BINA INSANI ICT JOURNAL
Vol. 11, No. 1, Juni 2024, 56 - 65
ISSN: 2355-3421 (Print) ISSN: 2527-9777 (Online)

Analisis Komparasi Kinerja Metode *Decision Tree* dan *Random Forest* dalam Klasifikasi Teks Data Kesehatan

Hardian Oktavianto[1,*], Henny Wahyu Sulistyo[1], Guruh Wijaya[1], Dudi Irawan[1], Ginanjar Abdurrahman[1]

[1] Teknik Informatika; Universitas Muhammadiyah Jember; Jl. Karimata No. 49 Jember - Jawa Timur - Indonesia, Telp. (0331) 336728 Fax. (0331) 337957; e-mail: hardian@unmuhjember.ac.id, henny.sulistyo@unmuhjember.ac.id, guruh.wijaya@unmuhjember.ac.id, dudi.irawan@unmuhjember.ac.id, abdurrahmanginanjar@unmuhjember.ac.id

Korespondensi: hardian@unmuhjember.ac.id

Diterima: 11 Juni 2024 ; Review: 27 Juni 2024; Disetujui: 30 Juni 2024

Cara sitasi: Oktavianto H, Sulistyo HW, Wijaya G, Irawan D, Abdurrahman G. 2024. Analisis Komparasi Kinerja Metode *Decision Tree* dan *Random Forest* dalam Klasifikasi Teks Data Kesehatan. Bina Insani ICT Journal. Vol. 11 (1): 56 - 65

Abstrak: Salah satu topik penelitian yang diminati para peneliti adalah klasifikasi teks otomatis. Klasifikasi teks pada saat ini sering digunakan karena semakin banyaknya jumlah dokumen teks yang harus kita tangani maupun kita gunakan setiap hari. Metode pengklasifikasi yang banyak digunakan diantaranya adalah *decision tree* dan *random forest*. *Decision Tree* umumnya dipilih memiliki kesederhanaan visual, konstruksi keputusan yang relatif cepat, dan tidak memerlukan asumsi sebelumnya tentang data. Algoritma *Random Forest* menggunakan konsep yang sama dengan *Decision Tree* kemudian melakukan agregasi hasil dari setiap pohon keputusan untuk memperoleh hasil. Algoritma *Random Forest* juga menggunakan pendekatan perluasan, yang disebut pendekatan bagging, yang mana fitur-fitur berbeda dari kumpulan data akan ditugaskan ke setiap pohon keputusan. Tujuan penelitian ini adalah melakukan kategorisasi dengan menggunakan teknik klasifikasi menggunakan algoritma *decision tree* dan *random forest*, yang merupakan algoritma pengklasifikasi yang banyak digunakan. Dengan menggunakan empat buah skenario uji dimana membagi data latih dan data uji dengan kombinasi 10%-90% 15%-85% 20%-80% dan 25%-75% menghasikan informasi bahwa Random Forest memiliki akurasi yang lebih baik. Akurasi Random Forest pada tiap skenario uji selalu lebih baik daripada nilai akurasi Decision Tree. Pada Decision Tree, menunjukkan nilai akurasi cukup stabil pada kisaran 75%, sedangkan pada Random Forest menunjukkan nilai akurasi lebih stabil pada nilai 99%.

Kata kunci: decision tree, random forest, klasifikasi teks

Abstract: *One of the research topics that researchers are interested in is automatic text classification. Text classification is currently often used because of the increasing number of text documents that we have to handle and use every day. Classifier algorithms that are widely used include decision trees and random forests. Decision Trees are generally chosen for their visual simplicity, relatively fast decision construction, and do not require prior assumptions about the data. The Random Forest algorithm uses the same concept as Decision Tree and then aggregates the results of each decision tree to obtain results. The Random Forest algorithm also uses an expansion approach, called the bagging approach, in which different features of the data set are assigned to each decision tree. The aim of this research is to carry out categorization using classification techniques using decision tree and random forest algorithms, which are widely used classifier algorithms. By using four test scenarios where dividing the training data and test data with a combination of 10%-90% 15%-85% 20%-80% and 25%-75% produces information that Random Forest has better accuracy. Random Forest accuracy in each test scenario is always better than the Decision Tree accuracy value. Decision Tree shows that the accuracy value is*

quite stable at around 75%, while Random Forest shows that the accuracy value is more stable at 99%.

***Keywords**: decision tree, random forest, text classification*

1. Pendahuluan

Klasifikasi teks berguna dalam banyak aplikasi *natural language processing* (NLP) dan *text mining* seperti pengambilan atau ekstraksi informasi. Klasifikasi teks digunakan sebagai dasar pekerjaan dalam pemrosesan teks, penambangan web, dan penyaringan teks, contohnya adalah pada penyaring *spam email*, pengorganisasian teks sebagai arsip dokumen, dan sebagainya [1]. Klasifikasi teks juga dikenal sebagai kategorisasi teks, penentuan topik, kategorisasi dokumen, atau klasifikasi dokumen, dimana inti proses yang dilakukan adalah menetapkan dan memberi label pada dokumen pada sekumpulan kategori yang telah ditentukan berdasarkan isi atau konten [2].

Klasifikasi teks otomatis menjadi salah satu implementasi dan topik penelitian yang diminati para peneliti. Saat ini, klasifikasi teks sering digunakan karena banyaknya jumlah dokumen teks yang harus kita tangani setiap hari. Umumnya sebagian besar data untuk klasifikasi teks dikumpulkan dari media sosial, melalui newsgroup, papan buletin, dan berita siaran atau cetak [3]. Data tersedia dari berbagai sumber sehingga memiliki dampak seperti format yang berbeda, kosa kata ideal yang berbeda, dan gaya penulisan yang berbeda juga. Machine learning adalah studi komputasi mengenai algoritma dan metode statistik yang digunakan komputer untuk melakukan tugas tertentu. Algoritma machine learning merancang model statistik yang dikenal sebagai data training, untuk membuat prediksi atau keputusan secara eksplisit untuk melakukan tugas tersebut [4].

