Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 890−900
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5652

Implementasi Metode *Resampling* Dalam Menangani *Data Imbalance* Pada Klasifikasi *Multiclass* Penyakit *Thyroid*

Najmi Cahaya Nugraha*, Hanny Hikmayanti, Jamaludin Indra, Ayu Ratna Juwita

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Buana Perjuangan Karawang, Karawang, Indonesia
Email: if20.najminugraha@mhs.ubpkarawang.ac.id, hanny.hikmayanti@ubpkarawang.ac.id, jamaludin.indra@ubpkarawang.ac.id, ayurj@ubpkarawang.ac.id
Email Penulis Korespondensi: if20.najminugraha@mhs.ubpkarawang.ac.id
Submitted: **22/08/2024**; Accepted: **09/09/2024**; Published**: 09/09/2024**

Abstrak−Diperkirakan setidaknya 17 juta orang Indonesia mengalami gangguan tiroid. Menariknya, hampir 60% dari mereka yang hidup dengan gangguan tiroid tidak mendapatkan diagnosis. Dengan demikian, perlu dilakukan penelitian yang mengaplikasikan metode-metode untuk memprediksi penyakit tiroid. Sebelum menggunakan metode prediksi, penting untuk menerapkan metode klasifikasi agar dapat memperoleh model prediksi yang akurat. Namun, agar hasil klasifikasi optimal dan untuk menghindari ketidakakuratan, diperlukan keseimbangan dalam data yang digunakan. *Data Imbalance* merupakan kondisi di mana rasio antara kelas pada data tidak seimbang yang dapat mengakibatkan model yang dihasilkan menjadi bias. Tujuan utama penelitian untuk menyajikan solusi yang dapat meningkatkan akurasi deteksi dini penyakit thyroid melalui penanganan ketidakseimbangan data dan penerapan algoritma klasifikasi yang tepat. Metode penelitian dimulai dengan pengumpulan dan analisis *dataset* sebanyak 9172 data dan dilakukan tahapan *preprocessing* sehingga dapat dibagi menjadi 5321 data latih dan 1331 data uji. Tahapan pengujian menggunakan 7 algoritma klasifikasi dengan 7 data *resampling* berbeda serta evaluasi menggunakan *confusion matrix*. *Penelitian* ini menghasilkan tingkat akurasi tertinggi yang diperoleh dari kombinasi Algoritma *Random Forest* dengan metode *Random Over Sampling* sebesar 98%. Dapat disimpulkan bahwa kombinasi Algoritma *Random* Forest dengan metode *resampling Random Over Sampling* dapat meningkatkan akurasi deteksi dini pada penyakit tiroid.

Kata Kunci: Klasifikasi, ketidakseimbangan data, metode *resampling*, tiroid

Abstract−It is estimated that at least 17 million Indonesians suffer from thyroid disorders. Interestingly, nearly 60% of those living with a thyroid disorder do not receive a diagnosis. Thus, it is necessary to carry out research that applies methods to predict thyroid disease. Before applying prediction methods, it is crucial to implement classification methods to obtain an accurate prediction model. However, to achieve optimal classification results and to avoid inaccuracies, a balance in the used data is required. Data imbalance is a condition where the ratio between classes in the data is uneven, which can result in the generated model becoming biased. The main objective of the research is to present a solution that can improve the accuracy of early detection of thyroid diseases through addressing data imbalance and implementing appropriate classification algorithms. The research methodology began with the collection and analysis of a dataset consisting of 9172 data points. Preprocessing was then performed, resulting in 5321 training data points and 1331 test data points. The testing phase employed 7 different classification algorithms with 7 different resampling methods and evaluation using a confusion matrix. This research achieved the highest accuracy rate of 98%, obtained from the combination of the Random Forest Algorithm and the Random Over Sampling method. It can be concluded that the combination of the Random Forest Algorithm with the Random Over Sampling resampling method can improve early detection accuracy for thyroid diseases.

Keywords: Classification,data imbalance, resampling method, thyroid

1. PENDAHULUAN

Kelenjar yang terletak di bawah jakun pada leher bagian depan biasa disebut kelenjar gondok/tiroid (*thyroidgland*) [1]. Kelenjar tiroid menghasilkan hormon yang dapat meningkatkan penggunaan oksigen serta mengatur metabolisme pada tubuh. Tergantung pada penyebabnya, gangguan kelenjar dan hormon yang diakibatkannya menyebabkan gejala penyakit tiroid yang berbeda-beda. Penyakit tiroid muncul ketika kelenjar tiroid mengalami perubahan dan menghasilkan hormon tiroid dalam jumlah yang kurang atau berlebihan [2].

Pada peringatan Pekan Kesadaran Tiroid Internasional (ITAW) ke-12, yang berlangsung dari 25 hingga 31 Mei 2020, diperkirakan setidaknya 17 juta orang Indonesia mengalami gangguan tiroid. Menariknya, hampir 60% dari mereka yang hidup dengan gangguan tiroid tidak mendapatkan diagnosis [3]. Gangguan fungsi tiroid seringkali sulit diidentifikasi karena gejalanya mirip dengan keluhan akibat gaya hidup modern. Oleh karena itu, gejala tersebut sering diabaikan . Akibatnya, banyak pasien yang tidak menyadari adanya masalah dalam tubuh mereka dan tidak melakukan pemeriksaan medis [4]. Dengan demikian, perlu dilakukan penelitian yang mengaplikasikan metode-metode untuk memprediksi penyakit tiroid. Hal ini akan membantu memudahkan pasien dalam mendiagnosis dan mendeteksi dini gangguan tiroid.

Sebelum menggunakan metode prediksi, penting untuk menerapkan metode klasifikasi agar dapat memperoleh model prediksi yang akurat [5]. Namun, agar hasil klasifikasi optimal dan untuk menghindari ketidakakuratan, diperlukan keseimbangan dalam data yang digunakan [6]. Data dianggap tidak seimbang apabila rasio objek pada suatu kelas data lebih dominan dibandingkan dengan kelas yang lain [7]. Diperlukan tindakan *resampling* dalam mengklasifikasikan data antara kelas minoritas dan mayoritas agar kinerja algoritma meningkat dan distribusi kelas menjadi lebih seimbang secara proporsional [8].

