JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

Seleksi Fitur Information Gain untuk Optimasi Klasifikasi Penyakit Tuberkulosis

Ardi Caesar Kurniawan*, Abu Salam

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1,*]111202012701@mhs.dinus.ac.id, [2]abu.salam@dsn.dinus.ac.id
Email Penulis Korespondensi: 111202012701@mhs.dinus.ac.id

Abstrak−Tuberkulosis (TB), penyakit yang disebabkan oleh Mycobacterium tuberculosis, adalah ancaman kesehatan global yang dapat menyebar melalui udara. Faktor-faktor seperti jenis kelamin, usia, dan lokasi geografis mempengaruhi penyebarannya. Indonesia, sebagai negara dengan jumlah kasus TB kedua terbanyak di dunia, mencatat peningkatan kasus TB yang signifikan dari tahun 2020 hingga 2022, terutama di Kota Semarang. Untuk meminimalisir dampak TB, penting untuk mengidentifikasi faktor-faktor yang mempengaruhi perkembangan penyakit ini. Teknik Machine Learning seperti seleksi fitur seperti Information Gain dan algoritma klasifikasi seperti Random Forest dapat digunakan. Seleksi fitur membantu menentukan faktor mana yang paling berpengaruh terhadap penyakit TB dengan perankingan bobot atribut, sementara Random Forest digunakan untuk klasifikasi. Teknik oversampling seperti Synthetic Minority Oversampling Technique (SMOTE) digunakan untuk menangani ketidakseimbangan data dan meningkatkan performa klasifikasi. Dari hasil penelitian, disimpulkan bahwa model klasifikasi Random Forest menunjukkan performa terbaik dengan menggunakan semua fitur atau atribut dari bobot terbesar hingga terkecil yaitu; 'tipe_diagnosis', 'jenis_fasyankes', 'usia', 'kelurahan_kecamatan', 'riwayat_dm', 'riwayat_HIV', 'tahun', 'paduan_OAT', 'status_pekerjaan', 'jenis_kelamin', 'tipe_TBC', 'riwayat_TBC', 'bulan' dan 'sumber_obat' pada dataset asli penyakit TB di Kota Semarang. Tingkat recall dan akurasi mencapai 75%. Hasil ini lebih baik dibandingkan dengan model klasifikasi TB di Kota Semarang yang menggunakan dataset oversampling dengan SMOTE dan hanya menggunakan 10-12 atribut teratas, dengan tingkat recall dan akurasi mencapai 74%. Penelitian ini menunjukkan bahwa teknik-teknik tertentu dalam Machine Learning dapat membantu memahami faktor-faktor yang mempengaruhi hasil pengobatan TB.

Kata Kunci: Tuberkulosis; Machine Learning; Information Gain; Random Forest; Synthetic Minority Oversampling Technique

Abstract−Tuberculosis (TB), caused by Mycobacterium tuberculosis, is a global health threat that spreads through the air. Factors such as gender, age, and geographical location influence its spread. Indonesia, the country with the second-highest number of TB cases globally, recorded a significant increase in TB cases from 2020 to 2022, especially in Semarang City. To minimize TB's impact, it's crucial to identify the factors influencing its progression. Machine Learning techniques like feature selection (Information Gain) and classification algorithms (Random Forest) can be utilized. Feature selection helps determine which factors most influence TB by ranking attribute weights, while Random Forest is used for classification. Oversampling techniques like Synthetic Minority Oversampling Technique (SMOTE) are used to handle data imbalance and improve classification performance. The study concluded that the Random Forest classification model showed the best performance using all features or attributes from the highest to the lowest weight namely; 'tipe_diagnosis', 'jenis_fasyankes', 'usia', 'kelurahan_kecamatan', 'riwayat_dm', 'riwayat_HIV', 'tahun', 'paduan_OAT', 'status_pekerjaan', 'jenis_kelamin', 'tipe_TBC', 'riwayat_TBC', 'bulan' and 'sumber_obat' on the original TB disease dataset in Semarang City. The recall and accuracy rate reached 75%. This result is better than the TB classification model in Semarang City that uses the oversampling dataset with SMOTE and only uses the top 10-12 attributes, with a recall and accuracy rate of 74%. This research shows that certain techniques in Machine Learning can help understand the factors influencing TB treatment outcomes.

Keywords: Tuberculosis; Machine Learning; Information Gain; Random Forest; Synthetic Minority Oversampling Technique

1. PENDAHULUAN

Tuberkulosis (TB), penyakit yang disebabkan oleh bakteri Mycobacterium tuberculosis, adalah ancaman kesehatan global yang dapat ditularkan melalui udara saat seseorang batuk, bersin, atau meludah [1]. Faktor-faktor seperti jenis kelamin, usia, lokasi geografis, dan riwayat kesehatan juga berperan dalam penyebaran TB [2]–[4]. Indonesia, sebagai negara dengan jumlah kasus TB kedua terbanyak di dunia, mencatat peningkatan kasus TB sebesar 17% dari tahun 2020 ke 2022, yaitu dari 824.000 menjadi 969.000 kasus [5], [6]. Di Kota Semarang, jumlah kasus TB pada tahun 2021 mencapai 3.221, dengan mayoritas penderita adalah laki-laki (55%). Angka ini terus meningkat dari tahun 2020 [7]. Faktor-faktor seperti jenis kelamin, usia, dan tipe diagnosis berkontribusi terhadap peningkatan ini [8], [9].

Untuk meminimalisir dampak TB, penting untuk mengidentifikasi faktor-faktor yang mempengaruhi perkembangan penyakit ini selain infeksi bakteri. Salah satu cara untuk mencapai ini adalah dengan menggunakan teknik Machine Learning, khususnya fitur seleksi, yang dapat membantu kita memahami faktor-faktor mana yang paling berpengaruh terhadap TB. Teknik Machine Learning lainnya seperti klasifikasi untuk membuat model yang dapat memprediksi kelas atau label suatu instance berdasarkan fitur-fiturnya dan oversampling untuk menangani ketidakseimbangan kelas dalam dataset. Analisis data yang cermat dan menyeluruh dengan menggunakan teknik ini dapat memberikan wawasan berharga untuk mengetahui faktor yang mempengaruhi penyakit TB, meningkatkan pengelolaan TB dan hasil pengobatan pasien [2], [10].

Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

Seleksi fitur adalah proses yang menentukan fitur mana yang paling relevan dalam suatu masalah. Information Gain adalah salah satu algoritma seleksi fitur paling sederhana yang digunakan untuk menentukan atribut terbaik dalam dataset berdasarkan perbedaan entropi sebelum dan sesudah pemisahan fitur. Atribut dengan nilai Information Gain tertinggi dianggap paling relevan dan penting dalam memprediksi kelas target. Metode ini membantu dalam menentukan berapa banyak atribut yang akan digunakan selanjutnya untuk proses klasifikasi [11]–[13].

