SKRIPSI

METODE *RESAMPLING* DAN *RANDOM FOREST* UNTUK KLASIFIKASI DATA

CICIANA
E0118001

PROGRAM STUDI MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SULAWESI BARAT
TAHUN 2024

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh :

Nama : Ciciana

NIM : E0118001

Judul : Metode *Resampling* dan *Random Forest* Untuk Klasifikasi Data

Telah berhasil dipertahankan di depan Tim Penguji (SK Nomor : 65/UN55.7/HK.04/2023) dan diterima sebagai bagian persyaratan memperoleh gelar sarjana Matematika (S.Mat.) pada Program Studi Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sulawesi Barat.

Disahkan oleh:

Dekan FMIPA

Universitas Sulawesi Barat

Musafira, S.Si., M.Sc.

NIP. 197709112006042002

Tim Penguji:

Ketua Penguji : Musafira, S.Si., M.Sc. (.....)

Sekretaris : Ahmad Ansar, S.Pd., M.Sc. (.....)

Pembimbing 1 : Rahmawati, S.Si., M.Si. (.....)

Pembimbing 2 : Laila Qadrini, S.Si., M.Stat. (.....)

Penguji 1 : Hikmah, S.Pd., M.Sc. (.....)

Penguji 2 : Apriyanto, S.Pd., M.Sc. (.....)

Penguji 3 : Darma Ekawati, S.Pd., M.Sc. (.....)

ABSTRAK

Klasifikasi data sendiri merupakan proses mengasosiasikan karakteristik metadata ke setiap aset di kawasan digital, yang mengidentifikasi jenis data yang terkait dengan aset tersebut.s Dalam mengklasifikasikan data, ada banyak metode yang dapat digunakan salah satunya adalah *random forest*. Metode *random forest* adalah metode yang dapat mengatasi masalah non-linear, tahan terhadap *outlier*, *noise*, mudah digunakan serta memberikan hasil klasifikasi yang baik. *Imbalance* merupakan keadaan data dengan sebaran kelas yang tidak seimbang, dimana jumlah kelas yang satu lebih banyak atau sedikit dari kelas yang lain. Dalam kondisi *imbalance* sebagian besar klasifikasi bias terhadap kelas mayoritas, selain itu *imbalance* dapat menyebabkan *overfitting*, model yang kurang baik dan cukup berperan terhadap ternyadinya *misklasifikasi*. Untuk mengatasi masalah ini dapat dilakukan *resampling*. SMOTE merupakan turunan dari *resampling*, lebih tepatnya *oversampling* dimana data pada kelas minoritas akan ditambah sehingga setara dengan kelas mayoritas dengan cara membangkitkan data *synthetic*. Tujuan dari penelitian ini untuk mengetahui nilai akurasi klasifikasi data menggunakan *random forest* dan mengetahui hasil dari penerapan *resampling* dan *random forest* dalam klasifikasi. Data yang digunakan pada penelitian ini adalah data *breast cancer* dan data pasien BBLR Puskesmas Banggae I Kabupaten Majene. Hasil analisis diperoleh akurasi data *breast cancer* 94,74%, *sensitivity* 93,33% dan *F1-Score* 95,89%. Hasil akurasi data BBLR adalah 73,75%, *sensitivity* 77,63% dan *F1-Score* 84, 89%.

Kata Kunci: BBLR, *Resampling*, *Random Forest*, SMOTE, *Imbalance*

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pada dasarnya data merupakan sekumpulan informasi atau keterangan dari suatu hal yang diperoleh melalui pengamatan atau pencarian ke sumber-sumber tertentu. Data yang diperoleh namun belum diolah lebih lanjut dapat menjadi sebuah fakta atau anggapan. Klasifikasi data sendiri merupakan proses mengasosiasikan karakteristik metadata ke setiap aset di kawasan digital, yang mengidentifikasi jenis data yang terkait dengan aset tersebut. Dalam mengklasifikasikan data ada banyak metode yang dapat digunakan, diantaranya adalah *random forest*.

Metode *random forest* merupakan metode yang dapat mengatasi masalah non-linier. *Random forest* memiliki banyak keunggulan seperti tahan terhadap *outlier, noise* dan mudah digunakan, selain itu *random forest* dapat memberikan hasil klasifikasi yang baik dengan error yang rendah dan efektif mengatasi masalah *missing* data (Lestari & Sirodj, 2021). *Imbalance* merupakan keadaan data yang mempunyai sebaran kelas yang tidak seimbang, dimana jumlah kelas yang satu lebih banyak atau lebih sedikit dari jumlah kelas yang lain. Dalam kondisi *Imbalance* sebagian besar klasifikasi bias terhadap kelas mayoritas, dengan mesin klasifikassi lebih cenderung memprediksi kelas mayoritas dan mengabaikan kelas minoritas (Japkowick & Stephan, 2002). Menurut penelitian yang dilakukan oleh Gong dan Kim pada tahun 2017, ketidakseimbangan kelas dalam suatu dataset dapat berpotensi mengakibatkan overfitting dan pembentukan model prediksi yang buruk. Selain itu, dalam penelitian lain oleh Mellor dan koleganya pada 2015 juga disebutkan bahwa ketidakseimbangan kelas memiliki peran yang cukup signifikan terhadap terjadinya kesalahan klasifikasi oleh suatu model. Untuk mengatasi masalah ini dapat dilakukan *resampling*/sampling data (Fadilah, 2018).

