Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 4 No. 4 (2020) 635 - 641 ISSN Media Elektronik: 2580-0760

Pengaruh Oversampling pada Klasifikasi Hipertensi dengan Algoritma Naïve Bayes, Decision Tree, dan Artificial Neural Network (ANN)

Nurul Chamidah[1], Mayanda Mega Santoni[2], Nurhafifah Matondang[3]
[1,2,3]Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta
[1]nurul.chamidah@upnvj.ac.id, [2]megasantoni@upnvj.ac.id, [3]nurhafifahmatondang@upnvj.ac.id

Abstract

Oversampling is a technique to balance the number of data records for each class by generating data with a small number of records in a class, so that the amount is balanced with data with a class with a large number of records. Oversampling in this study is applied to hypertension dataset where hypertensive class has a small number of records when compared to the number of records for non-hypertensive classes. This study aims to evaluate the effect of oversampling on the classification of hypertension dataset consisting of hypertensive and non-hypertensive classes by utilizing the Naïve Bayes, Decision Tree, and Artificial Neural Network (ANN) as well as finding the best model of the three algorithms. Evaluation of the use of oversampling on hypertension dataset is done by processing the data by imputing missing values, oversampling, and transforming data into the same range, then using the Naïve Bayes, Decision Tree, and ANN to build classification models. By dividing 80% of data as training data to build models and 20% as validation data for testing models, we had an increase in classification performance in the form of accuracy, precision, and recall of the oversampled data when compared without oversampling. The best performance in this study resulted in the highest accuracy using ANN with 0.91, precision 0.86 and recall 0.99.

Keywords: Oversampling, Hypertension, Naïve Bayes, Decision Tree, ANN

Abstrak

Oversampling merupakan teknik menyeimbangkan jumlah data dengan men-*generate* data dengan jumlah *record* yang sedikit pada suatu kelas, sehingga jumlahnya seimbang dengan data dengan kelas yang jumlah *record*-nya banyak. *Oversampling* pada penelitian ini diterapkan pada *dataset* hipertensi dimana kelas hipertensi memiliki jumlah *record* yang sedikit bila dibandingkan dengan jumlah *record* untuk kelas tidak hipertensi. Penelitian ini bertujuan untuk mengevaluasi pengaruh *oversampling* pada klasifikasi data hipertensi yang terdiri dari kelas hipertensi dan tidak hipertensi dengan memanfaatkan Algoritma *Naïve Bayes*, *Decision Tree*, dan *Artificial Neural Network* (ANN) sekaligus mencari model terbaik dari tiga algoritma tersebut. Evaluasi penggunaan *oversampling* pada data hipertensi ini dilakukan dengan memproses data dengan mengimputasi *missing value*, melakukan *oversampling*, dan mentransformasi data kedalam *range* yang sama, kemudian menggunakan algoritma *Naïve Bayes*, *Decision Tree*, dan ANN untuk membangun model klasifikasi. Dengan pembagian data 80% sebagai data training untuk membangun model dan 20% sebagai data validasi untuk menguji model, diperoleh peningkatan performa klasifikasi berupa akurasi, *precision*, dan *recall* pada data yang di-*oversampling* bila dibandingkan tanpa *oversampling*, Performa terbaik dalam penelitian ini menghasilkan akurasi tertinggi menggunakan algoritma ANN dengan 0.91, *precision* 0.86 dan *recall* 0.99.

Kata kunci: *Oversampling*, Hipertensi, *Naïve Bayes*, *Decision Tree*, ANN

1. Pendahuluan

Tekanan darah tinggi atau yang biasa dikenal dengan hipertensi merupakan salah satu penyakit tidak menular yang menjadi penyebab kematian nomor satu di dunia setiap tahun dan menjadi masalah kesehatan di Indonesia dan dunia karena merupakan faktor risiko penyakit seperti jantung, diabetes, gagal ginjal, dan stroke [1]. Pengecekan tekanan darah menjadi salah satu standar pada pelayanan kesehatan untuk mengetahui kondisi pasien mengingat pentingnya tekanan darah pada kondisi kesehatan.

Machine Learning (ML) merupakan sub bidang dari *Artificial Intelligence* (AI) yang memiliki tujuan untuk menangani dan mempelajari data dalam jumlah besar dimana data ini akan digunakan untuk membangun suatu model dengan mempelajari pola-pola pada data, yang diharapkan model tersebut dapat menangani kasus baru dari data yang telah dipelajari sebelumnya. Beberapa

Diterima Redaksi : 07-06-2020 | Selesai Revisi : 09-08-2020 | Diterbitkan Online : 20-08-2020

tipe-tipe model yang terdapat dalam *machine learning* diantaranya prediksi, *clustering* klasifikasi, dan *explanation*.

