ISSN: ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 9 , Nomor 3, Tahun 2020, Halaman 227 - 236
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PERBANDINGAN METODE SMOTE *RANDOM FOREST* DAN SMOTE XGBOOST UNTUK KLASIFIKASI TINGKAT PENYAKIT HEPATITIS C PADA *IMBALANCE CLASS DATA*

Muhamad Syukron[1], Rukun Santoso[2], Tatik Widiharih[3]
[1,2,3] Departemen Statistika FSM Universitas Diponegoro
rukunsantoso25@gmail.com

ABSTRACT

Hepatitis causes around 1.4 million people die every year. This number makes hepatitis to be the largest contagious disease in the number of deaths after tuberculosis. Liver biopsy is still the best method for diagnosing the stage of hepatitis C, but this method is an invasive, painful, expensive, and can cause complications. Non-invasively method needs to be developed, one of non-invasif method is machine learning. Random Forest and XGboost are classification methods that are often used, since they have many advantages over classical classification methods. The SMOTE algorithm can be used to improve the accuracy of predictions from imbalanced data. the data in this study have 24 independent variables in the form of patients self-data, hepatitis C symptoms, and laboratory test results. The dependent variable in this study is a binary category, namely the level of hepatitis C disease (fibrosis and cirrhosis). The results showed that the random forest and XGboost had an accuracy of around 74% but the recall value was less than 2%. SMOTE random forest dan SMOTE XGboost have an accuracy & recall value more than 75%. SMOTE random forest has a higher accuracy for predicting fibrosis class while SMOTE XGboost is better in cirrhosis class. Variables that are more influental in determining hepatitis C stage are variables from laboratory test.

Keyword : Fibrosis, Cirrhosis, Random Forest, SMOTE, XGboost

1. PENDAHULUAN

Pada tahun 2019, terdapat 325 juta orang di dunia hidup dengan mengidap virus hepatitis B dan C. Hepatitis C adalah penyakit hati yang disebabkan oleh virus hepatitis C (HCV). Virus hepatitis C adalah virus yang ditularkan melalui darah, misalnya dengan digunakannya peralatan kesehatan yang tidak aman, narkoba suntikan, tranfusi darah, serta praktik seksual yang mengarah pada paparan darah. Sebagian besar dari penderita yang terinfeksi hepatitis C kronis akan berkembang hingga menjadi sirosis atau kanker hati [9].

Sejauh ini biopsi hati yang merupakan prosedur medis dengan cara mengambil sebagian kecil jaringan hati masih menjadi metode terbaik untuk mendiagnosis dan menentukan stadium fibrosis hati pada pasien dewasa dan anak-anak. Namun, cara ini merupakan proses invasif, menyakitkan, mahal, dan dapat menyebabkan komplikasi. Pengembangan metode selain biopsi hati yaitu metode non-invasif perlu ditingkatkan[1]. Penentuan tingkat penyakit hepatitis C bisa dilakukan berdasarkan dari hasil tes laboratorium rutin, penanda serum biokimia, dan berbagai teknik pencitraan terutama untuk mengukur kekakuan hati, atau kombinasi dari beberapa atau semua cara tersebut. *Machine learning* dapat diterapkan untuk menjaga keakuratan hasil prediksi berdasarkan data hasil rekam medis pasien. XGboost dan *random forest* merupakan metode klasifikasi yang bisa diterapkan untuk prediksi tingkat penyakit hepatitis C yaitu masuk ke tahap fibrosis atau sirosis. Kedua metode tersebut memiliki banyak keunggulan dibanding metode klasifikasi lainnya karena lebih *robust* terhadap *outlier*, waktu komputasi yang kecil, serta hasil yang akurat.