Seperti pada penelitian sebelumnya, ketidakseimbangan kelas pada prediksi penyakit jantung diatasi menggunakan metode SMOTE yang kemudian dilakukan pemodelan menggunakan 8 algoritma. Data yang digunakan

Copyright © 2024 **Authors**, Page 890
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 890−900
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5652

dalam penelitian tersebut merupakan *dataset Heart Failure Prediction* yang terdiri dari 12 atribut utama dan 1 atribut kelas. Hasil pada penelitian tersebut, menunjukkan bahwa metode SMOTE dapat meningkatkan performa akurasi dari beberapa algoritma kecuali SVM dan *Adaboost*. Beberapa algoritma dengan peningkatan performa akurasi diantaranya C45, *Naïve Bayes*, *Random Forest*, *Neural Network*, KNN dan *Bagging* dimana *Random Forest* memiliki tingkat akurasi terbaik [9].

Pada penelitian untuk klasifikasi serangan jantung miokarditis, proporsi data yang digunakan tidak seimbang sehingga dilakukan proses *resampling* menggunakan metode *oversampling*. Penelitian tersebut menggunakan *dataset* sebanyak 1700 data dengan 124 fitur yang bersumber dari *Machine Learning Repository*. Hasil dari penelitian tersebut, metode SMOTE yang merupakan salah satu dari metode *oversampling* mampu meningkatkan tingkat akurasi prediksi dengan Algoritma Decision Tree, SVM dan Naïve Bayes [10].

Adapun penelitian serupa yang bertujuan untuk mengevaluasi kinerja 2 algoritma dalam memprediksi penyakit diabetes pada *dataset* yang tidak seimbang, diatasi dengan menggunakan metode *resampling* SMOTE. Penelitian tersebut menggunakan algoritma KNN dan Naïve Bayes dalam memprediksi penyakit diabetes. Hasil dari penelitian tersebut, metode *resampling* SMOTE hanya meningkatkan tingkat akurasi pada Algoritma Naïve Bayes sedangkan tingkat akurasi Algoritma KNN cenderung menurun[11].

Berikutnya, penelitian mengenai ketidakseimbangan data dilakukan dalam proses klasifikasi penyakit jantung tipe kardiovaskular. Penelitian tersebut menggunakan Algoritma *Extreme Gradiant Boosting* dengan kombinasi metode *Adaptive Synthetic Sampling* (ADASYN) dalam proses *resampling*-nya. *Heart Disease Dataset* yang bersumber dari website Kaggle digunakan dalam penelitian tersebut sebanyak 4238 *record* dengan 16 atribut. Penelitian tersebut mengungkapkan bahwa model yang menggunakan ADASYN mencapai nilai ROC-AUC sebesar 0.971 dan akurasi sebesar 0.916. Sebaliknya, model tanpa ADASYN hanya memperoleh nilai ROC-AUC sebesar 0.698 dan akurasi sebesar 0.841. Hal ini menunjukkan bahwa model XGBoost yang dikombinasikan dengan ADASYN memiliki kinerja yang lebih unggul [12].

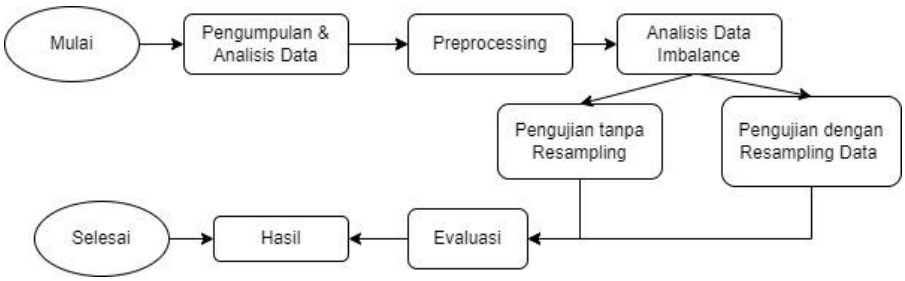
Selanjutnya, penelitian mengenai teknik *resampling data* juga dilakukan dalam klasifikasi risiko kesehatan ibu hamil menggunakan metode *random over sampling*. Penelitian ini menggunakan Algoritma *Ranfom Forest* dengan kombinasi metode *random over sampling* dan metode PSO. Data yang digunakan merupakan *dataset* dari UCI *Machine Learning* sebanyak 1.014 data dengan 6 atribut dan 1 label. Hasil dari penelitian tersebut menunjukkan bahwa klasifikasi menggunakan Algoritma *Random Forest* dengan kombinasi metode *random oversampling* meningkatkan tingkat akurasi sebesar 1.09% [13].

Dalam upaya mengembangkan model klasifikasi untuk deteksi dini penyakit thyroid, terdapat permasalahan ketidakseimbangan data yang signifikan dalam *dataset* yang digunakan sehingga dapat mempengaruhi tingkat akurasi model. Seperti pada beberapa penelitian sebelumnya, rumusan permasalahan utama melibatkan penanganan ketidakseimbangan data untuk meningkatkan kinerja model dengan mengeksplorasi metode-metode *resampling* untuk mengatasi ketidakseimbangan tersebut.

Tujuh algoritma klasifikasi yang berbeda akan diuji untuk mengevaluasi keefektifan masing-masing algoritma dalam menangani ketidakseimbangan data. Algoritma-algoritma tersebut melibatkan Algoritma *Support Vector Machine*, *K-Nearest Neighbor*, *Random Forest*, *Decision Tree*, *Naïve Bayes*, *Logistik Regression* dan *Neural Network*. Dengan demikian, tujuan utama penelitian ini adalah menyajikan solusi yang dapat meningkatkan akurasi deteksi dini penyakit thyroid melalui penanganan ketidakseimbangan data dan penerapan algoritma klasifikasi yang tepat.

2. METODOLOGI PENELITIAN

Penelitian ini bertujuan untuk menyajikan solusi yang dapat meningkatkan akurasi deteksi dini penyakit tiroid melalui penanganan ketidakseimbangan data dan penerapan algoritma klasifikasi yang tepat. Berikut alur penelitian dan metode algoritma yang akan diuji seperti pada Gambar 1 berikut:



Gambar 1. Tahapan penelitian

2.1 Pengumpulan dan Analisis Data

Tyroid Disease Data yang diperoleh dari *platform Kaggle* merupakan data rekam medis demografi pasien dan hasil tes darah bersamaan dengan diagnosis penyakit tiroid [14]. Sebanyak 9172 data yang terdiri dari 30 atribut utama yaitu: *age*, *sex*, *on_thyroxine*, *query_on_thyroxine*, *on_antithyroid_meds*, *sick*, *pregnant*, *thyroid_surgery*,

Copyright © 2024 **Authors**, Page 891
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 890−900
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5652

I131_treatment, query_hypothyroid, query_hyperthyroid, lithium, goitre, tumor, hypopituitary, psych, TSH_measured, TSH, T3_measured, T3, TT4_measured, TT4, T4U_measured, T4U, FTI_measured, FTI, TBG_measured, TBG, referral_source, patient_id dan 1 atribut *multiclass* yaitu Target.