Resampling secara luas digunakan untuk memecahkan masalah data yang tidak seimbang (*imbalance*) dengan mencoba menyeimbangkan data asli

berdasarkan serangkaian algoritma *sampling*, menyesuaikan jumlah sampel dalam kelas yang berbeda, kemudian melatih data "seimbang" baru dengan mengadopsi algoritma klasifikasi (Syukron & Subekti, 2018).

Pada penelitian sebelumnya oleh Muqiit, dkk (2020) dengan judul "Penerapan Metode Resampling dalam Mengatasi *Imbalanced* Data pada Determinan Kasus Diare pada Balita di Indonesia" yang dapat disimpulkan bahwa penerapan metode SMOTE sangat tepat digunakan untuk meningkatkan keakuratan analisis regresi logistik berganda serta dapat menghindari terjadinya *overfitting* pada data diare balita di Indonesia tahun 2017 yang memiliki karakteristik *imbalance*. Penelitian lainnya dengan judul "Klasifikasi Penipuan Transaksi Kartu Kredit Menggunakan Metode *Random Forest*" oleh (Lestari & Sirodj, 2021) mendapatkan hasil akurasi sebesar 97,275%, sensitivitas sebesar 98,795%, presisi sebesar 97,976%, *F-Measure* sebesar 98,384%, dan nilai AUC sebesar 94,065% yang termasuk dalam klasifikasi yang sangat baik karena dari keseluruhan hasil klasifikasi berada pada rentang 90-100%. Penelitian lainnya oleh Qadrini, dkk (2022) dengan judul "*Oversampling*, *Undersampling*, Smote SVM dan *Random Forest* pada Klasifikasi Penerima Bidikmisi Se Jawa Timur Tahun 2017" dengan kesimpulan bahwa penerapan *random sampling oversampling* dan SMOTE memberikan nilai AUC yang hampir sama dan dapat diterapkan untuk kasus data tak seimbang karena menyebabkan nilai akurasi, presisi, recall dan AUC yang tinggi, tidak *overfit* ataupun *underfit*. Berdasarkan pendahuluan dan penelitian-penelitian yang terdahulu maka pada penelitian ini, penulis mengangkat judul "Metode *Resampling* dan *Random Forest* untuk Klasifikasi Data".

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah pada penelitian ini adalah:

1. Bagaimana mengoptimalkan data menggunakan metode *resampling* dan *random forest* dalam klasifikasi data?
2. Bagaimana akurasi *random forest* untuk klasifikasi data?

3. Bagaimana hasil penerapan metode *resampling* dan *random forest* dalam klasifikasi ?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah maka tujuan penelitian ini adalah :

1. Mengetahui nilai optimal dari metode *resampling* dan *random forest* untuk klasifikasi data.
2. Mengetahui hasil akurasi klasifikasi data menggunakan *random forest*.
3. Mengetahui hasil dari penerapan metode *resampling* dan *random forest* dalam klasifikasi.

1.4 Manfaat Penelitian

Adapun manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut:

1. Menambah dan mencapai ilmu pengetahuan statistika yang berhubungan dengan metode resampling dan random forest serta menerapkannya untuk mengklasifikasi data.
2. Mengetahui bahwasanya *Random forest* dapat mencegah terjadinya *outlier* dan *noise* pada data dan *resampling* mengatasi bias, *overfitting*, dan *misklasifikasi*.
3. Diharapkan penelitian ini dapat menjadi bahan referensi dan rujukan pustaka untuk penelitian selanjutnya yang berkaitan dengan metode resampling dan metode random forest.

1.5 Batasan Masalah

Batasan masalah pada penelitian ini yaitu, data yang digunakan adalah data yang terdiri dari dua kelas atau data biner dan metode *resampling* yang digunakan adalah SMOTE.

BAB II
TINJAUAN PUSTAKA

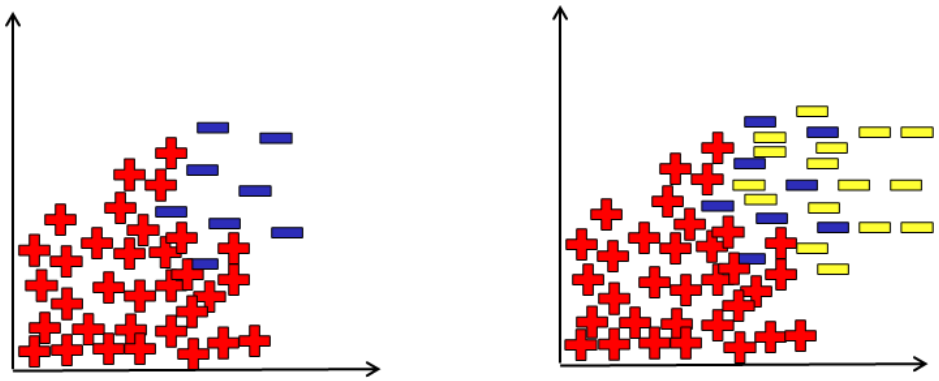
2.1 Resampling

Metode *resampling* adalah teknik pengambilan sampel ulang terhadap sampel yang sudah ada yang dilakukan secara acak dan bebas. Teknik ini memberikan peluang yang sama kepada sampel awal untuk diambil menjadi anggota sampel baru dengan ukuran lebih kecil atau lebih besar dari ukuran sampel awal. Metode *resampling* merupakan metode yang paling populer digunakan untuk mengatasi ketidakseimbangan kelas (*imbalance*). Terdapat dua jenis pendekatan pada kasus *imbalance*, diantaranya pendekatan pada level data dan level algoritma. Pada kasus *imbalance* ada beberapa permasalahan yang sering muncul yaitu (Choirunnisa, 2019):

1. *Outlier* yaitu ketika data yang bernilai ekstrim atau beda sangat jauh dengan mayoritas kelompoknya.
2. Banyak data antar kelas yang *overlap*. Apabila terdapat *overlapping*, maka *discriminative rule* akan sulit untuk diproses.
3. Terdapat beberapa data pad *sub-cluster* yang memiliki jarak terlalu rapat antar kedua kelas (*small disjunction*).

