JURNAL MATEMATIKA, STATISTIKA DAN KOMPUTASI e-ISSN: 2614-8811
Published by Departement of Mathematics, Hasanuddin University, Indonesia
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 21, No. 1, September 2024, pp. 88-102
DOI: 10.20956/j.v21i1.35552

Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models

Pengaruh Random Undersampling, Oversampling, dan SMOTE terhadap Kinerja Model Prediksi Penyakit Kardiovaskular

Uswatun Hasanah[1], Agus Mohamad Soleh[2], Kusman Sadik[3]
[1,2,3]Pascasarjana Statistika dan Sains Data, IPB University
***Email**: [1]21uswatun@apps.ipb.ac.id, [2]agusms@apps.ipb.ac.id, [3]kusmans@apps.ipb.ac.id*

Abstract

Cardiovascular Disease (CVD) or commonly known as Heart Disease is a leading cause of mortality globally, prompting extensive research into predictive models to assess individual risk and plan preventive measures. Machine learning approaches such as Random Forest, Support Vector Machine (SVM), and LASSO Logistic Regression have showed promise. Recent studies have indicated that traditional resampling methods like Random Oversampling, Random Undersampling, and SMOTE may not significantly improve model discrimination. This study aims to evaluate the impact of these techniques on the performance of Cardiovascular Disease (CVD) prediction models, utilizing data from the UCI Machine Learning Heart Disease database. By employing LASSO Logistic Regression, Random Forest, and Support Vector Machine (SVM) with resampling techniques, including Random Oversampling, Random Undersampling, and SMOTE. This research seeks to enhance understanding of model performance in addressing class imbalances within the dataset and contribute to refining cardiovascular disease (CVD) prediction strategies. This study demonstrates that the use of the SMOTE technique significantly enhances the performance of cardiovascular disease (CVD) prediction models. Specifically, when combined with the Random Forest algorithm, SMOTE achieves the best performance in terms of accuracy, sensitivity, and specificity. This highlights the importance of selecting appropriate resampling techniques to handle class imbalance in datasets. Consequently, this research contributes to refining CVD prediction strategies and provides new insights into improving prediction accuracy in imbalanced medical data.

Keywords: Cardiovascular Disease; Machine Learning; LASSO Logistic Regression; Random Forest; Support Vector Machine (SVM); Resampling Techniques

Abstrak

Penyakit Kardiovaskular (CVD) atau yang dikenal sebagai Penyakit Jantung merupakan penyebab kematian paling umum di seluruh dunia. Pendekatan pembelajaran mesin seperti Random Forest, Support Vector Machine (SVM), dan Regresi Logistik LASSO telah menunjukkan potensi untuk memodelkan prediktif guna menilai resiko individu dan

merencanakan langkah-langkah pencegahan. Studi terbaru mengindikasikan bahwa metode resampling tradisional seperti Random Oversampling, Random Undersampling, dan SMOTE mungkin tidak secara signifikan meningkatkan diskriminasi model. Penelitian ini bertujuan untuk mengevaluasi dampak teknik-teknik ini terhadap kinerja model prediksi Kardiovaskular (CVD), dengan menggunakan data dari database UCI Machine Learning. Dengan menerapkan Regresi Logistik LASSO, Random Forest, dan Support Vector Machine (SVM) dengan teknik resampling, termasuk Random Oversampling, Random Undersampling, dan SMOTE. Penelitian ini menunjukkan bahwa penggunaan teknik SMOTE secara signifikan meningkatkan kinerja model prediksi penyakit kardiovaskular (CVD). Teknik ini, terutama jika digunakan bersama dengan algoritma Random Forest, menghasilkan kinerja terbaik dalam hal akurasi, sensitivitas, dan spesifisitas. Hal ini menekankan pentingnya pemilihan teknik resampling yang tepat dalam menangani ketidakseimbangan kelas dalam dataset. Dengan demikian, penelitian ini berkontribusi pada penyempurnaan strategi prediksi CVD dan memberikan wawasan baru tentang cara meningkatkan akurasi prediksi dalam data medis yang tidak seimbang

Kata kunci: Penyakit Kardiovaskular; Pembelajaran Mesin; Regresi Logistik LASSO; Random Forest; Support Vector Machine (SVM); Teknik Resampling

1. PENDAHULUAN

Penyakit *Cardiovascular* (CVD) atau biasa disebut Penyakit Jantung merupakan salah satu penyakit kronis yang menjadi penyebab kematian paling umum di seluruh dunia. Penyakit *Cardiovascular* (CVD) menyebabkan lebih dari 17 juta kematian, kehilangan 330 juta tahun kehidupan, dan 35,6 juta tahun hidup cacat pada tahun 2017 di seluruh dunia [15]. Prediksi penyakit *Cardiovascular* (CVD) telah menjadi fokus utama dalam penelitian medis karena besarnya dampak yang ditimbulkan oleh penyakit tersebut. Prediksi penyakit *Cardiovascular* (CVD) yang dilakukan sedini mungkin dengan tujuan untuk mengembangkan model yang efektif guna menilai risiko seseorang terhadap penyakit *Cardiovascular* (CVD). Hal ini dilakukan untuk melakukan evaluasi lebih awal dan merencanakan tindakan pencegahan yang tepat guna menghindari dampak penyakit *Cardiovascular* (CVD).

Teknik pembelajaran mesin telah meningkat dari tahun ke tahun dalam pengembangan model prediksi penyakit *Cardiovascular*(CVD). Penanganan ketidakseimbangan kelas dalam data adalah bagian penting dari pengembangan model prediksi. Untuk mengatasi masalah ketidakseimbangan kelas, para peneliti telah mengembangkan teknik pembelajaran ketidakseimbangan untuk memproses dan mengekstraksi informasi dari data dengan distribusi yang sangat tidak merata [13]. Untuk mengatasi masalah kesalahan pelabelan yang sering terjadi pada data kelas minoritas pada model pengklasifikasi dibutuhkan teknik pengambilan ulang sampel seperti *undersampling* atau *oversampling*. Penelitian terbaru yang berfokus pada regresi logistik dengan teknik *resampling Random Oversampling* (ROS) dan *Random Undersampling* (RUS), dan SMOTE untuk memperbaiki ketidakseimbangan kelas menunjukkan bahwa penggunaan metode-metode ini tidak meningkatkan diskriminasi model [16]. Penelitian tersebut berfokus pada data dimensi rendah dengan ukuran sampel yang lebih kecil, sehingga dampak dari dari ketidakseimbangan kelas terhadap kinerja model prediksi yang dikembangkan dalam basis data penyakit *Cardiovascular* (CVD) belum terbukti.

