INOTEK, Vol.8 Agustus 2024
ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

Klasifikasi Risiko Kambuhnya Kanker Tiroid Menggunakan Algoritma *Random Forest*

Diterima:
10 Juni 2024
Revisi:
10 Juli 2024
Terbit:
1 Agustus 2024

[1]Muhammad Faruqziddan, [2]Ewanda Herdika Septa Aulia, [3]Salsabila Dini Azzahra, [4]Aidina Ristyawan, [5]Erna Daniati
[1]Fakultas Teknik dan Ilmu Komputer, [2]Sistem Informasi, [3]Universitas Nusantara PGRI Kediri
[1]faruqziddan@gmail.com,[2]ewandaherdika@gmail.com,
[3]salsazhrra1122@gmail.com,
[4]aidinaristi@unpkediri.ac.id,[5]ernadaniati@unpkediri.ac.id

Abstrak— Kanker Tiroid merupakan sebuah jenis kanker yang berkembang dalam kelenjar tiroid, organ kecil yang terletak di bagian depan leher. Meskipun tingkat kematian akan kanker jenis tersebut rendah tetapi risiko kambuhnya kanker tiroid menjadi salah satu masalah lain yang perlu diatasi. Untuk membantu mengevaluasi kambuhnya kanker tiroid pada pasien tujuan penelitian ini mengembangkan sebuah model algoritma dengan memanfaatkan dataset dari *UCI Machine Learning Repository*. Dataset tersebut termasuk kedalam kategori klasifikasi dan Algoritma yang akan digunakan adalah *Random Forest*. Setelah dilakukan penelitian sesuai dengan *Knowlegde Discovery in Databases* (KDD), algoritma *Random Forest* memiliki Sensitifitas sebesar 98,39%, Spesifisitas sebesar 96,77%, *Precision* sebesar 96,83%, *Area Under the Curve* (AUC) sebesar 97,6%, dan Accuracy sebesar 97,5%. Dengan hasil yang ditemukan, algoritma *Random Forest* telah terbukti efektif dalam mengembangkan model untuk membantu mengevaluasi risiko kambuhnya kanker tiroid pada pasien
Kata Kunci— KDD; Kanker Tiroid; *Random Forest*

Abstract— *Thyroid cancer is a type of cancer that develops in the thyroid gland, a small organ located at the front of the neck. Although the death rate of this type of cancer is low, the risk of recurrence of thyroid cancer is another problem that needs to be overcome. To help evaluate the recurrence of thyroid cancer in patients. The purpose of this study is to develop an algorithm model by utilizing datasets from the UCI Machine Learning Repository. The dataset is included in the classification category and the algorithm to be used is Random Forest. After conducting research in accordance with Knowledge Discovery in Databases (KDD), the Random Forest algorithm has a Sensitivity of 98.39%, Specificity of 96.77%, Precision of 96.83%, AUC of 97.6%, and Accuracy of 97.5%. With the results found, the Random Forest algorithm has proven effective in developing models to help evaluate the risk of thyroid cancer recurrence in patients*.
Keywords*— KDD; Thyroid Cancer; Random Forest*

Penulis Korespondensi:

Aidina Ristyawan,
Sistem Informasi,
Universitas Nusantara PGRI Kediri,
Email: adinaristi@unpkediri.ac.id
ID Orcid: https://orcid.org/0009-0003-2712-1507
Handphone: 081232624460

I. PENDAHULUAN

Kanker tiroid adalah jenis kanker yang mempengaruhi kelenjar tiroid, yang terletak di bagian depan leher sedikit di bawah laring dan memiliki bentuk seperti kupu-kupu. Prevalensi kanker tiroid berkisar antara 0,85% hingga 2,5% dari seluruh kasus kanker tiroid, dengan perbandingan 1:3 antara laki-laki dan perempuan [1]. Dapat diartikan bahwa kejadian kanker tiroid lebih umum pada perempuan. Meskipun umumnya terjadi pada rentang usia 20-50 tahun, kanker tiroid dapat terjadi pada semua rentang usia.

