Telematika: Jurnal Informatika dan Teknologi Informasi *ISSN: 1829-667X / E-ISSN: 2460-9021*
Vol. 20, No. 2, Juni 2023, pp.238-249 *DOI:10.31515/telematika.v20i2.9676*

SMOTE and K-Means Preprocessing for Classification by Logistic Regression on Pima Indian Diabetes Dataset

Prapemrosesan Menggunakan SMOTE dan K-means untuk Klasifikasi Regresi Logistik pada Data Pima Indian Diabetes

Ahmad Taufiq Akbar [1], Hari Prapcoyo [2], Rochmat Husaini [3]

[1,3] Informatika, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia
[2] Sistem Informasi, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia

[1*]ahmadtaufiq.akbar@upnyk.ac.id, [2] hari.prapcoyo@upnyk.ac.id, [3]husaini@upnyk.ac.id
**: Penulis korenspondensi* (*corresponding author*)

Informasi Artikel
Received: May 2023
Revised: May 2023
Accepted: June 2023
Published: June 2023

Abstract

Purpose: Our study aims to combine pre-processing methods to develop a training data model from the Indian diabetic Pima dataset so that it can improve the performance of machine learning in recognizing diabetes
Design/methodology/approach: This research was started through several stages such as collecting the Pima indian diabetes dataset, pre-processing including k-means clustering, oversampling using SMOTE, then undersampling the dataset whose cluster is a minority in each class. Furthermore, the dataset is classified using machine learning namely logistic regression through 10 cross validation
Findings/result: The results of this classification performance show that the accuracy reaches 99.5% and is higher than the method in previous studies.
Originality/value/state of the art:
The method in this study uses SMOTE to handle data imbalances and k-means clustering to remove outliers by removing labels that do not match the majority cluster in each class so that clean data is produced and validation using logistic regression is more accurate than previous studies.

Keywords:SMOTE; k-means; Logistic Regression
Kata kunci: SMOTE; k-means; Logistic Regression

Abstrak

Tujuan: Penelitian ini bertujuan untuk menerapkan metode pre-processing untuk membentuk model data latih dari dataset Pima Indian diabetes sehingga dapat meningkatkan performa mesin pembelajaran dalam mengenali diabetes.
Perancangan/metode/pendekatan: Riset ini dimulai melalui

beberapa tahap yakni pengumpulan dataset Pima Indian diabetes, pre-processing meliputi clustering, oversampling menggunakan SMOTE, kemudian undersampling pada dataset pada klaster minoritas pada setiap kelas. Selanjutnya dataset diklasifikasikan menggunakan machine learning yakni metode regresi logistik melalui 10 cross validation
Hasil: Hasil dari performa klasifikasi ini menunjukkan akurasi mencapai 99,5% dan lebih tinggi daripada metode pada penelitian sebelumnya.
Keaslian/ *state of the art*: Metode dalam penelitian ini menggunakan SMOTE untuk menangani ketidakseimbangan data dan k-means klastering untuk membuang outlier dengan cara menghapus label yang tidak sesuai dengan klaster mayoritas pada setiap kelas sehingga dihasilkan data yang bersih dan pada validasi menggunakan logistic regression lebih akurat daripada penelitian sebelumnya.

1. Pendahuluan

Data mining adalah penambangan data untuk menemukan informasi, pola, ataupun tren sehingga dapat menghasilkan model untuk melakukan prediksi, klasifikasi maupun estimasi [1]. Bidang data mining telah menjadi tren di berbagai bidang ilmu karena semakin maraknya perkembangan teknologi informasi sehingga muncul fenomena big data [2]. Data-data dalam bidang kesehatan yang meliputi rekam medis, klinik kesehatan, pasien, prognosis, diagnosis, dan data-data penyakit hingga kini telah berkembang dalam jumlah besar. Perkembangan ini sangat baik jika dapat dikelola untuk diekstraksi sehingga menjadi knowledge dan model untuk dapat menghasilkan informasi yang lebih bermanfaat [3].

Data mining akan menghasilkan model yang baik dalam klasifikasi, prediksi maupun estimasi jika performa machine learning yang digunakan juga handal. Performa machine learning yang baik akan sangat berguna untuk melakukan klasifikasi secara tepat. Tidak sedikit metode klasifikasi mengalami modifikasi agar menjadi handal dalam melakukan klasifikasi sehingga memberikan akurasi yang lebih baik. Namun seiring dengan meningkatnya data-data informasi yang direpresentasikan dalam berbagai format dan platform mengakibatkan peningkatan fitur data dan struktur data menjadi usang [2] [4]. Terkadang tidak semua fitur maupun sample data tersebut mendukung untuk klasifikasi secara akurat, sehingga mengurangi performa algoritma klasifikasi [3].