Metode Analisis data dengan algoritma *Machine Learning* telah meningkatkan hasil yang signifikan dalam akurasi dan waktu , serta memainkan peran yang penting dalam penelitian medis dalam 10 tahun terakhir [2]. Beberapa metode *Machine Learning* yang dapat digunakan adalah Regresi Logistik LASSO, *Random Forest*, dan *Suppot Vector Machine* (SVM). [18] menggunakan metode *Machine Learning* Regresi Logistik LASSO, *Random Forest*, dan XGBoost dengan teknik

resampling Random Oversampling, Random Undersampling, dan SMOTE menunjukkan bahwa Teknik *Resampling* tidak meningkatkan kinerja model prediksi yang dikembangkan pada data kesehatan observasional. Penelitian [20] menggunakan *hybrid* K-Means dengan *Support Vector Machine* (SVM) dengan penanganan data tidak seimbang dengan metode *Random Oversampling* dan menghasilkan bahwa *Random Oversampling* merupakan solusi yang efektif dan efisien untuk masalah ketidakseimbangan kelas dalam diagnosis kanker payudara. Tujuan dari penelitian ini adalah untuk melihat dampak *Random Oversampling, Random Undersampling*, dan SMOTE pada kinerja validasi internal dan eksternal dari model prediksi yang dikembangkan menggunakan data penyakit *Cardiovascular* (CVD). Pada penelitian digunakan tiga pendekatan klasifikasi yang berbeda, yaitu Regresi Logistik LASSO, *Random Forest*, dan *Support Vector Machine* (SVM) dengan penanganan ketidakseimbangan data menggunakan *Random Oversampling*, *Random Undersampling*, dan SMOTE.

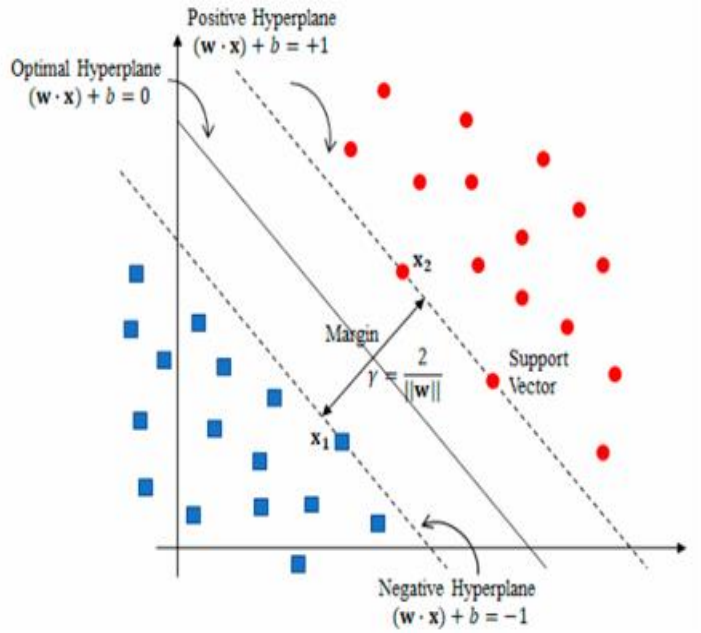
2. TINJAUAN PUSTAKA

2.1 Teknik *Resampling*

Terdapat dua teknik *resampling* data yakni *oversampling* dan *undersampling*. Pada penelitian ini, langkah pertama adalah tahap pengumpulan data *imbalanced* yang kemudian akan dibagi 80% untuk pemodelan atau pelatihan dan 20% untuk data pengujian. Kemudian, data pelatihan akan dilakukan teknik *oversampling* menggunakan metode *Random Oversampling* dan SMOTE, dan teknik *undersampling* menggunakan metode *Random Undersampling*. *Random undersampling* menangani data tidak seimbang dengan secara acak mengurangi jumlah sampel dari kelas mayoritas. *Random Oversampling* (ROS) bekerja dengan menduplikasi sampel data secara acak pada kelas minoritas dan menambahkannya ke dalam dataset pelatihan. ROS meningkatkan ukuran dataset pelatihan dengan mengulang sampel asli hingga distribusi kelas menjadi seimbang [10]. *Random Undersampling* menangani data tidak seimbang dengan secara acak meningkatkan jumlah sampel dari kelas minoritas dengan seacar acak mereplikasi sampel dari kelas minoritas. *Synthetic Minority Oversampling Technique* (SMOTE) menciptakan sampel baru dengan menginterpolasi sampel minoritas [17]. Mirip dengan ROS, SMOTE juga meningkatkan ukuran dan variasi dataset pelatihan dengan menghasilkan sampel buatan dalam dataset pelatihan melalui interpolasi antara titik data yang ada pada kelas minoritas yang berdekatan satu sama lain [5].

2.2 *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi [6]. Pendekatan SVM dilakukan dengan mencari sebuah *hyperplane* atau batas keputusan yang membedakan satu kelas dari kelas lainnya dicari. *Hyperplane* terbaik dapat ditemukan dengan mengukur lebar *hyperplane* (margin) dan menemukan titik maksimumnya.



Gambar 2.1. Ilustrasi dari Margin *Hyperplane*

Klasifikasi linear *hyperplane* SVM memiliki persamaan berikut:

$$f(x) = w^t x + b \tag{2.1}$$

dengan w^t adalah bobot vektor, x adalah peubah penjelas dan b adalah bias. Sehingga diperoleh persamaan berikut :

$$[\, w^t x + b \,] \geq 1 \text{ untuk } y_i = +1$$
$$[\, w^t x + b \,] \leq -1 \text{ untuk } y_i = -1 \tag{2.2}$$

2.3 *Random Forest*

Random forest dibangun dengan metode *bagging* dengan atribut acak yang merupakan pengembangan metode *Decision Tree*. Metode ini merupakan salah satu algoritma pembelajaran mesin dengan kinerja terbaik yang dikembangkan hingga ukuran pohon yang paling besar dan tidak akan dipangkas menggunakan Metode CART (*Classification and Regresion Tree*). Kemudian terbentuk kumpulan pohon yang kemudian disebut hutan (*forest*) [9]. *Random Forest* adalah pengklasifikasi yang terdiri dari sekumpulan pohon klasifikasi {h(x, S_b), b = 1, … , B} dengan {S_b}yang tidak berhubungan satu sama lain dan memiliki distribusi vektor acak yang sama. Untuk pohon yang masuk ke dalam kelas terbaik, voting x diberikan [19]. Proses membentuk algoritma random forest sebagai berikut: [8]

1. Sepertiga dari data sampel tidak digunakan ketika sampel *bootstrap* dibentuk dengan penggantian untuk setiap pohon keputusan.
2. Sampel yang tidak digunakan ini disebut sebagai OOB *(Out of Bag).*
3. Setiap pohon keputusan dalam hutan memiliki data OOB sendiri, yang digunakan untuk menghitung kesalahan masing-masing pohon keputusan. Estimasi ini disebut kesalahan OOB. Selain itu, *random forest* memiliki kemampuan untuk menghitung tingkat kepentingan dari masing-masing variabel dan prediksi yang dihasilkannya. Data yang hilang dan outlier diganti dengan prediksi yang dihasilkan.