Jika kelenjar tiroid tidak berfungsi dengan benar, hal itu dapat mengganggu fungsi organ tubuh lainnya. Fungsi kelenjar tiroid termasuk meningkatkan metabolisme kalori, mengubah makanan menjadi energi, dan mengatur detak jantung. Usia dan daerah endemik bukan satu-satunya faktor yang dapat mempengaruhi kanker tiroid, gen dan jenis kelamin juga dapat mempengaruhi [2].

Walaupun tingkat kematian akan kanker teroid ini tergolong rendah, akan tetapi mengetahui apakah pasien memiliki risiko kambuh kanker tiroid ini menjadi salah satu masalah yang perlu diselesaikan [3]. Untuk mempercepat proses identifikasi dan meningkatkan akurasi diagnosis, diperlukan sistem diagnosis yang baik dan dapat diandalkan [4]. Oleh karena itu pada penelitian ini akan dibangun sebuah model algoritma klasifikasi pasien yang memiliki risiko kambuh kanker tiroid atau tidak. Untuk melakukan klasifikasi sistem, metode yang tepat untuk mengelola pengetahuan yang dikumpulkan dari model algoritma diperlukan untuk mendapatkan hasil yang akurat [5].

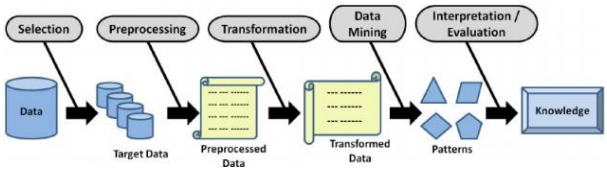
Data diperoleh dari pemeriksaan kanker tiroid dalam penelitian yang berlangsung selama 15 tahun yang diperoleh dari situs *UCI Machine Learning Repository* dengan judul *Differentiated Thyroid Cancer Recurrence* [6]. Metode pembuatan model algoritma membutuhkan data *training* untuk menghasilkan model yang tepat [7]. Klasifikasi dapat didefinisikan sebagai proses pelatihan atau pembelajaran objek data untuk mengategorikannya ke dalam salah satu dari beberapa kelas yang ada [8]. Metode penelitian yang digunakan adalah metode *Knowledge Discovery In Database* (KDD) karena metode tersebut sangat cocok digunakan dalam penelitian tentang data mining. *Knowledge Discovery In Database* (KDD) merupakan metode untuk memperoleh pengetahuan dari database yang ada [9].

Terdapat banyak algoritma pada klasifikasi salah satunya adalah algoritma *Random forest*. Algoritma *Random Forest* merupakan gabungan antara metode *Bagging* dan *Random Subspaces*. Algoritma ini telah terbukti berhasil dalam berbagai masalah regresi dan klasifikasi dalam beberapa tahun terakhir, dan menjadi salah satu algoritma *machine learning* paling efektif yang digunakan secara luas di berbagai bidang[10].

II. METODE

2.1 *Knowledge Discovery In Databases* (KDD)

Knowledge Discovery in Databases (KDD) adalah metodologi data mining yang digunakan untuk menghasilkan pengetahuan untuk membuat keputusan [11]. Untuk memberikan pemahaman yang lebih jelas tentang alur penelitian, ilustrasi grafis pada gambar 1:



Gambar 1. Proses *Knowledge Discovery In Databases* (KDD)

2.1.1. Data *Selection*

Karena atribut yang digunakan hanya berdampak pada kelas, proses data mining tidak menggunakan semua atribut database, perlu adanya analisis terhadap atribut apakah semua atribut tersebut berpengaruh atau tidak pada kelas dataset *Differentiated Thyroid Cancer Recurrence*. Fokus tahap ini adalah menentukan data target dan subset dari data sampel atau variabel [12].

