Class Imbalanced Learning Menggunakan Algoritma *Synthetic Minority Over-sampling Technique – Nominal* (SMOTE-N) pada *Dataset Tuberculosis* Anak

Yulia Ery Kurniawati
Program Studi Informatika, Fakultas Industri Kreatif, Institut Teknologi dan Bisnis Kalbis
Jl. Pulomas Selatan Kav.22 , Jakarta Timur 13210, DKI Jakarta, Indonesia
Email: yulia.kurniawati@kalbis.ac.id

Masuk: 28 Agustus 2019; Direvisi: 24 Oktober 2019; Diterima: 29 Oktober 2019

Abstract. *Class Imbalanced Learning (CIL) is the learning process for data representation and information extraction with severe data distribution to develop effective decisions supporting the decision-making process. SMOTE-N is one of the data level approach in CIL using over-sampling method. It generates synthetic instances to balance its minority class. This research applied SMOTE-N on Children Tuberculosis Dataset that has class imbalance. Over-sampling method is chosen to avoid important information loss because the Childhood Tuberculosis Dataset has a small number of instances. The Naïve Bayes Classifier has been applied to the balance dataset to evaluate its model. The results show that SMOTE-N can improve CIL performance metrics.*
Keywords: *Class Imbalance Learning, Over-sampling, SMOTE-N, Naïve Bayes Classifier*

Abstrak. *Class Imbalance Learning (CIL) merupakan proses pembelajaran untuk representasi data dan ekstraksi informasi dengan distribusi data yang buruk untuk mendukung pembuatan keputusan yang efektif dalam proses pengambilan keputusan. SMOTE-N adalah salah satu pendekatan data-level dalam CIL mengunakan metode over-sampling. SMOTE-N menghasilkan instance sintesis untuk menyeimbangkan jumlah instance pada kelas minoritasnya. Penelitian ini mengaplikasikan SMOTE-N pada dataset Tuberculosis Anak (TB Anak) yang memiliki ketidakseimbangan kelas. Metode over-sampling dipilih untuk menghindari kehilangan informasi yang penting dikarenakan dataset TB Anak memiliki jumlah instance yang sedikit. Naïve Bayes Classifier digunakan untuk mengevaluasi model dari dataset yang sudah seimbang. Hasilnya menunjukkan bahwa SMOTE-N dapat meningkatkan kinerja pada CIL.*
Kata Kunci: *Class Imbalance Learning, Over-sampling, SMOTE-N, Naïve Bayes Classifier*

1. Pendahuluan

Tuberculosis atau TB adalah suatu penyakit infeksi yang menular yang disebabkan oleh bakteri *Mycobacterium tuberculosis*, yang dapat menyerang berbagai organ terutama paru-paru [1]. Menurut hasil tahunan WHO tahun 2018, TB merupakan salah satu dari sepuluh besar penyebab kematian dan menjadi penyebab utama agen infeksi tunggal (di atas HIV/ AIDS) pada tahun 2017 [2], [3]. Penyakit *Tuberculosis* yang tidak mendapatkan pengobatan yang tuntas dapat menimbulkan komplikasi berbahaya hingga kematian.

TB tidak hanya dapat di derita oleh orang dewasa, tetapi juga dapat diderita oleh anak-anak. TB anak merupakan salah satu penyebab kesakitan dan kematian paling sering pada anak [4]. Hal ini dikarenakan gejala TB anak tidak khas. Gejala yang umum pada TB Anak yaitu penurunan berat badan, lemah, letih, dan lesu. Batuk yang biasa menjadi gejala utama TB pada orang dewasa, jarang menjadi gejala utama pada TB anak sehingga penderita TB anak jarang terdeteksi sebagai TB anak. Akibat gejala yang tidak memiliki khas dan kesulitan untuk mendeteksi TB anak, Kementrian Kesehatan dan Ikatan Dokter Anak Indonesia (IDAI) membuat pendekatan diagnosis TB anak menggunakan sistem skoring untuk melakukan diagnosis TB anak, [4].