Kehadiran Kecerdasan Buatan (*Artificial Intelligence*) di industri kesehatan dengan cepat meningkatkan kualitas pelayanan kesehatan. Dua model kecerdasan buatan seperti *machine learning* dan *deep learning* dengan mudah dan cepat dalam mempelajari dan mengolah data untuk memperoleh informasi yang dibutuhkan.

Penelitian-penelitian yang memanfaatkan *machine learning* dalam bidang medis seperti, klasifikasi untuk mendiagnosa diabetes menggunakan metode *Bayesian Regularization Neural Network* (RBNN) [2], klasifikasi kelainan jantung menggunakan *Learning Vector Quantization* (LVQ) berdasarkan citra digital electrocardiogram [3], dan lain sebagainya.

Pemanfaatan *machine learning* untuk deteksi hipertensi juga telah banyak dilakukan seperti mengevaluasi faktor risiko penting yang digunakan untuk memprediksi hipertensi atau tidak hipertensi dengan *Artificial Neural Network* (ANN) menghasilkan akurasi 82% dari *dataset* berjumlah 185.371 [4]. Model prediksi untuk penyakit hipertensi menggunakan *regresi logistic* dan A*rtificial Neural Network* .dari *dataset* berjumlah 308.711 pasien menunjukkan akurasi 72% [5]. Selain ANN, *Logistic Regression*, *Naïve Bayes* juga dilakukan untuk mendeteksi hipertensi [6], [7]. Metode *Decision Tree* juga dapat digunakan untuk mengetahui variabel yang berpengaruh terhadap prediksi hipertensi [8], serta memprediksi risiko hipertensi [9].

Data pada kasus domain dunia nyata menghasilkan sejumlah besar data dengan distribusi kelas yang tidak seimbang yakni proporsi satu kelas memiliki rasio yang lebih tinggi daripada kelas lainnya. Kelas yang memiliki banyak *instance* disebut kelas mayoritas dan yang memiliki jumlah *instance* yang lebih sedikit disebut kelas minoritas [10]. Dalam situasi kehidupan nyata kadang-kadang kelas minoritas lebih menarik daripada kelas mayoritas, misal dalam data medis pada kasus gagal jantung [11] bidang ekonomi seperti *credit scoring* [12], *credit cards fraud*, dimana penyalahgunaan kartu kredit lebih sedikit daripada yang tidak disalah gunakan. serta bidang-bidang lain seperti deteksi *spam* pada email dimana email berupa *spam* lebih sedikit daripada bukan *spam*. Pada kasus-kasus ini, mendeteksi kelas minoritas menjadi lebih penting dari pada kelas mayoritas.

Algoritma klasifikasi tradisional sulit mengklasifikasikan kelas minoritas dengan benar, kondisi ini disebut masalah *imbalance class* atau ketidakseimbangan kelas. Masalah ketidakseimbangan kelas secara signifikan mempengaruhi kinerja dan menimbulkan tantangan serius untuk teknik *machine learning* [13]

Terdapat tiga metode yang umum digunakan dalam mengatasi *imbalance class* atau ketidakseimbangan kelas. Metode pertama dengan menyeimbangkan distribusi data, yakni dengan *undersampling* dan *oversampling*. Pada teknik *undersampling*, data yang mayoritas dikurangi sehingga jumlahnya sama dengan kelas minoritas, sedangkan *oversampling* men-*generate* data baru untuk kelas minoritas sehingga jumlahnya seimbang dengan kelas mayoritas. Metode kedua dengan menerapkan modifikasi pada algoritma, misal dengan memberikan pembobotan yang lebih besar pada kelas minoritas. Metode ketiga dengan menggabungkan metode yang menyeimbangkan distribusi data dan algoritma [14].

Teknik mengatasi *imbalance class* dengan menyeimbangkan distribusi data telah banyak dilakukan pada penelitian sebelumnya dan menghasilkan performa yang cukup baik pada algoritma *machine learning*. Penelitian menggunakan *Oversampling Adaptive Synthetic* (ADASYN) digunakan untuk menyeimbnagkan data hasil tes lab *pap smear* untuk kanker serviks [15], Prediksi *churn* pada *customer* dengan ADASYN dan *Backpropagation* [16], deteksi penyakit kardiovaskular dengan *oversampling* SMOTE (*Synthetic Minority Oversampling Technique*) dan ADASYN [17], dimana penelitian-penelitian tersebut menunjukkan peningkatan performa pada algoritma klasifikasi.