2.2 *Preprocessing*

Tahapan ini terdiri dari beragam operasi untuk membersihkan dan mentransformasi data yang telah disiapkan untuk langkah pengujian berikutnya.

a. Eliminasi *missing value* menggunakan KNN *Imputer*
KNN *Imputer* digunakan untuk mengeliminasi nilai kosong atau nilai yang tidak tersedia dalam dataset.

$$\hat{x}_{i,m} = \frac{\sum_{j=1}^{k} \omega_{i.j} x_{j,m}}{\sum_{j=1}^{k} w_{i,j}} \quad (1)$$

b. Eliminasi duplikasi data
Duplikasi data merujuk pada adanya salinan identik atau serupa dari informasi dalam satu set data. Eliminasi duplikasi data merupakan langkah penting dalam analisis data untuk memastikan integritas dan akurasi data serta mencegah kesalahan interpretasi hasil analisis [15].

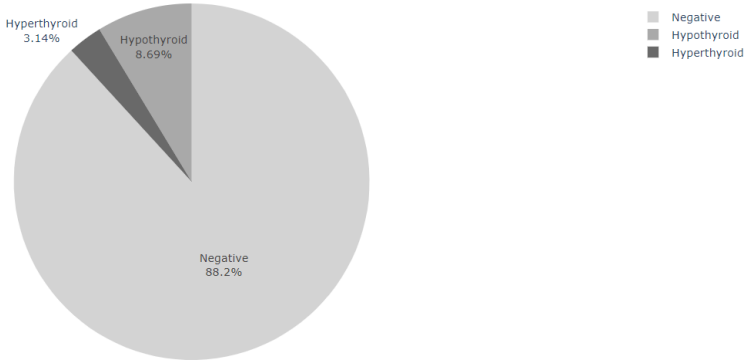
c. Eliminasi *outlier*
Outlier adalah suatu nilai yang jauh berbeda dari sebagian besar data dalam satu set data.
IQR = Q3 - Q1
Batas Bawah = Q1 - *k* . IQR (2)
Hitung Batas Atas = Q3 + *k* . *IQR*
Identifikasi dan hapus *outlier* dengan mengeliminasi nilai yang kurang dari Batas Bawah atau lebih dari Batas Atas.

d. Seleksi fitur dengan menentukan korelasi antar variabel menggunakan visualisasi matriks korelasi
Matriks korelasi mengukur sejauh mana dua variabel berkorelasi atau memiliki hubungan linier satu sama lain. Korelasi dapat berkisar dari -1 hingga 1, dengan interpretasi sebagai berikut:
1: Korelasi positif sempurna. Jika satu variabel naik, variabel lain juga naik secara proporsional.
0: Tidak ada korelasi. Variabel tidak memiliki hubungan linier satu sama lain.
-1: Korelasi negatif sempurna. Jika satu variabel naik, variabel lain turun secara proporsional

e. Transformasi dan Split Data
Pada penelitian ini, transformasi data digunakan dalam mengubah format target diagnosis pasien yang cukup banyak menjadi 3 kategori utama yaitu *Negative, Hypothyroid dan Hyperthyroid.* Proses pembagian data dalam penelitian ini melibatkan membagi dataset menjadi beberapa bagian untuk tujuan melatih dan menguji model klasifikasi.

2.3 Analisis Data Imbalance

Ketidakseimbangan data terjadi ketika rasio objek suatu kelas data lebih banyak dibandingkan dengan kelas lain. *Thyroid Disease Dataset* yang telah melewati tahapan *preprocessing* memiliki 3 kelas utama yaitu kelas *negative*, *hipothyroid* dan *hyperthyroid.* Pada Gambar 2, kelas *negative* pada *dataset* ini menjadi kelas mayor dimana rasio objek nya lebih banyak dibanding 2 kelas lainnya.



Gambar 2. Pie plot jumlah kelas

2.4 Pengujian

Metode *oversampling*, *undersampling* dan *hybrid* digunakan dalam penelitian ini untuk mengeksplorasi metode *resampling* yang lebih mempengaruhi tingkat akurasi suatu algoritma klasifikasi.

a. *Oversampling*
1. Random Over Sampling: menambahkan salinan acak dari sampel kelas minoritas [16].
Nbaru = (1+Tingkat oversampling) . Nminoritas (3)

Copyright © 2024 **Authors**, Page 892
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 890−900
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5652

a) *N*minoritas adalah jumlah awal sampel pada kelas minoritas
b) Tingkat *Oversampling* adalah seberapa banyak jumlah sampel pada kelas minoritas yang ingin ditingkatkan
c) Ambil sampel-sampel yang sudah ada pada kelas minoritas dan tambahkan ulang secara acak hingga mencapai *N*baru

2. SMOTE: membuat sampel sintetis baru berdasarkan jarak di antara sampel kelas minoritas [17].
   a) Pilih sampel minoritas x_i
   b) Temukan *k*-tetangga terdekat dari x_i.
   c) Hitung jarak antara x_i dengan tetangga terdekat
   d) Pilih jarak acak *r* dari jarak-jarak yang dihitung
   e) Hitung sampel sintetis baru: $\boldsymbol{x_{baru} = x_i + r(x_j - x_i)}$ (4)
   f) Iterasi untuk sampel lain sesuai kebutuhan
3. ADASYN: serupa dengan SMOTE, namun memberikan bobot lebih pada sampel yang lebih sulit dibedakan [18].
   a) Hitung jarak, $\boldsymbol{D_i^k = \{jarak(x_i, x_j) | x_j \in X_{m_i}^k\}}$ (5)
   b) Hitung rasio, $\boldsymbol{G_i = \frac{Jumlah\ Sampel\ Mayoritas\ dalam\ X_{m_i}^k}{k}}$
   c) Hitung bobot, $\boldsymbol{w_i = \frac{1}{1+G_i}}$
   d) Hitung jumlah sampel baru, $\boldsymbol{N_{baru} = int(w_i \,.\, tingkat\ oversampling)}$
   e) Generasi sampel sintetis seperti SMOTE pada sampel minoritas yang dipilih