Algoritma pengklasifikasi yang banyak digunakan diantaranya adalah *decision tree* dan *random forest*. *Decision Tree* adalah salah satu teknik yang paling sering digunakan dalam penelitian data mining dan dalam sistem pembelajaran adaptif berbasis kecerdasan buatan. Setiap node dari pohon keputusan berisi pengujian pada suatu atribut; setiap cabang dari sebuah node berhubungan dengan kemungkinan hasil pengujian; setiap daun berisi prediksi kelas. *Decision Tree* umumnya dipilih memiliki kesederhanaan visual, konstruksi keputusan yang relatif cepat, dan tidak memerlukan asumsi sebelumnya tentang data [5]. Beberapa penelitian tentang *decision tree* diantaranya adalah, penelitian tentang klasifikasi data dengan memanfaatkan decision tree sebagai seleksi fitur menghasilkan perbandingan antara model *Decision Tree* dan Naive Bayes, untuk ketiga dataset Iris, Glass, dan Credit, kinerja *Decision tree* sedikit lebih baik dari segi akurasi dibandingkan Naive Bayes, baik ketika menerapkan seleksi fitur ataupun tidak [6]. Penelitian tentang prediksi keberhasilan studi mahasiswa menggunakan decision tree C-4.5 yang menghasilkan bahwa prediksi prestasi masa studi dengan menerapkan decision tree C-4.5 dapat menghasilkan pendekatan atau faktor-faktor yang mempengaruhi kelulusan siswa dengan persentase sebesar 85%. [7]. Penelitian tentang klasifikasi peserta beasiswa keluarga miskin menggunakan decision tree C-4.5, adapun kontribusi penelitian ini adalah melakukan klasifikasi peserta beasiswa keluarga miskin dengan algoritma C 4.5 tanpa metode prunning, pre prunning, dan post prunning, berdasarkan skenario yang telah dilakukan nilai akurasi yang dihasilkan setelah dilakukan prunning lebih baik dibandingkan ketika tanpa dilakukan prunning [8]. Seleksi merchant otomatis menggunakan decision tree C-4.5 Hasil perbandingan menunjukkan bahwa algoritma C 4.5 merupakan model terbaik untuk menangani kasus kelayakan Merchant dalam Program Sponsor. Hal ini dapat dibuktikan dengan melihat tingkat akurasi yang dihasilkan pada proses pengujian dan validasi model. Kedua model memiliki nilai AUC yang sama namun algoritma C-4.5 menghasilkan nilai akurasi yang unggul dengan selisih 0,45% dibandingkan Naive Bayes [9].

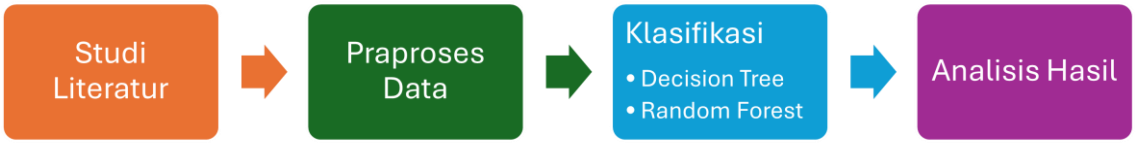
Algoritma *Random Forest* adalah metode klasifikasi yang melakukan proses serupa dengan Decision Tree. Konsep *Random Forest* pertama kali dikembangkan oleh Breiman dan dapat disebut sebagai *ensemble*. Konsep ini menggunakan konsep yang sama dengan Decision Tree kemudian melakukan agregasi hasil dari setiap pohon keputusan untuk memperoleh hasil klasifikasi. Algoritma Random Forest juga menggunakan pendekatan perluasan, yang disebut pendekatan bagging, yaitu fitur – fitur berbeda dari kumpulan data akan ditugaskan ke setiap pohon keputusan. Pengacakan fitur ini juga bisa disebut sebagai konsep "*bootstrapping*". Dengan demikian, hasil terbaik bisa didapat. Selain itu, Random Forest dapat mengurangi permasalahan "*overfitting*" berdasarkan kriteria pemilihan fitur yang dapat menciptakan saling ketergantungan.

Oleh karena itu, *Random Forest* dapat menghasilkan hasil klasifikasi yang baik [10]. Klasifikasi malware pada data *imbalance* menggunakan decision tree dan random forest, Random Forest mendapatkan akurasi paling besar dengan nilai 99,99% disusul oleh Decision Tree dengan akurasi 99,98%, KNN dengan akurasi 99,94%, dan Naive Bayes dengan akurasi 99,00% [11]. Prediksi kanker payudara menggunakan decision tree, random forest, dan *linear discriminant analysis*, Hasilnya menegaskan bahwa model Random Forest mampu mencapai akurasi tertinggi hingga 99,4% dan kemampuan generalisasi terbaik. Namun model Analisis Diskriminan Linier dapat menjaga stabilitas akurasi prediksi lebih baik dan memiliki running time tercepat [12].

Tujuan penelitian ini adalah melakukan kategorisasi dengan menerapkan teknik klasifikasi menggunakan algoritma decision tree dan random forest, kemudian melakukan analisis terhadap hasil klasifikasi yang terbentuk. Urgensi dari penelitian ini adalah membantu pihak pihak yang terkait permasalahan kanker dengan menyajikan hasil klasifikasi dari teks data kesehatan khususnya data kanker. Inovasi dari penelitian ini adalah menghasilkan kelompok data kanker dari sumber data teks kesehatan penyakit kanker dengan pendekatan text mining.