2.4 Regresi Logistik LASSO

Regresi logistik digunakan untuk mengetahui hubungan antara peubah respon Y yang bersifat biner dengan satu atau lebih peubah penjelas. Peubah respon dalam regresi logistik biner memiliki dua kemungkinan yakni y = 1 menunjukkan kemungkinan sukses, y = 0 menunjukkan kemungkinan gagal [1]. Estimasi parameter pada regresi logistik diperoleh dengan memaksimalkan fungsi log-likelihood sebagai berikut : [14]

$$l(\beta) = \sum_{i=1}^{n}[y_i\, log(\pi_i) + (1-y_i)\log(1-\pi_i)] = \sum_{i=1}^{n}\left[y_i\, log\left(\frac{\pi_i}{1-\pi_i}\right) + log(1-\pi_i)\right] = \sum_{i=1}^{n}\left[y_i x_i \beta + log\left(1+e^{x_i\beta}\right)\right] \quad (2.3)$$

Untuk mengubah model regresi logistik diatas menjadi model regresi logistik LASSO, batasan L_1 diterapkapkan pada parameter β dengan meminimalkan fungsi kemungkinan log negatif sebagai berikut : [7]

$$\hat{\beta}_{LASSO} = \min_{\beta}\left(\sum_{i=1}^{n}\left[log\left(1+e^{x_i\beta}\right) - y_i(\beta_0 + x_i\beta)\right] + \lambda\sum_{j=1}^{p}|\beta_j|\right) \quad (2.4)$$

Dengan syarat $\sum_{j=1}^{p}|\beta_j| \leq \lambda$ dan $\lambda > 0$. Nilai λ dipilih menggunakan cross-validation pada fungsi "cv.glmnet" menggunakan paket "glmnet" di R untuk mendapatkan estimator logistik LASSO [11].

2.5 Evaluasi Kinerja Model

Evaluasi kinerja model dilakukan melalui beberapa metrik yang dihasilka oleh *Confussion matrix*. *Confussion matrix* digunakan untuk menilai kinerja model klasifikasi. Berikut merupakan ilustrasi *Confussion matrix:*

Tabel 2.1. Ilustrasi Confusion matrix

Nilai Prediksi	**Nilai Aktual**	
	Positif (1)	**Negatif (0)**
Positif (1)	True Positive (TP)	False Positive (FP)
Negatif (0)	False Negative (FN)	True Negative (TN)

True Positive (TP) menjelaskan jumlah prediksi benar untuk kelas positif, *True Negative* (TN) menjelaskan mengenai jumlah prediksi benar untuk kelas negatif, *False Positive* (FP) menjelaskan terkait jumlah prediksi salah untuk kelas positif, dan *False Negative* (FN) menjelaskan jumlah prediksi salah untuk kelas negatif. Berdasarkan *Confussion Matrix* yang diperoleh, dihitung Akurasi, Sensitivitas, Spesivitas, dan AUC.

Akurasi merupakan proporsi prediksi yang benar (positif dan negatif) terhadap total prediksi. Perhitungan akurasi sebagai berikut :

$$Akurasi = \frac{TN+TP}{TN+TP+FN+FP}\ x\ 100\% \quad (2.5)$$

Sensitivitas atau yang biasa disebut *Recall* merupakan proporsi dari prediksi positif yang benar terhadap semua kasus positif yang sebenarnya. Perhitungan Sensitivitas sebagai berikut :

$$Sensitivitas = \frac{TP}{TP+FN} \quad (2.6)$$

Spesifitas mengukur proporsi kasus negatif yang sebenarnya yang diidentifikasi dengan benar oleh model. Perhitungan Spesivitas sebagai berikut :

$$Spesifitas = \frac{TN}{TN+FP} \quad (2.7)$$

Kurva ROC digunakan untuk memvalidasi kesesuaian model dan kinerja prediksi, dengan kurva dua dimensi yang dihasilkan menggunakan sumbu X untuk sensitivitas berdasarkan tingkat FP dan sumbu Y untuk 1-spesifitas berdasarkan tingkat TP [3]. Area di bawah kurva ROC (AUC), yang

paling umum digunakan dan berkisar antara 0.5 hingga 1, adalah ukuran akurasi dan keandalan model dalam memprediksi suatu kejadia, dengan nilai yang mendekati satu menunjukkan akurasi yang lebih tinggi dan nilai yang mendekati 0.5 menunjukkan ketidakakuratan [4]. AUC dihitung menggunakan persamaan berikut:

$$AUC\ ROC = \frac{\sum TP + \sum TN}{TN + TP + FN + FP} \tag{2.8}$$

3. METODOLOGI PENELITIAN

3.1 Data

Penelitian ini menggunakan data yang bersumber dari UCI Machine Learning yaitu Heart Disease yang terdiri dari 4 database: Cleveland, Hungary, Switzerland, dan VA Long Beach yang diunduh pada tanggal 6 Maret 2024. Data dapat diakses melalui https://archive.ics.uci.edu/dataset/45/heart+disease. Total data dalam penelitian ini adalah 918 dataset dengan 12 peubah yang digunakan sebagai berikut.