b. *Undersampling*
1. *Random Under Sampling*: mengurangi jumlah sampel dari kelas mayoritas secara acak [19].
   a) Tentukan jumlah sampel yang akan dihapus (N_hapus)
   b) Pilih secara acak (N_hapus) sampel dari kelas mayoritas untuk dihapus
   c) Hapus sampel yang dipilih
2. *Tomek Links*: menghapus pasangan data yang berdekatan, yang dapat membantu meningkatkan batas keputusan di sekitar kelas minoritas [20].
   a) Hitung jarak antara setiap sampel minoritas (x_i) dan mayoritas (x_j)
   b) Identifikasi *Tomek Links* dengan memukan pasangan sampel (x_i, x_j) yang memenuhi kriteria sebagai *Tomek links*,
   $$\boldsymbol{jarak(x_i, x_j) < \min\left(jarak(x_i, x_j)\right), \forall x_k \neq x_i, x_j} \quad (6)$$
   c) Hapus satu atau kedua sampel dari pasangan (x_i, x_j) yang membentuk *Tomek Links*
3. *Edited Neirest Neighbours* (ENN): mengurangi sampel kelas mayoritas yang terklasifikasi secara tidak benar oleh model [21].
   a) Hitung jarak antara setiap sampel (x_i) dan tetangga terdekatnya dalam kelas yang sama dan kelas yang berbeda untuk setiap kelas
   $$\boldsymbol{jarak(x_i, x_j), \forall x_j \in X\ dan\ y_i = y_j} \quad (7)$$
   $$\boldsymbol{jarak(x_i, x_k), \forall x_k \in X\ dan\ y_i \neq y_k}$$
   b) Temukan sampel (x_i) yang diklasifikasikan secara salah oleh tetangga terdekatnya dalam kelas yang berbeda untuk setiap kelas
   $$\boldsymbol{jarak(x_i, x_j) > \min(jarak(x_i, x_k)), \forall x_j \in X\ dan\ y_i = y_j\ serta\ \forall x_k \in X\ dan\ y_i \neq y_k} \quad (8)$$
   c) Hapus sampel yang diklasifikasikan secara salah dari *dataset*

c. *Hybrid*
SMOTEENN(SMOTE+ENN): kombinasi antara SMOTE dan ENN untuk menghapus *noisy sample* dan menyusun kembali *dataset* [22].

2.5 Evaluasi

Evaluasi metode *confusion matrix* digunakan untuk memperoleh tingkat keakuratan model yang diuji dengan menganalisa dengan baik kualitas *classifier* dalam mengenali tuple-tuple dari kelas yang ada [23]. *Confusion matrix* memperoleh hasil *accuracy*, *precission* dan *recall.*

$$Akurasi = \frac{TP+TN}{(TP+FP+TN+FN)} \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

$$F1 - Score = 2.\frac{Precision\,.Recall}{Precision+Recall} \quad (12)$$

Copyright © 2024 **Authors**, Page 893
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 890−900
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5652

Confusion matrix terdiri dari empat komponen utama:

a. *True Positive* (TP) mengacu pada kasus di mana nilai aktualnya positif dan model juga memprediksi nilai tersebut sebagai positif.
b. *True Negative* (TN) adalah ketika nilai aktualnya negatif dan model berhasil memprediksi nilai tersebut sebagai negative.
c. *False Positive* (FP) terjadi ketika nilai aktualnya negatif, tetapi model secara keliru memprediksinya sebagai positif.
d. *False Negative* (FN) adalah ketika nilai aktualnya positif, namun model memprediksi nilai tersebut sebagai negative [24].

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan dan Analisis Data

Dataset yang digunakan merupakan data rekam medis demografi pasien dan hasil tes darah bersamaan dengan diagnosis penyakit tiroid. Sebanyak 9172 data yang terdiri dari 30 atribut utama diantaranya seperti pada Tabel 1 berikut:

Tabel 1. *Dataset*

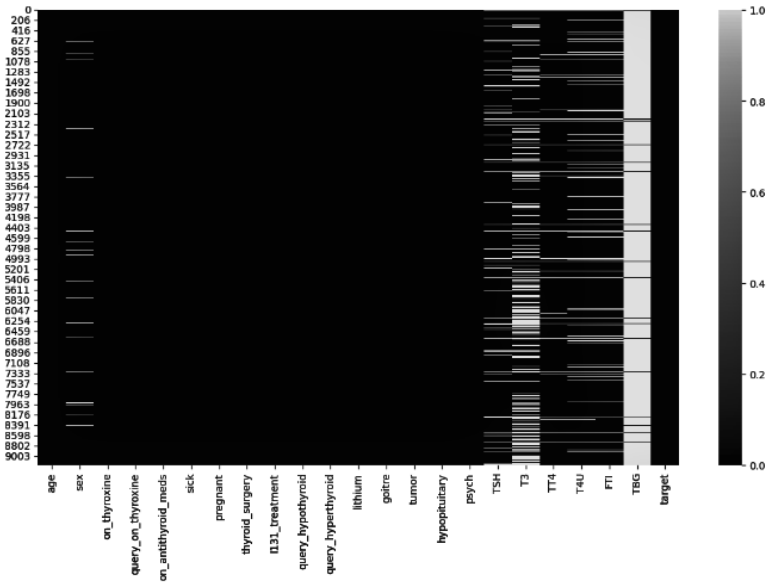
Nama Atribut	Keterangan	Nama Atribut	Keterangan
Age	Usia pasien	*Psych*	Apakah pasien memiliki kejiwaan
Sex	Jenis kelamin pasien	*TSH_measured*	Apakah TSH diukur
On_thyroxine	Apakah pasien menggunakan tiroksin	*TSH*	Tingkat TSH dalam darah
Query on Thyroxine	Pertanyaan apakah pasien menggunakan tiroksin	*T3_measured*	Apakah T3 diukur
On antithyroid meds	Apakah pasien sedang mengonsumsi obat antitiroid	*T3*	Tingkat T3 dalam darah
Sick	Apakah pasien sakit	*TT4_measured*	Apakah TT4 diukur
Pregnant	Apakah pasien hamil	*TT4*	Tingkat TT4 dalam darah
Thyroid_Surgery	Apakah pasien telah menjalani operasi tiroid	*T4U_measured*	Apakah T4U diukur
L131_treatment	Apakah pasien sedang menjalani pengobatan I131	*T4U*	Tingkat T4U dalam darah
Query_hypothyroid	Apakah pasien yakin mereka menderita hipotiroid	*FTI_measured*	Apakah FTI diukur
Query_hyperthyroid	Apakah pasien yakin mereka menderita hipertiroid	*FTI*	Tingkat FTI dalam darah
Lithium	Apakah pasien litium	*TBG_measured*	Apakah TBG diukur
Goitre	Apakah pasien menderita penyakit gondok	*TBG*	Tingkat TBG dalam darah
Tumor	Apakah pasien mengidap tumor	*Target*	Diagnosis medis hipertiroidisme
Hypopituitary	apakah pasien menderita hipopituitari	*Patient_id*	Identitas pasien

3.2 *Preprocessing*

a. Identifikasi *missing value*, duplikasi data dan *outlier*
Tahapan ini diawali dengan mendeteksi adanya *missing value* pada *dataset* guna memastikan integritas dan kualitas data, yang pada akhirnya akan berdampak pada tingkat akurasi.