Salah satu data kesehatan yang cukup banyak diteliti adalah Pima Indian Diabetes dari UCI Machine Learning Repository yang terdiri dari 8 fitur beserta 1 fitur sebagai kelas untuk status penyakit diabetes. Penyakit diabetes termasuk penyakit yang banyak diderita dan cukup mematikan sebagai akibat meningkatnya gula darah dalam tubuh [2]. Penyakit ini ditimbulkan karena tubuh tidak dapat menghasilkan insulin yang cukup untuk mengendalikan gula darah. Penyakit diabetes juga dapat menyebabkan komplikasi lain seperti serangan jantung, hipertensi, dan stroke [4]. Oleh karena itu data mining untuk mendeteksi seseorang cenderung

menderita diabetes akan sangat bermanfaat bila didukung dengan peningkatan akurasi. Disamping modifikasi algoritma, pendekatan lain yang dapat meningkatkan akurasi diantaranya dengan pre-processing data meliputi seleksi fitur, penambahan fitur, maupun modifikasi sampel data melalui oversample ataupun undersample [5].

Seleksi fitur merupakan tahap pengolahan data sebelum diklasifikasikan, sehingga hanya mempertahankan fitur-fitur yang signifikan relevan terhadap kelas. Dengan demikian akan meningkatkan performa klasifikasi dan mengurangi cost pada klasifikasi [6], [7]. Metode pemilihan fitur memiliki 3 basis yakni berbasis filter, wrapper, dan embedded [6]. Seleksi fitur berbasis wrapper dalam penelitian terdahulu terbukti memiliki performa yang terbaik daripada 2 metode lainnya.

Dataset Pima diabetes telah banyak dikaji pada sejumlah penelitian sebelumnya. Penelitian oleh Hairani dan tim mengenai klasifikasi diabetes menggunakan seleksi fitur berbasis wrapper dan algoritma correlated Naive Bayes yang telah dilengkapi bilangan laplacian untuk mencegah nilai probabilitas zero dan fitur R-square yang memuat nilai korelasi antara fitur-fitur terhadap kelas telah menghasilkan peningkatan akurasi hingga 4,1% pada dataset Pima Indian dibetes [4]. Penelitian oleh Santi Wulan dan tim [8] menggunakan metode Multiple Knots Spline Smooth Support Vector Machine (MKS-SSVM) yang merupakan pengembangan dari SVM untuk mengklasifikasikan diabetes pada dataset Pima secara 5 fold cross validation menghasilkan akurasi 93%. Rajni dan amandep melakukan penelitian pada klasifikasi dataset Pima diabetes dengan metode Rb-Bayes yang juga mampu menghindari probabilitas nol, telah menghasilkan akurasi yang lebih baik dari decision tree dan SVM standar [9]. Permana dan tim menggabungkan metode klasifikasi Linear Vector Quantization (LVQ) dengan genetic algorithm yang dapat mengoptimalkan pada bobot dalam LVQ sehingga menghasilkan peningkatan akurasi pada klasifikasi dataset pima diabetes mencapai 73.67% , lebih baik daripada dengan LVQ yang hanya 66.99% [10]. Adnan melakukan peneltian kalsifikasi pima diabetes menggunakan Self organizing map menghasilkan akurasi mencapai 85% lebih baik daripada random forest dengan 70% data training dan 30% data uji [11].