Sedikitnya kasus TB pada anak dan gejala yang tidak memiliki kekhasan memberikan tantangan tersendiri dalam penelitian *machine learning* dan *data mining*. Tantangan tersebut yaitu *imbalanced data* atau ketidakseimbangan *instance* pada setiap kelasnya. Ketidakseimbangan data terjadi pada *dataset* yang mempunyai rasio kelas/ kasus yang tidak seimbang antara kelas yang satu dengan kelas yang lain. *Instance* pada kelas minoritas dianggap sebagai *noisy data* atau data yang tidak berarti dan dieliminasi oleh *classifier* atau algoritma klasifikasi [5]. Hal ini mengakibatkan kerugian karena *machine learning* dalam *data mining* kesulitan mengklasifikasikan kelas minoritas atau kelas dengan jumlah *instance* yang kecil secara benar. Algoritma mengasumsiikan bahwa distribusi kelas yang diuji seimbang sehingga pada beberapa kasus akan salah dalam mengkasifikasikan hasil pada tiap kelas. Klasifikasi kelas minoritas akan menyebabkan *error* dikarenakan pada saat klasifikasi, *classifier* cenderung fokus pada kelas mayoritas dan mengabaikan kelas minoritas.

Pada penelitian ini, *dataset* yang digunakan adalah *dataset* TB anak yang digunakan oleh Sari [6]. *Dataset* tersebut memiliki ketidakseimbangan kelas dan keterbatasan jumlah data. Ada dua pendekatan dalam penanganan ketidakseimbangan kelas yaitu *data-level* dan pendekatan algoritma [7]. Pendekatan *data-level* yaitu menggunakan metode *sampling* data asli baik kelas minoritas (*over-sampling*) maupun kelas mayoritas (*under-sampling*). Sedangkan pendekatan algoritma yaitu dengan mendesain algoritma baru atau meningkatkan algoritma yang ada misalnya menggunakan *adaptive boosting* atau *bagging*. Pada penelitian ini pendekatan *data-level* yaitu *over-sampling* dipilih untuk menyelesaikan permasalahan dalam ketidaseimbangan kelas data TB anak tersebut. Pendekatan ini dipilih karena pendekatan *under-sampling* tidak cocok digunakan untuk data medis karena jumlah data medis yang biasanya terbatas. Dengan keterbatasan data ini, jika dilakukan *under-sampling* maka data akan kehilangan beberapa informasi penting untuk proses pengambilan keputusan oleh *machine-learning*. Algoritma *Synthetic Minority Over-sampling Tecnique* (SMOTE) yang diajukan oleh Chawla pada tahun 2002 [8] dipilih penanganan kelas minoritasnya. SMOTE dipilih karena SMOTE memberikan solusi untuk *overfitting* yang disebabkan oleh metode *over-sampling*. SMOTE memanfaatkan *nearest-neighbors* dan jumlah *over-sampling* yang bisa disesuaikan. Sedangkan pengujian dilakukan dengan menggunakan *Naïve Bayes Classifier* (NBC). NBC dipilih karena memiliki kecepatan dan ketepatan dengan menggunakan probabilitas bersyarat.

2. Tinjauan Pustaka

Class Imbalance Learning (CIL) merupakan pembelajaran dari data yang memiliki ketidakseimbangan kelas. Tidak ada perjanjian atau standar terkait *class imbalance* atau ketidakseimbangan kelas yang diperlukan untuk *dataset* yang dianggap benar-benar "*imbalanced*". *Dataset* dimana kelas yang paling umum kurang dari dua kali kelas yang sedikit hanya akan sedikit tidak seimbang, sedangkan *dataset* dengan *imbalance ratio* 10:1 akan tidak seimbang dan *dataset* dengan *imbalance ratio* 1000:1 sangat tidak seimbang [7]. *Dataset* yang tidak seimbang jika kategori klasifikasi tidak kurang lebih sama terwakili [8].

Dampak dari ketidakseimbangan dalam pembelajaran dan kemampuan dalam belajar kelas yang langka (*rare classes*). Ada dua aspek pendekatan dalam menangani *imbalance dataset* yaitu *data level* dan algoritma [7], [9], [10]. Pendekatan *data level* yaitu mengubah distribusi *dataset* dengan metode *sampling* yaitu *over-sampling* atau meningkatkan jumlah sampel pada kelas minoritas dan *under-sampling* atau mengurangi jumlah sampel dari kelas mayoritas, dan atau dengan kombinasi keduanya. *Over-sampling*, meningkatkan jumlah dari minoritas, tetapi menyebabkan *over-fitting* karena duplikasi data sedangkan *under-sampling*, mengurangi jumlah dari mayoritas, sehingga memungkinkan kehilangan informasi dari mayoritas dan menurunkan performa klasifikasi [11]. Sedangkan pendekatan algoritma yaitu dengan mendesign algoritma baru atau meningkatkan algoritma yang sudah ada. Teknik yang diklasifikasikan sebagai pendekatan *algorithm level* diantaranya adalah algoritma *adaptive boosting*, *bagging*, *cost-sensitive*, dan *active learning* [12].