Klasifikasi adalah proses pengelompokan objek berdasarkan atribut atau fitur yang dimilikinya untuk memprediksi kelas atau kategori objek tersebut [14], [15]. Salah satu teknik yang dapat digunakan adalah metode Random Forest. Random Forest merupakan algoritma machine learning yang digunakan untuk klasifikasi dan regresi. Algoritma ini bekerja dengan menggabungkan beberapa pohon keputusan yang dibuat secara acak dari subset data. Setiap pohon memberikan prediksi dan hasil akhir diperoleh dengan menggabungkan prediksi dari semua pohon. Kelebihan utama dari Random Forest adalah resistensi terhadap overfitting, akurasi yang tinggi, dan efisiensi pada dataset besar. Algoritma ini juga dapat diterapkan pada dataset dengan jumlah data tidak seimbang dengan hasil yang baik dan waktu eksekusi yang cepat [11], [14], [16]. Untuk meningkatkan hasil klasifikasi, dilakukan upaya untuk menyeimbangkan kelas pada dataset menggunakan teknik oversampling.

Oversampling adalah teknik yang digunakan untuk menangani ketidakseimbangan data dengan meningkatkan jumlah sampel dari kelas minoritas [17], [18]. Salah satu metode oversampling adalah Synthetic Minority Oversampling Technique (SMOTE), menciptakan data sintetis tambahan dari kelas minoritas dengan melakukan interpolasi antara titik-titik data yang telah ada. Ini membantu memperluas area keputusan kelas minoritas dan meningkatkan performa klasifikasi pada dataset yang tidak seimbang [19], [20].

Pada penelitian yang dilakukan oleh Elreedy dan Atiya [19], mengulas tentang penerapan metode oversampling SMOTE pada dataset yang tidak seimbang. Data yang digunakan dalam penelitian ini berasal dari UCI dataset. Performa yang diukur menggunakan metode klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) dalam penelitian tersebut adalah akurasi. Penelitian ini mengevaluasi seberapa dekat pola yang dihasilkan SMOTE dengan pola asli dan bagaimana faktor-faktor tertentu mempengaruhi distribusi pola tersebut. Hasilnya menunjukkan bahwa SMOTE meningkatkan akurasi klasifikasi.

Penelitian yang dilakukan oleh Nugroho dan Rilvani [14], membahas tentang penggunaan metode oversampling SMOTE untuk mengoptimalkan algoritma Random Forest terhadap prediksi kebangkrutan perusahaan. Data yang digunakan dalam penelitian ini data Bankruptcy dari Taiwan Economic Journal tahun 1999–2009. Performa yang diukur pada penelitian tersebut adalah akurasi, recall, dan F1 score. Hasil yang diperoleh menunjukkan bahwa penggunaan teknik oversampling SMOTE dapat meningkatkan nilai akurasi pada algoritma Random Forest Classifier untuk prediksi kebangkrutan perusahaan pada dataset tidak seimbang. Terbukti dengan peningkatan akurasi sebesar 7,40%, dari 88,30% menjadi 95,70%, sedangkan recall mengalami penurunan sebesar 24,20%, dari 74,20% menjadi 50,00%.

Penelitian yang dilakukan oleh Prakoso et al. [11], membahas tentang penggunaan fitur seleksi yaitu Information Gain untuk optimasi algoritma naïve bayes dan Random Forest pada klasifikasi berita. Data yang digunakan pada penelitian tersebut berasal dari situs pemberitaan online detik.com. Performa yang diukur pada penelitian tersebut mencakup akurasi, recall, dan presisi. Setelah menjalankan 6 skenario pengujian, model Remove Useless Attributes, Naive bayes Classifier-Multinomial, dan Random Forest-Feature Selection Information Gain, mendapatkan hasil evaluasi tertinggi. Hasil tersebut mencapai nilai accuracy 85.67%, nilai recall 85.67%, dan nilai precision 86.23%.

Penelitian yang dilakukan oleh Hasibuan dan Marji [21], mengulas tentang penerapan Information Gain untuk seleksi fitur yang akan digunakan dalam klasifikasi penyakit gagal ginjal dengan menggunakan metode MKNN. Data yang digunakan pada penelitian tersebut bersumber dari Chronic Kidney Disease database UCI Machine Learning Repository. Performa yang diukur pada penelitian tersebut adalah nilai akurasi. Hasil yang diperoleh berdasarkan uji variasi jumlah fitur setelah seleksi dan pengaruh variasi nilai K, menghasilkan akurasi tertinggi pada 4 fitur dengan nilai K = 2 dan 4 yang menghasilkan akurasi 97,7%, dan pada 6 fitur dengan nilai K = 2 dan 4 yang menghasilkan akurasi 97,7%. Untuk uji sistem menggunakan Information Gain menghasilkan akurasi 96,8%, sedangkan yang tidak menggunakan Information Gain menghasilkan akurasi 79,9%.

Penelitian yang dilakukan oleh Madaan et al. [16], membahas tentang penggunaan algoritma Random Forest dan Decision Tree untuk prediksi gagal bayar pinjaman di sektor perbankan. Data yang digunakan dalam penelitian ini adalah data Lending Club. Performa yang diukur pada penelitian tersebut adalah akurasi F1 score. Hasil yang didapatkan yaitu bahwa algoritma Random Forest mampu memberikan akurasi sebesar 80% sedangkan algoritma Decision Tree hanya mampu memberikan akurasi sebesar 73%, sehingga algoritma Random Forest menjadi pilihan yang lebih baik untuk jenis data tersebut.

Dengan latar belakang tersebut, penelitian ini bertujuan untuk mengidentifikasi model klasifikasi dan fitur seleksi yang paling efektif dalam memprediksi hasil pengobatan pasien TB. Dengan mempertimbangkan ketidakseimbangan dataset yang digunakan, penelitian ini berpotensi memberikan wawasan berharga dalam pengelolaan TB dan membantu penyedia layanan kesehatan membuat keputusan yang lebih tepat dalam merawat pasien TB.

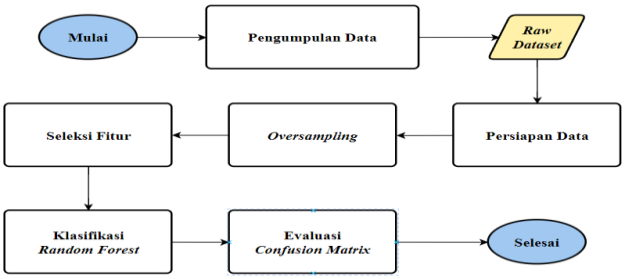
Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Proses penelitian dirancang sesuai dengan gambar 1, yang menggambarkan tahap-tahapan penelitian. Penjelasan mengenai langkah-langkah yang dilakukan untuk mencapai tujuan penelitian sebagai berikut.



Gambar 1. Skema Penelitian

Tahapan utama dalam metode penelitian terdiri dari pengumpulan data, persiapan data, oversampling, seleksi fitur, klasifikasi Random Forest dan evaluasi confusion matrix. Beberapa proses melibatkan tahapan-tahapan tersebut.

2.2 Pengumpulan Data

Data yang digunakan pada penelitian ini berasal dari Dinas Kesehatan Kota (DKK) Semarang. Data ini bernama 'Data Sistem Informasi Tuberkulosis(SITB)' tahun 2020-2022 dengan jumlah data sebanyak 6754. Atribut data yang digunakan dapat dilihat pada tabel 1 berikut ini.