Table 3.1. Peubah yang digunakan

No	Peubah	Keterangan
1	Umur	Umur Pasien dalam tahun
2	Jenis Kelamin	Jenis kelamin pasien
3	Tipe Nyeri Dada	Jenis nyeri dada yang dirasakan pasien
4	Tekanan Darah Istirahat	Tekanan darah pasien dalam keadaan istirahat
5	Kolesterol	Kadar kolesterol serum pasien
6	Gula darah Puasa	Kadar gula darah puasa pasien
7	Elektrokardiogram Istirahat	Hasil elektrokardiogram (EKG) pasien dalam keadaan istirahat
8	Denyut Jantung Maksimum	Denyut jantung maksimum yang dicapai pasien.
9	Angina Akibat Latihan Fisik	Angina yang diinduksi oleh latihan fisik
10	Depresi ST Lama	Jumlah penurunan segmen ST yang ditemukan pada elektrokardiogram (EKG) pasien selama latihan.
11	Kemiringan ST Segment Puncak Latihan	Kemiringan segmen ST puncak selama latihan
12	Penyakit Cardiovascular (CVD)	Indikasi pasien yang memiliki penyakit dan tidak memiliki penyakit Cardiovascular (CVD)

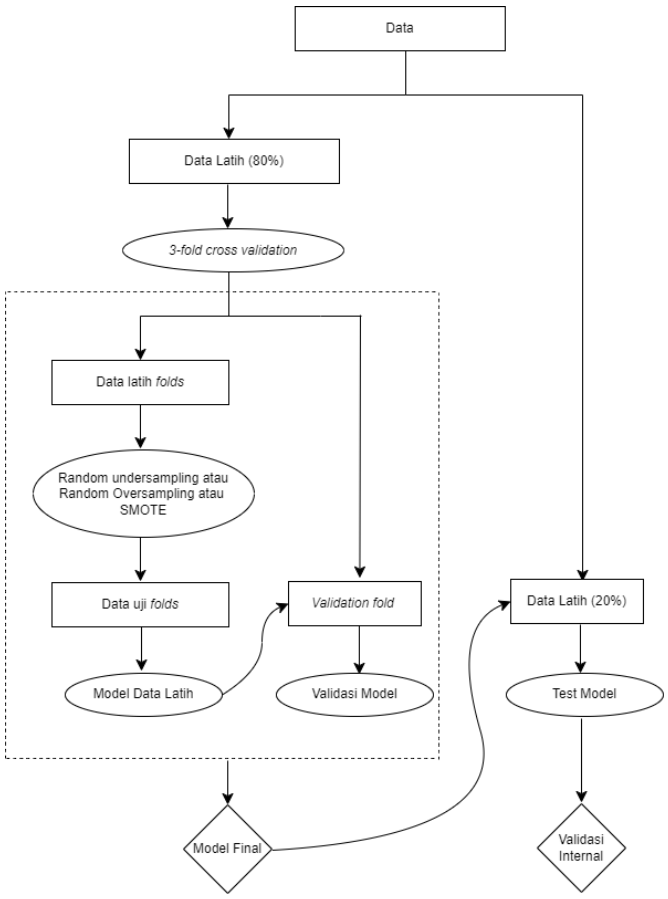
3.2 Prosedur Analisis Data

Tahapan analisis yang dilakukan pada penelitian ini sebagai berikut:

1. Melakukan ekplorasi data serta menghitung nilai korelasi antar peubah
2. Terdapat empat skema yang diberikan pada data, yaitu tanpa tahap data *pre-processing* dan melalui tahap *pre-processing*. Data *pre-processing* (1) menerapkan *random undersampling*, Data *pre-processing* (2) menerapkan *random oversampling*, dan Data *pre-processing* (3) menerapkan SMOTE. Ketiga skema tersebut menghasilkan data yang lebih seimbang.
3. Dataset Pertama (Tanpa *pre-processing):*
   a. Membagi data latih (80persen) dan data uji (20persen) dan 3-*Fold Cross Validation*

b. Membentuk Model klasifikasi dengan regresi logistik LASSO, *random forest*, dan *Support Vector Machine* (SVM)
c. Melakukan prediksi data latih dan data uji menggunakan model yang telah terbentuk
d. Evaluasi kinerja model menggunakan akurasi, sensitifitas, spesitifitas, dan AUC.

4. Dataset Kedua (Dengan *pre-processing*):
   a. Membagi data latih (80persen) dan data uji (20persen) dan 3-*Fold Cross Validation*
   b. Melakukan balancing pada data latih menggunakan algoritma *random undersampling, random oversampling,* dan SMOTE sehingga jumlah kelas untuk Pasien yang mempunyai penyakit *Cardiovascular* (CVD) dan tidak mempunyai penyakit *Cardiovascular* (CVD) memiliki jumlah kelas yang sama
   c. Membentuk Model klasifikasi dengan regresi logistik LASSO, *random forest*, dan *Support Vector Machine* (SVM)
   d. Melakukan prediksi data latih dan data uji menggunakan model yang telah terbentuk
   e. Evaluasi kinerja model menggunakan akurasi, sensitifitas, spesitifitas, dan AUC.
5. Tahap akhir penelitian membandingkan kinerja metode regresi logistik LASSO, *random forest*, dan *Support Vector Machine* (SVM) pada data tanpa *pre-processing* dan data yang diproses terlebih dahulu menggunakan *random undersampling, random oversampling,* dan SMOTE.