Synthetic Minority-Over Sampling Technique (SMOTE) merupakan pendekatan untuk menangani ketidakseimbangan kelas yang diajukan oleh Chawla dkk [13] sebagai perbaikan

metode *over-sampling*. SMOTE menggunakan pendekatan *over-sampling* pada kelas minoritas. Kelas minoritas dilakukan *over-sampling* dengan mengambil sampel pada kelas minoritas kemudian membuat sampel sintesis sepanjang garis segmen dimana melibatkan beberapa atau seluruh *k-neighbor* (sampel yang berdekatan). Nilai k disesuaikan dengan kebutuhan tingkat *over-sampling*, sebanyak tetangga yang akan diolah secara acak.

Rashu, dkk [14] membandingkan metode tiga metode *resampling* yaitu SMOTE, *Random Over Sampling* (ROS), dan *Random Under Sampling* (RUS). *Classifier* yang digunakan dalam penelitian yang dilakukan Rashu, dkk ini menggunakan NBC, Decision Tree, dan Neural Network. Dari ketiga metode penanganan ketidakseimbangan kelas yang dilakukan, SMOTE memberikan akurasi yang lebih tinggi daripada metode *resampling* yang lain pada ketiga *classifier* yang digunakan. Flores dkk [15] menggunakan SMOTE pada *sentiment analysis dataset* hasilnya *over-sampling* menggunakan SMOTE memberikan efek yang bagus pada *preprocessing sentiment analysis dataset* yang memiliki ketidakseimbangan pada kedua *classifier* yang digunakan yaitu SVM dan NBC. Sedangkan Ahsan dkk [16] melakukan penelitian untuk menunjukkan bagaimana pengaruh algoritma *machine learning* ketika menggunakan *imbalanced dataset* dan SMOTE pada *phising dataset*. Hasil dari pengujian menggunakan algoritma SVM, Random Forests, dan XGBoost menunjukkan penggunaan SMOTE dapat meningkatkan akurasi pada pembelajaran.

K-Nearest Neighbors adalah algoritma *supervised learning* dimana hasil dari *instance* yang baru diklasifikasikan berdasarkan mayoritas dari kategori k-tetangga terdekat. Menurut Gorunescu [17], algoritma KNN merupakan sebuah metode untuk klasifikasi terhadap objek dimana objek baru diberi label berdasarkan pada objek tetangga terdekatnya. KNN merupakan algoritma paling sederhana diantara algoritma *machine learning* yang lain karena sederhana dalam mengklasifikasikan objek berdasarkan *majority vote* dari tetangganya (*neighbors*). *Nearest neighbor* merupakan pendekatan untuk memberikan label berdasarkan pada kedekatan antara data yang baru dengan data yang lama berdasarkan pada kecocokan bobot dari fitur yang ada. Persamaan (1) merupakan rumus Euclidean distance untuk menghitung jarak antar dua data.

$$D(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad (1)$$

Dengan D adalah jarak antara titik x dan y yang diklasifikasikan, dimana $x = x_1, x_2, x_3, \dots, x_k$; $y = y_1, y_2, y_3, \dots, x_k$; k merepresentasikan nilai atribut, dan n merupakan dimensi atribut.

Sampel sintesis dibuat dengan menghitung nilai perbedaan (pengurangan) antaraSampel sintesis dibuat dengan menghitung nilai perbedaan (pengurangan) antara vektor *feature space* dengan vektor lain yang terletak berdekatan. Nilai pengurangan dikalikan dengan nomor acak antara nol dan satu, kemudian ditambahkan pada nilai vektor *feature space* yang telah dipilih. Proses ini menggambarkan seleksi poin secara acak pada garis segmen antara dua vektor *feature space*. Pendekatan tadi dapat memaksa area keputusan (*decision region*) dari kelas minoritas menjadi lebih umum.
Contoh:

Instance yang dipertimbangkan (6,4) dan (4,3) merupakan *nearest neighbor*-nya. (6,4) adalah sampel dimana KNN diidentifikasi, (4,3) adalah salah satu KNN-nya.
Sehingga,
f1_1 = 6 f2_1 = 4, f2_1 – f1_1 = -2
f1_2 = 4 f2_2 = 3, f2_2 – f1_2 = -1
Instance baru yang akan dibangkitkan sebagai berikut:
(f1',f2') = (6,4) + rand(0-1) * (-2,-1)
Dengan rand(0-1) merupakan angka yang dibangkitkan secara acak antara nol dan satu.

