Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 7, No. 7, Juli 2023, hlm. 3448-3456 http://j-ptiik.ub.ac.id

Prediksi Kanker Paru-Paru menggunakan Algoritme *Random Forest Decision Tree*

Rafly Dwi Marzuq[1], Satrio Agung Wicaksono[2], Nanang Yudi Setiawan[3]

Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]rdmarzuq@student.ub.ac.id, [2]satrio@ub.ac.id, [3]nanang@ub.ac.id

Abstrak

Teknologi telah berevolusi dari awal penciptaannya sampai sekarang dengan laju yang cepat. Salah satu aspek dari teknologi tersebut merupakan berkembanganya pertukaran data yang terjadi. Dengan skala pertukaran data yang besar tentu juga jumlah data yang berputar semakin besar. Untuk menggunakan dan menggali data tersebut menjadi informasi yang bisa digunakan terciptanya konsep Data Mining. Data Mining merupakan teknik untuk menemukan pola dan informasi dari data berjumlah besar. Data Mining bisa diimplementasikan pada banyak industri, salah satunya merupakan industri kesehatan. Penggunaan Data Mining untuk membantu riset dan penanganan kanker sedang sangat meningkat. Dengan munculnya penyakit kanker paru-paru di tubuh manusia terdapat beberapa gejala yang biasa dirasakan oleh kebanyakan pasien kanker paru-paru. Namun, gejala ini seringkali tidak dihiraukan dan tidak dicek oleh praktisi medis sehingga hanya 14% dari pasien yang didiagnosa kanker paru-paru sembuh dari penyakitnya lima tahun dari diagnosa. Menggunakan klasifikasi, terdapat cara untuk menganalisis data hasil gejala awal kanker paru-paru untuk menentukan *class label* dari hasil tersebut dengan tujuan membantu fasilitas kesehatan membuat keputusan medis terhadap calon pasien. Pada penelitian ini dilakukan implementasi klasifikasi menggunakan *Random Forest*, serta pengujian menggunakan *Confusion Matrix* dan *f-Fold Cross Validation*. Hasil dari pengujian menggunakan *Confusion Matrix* adalah ditemukan akurasi tertinggi sebesar 0,904 dan rata-rata akurasi sebesar 0,813. Hasil pengujian menggunakan *K-fold Cross Validation* adalah rata-rata akurasi tertinggi saat menggunakan 5-*fold cross validation* yaitu akurasi sebesar 0,889.

Kata kunci: *Klasifikasi, Kanker, Data Mining, Random Forest Decision Trees, Confusion Matrix, K-fold Cross Validaton*

Abstract

Technology has evolved rapidly since its inception. One aspect of this evolution is the development of data exchange. With the scale of data exchange increasing, the amount of data being circulated has also grown significantly. To utilize and extract useful information from this data, the concept of Data Mining has emerged. Data Mining is a technique used to discover patterns and information from large datasets. It can be implemented in various industries, including the healthcare sector. The use of Data Mining to aid cancer research and treatment is rapidly increasing. When it comes to lung cancer, there are common symptoms experienced by most patients. However, these symptoms are often overlooked and not examined by medical practitioners, resulting in only 14% of lung cancer patients being cured within five years of diagnosis. By employing classification, it is possible to analyze data on early lung cancer symptoms and determine the class labels associated with these results. This aims to assist healthcare facilities in making medical decisions regarding potential patients. In this study, the implementation of classification using Random Forest was conducted, along with testing using Confusion Matrix and f-Fold Cross Validation. The results obtained from the Confusion Matrix testing showed the highest accuracy of 0,904 and an average accuracy of 0,813. The results obtained from K-fold Cross Validation showed that the highest average accuracy was achieved when using 5-fold cross-validation, with an accuracy of 0,889.