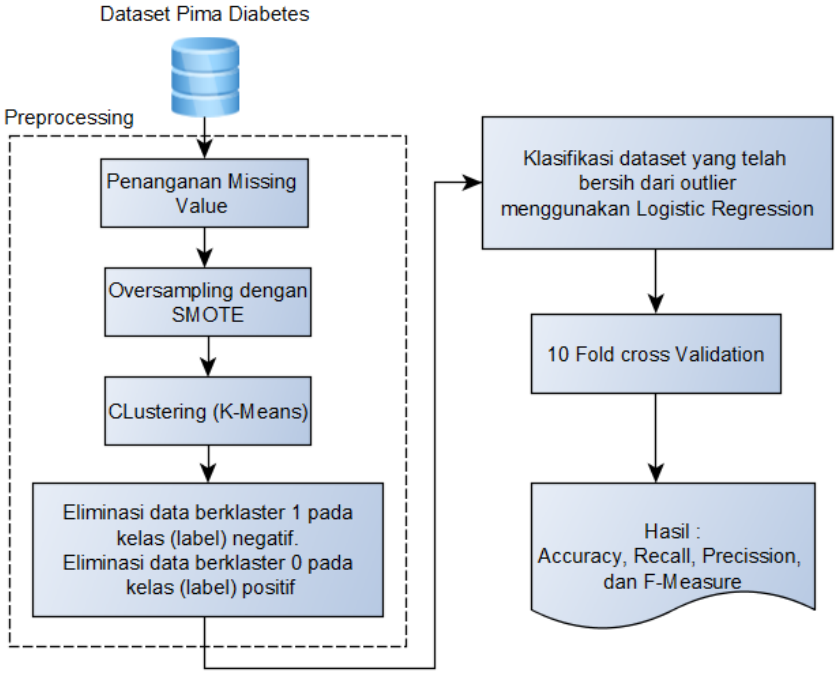
Metode pre-processing berupa modifikasi data melalui k-means-SMOTE kemudian diklasifikasikan dengan Naïve Bayes juga digunakan oleh Hairani et al. [12] dengan hasil akurasi mencapai 89 % lebih baik daripada metode seleksi fitur yang dilakukan oleh Hariani et al. di penelitian sebelumnya [4]. Ini menunjukkan pada beberapa kondisi, seleksi fitur kurang signifikan dalam memperbaiki akurasi klasifikasi dibanding dengan modifikasi data melalui resampling untuk menyeimbangkan data training sehingga performa klasifikasi meningkat. Penelitian oleh Akbar et al. juga mengenai klasifikasi pima diabetes menggunakan preprocessing supervised resample dan unsupervised resample, kemudian diklasifikasikan dengan k-NN menghasilkan akurasi mencapai 92,8% melalui 10 fold x validation [13]. Metode pre-processing selain seleksi fitur juga dilakukan pada penelitian oleh Barale et al. dengan melakukan imputasi missing value menggunakan K-NN, kemudian penghapusan outlier melalui boxplot dan k-means, sehingga diperoleh akurasi klasifikasi mencapai 99,3 % dengan algoritma classifier Logistic Regression [5]. Kaur meneliti klasifikasi pada dataset Pima diabetes dengan algoritma k-NN dan SVM, dimana akurasi klasifikasi k-NN mencapai 83 % mengungguli akurasi SVM yang berkisar 80% [2]

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v20i2.9676

Berdasarkan penelitian sebelumnya, pada beberapa kondisi, metode seleksi fitur dapat meningkatkan akurasi klasifikasi dibandingkan metode klasifikasi menggunakan machine learning default tanpa seleksi fitur. Pada penelitian yang dilengkapi pra-pengolahan dengan klastering (k-means) dan oversampling ternyata lebih meningkatkan akurasi daripada metode seleksi fitur. [12] [4] [5] [13]. Penelitian oleh Barale et al. memiliki akurasi tertinggi sebesar 99,3% dengan pre-processing berupa penghapusan outlier melalui k-means dan boxplot kemudian diklasifikasikan dengan Logistic Regression. Namun demikian, faktor ketidakseimbangan data yang mempengaruhi performa kalsifikasi tidak dipertimbangkan dalam penelitian tersebut. Oleh sebab itu, penelitian ini akan menggunakan metode pre-processing yang terdiri dari oversampling SMOTE, klasterisasi dengan k-means dan melakukan undersampling dengan menghapus data sampel pada klaster minoritas di setiap kelas. Dengan demikian dataset menjadi seimbang antar kelas dan lebih bersih karena klastering dengan k-means akan menampakkan batas setiap kelas sehingga data sampel outlier dapat dihilangkan [14]. Dengan preprocessing tersebut diharapkan performa algoritma yang mengklasifikasi dataset pima diabetes pada penelitian ini akan meningkat akurasinya.

2. Metode Penelitian

Peneitian ini dilakukan melalui beberapa tahap sebagaimana alur pada Gambar 1. Tahap pertama adalah pengumpulan dataset Pima Indian diabetes. Kedua adalah penanganan missing value. Ketiga dilakukan oversampling menggunakan SMOTE. Keempat k-means Clustering. Kelima eliminasi outlier (data sampel dari kelas yang tidak sesuai klasternya). Keenam dilakukan klasifikasi dengan Logistic Regression secara 10 fold cross validation. Terakhir adalah hasil berupa Akurasi Recall, dan F-Measure.