2. Metode Penelitian

Tahapan atau langkah – langkah penelitian yang akan dilakukan secara umum terdiri dari 4 buah tahapan, mulai dari studi literatur, kemudian pengumpulan dan pemrosesan dataset, dilanjutkan dengan implementasi klasifikasi, dan yang terakhir adalah penarikan kesimpulan atau analisis hasil. Tahapan atau langkah – langkah penelitian ini secara umum dapat dilihat pada gambar 1.



Sumber: Hasil Penelitian (2024)

Gambar 1. Tahapan Penelitian

Studi literatur merupakan langkah yang dilakukan untuk mempelajari referensi berupa jurnal penelitian, paper, buku-buku referensi yang lain terkait dengan penelitian untuk melengkapi pengetahuan awal, guna memahami teori yang dapat digunakan untuk menunjang penelitian.

Dataset yang digunakan adalah dokumen teks biomedis yang diambil dari abstrak artikel ilmiah yang mempunyai halaman lebih dari 6 halaman. Kumpulan data mencakup dokumen kanker untuk diklasifikasikan ke dalam 3 kategori seperti 'Kanker_Tiroid','Kanker_Usus Besar','Kanker_Paru-Paru'. Total publikasi sejumlah 7570. Data ini memiliki 3 label kelas dalam kumpulan data, adapun jumlah sampel di setiap kategori sebagai berikut, kanker usus besar 2580, kanker paru – paru 2180, kanker tiroid 2810.

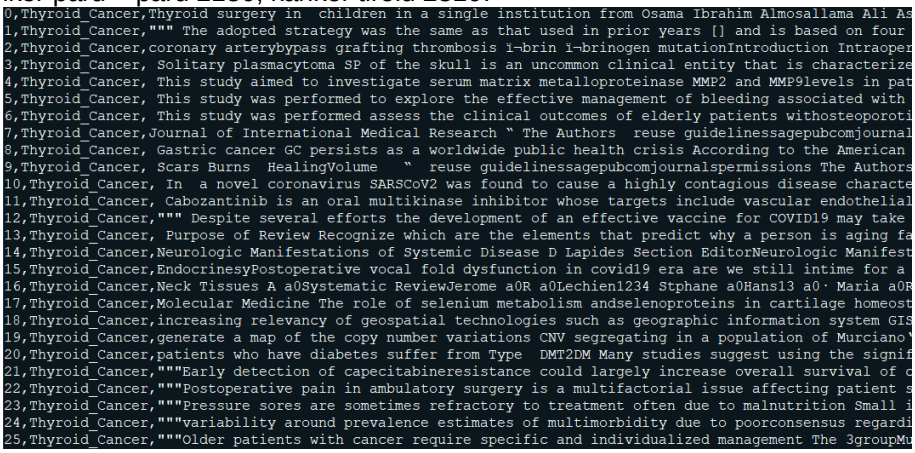
Klasifikasi pada data teks akan menggunakan algoritma Decision Tree dan Random Forest. Data masukan berupa data teks akan diproses Tokenizing, kemudian dilanjutkan dengan proses Stemming, setelah itu dilakukan proses Delete Stop Words, proses selanjutnya adalah membentuk vektor representasi dari masing – masing kata, kemudian selanjutnya adalah ekstraksi fitur, dan yang terakhir adalah proses klasifikasi menggunakan algoritma Decision Tree dan Random Forest.

Analisis hasil yaitu melakukan pendeskrispsian terhadap hasil pengelompokan yang terbentuk. Analisis hasil akan menjelaskan masing – masing keluaran dari Decision Tree dan Random Forest secara berurutan. Pada tahap ini juga dilakukan penghitungan akurasi untuk mengetahui seberapa baik performa atau kinerja dari metode yang diterapkan.

2.1. Dataset

Dataset yang dipakai pada penelitian ini adalah data teks abstrak penelitian di bidang medisk atau kesehatan, yang secara khusus membahas tentang penyakit kanker. Dataset bersifat bebas dipakai untuk digunakan dan diperoleh secara bebas dari situs kaggle https://www.kaggle.com/datasets/falgunipatel19/biomedical-text-publication-classification.
Dataset ini terdiri dari 7570 data dan memiliki 3 label kelas 'Thyroid_Cancer','Colon_Cancer',dan

