC
NN_0	94.05	0.00	100.00	0.00	-
NN_1	85.12	90.00	84.81	87.37	72.28
NN_2	**91.67**	**90.00**	**91.77**	**90.88**	**78.68**
NN_3	90.48	90.00	90.51	90.25	77.22
NN_4	91.67	90.00	91.77	90.88	78.68
NN_5	91.67	90.00	91.77	90.88	78.68
NN_6	91.67	90.00	91.77	90.88	78.68
NN_7	93.45	80.00	94.30	86.86	79.15
NN_8	**92.86**	**90.00**	**93.04**	**91.51**	**80.33**
NN_9	92.26	90.91	92.41	91.19	79.48
NN_10	91.67	90.00	91.77	90.88	78.68
NN_11	92.86	90.00	93.04	91.51	80.33
NN_12	**93.45**	**90.00**	**93.67**	**91.82**	**81.25**

Hasil ukuran kinerja model NN dengan kombinasi teknik *undersampling* dan *oversampling* yang diusulkan pada peneltian ini ditunjukkan pada nilai ukuran metriks kinerja model NN_4 sampai dengan NN_12. Model NN_4 sampai dengan NN_6 menghasilkan metriks ukuran kinerja yang sama dengan ukuran kinerja NN_2. Hal ini menunjukkan bahwa strategi penggabungan SMOTE dan ENN yang dilakukan hanya di kelas KIP dan sub kelas Non-KIP terdekat saja sudah mampu menghasilkan klasifikasi dengan tingkat kinerja yang sama dengan teknik penggabungan SMOTE dan ENN pada seluruh sampel pada kelas KIP dan kelas Non-KIP. Penghapusan sampel secara langsung pada formula Dataset_2 belum memiliki pengaruh terhadap kinerja model yang dihasilkan. Penghapusan langsung pada Dataset_2 dilakukan pada kelas Non-KIP yang berada di area *extreme* bukan pada area yang dijadikan batas pada penelitian ini.

Pengaruh penghapusan sampel secara langsung pada sub kelas Non_KIP baru terlihat pada Dataset_3 dan Dataset_4. Keseluruhan model yang dihasilkan dari formula Dataset_3 dan Dataset_4 mampu menaikkan nilai metriks kinerja klasifikasi dibandingkan dengan Dataset_2. NN_8 dan NN_12 merupakan model dengan nilai metriks kinerja klasifikasi terbaik pada masing-masing Dataset_3 dan Dataset_4. Model NN_9 mengalami penurunan nilai ukuran kinerja dibandingkan NN_8, namun masih lebih baik dibandingkan NN_7. Model NN_8 dihasilkan dari Dataset_3 dengan skema 3 dimana dilakukan metode penghapusan langsung pada sub kelas Non-KIP yang berada di area *extreme*. Hal ini menunjukkan bahwa penentuan batas pada area dimana diterapkan metode penghapusan sampel secara langsung perlu disimulasikan.

4. KESIMPULAN

Keseimbangkan distribusi antara kelas mayoritas dan kelas minoritas dengan rasio tertentu sebagai startegi *data processing level* dalam mengatasi permasalahan *imbalance class* merupakan hal yang mutlak diperlukan dalam membangun model klasifikasi menggunakan NN. Pengkombinasian teknik *oversampling* dan *cluster-based undersampling* berbasis jarak antar kelas minoritas dan sub-sub kelas mayoritas terpilih mampu menghasilkan model NN yang dapat memprediksi kelas minoritas dengan tepat secara signifikan. Penambahan metode penghapusan langsung pada sub kelas sampel mayoritas terpilih juga mampu menaikkan kinerja klasifikasi dalam memperediksi kelas mayoritas dengan tepat. Hal-hal yang berpengaruh terhadap kinerja model klasifikasi adalah pemilihan jumlah *clustering* kelas mayoritas, rasio sampling yang digunakan dalam menerapkan metode SMOTE dan ENN, dan pemilihan sub-kelas yang akan diterapkan metode penghapusan sampel secara langsung. Hal penting lain yang perlu diperhatikan dalam menghadapi *imbalance class* adalah metriks evaluasi kinerja yang digunakan untuk melihat bagaimana kinerja dari model *machine learning* yang dihasilkan.

Model NN_12 merupakan model terbaik yang dihasilkan pada penelitian ini yang berasal dari dataset dengan *clustering* sub kelas Non-KIP sebanyak 4 sub kelas dimana metode penghapusan sampel secara langsung diterapkan pada dua sub kelas yang keduanya berada di area lebih dari $6 \times IQR$ dari box plot kelas KIP. Model NN_12 mamiliki nilai *accuracy* 93.45%, TPR sebesar 90%, TNR sebesar 93.67%, G-Mean sebesar 91.82%, dan nMCC sebesar 81.25%. Model klasifikasi Penerima KIP Kuliah yang terbentuk dapat dimanfaatkan oleh institusi dalam menyeleksi dan mengevaluasi calon mahasiswa penerima KIP Kuliah agar tepat sasaran.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Pusat Penelitian dan Pengabdian Masyarakat PENS yang telah memberikan dukungan pendanaan dan fasilitas.