Berdasarkan deskripsi dan penelitian-penelitian sebelumnya tersebut, maka penelitian ini akan mengevaluasi penerapan *oversampling* untuk kasus data hipertensi dengan memanfaatkan algoritma *machine learning* untuk klasifikasi, yakni algoritma *Naïve Bayes*, *Decision Tree*, dan *Artificial Neural Network* (ANN) *backpropagation* kemudian membandingkannya dengan penerapan algoritma-algoritma tersebut ada data hipertensi tanpa *oversampling*.

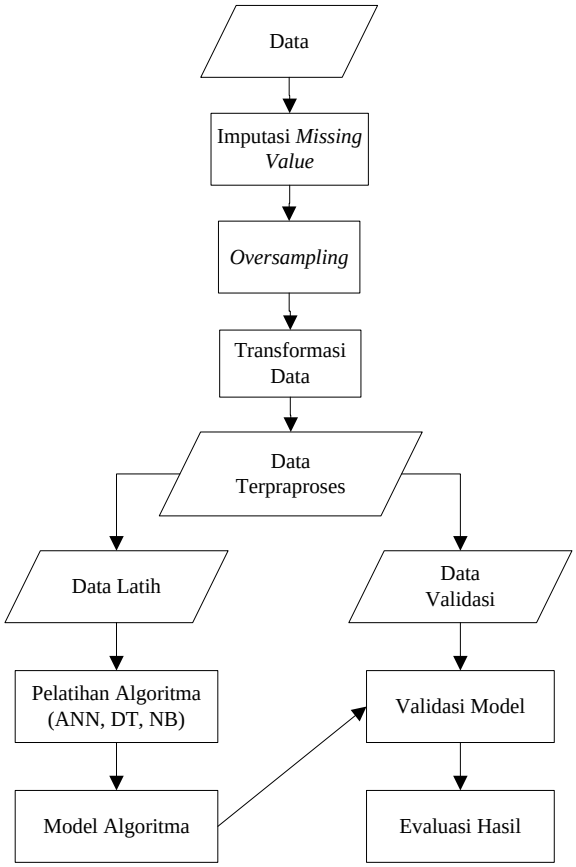
2. Metode Penelitian

Metodologi yang digunakan dalam penelitian ini dapat dilihat pada Gambar 1. Penelitian dimulai dengan mengumpulkan data berdasarkan kuesioner. Praproses data dengan imputasi *missing value*, kemudian dilakukan *oversampling* terhadap data yang telah diimputasi. Praproses selanjutnya adalah dengan transformasi data. Data yang telah di*oversampling* dan dipraproses, dilakukan pemodelan dengan algoritma klasifikasi dan diperoleh model klasifikasi. Model ini akan diujikan dengan testing data dan dilakukan evaluasi.

2.1. Data

Data diperoleh dari kuesioner dengan total responden sebanyak 274 orang. Pertanyaan yang digunakan sebanyak 26 dimana 25 pertanyaan digunakan sebagai fitur, dan 1 pertanyaan berupa kelas yang menyatakan seseorang memiliki riwayat menderita hipertensi atau tidak. Dari data tersebut diperoleh 40 *record* data berkelas hipertensi dan 234 *record* data tidak hipertensi.

Dua puluh lima fitur tersebut berasal dari 6 kelompok pertanyaan yang berupa identitas, gaya hidup, pola makan, kebiasaan istirahat, kebiasan merokok, dan riwayat kesehatan. Pemilihan pertanyaan yang dimanfaatkan menjadi fitur ini berdasarkan penelitian yang memanfaatkan kuesioner untuk memprediksi hipertensi [5][18].



Gambar 1. Metodologi Penelitian

Fitur identitas berupa jenis kelamin, usia, status pernikahan, tinggi badan, dan berat badan. Fitur gaya hidup berupa olahraga 30 menit perhari dan olahraga minimal 1 kali perminggu. Fitur pola makan berupa konsumsi daging <3 kali seminggu, konsumsi makanan berlemak tinggi <3 kali seminggu, konsumsi gorengan <3 kali seminggu, konsumsi makanan cepat saji <3 kali seminggu, konsumsi makanan yang diasinkan <3 kali seminggu, konsumsi sayuran >=3 kali seminggu, konsumsi buah-buahan >=3 kali seminggu, dan konsumsi mie instan > 2 bungkus seminggu.