Gambar 1. Alur Metode Penelitian

2.1. Pengumpulan Data

Tahap pertama adalah pengumpulan data Pima diabetes dari UCI Machine Learning Repository. Dataset ini memuat sebanyak 768 data pasien perempuan yang berdarah India, dengan usia rata-rata 20 tahun keatas. Dataset ini terdiri dari 8 atribut yang merupakan data medis bertipe numerik pada setiap pasien. Kelas atau output atribut berupa tested positive (diabetes) dan tested negative (tidak diabetes). Sebanyak 500 data merupakan data pasien yang negatif, dan sisanya sebanyak 268 adalah data pasien yang positif diabetes. Pada Tabel 1 menunjukkan spesifikasi dari dataset Pima diabetes yag bersumber dari Original.

Tabel 1. Deskripsi Dataset Pima Indian Diabetes

No	Atribut (Fitur)	Tipe
1	*pregnant*	*Numeric*
2	*glucose*	*Numeric*
3	*diastolic*	*Numeric*
4	*triceps*	*Numeric*
5	*insulin*	*Numeric*
6	*bmi*	*Numeric*
7	*diabetes*	*Numeric*
8	*age*	*Numeric*
9	*Test (class)*	*Nominal*

Source of Dataset: National Institute of Diabetes and Digestive and Kidney Diseases dan kini tersimpan di UCI Machine Learning Repository. Sebanyak 8 atribut dataset sebagaimana pada Tabel 1 ini merupakan indikasi berdasarkan standar WHO untuk pasien mengidap diabetes atau tidak terkena diabetes.

2.2. Penanganan Missing Value

Tahap kedua yakni penanganan Missing Value. Proses ini dilakukan dengan mengisikan nilai mean (rerata) dari nilai-nilai seluruh data pada atribut yang mengalami missing value tersebut atau dengan memberikan nilai 0. Pengisian nilai ini akan lebih baik daripada menghapus data-data yang mengandung missing value karena dapat menghilangkan informasi penting. Pada dataset Pima ini, beberapa atribut yang mengandung missing value dan diisikan dengan nilai 0 , diantaranya atribut: diastolic, triceps, dan insulin.

2.3. Penanganan Ketidakseimbangan Data

Datset Pima diabetes memiliki jumlah yang tidak seimbang antara data pada kelas positif sebanyak 268 data dengan kelas negatif sebanyak 500 data. Ketidakseimbangan data menyebabkan performa machine learning cenderung mengklasifikasikan kedalam kelas mayoritas (tested negative) karena referensi data latih yang kurang pada kelas positif (minoritas). Keadaan ini dapat berpotensi overfitting pada kelas mayoritas, sehingga proporsi kelas minoritas perlu untuk diseimbangkan. Dalam penelitian ini penyeimbangan data menggunakan SMOTE karena SMOTE memunculkan data unik dari data-data terdekat dengan data-data kelas minoritas dan tidak menghapus data-data asli [13].

Metode oversampling SMOTE terdiri dari beberapa proses sebagai berikut: Xnew adalah data sampel yang akan dibuat oleh metode SMOTE. Xminor adalah sample data dari kelas minoritas yang akan diperbanyak (oversampled). Xknn adalah data sampel dari k sampel terdekat pada Xminor. δ adalah nilai acak bilangan real antara 0 and 1. Sampel data Xnew

yang telah dibuat SMOTE akan dimasukkan pada sample-sample kelas minoritas untuk menyeimbangkan banyaknya data kelas minoritas terhadap banyaknya data sample kelas mayoritas [15].