Copyright © 2024 **Authors**, Page 894
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 890−900
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5652



Gambar 3. Sebaran *missing value*

Hasil deteksi *missing value* divisualisasikan seperti pada Gambar 3 dan metode KNN *imputer* digunakan untuk mengatasi sebaran *missing value*. Pada kasus ini, atribut TBG dieliminasi karena memiliki rasio *missing value* sebesar 96%.

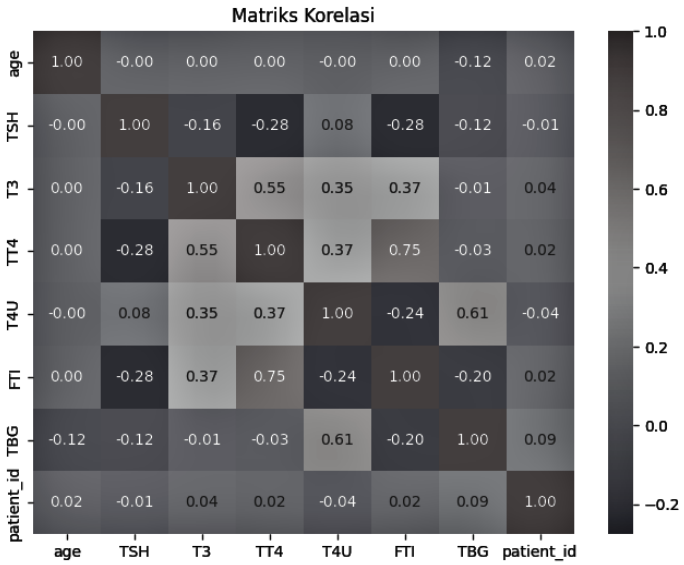
age	sex	on_thyrox	query_on	on_antith
65511	M	f	f	f
65512	M	f	f	f
65526	F	f	f	f

Gambar 4. Outlier

Pada Gambar 4, *outlier* yang terdeteksi tidak normal hanya pada atribut *age* (usia), dimana ada sejumlah data yang tidak realistis menunjukan usia manusia sebesar 6000 tahun.

b. Seleksi fitur menggunakan korelasi data

Penentuan atribut data untuk dilanjutkan ke tahap pengujian dilakukan dengan melihat korelasi atau hubungan antar atribut dengan visualisasi menggunakan *heatmap diagram.* Rentang korelasi dimulai dari -1 hingga 1 dimana jika semakin positif maka atribut akan semakin berkolerasi begitu juga sebaliknya.



Gambar 5. *Heatmap* korelasi data

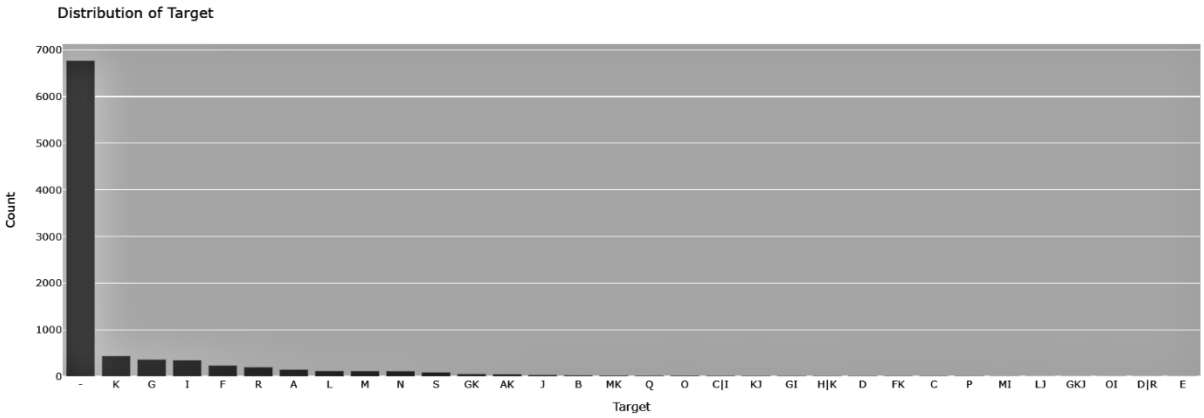
Pada Gambar 5, atribut TSH, T3, TT4, T4U dan FTI yang merupakan hormon kelenjar tiroid memiliki hubungan antar korelasi yang cukup positif dibanding dengan atribut lainnya, sehingga ke lima atribut tersebut digunakan dalam proses pengujian selanjutnya.

c. Transformasi data

Copyright © 2024 **Authors**, Page 895
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 890−900
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5652

Pada penelitian ini, transformasi data digunakan dalam mengubah format target diagnosis pasien yang cukup banyak menjadi 3 kategori utama.



Gambar 6. Sebaran target

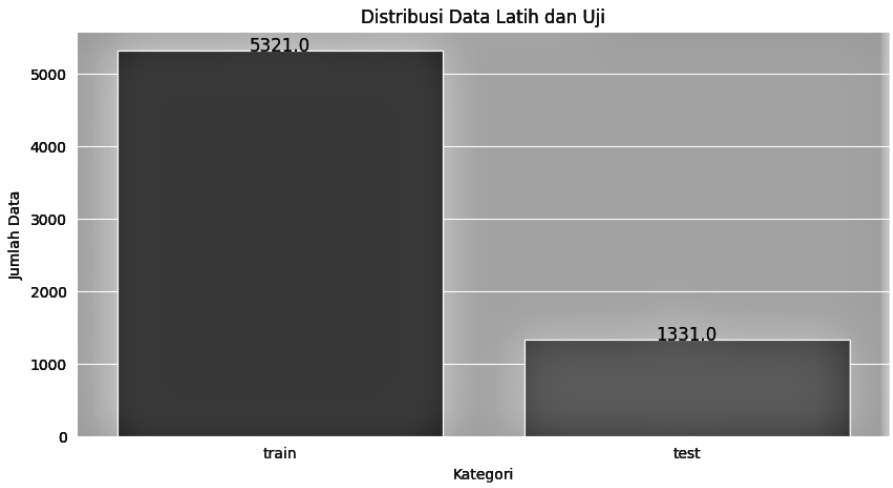
Pada *dataset* yang digunakan, atribut target terdiri dari beberapa macam kondisi yang menunjukkan diagnosis pasien penyakit tiroid. Pada Gambar 6, kategori yang menunjukkan pasien dengan diagnosis *hyperthyroid* merupakan huruf A, B, C dan D. Kategori E, F, G dan H menunjukkan pasien dengan diagnosis *hypothyroid* sedangkan untuk kategori lainnya merupakan atribut tambahan dalam diagnosis penyakit tiroid seperti I dan J untuk tingkat protein pengikat. Seluruh kondisi tersebut dirubah atau ditransformasikan menjadi 3 kategori utama yaitu *negative*, *hypothyroid* dan *hyperthyroid* seperti pada Gambar 7 dibawah ini:

	target
Negative	6771
Hypothyroid	667
Hyperthyroid	241

Gambar 7. Target utama

d. *Split data*

Dalam tujuan memungkinkan pengembangan model yang optimal, data perlu dipisahkan menjadi dua bagian utama yaitu data *training* dan data *testing*. Data *training* digunakan untuk pembelajaran sedangkan data *testing* digunakan untuk menguji model yang dibangun.