'Lung_Cancer', adapun jumlah sampel di setiap kategori sebagai berikut, kanker usus besar 2580, kanker paru – paru 2180, kanker tiroid 2810.



```
0,Thyroid_Cancer,Thyroid surgery in  children in a single institution from Osama Ibrahim Almosallama Ali As
1,Thyroid_Cancer,""" The adopted strategy was the same as that used in prior years [] and is based on four
2,Thyroid_Cancer,coronary arterybypass grafting thrombosis ï¬brin ï¬brinogen mutationIntroduction Intraoper
3,Thyroid_Cancer, Solitary plasmacytoma SP of the skull is an uncommon clinical entity that is characterize
4,Thyroid_Cancer, This study aimed to investigate serum matrix metalloproteinase MMP2 and MMP9levels in pat
5,Thyroid_Cancer, This study was performed to explore the effective management of bleeding associated with
6,Thyroid_Cancer, This study was performed assess the clinical outcomes of elderly patients withosteoporoti
7,Thyroid_Cancer,Journal of International Medical Research " The Authors  reuse guidelinessagepubcomjournal
8,Thyroid_Cancer, Gastric cancer GC persists as a worldwide public health crisis According to the American
9,Thyroid_Cancer, Scars Burns  HealingVolume  "  reuse guidelinessagepubcomjournalspermissions The Authors
10,Thyroid_Cancer, In  a novel coronavirus SARSCoV2 was found to cause a highly contagious disease characte
11,Thyroid_Cancer, Cabozantinib is an oral multikinase inhibitor whose targets include vascular endothelial
12,Thyroid_Cancer,""" Despite several efforts the development of an effective vaccine for COVID19 may take
13,Thyroid_Cancer, Purpose of Review Recognize which are the elements that predict why a person is aging fa
14,Thyroid_Cancer,Neurologic Manifestations of Systemic Disease D Lapides Section EditorNeurologic Manifest
15,Thyroid_Cancer,EndocrinesyPostoperative vocal fold dysfunction in covid19 era are we still intime for a
16,Thyroid_Cancer,Neck Tissues A a0Systematic ReviewJerome a0R a0Lechien1234 Stphane a0Hans13 a0 · Maria a0R
17,Thyroid_Cancer,Molecular Medicine The role of selenium metabolism andselenoproteins in cartilage homeost
18,Thyroid_Cancer,increasing relevancy of geospatial technologies such as geographic information system GIS
19,Thyroid_Cancer,generate a map of the copy number variations CNV segregating in a population of Murciano
20,Thyroid_Cancer,patients who have diabetes suffer from Type  DMT2DM Many studies suggest using the signif
21,Thyroid_Cancer,"""Early detection of capecitabineresistance could largely increase overall survival of c
22,Thyroid_Cancer,"""Postoperative pain in ambulatory surgery is a multifactorial issue affecting patient s
23,Thyroid_Cancer,"""Pressure sores are sometimes refractory to treatment often due to malnutrition Small i
24,Thyroid_Cancer,"""variability around prevalence estimates of multimorbidity due to poorconsensus regardi
25,Thyroid_Cancer,"""Older patients with cancer require specific and individualized management The 3groupMu
```

Sumber: Hasil Penelitian (2024)

Gambar 2. Potongan Dataset

2.2. Decision Tree

Proses rekursi digunakan untuk membuat struktur Decision Tree berdasarkan fitur yang ada pada dataset berdasarkan nilai *information gain* mana yang paling tinggi. *Information Gain* berfungsi sebagai kriteria penentu untuk memilih atribut klasifikasi. Decision Tree menggunakan tiga perhitungan utama yaitu:

1) Nilai entropi dataset;
2) Rata-rata nilai entropi dari atribut;
3) Nilai information gain dari setiap atribut.

Pertama, entropi seluruh dataset dihitung sebagai ukuran ketidakpastian data. Hal ini dicapai dengan mendefinisikan himpunan pelatihan matriks data sebagai S, di mana S berisi m label kelas dan Si adalah subset skenario dalam himpunan pelatihan S. Kemudian entropi S dihitung dengan rumus (1)

$$Entropy(S) = -\sum_{i=1}^{m} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \dots (1)$$

Kedua, Data S dipartisi menggunakan atribut A, dimana A memiliki *k* hasil yang berbeda. Partisi ini akan menghasilkan subset S_j dengan nilai j = 1 sampai k. Entropi informasi rata-rata untuk semua atribut (A_1 ... A_n) di dalam S_j dihitung dengan rumus (2)

$$Entropy(A) = \sum_{j=1}^{k} \frac{|S_j|}{|S|} Entropy(S_j) \dots (2)$$

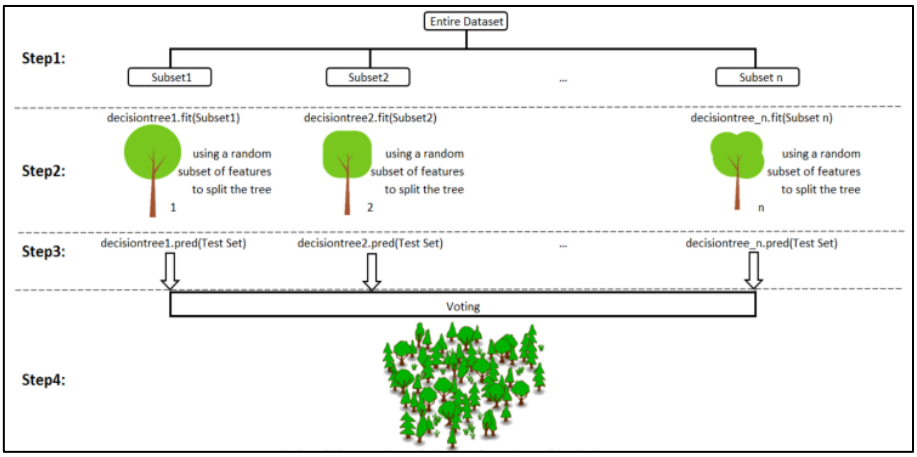
Terakhir, information gain, yaitu selisih entropi sebelum dan sesudah pemisahan dataset pada atribut A, dihitung untuk setiap atribut dalam matriks data dengan rumus (3)

$$Gain(A) = Entropy(S) - Entropy(A) \dots (3)$$

Atribut dengan information gain tertinggi dipilih sebagai simpul akar (root), yaitu titik yang memulai partisi data. Node akar mewakili atribut yang meminimalkan informasi yang dibutuhkan dan mengurangi keacakan partisi. Proses ini dilakukan berulang dengan membagi subset data pada setiap node internal hingga tidak ada atribut yang tersisa untuk klasifikasi, atau dataset kosong, atau data dalam setiap grup termasuk dalam kelas yang sama dan tidak diperlukan klasifikasi lebih lanjut. Sebuah pohon lengkap mempunyai cabang hingga simpul daun (leaf), yang mewakili label kelas.