2.1.2. Preprocessing

Preprocessing data adalah tahap pembersihan data dengan memeriksa duplikat data dan tingkat konsistensi [13]. Hal ini dilakukan agar data *Differentiated Thyroid Cancer Recurrence* lebih siap untuk proses data transformation dan penerapan algoritma *Random forest*, hal ini memastikan data menjadi lebih akurat dan relevan. Pada Dataset *Differentiated Thyroid Cancer Recurrence* akan dilakukan pengecekan missing value, pengecekan dan penanganan data duplikat serta penanganan balancing data menggunakan metode *SMOTEENN*.

2.1.3. Transformation

Tahapan selanjutnya setelah preprocessing adalah *transformation*, yaitu merubah format data agar dapat digunakan [14]. Tahap transformasi termasuk pemrosesan data skala, normalisasi, dan reduksi dimensi [15]. Proses encoding juga termasuk tahap transformasi data yaitu mengubah data nominal menjadi data numerik. Karena pada algoritma *Random Forest* sendiri tidak bisa menggunakan data bertipe nominal.

2.1.4. Data Mining

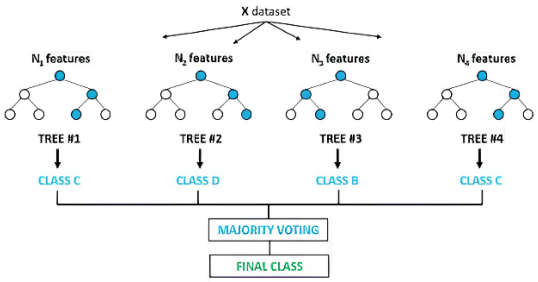
Pada tahap ini adalah melakukan pengujian dengan data uji untuk mengevaluasi keakuratan klasifikasi yang dilakukan [16]. Data yang digunakan yaitu dataset *Differentiated Thyroid Cancer Recurrence* yang telah melalui proses data *selection, preprocessing* dan *transformation.*

2.1.5. Evaluation

Tahap ini adalah proses evaluasi sekaligus tahap terakhir terhadap hasil dari proses yang sudah dilakukan [17].

2.2. Random Forest

Algoritma yang digunakan untuk mengklasifikasikan dataset *Differentiated Thyroid Cancer Recurrence* adalah algoritma *Random Forest. Random forest* adalah kumpulan dari pohon keputusan untuk regresi atau klasifikasi yang tidak dipangkas, dibentuk dengan cara memilih sampel acak dari data [18]. Pada gambar 2 merupakan ilustrasi Algoritma *Random Forest*.



Gambar 2. Ilustrasi Algoritma *Random Forest* [19]

Cara kerja Algoritma *Random Forest*; (1) menggunakan pengambilan sampel acak untuk membuat setiap pohon keputusan, (2) setiap pohon menggunakan subset fitur yang dipilih secara acak untuk prediksi, dan (3) menggabungkan hasil prediksi semua pohon dengan voting terbanyak untuk klasifikasi atau rata-rata untuk regresi [20].

Secara singkat Algoritma *Random Forest* bekerja dengan membangun banyak pohon keputusan dan menggabungkan hasil prediksi setiap pohon melalui voting mayoritas untuk klasifikasi atau rata-rata untuk regresi.

III. HASIL DAN PEMBAHASAN

Dalam konteks kanker tiroid, risiko kambuh menjadi salah satu masalah utama. Memahami faktor-faktor yang berkontribusi terhadap kekambuhan sangat penting untuk menyesuaikan manajemen pasien dan strategi tindak lanjut. Dataset ini menyediakan informasi klinis dan patologis yang rinci yang dapat digunakan untuk mengembangkan model prediktif untuk risiko kekambuhan, yang dapat membantu dalam mengidentifikasi pasien yang memiliki risiko kambuh. Tujuan dari penelitian ini adalah memanfaatkan dataset untuk membangun model algoritma yang dapat memprediksi kemungkinan kambuhnya kanker Tiroid.