Fitur kebiasaan istirahat berupa frekuensi terbangun <2 kali saat tidur malam, frekuensi susah tidur <2 kali dalam seminggu, tidur siang 1-2 jam >= 3 kali dalam seminggu, tidur teratur 6-8 jam pada malam hari. Fitur kebiasaan merokok berupa: kebiasaan merokok, merokok > 20 batang dalam sehari, kebiasaan minum minuman beralkohol. Fitur riwayat kesehatan berupa riwayat diabetes, riwayat kolesterol, dan hipertensi.

Tabel jumlah *missing value* dari setiap fitur dan distribusi data yang terdiri dari rata-rata, standar deviasi, minimum, dan maksimum dapat dilihat pada Tabel 1.

Tabel 1. Deskripsi Data

Fitur ke-	Jumlah *missing value*	Rata-rata	Stdev	Min	Max
1	6	34.18	10.59	16	63
2	2	1.56	0.50	1	2
3	0	1.44	0.55	1	3
4	2	161.86	11.75	70	268
5	5	65.72	15.62	39	175
6	0	1.76	0.43	1	2
7	0	1.27	0.44	1	2
8	1	1.43	0.50	1	2
9	0	1.51	0.50	1	2
10	1	1.28	0.45	1	2
11	1	1.46	0.50	1	2
12	0	1.59	0.49	1	2
13	2	1.66	0.47	1	2
14	0	1.09	0.29	1	2
15	0	1.25	0.43	1	2
16	0	1.75	0.43	1	2
17	0	1.51	0.50	1	2
18	0	1.65	0.48	1	2
19	1	1.66	0.48	1	2
20	0	1.36	0.48	1	2
21	1	1.87	0.33	1	2
22	1	1.98	0.15	1	2
23	2	1.83	0.38	1	2
24	0	1.94	0.24	1	2
25	0	1.81	0.39	1	2

2.2. Imputasi *Missing Value*

Imputasi *missing value* dilakukan untuk mengisi nilai yang kosong pada sampel data. Jumlah *missing value* pada setiap atribut dapat dilihat pada Tabel 1. Imputasi dilakukan dengan menggunakan *K-Nearest Neighbour* dengan jumlah K=1 atau 1-NN (1-*Nearest Neighbour*), yakni mengisi nilai yang kosong dengan mengambil nilai dari data lain yang paling mirip atau dekat dengan data sampel yang memiliki nilai kosong tersebut [19]. Formulasi 1-NN adalah:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2} \quad (1)$$

dengan $d(x_i, x_j)$ adalah jarak antar observasi target x_i dan observasi x_j, x_{ik} merupakan nilai pengamatan ke-k pada observasi target x_i, k=1, 2, … ,m, dan x_{jk} adalah nilai pengamatan ke-k pada observasi target x_j, k=1, 2, … ,m.

2.3. *Oversampling*

Oversampling merupakan teknik menyeimbangkan data dengan membangkitkan / men-*generate* data. Data yang digunakan pada penelitian ini tidak seimbang antara kelas tidak hipertensi dengan kelas hipertensi dengan perbandingan 234:40 atau 5,85:1 dimana kelas tidak hipertensi hampir 6 kali lebih besar dari kelas hipertensi, sehingga *oversampling* dilakukan agar perbandingan ini menjadi seimbang untuk dilakukan pembagian data.

Teknik *oversampling* yang digunakan adalah dengan ADASYN. ADASYN merupakan metode sampling pada *dataset* yang tidak seimbang jumlah sampel trainingnya. Metode ini diusulkan oleh [20].

2.4. Transformasi Data

Transformasi dalam penelitian ini adalah melakukan normalisasi data ke dalam rentang yang sama agar setiap fitur memiliki peran yang sama dalam menentukan hasil klasifikasi atau menghindari dominasi fitur tertentu. Teknik transformasi yang digunakan adalah transformasi min-max. Transformasi ini dilakukan pada setiap fitur secara terpisah ke dalam *range* 0-1. Berikut formula normalisasi min-max [21].

$$Xbaru = \frac{Xlama - \min}{maks - \min} \quad (2)$$

dengan Xbaru adalah nilai baru setelah transformasi, Xlama adalah nilai lama sebelum transformasi, min adalah nilai minimum pada fitur, dan maks adalah nilai maksimum pada fitur.