Tabel 1. Atribut pada dataset penyakit tb

No	Nama Atribut	Tipe Data
1	kode_unik	object
2	tahun	int64
3	bulan	object
4	jenis_kelamin	object
5	usia	int64
6	kelurahan	object
7	kecamatan	object
8	kab_kota	object
9	status_pekerjaan	object
10	jenis_fasyankes	object
11	tipe_diagnosis	object
12	tipe_TBC	object
13	riwayat_TBC	object
14	riwayat_dm	object
15	riwayat_HIV	object
16	paduan_OAT	object
17	sumber_obat	object
18	hasil_pengobatan	object

2.3 Persiapan Data

Persiapan data adalah proses transformasi data mentah menjadi format terorganisir untuk analisis dengan algoritma machine learning, yang merupakan langkah penting dalam mempersiapkan data untuk pelatihan model dalam domain data mining [22], [23]. Fase persiapan data terbagi menjadi dua tahap; tahap pertama adalah Exploratory Data Analysis (EDA) dan tahap kedua adalah data preprocessing. EDA pertama-tama dilakukan untuk mendapatkan pemahaman yang lebih baik tentang setiap fitur dalam dataset yang digunakan. Setelah mendapatkan hasil EDA, data disiapkan dan ditransformasi selama tahap data preprocessing. Tahap data preprocessing tersusun dari proses remove unused attributes and merge attributes sampai dataset details.

a. Exploratory Data Analysis (EDA): EDA dijalankan dalam satu tahap, yaitu univariate analysis, yang hanya melibatkan satu fitur.
b. remove unused attributes and merge attributes: Pada tahap ini, atribut yang tidak digunakan dihapus. Kemudian menggabungkan dua atau lebih atribut agar menjadi satu aribut saja. Selanjutnya posisi atribut diatur ulang agar sama seperti format awalnya.

Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

c. dataset transformation: Setelah urutan atribut disusun kembali, atribut yang terdeteksi memiliki jenis data kategorikal akan diubah menjadi jenis data numerik sesuai dengan format yang ada dalam dataset.
d. dataset details: Proses yang dilakukan pada tahapan ini yaitu memvisualisasikan dan mengecek deskripsi, info, serta shape dari dataset. Pada tahap ini juga, atribut dan target dipisahkan dan dibagi menjadi data latih dan data uji.

2.4 Oversampling

Oversampling adalah tahapan selanjutnya dalam proses ini, yang merupakan metode untuk menyelesaikan masalah ketidakseimbangan dalam data dengan meningkatkan jumlah sampel dari kelas minoritas, sehingga menyeimbangkan hasil klasifikasi [17], [24]. Salah satu metode yang sering digunakan adalah Synthetic Minority Oversampling Technique (SMOTE), khususnya pada metode Oversampling. SMOTE bekerja dengan menghasilkan data sintetis tambahan dari kelas minoritas melalui teknik interpolasi antara titik-titik data yang sudah ada. Pendekatan ini melibatkan penambahan contoh dari kelas minoritas dengan mengekstraksi sampel data minoritas yang ada menggunakan sampel acak dari nilai k tetangga terdekat [25]. Proses ini dapat dijelaskan dengan rumus berikut [19]:

$$Z = X0 + w(X - X0) \quad (1)$$

Di mana Z adalah pola baru, $X0$ adalah pola dari kelas minoritas, X adalah salah satu dari k tetangga terdekat dari $X0$, dan w adalah variabel acak seragam dalam rentang $[0,1]$. Dengan kata lain, z adalah titik di sepanjang garis antara $X0$ dan X, dan posisinya ditentukan oleh nilai acak w.

Dengan demikian, SMOTE dapat memperluas area keputusan dari kelas minoritas dan meningkatkan performa klasifikasi pada dataset yang tidak seimbang [14], [19]. Ini membantu dalam meningkatkan akurasi dan performa model klasifikasi pada dataset yang tidak seimbang pada penelitian ini.

2.5 Seleksi Fitur

Dataset terdiri dari sejumlah atribut atau fitur yang jumlahnya dapat mempengaruhi hasil pengambilan keputusan, dan penggunaan atribut yang lebih sedikit cenderung menghasilkan keputusan yang lebih baik. Oleh karena itu, seleksi atribut atau fitur penting untuk menentukan atribut yang paling berpengaruh. Salah satu algoritma untuk melakukan seleksi fitur adalah Information Gain, yang merangking atribut menggunakan konsep entropy, ukuran ketidakpastian berdasarkan probabilitas suatu kejadian atau atribut tertentu. Dengan kata lain, Information Gain mengukur sejauh mana setiap fitur mempengaruhi suatu dataset dan jumlah atribut yang akan digunakan [11], [12], [26], [27]. Dalam penelitian ini, metode Information Gain digunakan untuk seleksi fitur. Proses ini dilakukan dalam dua konfigurasi, yaitu dengan menggunakan dataset asli sebagai data latih dan data latih yang telah melalui proses oversampling pada dataset.

2.6 Klasifikasi Random Forest

Klasifikasi adalah teknik data mining yang mencari pola untuk memisahkan jenis data, dengan tujuan mengelompokkan objek ke dalam kategori berdasarkan perilaku dan atribut grup yang telah ditetapkan. Algoritma klasifikasi yang digunakan adalah Random Forest, algoritma pembelajaran mesin yang diperkenalkan oleh Breiman dan terdiri dari banyak pohon keputusan. Algoritma ini adalah gabungan dari metode Bagging dan Random Subspaces. Dalam Random Forest, dataset dibagi secara acak menjadi dua bagian: data pelatihan dan data validasi. Banyak pohon keputusan dibuat secara acak dengan "bootstrap sample" dari dataset, dan percabangan setiap pohon ditentukan oleh prediktor yang dipilih secara acak. Estimasi akhir Random Forest adalah rata-rata dari semua hasil setiap pohon, sehingga setiap pohon mempengaruhi estimasi dengan bobot tertentu. Algoritma ini kuat dan dapat mencegah overfitting [28], [29].

Pada tahapan ini, terdapat empat tahapan klasifikasi dengan mempertimbangkan konfigurasi yang berbeda pada tiap tahapannya. Empat pengujian model dilakukan dengan variasi pada dataset asli dan dataset yang telah melalui proses SMOTE, serta penggunaan atau ketiadaan metode Information Gain. Pengujian pertama pada model Random Forest dievaluasi menggunakan dataset asli tanpa menerapkan metode Information Gain, pengujian kedua dilakukan dengan mempertimbangkan dataset asli dan menerapkan metode Information Gain pada proses klasifikasi, pengujian ketiga melibatkan dataset yang telah melalui proses SMOTE untuk menangani ketidakseimbangan kelas, namun tanpa menggunakan Information Gain, pada pengujian terakhir, model dievaluasi menggunakan dataset SMOTE dengan penerapan metode Information Gain pada tahapan klasifikasi. Hyperparameter yang digunakan untuk mengontrol proses pelatihan dan meningkatkan performa model dapat diperiksa pada tabel 2 berikut.

Tabel 2. Mendefinisi hyperparameter random forest

Nama Hyperparameter	Nilai Hyperparameter
n_estimators	100
max_depth	10
random_state	42

Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

2.7 Evaluasi Confusion Matrix

Confusion Matrix merupakan sebuah instrumen yang digunakan untuk menilai efisiensi klasifikasi, memberikan insight tentang sejauh mana sistem tersebut berhasil dalam mengelompokkan data. Evaluasi kinerja sistem klasifikasi melalui Confusion Matrix terwakili oleh empat konsep, yakni True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) [30]. Analisis lebih lanjut dapat ditemukan dalam tabel 3.