Tabel 1. Oversampling dengan SMOTE

Metode *Oversampling* SMOTE	
1	Ukur jarak Euclidean antar data pada kelas minoritas
2	Tentukan persentase data yang akan dikenakan *oversampling* dengan SMOTE
3	Tentukan banyaknya k data terdekat dalam kelas minoritas
4	Hasilkan data baru melalui persamaan berikut; Xnew=Xminor+(Xknn- Xminor)× δ

2.4. Klastering Data

Data yag telah dilakukan oversampling melalui SMOTE kemudian diklasterisasi menggunakna k-means. K-means merupakaan metode klaster dengan mecari centroid atau pusat klaster berdasarkan pengukuran jarak antara dua data. Algoritma Klasterisasi k-means terdiri dari beberapa tahap sebagai berikut, tentukan k klaster yang akan digunakan. Tentukan sentroid awal setiap klaster secara acak. Iterasikan pada tahap penentuan setiap data samepl yang terdekat dengan masing-masing sentroid dan perhitungan centroid baru pada setiap klaster. Iterasi berhenti jika tidak ada lagi perubahan sentroid pada setiap klaster [14]. Penelitian ini lebih mempertimbangkan pada proporsi klaster dalam setiap kelas sehingga akan terlihat data-data berklaster mayoritas pada setiap kelas. Dengan demikian dapat terlihat batas pada setiap kelas sehingga data-data berklaster minoritas pada stiap kelas dapat kita eliminasi untuk menghindari outlier dan mempertegas batas antarkelas [14]. Batas antar kelas perlu diperjelas karena oversampling dengan SMOTE memiliki kekurangan berupa ketidakjelasan batas antar kelas [16] [17] Setelah tahap tersebut maka akan didapatkan dataset yang paling dekat dengan sentroid masing-masing kelas.

2.5. Klasifikasi

Metode klasifikasi yang digunakan dalam penelitian ini adalah Logistic Regression. Logistic Regression sangat berguna dalam mengklasifikasikan kelas biner. dan mengetahui hubungan antara atribut respon (kelas) dengan atribut predictor yang dapat bersifat metric ataupun non-metrik. Logistic Regression yang diterapkan dalam penelitian ini menggunakan multinomial Logistic Regression yang terdapat pada library weka. Metode ini menggunakan ridge estimator yang dapat dikalikan dengan nilai nilai koefisien fitur (B). Jika terdapat sebanyak k kelas (label) dan n baris data (instances), maka koefisien B yang direpresentasikan dalam matriks dapat dihitung sebagai matriks berukuran m×(k-1). Probabilitas kelas j selain kelas terakhir dapat dihitung dengan rumus:

$$P_j\,(x_i) = \frac{e^{(x_i B_j)}}{1 + e^{\sum_{j=1}^{k-1}(x_i B_j)}} \quad (1)$$

Untuk probabilitas kelas terakhir dapat dhitung dengan rumus:

$$P_{jLast}\,(x_i) = 1 - \sum_{j=1}^{k-1} \Big(P_j\,(x_i) \Big) \quad (2)$$

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v20i2.9676

Sehingga

$$P_{jLast}\,(x_i) = \frac{1}{1+ e^{\sum_{j=1}^{k-1}(x_i B_j)}} \quad (3)$$

Fungsi negatif multinomial log-likelihood dinotasikan sebagai berikut:

$$L = -\sum_{i=1}^{n}\left(\sum_{j=1}^{k-1}\left(Y_{ij} \times \ln\left(P_j\,(x_i)\right)\right) + \left(1 - \sum_{j=1}^{k-1} Y_{ij}\right) \times \ln\left(1 - \sum_{j=1}^{k-1} P_j\,(x_i)\right)\right) + ridge \times B^2 \quad (4)$$

Untuk menemukan matriks B ketika L minimum, digunakan metode Quasi-Newton sehingga dapat diperoleh nilai optimal dari matriks B berukuran m*(k-1). Sebelum menggunakan metode miltinomial Logistic Regression ini, fitur yang bernilai nominal perlu ditransformasikan kedalam fitur numerik [18].

3. Hasil dan Pembahasan

Dataset Pima Indian Diabetes terdiri dari 768 instances (sampel data) dengan 8 atribut dan 1 atribut sebagai kelas. Struktur Dataset ini dapat dilihat pada Tabel 1. Terdapat missing value pada atribut insulin di sejumlah instances. Sehingga dataset ini dilakukan imputasi dengan nilai 0. Imputasi dengan nilai 0 ini untuk menyesuaikan dengan kondisi real bahwa nilai tersebut memang tidak muncul. Dengan imputasi ini, maka semua atribut telah terisi dengan nilai numerik yakni 0 kecuali atribut kelas yang memuat nilai nominal “positive” atau “ negative”. Pada Barele et al, imputasi dilakukan dengan metode k-*nearest neighbor* dengan rerata dari sample terdekat untuk atribut numerik dan dengan cara *majority vote* untuk atribut kategori(nominal) [5]. Sebagaimana pada Gambar 2 kita melihat bahwa proporsi kelas pada dataset diabetes ini tidak seimbang. Jumlah instance kelas tested-negative (berwarna biru) adalah 500, sedangkan jumlah kelas tested-positive (berwarna merah) adalah 268.