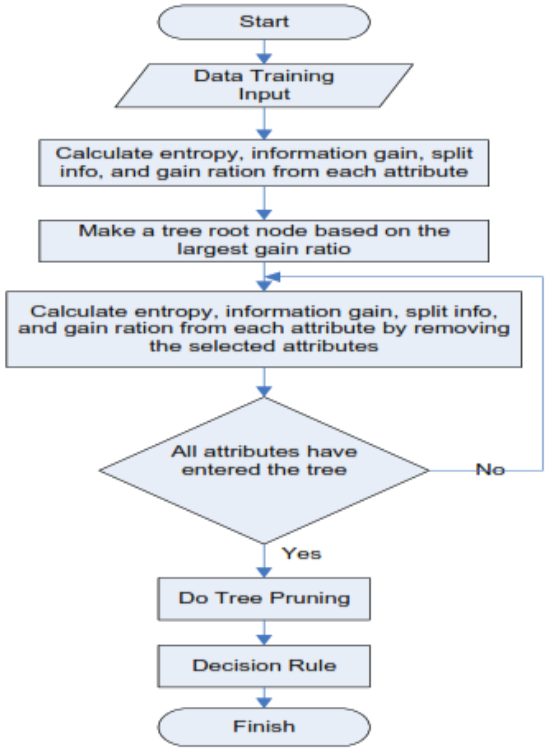
2.3. Random Forest

Random forest adalah metode pembelajaran *ensemble* berbasis pohon dengan setiap pohon bergantung pada beberapa variabel yang dipilih secara acak. Selain itu, *random forest* dianggap sebagai perpanjangan dari *bagging* dan dianggap sebagai kompetitor boosting. Gambar 3 menggambarkan prinsip Random Forest:



Sumber: Hasil Penelitian (2024)

Gambar 3. Prosedur Decision Tree



Sumber: Hasil Penelitian (2024)

Gambar 4. Konsep Random Forest

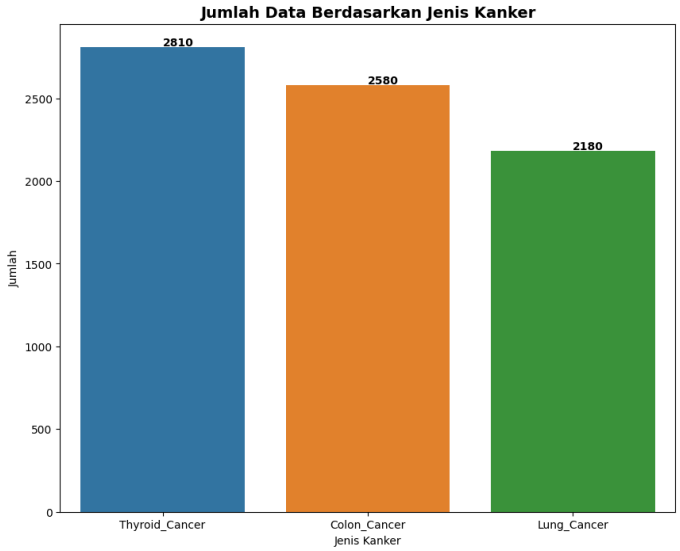
Tata cara pemodelan Random Forest yang diilustrasikan dari gambar 4 adalah sebagai berikut:

1. Pilih fitur m dari fitur M yang bersifat acak. Dengan jumlah m tidak lebih dari M.
2. Hitung titik pemisahan terbaik untuk pohon k berdasarkan metrik pemisahan (Gini impurity, dll.) dan pisahkan node saat ini menjadi node anak dan kurangi jumlah fitur M dari node ini.
3. Ulangi langkah 1 dan 2 hingga kedalaman pohon maksimum l tercapai atau matriks pemisahan mencapai titik ekstrim.

4. Ulangi langkah 1 sampai 3 untuk setiap pohon.
5. Pilih hasil setiap pohon.

3. Hasil dan Pembahasan

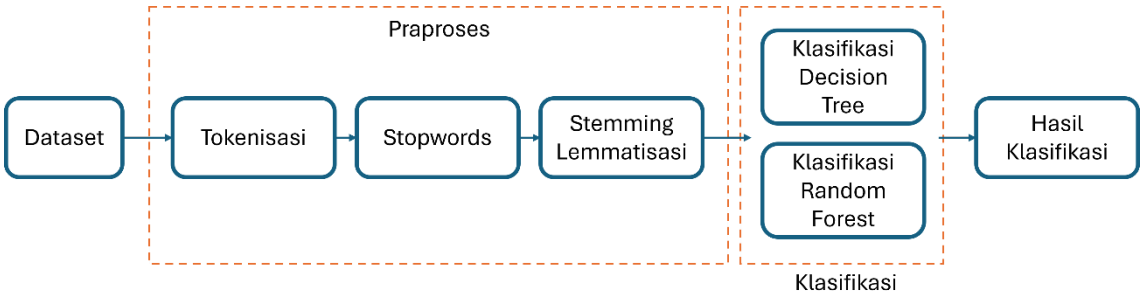
Dataset pada penelitian ini adalah dokumen abstrak dari artikel di bidang kesehatan khususnya bidang kanker, yang meliputi kanker tiroid, kanker usus besar, dan kanker paru – paru. Jumlah total dataset adalah 7570 buah, kanker tiroid 2810 data, kanker usus besar 2580 data, kanker paru – paru 2180 data. Gambar 5 menunjukkan perbandingan jumlah data teks untuk masing – masing kelas dataset. Data teks yang dipakai memiliki format dalam bentuk csv, dimana format csv ini mirip dengan data bentuk tabel biasa, hanya saja pada format csv, kolom dipisahkan oleh tanda baca koma (,). Dataset yang dipakai ini apabila diperhatikan dengan seksama memiliki struktur nomor baris, label kelas, dan abstrak