2.5. Pembagian Data

Data yang telah dipraproses dibagi menjadi data latih dan data validasi. Data latih digunakan untuk membangun model *Artificial Neural Network*, *Decision tree*, dan *Naïve Bayes* melalui proses pelatihan, sedangkan data validasi digunakan untuk memvalidasi model yang telah dibangun pada proses pelatihan sebelumnya. Data yang digunakan untuk pelatihan adalah 80% dan sisanya sebesar 20% digunakan untuk validasi. Pemilihan data untuk pelatihan dan validasi dilakukan secara acak.

2.6. *Artificial Neural Network*

Pelatihan dengan *Artificial Neural Network* dilakukan untuk memperoleh model ANN. Algoritma ANN yang digunakan adalah *feed forward backpropagation*. Arsitektur ANN yang digunakan dalam penelitian ini adalah jumlah input neuron adalah 25 sesuai dengan jumlah fitur, satu *hidden layer*, *hidden neuron* 50, dan satu *output neuron* untuk kelas hipertensi (1) atau tidak hipertensi (0), maksimum *epoch* 1000, *learning rate* 0.1, dan fungsi aktivasi *sigmoid*.

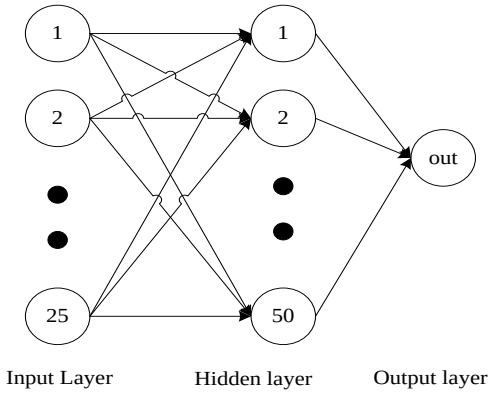
2.6. *Decision Tree*

Decision tree atau pohon keputusan merupakan salah satu algoritma dalam metode klasifikasi yang sering digunakan karena kesederhanaan cara kerjanya. Cara kerja dari algoritma ini adalah dengan mengubah fakta menjadi pohon keputusan yang merepresentasikan suatu rule/aturan, dimana rule ini nantinya akan dengan mudah diinterpretasikan oleh manusia.

Pada penelitian ini, *Gini Index* digunakan untuk mengevaluasi pemilihan *node*. *Gini index* dari tabel yang *pure* (terdiri dari 1 kelas) adalah nol, karena probabilitasnya = 1 dan 1-12 = 0. *Gini index* juga mencapai nilai maksimum ketika semua kelas memiliki probabilitas yang sama. Formulasi *gini index* dapat dilihat pada Formula 2

$$Gini(t) = 1 - \sum_{i=1}^{c}[p|t]^2 \quad (3)$$

dengan nilai c merupakan jumlah kelas pada tabel, dan p|t merupakan probabilitas kelas di dalam tabel.



Gambar 2. Arsitektur ANN

2.7. *Naïve Bayes*

Algoritma *Naïve Bayes* adalah probabilitas sederhana *classifier* yang menghitung satu set probabilitas dengan menghitung frekuensi dan kombinasi nilai dalam satu *set* data yang diberikan. Algoritma menggunakan teorema *Bayes* dan menganggap semua atribut bersifat independen terhadap nilai variabel kelas. Meskipun asumsi independensi antar variabel ini jarang terjadi, tapi *Naïve Bayes* berkinerja baik dan belajar dengan cepat pada berbagai masalah klasifikasi [22].

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (4)$$

Dengan X merupakan *evidence*, H merupakan hipotesis, $P(H|X)$ Probabilitas bahwa H benar untuk *evidence* X (*posterior probability*), $P(H)$ *prior probability* untuk hipotesis H, $P(X|H)$ probabilitas bahwa X benar untuk hipotesis H (*likelihood*), $P(X)$ Probabilitas untuk *evidence* X.

2.8. Validasi Model

Validasi dilakukan untuk menguji setiap model algoritma yang dibangun saat pelatihan. Validasi ini dilakukan dengan melakukan *testing* (pengujian) dengan data validasi, yakni data yang tidak digunakan pada saat pelatihan supaya hasil evaluasi lebih obyektif.

2.9 Evaluasi.

Hasil pengujian dari proses validasi selanjutnya akan dipetakan kedalam suatu *confussion matrix* yang dapat dilihat pada Tabel 2. Dimana TP (*True Positive)* merupakan jumlah kelas positif (hipertensi) diprediksi benar sebagai kelas positif (hipertensi), FP (*False Positive)* merupakan jumlah kelas negatif (tidak hipertensi) yang salah diprediksi sebagai kelas positif (hipertensi), TN (*True Negative)* adalah kelas realnya

negatif (tidak hipertensi) dan diprediksi benar sebagai kelas negatif (tidak hipertensi), FN (*False Negative)* merupakan jumlah kelas positif (hipertensi) yang salah diprediksi sebagai kelas negatif (tidak hipertensi).