Keywords: *Classification, Cancer, Data Mining, Random Forest Decision Trees, Confusion Matrix, K-fold Cross Validaton*

1. PENDAHULUAN

Salah satu penyebab kematian terbesar di dunia adalah penyakit kanker. Kanker paru-paru merupakan salah satu kanker yang paling sering terjadi di seluruh dunia dan mencapai 19,4% dari total kematian dari penyakit kanker (Yang & Chen, 2015). Menurut proyeksi World Health Organization (WHO), pada tahun 2040 angka kasus kanker paru-paru di Indonesia akan meningkat 80% dari 2018 dengan penyebab tertinggi kematian dari kanker merokok (WHO, 2020).

Dengan munculnya penyakit kanker paru-paru di tubuh manusia terdapat beberapa gejala yang biasa dirasakan oleh kebanyakan pasien kanker paru-paru. Namun, gejala ini sering kali tidak dihiraukan dan tidak dicek oleh praktisi medis atau dokter sehingga hanya 14% dari pasien yang didiagnosis kanker paru-paru sembuh dari penyakitnya lima tahun dari diagnosis (Krishnaiah, Narsimha, & Chandra, 2013). Berdasarkan penelitian yang dilakukan di RS Paru Dr. HA Rotinsulu Bandung, sebagian besar pasien yang didiagnosis oleh kanker paru-paru sudah berada di stadium III hingga stadium IV (Sholih, et al., 2019). Oleh karena itu, penting diadakannya prediksi kanker paru-paru menggunakan gejala awal dan faktor risiko untuk mencegah hal tersebut.

Teknologi telah berevolusi dari awal penciptaannya sampai sekarang dengan laju yang cepat. Salah satu aspek dari evolusi teknologi tersebut merupakan berkembangnya pertukaran data yang terjadi. Dengan skala pertukaran data yang besar tentu jumlah data yang berputar semakin besar. Untuk menggunakan dan menggali data tersebut menjadi informasi yang bisa digunakan terciptanya konsep *data mining* yang adalah teknik untuk mencari pola dan informasi dari data dengan jumlah besar (Han, Kamber, & Pei, 2012).

Data mining bisa diimplementasikan pada banyak industri, salah satunya merupakan industri kesehatan. Penggunaan *data mining* untuk membantu riset dan penanganan kanker sedang sangat meningkat. Klasifikasi merupakan bentuk analisis data dimana terdapat model atau *classifier* yang digunakan untuk memprediksi *class labels* (Han, Kamber, & Pei, 2012). Menggunakan klasifikasi, terdapat cara untuk menganalisis data hasil gejala awal dan faktor risiko kanker paru-paru untuk menentukan *class label* dari data tersebut.

Dalam penggunaan klasifikasi untuk riset dalam industri kesehatan, terdapat beberapa model yang sering digunakan oleh peneliti. Penelitian yang dilakukan oleh Fadlilah, Wihandika, & Rahayudi (2019) melakukan klasifikasi dengan objek fungsi kognitif pasien stroke dan menggunakan metode Random Forest Decision Trees. Hasil dari penelitian Fadlilah, Wihandika, & Rahayudi (2019) merupakan hasil rata-rata akurasi 53,094% dengan jumlah fitur optimal 13 dan jumlah *trees* optimal 100 *tree*. Pengujian dilakukan menggunakan K-Fold Cross Validation. (Fadlilah, Wihandika, & Rahayudi, 2019). Dalam penelitian yang dilakukan Yang & Chen (2015) dilakukan prediksi stadium kanker paru-paru menggunakan informasi klinis dan patologi seperti rontgen, *CT Scan*, dan biopsi. Penelitian Yang & Chen (2015) menggunakan metode Decision Trees dengan hasil rata-rata akurasi 81,97& (Yang & Chen, 2015).

Penggunaan metode Random Forest Decision Trees pada penelitian ini didukung oleh Denisko dan Hoffman (2018) yang menyimpulkan bahwa dalam penggunaan klasifikasi di industri medis, metode Random Forest Decision Trees memiliki kelebihan seperti performa yang tinggi, bisa ditemukannya kepentingan dari setiap fitur, dan interpretasi yang mudah (Denisko & Hoffman, 2018). Pengambilan fitur yang digunakan pada penelitian ini juga didasarkan dari penelitian yang dilakukan Sholih (2019) yang menjabarkan gejala-gejala dan faktor risiko kanker paru-paru di Indonesia. Oleh karena itu, penelitian ini akan melakukan prediksi kanker paru-paru menggunakan metode Random Forest Decision Trees dengan metode Random Forest Decision Trees dan pemilihan fitur akan berdasarkan gejala-gejala dan faktor risiko kanker paru-paru di Indonesia.