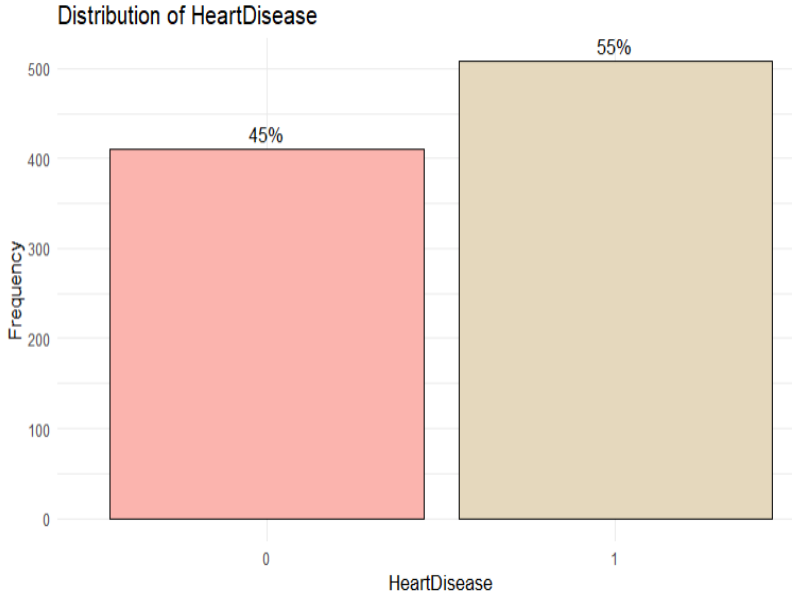


Gambar 3.1. Diagram Alir

4. HASIL DAN PEMBAHASAN

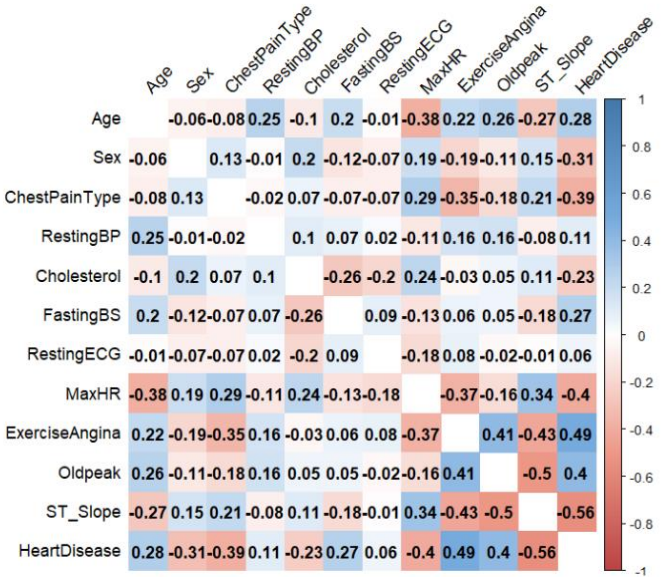
4.1 Eksplorasi Data

Data penelitian ini merupakan data penyakit *Cardiovascular* (CVD) yang terdiri dan 11 peubah lainnya. Setiap peubah memiliki hubungan dengan penyakit *Cardiovascular* (CVD) yang merupakan peubah respon. Data diperiksa agar dapat diterapkan dengan efisiensi pada algoritma pembelajaran mesin. Peubah respon (Penyakit *Cardiovascular* (CVD)) merupakan peubah biner yang mengandung dua kelas (1,0), 1 jika Pasien memiliki penyakit *Cardiovascular* (CVD) dan 0 jika Pasien tidak memiliki penyakit *Cardiovascular* (CVD).



Gambar 4.1. Histogram dari distribusi peubah Penyakit *Cardiovascular* (CVD)

Gambar 4.1. menunjukkan peubah respon tidak seimbang, dengan kelas 0 memiliki 45% observasi dan kelas 1 memiliki 55% observasi. Algoritma *Machine Learning* tidak bekerja secara efisien untuk data yang tidak seimbang. Oleh karena itu, diterapkan penanganan kelas data tidak seimbang dengan *Random Undersampling*, *Random Oversampling*, dan SMOTE. Data menggunakan fungsi package ROSE dari program R. Dalam keberadaan kelas yang tidak seimbang, paket ini menawarkan fungsi untuk menangani masalah klasifikasi biner. Menurut ROSE, sampel seimbang dihasilkan dengan *oversampling* acak dari contoh minoritas, undersampling dari contoh mayoritas, atau kombinasi *oversampling* dan *undersampling* [12].



Gambar 4.2. Korelasi Antar Peubah

Gambar 4.2. menjelaskan matriks korelasi yang ditampilkan, beberapa faktor yang memiliki pengaruh signifikan terhadap penyakit *Cardiovascular* (CVD). Hasil korelasi menunjukkan beberapa hubungan penting antara variabel dan penyakit *Cardiovascular* (CVD). Usia memiliki korelasi positif lemah (0.28), menunjukkan bahwa semakin tua seseorang, semakin tinggi kemungkinan menderita penyakit *Cardiovascular* (CVD). Jenis kelamin (-0.31) dan jenis nyeri dada (-0.39) memiliki korelasi negatif lemah, mengindikasikan beberapa perbedaan dalam kemungkinan menderita penyakit *Cardiovascular*(CVD) berdasarkan faktor-faktor tersebut. Tekanan darah saat istirahat (0.11), kolesterol (-0.23), dan gula darah puasa (0.27) juga menunjukkan korelasi lemah. Tekanan Darah Istirahat hampir tidak berhubungan dengan penyakit *Cardiovascular* (CVD) (0.06). Detak jantung maksimum memiliki korelasi negatif sedang (-0.40), menunjukkan detak jantung maksimum yang lebih tinggi berhubungan dengan penurunan risiko penyakit *Cardiovascular* (CVD). Angina Akibat Latihan Fisik (0.49) dan nilai Oldpeak (0.4) menunjukkan korelasi positif yang lebih kuat, sementara kemiringan segmen ST memiliki korelasi negatif paling kuat (-0.56), menunjukkan bahwa kemiringan ST yang lebih rendah berhubungan dengan peningkatan kemungkinan penyakit *Cardiovascular* (CVD). Ini menunjukkan bahwa beberapa faktor seperti Angina Akibat Latihan Fisik, Oldpeak, dan kemiringan ST memiliki korelasi yang lebih kuat dengan penyakit *Cardiovascular* (CVD) dibandingkan faktor lainnya.

4.2 Klasifikasi Tanpa *Pre-Processing Data*

Metode klasifikasi yang diterapkan dalam penelitian ini adalah metode random forest, regresi logistik LASSO, dan *Support Vector Machine* (SVM). Hasil dari ketiga pendekatan akan dibandingkan dengan berbagai perlakuan untuk menangani data. Penelitian ini menggunakan metode *cross validation* untuk membagi data menjadi data latih (80persen) dan data uji (20persen). Metode cross validation digunakan untuk mengurangi bias yang terkait dengan pengambilan sampel acak dengan memilih fold sebanyak 3 sehingga data penyakit *Cardiovascular* (CVD) dibagi menjadi 3 bagian yang ukurannya yang hampir sama, sehingga data Penyakit *Cardiovascular* (CVD) akan pernah menjadi data latih dan data validasi.

Langkah pertama dalam metode klasifikasi *random forest* adalah melakukan pengambilan sampel acak dari sebelas peubah bebas. Selanjutnya dibangun sebanyak 10 pohon keputusan. Proses

ini dilakukan dengan menggunakan aplikasi R melalui proses *looping* untuk mendapatkan pohon keputusan. Hasilnya dapat dilihat pada Tabel 4.1. berupa nilai akurasi, sensitivitas, spesifisitas, dan AUC pada data penyakit Cardiovascular (CVD). Langkah pertama dalam metode klasifikasi regresi logistik LASSO adalah mengestimasi parameter β menggunakan metode maksimum likelihood dengan penalti L1 (LASSO). Setelah itu, dilakukan klasifikasi menggunakan parameter β yang telah diestimasi dengan bantuan aplikasi R. Nilai akurasi, sensitifitas, spesifitas, dan AUC yang dihasilkan menunjukkan seberapa baik model regresi logistik LASSO bekerja. Proses pelatiahan dengan model *Support Vector Machine* (SVM) menggunakan kernel linier. Parameter model dituning secara otomatis untuk menemukan kombinasi parameter terbaik berdasarkan metrik AUC. Skema validasi silang digunakan untuk mengevaluasi kinerja model secara lebih akurat dan menghindari overfitting. Dari model ini, didapatkan nilai akurasi, sensitivitas, spesifisitas, dan AUC yang dapat digunakan untuk mengevaluasi performa model *Support Vector Machine* (SVM).