Teknik *resampling* secara umum, terbagi menjadi tiga yaitu *oversampling*, *undersampling* dan gabungan dari keduanya (*hybrid*).

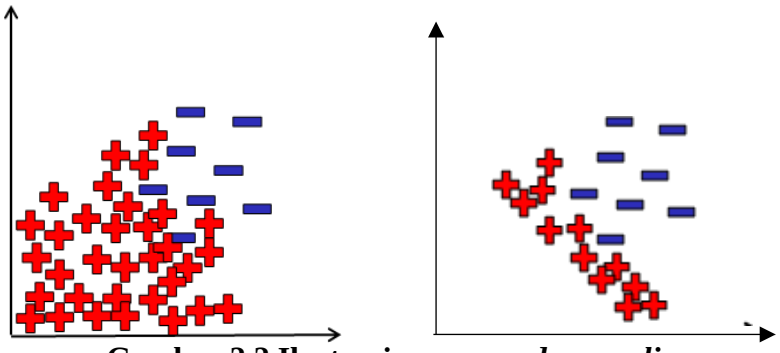
1. Metode *oversampling*, merupakan metode *sampling* dengan menambahkan jumlah data pada kelas minoritas sehingga dapat mengimbangi atau mendekati jumlah data pada kelas mayoritas. Proses pengambilan sampel dengan teknik *oversampling* ini adalah dengan menduplikasi kelas positif dan menyeimbangkan kelas secara acak. Namun, karena metode ini menduplikasi kelas positif yang ada di kelas minoritas, kemungkinan terjadinya *overfitting* (Rianto, 2015). Ilustrasi dari *Oversampling* dapat dilihat pada gambar 2.1



Gambar 2.1 Ilustrasi data menggunakan *oversampling*.

Sumber (Choirunnisa, 2019)

2. Metode *undersampling*, teknik *undersampling* merupakan proses sampling yang dilakukan dengan mengurangi atau mengeliminasi sebagian data pada kelas mayoritas pada data. Selama proses, kelas mayoritas akan dihapuskan sehingga didapat jumlah yang sama dengan kelas minoritas. Ilustrasi dari *Undersampling* dapat dilihat pada gambar 2.2.

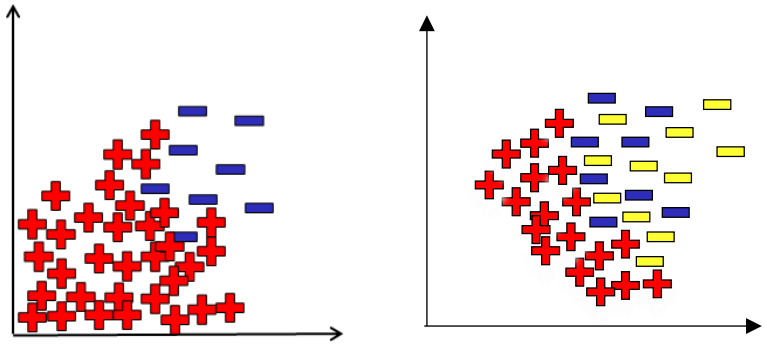


Gambar 2.2 Ilustrasi proses *undersampling*.

Sumber (Choirunnisa, 2019)

3. Metode *hybrid oversampling* dan *undersampling*, metode ini merupakan metode *blanching* dengan menggabungkan metode *undersampling* dan *oversampling*. Jumlah data pada kelas minoritas ditambahkan dengan metode *oversampling*, pada kelas mayoritas jumlah data akan dikurangi atau dibersihkan dari data *noise* menggunakan metode *undersampling*. Ilustrasi

dari metode *hybrid oversampling* dan *undersampling* dapat dilihat pada Gambar 2.3.



Gambar 2.3 Ilustrasi metode *hybrid oversampling* dan *undersampling*

2.1.1 SMOTE (*Synthetic Minority Over-sampling Technique*)

SMOTE (*Synthetic Minority Oversampling Technique*) merupakan turunan dari *oversampling* (Siringoringo, 2018). SMOTE merupakan metode pengelolaan data tidak seimbang yang diperkenalkan dan diusulkan pertama kali oleh Chawla dkk (2002). Ide dasar dari SMOTE untuk menambah jumlah sampel pada kelas minoritas agar setara dengan kelas mayoritas dengan cara membangkitkan data *synthetic* berdasarkan tetangga terdekat *k-nearest neighbour* dimana tetangga terdekat dipilih berdasarkan jarak *euclidean* antara kedua data (Chawla dkk, 2002). Misalkan diberikan data dengan p variabel yaitu $x^T = \left[x_1, x_2, ..., x_p \right]$ dan $z^T = \left[z_1, z_2, ..., z_p \right]$ maka jarak *euclidean* $d\left(x, z\right)$ secara umum sebagai berikut:

$$d\left(x, z\right) = \sqrt{\left(x_1 - z_1\right)^2 + \left(x_2 - z_2\right)^2 + ... + \left(x_p - z_p\right)^2} \tag{2.1}$$