Tabel 3. Pengujian confusion matrix

Klasifikasi		Kondisi Prediksi	
		TRUE	**FALSE**
Kondisi Aktual	TRUE	True Positif (TP)	False Negatif (FN)
	FALSE	False Positif (FP)	True Negatif (TN)

Penjelasannya :

a. TP (True Positive), mencerminkan jumlah data positif yang berhasil terdeteksi dengan benar oleh sistem.
b. TN (True Negative), mencakup jumlah data negatif yang berhasil terdeteksi dengan benar oleh sistem..
c. FP (False Positive), mencerminkan jumlah data positif yang tidak terdeteksi dengan benar oleh sistem.
d. FN (False Negative), mengindikasikan jumlah data negatif yang tidak terdeteksi dengan benar oleh sistem.

Dengan memanfaatkan Confusion Matrix, kita dapat memperoleh informasi tambahan yang penting untuk mengevaluasi kinerja algoritma atau model yang digunakan. Informasi ini mencakup:

a. Accuracy
Merupakan nilai yang diperoleh dari sejauh mana kesamaan antara nilai prediksi dan nilai aktual. Rumus untuk menghitung nilai akurasi dapat dirumuskan sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

b. Recall / Sensitivity
Merupakan sebuah metrik evaluasi untuk model klasifikasi yang mengukur seberapa tepat model dalam mengenali kelas positif secara keseluruhan. Untuk menghitung recall dan sensitivity, caranya adalah dengan membagi jumlah prediksi benar positif dengan total data positif yang sebenarnya. Rumus untuk menghitung nilai recall dan sensitivity dapat dirumuskan sebagai berikut:

$$Recall\ /\ Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

c. Precission
Atau presisi merupakan suatu metrik evaluasi kinerja dari model klasifikasi yang mengukur seberapa tepat model dalam mengenali kelas positif. Nilai presisi dihitung dengan membagi jumlah prediksi benar positif dengan total jumlah prediksi positif. Rumus untuk menghitung nilai presisi dapat dirumuskan sebagai berikut:

$$Precission = \frac{TP}{TP+FP} \quad (4)$$

d. F1_score
Merupakan nilai rata-rata harmonik antara presisi dan recall. Perhitungan F1 Score dapat diwakili oleh formula berikut::

$$F1\ score = \frac{2(precision \times recall)}{precision+recall} \quad (5)$$

e. Gmean
Geometric Mean adalah ukuran yang mencoba untuk memaksimalkan keseimbangan antara Sensitivity dan Specificity, terutama untuk dataset yang tidak seimbang. Ada dua definisi yang umum digunakan, tergantung pada konteksnya:

1. Berdasarkan Precision dan Recall:

$$g_{PR} = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \quad (6)$$

2. Berdasarkan Sensitivity dan Specificity:

$$g_{SS} = \frac{\sqrt{TP*TN}}{\sqrt{(TP+FN)(TN+FP)}} \quad (7)$$

f. Specificity
juga dikenal sebagai True Negative Rate, adalah ukuran seberapa baik model kita dalam mengidentifikasi hasil negatif. Rumus untuk menghitung nilai specificity dapat dirumuskan sebagai berikut:

$$Specificity = \frac{TN}{TN+FP} \quad (8)$$

Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

3. HASIL DAN PEMBAHASAN

Bagian ini menggambarkan output yang diperoleh dari setiap tahapan yang telah dijelaskan dalam bagian metode, termasuk persiapan data, oversampling, seleksi fitur, klasifikasi Random Forest dan evaluasi confusion matrix.

3.1 Persiapan Data

Hasil yang diperoleh pada tahap persiapan data mencakup proses EDA, remove unused attributes and merge attributes, dataset transformation dan dataset details.

a. Exploratory Data Analysis (EDA): Dari hasil analisis univariat pada dataset penyakit TB yang digunakan, ditemukan bahwa label kelas tidak seimbang dan ada satu label kelas, yaitu 'gagal', yang memiliki jumlah data di bawah 1%, seperti yang ditunjukkan pada Gambar 2 berikut.

Gambar 2. Ketidakseimbangan Data pada Label Kelas

b. remove unused attributes and merge attributes: Atribut 'kode_unik' dan 'kab_kota' dieliminasi dari dataset karena kedua atribut tersebut tidak digunakan dalam analisis. Selanjutnya, atribut 'kelurahan' dan 'kecamatan' digabung menjadi satu atribut baru yang disebut "kelurahan_kecamatan" karena kedua atribut tersebut saling terkait. Setelah proses tersebut, urutan atribut diatur kembali untuk memastikan struktur data kembali seperti semula.

c. dataset transformation: Tahap selanjutnya melibatkan konversi fitur yang masih bersifat kategorikal menjadi format numerikal. Hal ini diperlukan karena model Random Forest hanya menerima input berupa tipe data numerikal. Proses perubahan jenis data dilakukan dengan menggunakan metode label encoding. Untuk fitur seperti 'bulan', 'jenis_kelamin', 'status_pekerjaan', 'jenis_fasyankes', 'tipe_diagnosis', 'tipe_TBC', 'riwayat_TBC', 'riwayat_dm', 'riwayat_HIV', 'paduan_OAT', 'sumber_obat' dan 'hasil_pengobatan', proses encoding dilakukan secara manual sesuai dengan yang dijelaskan dalam tabel 4.

Tabel 4. Penjelasan pengkodean fitur secara manual

No.	Nama Atribut	Deskripsi
1.	bulan	1: Januari 2: Februari 3: Maret 4: April 5: Mei 6: Juni 7: Juli 8: Agustus 9: September 10: Oktober 11: November 12: Desember
2.	jenis_kelamin	0: l (laki-laki) 1: p (perempuan)
3.	status_pekerjaan	0: tidak bekerja 1: pegawai swasta 2: pelajar/mahasiswa 3: pekerja lepas 4: wiraswasta 5: asn 6: tenaga profesional 7: anak-anak

Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

No.	Nama Atribut	Deskripsi
		8: pensiunan
		9: lainnya
4.	jenis_fasyankes	0: rumah sakit
		1: puskesmas
		2: bp4/ bbkpm/ bkpm
		3: lapas/rutan
		4: klinik
5.	tipe_diagnosis	0: terdiagnosis klinis
		1: terkonfirmasi bakteriologis
6.	tipe_TBC	0: tbc paru
		1: tbc ekstraparu
7.	riwayat_TBC	0: baru
		1: tidak diketahui
		2: kambuh
		3: diobati setelah putus berobat
		4: lain-lain
		5: diobati setelah gagal kategori 1
		6: pernah diobati tidak diketahui hasilnya
		7: diobati setelah gagal
		8: diobati setelah gagal kategori 2
8.	riwayat_dm	0: tidak diketahui
		1: tidak
		2: ya
9.	riwayat_HIV	0: tidak diketahui
		1: tidak
		2: ya
10.	paduan_OAT	0: kategori 1
		1: kategori anak
		2: kategori 2
		3: paduan tidak standar tb so
11.	sumber_obat	0: program tbc
		1: bayar sendiri
		2: asuransi
		3: lain-lain
12.	hasil_pengobatan	0: pengobatan lengkap
		1: sembuh
		2: putus berobat
		3: meninggal
		4: gagal

d. dataset details: Setelah tahap dataset transformation dilakukan, selanjutnya yaitu memvisualisasikan dan mengecek deskripsi, info dan juga shape data yang telah dirubah kedalam bentuk numerikal agar memudahkan dalam proses pengecekan data. Shape data dapat dilihat pada tabel 5. Pada tahap ini, atribut dan target dipecah dan dibagi menjadi data pelatihan dan data pengujian dengan rasio 80:20, yaitu 80% untuk data pelatihan dan 20% untuk data pengujian. Dalam penelitian ini, target atau label yang akan digunakan adalah 'hasil_pengobatan'. Label ini memiliki lima kelas, yaitu 0 yang berarti 'pengobatan lengkap', 1 yang berarti 'sembuh', 2 yang berarti 'putus berobat', 3 yang berarti 'meninggal', dan 4 yang berarti 'gagal'.