Gambar 8. *Split data*

Pada Gambar 8, *split data* dilakukan dalam proses *preprocessing* dengan membagi data menjadi data latih dan data uji dengan rasio 80:20 agar terjadi keseimbangan antara pelatihan dan pengujian juga untuk menghindari *overfitting.*

Copyright © 2024 **Authors**, Page 896
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 890−900
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5652

3.3 *Resampling Data*

Resampling data dapat dilakukan setelah proses *split data* karena *resampling* hanya mempengaruhi data *training*. Hal ini memastikan bahwa evaluasi model dilakukan pada data yang belum diubah dan masih merepresentasikan distribusi kelas yang asli.

Tabel 2. Hasil *resampling data*

No	Metode Resampling	Jumlah Awal	Setelah Resampling
1	*Random Over Sampling*	5321	13971
2	*Smote*		13971
3	*Adasyn*		13981
4	*Random Under Sampling*		537
5	*Tomeks Links*		5248
6	*Edited Neirest Neighbours*		4853
7	*Smotenn*		13189

Pada Tabel 2, metode *oversampling* dengan *Random Over Sampling*, SMOTE dan ADASYN mengubah data latih menjadi lebih besar dengan jumlah yang hampir sama. Metode *undersampling* mengubah data latih menjadi lebih kecil dengan *Random Under Sampling*, sedangkan metode *Tomek Links* dan *Edited Neirest Neighbours* hanya sedikit mengubah data latih menjadi lebih kecil. Metode *hybrid* dengan SMOTENN mengubah data latih menjadi lebih besar tetapi tidak lebih dari metode *oversampling*.

3.4 Pengujian

a. Algoritma SVM

'Hasil Evaluasi untuk Algoritma SVM'

	Metode Resampling	Akurasi	Presisi	Recall	F1-Score
0	None	0.949662	0.947523	0.949662	0.947072
1	SMOTE	0.939895	0.962169	0.939895	0.946315
2	ADASYN	0.835462	0.948491	0.835462	0.873630
3	RandomOverSampling	0.941397	0.963371	0.941397	0.947893
4	RandomUnderSampling	0.915853	0.955399	0.915853	0.927349
5	TomekLinks	0.951916	0.949995	0.951916	0.949692
6	ENN	0.951916	0.950796	0.951916	0.949535
7	SMOTEENN	0.938392	0.963332	0.938392	0.945664

Gambar 9. Hasil evaluasi Algoritma SVM

Pada Gambar 9, pengujian algoritma SVM mengalami peningkatan akurasi, *recall* dan *f1 score* ketika dikombinasikan dengan *Tomek Links* dan ENN dengan hasil akurasi yang sama dari 94.9% menjadi 95.1%. Tingkat presisi SVM hanya meningkat dengan *Random Over Sampling* sebesar 94% menjadi 96% dibandingkan dengan kombinasi metode lainnya yang mengalami penurunan seperti pada metode ADASYN.

b. Algoritma *Decision Tree*

	Metode Resampling	Akurasi	Presisi	Recall	F1-Score
0	None	0.976709	0.976754	0.976709	0.976704
1	SMOTE	0.975958	0.978540	0.975958	0.976843
2	ADASYN	0.969947	0.972203	0.969947	0.970822
3	RandomOverSampling	0.974455	0.975350	0.974455	0.974838
4	RandomUnderSampling	0.948159	0.970459	0.948159	0.955228
5	TomekLinks	0.975958	0.976517	0.975958	0.976201
6	ENN	0.959429	0.960059	0.959429	0.958332
7	SMOTEENN	0.978212	0.981625	0.978212	0.979202

Gambar 10. Hasil evaluasi Algoritma *Decision Tree*

Copyright © 2024 **Authors**, Page 897
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 890−900
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5652

Pada Gambar 10, pengujian Algoritma *Decision Tree* mengalami peningkatan keseluruhan ketika dikombinasikan dengan metode SMOTEENN dengan hasil akurasi sebesar 97.6% menjadi 97.8, sedangkan untuk kombinasi metode lainnya mengalami penurunan tetapi tidak terlalu signifikan.

c. Algoritma *Neural Network*

	Metode Resampling	Akurasi	Presisi	Recall	F1-Score
0	None	0.969196	0.968324	0.969196	0.968139
1	SMOTE	0.936890	0.963044	0.936890	0.944579
2	ADASYN	0.924869	0.955744	0.924869	0.934210
3	RandomOverSampling	0.963186	0.968896	0.963186	0.964627
4	RandomUnderSampling	0.862509	0.942502	0.862509	0.891397
5	TomekLinks	0.966191	0.964598	0.966191	0.964347
6	ENN	0.950413	0.948812	0.950413	0.947320
7	SMOTEENN	0.942149	0.964682	0.942149	0.949052

Gambar 11. Hasil evaluasi Algoritma *Neural Network*

Pada Gambar 11, pengujian Algoritma *Neural Network* menghasilkan tingkat akurasi sebesar 96.9% dan ketika dikombinasikan menggunakan beberapa metode *resampling* hanya menghasilkan penurunan tingkat akurasi. Pada pengujian ini, hanya terjadi perubahan kecil pada tingkat presisi ketika dikombinasikan dengan metode *Random Over Sampling.*

d. Algoritma *Random Forest*

	Metode Resampling	Akurasi	Presisi	Recall	F1-Score
0	None	0.984974	0.985521	0.984974	0.985168
1	SMOTE	0.981968	0.983958	0.981968	0.982591
2	ADASYN	0.981217	0.983512	0.981217	0.981935
3	RandomOverSampling	0.985725	0.986415	0.985725	0.985931
4	RandomUnderSampling	0.954170	0.974254	0.954170	0.960373
5	TomekLinks	0.984974	0.985631	0.984974	0.985186
6	ENN	0.959429	0.958955	0.959429	0.958236
7	SMOTEENN	0.978212	0.983199	0.978212	0.979536

Gambar 12. Hasil evaluasi Algoritma *Random Forest*.