Data yang dimiliki peneliti tidak selalu bisa diolah secara langsung, adakalanya data tersebut memiliki masalah seperti adanya *imbalance class data*. Dataset dikategorikan sebagai *imbalance class data* apabila proporsi antar kelas respon tidak ekuivalen[2]. Kelas respon tidak seimbang menyebabkan hasil prediksi akan akurat hanya pada satu kelas tertentu yaitu kelas dengan respon terbanyak. Penanganan *imbalance class data* bisa dilakukan dengan algoritma SMOTE yaitu membangkitkan data *syntetic* dari kelas minor.

2. TINJAUAN PUSTAKA

2.1 Hepatitis C

Hepatitis C merupakan penyakit yang dapat menyerang organ hati. Tahap hepatitis C diantaranya adalah inflamasi, fibrosis, sirosis, dan gagal hati atau kanker hati. Seseorang yang mengalami hepatitis C pada usia tua cenderung lebih susah untuk disembuhkan. Wanita secara biologis lebih mudah sembuh dengan sendirinya karena memiliki sistem imun lebih baik. Gejala yang mungkin dirasakan oleh seseorang yang mengidap penyakit hepatitis C adalah rasa lelah, demam, diare, mual, dan bola mata atau kulitnya akan berwarna kekuningan [5]. Deteksi penyakit hepatitis C bisa dilakukan dengan mengambil sampel darah yaitu untuk melihat jumlah sel darah merah, sel darah putih, hemoglobin, dan trombosit. Selain itu perlu juga untuk melihat jumlah enzim *aspartate aminotransferase* (AST) dan *alanaine aminotransferase* (ALT) dalam darah. Meningkatnya enzim AST dan ALT bisa mengindikasikan adanya gangguan organ hati[8]. Jumlah HCV RNA juga perlu diukur untuk melihat banyaknya virus hepatitis C dalam tubuh seseorang.

2.2 *Pre-Processing* Data

Data yang diambil dari lapangan tidak bisa digunakan secara langsung karena bisa saja data tersebut memiliki beberapa masalah yang perlu diselesaikan agar menjadi data yang bersih dari *noise*. Data yang tidak bersih artinya ada masalah seperti *missing value* (data hilang) serta adanya *outlier*. Keberadaan *missing value* bisa diatasi dengan menghilangkan observasi yang memiliki data hilang tersebut atau mengganti data dengan suatu estimasi nilai, misalkan dengan nilai modus apabila data berbentuk kategorik dan dengan nilai rata-rata apabila data memiliki nilai kontinyu[6]. *Outlier* bisa dideteksi dengan boxplot. Data dikategorikan sebagai *outlier* apabila berada diluar tubuh dan *whisker* boxplot yaitu nilainya lebih besar dari Q3+1,5*IQR atau kurang dari Q1+1,5*IQR. Q1 adalah kuartil bawah, Q3 adalah kuartil atas, serta IQR merupakan *interquartile* (Q3-Q1)[4].

2.3 *Random Search & Grid Search*

Random forest dan XGboost memiliki beberapa parameter yang perlu diatur agar menghasilkan kumpulan pohon keputusan yang dapat memprediksi tingkat penyakit hepatitis C secara akurat. Parameter terbaik bisa dicari dengan beberapa algoritma seperti *random search* dan *grid search*. Algoritma *grid search* akan mencobakan semua parameter yang ditentukan oleh peneliti sedangkan *random search* hanya akan mencoba beberapa kombinasi yang jumlah kombinasinya sudah ditentukan. *Random search* akan memiliki *range* percobaan yang lebih besar dibandingkan *grid search* jika jumlah percobaan yang ditentukan sama. *Random search* juga akan efektif digunakan apabila jumlah dimensi parameter yang dicobakan besar.