3.1. Data *Selection*

Dataset tersebut terdiri dari data pasien sebanyak 383 pasien dimana sebanyak 108 pasien terdeteksi mengalami kambuh kanker tiroid dan 275 pasien terdeteksi tidak mengalami kambuh. Dari 383 pasien terdiri dari 17 atribut. Penelitian ini dilakukan untuk meningkatkan pemahaman tentang risiko kambuhnya kanker tiroid. Deskripsi dataset dapat dilihat pada Tabel 1.

Tabel 1. Deskripsi Dataset

Nama Atribut	Penjelasan Atribut	Sumber
Age	Usia pasien pada saat diagnosis atau pengobatan.	*UCI Repository*
Gender	Jenis kelamin pasien	*UCI Repository*
Smoking	Status merokok pasien	*UCI Repository*
Hx Smoking	Riwayat merokok pasien	*UCI Repository*
Hx Radiothreapy	Riwayat pengobatan radioterapi	*UCI Repository*
Thyroid Function	Status fungsi tiroid	*UCI Repository*
Physical Examination	Temuan dari pemeriksaan fisik pasien, yang mungkin mencakup palpasi kelenjar tiroid dan struktur sekitarnya.	*UCI Repository*
Adenopathy	Ada tidaknya pembesaran kelenjar getah bening	*UCI Repository*
Pathology	Jenis spesifik kanker tiroid yang ditentukan oleh pemeriksaan patologi dari sampel biopsi.	*UCI Repository*
Focality	Apakah kanker bersifat unifokal (terbatas pada satu lokasi) atau multifokal (hadir di beberapa lokasi).	*UCI Repository*
Risk	Kategori risiko kanker berdasarkan berbagai faktor.	*UCI Repository*
T	Klasifikasi tumor berdasarkan ukuran dan tingkat invasi ke struktur terdekat.	*UCI Repository*
N	Keterlibatan kelenjar getah bening.	*UCI Repository*
M	Klasifikasi metastasis yang menunjukkan keberadaan atau ketiadaan metastasis jauh.	*UCI Repository*
Stage	Stadium kanker berdasarkan klasifikasi TNM	*UCI Repository*
Response	Respon pasien terhadap pengobatan	*UCI Repository*
Recurred	Mengalami kambuh atau tidak.	*UCI Repository*

Setelah melakukan Analisa terhadap seluruh atribut, ditemukan kesimpulan bahwa seluruh atribut akan digunakan dalam proses data mining. Karena atribut-atribut yang tersedia semuanya memiliki kemungkinan untuk menjadi penyebab kambuhnya kanker tiroid. 4 baris pertama dataset dapat dilihat pada gambar 3.

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Thyroid Function	Physical Examination	Adenopathy	Pathology	Focality	Risk	T	N	M	Stage	Response	Recurred
0	27	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Indeterminate	No
1	34	F	No	Yes	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
2	30	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
3	62	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
4	62	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Excellent	No

Gambar 3. 4 baris pertama Dataset.

3.2. *Preprocessing*.

Pada tahap ini akan dilakukan pengecekan terhadap *missing value*, data duplikat pada dataset *Differentiated Thyroid Cancer Recurrence*. Setelah dilakukan analisis ada tahapan

preprocessing yang akan dilakukan yaitu pengecekan dan penanganan data duplikat, *missing value*, dan data *imbalance*.

3.2.1 Data Duplikat

Tahap ini dilakukan untuk memastikan pada dataset yang dipakai tidak ada duplikasi, hal ini bertujuan untuk mempermudah dalam pengolahan data tersebut. Pengecekan data duplikat dapat dilihat pada gambar 4.

```
duplikat =dfdata.duplicated()
print(f"\nJumlah data dengan duplikat : {len(dfdata)}")
print(f"\nJumlah baris duplikat: {duplikat.sum()}")
```

```
Jumlah data dengan duplikat : 383

Jumlah baris duplikat: 19
```

Gambar 4. Pengecekan data duplikat

Ditemukan hasil bahwa dataset memiliki 19 data duplikat dari total 383 data. Langkah berikutnya adalah penanganan data duplikat yaitu melakukan penanganan data duplikat tersebut yang dapat dilihat pada gambar 5.

```
data_bersih = dfdata.drop_duplicates()
print(f"\nJumlah data tanpa duplikat : {len(data_bersih)}")
```

```
Jumlah data tanpa duplikat : 364
```

Gambar 5. Penanganan data duplikat

Penanganan data duplikat merupakan penghapusan data berulang pada dataset, setelah dilakukan penanganan Dataset menjadi berjumlah 364.