DAFTAR PUSTAKA

BACH, M., WERNER, A., ŻYWIEC, J., & PLUSKIEWICZ, W., 2017. The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. *Information Sciences*, *384*. https://doi.org/10.1016/j.ins.2016.09.038.

CHICCO, D., & JURMAN, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7.

DENG, M., GUO, Y., WANG, C., & WU, F., 2021. An oversampling method for multi-class imbalanced data based on composite weights. *PloS One*, *16*(11), e0259227. https://doi.org/10.1371/journal.pone.0259227.

DEVI, DEBASHREE, BISWAS, SAROJ KR., PURKAYASTHA, B., (2020). A Review on Solution to Class Imbalance Problem: Undersampling Approaches. *2020 International Conference on Computational Performance Evaluation (ComPE)*, 626–631. http://compe2020.com/.

FARIZAWANI, A. G., PUTEH, M., MARINA, Y., & RIVAIE, A., 2020. A review of artificial neural network learning rule based on multiple variant of conjugate gradient approaches. *Journal of Physics Conference Series*, *1529*, 22040. https://doi.org/10.1088/1742-6596/1529/2/022040.

FERNÁNDEZ, A., GARCIA, S., HERRERA, F., & CHAWLA, N., 2018. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192.

HANCOCK, J. T., & KHOSHGOFTAAR, T. M., 2020. Survey on categorical data for neural networks. *Journal of Big Data*, *7*, 1–41.

HASSANAT, A. B., TARAWNEH, A. S., ABED, S. S., ALTARAWNEH, G. A., ALRASHIDI, M., & ALGHAMDI, M., 2022. RDPVR: Random Data Partitioning with Voting Rule for Machine Learning from Class-Imbalanced Datasets. *Electronics (Switzerland)*, *11*(2). https://doi.org/10.3390/electronics11020228.

HASSANAT, A. B., TARAWNEH, A. S., ALTARAWNEH, G. A., & ALMUHAIMEED, A., 2022. Stop Oversampling for Class Imbalance Learning: A Review. *IEEE Access*, *10*, 47643–47660. https://doi.org/10.1109/ACCESS.2022.3169512.

KOZIARSKI, M., 2021. CSMOUTE: Combined Synthetic Oversampling and Undersampling Technique for Imbalanced Data Classification. *Proceedings of the International Joint Conference on Neural Networks*, *2021-July*. https://doi.org/10.1109/IJCNN52387.2021.9533415.

KUMAR, P., BHATNAGAR, R., GAUR, K., & BHATNAGAR, A., 2021. Classification of Imbalanced Data:Review of Methods and Applications. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 12077. https://doi.org/10.1088/1757-899X/1099/1/012077.

LEMAITRE, G., NOGUEIRA, F., & ARIDAS, C. K., 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *18*, 1–5. https://hal.inria.fr/hal-01516244.

LUQUE, A., CARRASCO, A., MART\'\IN, A., & DE LAS HERAS, A., 2019. The Impact of Class Imbalance in Classification Performance Metrics Based on the Binary Confusion Matrix. *Pattern Recogn.*, *91*(C), 216–231. https://doi.org/10.1016/j.patcog.2019.02.023.

NUGRAHA, W., MAULANA, M. S., & SASONGKO, A., 2020. Clustering Based Undersampling for Handling Class Imbalance in C4.5 Classification Algorithm. *Journal of Physics: Conference Series*, *1641*(1), 12014. https://doi.org/10.1088/1742-6596/1641/1/012014.

PUSLAPDIK, K. R., 2022. *Pedoman Pendaftaran Kartu Indonesia Pintar Kuliah - KIP Kuliah Merdeka*. https://kip-kuliah.kemdikbud.go.id/uploads/Pedoman-Pendaftaran-KIP-K-2022-ver-20220202---final_cd9b5e.pdf.

SARITAS, M. M., & YASAR, A., 2019. Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(2), 88–91. https://doi.org/10.18201//ijisae.2019252786.

SUSETYOKO, R., YUWONO, W., and PURWANTINI, E., 2022. Model Klasifikasi Pada Seleksi Mahasiswa Baru Penerima KIP Kuliah Menggunakan Regresi Logistik Biner, *JIP*, vol. 8, no. 4, pp. 31-40, Aug. 2022.

TSAI, C.-F., LIN, W.-C., HU, Y.-H., & YAO, G.-T., 2019. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, *477*, 47–54, https://doi.org/https://doi.org/10.1016/j.ins.2018.10.029.

XU, Z., SHEN, D., NIE, T., & KOU, Y., 2020. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *Journal of Biomedical Informatics*, *107*, 103465. https://doi.org/10.1016/j.jbi.2020.103465.

XU, Z., SHEN, D., NIE, T., KOU, Y., YIN, N., & HAN, X., 2021. A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data. *Inf. Sci.*, *572*, 574–589.

YANUAR, 2022. *Mahasiswa Penerima KIP Kuliah Dapat Diganti Bila Memenuhi Syarat Ini*. https://puslapdik.kemdikbud.go.id.