2.4 Holdout Validation & K-Fold Validation

Data yang ada biasanya tidak seluruhnya digunakan dalam proses pelatihan. Untuk menghindari *overfitting* maka model harus dibangun sedemikian rupa sehingga ketika ada data baru bisa diprediksi sama baiknya dengan menggunakan data pada proses pelatihan. Pengujian model bisa dilakukan dengan data test yang telah memiliki label kelas respon. Pembagian data train dan data test bisa dilakukan dengan *holdout validation* yaitu membagi data menjadi 2 bagian dengan proporsi tertentu yang ditentukan oleh peneliti. Proporsi yang biasa digunakan oleh peneliti adalah 60/40, 70/30, atau 80/20 [7]. Selain *holdout validation*, data juga bisa dibagi menjadi data latih dan data uji dengan metode K-fold *cross validation*. Metode ini membagi data latih dan data uji sebanyak "k" kelompok, sehingga proses pelatihan akan menjadi sebanyak "k" kemudian *performance* dari model merupakan rata-rata dari semua proses pelatihan tersebut.

2.5 XGboost

XGboost merupakan salah satu metode *boosting* yaitu kumpulan *decission tree* yang pembangungan pohon berikutnya akan bergantung pada pohon sebelumnya. Pohon pertama dalam XGboost akan lemah dalam melakukan klasifikasi dengan inisialisasi *probability* yang ditentukan oleh peneliti dan kemudian akan dilakukan *update* bobot pada setiap pohon yang dibangun sehingga menghasilkan kumpulan pohon klasifikasi yang kuat. Prediksi dilakukan dengan menjumlahkan seluruh bobot yang ada di setiap pohon dan kemudian memasukkan nilai tesebut ke fungsi logistik. XGboost akan meminimumkan fungsi objektif sebagai berikut:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i{}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

Loss function pada klasifikasi kelas respon biner bisa menggunakan *log loss*. Persamaan *Omega* merupakan parameter regularisasi yang akan membuat model berusaha menghindari *overfitting*. Nilai gain bisa ditentukan untuk penentuan *splitting node*. Berikut ini adalah rumus untuk mencari nilai gain pada XGboost:

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_L} h_R + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma$$

Nilai $g_i$dan h_i merupakan turunan pertama dan kedua *loss function* pada XGboost[3].

2.6 *Random Forest*

Random Forest (RF) adalah metode klasifikasi dan regresi yang berbentuk kumpulan pohon keputusan. Dalam model *Random Forest*, masing-masing pohon adalah *Classification and Regression Trees* (CART) yang menggunakan *Decrease Gini Impurity* dalam pemilihan prediktor pemisah dari *subset* yang dipilih secara acak dari semua variabel prediktor yang tersedia. Setiap pohon juga tidak menggunakan semua data asli melainkan menggunakan data sampel *bootstrap* dengan pengembalian. Penentuan kelas diambil berdasarkan mayoritas hasil *vote* dari semua pohon yang terbentuk. Beberapa parameter *random forest* adalah *mtry* yaitu jumlah *feature* yang dicobakan dalam proses pemilahan serta *ntree* yang merupakan jumlah pohon yang dibangun dalam suatu model.

2.7 SMOTE

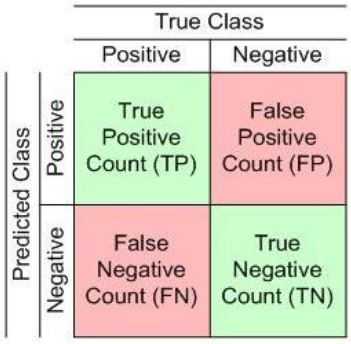
Metode SMOTE menggunakan prinsip *oversampling* yaitu menambah data dari kelas minor agar jumlahnya seimbang dengan data dari kelas mayor. SMOTE akan membangkitkan data dari kelas minor dengan pendekatan ketetanggan[2]. Misalkan diberikan data dengan jumlah variabel p maka jarak antara $x^T = [x_1, x_2, \dots, x_p]$ dan $z^T = [z_1, z_2, \dots, z_p]$ adalah $d(x, y) = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \cdots + (x_p - z_p)^2}$. Untuk membangkitkan data dengan metode SMOTE maka digunakan persamaan sebagai berikut:

$$x_{syn} = x_i + (x_{knn} - x_i)\gamma$$

x_{syn} merupakan pengamatan baru hasil pembangkitan, x_i adalah pengamatan ke-i, xknn merupakan x terdekat dari x_i, serta gamma merupakan bilangan acak antara 0 dan 1. Untuk data nominal maka akan diisi dengan nilai mayoritas pada k-tetangga terdekat. Perhitungan jarak pada SMOTE apabila ada variabel kategorik maka akan diganti dengan kuadrat median standar deviasi variabel kontinyu kelas minoritas jika nilai kategorik pada pengamatan ke-i dan j berbeda.

2.8 Pengukuran Hasil Prediksi

Ukuran kinerja dari algoritma *machine learning* biasanya dievaluasi dengan *confusion matrix* seperti tampak pada gambar dibawah ini:



Gambar 1 Confusion Matrix Klasifikasi Biner

Nilai akurasi merupakan keakuratan prediksi secara keseluruhan sedangkan *recall* merupakan ukuran kebaikan model yang bisa digunakan untuk melihat keakuratan pada satu kelas terutama pada kasus dataset yang tidak seimbang. Nilai akurasi dan *recall* dapat dihitung dengan rumus sebagai berikut

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, recall = \frac{TP}{TP + FN}$$

3. METODOLOGI PENELITIAN

Data yang digunakan dalam penelitian ini adalah data sekunder, yaitu data pasien hepatitis C yang diperoleh dari *UCI Machine Learning Repository* sebanyak 1385 data. Variabel data yang digunakan dalam penelitian ini adalah tingkat penyakit hepatitis C (fibrosis dan sirosis) sebagai variabel bebas dan terdapat 24 variabel bebas. Variabel bebas penelitian ini berupa data diri pasien, gejala, serta hasil uji laboratorium yaitu sebagai berikut: usia (x_1), jenis kelamin (x_2), indeks masa tubuh (x_3), demam (x_4), mual (x_5), sakit kepala (x_6), diare (x_7), rasa lelah (x_8), penyakit kuning (x_9), nyeri ulu hati (x_{10}), sel darah putih (x_{11}), sel darah merah (x_{12}), hemoglobin (x_{13}), trombosit (x_{14}), enzim AST *week 1* (x_{15}), enzim ALT *week 1* (x_{16}), ALT *week 4* (x_{17}), ALT *week 12* (x_{18}), ALT *week 24* (x_{19}), ALT *week 36* (x_{20}), ALT *week 48* (x_{21}), HCV RNA *week 4* (x_{22}), HCV RNA *week 12* (x_{23}), HCV RNA *end of treatment* (x_{24}). Variabel gejala bernilai kategorik biner dengan 0 berarti tidak mengalami gejala dan 1 berarti mengalami gejala.

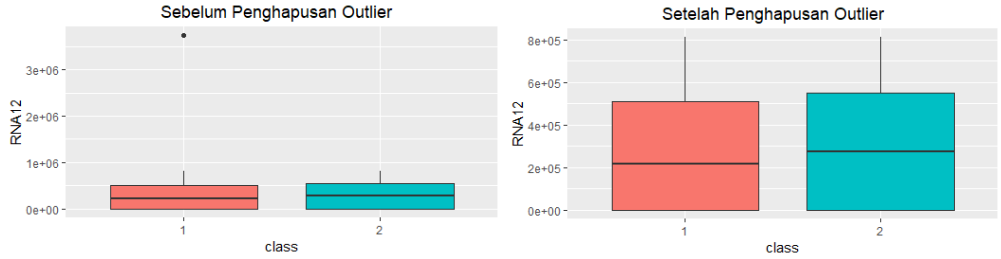
Data dalam penelitian ini diolah dengan menggunakan *python 3.7* dan *R studio*. Langkah-langkah analisis dalam menyelesaikan penelitian ini adalah sebagai berikut:

1. Memasukkan data penelitian ke R studio untuk visualisasi data dan ke dalam python 3.7 untuk pemodelan secara lengkap
2. Melakukan *pre-processing*
   a. Cek keberadaan *missing value*. Apabila data yang digunakan dalam penelitian terdapat *missing value* maka akan dilakukan *imputasi* menggunakan nilai rata-rata (variabel kontinyu) atau menggunakan nilai modus (variabel diskrit). Apabila tidak ada *missing value* maka akan dilanjutkan dengan deteksi *outliers.*
   b. Deteksi *outlier* menggunakan boxplot. Apabila terdapat *outlier* pada data penelitian maka akan dihilangkan pengamatan yang mengandung *outlier*.
3. Melakukan *copy* dataset sehingga terdapat 2 dataset yang sama persis
   - Dataset pertama tidak dilakukan *balancing* sehingga seperti data semula setelah *pre-processing*. Kemudian dilakukan langkah-langkah penelitian sebagai berikut:
     a. Melakukan *tuning hyperparameter random forest* dan XGboost
     b. Melakukan pembagian data train dan data test dengan *holdout validation* dan 5-Fold *cross validation*
     c. Membentuk model pohon klasifikasi *random forest* dan XGBoost
     d. Melakukan prediksi untuk data train dan data test dengan model pohon klasifikasi *random forest* dan XGboost yang telah terbentuk
     e. Melakukan pengukuran akurasi dan *recall*
   - Dataset kedua dilakukan *balancing* sehingga jumlah data untuk kelas fibrosis dan sirosis sama. *Balancing* data dilakukan dengan algoritma SMOTE. Selanjutnya dilakukan langkah penelitian sebagai berikut:
     a. Melakukan *tuning hyperparameter* SMOTE *Random Forest* dan SMOTE XGboost
     b. Melakukan pembagian data train dan data test dengan holdout validation dan 5-Fold cross validation
     c. Membentuk model pohon klasifikasi SMOTE random forest dan SMOTE XGBoost
     d. Melakukan prediksi untuk data train dan data test dengan model pohon klasifikasi SMOTE random forest dan SMOTE XGboost yang telah terbentuk
     e. Melakukan pengukuran akurasi dan recall
4. Menentukan model pohon keputusan terbaik diantara *random forest*, XGboost, SMOTE *random forest*, dan SMOTE XGboost berdasarkan nilai akurasi dan recall

4. ANALISIS DAN PEMBAHASAN

4.1 Pre-Processing

Pengecekan adanya *missing value* dilakukan pada dataset awal. Setelah diamati maka dapat diketahui tidak ada data yang hilang kemudian dilanjutkan deteksi *outlier* dengan boxplot. Terdapat satu *outlier* pada kelas fibrosis pada variabel HCV RNA *week 12*. *Outlier* sangat jauh dari pengamatan lain dan hanya berjumlah satu sehingga dilakukan penghapusan observasi yang mengandung *outlier* sehingga data menjadi 1384 baris.



Gambar 2 Deteksi Outlier dengan Boxplot

4.2 *Random Forest*

Penentuan parameter terbaik dari model dilakukan *tuning* secara simultan dengan *grid search*. Berikut ini adalah parameter yang dicobakan:

Tabel 1 Parameter *Random Forest*

Parameter	Nilai Parameter
Mtry	3, 4, 5, 6, 7,8,9,10, 11, 12
Ntree	50, 75, 100, 150, 200, 250, 300, 500, 750, 1000

Tuning parameter ini menggunakan 5-fold *cross validation* dan didapatkan parameter terbaik yaitu *mtry* = 3 dan *ntree* = 100. Berikut ini adalah ukuran kebaikan model dari *random forest*.

