Jurnal Lingkar Pembelajaran Inovatif

Volume 5 Nomor 12, Desember 2024

KLASIFIKASI KELAS PADA DATA TIDAK SEIMBANG DALAM DETEKSI MIKROKALSIFIKASI MENGGUNAKAN SMOTE-ENN DAN XGBOOST DENGAN OPTIMASI BAYESIAN

Cindy Novita Yolanda Panjaitan[1], Sutarman[2]
cindypanjaitan1310@gmail.com[1], sutarman@usu.ac.id[2]
Universitas Sumatera Utara

ABSTRAK

Data tidak seimbang merupakan masalah umum dalam klasifikasi. Klasifikasi kelas pada data tidak seimbang dapat ditangani dengan dua pendekatan, yaitu level data dan level algoritma. Level data digunakan untuk menyeimbangkan distribusi kelas. Level algoritma digunakan untuk memperbaiki algoritma klasifikasi. Penelitian ini bertujuan untuk menangani klasifikasi kelas pada data tidak seimbang dan mengetahui efektivitas optimasi hyperparameter. Pada penelitian ini dilakukan analisis klasifikasi deteksi mikrokalsifikasi yang mengalami ketidakseimbangan kelas sebesar 98%. Metode-metode yang digunakan dalam penelitian ini adalah SMOTE-ENN sebagai metode pada pendekatan level data, XGBoost sebagai metode pada pendekatan level algoritma, dan optimasi Bayesian sebagai metode optimasi hyperparameter. Evaluasi performa kinerja klasifikasi dilakukan dengan membandingkan XGBoost yang menggunakan nilai default dan XGBoost yang menggunakan optimasi Bayesian. Hasil penelitian menunjukkan bahwa metode SMOTE-ENN mampu menyeimbangkan distribusi kelas data yang sangat tidak seimbang. Metode XGBoost mampu membentuk model klasifikasi dengan nilai accuracy yang tinggi, namun cukup rendah pada nilai precision dan F-measure. Metode Optimasi Bayesian mampu meningkatkan performa kinerja klasifikasi secara signifikan, di mana berhasil meningkatkan nilai accuracy, specificity, precision, dan F-measure, tetapi mengalami penurunan pada nilai recall. Berdasarkan hasil analisis pada penelitian ini, diperoleh bahwa metode XGBoost dengan menggunakan optimasi Bayesian menghasilkan evaluasi performa kinerja yang lebih baik dibandingkan metode XGBoost dengan menggunakan nilai default.

Kata Kunci: Data Tidak Seimbang, Klasifikasi, Optimasi Bayesian, SMOTE-ENN, XGBoost.

PENDAHULUAN

Klasifikasi merupakan salah satu topik umum pada penelitian ilmiah. Klasifikasi adalah cara mendapatkan model atau fungsi yang mendeskripsikan dan membedakan data ke dalam beberapa kelas dengan label kelas yang tidak diketahui (Han et al., 2011). Klasifikasi dapat diterapkan dalam berbagai bidang dan ilmu pengetahuan. Pada bidang teknologi digunakan untuk pengelolaan basis data. Pada bidang ekonomi digunakan untuk analisis pasar dan perencanaan ekonomi. Pada bidang medis digunakan untuk deteksi dan pengobatan penyakit, salah satunya adalah deteksi mikrokalsifikasi.

Mikrokalsifikasi adalah endapan kalsium yang dianggap sebagai tanda awal kanker payudara. Seorang dokter radiologi dapat menggunakan klasifikasi dalam membedakan antara mikrokalsifikasi dan non-mikrokalsifikasi. Hal ini berperan penting dalam diagnosis dan penentuan pengobatan yang lebih efektif untuk meningkatkan harapan hidup pasien. Jaringan di sekitar payudara, variasi bentuk, orientasi, kecerahan, dan ukuran diameter membuat mikrokalsifikasi terkadang sulit dideteksi (Quintanilla et al., 2018). Hal ini mengakibatkan kelas data pada klasifikasi deteksi mikrokalsifikasi menjadi tidak seimbang.

Data tidak seimbang adalah keadaan di mana jumlah kelas negatif (mayoritas) jauh lebih besar dibandingkan jumlah kelas positif (minoritas) (Ali et al., 2015). Data tidak seimbang terjadi karena adanya overfitting, di mana model klasifikasi tidak dapat melakukan generalisasi dengan baik pada data baru. Hal ini menyebabkan model klasifikasi mengabaikan kelas minoritas dan klasifikasi menjadi salah.