Tabel 5. Bentuk data

Bentuk Data	Jumlah Bentuk Data
Dimensi dari X	(6754, 14)
Dimensi dari y	(6754,)
Dimensi dari X_train	(5403, 14)
Dimensi dari y_train	(5403,)
Dimensi dari X_test	(1351, 14)
Dimensi dari y_test	(1351,)

3.2 Oversampling

Seperti yang ditunjukan pada tabel 6, dapat terlihat bahwa kelas '0' yang mewakili 'pengobatan lengkap' memiliki jumlah sampel sebanyak 3588 data sedangkan pada kelas '1' yang mewakili 'sembuh' memiliki jumlah sampel sebanyak 934 data. Dapat terlihat bahwa jumlah sampel data yang dimiliki masing-masing kelas memiliki

Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

perbedaan yang sangat jauh sehingga menyebabkan data yang dimiliki mengalami imbalance atau ketidakseimbangan.

Tabel 6. Jumlah sampel data latih per kelas sebelum oversampling

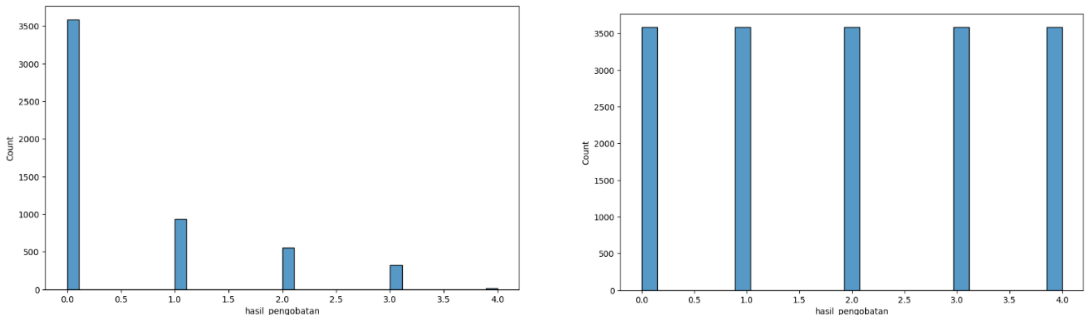
Kelas	Jumlah Sampel Data
0	3588
1	934
2	548
3	319
4	14

Oleh karena itu, penerapan oversampling menggunakan metode SMOTE menjadi suatu kebutuhan untuk mengatasi permasalahan tersebut. Oversampling dilaksanakan dengan tujuan untuk meningkatkan jumlah sampel dari kelas minoritas guna mencapai keseimbangan dalam hasil klasifikasi. Proses ini melibatkan penciptaan data sintetis baru dengan mengekstraksi sampel data minoritas yang ada, menggunakan sampel acak dari nilai k tetangga terdekat. Dengan demikian, hasil dari proses tersebut adalah peningkatan jumlah sampel data pada kelas minoritas, menyamai jumlah sampel data tertinggi, yaitu 3588 data. Rincian hasil oversampling dapat ditemukan pada tabel 7.

Tabel 7. Jumlah sampel data latih per kelas setelah oversampling

Kelas	Jumlah Sampel Data
0	3588
1	3588
2	3588
3	3588
4	3588

Hasil Visualisasi jumlah sampel data per kelas sebelum dan sesudah oversampling dapat diamati pada gambar 3.

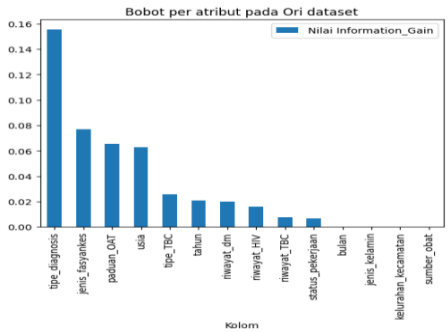


(a) (b)

Gambar 3. Visualisasi Jumlah Data Latih (a)sebelum oversampling, (b)sesudah oversampling

3.3 Seleksi Fitur

Seleksi fitur dilakukan setelah proses oversampling. Dari data latih original dataset, 10 dari 14 fitur dengan bobot nilai tertinggi diperoleh, yaitu; ‘tipe_diagnosis’, ‘jenis_fasyankes’, ‘paduan_OAT’, ‘usia’, ‘tipe_TBC’, ‘tahun’, ‘riwayat_dm’, ‘riwayat_HIV’, ‘riwayat_TBC’, dan ‘status_pekerjaan’. Fitur lainnya seperti ‘bulan’, ‘jenis_kelamin’, ‘kelurahan_kecamatan’ dan ‘sumber_obat’ mengikuti setelahnya. Detail ini dapat ditemukan pada gambar 4.

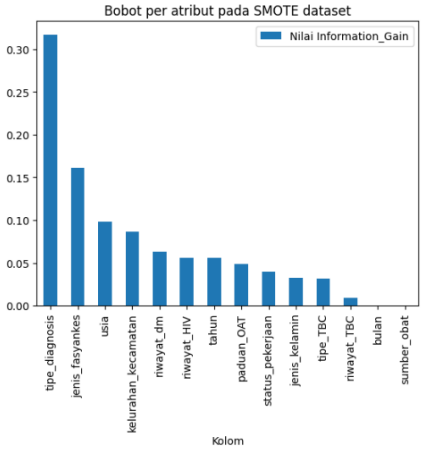


Gambar 4. Bobot Atribut pada Original Dataset

Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

Sedangkan hasil dari seleksi fitur pada data latih yang telah melalui proses oversampling pada dataset diperoleh 12 dari 14 fitur dengan bobot nilai tertinggi yaitu; 'tipe_diagnosis', 'jenis_fasyankes', 'usia', 'kelurahan_kecamatan', 'riwayat_dm', 'riwayat_HIV', 'tahun', 'paduan_OAT', 'status_pekerjaan', 'jenis_kelamin', 'tipe_TBC' dan 'riwayat_TBC'. Fitur lainnya seperti 'bulan' dan 'sumber_obat' mengikuti setelahnya. Hasil ini dapat diamati pada gambar 5.



Gambar 5. Bobot Atribut dalam dataset yang dihasilkan oleh metode SMOTE

Berdasarkan dua konfigurasi yang telah diuji dalam proses seleksi fitur, ada dua fitur, yaitu 'tipe_diagnosis' dan 'jenis_fasyankes', yang selalu mendapatkan peringkat yang sama secara berurutan. Ini menunjukkan bahwa kedua fitur tersebut memiliki signifikansi yang konsisten dalam kedua konfigurasi tersebut.