Gambar 2. Proporsi kelas dataset Pima

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v20i2.9676

Sehingga perlu dilakukan penyeimbangan jumlah instance pada kedua kelas, agar machine learning dapat mengklasifikasikan melalui pembelajaran dari data training yang setara. Dalam menyeimbangkan dataset in maka dilakukan oversampling dengan metode SMOTE untuk membangkitkan instance (data sampel) unik dari instance pada kelas minoritas, sehingga jumlah instance kelas minoritas menjadi seimbang dengan jumlah instance kelas mayoritas. Setelah proses oversampling dengan SMOTE ,jumlah instance kelas positif meningkat sebagaimana pada Gambar 3.

Gambar 3. Dataset yang telah seimbang

Dengan jumlah instance total menjadi 1036, yang terdiri dari kelas tested-negative sebanyak 500 dan kelas positif menjadi sebanyak 536. Dataset hasil SMOTE ini diklasterisasi menggunakan k-means sehingga hasil klasterisasi dapat dilihat pada Gambar 4.

Gambar 4. Dataset setelah diklasterisasi

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v20i2.9676

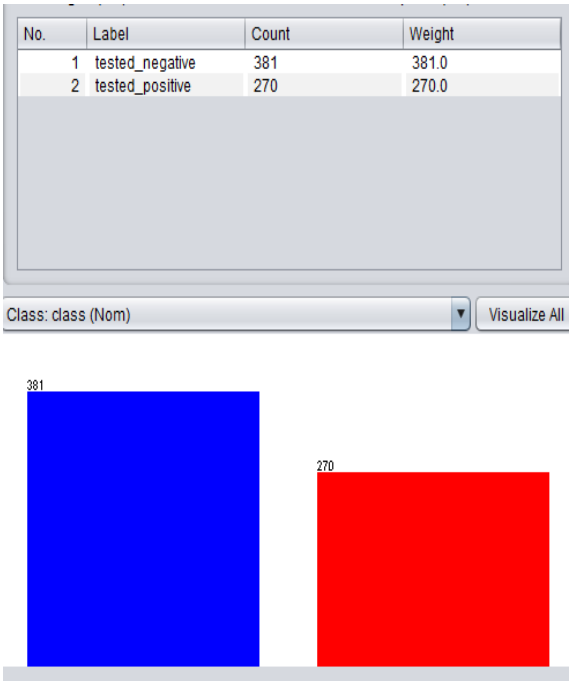
Terlihat pada Gambar 4 bahwa warna biru untuk jumlah instance dari kelas positif maupun negatif yang termasuk dalam klaster 0 sebanyak 647 instance. Pada warna merah adalah jumlah instance dari kelas positif maupun negatif yang termasuk dalam klaster 1 sebanyak 389 instance. Jika kita melihat distribusi klaster 0 dan klaster 1 pada setiap kelas maka kita akan melihat batas antara outlier dan data yang baik. Sebagaimana pada Gambar 4. Selanjutnya kita akan melakukan undersampling dengan penghapusan sejumlah instance untuk menyeleksi dan mempertahankan sejumlah instances yang berada pada klaster mayoritas di setiap kelasnya, dengan demikian data sampel outlier dapat terihat pada sejumlah instance diluar n instance yang terseleksi sebagaimana Gambar 5.

Gambar 5. Outlier berupa klaster minoritas di setiap kelas

Pada Gambar 5, terlihat bahwa di dalam kelas tested-negative yang terdiri dari 500 instances terdapat sebagian besar jumlah instance yang berada di klaster 0 (biru). Dengan demikian pada kelas tested-negative ini, klaster 0 (biru) adalah mayoritas sehingga kita eliminasi sejumlah instance pada kelas tested negative yang berklaster 1 (minoritas) yang berwarna merah. Selanjutnya pada kelas tested-positive yang terdiri dari 536 instance terdapat sebagian besar jumlah instance berada di klaster 1 (merah). Sehingga pada kelas positif ini klaster 1 (merah) adalah mayoritas, oleh sebab itu kita eliminasi sejumlah instance pada kelas positif yang berklaster 0 (minoritas) yang berwarna biru.