3.2.2 Missing Value

Selanjutnya memastikan dataset tidak memiliki *Missing value*, Hasil dari pengecekan *missing value* dapat dilihat pada gambar 6.

```
Jumlah missing values per kolom:
 Age                     0
Gender                   0
Smoking                  0
Hx Smoking               0
Hx Radiothreapy          0
Thyroid Function         0
Physical Examination     0
Adenopathy               0
Pathology                0
Focality                 0
Risk                     0
T                        0
N                        0
M                        0
Stage                    0
Response                 0
Recurred                 0
dtype: int64
```

Gambar 6. Pengecekan *missing value*

3.2.3 *Imbalance* Dataset

Langah berikutnya kita akan melakukan *balancing* data dengan memastikan jumlah kelas pada dataset adalah sama. Setelah dilakukan pengecekan ternyata dataset tersebut merupakan data *imbalance.* Hasil dari pengecekan dataset dapat dilihat pada gambar 7.

```
Recurred
{'No'} {256}
Recurred
{'Yes'} {108}
```

Gambar 7. Hasil pengecekan dataset *imbalance*

Untuk metode *balancing* data sendiri akan menggunakan metode *SMOTEENN.* Metode *SMOTEENN* merupakan metode *hybrid sampling* dengan menggabungkan metode *Synthetic Minority Oversampling Technique* (SMOTE) untuk meningkatkan jumlah kelas minoritas dan *Edited Nearest Neighbors* (ENN) untuk mengurangi jumlah kelas mayoritas [21].

SMOTE adalah metode over-sampling dimana data pada kelas minoritas diperbanyak dengan menggunakan data sintetik yang berasal dari replikasi data pada kelas minoritas [22]. *Edited Nearest Neighbors* (ENN) adalah salah satu metode *under-sampling* dimana kelas mayoritas akan dikurangi dengan menghapus sampel dari kelas mayoritas yang memiliki label berbeda dari mayoritas tetapi memiliki tetangga-tetangga terdekat yang berbeda kelas [23]. Gambar 8 menunjukan bahwa kelas pada dataset sudah sama jumlahnya setelah dilakukan *balancing* data.

```
Recurred
{'No'} {206}
Recurred
{'Yes'} {206}
```

Gambar 8. Dataset *balance*

3.3. Transformation

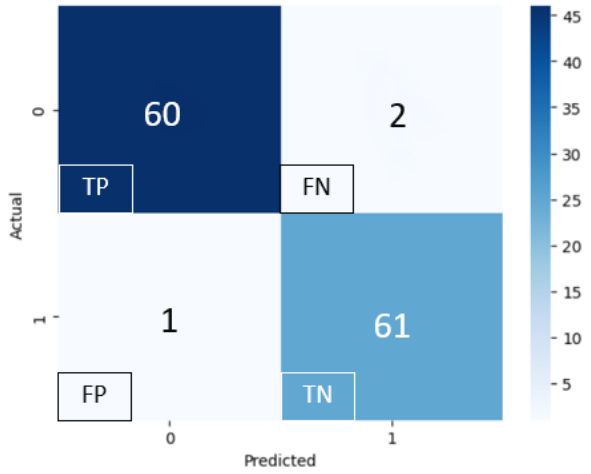
Label *encoding* adalah proses untuk mengkonversi data nominal menjadi data numerik. Alasan dilakukanya *encoding* karena penelitian ini menggunakan algoritma *Random Forest* yang hanya bisa diterapkan pada data berbentuk numerik. Tidak hanya itu pada proses *imbalance* Dataset menggunakan *SMOTEENN* juga hanya bisa diterapkan pada tipe data numerik. Proses *encoding* sendiri hanya dilakukan pada atribut independen dan tidak dilakukan pada kelas. Hasil *encoding* dapat dilihat pada gambar 9.