3.4 Klasifikasi Random Forest

Hasil dari empat pengujian model telah dilakukan dengan variasi pada dataset asli dan dataset yang telah melalui proses SMOTE, serta dengan dan tanpa penggunaan metode Information Gain. Pengujian pertama adalah model Random Forest yang dievaluasi menggunakan dataset asli tanpa menerapkan metode Information Gain. Pengujian kedua adalah model Random Forest yang dievaluasi dengan mempertimbangkan dataset asli dan menerapkan metode Information Gain pada proses seleksi fitur. Pengujian ketiga adalah model Random Forest yang melibatkan dataset yang telah melalui proses SMOTE untuk menangani ketidakseimbangan kelas, namun tanpa menggunakan Information Gain. Dan pengujian terakhir adalah model Random Forest yang dievaluasi menggunakan dataset SMOTE dengan penerapan metode Information Gain. Pada proses ini, confusion matrix digunakan untuk mengevaluasi hasil model. Selanjutnya, pemilihan atribut dilakukan berdasarkan peringkat yang diperoleh dari seleksi fitur, dari yang tertinggi hingga terendah. Dengan cara ini, didapatkan top 2 fitur, top 4 fitur, top 6 fitur, top 8 fitur, top 10 fitur, top 12 fitur, dan top 14 fitur. Proses ini diterapkan pada dataset asli yang digunakan sebagai data pelatihan dan juga pada data pelatihan yang telah melalui proses oversampling. Hal ini dilakukan untuk menjamin variasi dalam data pelatihan.

Berdasarkan pengujian pertama, hasil yang didapatkan adalah sebagai berikut: akurasi sebesar 74%, recall sebesar74%, precision sebesar 68%, f1 score sebesar 69%, gmean score sebesar 69%, sensitivity sebesar 74% dan specifity sebesar 64%. Berdasarkan pengujian kedua, hasil yang didapatkan adalah sebagai berikut: akurasi tertinggi terdapat pada top 14 fitur sebesar 75%, recall tertinggi terdapat pada top 14 fitur sebesar 75%, precision tertinggi terdapat pada top 14 fitur sebesar 72%, f1 score tertinggi terdapat pada top 12 fitur sebesar 69%, gmean score tertinggi terdapat pada top 12 fitur sebesar 70%, sensitivity tertinggi terdapat pada top 14 sebesar 75% dan specifity tertinggi terdapat pada top 2 sebesar 67%. Berdasarkan pengujian ketiga, hasil yang didapatkan adalah sebagai berikut: akurasi sebesar 60%, recall sebesar 60%, precision sebesar 69%, f1 score sebesar 63%, gmean score sebesar 70%, sensitivity sebesar 60% dan specifity sebesar 81%. Berdasarkan pengujian keempat, hasil yang didapatkan adalah sebagai berikut: akurasi tertinggi terdapat pada top 10 dan top 12 fitur sebesar 75%, recall tertinggi terdapat pada top 10 dan top 12 fitur sebesar 75%, precision tertinggi terdapat pada top 12 fitur sebesar 72%, f1 score tertinggi terdapat pada top 10 fitur sebesar 69%, gmean score tertinggi terdapat pada top 2 fitur sebesar 70%, sensitivity tertinggi terdapat pada top 10 dan top 12 fitur sebesar 75% dan specifity tertinggi terdapat pada top 2 fitur sebesar 67%.

3.5 Evaluasi Confusion Matrix

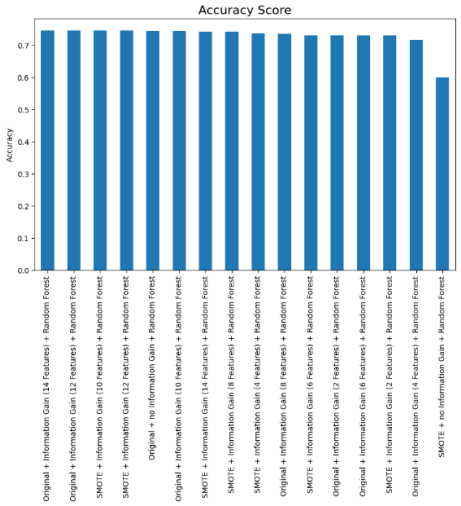
Setelah mendapatkan hasil model dan metrik evaluasi dari setiap pengujian yang telah dilakukan, evaluasi menyeluruh dari confusion matrix sangat diperlukan. Evaluasi ini memberikan gambaran komprehensif tentang kinerja model dalam berbagai aspek, termasuk akurasi model dalam memprediksi kelas yang benar, sensitivitas model dalam mengidentifikasi kelas positif, dan kemampuan model dalam memprediksi kelas negatif. Evaluasi

Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

ini sangat penting untuk memastikan bahwa model memiliki kinerja yang baik dan dapat diandalkan dalam memprediksi hasil, serta membantu dalam peningkatan dan penyesuaian model di masa mendatang.

Berdasarkan gambar 6, hasil akurasi tertinggi diperoleh pada model 'Original + Information Gain (14 Features) + Random Forest' sebesar 0.746114 atau 74,6% dan diikuti oleh model 'Original + Information Gain (12 Features) + Random Forest' sebesar 0.745374 atau 74,5%. Sedangkan untuk hasil akurasi terendah diperoleh pada model 'SMOTE + no Information Gain + Random Forest' sebesar 0.598816 atau 60%.



Gambar 6. Hasil Akurasi Model dari yang Tertinggi hingga Terendah

Berdasarkan tabel 8, yang merupakan hasil dari pengujian keseluruhan, didapatkan hasil sebagai berikut: akurasi tertinggi terdapat pada model 'Original + Information Gain (14 Features) + Random Forest' sebesar 0.746114 atau 75%, recall tertinggi terdapat pada model 'Original + Information Gain (14 Features) + Random Forest' sebesar 0.746114 atau 75%, precision tertinggi terdapat pada model 'Original + Information Gain (14 Features) + Random Forest' sebesar 0.723686 atau 72%, f1 score tertinggi terdapat pada model 'Original + Information Gain (12 Features) + Random Forest' sebesar 0.691553 atau 69%, gmean score tertinggi terdapat pada model 'Original + Information Gain (12 Features) + Random Forest' sebesar 0.701762 atau 70%, sensitivity tertinggi terdapat pada model 'Original + Information Gain (14 Features) + Random Forest' sebesar 0.746114 atau 75% dan specifity tertinggi terdapat pada model 'SMOTE + no Information Gain + Random Forest' sebesar 0.807731 atau 81%.