Tabel 2. *Confussion Matrix*

		Kelas Prediksi	
		+	-
Kelas Real	+	TP	FN
	-	FP	TN

Dari *confussion matrix* tersebut selanjutnya dapat dievaluasi performa dari pengujian yakni dengan menghitung akurasi kinerja sistem, *precision*, dan *recall*. Akurasi merupakan rasio antara jumlah prediksi benar dari kelas hipertensi maupun tidak hipertensi dibandingkan terhadap seluruh data validasi. Rumus untuk akurasi dapat dihitung melalui persamaan (5), *Precision* merupakan rasio antara kelas hipertensi yang diprediksi dengan benar dibandingkan dengan hasil prediksi sistem terhadap kelas hipertensi baik benar (TP) maupun salah (FP). Formulasi untuk *precision* dapat dilihat pada persamaan (6). *Recall* merupakan suatu rasio antara kelas hipertensi yang diprediksi benar oleh sistem dibandingkan dengan kelas hipertensi yang sesungguhnya. Formulasi untuk *recall* yang dapat dilihat pada persamaan (7).

Performa kinerja sistem yakni akurasi, *precision* dan *recall* dapat dihitung melalui persamaan berikut:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

3. Hasil dan Pembahasan

Data hipertensi yang telah dipraproses dengan melakukan imputasi *missing value*, kemudian dilakukan *oversampling* dari kelas hipertensi sebanyak 40 data menjadi 234 data yakni sama dengan kelas tidak hipertensi, sehingga jumlah total data setelah *oversampling* untuk kelas hipertensi adalah 234 *record* dan kelas tidak hipertensi adalah 234 *record*, dengan total keseluruhan 468 *record* data. Tabel 3 menunjukkan distribusi data setelah dilakukan *oversampling*. Praproses berikutnya setelah *oversampling* adalah dengan melakukan transformasi data kedalam *range* 0-1 menggunakan metode min-max. Setelah praproses ini selesai, tahap selanjutnya adalah pembagian data.

Pembagian data dilakukan dengan tujuan memisahkan data latih dan data uji, yakni data latih sebesar 80% dan data uji sebesar 20% dari 468 *record*. Pembagian data dilakukan secara acak dengan menyamakan proporsi antara kelas hipertensi dan kelas tidak hipertensi. Pada pembagian ini diperoleh data training sebanyak 374 *record* yang terdiri dari 187 *record* kelas hipertensi dan 187 *record* kelas tidak hipertensi, serta data *testing* sebanyak 94 *record* yang terdiri dari 47 *record* kelas hipertensi dan 47 *record* kelas tidak hipertensi.

Tabel 3. Distribusi Data setelah *Oversampling*

fitur ke-	rata-rata	stdev	Min	max
1	36.10	10.86	16	63
2	1.48	0.48	1	2
3	1.34	0.51	1	3
4	162.35	11.39	70	268
5	68.39	14.72	39	175
6	1.81	0.37	1	2
7	1.27	0.42	1	2
8	1.48	0.47	1	2
9	1.47	0.47	1	2
10	1.32	0.43	1	2
11	1.51	0.47	1	2
12	1.55	0.47	1	2
13	1.61	0.45	1	2
14	1.09	0.27	1	2
15	1.27	0.41	1	2
16	1.76	0.40	1	2
17	1.45	0.47	1	2
18	1.65	0.45	1	2
19	1.70	0.43	1	2
20	1.29	0.43	1	2
21	1.86	0.33	1	2
22	1.99	0.11	1	2
23	1.82	0.36	1	2
24	1.94	0.23	1	2
25	1.65	0.44	1	2

Eksperimen dilakukan dengan melakukan pengulangan sebanyak 10 kali, dimana setiap pengulangan menggunakan data train dan tes yang sama untuk diuji dengan 3 algoritma ANN, *Decision Tree*, dan *Naïve Bayes*. Analisis dari eksperimen ini dibandingkan dengan hasil pengujian dengan metode dan pengulangan yang sama tapi tanpa melakukan *oversampling*.

Tabel 4 menunjukkan hasil perbandingan eksperimen yang dilakukan dengan *oversampling* dan tanpa *oversampling*. Jika hanya dievaluasi dari akurasi, maka akan terlihat tanpa *oversampling*, algoritma ANN, *Decision tree*, dan *Naïve Bayes* memiliki akurasi yang cukup baik yakni di atas 0.80, tapi mengingat data hipertensi ini imbalance, *precision* dan *recall*-nya sangat perlu untuk dievaluasi.