Pada Gambar 12, pengujian Algoritma *Random Forest* mengalami peningkatan keseluruhan dengan kombinasi metode *Random Over Sampling* dengan hasil sebesar 98.4% menjadi 98.5%, sedangkan untuk kombinasi metode lainnya mengalami penurunan tetapi tidak terlalu signifikan.

e. Algoritma *Logistic Regression*

	Metode Resampling	Akurasi	Presisi	Recall	F1-Score
0	None	0.943651	0.940573	0.943651	0.940017
1	SMOTE	0.935387	0.961911	0.935387	0.943347
2	ADASYN	0.924869	0.955691	0.924869	0.934261
3	RandomOverSampling	0.939895	0.963128	0.939895	0.946757
4	RandomUnderSampling	0.923366	0.956097	0.923366	0.933082
5	TomekLinks	0.947408	0.945488	0.947408	0.944368
6	ENN	0.951165	0.949853	0.951165	0.948818
7	SMOTEENN	0.927122	0.956403	0.927122	0.935677

Gambar 13. Hasil evaluasi Algoritma *Logistic Regression*

Pada Gambar 13, pengujian Algoritma *Logistic Regression* mengalami peningkatan akurasi, recall dan f1 score pada kombinasi metode ENN dengan hasil akurasi sebesar 94.3% menjadi 95.1%, tetapi untuk peningkatan hasil presisi nya lebih optimal pada metode *Random Over Sampling* dengan hasil 94% menjadi 96.3%.

Copyright © 2024 **Authors**, Page 898
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 890−900
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5652

f. Algoritma *Naive Bayes*

	Metode Resampling	Akurasi	Presisi	Recall	F1-Score
0	None	0.915101	0.925025	0.915101	0.918834
1	SMOTE	0.835462	0.917149	0.835462	0.858152
2	ADASYN	0.184072	0.896218	0.184072	0.145280
3	RandomOverSampling	0.851240	0.921075	0.851240	0.870820
4	RandomUnderSampling	0.839219	0.910670	0.839219	0.860150
5	TomekLinks	0.918858	0.927026	0.918858	0.921997
6	ENN	0.954170	0.960698	0.954170	0.956177
7	SMOTEENN	0.874530	0.935143	0.874530	0.890204

Gambar 14. Hasil evaluasi Algoritma *Naïve bayes*

Pada Gambar 14, pengujian Algoritma *Naive Bayes* mengalami peningkatan keseluruhan ketika menggunakan metode *Tomek Links* dan ENN. Kombinasi dengan metode ENN menghasilkan tingkat akurasi paling signifikan sebesar 91.5% menjadi 95.4%, sedangkan untuk beberapa kombinasi metode yang lain mengalami penurunan tingkat akurasi yang cukup signifikan.

g. Algoritma *K-Nearest Neighbours*

	Metode Resampling	Akurasi	Presisi	Recall	F1-Score
0	None	0.945905	0.945404	0.945905	0.940903
1	SMOTE	0.901578	0.932824	0.901578	0.912804
2	ADASYN	0.876033	0.931474	0.876033	0.895721
3	RandomOverSampling	0.926371	0.935414	0.926371	0.929439
4	RandomUnderSampling	0.812171	0.893149	0.812171	0.838512
5	TomekLinks	0.945154	0.945118	0.945154	0.939686
6	ENN	0.936138	0.937216	0.936138	0.928669
7	SMOTEENN	0.885800	0.933898	0.885800	0.902799

Gambar 15. Hasil evaluasi Algoritma *K-Nearest Neighbors*

Sama seperti Algoritma *Neural Network*, KNN ketika dikombinasikan menggunakan beberapa metode resampling hanya menghasilkan penurunan tingkat akurasi seperti pada Gambar 15.

4. KESIMPULAN

Berdasarkan hasil pengujian, kombinasi algoritma klasifikasi dan metode *resampling* untuk meningkatkan kinerja algoritma pada *dataset* penyakit tiroid ini dapat disimpulkan untuk Algoritma SVM menunjukkan peningkatan yang signifikan dalam akurasi, recall, dan f1 score ketika dikombinasikan dengan *Tomek Links* dan ENN, sementara untuk meningkatkan presisi, *Random Over Sampling* lebih efektif. Hasil pengujian Algoritma *Decision Tree* menunjukkan hasil terbaik dengan kombinasi SMOTEENN yang signifikan dalam peningkatan akurasi. Pengujian Algoritma *Neural Network* menunjukkan peningkatan presisi dengan *Random Over Sampling*, namun dengan kombinasi metode *resampling* lainnya dapat menurunkan akurasi. Pengujian Algoritma *Random Forest* memberikan hasil terbaik dengan *Random Over Sampling* untuk meningkatkan akurasi. Hasil dari pengujian Algoritma *Logistic Regression* menunjukkan peningkatan akurasi, recall, dan f1 score dengan kombinasi ENN, sementara presisi meningkat lebih baik dengan *Random Over Sampling.* Pengujian Algoritma *Naive Bayes* menunjukkan peningkatan yang signifikan dalam akurasi dengan kombinasi *Tomek Links* dan ENN, terutama dengan penggunaan ENN. Sama hal nya dengan Algoritma *Neural Network*, KNN tidak menunjukkan peningkatan yang signifikan dalam akurasi dengan metode *resampling*. Seluruh hasil ini merupakan pengujian dari 7 algoritma dengan karakteristik dan pendekatan yang berbeda. Pada penelitian ini, Algoritma *Random Forest* menghasilkan tingkat akurasi tertinggi di antara algoritma yang lain dikarenakan sangat efektif dalam proses klasifikasi dengan data *multiclass.* Dengan demikian, kombinasi Algoritma *Random Forest* dengan metode *resampling Random Over Sampling* dapat meningkatkan akurasi deteksi dini pada penyakit tiroid.

REFERENCES

[1] W. E. Ariawan, I. Made, and A. W. Putra, "Sistem Pakar Mendiagnosa Penyakit Tiroid Menggunakan Metode Certainty Factor Berbasis Web," *Jurnal Sutasoma*, vol. 01, no. 01, pp. 104–110, 2022, [Online]. Available: https://s.id/jurnalsutasoma

Copyright © 2024 **Authors**, Page 899
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 890−900
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5652

[2] R. S. Tantika and A. Kudus, "Penggunaan Metode Support Vector Machine Klasifikasi Multiclass pada Data Pasien Penyakit Tiroid," *Bandung Conference Series: Statistics*, vol. 2, no. 2, pp. 159–166, Jul. 2022, doi: 10.29313/bcss.v2i2.3590.