Pembangkitan data *synthetic* dilakukan dengan menggunakan persamaan sebagai berikut:

$$X_{syn} = X_i + \left(X_{knn} - X_i\right)\gamma \tag{2.2}$$

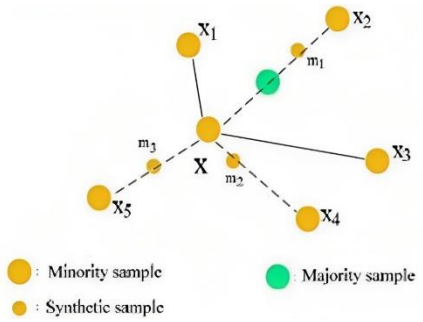
Dimana : X_{syn} adalah data *synthetic*

X_i adalah data ke-*i* dari kelas minor

X_{knn} adalah data dari kelas minor yang memiliki jarang terdekat dari X_i

γ adalah bilangan random antara 0 dan 1

Nilai X pada Gambar 2.4 menyatakan sebuah sampel dan X_1, X_2, X_3, X_4 dan X_5 adalah tetangga terdekat dari sampel X . SMOTE membangkitkan data baru (data *synthetic*) m_1, m_2 dan m_3 melalui sebuah garis di antara X dan masing-masing tetangga terdekatnya.



Gambar 2.4 Ilustrasi SMOTE

Sumber (Pangastuti, 2018)

2.2 *Machine Learning*

Pada tahun 1959, Arthur Samuel memperkenalkan istilah *machine learning* melalui jurnalnya yang berjudul "*Some Studies in Machine Learning Using the Game of Checkers*". Tujuannya adalah membuat komputer menjadi lebih baik dalam bermain catur dari dirinya. Pada tahun 1962 tujuannya tercapai, program buatannya dapat mengalahkan juara catur dari negara bagian *Connecticut* (Primartha, 2018).

Machine learning memerlukan sebuah model yang didefinisikan berdasarkan parameter-parameter tertentu. Proses learning adalah eksekusi program komputer untuk mengoptimasi parameter-parameter tersebut dengan memanfaatkan data *training* atau *past experience*. Jadi, *machine learning* secara sederhana merupakan pemrograman komputer untuk mencapai kriteria/performa tertentu dengan menggunakan sekumpulan data *training* atau *past experience* (Primartha, 2018).

Secara umum algoritma *machine learning* dapat dikelompokkan menjadi 4 kategori:

1. *Supervised learning*

   Secara keseluruhan *supervised learning* adalah pembelajaran dari contoh-contoh data yang diberikan label sebelumnya. *Supervised learning* membutuhkan data berlabel untuk dapat melakukan pelatihan data, yang disebut modelnya (Firmansyah, 2021). Terkait *supervised learning*, ada beberapa algoritma yang sudah dikembangkan yaitu, SVM, *naive bayes classifier*, *decision tree* dan sebagainya.

2. *Unsupervised learning*

   Unsupervised learning adalah tentang memodelkan data yang di input tanpa label. Dengan data yang cukup, dimungkinkan untuk menemukan pola dan struktur dari data. Ada dua permasalahan seputar *unsupervised learning* yaitu, *clustering* (pengelompokkan) dan *dimensionality reduction* (pengurangan dimensi). Ada banyak algoritma dalam *Unsupervised learning* antara lain, *hierarchical clustering*, k-Means dan lain sebagainya.

3. *Reinforcement learning*

   Reinforcement learning merupakan metode learning yang dipengaruhi oleh feedback dari lingkungan dengan teknik learning yang berulang-ulang dan menyesuaikan. Dalam *Reinforcement learning* ada empat algoritma yaitu, *genetic algorithm* (GA), *dynamic programming* (DP) *generalized policy iteration* (GIP) dan *monte carlo.*

4. *Deep learning*

   Deep learning merupakan metode learning yang memanfaatkan artificial *neural network* yang *multi layer* (berlapis-lapis) (Primartha, 2018). *Deep learning* memungkinkan model komputasi yang terdiri atas beberapa lapisan pemrosesan untuk mempelajari representasi data dengan berbagai tingkat abstraksi. Beberapa algoritma yang termasuk dalam kategori *deep learning* diantaranya, *convolutional network*, *restricted boltzmann machine* (RBM), *deep belief network* (DBN) dan *stacked autoencoders.*

2.2.1 Data *processing*

Data *processing* mengacu pada ekstraksi informasi melalui pengorganisasian, pengindeksan dan manipulasi data. Informasi di sini berarti hubungan dan pola berharga yang dapat membantu memecahkan masalah yang diminati. Sejalan dengan kemajuan teknologi, sejarah data *processing* dibagi menjadi tiga tahap: manual data *processing* (memproses data tanpa bantuan mesin), mekanik data *processing* (mengolah data dengan bantuan alat mekanik) dan elektronik data *processing* (mengolah data menggunakan komputer canggih) (Huang, 2019).