Tabel 4. Hasil Eksperimen

		ANN	*Decision tree*	*Naïve Bayes*
Dengan *Oversampling*	akurasi	0.91	0.86	0.71
	precision	0.86	0.86	0.72
	recall	0.99	0.87	0.70
Tanpa *Oversampling*	akurasi	0.84	0.82	0.85
	precision	0.44	0.36	0.00
	recall	0.35	0.33	0.00

Naïve Bayes tanpa *oversampling* tidak dapat mengklasifikasi kelas hipertensi sama sekali dengan nilai *precision* dan *recall* 0.00, sedangkan nilai akurasi 0.85 diperoleh dari semua data validasi yang diklasifikasikan ke dalam kelas tidak hipertensi. Setelah dilakukan *oversampling*, kemampuan *Naïve Bayes* dalam memprediksi hipertensi ini naik dengan *precision*

0.72dan *recall* 0.70, sedangkan akurasi keseluruhan adalah 0.71.

Decision Tree tanpa *oversampling* memiliki *precision* 0.36 dan *recall* 0.33 dalam memprediksi kelas hipertensi. Kemampuan algoritma ini dalam melakukan generalisasi masih tidak baik karena kelas hipertensi memiliki jumlah lebih sedikit dibandingkan kelas tidak hipertensi. Meskipun akurasi pada algoritma ini mencapai 0.82, namun ini tidak menunjukkan performa algoritma sesungguhnya karena mayoritas hasil klasifikasi dianggap sebagai kelas tidak hipertensi. *Oversampling* yang dilakukan pada penelitian ini sukses menaikkan akurasi, *precision*, dan *recall* dari *Decision tree* dimana akurasi dan *precision* mencapai 0.86 dan *Recall* 0.87.

ANN memiliki akurasi hingga 0.91 setelah di *oversampling* dari sebelum *oversampling* 0.84. Sementara itu, untuk nilai *precision* dan *recall* berturut-turut 0.86 dan 0.99. Nilai *recall* yang tinggi ini menunjukkan kemampuan ANN dalam menemukan kelas hipertensi cukup tinggi, yakni lebih tinggi dari *precision* 0.86. Hal ini menunjukkan bahwa kemampuan ANN untuk memprediksi kelas hipertensi cukup tinggi, tapi kesalahan terjadi karena kelas tidak hipertensi diprediksi sebagai kelas hipertensi (*false alarm*).

4. Kesimpulan

Pada penelitian ini, dilakukan *oversampling* pada kelas *imbalance* menggunakan ADASYN pada data hipertensi dengan menyeimbangkan kelas hipertensi dan kelas tidak hipertensi. Data yang telah di-*oversampling* diklasifikasi menggunakan algoritma *Naïve Bayes*, *Decision Tree*, dan *Artificial Neural Network* (ANN) dengan *Feed Forward Backpropagation*.

Berdasarkan hasil dan analisis yang dilakukan dalam penelitian ini, *oversampling* dengan ADASYN dapat meningkatkan kemampuan algoritma dalam mengklasifikasi kelas hipertensi secara signifikan. *Oversampling* men-*generate* data kelas tidak hipertensi dari 40 *record* data menjadi 234 dan dievaluasi dengan menggunakan tiga pemodelan yakni *Naïve Bayes*, *Decision Tree*, dan ANN. Eksperimen yang dilakukan menghasilkan akurasi, *precision*, dan *recall* terbaik dengan Algoritma ANN yakni akurasi 0.91, *Precision* 0.86, dan *recall* mencapai 0.99 yang berarti hampir semua kelas hipertensi dapat dideteksi.

Penelitian ini menunjukkan bahwa *oversampling* efektif digunakan untuk menaikkan performa secara signifikan pada algoritma *Naïve Bayes*, *Decision Tree*, dan ANN pada kasus kelas imbalance bila dibandingkan dengan tanpa *oversampling*,

Penggunaan teknik *oversampling* ini cukup menjanjikan untuk dimanfaatkan dalam berbagai bidang meskipun besarnya data setelah *oversampling* akan menjadi masalah komputasi berikutnya jika data yang di *oversampling* berukuran besar. Meskipun demikian, teknik-teknik lain dapat digunakan misal dengan seleksi fitur untuk menentukan fitur mana saja yang berpengaruh pada performa klasifikasi dan mana yang tidak berpengaruh atau menurunkan performa, sehingga penelitian di bidang ini masih dapat terus dikembangkan.