Sumber : hasil penelitian (2024)

Gambar 5. Perbandingan Jumlah Dataset

Praproses data dilakukan dengan memanfaatkan *library* nltk. *Library* nltk tersebut memiliki 'wordnet' yang digunakan untuk tokenisasi, 'stopwords' dan 'punkt' untuk melakukan identifikasi karakter yang akan dibuang atau dihapus. Pada tahap praproses ini juga dilakukan penghapusan spasi dan karakter kosong lainnya. Praproses teks adalah langkah penting dalam melakukan analisis sentimen, karena membantu membersihkan dan menormalkan data teks, sehingga lebih mudah untuk dianalisis. Langkah praproses umumnya melakukan tokenisasi, menghapus stopwords, dan stemming dan lemmatization, dalam membantu mengubah data teks mentah menjadi bentuk yang dapat digunakan untuk analisis.



Sumber: Hasil Penelitian (2024)

Gambar 6. Tahapan Klasifikasi

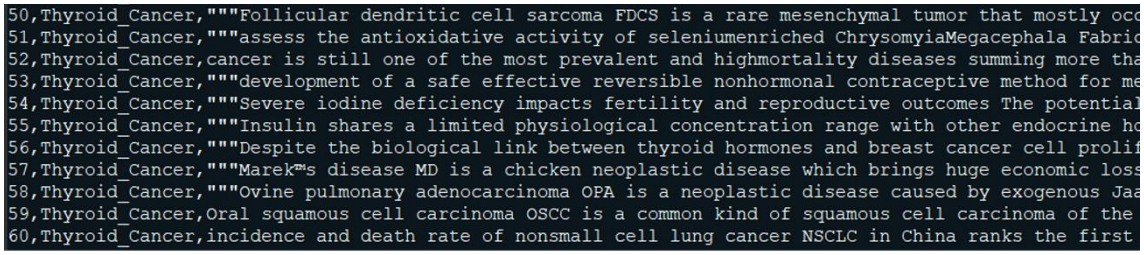
Tokenisasi adalah langkah pemrosesan awal teks dalam analisis sentimen yang melibatkan penguraian teks menjadi kata atau token individual. Ini adalah langkah penting dalam

menganalisis data teks karena membantu memisahkan setiap kata dari teks mentah, sehingga lebih mudah untuk dianalisis dan dipahami.

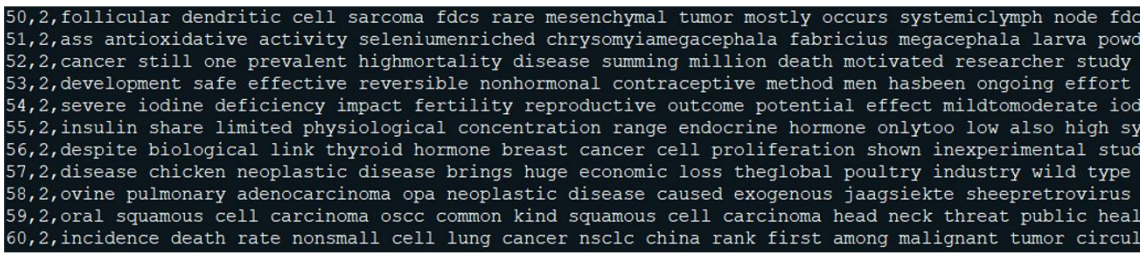
Penghapusan Stopwords adalah langkah praproses teks dalam analisis sentimen yang melibatkan penghapusan kata-kata umum dan tidak relevan yang kemungkinan besar tidak menyampaikan banyak sentimen. Kata-kata berhenti adalah kata-kata yang sangat umum dalam suatu bahasa dan tidak memiliki banyak arti, seperti "and", "which", "from", dan "that". Kata-kata ini dapat menimbulkan gangguan dan mengganggu analisis jika tidak dihilangkan. Dengan menghilangkan Stopwords, kata-kata yang tersisa dalam teks lebih cenderung menunjukkan sentimen yang diungkapkan. Hal ini dapat membantu meningkatkan akurasi analisis sentimen. NLTK menyediakan daftar Stopwords untuk beberapa bahasa, yang dapat digunakan untuk menyaring kata-kata ini dari data teks.

Stemming dan lemmatization adalah teknik yang digunakan untuk mereduksi kata menjadi bentuk akarnya. Stemming melibatkan penghapusan sufiks dari kata-kata, seperti "ing" atau "ed", untuk mereduksinya ke bentuk dasarnya. Misalnya, kata "melompat" akan dibentuk menjadi "melompat". Lemmatisasi, bagaimanapun, melibatkan pengurangan kata-kata ke bentuk dasarnya berdasarkan bagian ucapannya. Misalnya, kata "jumped" akan diberi lemmatisasi menjadi "jump", namun kata "jumping" akan diberi lemmatisasi menjadi "jumping" karena merupakan sebuah *present participle*.