Berdasarkan jenis data dan pola minat, data *processing* dibagi menjadi beberapa metode, seperti (Huang, 2019):

1. *Classification*

   Merupakan metode yang menggunakan pengklasifikasian untuk memasukkan data yang tidak terklasifikasi kedalam kategori yang ada. Pengklasifikasi dilatih menggunakan data yang dikategorikan yang diberi label oleh para ahli, sehingga merupakan salah satu jenis pembelajaran terawasi dalam terminologi pembelajaran mesin. Klasifikasi bekerja dengan baik dengan data kategorik.
2. *Regression*

   Regression merupakan metode untuk mempelajari hubungan antara variabel dependen dan variabel independen lainnya. Hubungan tersebut dapat digunakan untuk memprediksi hasil di masa mendatang. Regresi biasanya menggunakan data numerik.
3. *Clustering*

   Clustering merupakan metode untuk menemukan kelompok data yang berbeda berdasarkan karakteristiknya. Perusahaan media sosial biasanya menggunakannya untuk mengidentifikasi orang-orang dengan minat yang sama. Ini adalah jenis pembelajaran tanpa pengawasan dan bekerja dengan data kualitatif dan kuantitatif.

4. *Association Rule Mining*

   Association Rule Mining merupakan metode untuk menemukan hubungan antar variabel. Awalnya, Metode ini dikembangkan untuk melakukan *market basket analysis*.

5. *Outlier Analysis*

   Outlier Analysis disebut juga deteksi anomali, adalah metode untuk menemukan item data yang berbeda dari sebagian besar data.

6. *Times Series Analysis*

   Times Series Analysis merupakan seperangkat metode untuk mendeteksi tren dan pola dari data deret waktu. Data deret waktu adalah kumpulan data yang diindeks dengan waktu, dan jenis data ini digunakan di berbagai domain.

2.2.2 Metode Klasifikasi

Klasifikasi adalah suatu bentuk dari analisis data yang mengekstraksi model untuk menggambarkan atau mengkategorikan kelas dari data. Dalam klasifikasi, pengklasifikasi atau model dibangun untuk memprediksi label kelas (kategorikal). Kategori-kategori ini dapat diwakili oleh nilai diskrit, urutan antara nilai tidak ada artinya. Klasifikasi terdiri antar dua proses, proses yang pertama adalah proses pembelajaran (proses pengklasifikasian dibangun) dan proses kedua adalah proses klasifikasi (model yang dibangun digunakan untuk memprediksi label dari data yang telah diberikan).

Klasifikasi memiliki dua kegunaan. Pertama, model deskriptif dimana model ini berperan sebagai alat penjelas untuk membedakan objek-objek dari kelas yang berbeda. Kedua, model prediktif dimana klasifikasi ini dapat digunakan untuk memprediksi label kelas dari *record* yang tidak diketahui.

Gambar 2.5 Klasifikasi sebagai pemetaan sebuah himpunan atribut input (x) ke dalam label kelasnya (y).

2.3 *Random Forest*

Random forest pertama kali diperkenalkan oleh Leo Breiman tahun 2001. *Random forest* merupakan modifikasi dari *bangging*. Pada *random forest* dilakukan penambahan pada *random sub sampling* atau pemilihan *m* variabel yang digunakan dalam membangun pohon. Proses pembentukan pohon dalam *random forest* tidak dilakukan pemangkasan (*pruning*).

Random forest merupakan suatu metode klasifikasi yang berisi koleksi dari pohon klasifikasi. Misalkan $\{h(x,\theta_k), k=1,\ldots\}$ dimana $\{\theta_k\}$ merupakan vektor random yang iid (*independent identically distributed*) dan tiap pohon memilih kelas yang paling banyak dari data (*majority vote*) (Breiman, 2001). Misalkan suatu *ensemble* $h_1(x), h_2(x), \ldots, h_k(x)$ dengan data *training* dipilih secara random dari distribusi vektor *random y* dan *x*, fungsi margin $(mg(x,y))$ dari *random forest* didefinisikan sebagai berikut (Breiman, 2001):

$$mg(x,y) = \frac{\sum_1^K I\left(h_k(x)=y\right)}{K} - \max_{j\neq y}\left[\frac{\sum_1^K I\left(h_k(x)=j\right)}{K}\right] \tag{2.3}$$

dimana I adalah fungsi indikasi dan K adalah banyaknya pohon. Fungsi margin digunakan untuk mengukur tingkat banyaknya jumlah vote pada x dan y rata-rata vote dari kelas yang lain.

Kekuatan (s = *strength*) adalah rata-rata ukuran kekuatan akurasi pohon tunggal. Nilai s yang semakin besar menunjukkan bahwa akurasi prediksi semakin baik. Nilai s didefinisikan sebagai berikut (Breiman, 2001):

$$s = E_{x,y} mg(x,y) \tag{2.4}$$

Rata-rata korelasi $\bar{\rho}$ antara pasangan dugaan dari dua pohon tunggal dalam *random forest* didefinisikan sebagai berikut (Breiman, 2001):

$$\bar{\rho} = \frac{E_{\theta,\theta'}\left(\rho(\theta,\theta')sd(\theta)sd(\theta')\right)}{E_{\theta,\theta'}\left(sd(\theta)sd(\theta')\right)} \tag{2.5}$$

dimana $\rho(\theta,\theta')sd(\theta)sd(\theta')$ merupakan korelasi antar pohon