Terdapat berbagai metode untuk mengklasifikasikan data, baik metode klasik ataupun metode modern. Metode klasik yang sering digunakan adalah analisis diskriminan dan regresi logistik. Metode modern yang sering digunakan adalah Support Vector Machine, Genetic Algorithm, Decision Tree, Neural Network, Naive Bayes, Learning Vector Quantization, Fuzzy Sets, dan Rough Sets (Samosir et al., 2015). Metode modern umumnya digunakan untuk klasifikasi kelas pada data tidak seimbang karena algoritmanya yang dapat disesuaikan dan dioptimalkan untuk berbagai jenis data dan permasalahan. Kelas pada data tidak seimbang dapat diatasi dengan dua pendekatan, yaitu level data dan level algoritma (Angeli et al., 2022).

Pendekatan level data menggunakan metode resampling untuk mengurangi ketidakseimbangan kelas data. Metode resampling terdiri dari undersampling, oversampling, dan hybrid. Hampir pada semua kasus, metode hybrid lebih unggul dibandingkan metode undersampling dan oversampling (Wang et al., 2021b).

Pendekatan level algoritma biasanya menggunakan ensemble learning untuk memperbaiki algoritma klasifikasi agar diperoleh model yang lebih baik. Boosting dan bagging merupakan dua ensemble learning yang paling populer. Boosting umumnya lebih baik dibandingkan bagging karena telah terbukti mampu meningkatkan kinerja klasifikasi di berbagai keadaan, salah satunya ketika data tidak seimbang (Seiffert et al., 2008).

Terdapat beberapa penelitian mengenai klasifikasi kelas pada data tidak seimbang dengan pendekatan level data dan level algoritma. Salah satu diantaranya dilakukan oleh (Wang et al., 2021a) yang membuat model klasifikasi efek samping gagal jantung dan menemukan faktor-faktor yang memengaruhi prediksi perkembangan gagal jantung. Dalam penelitian ini, SMOTE-ENN digunakan sebagai pendekatan level data. Pada level algoritma digunakan Regresi logistik, Support Vector Machine, K-Nearest Neighbor, Random Forest, dan XGBoost. Berdasarkan temuan penelitian, SMOTE-ENN dan XGBoost memiliki kinerja terbaik dibandingkan metode lainnya.

Salah satu penelitian mengenai SMOTE-ENN dilakukan oleh (Li dan Wu, 2022) yang membuat klasifikasi untuk memprediksi apakah pelanggan suatu bank di Portugis akan berlangganan deposito berjangka atau tidak. Dalam penelitian ini dilakukan perbandingan kinerja klasifikasi SMOTE-ENN XGBoost dengan Decision Tree, AdaBoost, XGBoost, dan

SMOTE-XGBoost. Berdasarkan temuan penelitian, SMOTE-ENN XGBoost memiliki kinerja terbaik dibandingkan metode lainnya.

Salah satu penelitian mengenai XGBoost dilakukan oleh (Noviandy et al., 2023) yang membuat klasifikasi deteksi penipuan kartu kredit menggunakan XGBoost dengan beberapa pendekatan level data, yaitu SMOTE, SMOTE-TOMEK, SMOTE-ENN, dan ADASYN. Berdasarkan temuan penelitian, pendekatan paling efektif adalah XGBoost dan SMOTE-ENN.

SMOTE-ENN dan XGBoost merupakan metode yang unggul dalam menangani klasifikasi kelas pada data tidak seimbang. Hal ini dikarenakan kemampuan SMOTE-ENN dalam menghapus sampel noise dan XGBoost dapat mengklasifikasikan data dengan efisien dan cepat. Namun ditemukan kelemahan, yaitu cenderung menghasilkan nilai accuracy yang tinggi, tetapi rendah pada nilai performa klasifikasi lain. Hal ini dikarenakan kompleksitas hyperparameter XGBoost. Kelemahan tersebut dapat diatasi dengan penyetelan hyperparameter. Penyetelan Hyperparameter merupakan proses penentuan nilai optimal dari satu kumpulan hyperparameter (Ghawi dan Pfeffer, 2019).