Tabel 4.1. Perbandingan metode klasifikasi tanpa *pre-processing* data

No	Metode	Akurasi (%)	Sensitifitas (%)	Spesifisitas (%)	AUC (%)
1	Random Forest	87.41%	88.78%	86.05%	53%
2	Regresi Logistik LASSO	87.34%	89.80%	84.88%	50.30%
3	SVM	86.83%	88.78%	84.88%	50.10%

Berdasarkan Tabel 4.1. dapat dilihat metode klasifikasi *random forest* tanpa *pre-processing* data menunjukkan nilai akurasi sebesar 87.41% dengan nilai AUC sebesar 53%. Metode klasifikasi ini dapat memprediksi Penyakit *Cardiovascular* dengan sensitifitas sebesar 88.78% dan spesifisitas sebesar 86.05%. Sementara metode klasifikasi regresi logistik LASSO menunjukkan nilai akurasi sebesar 87.34% dengan nilai AUC sebesar 50.30%. Metode regresi logistik LASSO memiliki sensitifitas sebesar 89.80%, yang lebih tinggi daripada random forest, tetapi spesifisitasnya sedikit lebih rendah, yaitu 84.88%. Metode SVM menghasilkan nilai akurasi sebesar 86.83%, sensitifitas sebesar 88.78%, dan spesifisitas sebesar 84.88% dengan nilai AUC sebesar 50.10%. Sensitifitas dan spesifisitas SVM serupa dengan regresi logistik LASSO, namun akurasinya sedikit lebih rendah dan AUC-nya juga rendah, menunjukkan bahwa SVM mungkin tidak cukup baik dalam memisahkan kelas-kelas. Berdasarkan nilai sensitifitas, dapat disimpulkan bahwa metode regresi logistik LASSO sedikit lebih baik dalam mengidentifikasi kasus Penyakit *Cardiovascular* (sensitifitas lebih tinggi), namun metode *random forest* memiliki spesifisitas yang sedikit lebih baik. Kedua metode memiliki nilai AUC yang cukup rendah, terutama regresi logistik LASSO, yang menunjukkan bahwa model tersebut mungkin tidak terlalu baik dalam memisahkan kelas-kelas dengan baik. Oleh karena itu, penelitian lebih lanjut dengan berbagai teknik pre-processing data mungkin diperlukan untuk meningkatkan performa model secara keseluruhan. Selain itu, metode SVM juga membutuhkan perhatian karena performanya yang serupa dengan regresi logistik LASSO namun dengan akurasi dan AUC yang sedikit lebih rendah.

4.3 Klasifikasi dengan Teknik *Pre-Processing Data*

Pre-processing data untuk meningkatkan sensitifitas pada setiap model klasifikasi. Peneliti menggunakan teknik *resampling random undersampling, random oversampling*, dan SMOTE untuk mengatasi kelas yang tidak seimbang karena data yang digunakan terdiri dari 508 pasien—410 yang tidak memiliki penyakit *Cardiovascular* (CVD) dan 508 yang memiliki penyakit *Cardiovascular* (CVD). Tabel 4.2. menunjukkan perbandingan metode klasifikasi berdasarkan nilai akurasi, sensitifitas, spesifisitas, dan nilai AUC dari model klasifikasi yang digunakan dengan teknik *resampling*.

Tabel 4.2. Perbandingan metode klasifikasi dengan *pre-processing* data

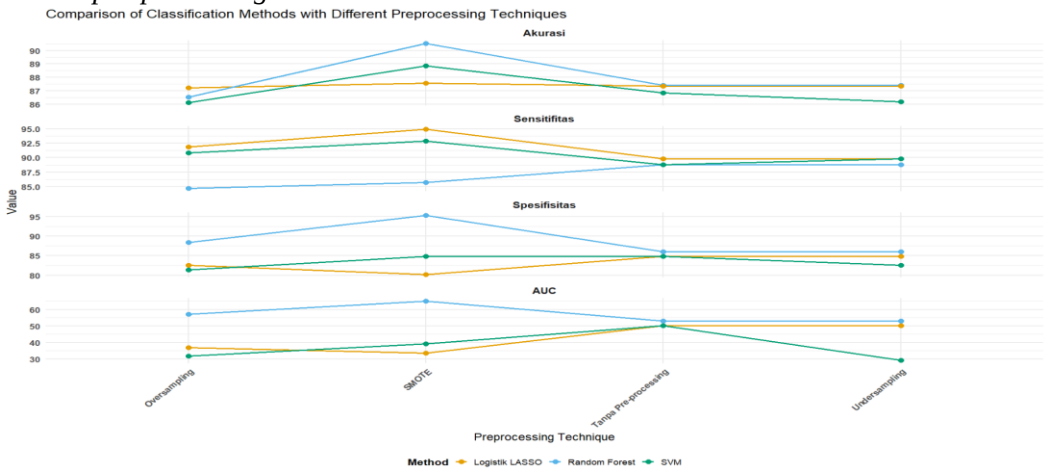
No	Metode	Akurasi (%)	Sensitifitas (%)	Spesifisitas (%)	AUC (%)
1	*Random Forest Undersampling*	87.41%	88.78%	86.05%	53.00%
2	Regresi Logistik LASSO *Undersampling*	87.34%	89.80%	84.88%	50.30%
3	SVM *Undersampling*	86.18%	89.80%	82.56%	29.10%
4	*Random Forest Oversampling*	86.53%	84.69%	88.37%	57.10%
5	Regresi Logistik LASSO *Oversampling*	87.20%	91.84%	82.56%	36.80%
6	SVM *Oversampling*	86.11%	90.82%	81.40%	31.60%
7	*Random Forest* SMOTE	90.53%	85.71%	95.35%	65.10%
8	Regresi Logistik LASSO SMOTE	87.57%	94.90%	80.23%	33.50%
9	SVM SMOTE	88.87%	92.86%	84.88%	39.10%