Setelah dilakukan penghapusan pada instance yang berada dalam klaster minoritas di setiap kelas, diperoleh sisa dataset sebanyak 651 instance data, dengan 381 untuk kelas negative dan 270 untuk kelas positif sebagaimana pada Gambar 6.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v20i2.9676



Gambar 6. Dataset setelah eliminasi outlier

Jika kita perhatikan pad Gambar 6, sebanyak 270 instance kelas positif dan 381 instance kelas negative masih relatif seimbang karena rasio kelas tersebut adalah 270:381 yakni 70.8%. Sedangkan pada data asli sebelum dilakukan oversampling, klasterisasi, dan undersampling rasio kelas adalah 268 : 500 yakni 53%. Dengan demikian telah diperoleh model dataset yang lebih bersih untuk diklasifikasikan.

Untuk menguji model yang telah dibentuk, maka dilakukan klasifikasi Logistic Regression dengan 10 fold cross validation. Dengan validasi ini, Logistic Regression mengklasifikasikan dataset dengan proporsi 90% data training dan 10% data testing pada setiap fold, pada fold berikutnya 10 % data testing tersebut bergantian diklasifikasikan hingga 10 fold. Hasil dari pengujian 10-fold cross-validation dalam penelitian ini mencapai akurasi 99,6%.

Nilai akurasi dalam penelitian ini lebih tinggi daripada hasil yang diperoleh Barale et al., yakni 99,4% [5], karena pada riset tersebut tidak mempertimbangkan penanganan ketidakseimbangan data instance melalui oversampling. Hasil akurasi dalam studi ini juga menegaskan bahwa klasterisasi k-means terhadap dataset Pima Indian diabetes yang telah mengalami oversampling SMOTE dapat membantu membersihkan data dari outlier sehingga pada setiap kelas hanya terdiri dari data milik klaster mayoritas. Dengan demikian, neighborhood instance pada setiap kelas menjadi semakin kuat, sehingga batas antar kelas menjadi jelas. Batas yang jelas pada setiap kelas, keseimbangan kelas, dan neighborhood yang kuat dapat meningkatkan performa algoritma klasifikasi sebagaimana pada penelitian sebelumnya [17] [13].

Performa klasifikasi bergantung pada representasi data. Ketika preprocessing dilakukan dengan tepat maka dapat meningkatkan akurasi klasifikasi. Data yang digunakan pada klasifikasi dalam penelitian ini diseimbangkan dahulu dengan SMOTE kemudian diseleksi melalui klustering untuk mengelompokkan sample-sampel yang paling

dipertimbangkan dalam klasifikasi. Perbandingan akurasi dari hasil penelitian ini dengan penelitian barele et al 2016 [5] dapat disajikan pada Tabel 2 berikut.

Tabel 2. Perbandingan akurasi klasifikasi dari model yang diusulkan dengan model lain

Method	Accuracy dalam %	Referensi
k-means+LR *	99.33	Barele et al. 2016 [5]
k-means+ANN(MFP) *	98.57	Barele et al. 2016 [5]
k-means+DT *	97.99	Barele et al. 2016 [5]
k-means+SVM *	97.13	Barele et al. 2016 [5]
data K-means+ANN (RBF)*	95.70	Barele et al. 2016 [5]
SMOTE + K-means + LR	99.60	Paper ini, 2023

4. Kesimpulan

Prapemrosesan menggunakan SMOTE dapat menangani ketidakseimbangan data. Metode klasterisasi k-means dapat mendeteksi outlier dari hasil oversampling dengan SMOTE sehingga memudahkan untuk penghapusan data-data yang klasternya tidak sesuai dengan kelasnya. Prapemrosesan menggunakan SMOTE dan penghapusan outlier melalui k-means ini dapat meningkatkan performa klasifikasi menggunakan Logistic Regression dengan hasil akurasi 99,6% atau lebih tinggi 2% dari akurasi yang dicapai pada penelitian sebelumnya pada dataset yang sama.

Daftar Pustaka

[1] M. Lestandy, A. Faruq, and A. Faruq, "Klasifikasi pendonor darah potensial menggunakan pendekatan algoritme pembelajaran mesin," vol. 8, no. July, pp. 217–221, 2020, doi: 10.14710/jtsiskom.2020.13619.

[2]. R. Kaur, "Predicting diabetes by adopting classification approach in data mining," Int. J. Informatics Vis., vol. 3, no. 2–2, pp. 218–221, 2019, doi: 10.30630/joiv.3.2-2.229.