Terdapat beberapa penelitian yang membandingkan beberapa penyetelan hyperparameter. Penelitian yang dilakukan oleh (Xia et al., 2017) yang membandingkan Optimasi Bayesian, Grid Search, Manual Search, dan Random Search menemukan bahwa Optimasi Bayesian adalah yang terbaik. Penelitian yang dilakukan oleh (Hnin dan Jeenanunta, 2019) yang membandingkan Optimasi Bayesian dengan Particle Swarm Optimization dan Genetic Algorithm juga menemukan bahwa Optimasi Bayesian adalah yang paling unggul.

Optimasi Bayesian unggul dalam meningkatkan performa klasifikasi karena menggunakan pendekatan probabilistik dengan menyesuaikan hyperparameter terhadap hasil evaluasi sebelumnya sehingga diperoleh hyperparameter optimal. Tanpa kemampuan penyesuaian seperti optimasi Bayesian, metode penyetelan hyperparameter lain lambat dalam pencarian hyperparameter dan tidak optimal sehingga mengurangi performa kinerja model.

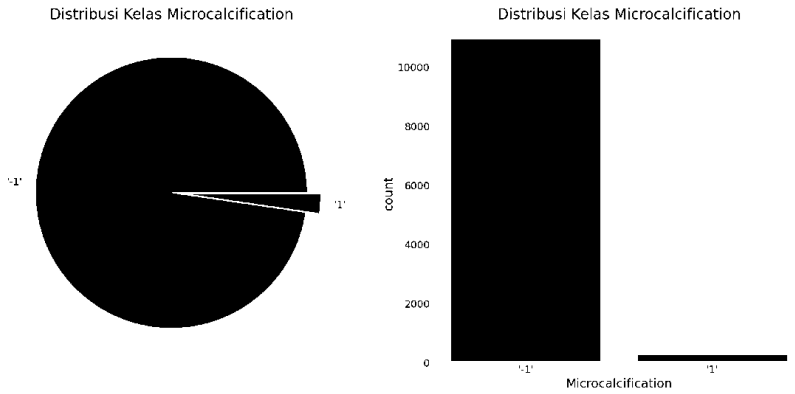
METODE

Jenis penelitian yang digunakan adalah studi literatur, yaitu melakukan studi kepustakaan untuk mencari referensi yang berkaitan dengan permasalahan dan pembahasan penelitian ini.

PEMBAHASAN

1. Statistik Deskriptif

Microcalcification classification dataset berisikan informasi sejumlah 11.183 sampel dari individu yang melakukan pendeteksian kanker payudara dengan enam variabel independen dan satu variabel dependen. Variabel dependen menggambarkan apakah pada suatu individu terdapat mikrokalsifikasi atau tidak. Variabel dependen bernilai “-1” menggambarkan bahwa pada suatu individu terdapat mikrokalsifikasi dan bernilai “1” menggambarkan bahwa pada suatu individu tidak terdapat mikrokalsifikasi *(non-microcalcification)*. Pada variabel dependen, kelas “-1” berjumlah 10.923 observasi (97,7%), sedangkan kelas “1” berjumlah 260 observasi (2,3%).



Gambar 1. Distribusi kelas variabel dependen

Data mikrokalsifikasi memiliki distribusi kelas yang tidak seimbang. Tingkat ketidakseimbangan data mikrokalsifikasi termasuk dalam kategori sedang karena persentase jumlah sampel kelas minoritas adalah 2% dari kumpulan data.

$$Kategori\ tingkat\ ketidakseimbangan = \frac{260}{11.183} \times 100 = 2{,}32\%$$

Statistik deskriptif menggambarkan informasi mengenai karakterisik data seperti jumlah observasi, mean, standar deviasi, nilai minimum, kuartil, dan nilai maksimum.