Synthetic Minority-Over Sampling Technique – Nominal (SMOTE-N) merupakan pengembangan dari SMOTE yang digunakan untuk fitur nominal dengan fitur nominal yang diajukan Chawla sebagai pengembangan dari SMOTE [13]. Pada SMOTE-N, *nearest neighbor* dihitung menggunakan versi modifikasi dari *Value Difference Metric* (VDM) yang diajukan oleh Cost dan Salzberg. VDM melihat pada nilai fitur yang overlap terhadap semua vektor fitur. Matriks mendefinisikan jarak antara nilai fitur yang sesuai untuk vektor fitur yang dibuat. Jarak δ antara dua nilai fitur yang sesuai didefinisikan sebagai berikut:

$$\delta(F_1, F_0) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{0i}}{C_0} \right|^k \quad (2)$$

Pada Persamaan (2), F_1 dan F_0 adalah dua nilai fitur yang sesuai. C_1 adalah total jumlah kemunculan dari nilai fitur F_1 dan C_{1i} adalah jumlah kemunculan dari nilai fitur F_1 untuk kelas i. Konvensi yang sama juga dapat diterapkan pada C_{0i} dan C_0. k merupakan konstanta yang biasanya bernilai 1. Persamaan (2) digunakan untuk menghitung matriks dari perbedaan nilai untuk setiap fitur nominal yang diberikan pada vektor fitur dan memberikan jarak geometris yang pasti, nilai himpunan berhingga.

Contoh:

F1 = A B C D E
F2 = A F C G N
F3 = H B C D N

F1 adalah vektor fitur yang dipertimbangkan sedangkan F2 dan F3 merupakan 2 nearest neighbor. Sehingga fitur sintesisnya FS = A B C D N.

Klasifikasi merupakan proses untuk mencari model atau fungsi yang berasaldari analisis kumpulan data *training* (data yang label kelasnya diketahui). Tujuannya untuk menggunakan model tersebut sebagai model untuk melakukan prediksi data dari kelas yang tidak diketahui label kelasnya[18]. Salah satu algoritma klasifikasi adalah NBC. NBC merupakan algoritma pada metode klasifikasi yang menggunakan probabilitas sederhana yang mengaplikasikan Teorema Bayes. Menggunakan asumsi ketidaktergantungan (independent) yang tinggi yaitu nilai antar atribut serta asumsi bahwa tidak ada atribut yang tersembunyi yang dapat mempegaruhi proses prediksi. NBC memiliki algoritma yang efisien dan sederhana. Efisiensi ini ditunjukkan dengan hanya memeriksa data *training* sekali dan data *training* sedikit untuk mengestimasi parameter seperti rerata dan variansi variabel. Naïve Bayes dirumuskan dengan Persamaan (3).

$$P(c|x) = \frac{P(X|C).P(c)}{P(x)} \quad (3)$$

Nilai NBC didasari pada perhitungan *likelihood* $P(x|c)$ yaitu kemungkinan suatu atribut (x) terhadap kelasnya, nilai $P(c)$ yaitu kemungkinan probabilitas kelas dan dibagi dengan nilai $P(x)$ yaitu kemungkinan probabilitas atribut. Nilai Naïve Bayes disebut pula sebagai probabilitas posterior yang memprediksi kelas berdasar atribut yang diuji $P(c|x)$.

3. Metodologi Penelitian

Secara umum, skema metodologi pelaksanaan penelitian ini dapat dilihat pada Gambar 1. *Dataset* yang digunakan dalam penelitian ini adalah *dataset* TB Anak yang digunakan oleh Sari dkk [6]. Dari *dataset* tersebut, kemudian dilakukan pemisahan kelas mayoritas dan minoritas. *Dataset* kelas minoritas adalah *dataset* yang merupakan kelas dengan jumlah *instance* yang paling sedikit yaitu kelas negatif. Sedangkan kelas mayoritas adalah *dataset* yang merupakan kelas dengan jumlah *instance* terbanyak yaitu kelas positif.

Gambar 1. Skema Metodologi Penilitian

Setelah *dataset* kelas mayoritas dan minoritas dipisahkan, selanjutnya *dataset* kelas minoritas dilakukan penyeimbangan jumlah agar memiliki jumlah *instance* yang sama atau mendekati kelas mayoritasnya. Jumlah *instance* dalam *dataset* TB Anak yang terbatas syaitu 38 *instance* sehingga untuk penanganan kelas minoritasnya akan menggunakan metode *over-sampling*. Metode ini dipilih agar tidak kehilangan informasi yang penting karena pengurangan *instance* yang dilakukan oleh metode *under-sampling*. SMOTE-N merupakan salah satu algoritma *over-sampling* yang memperbaiki algoritma *Random Over Sampling* (ROS) untuk memperoleh *instance* sintesis. SMOTE-N merupakan algoritma yang diajukan Chawla [13] yang menggunakan pendekatan dengan sistem kerja *over-sampling* pada kelas minoritasnya dan memperbaiki *over-fitting* pada *over-sampling* yang hanya menggandakan *instance* sedangkan SMOTE-N membuat *instance* sintesis berdasarkan kemiripan data dengan menggunakan *K-Nearest Neighbor* (KNN).