[3]. S. N. Khan et al., "Comparative analysis for heart disease prediction," Int. J. Informatics Vis., vol. 1, no. 4–2, pp. 227–231, 2017, doi: 10.30630/joiv.1.4-2.66.

[4]. H. Hairani and M. Innuddin, "Kombinasi Metode Correlated Naive Bayes dan Metode Seleksi Fitur Wrapper untuk Klasifikasi Data Kesehatan," J. Tek. Elektro, vol. 11, no. 2, pp. 50–55, 2019, doi: 10.15294/jte.v11i2.23693.

[5]. M. S. Barale and D. T. Shirke, "Cascaded Modeling for PIMA Indian Diabetes Data," Int. J. Comput. Appl., vol. 139, no. 11, pp. 1–4, 2016, doi: 10.5120/ijca2016909426.

[6]. J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 13, no. 5, pp. 971–989, 2016, doi: 10.1109/TCBB.2015.2478454.

[7]. N. K. Suchetha, A. Nikhil, and P. Hrudya, "Comparing the wrapper feature selection evaluators on twitter sentiment classification," ICCIDS 2019 - 2nd Int. Conf. Comput. Intell. Data Sci. Proc., pp. 1–6, 2019, doi: 10.1109/ICCIDS.2019.8862033.

[8]. S. W. Purnami, A. Embong, J. M. Zain, and S. P. Rahayu, "A new smooth support

vector machine and its applications in diabetes disease diagnosis," J. Comput. Sci., vol. 5, no. 12, pp. 1003–1008, 2009, doi: 10.3844/jcssp.2009.1003.1008.

[9]. R. Bhalla and A. Bagga, "Opinion mining framework using proposed rb-bayes model for text classification," Int. J. Electr. Comput. Eng., vol. 9, no. 1, pp. 477–484, 2019, doi: 10.11591/ijece.v9i1.pp477-484.

[10]. I. Permana, N. E. Rozanda, F. Syafria, and F. N. Salisah, "Optimization learning vector quantization using genetic algorithm for detection of diabetics," Indones. J. Electr. Eng. Comput. Sci., vol. 12, no. 3, pp. 1111–1116, 2018, doi: 10.11591/ijeecs.v12.i3.pp1111-1116.

[11]. S. A. D. Alalwan, "Diabetic analytics: Proposed conceptual data mining approaches in type 2 diabetes dataset," Indones. J. Electr. Eng. Comput. Sci., vol. 14, no. 1, pp. 85–95, 2019, doi: 10.11591/ijeecs.v14.i1.pp88-95.

[12]. H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," J. Teknol. dan Sist. Komput., vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[13]. A. T. Akbar, R. Husaini, B. M. Akbar, and S. Saifullah, "A proposed method for handling an imbalance data in classification of blood type based on Myers-Briggs type indicator," J. Teknol. dan Sist. Komput., vol. 8, no. 4, pp. 276–283, 2020, doi: 10.14710/jtsiskom.2020.13625.

[14]. S. Sugriyono and M. U. Siregar, "Preprocessing kNN algorithm classification using K-means and distance matrix with students' academic performance dataset," J. Teknol. dan Sist. Komput., vol. 8, no. 4, pp. 311–316, 2020, doi: 10.14710/jtsiskom.2020.13874.

[15]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Comparison of Balancing Techniques for Unbalanced Datasets," Mach. Learn. Gr. Univ. Libr. Bruxelles Belgium, vol. 16, no. 1, pp. 321–357, 2002, doi: 10.1613/jair.953.

[16]. T. E. Tallo and A. Musdholifah, "The Implementation of Genetic Algorithm in Smote (Synthetic Minority Oversampling Technique) for Handling Imbalanced Dataset Problem," Proc. - 2018 4th Int. Conf. Sci. Technol. ICST 2018, vol. 1, pp. 1–4, 2018, doi: 10.1109/ICSTC.2018.8528591.

[17]. N. Cahyana, S. Khomsah, and A. S. Aribowo, "Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting," Proceeding - 2019 5th Int. Conf. Sci. Inf. Technol. Embrac. Ind. 4.0 Towar. Innov. Cyber Phys. Syst. ICSITech 2019, pp. 217–222, 2019, doi: 10.1109/ICSITech46713.2019.8987499.

[18]. X. Xu and E. Frank, "Logistic regression and boosting for labeled bags of instances," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3056, pp. 272–281, 2004, doi: 10.1007/978-3-540-24775-3_35..