Tabel 1. Statistik deskriptif

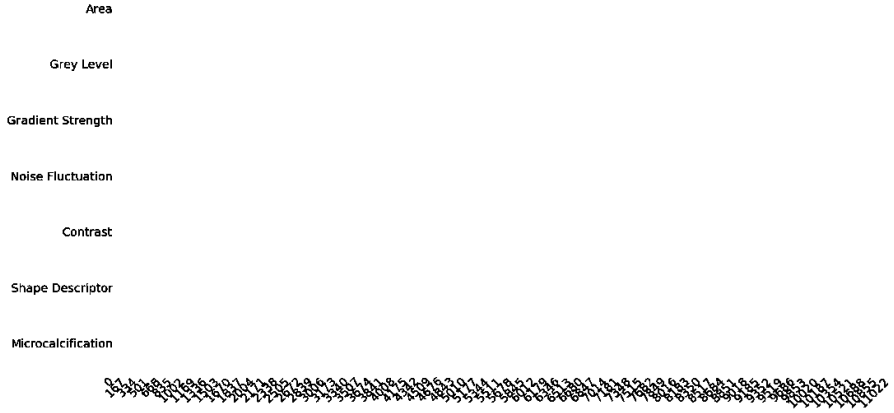
	count	*mean*	*std*	*min*	25%	50%	75%	*max*
Area	11.183	1,104e-10	1,0	-0,78	-0,74	-0,10	0,31	31,50
Grey Level	11.183	1,396e-09	1,0	-0,47	-0,47	-0,39	-0,07	5,08
Gradient Strength	11.183	5,619e-10	1,0	-0,59	-0,59	-0,23	0,21	29,47
Noise Fluctuation	11.183	-2,435e-09	1,0	-0,85	-0,85	-0,85	0,82	9,59
Contrast	11.183	-1,120e-09	1,0	-0,37	-0,37	-0,37	-0,37	23,61
Shape Descriptor	11.183	1,459e-09	1,0	-0,94	-0,94	-0,94	1,01	1,94

2. Pra-pemrosesan Data

Pra-pemrosesan data sangat penting untuk dilakukan sebelum membuat model klasifikasi. Tahapan ini dilakukan agar data yang akan diolah dapat menghasilkan hasil yang berkualitas. Pada penelitian ini, pra-pemrosesan data terdiri dari pengecekan *missing values*, pengubahan label kelas variabel dependen, partisi data, dan transformasi data.

a. Pengecekan Missing Values

Pengecekan *missing values* bertujuan untuk mengetahui kelengkapan data. Pada *microcalcification classification dataset* tidak ditemukan adanya *missing value.* Dengan fungsi *plot_missing_value* pada *library jcopml* diperoleh grafik sebagai berikut:



Gambar 2. Plot *missing values*

b. Pengubahan Label Kelas Variabel Dependen

Mengonversi variabel dependen menjadi format numerik agar dapat digunakan dalam algoritma klasifikasi. Pada penelitian ini digunakan fungsi *str.replace* yang mengubah label kelas variabel dependen berlabel "-1" dan "1" menjadi label baru dengan nilai biner 0 dan 1.

c. Partisi Data

Membagi data menjadi data *training* dan data *testing.* Data *training* digunakan untuk membangun model klasifikasi, sedangkan data *testing* digunakan untuk mengevaluasi model klasifikasi. Pada penelitian ini, proporsi pembagian *dataset* adalah 80% data *training* dan 20% data *testing*

Tabel 2. Jumlah proporsi data *training* dan data *testing*

Jumlah Data *Training*	Jumlah Data *Testing*	Total
8.946	2.237	11.183

d. Transformasi Data

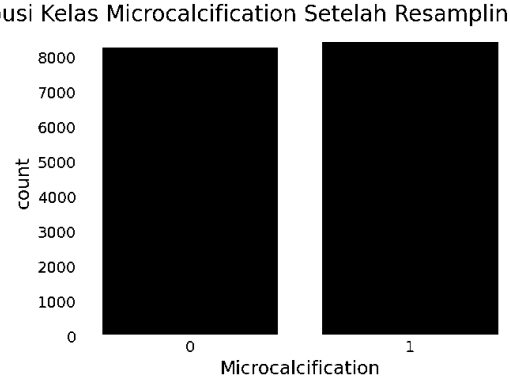
Memastikan data yang digunakan dalam analisis sesuai dengan format dari algoritma klasifikasi. Pada penelitian ini, transformasi data terdiri dari standardisasi dan normalisasi. Standardidasi bertujuan untuk mengubah data agar berdistribusi normal dengan mean 0 dan standar deviasi 1, sedangkan normalisasi bertujuan untuk membuat setiap variabel berada pada skala 0-1.