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Thyroid Function	Physical Examination	Adenopathy	Pathology	Focality	Risk	T	N	M	Stage	Response
0	27	0	0	0	0	2	3	3	2	1	2	0	0	0	0	2
1	34	0	0	1	0	2	1	3	2	1	2	0	0	0	0	1
2	30	0	0	0	0	2	4	3	2	1	2	0	0	0	0	1
3	62	0	0	0	0	2	4	3	2	1	2	0	0	0	0	1
4	62	0	0	0	0	2	1	3	2	0	2	0	0	0	0	1

Gambar 9. Hasil *encoding* Dataset

3.4. Data Mining

Dataset *Differentiated Thyroid Cancer Recurrence* masuk kedalam tipe klasifikasi karena atribut independent dalam dataset tersebut berupa numerik dan nominal serta memiliki label nominal. Selain itu data tersebut bertujuan untuk memprediksi atau mengklasifikasikan pasien ke dalam kategori tertentu berdasarkan atribut-atribut yang ada dalam dataset. Dalam model kali ini data *training* sebesar 70% dari total data. Algoritma yang akan digunakan dalam klasifikasi kali ini adalah *Random Forest*, tahap berikutnya adalah implementasi algoritma *Random Forest* ke dataset yang telah dilakukan proses *preprocessing*. *Confusion matrix* yang dihasilkan dapat dilihat pada gambar 10.



Gambar 10. *Confusion matrix*

Dimana:

"0" = Kelas *No*

"1" = Kelas *Yes*

True Positive (TP) = jumlah dari kelas *No* yang benar diklasifikasikan sebagai kelas *No*

False Negative (FN) = jumlah dari kelas *No* yang salah diklasifikasikan sebagai kelas *Yes*

False Positive (FP) = jumlah dari kelas *Yes* yang salah diklasifikasikan sebagai kelas *No*

True Negative (TN) = jumlah dari kelas *Yes* yang benar diklasifikasikan sebagai kelas *Yes*

3.5. Evaluation

Selanjutnya dilakukan pengukuran Sensitifitas, Spesifisitas, *Precision*, *Area Under the Curve* (AUC), *Accuracy* terhadap algoritma *Random Forest* yang telah diterapkan, Tabel 2 merupakan hasil pengukuran dari pengujian terhadap algpritma *Random Forest*.

Tabel 2. Perhitungan *confusion matrix*

Hasil Pengukuran	Skor
Sensitifitas	98,39%
Spesifisitas	96,77%
Precision	96,83%
AUC	97,6%
Accuracy	97,5%

Dalam mengevaluasi algoritma *Random Forest*, hasil pengukuran menunjukkan bahwa model tersebut memiliki kinerja yang sangat baik dengan Sensitifitas sebesar 98,39%, Spesifisitas sebesar 96,77%, *Precision* sebesar 96,83%, *Area Under the Curve* (AUC) sebesar 97,6%, dan Accuracy sebesar 97,5%. Berdasarkan analisis yang dilakukan, hasil perhitungan menegaskan bahwa algoritma *Random Forest* menunjukkan tingkat efektivitas yang sangat tinggi dalam memprediksi risiko kambuhnya penyakit kanker tiroid pada pasien, memberikan kepercayaan yang lebih kuat terhadap hasil prediksi dan memungkinkan penanganan yang lebih tepat terhadap kondisi kesehatan pasien.