Tabel 2 Ukuran Kebaikan Model *Random Forest*

	Train-Test (70%-30%)	Train-Test (75%-25%)	Train-Test (80%-20%)
Akurasi	75%	75,72%	73,65%
Recall	0%	0%	0%

Tabel diatas menunjukkan nilai akurasi yang dihasilkan lebih dari 70% namun nilai *recall* yang dihasilkan bernilai 0%. Nilai *recall* yang kecil mengindikasikan model tidak mampu memprediksi dengan benar pasien yang berada di tingkat sirosis. Pengembangan model yang dilakukan adalah dengan menambah algoritma SMOTE sehingga menjadi model SMOTE *Random Forest*.

4.3 XGboost

Penentuan parameter terbaik dari model dilakukan *tuning* secara simultan dengan *random search* 1000 percobaan. Berikut ini adalah parameter yang dicobakan:

Tabel 3 Parameter XGboost

Parameter	Nilai Parameter
Learning rate	0.01, 0.02, 0.03, 0.04,0.05
	0.06, 0.08, 0.09, 0.1, 0.12
Max depth	6, 7, 8, 9
Min child weight	0.7, 0.8, 0.9, 1, 1.2
Gamma	0.1, 0.2, 0.3, 0.4
Colsample by tree	0.5, 0.6, 0.7, 0.8
N estimators	30, 50, 75, 100, 125

Tuning parameter ini menggunakan 5-fold *cross validation* dan didapatkan parameter terbaik yaitu *min child weight* = 0.7, *max depth* = 7, *learning rate* = 0.01, gamma = 0.1, *colsample bytree* = 0.5, *n_estimators* = 30. Berikut ini adalah ukuran kebaikan model dari XGboost.

Tabel 4 Ukuran Kebaikan Model XGboost

	Train-Test (70%-30%)	Train-Test (75%-25%)	Train-Test (80%-20%)
Akurasi	75,24%	75,14%	73,65%
Recall	1,94%	0%	0%

Tabel diatas menunjukkan nilai akurasi yang dihasilkan lebih dari 70% namun nilai *recall yang* dihasilkan kurang dari 2%. Nilai *recall* yang kecil mengindikasikan model tidak mampu memprediksi dengan benar pasien yang berada di tingkat sirosis. Pengembangan model yang dilakukan adalah dengan menambah algoritma SMOTE sehingga menjadi model SMOTE XGboost.

4.4 SMOTE *Random Forest*

Penggunaan algoritma SMOTE membuat data dengan label sirosis menjadi lebih banyak *dan* jumlahnya sama dengan kelas fibrosis yaitu 1022 data.

Tabel 5 Jumlah Kelas Setelah *Oversampling SMOTE*

Tingkat Penyakit Hepatitis C	Jumlah Data
Fibrosis	1022
Sirosis	1022

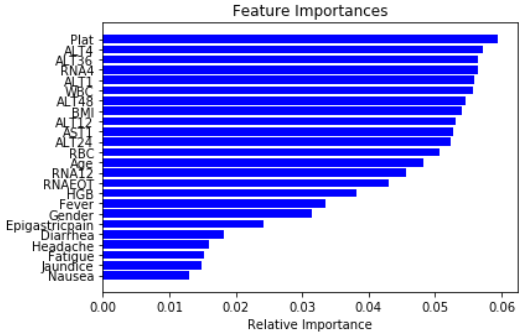
Setelah dilakukan *balacing data* maka dilakukan *tuning parameter* dengan parameter yang dicobakan sama dengan model *random forest* sebelumnya. Parameter terbaik untuk SMOTE *random forest* yaitu *mtry = 3* dan *ntree* = 100.

Tabel 6 Ukuran Kebaikan Model SMOTE *Random Forest*

	Train-Test (70%-30%)	Train-Test (75%-25%)	Train-Test (80%-20%)
Akurasi	79,64%	81,60%	81,66%
Recall	72,89%	78,16%	75,59%

Tabel diatas menunjukkan ukuran kebaikan model dengan *holdout validation.* Nilai akurasi yang dihasilkan lebih dari 79% dan nilai *recall* yang dihasilkan lebih dari 72%. Hasil rata-rata akurasi dan *recall* SMOTE *random forest* dengan *spliting* 5-fold CV berturut-turut yaitu sebesar 80,04% dan 75,34%. Nilai *recall* yang tidak jauh berbeda dengan akurasi total mengindikasikan model telah mampu memprediksi dengan benar untuk kedua kelas.