Dataset kelas minoritas akan dibuat *instance* sintesis sehingga jumlah *instance* dalam *dataset* kelas minoritas akan sama atau mendekati jumlah *instance* dalam *dataset* kelas mayoritas. Hal ini bertujuan agar rasio antar kelas menjadi seimbang yaitu satu banding satu atau minimal mendekati rasio satu banding satu untuk memperbaiki hasil pembelajaran. Konfigurasi untuk algoritma SMOTE-N yaitu KNN yang digunakan sebanyak lima *nearest neighbors* dan %N atau persen jumlah *over-sampling* adalah 200% yang berarti kelas mayoritas akan digandakan sebanyak dua kali jumlah sebelumnya.

Lima *nearest neighbor* dipilih karena jika terlalu sedikit maka menghilangkan kemungkinan data lain yang memiliki kemiripan sedangkan jika lebih dari lima akan memperberat komputasi sehingga membutuhkan waktu yang lama untuk menjalankan algoritmanya. Pada Gambar 2 dapat dilihat bahwa jumlah *instance* pada kelas mayoritas adalah

25 sedangkan jumlah *instance* pada kelas minoritas adalah 13. Sehingga, jumlah *instance* kelas minoritas akan dibuatkan *instance* sintesis sehingga jumlah *instance* kelas minoritas akan sama dengan atau mendekati jumlah *instance* kelas mayoritas yaitu 25. Jumlah *instance* minoritas akan dilakukan penambahan sebanyak dua kali lipat dari jumlah *instance* saat ini agar jumlahnya sama dengan atau mendekati 25 yaitu 26. Oleh karena itu %N yang digunakan dalam pengujian ini adalah sebanyak 200% karena menginginkan jumlah *instance* kelas minoritas membutuhkan jumlah dua kali lipat dari jumlah *instance* saat ini.

Dataset kelas minoritas dan kelas mayoritas kemudian digabungkan menjadi *dataset* baru dengan jumlah *instance* kelas minoritas dan mayoritasnya sudah memiliki rasio yang mendekati rasio satu banding satu atau memiliki rasio satu banding satu. Setelah itu, dilakukan pengujian dengan menggunakan metode klasifikasi NBC. Implementasi dengan *classifier* tersebut dilakukan menggunakan 30-*Stages* 10-*Cross Fold Validation.* Yang dimaksud dengan 30-*Stages* adalah *dataset* diacak sebanyak 30 kali dengan pembangkit angka acak yang ada dalam perangkat lunak WEKA yang disebut dengan *random seed*. Kemudian, setiap *dataset* yang telah diacak akan dilakukan validasi dengan 10-*Cross Fold Validation,* yaitu membagi *dataset* menjadi 10 bagian, dimana satu bagian akan menjadi *testing set* dan sembilan bagian sisanya digunakan sebagai *training set*, hal ini dilakukan bergantian sebanyak sepuluh kali.

Evaluasi hasil klasifikasi *dataset* yang telah dilakukan penyeimbangan kelas dinilai dengan akurasi, presisi, *recall*, *f-score*, dan *ROC area*. Akurasi merupakan standar pegukuran kinerja klasifikasi yang biasa digunakan, diperoleh dari probabilitas *instance* yang diklasifikasikan secara benar. Akurasi sebenarnya tidak tepat digunakan jika *dataset* tidak seimbang, karena akurasi diperoleh hanya dengan memprediksi semua *instance* sebagai kelas mayoritas [7], [13], [9].

Sedangkan presisi digunakan untuk menghitung seberapa sering sebuah *instance* diprediksi positif dengan benar dan *recall* adalah seberapa sering sebuah *instance* kelas positif dalam *dataset* diprediksi sebagai kelas *instance* positif. *F-score* atau yang biasa disebut dengan *f-measure* merupakan metrik kinerja yang umum digunakan untuk *imbalance learning* karena *f-score* mengombinasikan perhitungan presisi dan *recall* dengan keluaran nilai tunggal yang mencerminkan kebaikan *classifier* pada kelas yang sedikit. *Receiver Operating Characteristic* (ROC) area atau yang biasa dikenal dengan *Area Under Curve* (AUC) merupakan standar teknik untuk evaluasi *classifier.* Semakin luas area AUC maka semakin baik modelnya dan memiliki intrepetasi yang bagus sebagai probabilitas bahwa *classifier* memeringkatkan *instance* positif yang dipilih secara acak di atas *instance* negatif yang dipilih secara acak [19].