IV. KESIMPULAN

Penelitian ini bertujuan untuk mengembangkan model klasifikasi risiko kambuhnya kanker tiroid menggunakan dataset *Differentiated Thyroid Cancer Recurrence* yang terdiri dari 383 pasien. Melalui berbagai tahapan preprocessing seperti penanganan data duplikat, pengecekan dan penanganan *missing value*, serta penyeimbangan dataset yang mengalami ketidak seimbangan kelas menggunakan metode *SMOTEENN*. Transformasi data dilakukan melalui label *encoding* agar dapat diterapkan pada algoritma *Random Forest*. Setelah pembagian data *training* sebesar 70%, model *Random Forest* diimplementasikan dan menunjukkan performa yang sangat baik dengan Sensitifitas sebesar 98,39%, Spesifisitas sebesar 96,77%, *Precision* sebesar 96,83%, *Area Under the Curve* (AUC) sebesar 97,6%, dan *Accuracy* sebesar 97,5%.

Hasil evaluasi ini menunjukkan bahwa algoritma *Random Forest* sangat efektif dalam memprediksi risiko kambuhnya kanker tiroid. Tingkat akurasi yang tinggi memberikan

kepercayaan lebih besar terhadap hasil prediksi model, sehingga dapat digunakan untuk mendukung keputusan klinis dan strategi tindak lanjut yang lebih tepat bagi pasien.

UCAPAN TERIMAKASIH

Puji dan syukur Kami panjatkan ke hadirat Tuhan Yang Maha Esa, karena atas berkat dan rahmat-Nya, Kami dapat menyelesaikan jurnal ini dengan baik. Kami ingin menyampaikan terima kasih kepada berbagai pihak yang telah memberikan dukungan, bimbingan, dan bantuan selama proses penulisan jurnal ini:

1. Kami ucapkan terima kasih kepada Universitas Nusantara PGRI Kediri yang telah menyediakan fasilitas yang memadai sehingga jurnal ini dapat terselesaikan dengan baik.
2. Terima kasih Kami sampaikan kepada Fakultas Teknik dan Ilmu Komputer atas segala dukungan akademis dan fasilitas yang telah diberikan.

DAFTAR PUSTAKA

[1] Y. Parura, V. Pontoh, and M. Werung, "Pola kanker tiroid periode Juli 2013 – Juni 2016 di RSUP Prof. Dr. R. D Kandou Manado," *e-CliniC*, vol. 4, no. 2, 2016, doi: 10.35790/ecl.4.2.2016.14475.

[2] A. Nur, A. Santosa, and A. S. Komariyah, "Karakteristik Kanker Tiroid Di Maluku Utara Tahun 2017-2020," *J. Endur.*, vol. 8, no. 2, pp. 246–252, May 2023, doi: 10.22216/JEN.V8I2.2161.

[3] S. Borzooei, G. Briganti, M. Golparian, J. R. Lechien, and A. Tarokhian, "Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study," *Eur. Arch. Oto-Rhino-Laryngology*, vol. 281, no. 4, pp. 2095–2104, 2024, doi: 10.1007/s00405-023-08299-w.

[4] B. Wijonarko, "Perbandingan Algoritma Data Mining Naïve Bayes Dan Bayes Network Untuk Mengidentifikasi Penyakit Tiroid," *J. Pilar Nusa Mandiri*, vol. 14, no. 1, pp. 21–26, Mar. 2018, doi: 10.33480/PILAR.V14I1.83.

[5] D. Galih Pradana, M. L. Alghifari, M. Farhan Juna, and S. Dwisiwi Palaguna, "Klasifikasi Penyakit Jantung Menggunakan Metode Artificial Neural Network," *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 55–60, Jul. 2022, doi: 10.56705/IJODAS.V3I2.35.

[6] S. Borzooei and A. Tarokhian, "Differentiated Thyroid Cancer Recurrence," *UCI Mach. Learn. Repos.*, 2023, doi: https://doi.org/10.24432/C5632J.

[7] E. Daniati and H. Utama, "Decision Making Framework Based on Sentiment Analysis in Twitter Using SAW and Machine Learning Approach," *2020 3rd Int. Conf. Inf. Commun.*

Technol. ICOIACT 2020, pp. 218–222, Nov. 2020, doi: 10.1109/ICOIACT50329.2020.9331998.