Ucapan Terimakasih

Terima kasih kepada Universitas Pembangunan Nasional Veteran Jakarta yang telah mendanai penelitian ini.

Daftar Rujukan

[1] K. K. R. Indonesia, "Hipertensi Penyakit Paling Banyak Diidap Masyarakat," 2019. [Online]. Available: https://www.kemkes.go.id/article/view/19051700002/hipertensi-penyakit-paling-banyak-diidap-masyarakat.html. [Accessed: 04-Jun-2020].

[2] M. F. Rahman, M. Ilham Darmawidjadja, and D. Alamsah, "KLASIFIKASI UNTUK DIAGNOSA DIABETES MENGGUNAKAN METODE BAYESIAN REGULARIZATION NEURAL NETWORK (RBNN)," 2017.

[3] Y. Arum Sari and A. Arwan, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes Human Detection and Tracking View project Smart Nutrition Box View project," 2018.

[4] D. Lafreniere, F. Zulkernine, D. Barber, and K. Martin, "Using machine learning to predict hypertension from a clinical dataset," in *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*, 2017.

[5] A. Wang, N. An, G. Chen, L. Li, and G. Alterovitz, "Predicting hypertension without measurement: A non-invasive, questionnaire-based approach," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 7601–7609, Jun. 2015.

[6] B. O. Afeni, T. I. Aruleba, and I. A. Oloyede, "Hypertension Prediction System Using Naive Bayes Classifier," *J. Adv. Math. Comput. Sci.*, pp. 1–11, Sep. 2017.

[7] M. K. Kanwar *et al.*, "Risk stratification in pulmonary arterial hypertension using Bayesian analysis," *Eur. Respir. J.*, p. 2000008, May 2020.

[8] M. Tayefi *et al.*, "The application of a decision tree to establish the parameters associated with hypertension," *Comput. Methods Programs Biomed.*, vol. 139, pp. 83–91, Feb. 2017.

[9] I. A.-A. J. of M. and Computer and undefined 2017, "Predictive Model for the Classification of Hypertension Risk Using Decision Trees Algorithm," *academia.edu*.

[10] O. M. Olaitan and H. L. Viktor, "SCUT-DS: Learning from multi-class imbalanced canadian weather data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11177 LNAI, pp. 291–301.

[11] M. Khaldy, & C. K.-I. R., and undefined 2018, "Resampling imbalanced class and the effectiveness of feature selection methods for heart failure dataset," *pdfs.semanticscholar.org*.

[12] O. Heranova, "Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 443–450, Dec. 2019.

[13] S. Maheshwari, J. Agrawal, and S. Sharma, "A New approach for Classification of Highly Imbalanced Datasets using Evolutionary Algorithms," *Int. J. Sci. Eng. Res.*, vol. 2, no. 7, 2011.

[14] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," in *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, 2017, vol.

2017-January, pp. 79–85.

[15] Y. E. Kurniawati, A. E. Permanasari, and S. Fauziati, "Adaptive Synthetic-Nominal (ADASYN-N) and Adaptive Synthetic-KNN (ADASYN-KNN) for Multiclass Imbalance Learning on Laboratory Test Data," in *Proceedings - 2018 4th International Conference on Science and Technology, ICST 2018*, 2018.

[16] A. Aditsania, Adiwijaya, and A. L. Saonard, "Handling imbalanced data in churn prediction using ADASYN and backpropagation algorithm," in *Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017*, 2017, vol. 2018-January, pp. 533–536.

[17] J. L. P. Lima, D. MacEdo, and C. Zanchettin, "Heartbeat Anomaly Detection using Adversarial Oversampling," in *Proceedings of the International Joint Conference on Neural Networks*, 2019, vol. 2019-July.

[18] R. K. Sari and L. PH, "FAKTOR- FAKTOR YANG MEMPENGARUHI HIPERTENSI," *J. Ilm. Permas J. Ilm. STIKES Kendal*, vol. 6, no. 1, pp. 1–10, 2016.

[19] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *J. Syst. Softw.*, vol. 85, no. 11, pp. 2541–2552, Nov. 2012.

[20] H. He, H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *IEEE Int. Jt. Conf. NEURAL NETWORKS (IEEE WORLD Congr. Comput. Intell. IJCNN 2008*, pp. 1322--1328, 2008.

[21] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[22] M. M. Saritas and A. Yasar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, no. 2, pp. 88–91, Jun. 2019.