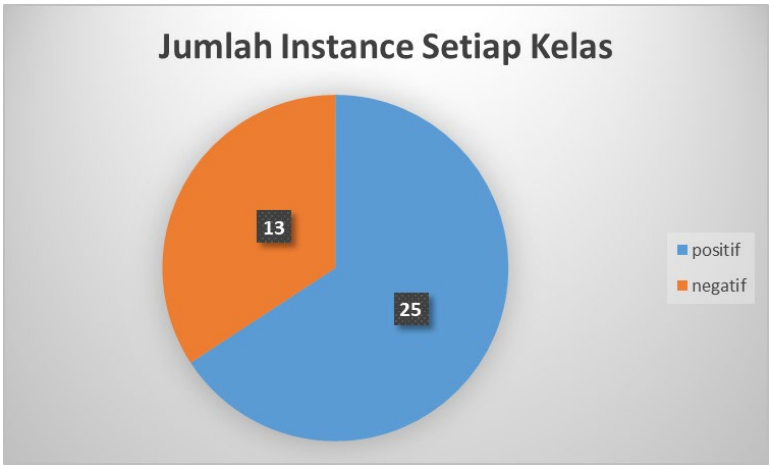
4. Hasil dan Diskusi

Dataset yang memiliki ketidakseimbangan kelas adalah *dataset* yang mempunyai rasio kelas/ kasus yang tidak seimbang antara kelas yang satu dengan kelas yang lainnya. *Dataset* yang digunakan dalam penelitian ini adalah *dataset* TB anak yang digunakan dalam penelitian yang dilakukan oleh Sari, dkk [6]. Dalam *dataset* tersebut terdapat tujuh fitur dan dua kelas. Ketujuh fitur yang ada dalam *dataset* tersebut terdapat pada Tabel 1. Namun, dalam penelitian ini hanya digunakan enam fitur yaitu Kontak TBC, Uji Tiberkulin, Demam, Batuk, Status Gizi, dan Pembesaran Kelenjar. Fitur Pembengkakan Tulang tidak digunakan dalam penelitian ini dikarenakan hasil tes pembengkakan tulang negatif semua.

Tabel 1. Fitur *Dataset* TB Anak [6]

No	Fitur
1	Kontak TBC
2	Uji Tiberkulin
3	Demam
4	Batuk
5	Status Gizi
6	Pembesaran Kelenjar
7	Pembengkakan Tulang

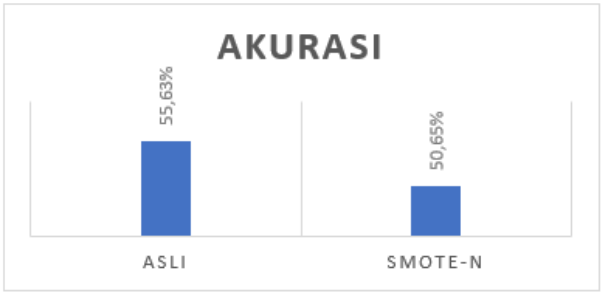
Terdapat dua kelas dalam *dataset* TB anak yaitu kelas positif yaitu kelas yang positif menderita TB anak dan kelas negatif yaitu kelas yang negatif menderita TB anak. Gambar 2 menunjukkan rasio antara hasil positif dan negatif menderita TB Anak dalam *dataset* yaitu 13:25. Ketidakseimbangan kelas ini akan berpengaruh pada hasil klasifikasi karena mengakibatkan *error* pada klasifikasi kelas minoritas yaitu kelas positif yang dikarenakan ketidakseimbangan kelas yang cenderung fokus pada kelas mayoritas yaitu kelas negatif dan mengabaikan kelas minoritas pada saat klasifikasi.