Batas besarnya kesalahan prediksi (ε_{RF}) oleh *random forest* adalah:

$$\varepsilon_{RF} \le \bar{\rho}\left(\frac{1-s^2}{s^2}\right) \tag{2.6}$$

Dari persamaan tersebut dapat dikatakan bahwa untuk menghasilkan error yang kecil maka harus memiliki korelasi yang kecil dan memperkuat *strength*. Oleh karena itu perlu dilakukan modifikasi nilai *mtry* dan *ntree*. Dengan menurunkan nilai *mtry*, maka menurunkan pula korelasi dan *strength*. Hal yang sama berlaku untuk nilai *ntree*. Jika *ntree* besar berarti kemiripan data diantara setiap pohon sangat tinggi. Akan tetapi, jika pemilihan *mtry* dan *ntree* sangat rendah, mengartikan setiap pohon akan kehilangan beberapa informasi penting dan akan menaikkan nilai error. Sehingga pemilihan *mtry* dan *ntree* dalam *random forest* sangatlah berpengaruh. Menurut Breiman (1996), nilai *ntree* = 50 telah memberikan hasil klasifikasi yang memuaskan, lain halnya dengan Sutton (2005) menyarankan nilai *ntree* besar dari 100 karena dengan nilai *ntree* tersebut cenderung menghasilkan misklasifikasi yang konstan.

Berikut ini adalah algoritma *random forest*:

1. Buat suatu *bootstrap* sampel atau pengambilan sampel Z dengan pengembalian (*replacement*) dari suatu ukuran *N* dari gugus data.
2. Pilih *mtry* variabel secara random dari *p* variabel, dimana $m \le p$. Biasanya ukuran m terbaik dipilih melalui aproksimasi dari akar kuadrat dari total jumlah *p* variabel, yaitu $\left|\sqrt{p}\right|$. Menurut Leo Breiman, nilai *m* juga dapat diperoleh dari dua kali nilai akar kuadrat dari total jumlah *p* variabel $\left(m = 2\left|\sqrt{p}\right|\right)$ dan setengah dari nilai akar kuadrat dari total jumlah *p* variabel $\left(m = \frac{1}{2}\left|\sqrt{p}\right|\right)$.

3. Setelah dilakukan pemilihan *m* secara random, maka pohon ditumbuhkan tanpa pemangkasan. Pemecahan simpul terbaik dalam suatu pohon dilakukan dengan menggunakan indeks gini.
4. Langkah 1-3 dilakukan sebanyak *n* kali hingga terbentuk suatu *forest* (klasifikasi) sebanyak *n* pohon.
5. Setelah terbentuk *forest*, kemudian dicari nilai misklasifikasi error (*Out of Bag Error*) untuk mendapatkan *mtry* optimal dan diperoleh tingkat kepentingan variabel yang lebih stabil.
6. Untuk prediksi suatu kelas dilakukan dengan *majority vote* (suara terbanyak).

2.4 Uji Kebaikan

2.4.1 *Confusion Matrix*

Confusion matrix merupakan salah satu metode yang digunakan untuk melakukan perhitungan *accurasy* dan *error rate*. Dimana, *accurasy* merupakan perbandingan kasus yang diidentifikasi benar dengan jumlah seluruh kasus. *Error rate* merupakan kasus yang diidentifikasi salah dengan jumlah seluruh kasus (Alber, 2021). Melalui *confusion matrix*, keakuratan, tingkat kesalahan, ketepatan dan nilai penarikan dapat diketahui. Pada pengukuran kinerja menggunakan *confusion matrix*, terdapat empat istilah representasi hasil proses klasifikasi yaitu, *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), *False Negative* (FN). Nilai *True Positive* (TP) merupakan data positif yang terdeteksi benar, *True Negative* (TN) merupakan jumlah data negatif yang terdeteksi benar, sedangkan *False Positive* (FP) merupakan data negatif namun terdeteksi sebagai data positif dan *False Negative* (FN) merupakan kebalikan dari *true positive*, sehingga data positif, namun terdeteksi sebagai data negatif (Saifullah, 2019). Adapun *confusion matrix* untuk kelas biner, yaitu dataset dengan dua jenis kelas saja dapat dilihat pada tabel 2.1 (Siringoringo, 2018).

Tabel 2.1 *Confusion matrix*.

		Prediksi	
		Positif	Negatif
Aktual	Positif	TP	FN
	Negatif	FP	TN

Untuk mengetahui nilai akurasi dapat diperoleh dengan persamaan berikut: (Saifullah, 2019).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.7}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.8}$$

$$F1 - Score = 2 \times \frac{presisi \times recal}{presisi + recal} \tag{2.9}$$

Untuk mendapat nilai dari *F1-Score* kita membutuhkan nilai dari presisi dan recal. Berikut ini rumus yang dapat digunakan untuk menghitung presisi.

$$Precision = \frac{TP}{FP + TP} \tag{2.10}$$

Adapun *recall* merupakan rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. Untuk menghitung recall dapat menggunakan rumus berikut:

$$Recall = \frac{TP}{TP + FN} \tag{2.11}$$

dimana:

TP (*True Positive*): jumlah data positif yang terklasifikasi dengan benar.

TN (*True Negative*): jumlah data negatif yang terklasifiksi dengan benar.

FN (*False Negative*): jumlah data negatif namun terklasifikasi salah.

FP (*False positive*): jumlah data positif namun terklasifikasi salah.

Untuk menilai kebaikan model klasifikasi, nilai kategorinya dapat dilihat pada Tabel 2.2 (Gorunescu, 2011). Menurut Qadrini, dkk (2001) apabila akurasi model yang di dapatkan diatas 70%, dalam bidang data sains digolongkan dalam kinerja model yang cukup baik.