Gambar 7 menunjukkan perbedaan kondisi dataset sebelum praproses dengan sesudah praproses. Label kelas yang semula “Colon_Cancer”, “Lung_Cancer”, dan “Thyroid_Cancer” diganti menjadi “0”, “1”, dan “2”. Huruf – huruf pada setiap kalimat dirubah menjadi huruf kecil, dan kemudian dilakukan penghapusan kata – kata yang tidak bermakna atau tidak berarti menggunakan acuan kamus data dari *library* nltk. Pada tahap tokenisasi, dilakukan penghapusan angka, tanda baca, dan karakter yang tidak mempunyai arti, dengan memanfaatkan *library* punkt dari nltk, hasil ditunjukkan pada gambar 7 yaitu hilangnya tanda baca ”””” dan juga karakter lainnya. Hasil praproses lainnya yaitu dari proses stemming dan lemmatization, yaitu dari setiap kata yang bukan kata dasar akan dikembalikan ke bentuk kata dasar, misalnya children menjadi child, studies menjadi study, risks menjadi risk, dan sebagainya. Gambar 8 adalah daftar kata – kata yang termasuk stopwords



(a)



(b)

Sumber: Hasil Penelitian (2024)

Gambar 7. (a) Sebelum Praproses (b) Setelah Praproses

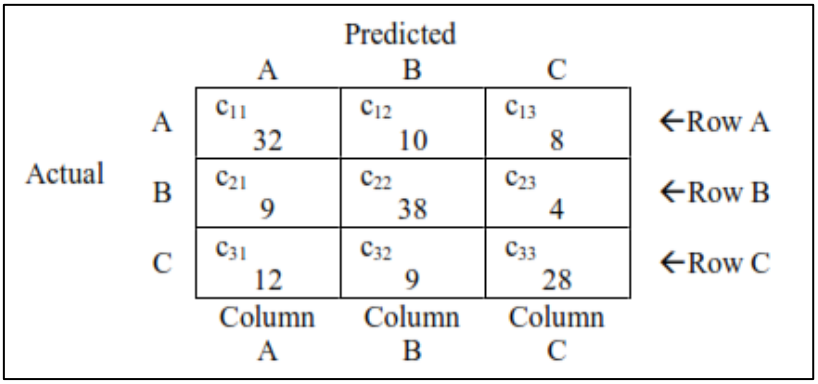
i	his	which	have	because	before	further	most	t
me	himself	who	has	as	after	then	other	can
my	she	whom	had	until	above	once	some	will
myself	her	this	having	while	below	here	such	just
we	hers	that	do	of	to	there	no	don
our	herself	these	does	at	from	when	nor	should
ours	it	those	did	by	up	where	not	now
ourselves	its	am	doing	for	down	why	only	
you	itself	is	a	with	in	how	own	
your	they	are	an	about	out	all	same	
yours	them	was	the	against	on	any	so	
yourself	their	were	and	between	off	both	than	
yourselves	theirs	be	but	into	over	each	too	
he	themselves	been	if	through	under	few	very	
him	what	being	or	during	again	more	s	

Sumber: Hasil Penelitian (2024)

Gambar 8. Daftar Stopwords

Analisis hasil klasifikasi dilakukan dengan membandingkan klasifikasi aktual dengan klasifikasi yang diprediksi, umumnya akan ada empat hasil berbeda [13]:

- Klasifikasi sebenarnya adalah positif, begitu pula klasifikasi yang diprediksi. Hal ini dikenal sebagai 'True Positive', disingkat TP, karena pengklasifikasi mengidentifikasi sampel positif dengan benar.
- Klasifikasi aktualnya negatif, dan klasifikasi prediksinya negatif. Ini adalah hasil "True Negative" (TN) karena pengklasifikasi mengidentifikasi sampel negatif dengan benar.
- Klasifikasi prediksinya positif, sedangkan klasifikasi sebenarnya negatif. Ini adalah hasil 'False Positive' (FP) karena pengklasifikasi salah mengidentifikasi sampel negatif sebagai positif.
- Klasifikasi prediksinya negatif, sedangkan klasifikasi sebenarnya positif. Ini adalah hasil 'False Negative' (FN) karena pengklasifikasi salah mengidentifikasi sampel positif sebagai sampel negatif.



Sumber: Hasil Penelitian (2024)

Gambar 9. Confusion Matrix

Pada tahap klasifikasi, dua buah algoritma diterapkan menggunakan bahasa pemrograman python dan dengan memanfaatkan *library* sklearn. Sedangkan untuk pembagian data latih dan data uji, skenario uji menggunakan kombinasi jumlah data latih dan data uji yang berbeda, mulai dari 10% 15% 20% dan 25% dari total dataset digunakan sebagai data uji, dan sisanya sejumlah 90% 85% 80% dan 75% persen digunakan sebagai data latih. Partisi data latih dengan data uji dilakukan di dalam *source code* python yang digunakan pada proses klasifikasi.

Tabel 1. Hasil Akurasi Uji Coba

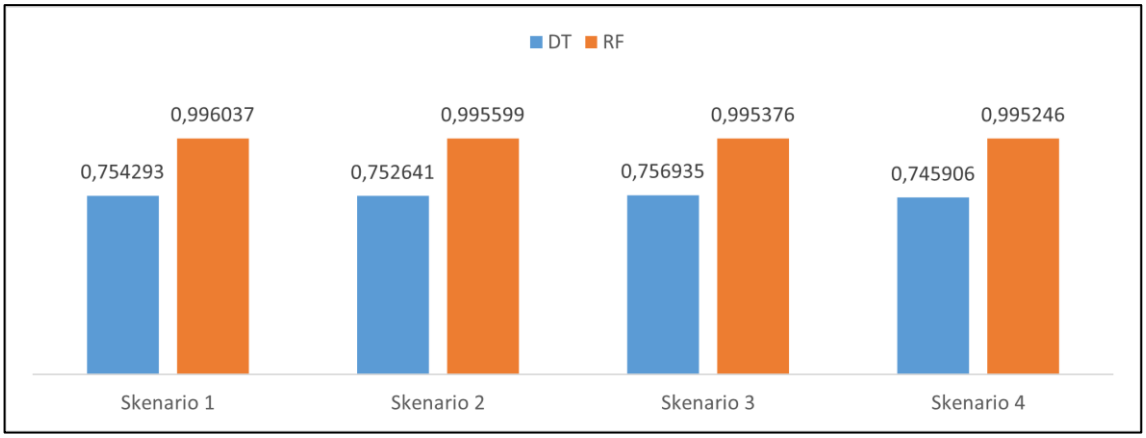
Model	Skenario 1	Skenario 2	Skenario 3	Skenario 4
DT	0,754293	0,752641	0,756935	0,745906
RF	0,996037	0,995599	0,995376	0,995246