Gambar 2. Jumlah *Instance* Setiap Kelas

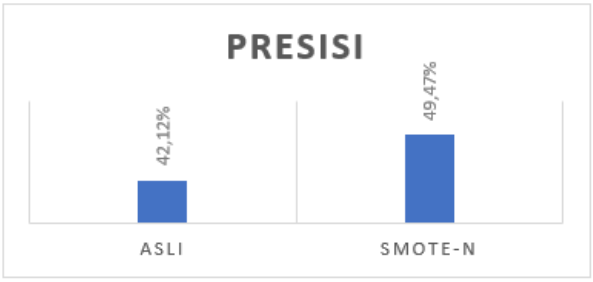
Setelah dibagi berdasarkan kelas mayoritas dan minoritasnya, kemudian dipisahkan menjadi dua *dataset* mayoritas dan minoritas. *Instance* pada kelas minoritas dilakukan penyeimbangan jumlah *instance* menggunakan SMOTE-N untuk mendapatkan rasio yang mendekati satu banding satu. Setelah itu, *dataset* kelas minoritas dan mayoritas digabungkan kembali menjadi *dataset* baru. Jumlah *instance* setelah dilakukan penyeimbangan kelas dengan SMOTE-N yaitu 25 kelas negatif dan 26 kelas positif atau rasionya menjadi 25:26 yaitu mendekati rasio satu banding satu.

Dataset hasil *over-sampling* dengan SMOTE-N kemudian diuji dengan mengevaluasi kinerja dari algoritma klasifikasi NBC berdasarkan pada akurasi, presisi, *recall*, *f-score*, dan *ROC Area*. Pengujian dilakukan dengan 30-*stages* 10-*cross fold validation* kemudian hasilnya dilakukan rata-rata.

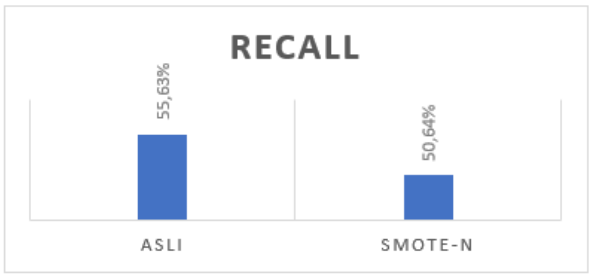


Gambar 3. Akurasi

Berdasarkan pada Gambar 3, hasil rata-rata akurasi dari data asli mengalami penurunan setelah dilakukan penyeimbangan kelas yaitu dari 55,6% menjadi 50,7%.

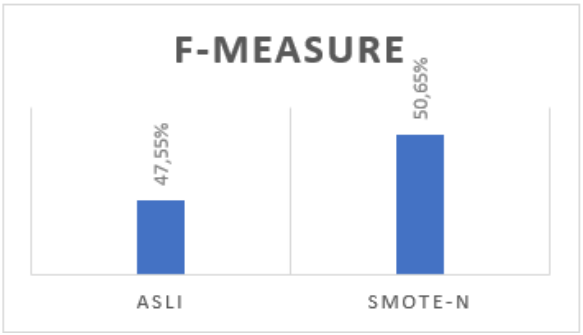


Gambar 4. Presisi

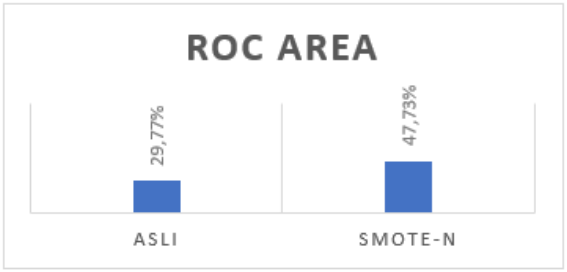


Gambar 5. *Recall*

Berdasarkan pada Gambar 4 nilai presisi mengalami kenaikan dari 42,1% menjadi 49,5% dan *recall* pada Gambar 5 mengalami penurunan dari 55,6% menjadi 50,6%.



Gambar 6. *F-Measure*



Gambar 7. *ROC Area*

Gambar 6 menunjukkan nilai *f-measure* dan Gambar 7 menunjukkan nilai *ROC Area* setelah dilakukan penyeimbangan kelas dengan SMOTE-N mengalami kenaikan yaitu *f-measure* dari 47,5% menjadi 50,7% dan *ROC Area* dari 29,8% menjadi 47,7%.