Tabel 2.2 Parameter Nilai Klasifikasi

Rentang	Klasifikasi
0,90 - 1,00	Klasifikasi sangat baik
0,80 - 0,90	Klasifikasi baik
0,70 - 0,80	Klasifikasi cukup
0,60 - 0,70	Klasifikasi buruk
0,50 - 0,60	Klasifikasi salah/gagal

2.5 Bayi Berat Lahir Rendah

Menurut Saifuddin (2009) BBLR adalah berat badan dari bayi baru lahir yang kurang dari 2500 gram. Sedangkan menurut Depkes RI (2009) BBLR adalah bayi yang lahir dengan berat lahir kurang dari 2500 gram tanpa memandang masa kehamilan.

Ciri-ciri BBLR menurut Manuaba (2006) adalah berat badan kurang dari 2.500 gram, lingkar dada kurang dari 30 cm, panjang badan kurang dari 45 cm, lingkar kepala kurang dari 33 cm, usia kehamilan kurang dari 37 minggu , kepala tidak mampu tegak, kepala relatif lebih besar, rambut lanugo banyak, kulit terlihat tipis transparan, lemak kulit kurang, otot hipotonik-lemah, pernapasan tidak teratur dapat terjadi apnea (gagal napas), frekuensi napas sekitar 45-50 kali per menit, frekuensi denyut nadi 100-140 kali per menit, sendi lutut/kaki fleksi-lurus, dan paha terlihat abduksi.

Faktor-faktor yang dapat menyebabkan terjadinya BBLR atau Bayi Berat Lahir Rendah adalah:

1. Umur ibu

   Usia dapat mempengaruhi kejadian BBLR dan memiliki peranan yang sangat penting terhadap kesehatan ibu hamil dan bayinya. Perencanaan kehamilan sebaiknya dilakukan antara usia 20-30 tahun (Setianingrum, 2005).

2. Paritas

   Paritas atau jumlah kelahiran adalah banyaknya kelahiran hidup yang dimiliki oleh seorang wanita. Paritas memiliki hubungan dengan BBLR disebabkan banyak pasangan yang tidak mengikuti program KB, sehingga jika paritas banyak akan beresiko prematur, BBLR dan kematian yang tinggi (Setiati & Rahayu, 2017).

3. Abortus

   Abortus adalah dikeluarkannya hasil konsepsi sebelum mampu hidup di luar rahim dengan berat badan kurang dari 1000 gram atau usia kehamilan kurang dari 28 minggu pada kehamilan sebelumnya (Manuaba, 2010). Riwayat abortus adalah riwayat keluarnya hasil konsepsi sebelum mampu hidup di luar kandungan dengan berat badan kurang dari 1000 gram atau usia kehamilan kurang dari 28 minggu pada kehamilan sebelumnya. Ibu yang mempunyai riwayat abortus 29,0% melahirkan bayi BBLR, sedangkan 12,9% tidak melahirkan bayi BBLR. Kejadian BBLR pada ibu yang memiliki riwayat abortus mempunyai risiko 1,79 kali lebih besar melahirkan bayi dengan berat badan lahir rendah dibandingkan dengan ibu tanpa riwayat abortus (Lestariningsih, 2014).

4. Gravida

   Menurut Manuaba (2008) *grandemultipara* dapat menyebabkan terjadinya BBLR, karena ibu yang melahirkan lebih dari 5 kali rentan mengalami anemia sehingga berdampak pada tumbuh kembang bayi dalam kandungan dan ibu dengan *grandemultipara* lebih rentan melahirkan bayi kurang bulan yang pasti memiliki berat badan bayi rendah.

DAFTAR PUSTAKA

Alber, J., 2021, Klasifikasi Data Mining Untuk Menentukan Tingkat Kepuasan Penggua Transaksi Bus Trans Metro Pekanbaru Menggunakan Metode *Naive Bayes*, *Skripsi*, Program Pasca Sarjana Teknik, Universitas Islam Riau, Pekanbaru.

Breiman, L., 1996, *Bagging Predictors, Machine Learning*, 24, 123-140.

Breiman, L., 2001, *Random Forest, Machine Learning*, 45, 5-32.

Chawla, N.V. dkk, 2002, SMOTE Boast: Improving Prediction Of The Minority Class In Boosting, Proc. Knowl, Discov, PP, Hal: 107-119.

Choirunnisa, S., 2019, Metode Hibrida Oversampling dan Undersampling Untuk Menangani Ketidakseimbangan Data Kegagalan Akademik Universitas XYZ, *Tesis*, Program Magister Komputer, Institut Teknologi Sepuluh Nopember, Surabaya.

Depkes RI, 2009, Pedoman Pelayanan Kesehatan Bayi berqat Lahir Rendah (BBLR) Dengan perawatan Metode Kanguru Di Rumah Sakit dan Jejaringannya, Jakarta: Bakti Husada.

Fadilah, L., 2018, Klasifikasi *Rarendomndom Forest* Pada Data *Imbalance*, *Skripsi*, Program Pasca Sarjana Matematika, UIN Syarif Hidayatullah, Jakarta.

Firmansyah, R., 2018, Implementasi *Deep Learning* Menggunakan *Convolusional Neural Network* Untuk Klasifikasi Bunga, *Skripsi*, Program Pasca Sarjana Sistem Imformassi, UIN Syarif Hidayatullah, Jakarta.