Tabel 8. Hasil pengujian model klasifikasi

Model Klasifikasi	accuracy score	recall score	precision score	f1 score	gmean score	sensitivity score	Specifity score
Original + no Information Gain + Random Forest	0.743153	0.743153	0.680918	0.687259	0.690499	0.743153	0.641576
Original + Information Gain (2 Features) + Random Forest	0.729830	0.729830	0.628135	0.673207	0.700593	0.729830	0.672528
Original + Information Gain (4 Features) + Random Forest	0.716506	0.716506	0.635976	0.666923	0.688428	0.716506	0.661451
Original + Information Gain (6 Features) + Random Forest	0.729830	0.729830	0.660699	0.679467	0.697546	0.729830	0.666689
Original + Information Gain (8 Features) + Random Forest	0.735751	0.735751	0.659085	0.684370	0.691995	0.735751	0.650841
Original + Information Gain (10 Features) + Random Forest	0.743153	0.743153	0.686785	0.689329	0.700616	0.743153	0.660513

Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

Model Klasifikasi	accuracy score	recall score	precision score	f1 score	gmean score	sensitivity score	Specifity score
Original + Information Gain (12 Features) + Random Forest	0.745374	0.745374	0.706575	0.691553	0.701762	0.745374	0.660702
Original + Information Gain (14 Features) + Random Forest	0.746114	0.746114	0.723686	0.690122	0.695512	0.746114	0.648341
SMOTE + no Information Gain + Random Forest	0.598816	0.598816	0.687027	0.626250	0.695472	0.598816	0.807731
SMOTE + Information Gain (2 Features) + Random Forest	0.729830	0.729830	0.628135	0.673207	0.700593	0.729830	0.672528
SMOTE + Information Gain (4 Features) + Random Forest	0.737232	0.737232	0.672931	0.686007	0.697374	0.737232	0.659671
SMOTE + Information Gain (6 Features) + Random Forest	0.730570	0.730570	0.650984	0.676624	0.680369	0.730570	0.633617
SMOTE + Information Gain (8 Features) + Random Forest	0.741673	0.741673	0.680627	0.686465	0.689736	0.741673	0.641435
SMOTE + Information Gain (10 Features) + Random Forest	0.744634	0.744634	0.696585	0.689935	0.696828	0.744634	0.652092
SMOTE + Information Gain (12 Features) + Random Forest	0.744634	0.744634	0.723565	0.687841	0.693958	0.744634	0.646731
SMOTE + Information Gain (14 Features) + Random Forest	0.742413	0.742413	0.692475	0.685554	0.691918	0.742413	0.644857

4. KESIMPULAN

Dari hasil penelitian terkait identifikasi model klasifikasi menggunakan Random Forest dan fitur seleksi Information Gain yang paling efektif dalam memprediksi faktor-faktor yang mempengaruhi hasil pengobatan pasien penyakit Tuberkulosis (TB), dengan mempertimbangkan ketidakseimbangan dataset yang kemudian diseimbangkan menggunakan teknik oversampling yaitu Synthetic Minority Oversampling Technique (SMOTE), dapat ditarik kesimpulan bahwa model klasifikasi Random Forest menunjukkan performa terbaik dengan menggunakan seluruh fitur atau atribut dari bobot terbesar hingga terkecil yaitu; 'tipe_diagnosis', 'jenis_fasyankes', 'usia', 'kelurahan_kecamatan', 'riwayat_dm', 'riwayat_HIV', 'tahun', 'paduan_OAT', 'status_pekerjaan', 'jenis_kelamin', 'tipe_TBC', 'riwayat_TBC', 'bulan' dan 'sumber_obat' pada dataset asli penyakit TB di Kota Semarang. Tingkat recall dan akurasi mencapai 75%. Hasil tersebut lebih unggul dibandingkan model klasifikasi TB di Kota Semarang yang memanfaatkan dataset oversampling menggunakan SMOTE dan hanya menggunakan 10-12 atribut teratas, dengan hasil recall dan akurasi mencapai 74%. Adapun saran untuk penelitian yang ingin mengembangkan labih lanjut adalah pada penelitian selanjutnya untuk mengintegrasikan data pasien penyakit TB di Kota Semarang dengan data tambahan lain seperti data sosial, pendapatan, data epidemiologi dan riwayat pengobatan pasien. Hal ini diharapkan dapat membuat dataset lebih optimal dan seimbang. Serta pada penelitian selanjutnya juga disarankan untuk mencoba metode pengolahan data yang berbeda dan menganalisis faktor-faktor apa saja yang mempengaruhi penyakit TB di Kota Semarang. Diharapkan hal ini dapat meningkatkan performa model klasifikasi dan pemahaman tentang penyakit tersebut.

Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

Penelitian ini dapat menjadi sebuah upaya atau langkah dalam meningkatkan standar kesehatan masyarakat di Kota Semarang dan secara lebih luas di Indonesia.

UCAPAN TERIMAKASIH

Dengan rasa syukur kepada Tuhan yang Maha Esa, atas segala rahmat, petunjuk, dan perlindungan-Nya, penelitian ini telah selesai, berkat dukungan yang sangat berharga dari banyak pihak. Oleh karena itu, peneliti ingin menyampaikan terima kasih kepada: Pihak kampus yang telah menyediakan sarana dan prasaran yang diperlukan oleh peneliti; Abu Salam, M.Kom, selaku dosen pembimbing yang membantu dalam penelitian ini serta para dosen lainnya yang ikut membantu; dan kedua orang tua yang saya cintai yang telah memberikan dorongan semangat dan doa kepada saya agar tidak mudah menyerah dalam menyelesaikan penelitian ini. Terakhir, terima kasih kepada peneliti sendiri yang telah menyelesaikan penelitian ini. Semoga Tuhan Yang Maha Esa memberikan ganjaran yang lebih besar kepada mereka semua, dan pada akhirnya, peneliti berharap bahwa laporan tugas akhir ini dapat memberikan manfaat dan nilai sesuai dengan tujuannya.