Tabel 3. Data setelah transformasi

	Area	*Grey Level*	*Gradient Strength*	*Noise Fluctuation*	*Contrast*	*Shape Descriptor*
1	-0,8036	-0,3800	-0,5856	-0,4690	-0,8574	-0,9473
2	-0,0995	-0,3800	0,6678	-0,4283	-0,8574	-0,9473
3	0,0911	-0,3800	-0,0932	-0,1752	0,6435	0,9345
			⋮			
11181	-0,1579	-0,3800	0,2649	-0,3515	-0,8574	-0,9473
11182	1,0198	-0,3800	-0,4961	-0,4464	-0,8574	-0,9473
11183	0,0013	-0,3800	0,8469	-0,2882	-0,8574	-0,9473

3. *Resampling* SMOTE-ENN

Resampling SMOTE-ENN dilakukan untuk menyeimbangkan distribusi kelas pada data tidak seimbang. SMOTE digunakan untuk menghasilkan sampel sintetis kelas minoritas dan ENN digunakan untuk menghapus sampel yang diduga *noise*. Penerapan SMOTE-ENN pada Python adalah menggunakan *package imblearn.combine.* Berikut ini hasil penerapan SMOTE-ENN pada *microcalcification classification dataset.*



Gambar 3. Disribusi kelas variabel dependen setelah SMOTE-ENN

Berdasarkan *bar plot* di atas dapat diketahui bahwa distribusi kelas mikrokalsifikasi setelah dilakukan resampling SMOTE-ENN menjadi lebih seimbang. Tahapan selanjutnya adalah membuat model klasifikasi menggunakan metode XGBoost dengan data yang dihasilkan oleh metode SMOTE-ENN.

4. Model Klasifikasi XGBoost

Model klasifikasi XGBoost dengan program Python menggunakan *package* XGBoost dalam *module* XGBClassifier. Berikut diperoleh plot pohon XGBoost dengan *default hyperparameter.*

Gambar 4. Plot pohon XGBoost

5. Penyetelan *Hyperparameter* Optimasi Bayesian

Penyetelan *hyperparameter* XGBoost dengan optimasi Bayesian diterapkan menggunakan *library* Optuna. Pencarian *hyperparameter* optimal XGBoost dilakukan dengan membuat fungsi objektif dalam ruang pencarian *hyperparameter* dan menjalankan proses optimasi. Pada penelitian ini, proses optimasi dilakukan sebanyak 50 kali.

a. Ruang Pencarian *Hyperparameter*

Terdapat delapan *hyperparameter* pada pencarian nilai *hyperparameter* optimal dalam penelitian ini. Langkah pertama untuk melakukan penyetelan *hyperparameter* adalah mendefinisikan ruang pencarian. Berikut ini penetapan ruang pencarian *hyperparameter* pada penelitan ini.

Tabel 4. Ruang pencarian penyetelan *hyperparameter* Optimasi Bayesian

Hyperparameter	Ruang Pencarian
max_depth	[3, 10]
learning_rate	[0,01; 0,3]
n_estimators	[100, 1000]
subsample	[0,5; 1,0]
colsample_bytree	[0,5; 1,0]
min_child_weight	[1, 10]
scale_pos_weight	[1, 10]
Gamma	[0; 0,5]

b. Proses Optimasi

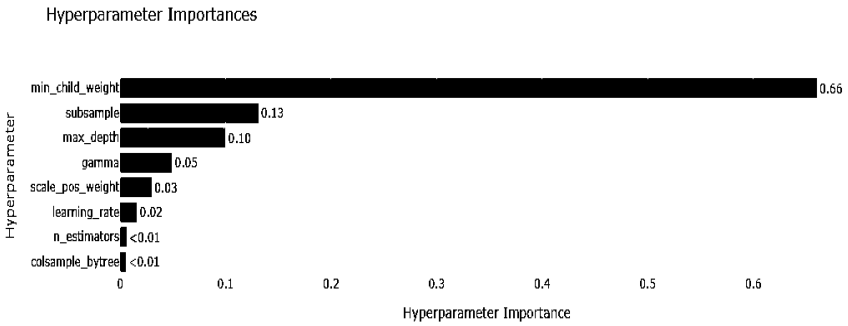
Proses optimasi pada penelitian ini dilakukan sebanyak 50 kali percobaan dengan menggunakan *library* Optuna dan parameter *default* diperoleh *hyperparameter* optimal dari model XGBoost sebagai berikut:

Tabel 5. *Hyperparameter* optimal Optimasi Bayesian

Hyperparameter	Ruang Pencarian
max_depth	6,000000
learning_rate	0,062846
n_estimators	838,000000
subsample	0,727315
colsample_bytree	0,956127
min_child_weight	1,000000
scale_pos_weight	3,000000
gamma	0,201671

c. *Hyperparameter Importances*

Memperhitungkan seberapa penting setiap *hyperparameter* dalam memengaruhi performa model klasifikasi. Berikut ini *hyperparameter importances* dari delapan *hyperparameter* yang digunakan pada penelitian ini.