Berdasarkan Tabel 4.2. dapat disimpulkan bahwa metode klasifikasi random forest dengan random undersampling menghasilkan akurasi 87.41% dengan AUC sebesar 53.00%, dapat memprediksi penyakit *Cardiovascular* (CVD) dengan sensitifitas 88.78% dan spesifisitas 86.05%. Metode regresi logistik LASSO dengan *random undersampling* menghasilkan akurasi 87.34% dengan AUC sebesar 50.30%, memiliki sensitifitas 89.80% dan spesifisitas 84.88%. *Random forest* dengan *random oversampling* menghasilkan akurasi 86.53%, dengan AUC sebesar 57.10%, sensitifitas 84.69% dan spesifisitas 88.37%. Sedangkan regresi logistik LASSO dengan *random oversampling* menghasilkan akurasi 87.20%, AUC 36.80%, sensitifitas 91.84% dan spesifisitas 82.56%. Metode *random forest* dengan SMOTE menunjukkan hasil terbaik dengan akurasi 90.53%, AUC 65.10%, sensitifitas 85.71% dan spesifisitas 95.35%. Metode regresi logistik LASSO dengan SMOTE menghasilkan akurasi 87.57%, AUC 33.50%, sensitifitas tertinggi 94.90%, dan spesifisitas 80.23%. Metode SVM dengan *random undersampling* menghasilkan akurasi 86.18%, sensitifitas 89.80%, spesifisitas 82.56%, dan AUC 29.10%. SVM dengan *random oversampling* menghasilkan akurasi 86.11%, sensitifitas 90.82%, spesifisitas 81.40%, dan AUC 31.60%. SVM dengan SMOTE menghasilkan akurasi 88.87%, sensitifitas 92.86%, spesifisitas 84.88%, dan AUC 39.10%.

Berdasarkan tabel tersebut, metode *random forest* dengan SMOTE menghasilkan performa terbaik dengan akurasi dan spesifisitas tertinggi serta AUC yang lebih baik (65.10%), menunjukkan kemampuan prediksi yang kuat baik untuk mendeteksi adanya penyakit *Cardiovascular* (CVD) maupun tidak. Namun, regresi logistik LASSO dengan SMOTE memiliki sensitifitas tertinggi (94.90%), yang menunjukkan kemampuan sangat baik dalam mendeteksi semua kasus penyakit *Cardiovascular* (CVD). Meskipun demikian, metode random forest dengan SMOTE tetap menjadi pilihan terbaik secara keseluruhan karena keseimbangan antara akurasi, sensitifitas, dan spesifisitas yang tinggi. Sementara itu, metode SVM menunjukkan hasil yang lebih rendah secara keseluruhan dibandingkan dengan *random forest* dan regresi logistik LASSO, baik dengan *random undersampling*, *random oversampling*, maupun SMOTE. Meskipun SVM dengan SMOTE menunjukkan peningkatan sensitifitas, akurasi dan AUC-nya masih lebih rendah dibandingkan dengan metode lain, menunjukkan bahwa SVM mungkin memerlukan penyesuaian lebih lanjut atau metode pre-processing yang lebih baik untuk meningkatkan performanya.

4.4 Perbandingan Performa Model Klasifikasi

Perbandingan performa model klasifikasi SVM, *random forest*, dan regresi logistik Lasso tanpa teknik *pre-processing* dan dengan teknik *pre-processing*. Berikut grafik perbandingan nilai akurasi, sensitifitas, spesifisitas, dan AUC dari berbagai metode klasifikasi tanpa teknik *pre-processing* dan dengan teknik *pre-processing.*



Gambar 4.3. Perbandingan metode klasifikasi tanpa *pre-processing* dan dengan teknik *preprocessing* data

Berdasarkan hasil evaluasi yang dilakukan, terlihat perbedaan kinerja yang signifikan antara model yang menggunakan teknik *pre-processing* data dengan teknik tidak menggunakan *pre-processing.* Ketika tidak ada *pre-processing* data yang dilakukan, baik *random forest*, regresi logistik LASSO, maupun SVM menunjukkan kinerja yang cukup baik dalam memprediksi penyakit *Cardiovascular* (CVD). Meskipun demikian, nilai AUC untuk semua metode cenderung rendah, dengan *random forest* mencapai 53.00%, regresi logistik LASSO 50.30%, dan SVM 50.10%. Meskipun sensitivitas dan spesifisitasnya relatif tinggi, AUC yang rendah menunjukkan bahwa model-model tersebut mungkin tidak dapat memisahkan kelas-kelas dengan baik tanpa *pre-processing.* Ketika *pre-processing* data diterapkan, terlihat peningkatan yang signifikan dalam kinerja model *random forest* dengan SMOTE menunjukkan hasil terbaik dengan AUC mencapai 65.10%, diikuti oleh regresi logistik LASSO dengan SMOTE (AUC 33.50%) dan *random forest* dengan *random oversampling* (AUC 57.10%). Teknik SMOTE secara khusus menunjukkan peningkatan yang cukup signifikan dalam kinerja model, meningkatkan kemampuan untuk memisahkan kelas-kelas dengan lebih baik. Penerapan *pre-processing* data, terutama dengan menggunakan teknik seperti SMOTE, sangat penting dalam meningkatkan kinerja model dalam memprediksi penyakit *Cardiovascular* (CVD). Tanpa *pre-processing*, model cenderung memiliki keterbatasan dalam memisahkan kelas-kelas dengan baik, seperti yang tercermin dari nilai AUC yang rendah.

4.5 Pembahasan

Pembahasan ini menjelaskan perbedaan dan kesesuaian hasil-hasil yang diperoleh dari penelitian dengan hasil dari literatur sebelumnya. Penelitian ini menunjukkan bahwa penggunaan teknik SMOTE memberikan peningkatan yang signifikan dalam kinerja model prediksi penyakit *Cardiovascular* (CVD). Hasil ini sesuai dengan penelitian [5] yang menunjukkan bahwa SMOTE dapat meningkatkan kinerja model dengan ketidakseimbangan kelas. Namun, dalam penelitian [16], penggunaan teknik *resampling* tradisional seperti *Random Oversampling* dan *Random Undersampling* tidak memberikan peningkatan yang signifikan dalam diskriminasi model, yang

juga terlihat dalam penelitian ini dimana nilai AUC tidak meningkat secara signifikan dengan teknik tersebut.

Penelitian ini juga menemukan bahwa metode *random forest* dengan SMOTE memiliki performa terbaik, yang konsisten penelitian [17] dan [19] yang menunjukkan bahwa *random forest* efektif dalam menangani data yang tidak seimbang. Selain itu, penelitian ini mengungkapkan bahwa metode SVM menunjukkan performa yang lebih rendah secara keseluruhan dibandingkan dengan *random forest* dan regresi logistik LASSO, bahkan setelah menggunakan teknik SMOTE. Penelitian [18] menunjukkan bahwa kombinasi *random oversampling* dengan SVM dapat memberikan hasil yang baik, yang berbeda dengan hasil penelitian ini dimana SVM tidak menunjukkan performa yang optimal meskipun menggunakan SMOTE. Hasil ini sejalan dengan penelitian [8] yang menyatakan bahwa SVM mungkin memerlukan penyesuaian lebih lanjut atau metode pre-processing tambahan untuk mencapai performa optimal.