[3] K. Anda and M. Sandrianti, "Siaran Pers Survei Mengungkap Kurangnya Pengetahuan Tentang Dampak Gangguan Tiroid Terhadap Kesuburan," 2020. [Online]. Available: https://www.healthywomen.org/content/article/thyroid-

[4] S. Agustiani, A. Mustopa, A. Saryoko, W. Gata, S. Khotimatul Wildah, and S. Nusa Mandiri, "Penerapan Algoritma J48 Untuk Deteksi Penyakit Tiroid," *Paradigma - Jurnal Informatika dan Komputer*, vol. 22, no. 2, pp. 153–160, 2020, doi: 10.31294/p.v21i2.

[5] T. Okta Bagaskara, M. Izman Herdiansyah, T. Sutabri, and E. Surya Negara, "Model Prediksi Menggunakan Teknik Machine Learning untuk Penjualan terhadap Produksi Kain Jumputan pada Pengerajin Batiq Colet Jumputan Palembang," *PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika*, vol. 16, no. 2, pp. 189–199, 2023, doi: https://doi.org/10.33322/petir.v16i2.2187.

[6] S. Muhammad and W. Nugraha, "Mwmote Dalam Mengatasi Ketidakseimbangan Kelas Pada Prediksi Churn Menggunakan Klasifikasi C4.5," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 54–62, Feb. 2023.

[7] A. Nugroho and Y. Religia, "Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 504–510, Jun. 2021, doi: 10.29207/resti.v5i3.3067.

[8] G. Gumelar, Q. Ain, R. Marsuciati, S. Agustanti Bambang, A. Sunyoto, and M. Syukri Mustafa, "Kombinasi Algoritma Sampling dengan Algoritma Klasifikasi untuk Meningkatkan Performa Klasifikasi Dataset Imbalance," *SISFOTEK-Sistem Informasi dan Teknologi*, pp. 250–255, 2021.

[9] A. Syukron, E. Saputro, and P. Widodo, "Penerapan Metode Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung," 2023. [Online]. Available: https://doi.org/10/25047/jtit.v10i1.312

[10] N. Yudistira and A. F. Putra, "Algoritma Decision Tree Dan Smote Untuk Klasifikasi Serangan Jantung Miokarditis Yang Imbalance," *Jurnal Litbang Edusaintech*, vol. 2, no. 2, pp. 112–122, Dec. 2021, doi: 10.51402/jle.v2i2.48.

[11] M. R. Hunafa and A. Hermawan, "Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Imbalace Class Dataset Penyakit Diabetes," *KLIK: Kajian Ilmiah Informatika dan Komputer* , vol. 4, no. 3, pp. 1551–1561, 2023, doi: 10.30865/klik.v4i3.1486.

[12] A. Handika Permana, F. Rakhmat Umbara, and F. Kasyidi, "Klasifikasi Penyakit Jantung Tipe Kardiovaskular Menggunakan Adaptive Synthetic Sampling dan Algoritma Extreme Gradient Boosting," *Technology and Science (BITS)*, vol. 6, no. 1, 2024, doi: 10.47065/bits.v6i1.5421.

[13] R. Aryanti, T. Misriati, and R. Hidayat, "Klasifikasi Risiko Kesehatan Ibu Hamil Menggunakan Random Oversampling Untuk Mengatasi Ketidakseimbangan Data," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 3, no. 5, pp. 409–416, 2023, [Online]. Available: https://djournals.com/klik

[14] L. Mutawali, W. Murniati, and K. Kunci, "PENERAPAN KNNIMPUTER DALAM MENGOLAH DATA MISSING VALUE UNTUK MEMBANTU MENINGKATKAN AKURASI SUPPORT VECTOR MACHINE KLASIFIKASI PENYAKIT TIROID," 2022. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/thyroid+diseas

[15] U. Erdiansyah, A. Irmansyah Lubis, and K. Erwansyah, "Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 1, p. 208, Jan. 2022, doi: 10.30865/mib.v6i1.3373.

[16] M. Hayaty, S. Muthmainah, and S. M. Ghufran, "Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification," *International Journal of Artificial Intelligence Research*, vol. 4, no. 2, p. 86, Jan. 2021, doi: 10.29099/ijair.v4i2.152.

[17] Y. T. Samuel, C. Beatrix, and A. Nahuway, "Prediksi Indeks Prestasi Mahasiswa Yang Berkuliah Sambil Bekerja Di Universitas Advent Indonesia Dengan Menggunakan Metode Decision Tree C4.5 Dan SMOTE," *Jurnal TeIKa*, vol. 10, no. 1, pp. 69–77, 2020.

[18] U. Ungkawa and M. A. Rafi, "Data Balancing Techniques Using the PCA-KMeans and ADASYN for Possible Stroke Disease Cases," *Jurnal Online Informatika*, vol. 9, no. 1, pp. 138–147, Jun. 2024, doi: 10.15575/join.v9i1.1293.

[19] W. I. Sabilla and C. B. Vista, "Implementasi SMOTE dan Under Sampling pada Imbalanced Dataset untuk Prediksi Kebangkrutan Perusahaan," *Jurnal Komputer Terapan*, vol. 7, no. 2, pp. 329–339, 2021, [Online]. Available: https://jurnal.pcr.ac.id/index.php/jkt/

[20] H. Utami, "Analisis Sentimen dari Aplikasi Shopee Indonesia Menggunakan Metode Recurrent Neural Network," *Indonesian Journal of Applied Statistics*, vol. 5, no. 1, p. 31, May 2022, doi: 10.13057/ijas.v5i1.56825.

[21] H. Wang and X. Liu, "Undersampling bankruptcy prediction: Taiwan bankruptcy data," *PLoS One*, vol. 16, no. 7 July, Jul. 2021, doi: 10.1371/journal.pone.0254030.

[22] F. Yang, K. Wang, L. Sun, M. Zhai, J. Song, and H. Wang, "A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis," *BMC Med Inform Decis Mak*, vol. 22, no. 1, p. 344, Dec. 2022, doi: 10.1186/s12911-022-02075-2.

[23] K. Akbar and M. Hayaty, "Data Balancing untuk Mengatasi Imbalance Dataset pada Prediksi Produksi Padi Balancing Data to Overcome Imbalance Dataset on Rice Production Prediction," *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, vol. 2, no. 02, pp. 1–14, 2020.

[24] H. Apriyani, "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus," 2020. [Online]. Available: https://journal-computing.org/index.php/journal-ita/index

Copyright © 2024 **Authors**, Page 900
This Journal is licensed under a Creative Commons Attribution 4.0 International License