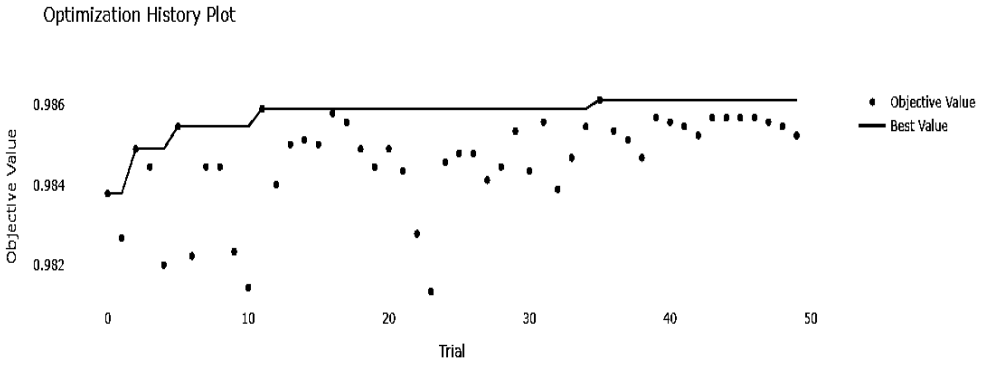


Gambar 5. *Hyperparameter importances*

Berdasarkan Gambar 5. terlihat bahwa *hyperparameter min_child_weight* adalah *hyperparameter* terpenting yang paling memengaruhi performa model agar optimal. *Hyperparameter* yang tidak terlalu memengaruhi performa model adalah *n_estimators* dan *colsample_bytree*, di mana kedua *hyperparameter* ini bernilai $< 0{,}01$.

d. Riwayat Optimasi *Hyperparameter*

Memvisualisasikan proses kinerja optimasi selama 50 kali percobaan yang dilakukan. Plot ini menggambarkan bagaimana fungsi objektif berubah seiring dilakukannya pencarian *hyperparameter* optimal.



Gambar 6. Plot Riwayat Optimasi

6. Performa Kinerja Klasifikasi

Penilaian performa kinerja klasifikasi diperoleh berdasarkan *confusion matrix* dari model yang telah dibentuk pada data *testing.* Berikut ini hasil *confusion matrix* untuk model klasifikasi, yaitu XGBoost dan XGBoost dengan Optimasi Bayesian.

Tabel 6. Hasil confusion matrix model XGBoost dengan nilai default dan XGBoost dengan optimasi Bayesian

Model	TP	FN	FP	TN
XGBoost dengan nilai *default*	2.123	64	7	43
XGBoost dengan Optimasi Bayesian	2.180	7	12	38

Berdasarkan tabel 7 terlihat bahwa nilai *True Positive* (TP) dan *False Positive* (FP) mengalami peningkatan, sedangkan nilai *False Negative (*FN) dan *True Negative* (TN) mengalami penurunan. Nilai TP, FN, FP, dan TN digunakan untuk menghasilkan *accuracy, specificity, recall, precision,* dan *F-measure.* Berikut ini hasil perhitungan performa kinerja klasifikasi dari model klasifikasi XGBoost dan XGBoost dengan Optimasi Bayesian.

Tabel 7. Nilai performa klasifikasi model XGBoost dengan nilai default dan XGBoost dengan Optimasi Bayesian

Model	*Accuracy*	*Specificity*	*Recall*	*Precision*	*F-measure*
XGBoost dengan nilai *default*	0,968	0,971	0,860	0,402	0,548
XGBoost dengan Optimasi Bayesian	0,992	0,995	0,760	0,844	0,800