Gong, J. & Kim, H., 2017, RHBoost: Improving Classification Performance in Imbalance Data, *Computational Statistics and Data Analysis*, Vol. 105, Hal: 1-13

Gorunescu, F., 2011, *Data Mining: Concepts, Models, and Techniques,*

Huang, F., 2019, *Data Prosesing*.

Hutomo, A.W., 2020, Perbandingan Kinerja *AUC* dan G-Means pada *Machine Learning* Berbasis Seleksi Fitur Algoritma Genetik untuk Prediksi Cacat Perangkat Lunak, *Tesis*, Program Magister Komputer, UIN Syarif Hidayatullah, Jakarta.

Japkowicz, N. & Stephan, S., 2002, The Class Imbalance: A Systematic Study, *Intelligent Data Analysis*, No. 5, Vol. 6, Hal: 203-231.

Lestari, T.S. & Agustin Nuriani Sirodj, D., 2021, Klasifikasi Penipuan Transaksi Kartu Kredit Menggunakan Metode Random Forest, *Jurnal Riset Statistik*, No. 2, Volume 1, Hal: 160-167.

Lestariningsih, S. & Arta Budi Sisila, D., 2014, Hubungan Preeklasia Dalam Kehamilan Dengan Kejadian BBLR di RSUD Jenderal Ahmad Yani Kuta Metro Tahun 2011, Jurnal Kesehatan Masyarakat, No. 1, Vol. 8, Hal: 32-39.

Manuaba, I.A.C. dkk., 2010, *Ilmu Kebidanan, Penyakit kandungan dan KB*, EGC, Jakarta.

Manuaba, I.A.C., 2006, *Buku Ajar Patologi Obstetri: untuk Mahasiswa Kebidanan*, EGC, Jakarta.

Manuaba, I.A.C., 2008, *Gawat Darurat Obstetri Ginekologi Dan Obsetri Ginekologi Sosial Untuk Profesi Bidan*, EGC, Jakarta.

Mellor, A. dkk, 2015, Exploring Issues Of Training Data Imbalance And Mislabelling On Random Forest Performance For Large Area Land Cover Classification Using The Ensemble Margin, *Journal of Photogrammetry and Remote Sensing*, Vol. 105, Hal: 155-168.

Muqiit WS, A. dkk., 2020, Penerapan Metode Resampling Dalam Mengatasi *Imbalance* Data Pada Determinan Kasus Diare Pada Balita di Indonesia, *Jurnal Matematika dan Statistika Serta Aplikasinya*, No. 1, Vol. 8, Hal: 19-27.

Pangastuti, SP., 2018, Perbandingan Metode *Ensemble Random Forest* Dengan *Smote-Boosting* Dan *Smote-Bagging* Pada Klasifikasi Data Mining Untuk Kelas Imbalance (Studi Kasus : Data Beasiswa Bidikmisi Tahun 2017 di Jawa Timur), *Tesis*, Program Magister sains, Institut Teknologi Sepuluh Nopember, Surabaya.

Primartha, R., 2018, *Belajar Machine Learning Teori dan Praktek*, Informatika Bandung, Bandung.

Qadrini, L. dkk., 2022, *Oversampling*, *Undersampling*, SMOTE SVM dan *Random Forest* Pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017, No. 4, Vol.3, Hal: 386-391.

Qadrini, L. Seppewali, A, Aina, A. (2021). Decision Tree dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial, *Jurnal Inovasi Penelitian*. 2(7): 2722-9475.

Rianto, H., Wahono, R.S., 2015, *Resampling Logistik Regression* Untuk Penanganan Ketidakseimbangan *Class* Pada Prediksi Cacat *Software*, *Jurnal Of Software Engineering*, No. 1, Vol. 1, Hal: 46-53.

Saifuddin, A.B., 2009, Panduan Praktis Pelayanan Kesehatan Maternal dan Neonatal, Jakarta: EGC.

Saifullah, 2019, Deteksi Kelayakan Fisik Air Untuk Konsumsi Menggunakan *Naive Bayes Clasifier*, *Skripsi*, Program Pasca Sarjana Komputer, UIN Maulana Malik Ibrahim, Malang.

Setianingrum, S., 2005, Hubungan Antara Kenaikan Berat Badan, Lingkar Lengan Atas, Kadar Hemoglobin Ibu Hamil Trimester III Dengan Berat Bayi Lahir di Puskesmas Ampel Boyolali, Jurnal Semarang.3

Setiati, A.R. & Rahayu, S., 2017, Faktor Yang Mempengaruhi Kejadian BBLR (Berat Badan Lahir Rendah) Di Ruang perawatan Intensif Neonatus RSUD DR Moewardi Di Surakarta, *Jurnal Keperawatan Global*, No. 1, Vol. 2, Hal: 1-61.

Siringoringo, R., 2018, Klasifikasi Data Tidak Seimbang Menggunakan Algorotma SMOTE dan k-Nearest Neighbor, Jurnal ISD, No. 1, Vol. 3, Hal: 44-49.

Sutton, C.D., 2005, Classification and Regression Trees, Bagging, and Boosting, *Handbook of* Stati*s*tics, Vol. 24, Hal: 303-329.

Syukron, A. & Subekti, D., 2018, Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilian Kredit, *Jurnal Informatika*, No. 2, Vol. 5, Hal: 175-185.