Penentuan variabel yang lebih berpengaruh dalam pemodelan bisa menggunakan *mean decrease gini* (MDG). Nilai MDG dihitung dengan rata-rata gini sebuah variabel ketika variabel tersebut terpilih sebagai *splitting node*. *Variable importance* pada SMOTE *random forest* menunjukkan bahwa variabel dari uji laboratorium seperti jumlah trombosit dan enzim ALT lebih berpengaruh dibanding variabel data diri pasien maupun gejala.



Gambar 3 *Feature Importance* SMOTE *Random Forest*

4.5 SMOTE XGboost

Penentuan parameter terbaik dari model dilakukan *tuning* secara simultan dengan *random search* 1000 percobaan. Berikut ini adalah parameter yang dicobakan:

Tabel 7 Parameter SMOTE XGboost

Parameter	Nilai Parameter
Learning rate	0.09, 0.1, 0.11, 0.12, 0.13
Max depth	5, 6, 7, 8
Min child weight	0.6, 0.7, 0.8, 0.9
Gamma	0.1, 0.2, 0.3
Colsample by tree	0.6, 0.7, 0.8
N estimators	50, 75, 100

Tuning parameter ini menggunakan 5-fold *cross validation* dan didapatkan parameter terbaik yaitu *min child weight* = 0.6, *max depth* = 7, *learning rate* = 0.1, gamma = 0.2, *colsample bytree* = 0.7, *n_estimators* = 100

Tabel 8 Ukuran Kebaikan Model SMOTE XGboost

	Train-Test (70%-30%)	Train-Test (75%-25%)	Train-Test (80%-20%)
Akurasi	77,02%	79,84%	79,32%
Recall	76,06%	77,77%	76,64%

Tabel diatas menunjukkan ukuran kebaikan model dengan *holdout validation.* Nilai akurasi yang dihasilkan lebih dari 77% dan nilai *recall* yang dihasilkan lebih dari 76%. Hasil rata-rata akurasi dan *recall* SMOTE *random forest* dengan *spliting* 5-fold CV berturut-turut yaitu sebesar 78,72% dan 78,88%. Nilai *recall* yang tidak jauh berbeda dengan akurasi total mengindikasikan model telah mampu memprediksi dengan benar untuk kedua kelas.

Penentuan variabel yang lebih berpengaruh dalam pemodelan bisa menggunakan frekuensi terpilihnya variabel tersebut sebagai pemilah, semakin besar frekuensi suatu variabel terpilih maka akan semakin besar pengaruhnya pada pemodelan. *Variable importance* pada SMOTE XGboost menunjukkan bahwa variabel dari uji laboratorium seperti jumlah sel darah merah dan jumlah HCV RNA lebih berpengaruh dibanding variabel data diri pasien maupun gejala.