5. Kesimpulan dan Saran

Dari eksperimen yang telah dilakukan dalam penelitian ini dapat diambil beberapa kesimpulan. Pertama, nilai akurasi mengalamani penurunan setelah dilakukan penyeimbangan kelas dengan SMOTE-N dikarenakan *instance* sintesis yang dihasilkan memiliki tingakat kemiripan yang tinggi sehingga menurunkan hasil akurasinya. Kedua, nilai *recall* mengalami penurunan yang disebabkan penurunan *instance* kelas positif (kelas mayoritas) dalam *dataset* yang diprediksi sebagai kelas *instance* positif yang disebabkan penambahan *instance* kelas negatif (minoritas). Ketiga, nilai *f-measure* mengalami kenaikan hal ini menunjukkan bahwa penggunaan SMOTE-N untuk penyeimbangan kelas pada *dataset* TB Anak berhasil karena nilai *f-measure* menunjukkan kebaikan *classifier* pada kelas yang sedikit (minoritas). Keempat, kenaikan nilai *ROC Area* menunjukkan semakin membaiknya model dan intrepetasi yang bagus sebagai probabilitas bahwa *classifier* memeringkatkan *instance* positif yang dipilih secara acak di atas *instance* negatif yang dipilih secara acak. Kelima, penggunaan SMOTE-N dapat meningkatkan performa *machine learning* pada *dataset* TB anak yang memiliki ketidakseimbangan kelas.

Saran untuk penelitian selanjutnya adalah pertama membandingkan dengan algoritma penanganan ketidakseimbangan kelas yang lain sebagai pembanding untuk mendapatkan model yang terbaik. Kedua, penambahan *instance* pada *dataset* TB Anak untuk mendapatkan informasi yang lebih banyak dan pembelajaran *machine learning* yang lebih baik.

Referensi

[1] Pusat Data dan Informasi Kementrian Kesehatan, *Info Data dan Informasi Tuberkulosis*. 2015.

[2] World Health Organization, "Global tuberculosis report 2018," *WHO*, 2018.

[3] Kementrian Kesehatan Republik Indonesia, "Peduli TBC, Indonesia Sehat," 2018. [Online]. Available: http://www.depkes.go.id/article/view/18032100002/peduli-tbc-indonesia-sehat.html. [Accessed: 22-Oct-2018].

[4] Kementrian Kesehatan Republik Indonesia, "TB Anak : TB Indonesia." [Online]. Available: http://www.tbindonesia.or.id/tb-anak/. [Accessed: 19-Oct-2018].

[5] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," *Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018*, pp. 1–11, 2018.

[6] W. E. Sari, O. Wahyunggoro, and S. Fauziati, "A Comparative Study on Fuzzy Mamdani-Sugeno-Tsukamoto for The Childhood Tuberculosis Diagnosis," *AIP Conf. Proc.*, vol. 1755, no. 1, p. 70003, 2016.

[7] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*, 1st ed. Wiley-IEEE Press, 2013.

[8] A. Dal Pozzolo, O. Caelen, and G. Bontempi, "Comparison of Balancing Techniques for Unbalanced Datasets," *Mach. Learn. Gr. Univ. Libr. Bruxelles Belgium*, vol. 16, no. 1, pp. 732–735, 2010.

[9] K. Li, W. Zhang, Q. Lu, and X. Fang, "An Improved SMOTE Imbalanced Data Classification Method Based on Support Degree," in *International Conference on Identification, Information and Knowledge in Internet of Things*, 2014, pp. 34–38.

[10] F. Koto, "SMOTE-Out, SMOTE-Cosine, and Selected-SMOTE : An Enhancement Strategy to Handle Imbalance in Data Level," *ICACSIS*, pp. 280–284, 2014.

[11] J. Li, H. Li, and J.-L. Yu, "Application of Random-SMOTE on Imbalanced Data Mining," *Fourth Int. Conf. Bus. Intell. Financ. Eng.*, pp. 130–133, 2011.

[12] G. I. Winata and M. L. Khodra, "Handling imbalanced dataset in multi-label text categorization using Bagging and Adaptive Boosting," in *Proceedings - 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community, ICEEI 2015*, 2015, pp. 500–505.

[13] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[14] R. I. Rashu, N. Haq, and R. M. Rahman, "Data Mining Approaches to Predict Final Grade by Overcoming Class Imbalance Problem," in *2014 17th International Conference on*

Computer and Information Technology, ICCIT 2014, 2014, pp. 14–19.

[15] A. C. Flores, R. I. Icoy, C. F. Peña, and K. D. Gorro, "An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set," in *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, 2018, pp. 1–4.

[16] M. Ahsan, R. Gomes, and A. Denton, "SMOTE Implementation on Phishing Data to Enhance Cybersecurity," in *2018 IEEE International Conference on Electro/Information Technology (EIT)*, 2018, pp. 531–536.

[17] F. Gorunescu, *Data mining: Concepts, models and techniques*, vol. 12. Berlin: Springer, 2011.

[18] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," *Ann. Phys. (N. Y).*, vol. 54, p. 770, 2006.

[19] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Third. Amsterdam: Morgan Kaufmann, 2011.