Berdasarkan tabel 8 terlihat bahwa model XGBoost dengan Optimasi Bayesian unggul pada setiap penilaian kecuali pada nilai *recall.* Diperoleh nilai *accuracy* sebesar 0,992 menunjukkan bahwa model XGBoost dengan Optimasi Bayesian dapat memprediksi *microcalcification classification dataset* dengan benar sebesar 99,2%. Nilai *specificity* sebesar 0,995 menunjukkan bahwa model XGBoost dengan Optimasi Bayesian dapat memprediksi *non-microcalcification* dengan benar sebesar 99,5%. Nilai *recall* sebesar 0,760 menunjukkan bahwa model XGBoost dengan Optimasi Bayesian dapat memprediksi *microcalcification* dengan benar sebesar 76%. Nilai *precision* sebesar 0,844 menunjukkan bahwa model XGBoost dengan Optimasi Bayesian dapat memprediksi *microcalcification* secara benar sebesar 84,4% dari semua prediksi *microcalcification.* Nilai *F-measure* sebesar 0,800 menunjukkan keseimbangan prediksi antara *recall* dan *precision* sebesar 80%. Berdasarkan hasil performa klasifikasi, model XGBoost dengan Optimasi Bayesian terbukti mampu meningkatkan performa klasifikasi model XGBoost.

KESIMPULAN

Berdasarkan analisis dari microcalcification classification dataset yang berukuran besar dengan ketidakseimbangan kelas 98% membuktikan bahwa metode SMOTE-ENN mampu membentuk distribusi kelas menjadi seimbang. Hasil evaluasi penilaian kinerja memperlihatkan bahwa model klasifikasi XGBoost dengan optimasi Bayesian mengalami peningkatan pada setiap penilaian performa kinerja, kecuali pada nilai recall. Diperolehnya nilai accuracy 99,2%, specificity 99,5%, precision 84,4%, dan F-measure 80% menunjukkan bahwa model klasifikasi XGBoost dengan optimasi Bayesian menggunakan library Optuna secara signifikan lebih unggul dibandingkan XGBoost dengan nilai default.

DAFTAR PUSTAKA

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with Class Imbalance Problem: A Review. International Journal of Advances in Soft Computing and Its Applications, 7(3), 176–204.

Angeli, K. De, Gao, S., Danciu, I., Durbin, E. B., Wu, X. C., Stroup, A., Doherty, J., Schwartz, S., Wiggins, C., Damesyn, M., Coyle, L., Penberthy, L., Tourassi, G. D., & Yoon, H. J. (2022). Class Imbalance in out-of-distribution Datasets: Improving the Robustness of The TextCNN for The classification of Rare Cancer Types. Journal of Biomedical Informatics, 125, 1–11. https://doi.org/10.1016/j.jbi.2021.103957

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(June 2002), 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 19(6).

Dwinanda, M. W., Satyahadewi, N., & Andani, W. (2023). Classification of Student Graduation Status Using. 17(3), 1785–1794.

Ghawi, R., & Pfeffer, J. (2019). Efficient Hyperparameter Tuning with Grid Search for Text Categorization using kNN Approach with BM25 Similarity. Open Computer Science, 9(1), 160–180. https://doi.org/10.1515/comp-2019-0011

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from Class-Imbalanced Data: Review of Methods and Applications. Expert Systems with Applications, 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Han, J., Kamber, M., & Pei, J. (2011). Data Mining : Concepts and Techniques (Third Edit). Elsevier.

Hnin, S. W., & Jeenanunta, C. (2019). Bayesian Optimization in A Support Vector Regression Model for Short-Term Electricity Load Forecasting. Engineering and Applied Science Research, 46(3), 267–275. https://doi.org/10.14456/easr.2019.30

Hossin, M. b, & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification

Evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1–11. https://doi.org/10.5121/ijdkp.2015.5201

Islam, S. F. N., Sholahuddin, A., & Abdullah, A. S. (2021). Extreme Gradient Boosting (XGBoost) Method in Making Forecasting Application and Analysis of USD Exchange Rates Against Rupiah. Journal of Physics: Conference Series, 1722(1), 1–11. https://doi.org/10.1088/1742-6596/1722/1/012016

Jha, A., Dave, M., & Madan, S. (2019). Comparison of Binary Class and Multi-Class Classifier Using Different Data Mining Classification Techniques. SSRN Electronic Journal, 894–903. https://doi.org/10.2139/ssrn.3464211

Kumar, M. S., Srivastava, D. M., & Prakash, D. V. (2024). Advanced Hybrid Prediction Model: Optimizing Lightgbm, Xgboost, Lasso Regression, and Random Forest With Bayesian Optimization. Journal of Theoretical and Applied Information Technology, 15(9), 4103–4115. www.jatit.org

Li, Y., & Wu, Z. (2022). Prediction of Customers ' Subscription to Time Deposits Based on SMOTEENN-XGBoost Model Prediction of Customers ' Subscription to Time Deposits Based on SMOTEENN-XGBoost Model. AIS Electronic Library (AISeL), 632–642.