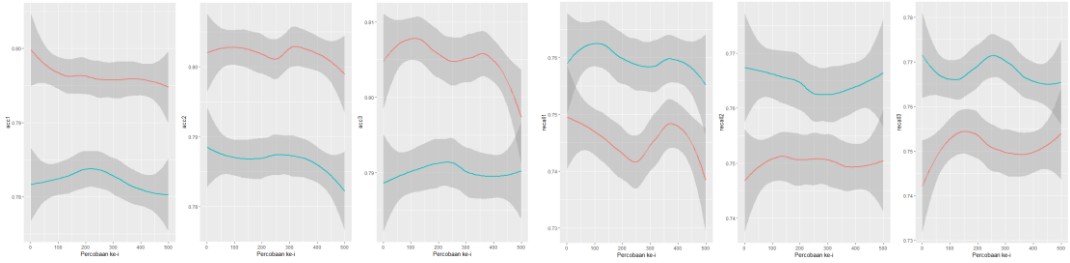
4.6 Perbandingan Metode Klasifikasi

Nilai akurasi dari keempat metode yaitu *random forest,* XGboost, SMOTE *random forest,* dan SMOTE XGboost tidak jauh berbeda yaitu diatas 70%, namun nilai *recall* antara model sebelum dan sesudah menggunakan SMOTE terlihat berbeda. Penggunaan SMOTE menyebabkan nilai *recall random forest* maupun XGboost yang dibawah 2% menjadi lebih dari 70%.

Tabel 9 Perbandingan Akurasi dan *Recall* 4 Model

	Random Forest	XGboost	SMOTE RF	SMOTE XGB
Akurasi	74,79%	74,68%	80,97%	78,63%
Recall	0%	0,65%	75,55%	76,82%

Gambar 4 menampilkan nilai akurasi (gambar kiri) dan *recall* (gambar kanan) data test pada model SMOTE *random forest* (merah) dan SMOTE XGboost (biru). Nilai akurasi dan *recall* data test tersebut berjumlah 500 untuk tiap metode yang dijalankan secara *looping*. Gambar paling kiri, tengah, dan kanan secara berurutan merupakan nilai akurasi serta *recall* data test pada proporsi split 70%:30%, 70%:25%, dan 80%:20%. SMOTE *Random Forest* cenderung memiliki akurasi yang lebih tinggi dibandingkan dengan SMOTE XGboost akan tetapi SMOTE XGboost memiliki nilai *recall* yang lebih tinggi dibandingkan SMOTE *Random Forest*.



Gambar 4 Akurasi dan Recall 500 Percobaan SMOTE RF vs SMOTE XGboost

5. KESIMPULAN

Berdasarkan hasil pembahasan maka dapat diketahui bahwa algoritma SMOTE dapat memperbaiki model sehingga dapat memprediksi dengan akurat pada semua kelas respon. SMOTE *Random Forest* memiliki akurasi keseluruhan yang lebih tinggi dibanding SMOTE

XGboost, namun XGboost memiliki *recall* kelas sirosis yang lebih baik. Variabel yang berpengaruh besar pada pemodelan SMOTE *Random Forest* dan SMOTE XGboost adalah variabel hasil uji laboratorium seperti jumlah sel darah, jumlah enzim ALT, serta jumlah HCV RNA dalam tubuh pasien.

DAFTAR PUSTAKA

[1] Barakat, N. H., Barakat, S. H., & Ahmed, N., 2019. Prediction and Staging of Hepatic Fibrosis in Children with Hepatitis C Virus: A Machine Learning Approach. *Healthcare Informatics Research,* Volume 25,p. 173.

[2] Chawla, N.V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research,* Volume 16, p. 321-357

[3] Chen, T. & Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, Knowledge Discovery and Data Mining.

[4] Jajo, N. & Matawie, K. M., 2019. Outlier Detection Using Boxplot. *International Journal of Ecology and Development,* Volume 13, pp. 116-122.

[5] John, T. M. S., 2008. Signs and Symptomps that May be Associated with Hepatitis C. *Hepatitis C Choices.* Caring Ambassadors Program, Inc., pp. 71-80

[6] Kotsiantis, S., Pintelas, P. E., & Kanellopoulus, D., 2006. Data Preprocessing for Supervised Learning. *International Journal of Computer Science,* Volume 1, pp. 111-117.

[7] Raschka, S., 2018. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning.

[8] Sandt, L., 2008. Understanding Hepatitis C disease. *Hepatitis C Choices.* Caring Ambassadors Program, Inc., pp. 23-42.

[9] World Health Organization (WHO), 2019. Hepatitis C. Retrieved from https://www.who.int/news=room/fact-sheets/detail/hepatitis-c