Sumber: Hasil Penelitian (2024)

Tabel 2. Rata – Rata Akurasi

Model	Rata - rata Akurasi
Decision Tree	0,7524438
Random Forest	0,9955645

Sumber: Hasil Penelitian (2024)



Sumber: Hasil Penelitian (2024)

Gambar 10. Perbandingan Akurasi DT dan RF

Hasil akurasi ditunjukkan pada gambar 10, yang menjelaskan hasil uji dari Decision Tree secara berturut turut sesuai skenario adalah 75.4% 75.2% 75.6% dan 74.6 %, sedangkan hasil akurasi dari Random Forest secara berturut turut sesuai skenario adalah 99.6% 99.5% 99.5% dan 99.2%. Hal ini menunjukkan bahwa algoritma Random Forest bekerja dengan lebih baik secara akurasi untuk pengelompokkan pada dataset kesehatan. Pada Decision Tree, menunjukkan nilai akurasi cukup stabil pada kisaran 75 persen, sedangkan pada Random Forest menunjukkan nilai akurasi lebih stabil pada nilai 99%.

4. Kesimpulan

Decision Tree dan Random Forest telah diterapkan dalam klasifikasi dokumen kesehatan. Dokumen kesehatan yang menjadi dataset adalah abstrak dari artikel di bidang kesehatan khususnya bidang kanker, yang meliputi kanker tiroid, kanker usus besar, dan kanker paru – paru. Jumlah total dataset adalah 7570 buah, kanker tiroid 2810 data, kanker usus besar 2580 data, kanker paru – paru 2180 data. Dengan menggunakan pengukuran akurasi berdasarkan confusion matrix, dan empat buah skenario uji dimana membagi data latih dan data uji dengan kombinasi 10%-90% 15%-85% 20%-80% dan 25%-75% menghasikan informasi bahwa Random Forest memiliki akurasi yang lebih baik. Akurasi Random Forest pada tiap skenario uji selalu lebih baik daripada nilai akurasi Decision Tree. Pada Decision Tree, menunjukkan nilai akurasi cukup stabil pada kisaran 75%, sedangkan pada Random Forest menunjukkan nilai akurasi lebih stabil pada nilai 99%.

Referensi

[1] C.-Y. Ho, M. Syamsudin e Y.-C. Shen, “Cancer Literature Classification Methods Performance,” em *2020 International Conference on Decision Aid Sciences and Application*, 2020.

[2] Z. A. Abutiheen, A. H. Aliwy e K. B. S. Aljanabi, “Arabic text classification using master-slaves technique,” em *J. Phys.: Conf. Ser. 1032 012052*, 2018.

[3] Z. Hamid e H. K. Khafaji, “Classification of Arabic Documents depending on Maximal Frequent Itemsets,” em *J. Phys.: Conf. Ser. 1804 012009*, 2021.

[4] W. Dai, “Classification and analysis of literary works based on distribution weighted term frequency-inverse document frequency,” em *J. Phys.: Conf. Ser. 1941 012018*, 2021.

[5] J. Smith, M. Musharraf, B. Veitch e F. Khan, “Pilot Study Using Decision Trees to Diagnose the Efficacy of Virtual Offshore Egress Training,” *IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 15, NO. 6,* 2022.

[6] S. Rahmadani, A. Dongoran, M. Zarlis e Zakarias, “Comparison of Naive Bayes and Decision Tree on Feature Selection Using Genetic Algorithm for Classification Problem,” em *J. Phys.: Conf. Ser. 978 012087*, 2018.

[7] A. S. Fitrani, “Prediction of Study Period Students (Bachelor Degree) Muhammadiyah University of Sidoarjo Based on Decision Tree Method using C4.5 Algorithm,” em *J. Phys.: Conf. Ser. 1179 012033*, 2019.

[8] Y. Kustiyahningsih, B. K. Khotimah, D. R. Anamisa, M. Yusuf, T. Rahayu e J. Purnama, “Decision Tree C 4.5 Algorithm for Classification of Poor Family Scholarship Recipients,” em *IOP Conf. Ser.: Mater. Sci. Eng. 1125 012048*, 2021.

[9] Normah, I. Yulianti, D. Novianti, M. N. Winnarto, A. Zumarniansyah e S. Linawati, “Comparison of Classification C4.5 Algorithms and Naive Bayes Classifier in Determining Merchant Acceptance on Sponsorship Program,” em *J. Phys.: Conf. Ser. 1641 012006*, 2020.

[10] K. Saengtabtim, N. Leelawat, J. Tang, W. Treeranurat, N. Wisittiwong, A. Suppasri, K. Pakoksung, F. Imamura, N. Takahashi e I. Charvet, “Predictive Analysis of the Building Damage from the 2011 Great East Japan Tsunami Using Decision Tree Classification Related Algorithms,” *IEEE Access,* 2021.

[11] Irfan, I. M. Wildani e I. N. Yulita, “Classifying Botnet Attack on Internet of Things Device Using Random Forest,” em *IOP Conf. Ser.: Earth Environ. Sci. 248 012002*, 2019.

[12] R. Wang, “Comparison of Decision Tree, Random Forest and Linear Discriminant Analysis Models in Breast Cancer Prediction,” em *J. Phys.: Conf. Ser. 2386 012043*, 2022.

[13] M. F. Amin, “Confusion Matrix in Three-class Classification Problems: A Step-by-StepTutorial,” *Journal of Engineering Research,* vol. 7, nº 1, 2023.