REFERENCES

[1] World Health Organization, “Global tuberculosis report 2022,” 2022. Accessed: Nov. 06, 2023. [Online]. Available: https://www.who.int/publications/i/item/9789240061729
[2] S. Adhanty and S. Syarif, “Kepatuhan Pengobatan pada Pasien Tuberkulosis dan Faktor-Faktor yang Mempengaruhinya: Tinjauan Sistematis,” Jurnal Epidemiologi Kesehatan Indonesia, vol. 7, no. 1, p. 7, Jun. 2023, doi: 10.7454/epidkes.v7i1.6571.
[3] S. D. Pralambang and S. Setiawan, “Faktor Risiko Kejadian Tuberkulosis di Indonesia,” Jurnal Biostatistik, Kependudukan, dan Informatika Kesehatan, vol. 2, no. 1, p. 60, Nov. 2021, doi: 10.51181/bikfokes.v2i1.4660.
[4] M. S. D. Wijaya, M. F. J. Mantik, and N. H. Rampengan, “Faktor Risiko Tuberkulosis pada Anak,” e-CliniC, vol. 9, no. 1, Jan. 2021, doi: 10.35790/ecl.v9i1.32117.
[5] “Laporan Kasus Tuberkulosis (TBC) Global Dan Indonesia 2022,” Yayasan KNCV Indonesia. Accessed: Nov. 06, 2023. [Online]. Available: https://yki4tbc.org/laporan-kasus-tbc-global-dan-indonesia-2022/
[6] S. M. E. dan T. Sulistyo, Laporan Program Penanggulangan Tuberkulosis Tahun 2021 – Perpustakaan Kemenkes RI, 2022nd ed. Jakarta: Kementerian Kesehatan Republik Indonesia, 2021. Accessed: Nov. 06, 2023. [Online]. Available: https://perpustakaan.kemkes.go.id/books/laporan-program-penanggulangan-tuberkulosis-tahun-2021/
[7] Sp. P. dr. Mochamad Abdul Hakam, M. K. dr. Noegroho Edy Rijanto, S. M. K. M. S. Hanif Pandu Suhito, and S. K. M. Prahita Indriana Raniasmi, “PROFIL KESEHATAN KOTA SEMARANG 2021,” Semarang, Jun. 2022. [Online]. Available: www.dinkes.semarangkota.go.id
[8] “Data_SITB_Bersih.” Dinas Kesehatan Kota Semarang, Semarang, Nov. 10, 2023.
[9] T. D. Hartanto, L. D. Saraswati, M. S. Adi, and A. Udiyono, “Analisis Spasial Persebaran Kasus Tuberkulosis Paru Di Kota Semarang Tahun 2018,” Jurnal Kesehatan Masyarakat, vol. 7, no. 4, pp. 719–727, Oct. 2019, doi: 10.14710/JKM.V7I4.25123.
[10] B. Remeseiro and V. Bolon-Canedo, “A review of feature selection methods in medical applications,” Comput Biol Med, vol. 112, p. 103375, Sep. 2019, doi: 10.1016/j.compbiomed.2019.103375.
[11] B. S. Prakoso, D. Rosiyadi, D. Aridarma, H. S. Utama, F. Fauzi, and M. A. N. Qhomar, “OPTIMALISASI KLASIFIKASI BERITA MENGGUNAKAN FEATURE INFORMATION GAIN UNTUK ALGORITMA NAIVE BAYES TERHUBUNG RANDOM FOREST,” Jurnal Pilar Nusa Mandiri, vol. 15, no. 2, pp. 211–218, Sep. 2019, doi: 10.33480/pilar.v15i2.684.
[12] M. Muqorobin, K. Kusrini, and E. T. Luthfi, “OPTIMASI METODE NAIVE BAYES DENGAN FEATURE SELECTION INFORMATION GAIN UNTUK PREDIKSI KETERLAMBATAN PEMBAYARAN SPP SEKOLAH,” Jurnal Ilmiah SINUS, vol. 17, no. 1, p. 1, Jan. 2019, doi: 10.30646/sinus.v17i1.378.
[13] A. Kulsumarwati, I. Purnamasari, and B. A. Darmawan, “Penerapan SVM dan Information Gain Pada Analisis Sentimen Pelaksanaan Pilkada Saat Pandemi,” Jurnal Teknologi Informatika dan Komputer, vol. 7, no. 2, pp. 101–109, Sep. 2021, doi: 10.37012/jtik.v7i2.641.
[14] A. Nugroho and E. Rilvani, “Penerapan Metode Oversampling SMOTE Pada Algoritma Random Forest Untuk Prediksi Kebangkrutan Perusahaan,” Techno.Com, vol. 22, no. 1, pp. 207–214, Feb. 2023, doi: 10.33633/tc.v22i1.7527.
[15] N. Hidayati, A. I. Rizmayanti, C. B. S. Dewi, R. Fatmasari, and W. Gata, “Penerapan Algoritma Klasterisasi dan Klasifikasi pada Tingkat Kepentingan Sistem Pembelajaran di Universitas Terbuka,” Swabumi, vol. 8, no. 2, pp. 134–142, Sep. 2020, doi: 10.31294/swabumi.v8i2.8385.
[16] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, “Loan default prediction using decision trees and random forest: A comparative study,” IOP Conf Ser Mater Sci Eng, vol. 1022, no. 1, p. 012042, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012042.
[17] S. Diantika, “PENERAPAN TEKNIK RANDOM OVERSAMPLING UNTUK MENGATASI IMBALANCE CLASS DALAM KLASIFIKASI WEBSITE PHISHING MENGGUNAKAN ALGORITMA LIGHTGBM,” JATI (Jurnal Mahasiswa Teknik Informatika), vol. 7, no. 1, pp. 19–25, Jan. 2023, doi: 10.36040/jati.v7i1.6006.
[18] R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results,” in 2020 11th International Conference on Information and Communication Systems (ICICS), IEEE, Apr. 2020, pp. 243–248. doi: 10.1109/ICICS49469.2020.239556.
[19] D. Elreedy and A. F. Atiya, “A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance,” Inf Sci (N Y), vol. 505, pp. 32–64, Dec. 2019, doi: 10.1016/j.ins.2019.07.070.

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 70-82
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7122

[20] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," Appl Soft Comput, vol. 76, pp. 380–389, Mar. 2019, doi: 10.1016/j.asoc.2018.12.024.
[21] M. R. Hasibuan and M. Marji, "Pemilihan Fitur dengan Information Gain untuk Klasifikasi Penyakit Gagal Ginjal menggunakan Metode Modified K-Nearest Neighbor (MKNN)," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 3, no. 11, pp. 10435–10443, Jan. 2020, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/6691
[22] anikakapoor, "ML | Data Preprocessing in Python - GeeksforGeeks," GeeksforGeeks. Accessed: Nov. 10, 2023. [Online]. Available: https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/
[23] R. Ordila, R. Wahyuni, Y. Irawan, and M. Yulia Sari, "PENERAPAN DATA MINING UNTUK PENGELOMPOKAN DATA REKAM MEDIS PASIEN BERDASARKAN JENIS PENYAKIT DENGAN ALGORITMA CLUSTERING (Studi Kasus : Poli Klinik PT.Inecda)," Jurnal Ilmu Komputer, vol. 9, no. 2, pp. 148–153, Oct. 2020, doi: 10.33060/JIK/2020/Vol9.Iss2.181.
[24] L. Qadrini, H. Hikmah, and M. Megasari, "Oversampling, Undersampling, Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017," Journal of Computer System and Informatics (JoSYC), vol. 3, no. 4, pp. 386–391, Sep. 2022, doi: 10.47065/josyc.v3i4.2154.
[25] A. Muqtadir and D. S. Purnia, "Pemanfaatan Metode SMOTE dan PSO Untuk Mengoptimalkan Tingkat Akurasi Klasifikasi Kepuasan Pelanggan," IJCIT (Indonesian Journal on Computer and Information Technology), vol. 8, no. 1, Aug. 2023, doi: 10.31294/ijcit.v8i1.11657.
[26] F. A. Fauzi, M. T. Furqon, and N. Yudistira, "Klasifikasi Jenis Tanaman Tembakau di Indonesia menggunakan Naïve Bayes dengan Seleksi Fitur Information Gain," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 5, no. 2, pp. 698–703, Feb. 2021, [Online]. Available: http://j-ptiik.ub.ac.id
[27] S. Zulaikhah Hariyanti Rukmana, A. Aziz, and W. Harianto, "OPTIMASI ALGORITMA K-NEAREST NEIGHBOR (KNN) DENGAN NORMALISASI DAN SELEKSI FITUR UNTUK KLASIFIKASI PENYAKIT LIVER," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 6, no. 2, pp. 439–445, Aug. 2022, doi: 10.36040/jati.v6i2.4722.
[28] C. M. Yeşilkanat, "Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm," Chaos Solitons Fractals, vol. 140, p. 110210, Nov. 2020, doi: 10.1016/j.chaos.2020.110210.
[29] I. Romli and A. T. Zy, "Penentuan Jadwal Overtime Dengan Klasifikasi Data Karyawan Menggunakan Algoritma C4.5," Jurnal Sains Komputer & Informatika (J-SAKTI, vol. 4, no. 2, pp. 694–702, 2020.
[30] N. Hadianto, H. B. Novitasari, and A. Rahmawati, "KLASIFIKASI PEMINJAMAN NASABAH BANK MENGGUNAKAN METODE NEURAL NETWORK," Jurnal Pilar Nusa Mandiri, vol. 15, no. 2, pp. 163–170, Sep. 2019, doi: 10.33480/pilar.v15i2.658.

Submitted: **12/12/2023**; Accepted: **05/01/2024**; Published: **09/01/2024**