Lin, G. M., & Zeng, H. C. (2021). Electrocardiographic Machine Learning to Predict Mitral Valve Prolapse in Young Adults. IEEE Access, 9, 103132–103140. https://doi.org/10.1109/ACCESS.2021.3098039

Noviandy, T. R., Idroes, G. M., Maulana, A., Hardi, I., Ringga, E. S., & Idroes, R. (2023). Credit Card Fraud Detection for Contemporary Financial Management Using XGBoost-Driven Machine Learning and Data Augmentation Techniques. Indatu Journal of Management and Accounting, 1(1), 29–35. https://doi.org/10.60084/ijma.v1i1.78

Quintanilla-Domínguez, J., Ruiz-Pinales, J., Barrón-Adame, J. M., & Guzmán-Cabrera, R. (2018). Microcalcifications Detection Using Image Processing. Computacion y Sistemas, 22(1), 291–300. https://doi.org/10.13053/CyS-22-1-2560

Sá, J. A. S., Almeida, A. C., Rocha, B. R. P., Mota, M. A. S., Souza, J. R. S., & Dentel, L. M. (2016). Lightning Forecast Using Data Mining Techniques on Hourly Evolution of The Convective Available Potential Energy. Brazilian Society on Computational Intelligence (SBIC), March, 1–5. https://doi.org/10.21528/cbic2011-27.1

Samosir, R. O., Wilandari, Y., & Yasin, H. (2015). Perbandingan Metode Klasifikasi Regresi Logistik Biner dan Radial Basis Function Network pada Berat Bayi Lahir Rendah. Jurnal Gaussian, 4, 997–1005.

Sang, A. I., Sutoyo, E., & Darmawan, I. (2021). Analisis Data Mining untuk Klasifikasi Data Kualitas Udara DKI Jakarta Menggunakan Algoritma Decision Tree dan Support Vector Machine. E-Proceeding of Engineering, 8(5), 8954–8963.

Santoso, B., Wijayanto, H., Notodiputro, K. A., & Sartono, B. (2017). Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : A Review. IOP Conference Series: Earth and Environmental Science, 58, 1–8. https://doi.org/10.1088/1755-1315/5

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2008). Resampling or Reweighting: A Comparison of Boosting Implementations. Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 1, 445–451. https://doi.org/10.1109/ICTAI.2008.59

Shi, K., Shi, R., Fu, T., Lu, Z., & Zhang, J. (2024). applied sciences A Novel Identification Approach Using RFECV – Optuna – XGBoost for Assessing Surrounding Rock Grade of Tunnel Boring Machine Based on Tunneling Parameters. Applied Sciences, 14, 1–23.

Wang, K., Tian, J., Zheng, C., Yang, H., Ren, J., Li, C., Han, Q., & Zhang, Y. (2021a). Improving Risk Identification of Adverse Outcomes in Chronic Heart Failure Using Smote+Enn and Machine Learning. Risk Management and Healthcare Policy, 14(May), 2453–2463. https://doi.org/10.2147/RMHP.S310295

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021b). Review of Classification Methods on Unbalanced Data Sets. IEEE Access, 9, 64606–64628. https://doi.org/10.1109/ACCESS.2021.3074243

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man and Cybernetics, 2(3), 408–421.

https://doi.org/10.1109/TSMC.1972.4309137
Xia, Y., Liu, C., Li, Y. Y., & Liu, N. (2017). A Boosted Decision Tree Approach Using Bayesian Hyper-Parameter Optimization for Credit Scoring. Expert Systems with Applications, 78, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017
Yang, F., Wang, K., Sun, L., Zhai, M., Song, J., & Wang, H. (2022). A Hybrid Sampling Algorithm Combining Synthetic Minority Over-Sampling Technique and Edited Nearest Neighbor for Missed Abortion Diagnosis. BMC Medical Informatics and Decision Making, 22(1), 1–14. https://doi.org/10.1186/s12911-022-02075-2
Yap, B. W., Rani, K. A., Abd Rahman, H. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. Lecture Notes in Electrical Engineering, 285 LNEE, 13–22. https://doi.org/10.1007/978-981-4585-18-7_2.