Secara keseluruhan, hasil-hasil ini memperkuat pentingnya memilih teknik *resampling* yang tepat dan metode klasifikasi yang sesuai untuk kelas data yang tidak seimbang. Penelitian ini memberikan kontribusi baru dengan menunjukkan bahwa SMOTE, khususnya dalam kombinasi dengan *random forest*, dapat memberikan hasil yang lebih baik dibandingkan dengan metode *resampling* lainnya. Ini menunjukkan bahwa teknik ini tidak hanya merupakan replikasi dari penelitian sebelumnya tetapi juga menawarkan wawasan baru dalam aplikasi prediksi penyakit *Cardiovascular* (CVD).

5. KESIMPULAN

Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa untuk memprediksi penyakit *Cardiovascular* (CVD), metode klasifikasi yang optimal adalah *random forest* dengan menggunakan teknik *pre-processing* data, terutama dengan menggunakan metode SMOTE. *Pre-processing* data, termasuk teknik SMOTE, terbukti meningkatkan kinerja model dalam memisahkan kelas-kelas dengan baik, seperti yang tercermin dari peningkatan signifikan dalam nilai AUC. Selain itu, metode SVM menunjukkan performa yang lebih rendah secara keseluruhan dibandingkan dengan random forest dan regresi logistik LASSO, bahkan setelah menggunakan teknik SMOTE. Hal ini sejalan dengan hasil penelitian yang menyatakan bahwa SVM mungkin memerlukan penyesuaian lebih lanjut atau metode pre-processing tambahan untuk mencapai performa optimal.

Secara keseluruhan, temuan ini menekankan pentingnya pemilihan teknik resampling yang tepat dan metode klasifikasi yang sesuai untuk data yang tidak seimbang. Penelitian ini memberikan wawasan baru dengan menunjukkan bahwa penggunaan SMOTE, terutama dalam kombinasi dengan random forest, dapat menghasilkan kinerja yang lebih baik dibandingkan dengan metode resampling lainnya. Ini menegaskan bahwa teknik ini tidak hanya mengulangi hasil penelitian sebelumnya tetapi juga memberikan pandangan baru dalam penerapan prediksi penyakit Cardiovascular (CVD). Oleh karena itu, untuk mendapatkan prediksi yang akurat, penting untuk melakukan pre-processing data, khususnya dengan menggunakan teknik SMOTE, sebelum menerapkan model random forest. Kesimpulan dari pembahasan ini menunjukkan bahwa teknik SMOTE memberikan keuntungan signifikan dalam meningkatkan kinerja model prediksi, khususnya dengan metode random forest, dan menunjukkan potensi yang lebih besar dibandingkan metode resampling lainnya.

REFERENSI

[1] Agresti, A., 2002. *Categorical Data Analysis Second Edition*. John Wiley & Sons Inc., New York.

[2] Alkhalaf, M., Yu, P., Shen, J., & Deng, C., 2022. A review of the application of machine learning in adult obesity studies. *Applied Computing and Intelligence*, 2(1), 32–48. https://doi.org/10.3934/aci.2022002

[3] Arabameri, A., Saha, S., Chen, W., Roy, J., Pradhan, B., & Bui, D. T. (2020). Flash flood susceptibility modelling using functional tree and hybrid ensemble techniques. *Journal of Hydrology*, 587, 125007. https://doi.org/10.1016/j.jhydrol.2020.125007

[4] Bammou, L., Kharchouf, M., Boughanem, H., Douik, A., & El Fazziki, A. (2024). Predictive models for gully erosion susceptibility using machine learning techniques. *Environmental Earth Sciences*, 83(5), 283. https://doi.org/10.1007/s12665-024-10023-4

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357. https://doi.org/10.1613/jair.953

[6] Cortes, C., & Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20(3), 273-297. https://doi.org/10.1007/BF00994018

[7] Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1). PMID: 20808728 https://doi.org/10.18637/jss.v033.i01

[8] Goel, E. & Abhilasha, E., 2017. Random Forest: A Review. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, 7(1), 251-257. https://doi.org/10.23956/ijarcsse.v7i1.006

[9] Han, J., Kamber, M., & Pei, J., 2012. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publisher.

[10] Indrawati, A., Subagyo, H., Sihombing, A., Wagiyah, & Afandi, S., 2020. Analyzing the impact of resampling method for imbalanced data text in Indonesian scientific articles categorization. *Jurnal Baca*, 41(2). https://doi.org/10.14203/j.baca.v41i2.563

[11] Kim, S. M., Kim, Y., Jeong, K., Jeong, H., & Kim, J., 2018. Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography. *Ultrasonography*, 37(1), 36-42. https://doi.org/10.14366/usg.17054

[12] Lunardon, N., Menardi, G., & Torelli, N., 2014. ROSE: A Package for Binary Imbalanced Learning. *R Journal*, 6, 79–89. https://doi.org/10.32614/RJ-2014-008

[13] Ma, Y., & He, H., 2013. *Imbalanced Learning: Foundations, Algorithms, and Application*s. John Wiley & Sons, Hoboken, NJ, USA.

[14] Pereira, J. M., Basto, M., & Ferreira da Silva, A., 2016. The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. *Procedia Economics and Finance*, 39, 634-641. https://doi.org/10.1016/S2212-5671(16)30292-2

[15] Roth, G. A., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., et al., 2018. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *Lancet*, 392(10159), 1736–88. https://doi.org/10.1016/S0140-6736(18)32203-7

[16] Van Goorbergh, R. vd M., Timmerman, D., & Van Calster, B., 2022. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *arXiv Preprint arXiv:220209101*. https://doi.org/10.48550/arXiv.2202.09101

[17] Wongvorachan, T., He, S., & Bulut, O., 2023. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*, 14, 54. https://doi.org/10.3390/info14010054

[18] Yang, C., Fridgeirsson, E. A., Kors, J. A., et al., 2024. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *J Big Data*, 11, 7. https://doi.org/10.1186/s40537-023-00857-7

[19] Zailani, A. U., & Hanun, N. L., 2020. Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera. *Infotech: Journal of Technology Information*, 6(1), 7-14. https://doi.org/10.37365/jti.v6i1.61

[20] Zhang, J., & Chen, L., 2019. Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. *Computer Assisted Surgery*, 24(sup2), 62–72. https://doi.org/10.1080/24699322.2019.1649074