[8] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 4, no. 2, p. 437, 2020, doi: 10.30865/mib.v4i2.2080.

[9] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017, doi: 10.22202/ei.2016.v2i2.1465.

[10] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *Sistemasi*, vol. 10, no. 1, p. 163, 2021, doi: 10.32520/stmsi.v10i1.1129.

[11] A. Aqillah Fadia Haya, Reynaldi Azhar, Muhamad Khandava Mulyadien, and Betha Nurina Sari, "Klasifikasi Minat Beli Pelanggan Terhadap Udang Vaname Menggunakan Algoritma Naïve Bayes," *J. Ilm. Betrik*, vol. 13, no. 1, pp. 59–65, 2022, doi: 10.36050/betrik.v13i1.452.

[12] I. A. Nikmatun and I. Waspada, "Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 10, no. 2, pp. 421–432, Nov. 2019, doi: 10.24176/SIMET.V10I2.2882.

[13] L. B. Adzy, A. Asriyanik, and A. Pambudi, "Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima," vol. 6, no. 1, pp. 1–10, 2023.

[14] A. C. Pradikdo and A. Ristyawan, "Model Klasifikasi Abstrak Skripsi Menggunakan Text Mining Untuk Pengkategorian Skripsi Sesuai Bidang Kajian," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 9, no. 2, pp. 1091–1098, Nov. 2018, doi: 10.24176/SIMET.V9I2.2513.

[15] Khoirunnisa Hamidah and A. Voutama, "Analisis Faktor Tingkat Kebahagiaan Negara Menggunakan Data World Happiness Report dengan Metode Regresi Linier.," *Explor. IT J. Keilmuan dan Apl. Tek. Inform.*, vol. 15, no. 1, pp. 1–7, 2023, doi: 10.35891/explorit.v15i1.3874.

[16] A. Karimah, G. Dwilestari, and M. Mulyawan, "Analisis Sentimen Komentar Video Mobil Listrik Di Platform Youtube Dengan Metode Naive Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 767–737, 2024, doi: 10.36040/jati.v8i1.8373.

[17] D. Akbar Baturaja, D. Juardi, and A. Susilo Yuda Irawan, "Analisis Sentimen Masyarakat Terhadap Dugaan Kontroversi Pondok Pesantren Al-Zaytun Menggunakan Naïve Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 4, pp. 2775–2782, 2024, doi: 10.36040/jati.v7i4.7198.

[18] N. A. Prakoso Indaryono, "Analisa Perbandingan Algoritma Random Forest Dan Naïve Bayes Untuk Klasifikasi Curah Hujan Berdasarkan Iklim Di Indonesia," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 1, pp. 158–167, 2024, doi: 10.29100/jipi.v9i1.4421.

[19] O. Wira Yuda, D. Tuti, and L. Sheih Yee, "Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest," *SATIN - Sains dan Teknol. Inf.*, vol. 8, no. 2, pp. 122–131, Dec. 2022, doi: 10.33372/STN.V8I2.885.

[20] M. R. Adrian, M. P. Putra, M. H. Rafialdy, and N. A. Rakhmawati, "Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB," *J. Inform. Upgris*, vol. 7, no. 1, Jun. 2021, doi: 10.26877/JIU.V7I1.7099.

[21] Y. R. Saputra, S. Syafriandi, D. Permana, and Zilrahmi, "Classification of Program Keluarga Harapan Recipient Households in Padang City Using K-Nearest Neighbors," *UNP J. Stat. Data Sci.*, vol. 2, no. 2, pp. 187–195, May 2024, doi: 10.24036/UJSDS/VOL2-ISS2/167.

[22] E. Sutoyo and M. Asri Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 6, no. 3, pp. 379–385, Dec. 2020, doi: 10.26418/JP.V6I3.42896.

[23] A. Indrawati, "Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 1, pp. 38–43, Apr. 2021, doi: 10.33387/JIKO.V4I1.2561.