2. LANDASAN KEPUSTAKAAN

2.1 Random Forest

Metode Random Forest Decision Trees adalah pengembangan lebih dalam dari metode CART. Yang membedakan metode Random Forest Decision Trees dengan metode CART adalah jumlah *tree construction* saat menggunakan metode Random Forest Decision Trees lebih dari satu. *Output* atau keluaran

klasifikasi atau hasil dari regresi dari setiap *tree* akan di*voting* dengan tujuan menemukan kelas atau label yang paling banyak dihasilkan. Metode Random Forest Decision Trees bisa disebut *random* karena dalam pembuatan *tree* data yang digunakan akan diacak secara tidak beraturan terlebih dahulu (Fadlilah, Wihandika, & Rahayudi, 2019).

1. Pengacakan Data

Pembangunan CART menggunakan nilai entropi dan *information gain*. Perhitungan tersebut ada pada persamaan 2.1 dan 2.2

$$Entropi\ (S) = \sum_{j=1}^{k} -p_j * log2p_j \quad (1)$$

Nilai *entropy* fitur merupakan jumlah total dari minus jumlah terjadi dibagi dengan jumlah kasus dan dikali dengan log2 jumlah terjadi dibagi dengan jumlah kasus.

2. $Gain\ (S, A) = Entropy\ (S) - \sum_{j=1}^{k} Entropy\ (A)$ (2)

Nilai *information gain* merupakan hasil pengurangan dari entropi kelas dan entropi fitur.

3. *Voting*
4. Eror Random Forest
5. Kondisi Berhenti

2.2 Confusion Matrix

Confusion Matrix adalah alat yang bisa digunakan untuk mengevaluasi model algoritme. Dalam klasifikasi dengan dua kelas atau Confusion Matrix 2x2 maka bisa ditunjukkan:

1. *True Positives* (TP): TP merepresentasikan data uji positif yang secara benar dimasukkan ke dalam kelas positif.
2. *True Negatives* (TN): TN merepresentasikan data uji negatif yang secara benar dimasukkan ke dalam kelas negatif.
3. *False Positives* (FP): FP merepresentasikan data uji negatif yang secara salah dimasukkan ke dalam kelas positif.
4. *False Negative* (FN): FN merepresentasikan data uji positif yang secara salah dimasukkan ke dalam kelas negatif.

Untuk menyimpulkan hasil dari Confusion Matrix bisa menggunakan *accuracy*, *precision*, dan *recall*. Perumusan untuk *accuracy*, *precision*, *recall*, dan *f1-score* akan ada pada persamaan 3, 4, 5, dan 6 (Yun, 2021).

$$Accuracy = \frac{TP + TN}{TP+TN+FP+FN} \quad (3)$$

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

$$Recall = \frac{TP}{TP+ FN} \quad (5)$$

$$f1 - score = 2(\frac{(Precision)(Recall)}{Precision+Recall}) \quad (6)$$

Keterangan:

Accuracy: Proporsi jumlah total prediksi yang benar dalam bentuk persen (%)

Precision: Proporsi *True Positive* yang diprediksi dan benar. Jumlah yang diklasifikasikan dengan benar sebagai positif dibagi jumlah total yang diklasifikasikan sebagai positif.

Recall: Proporsi *True Positive* yang diidentifikasi dengan benar. Jumlah yang diklasifikasikan dengan benar sebagai positif dibagi dengan jumlah contoh positif aktual.

F1-score: Rata-rata *harmonic* dari *precision* dan *recall*. Nilai terbaik *f1-score* adalah 1.0 dan nilai terburuknya adalah 0.

2.3 K-Fold Cross Validation

K-fold Cross Validation adalah metode pengujian sebuah model algoritme dengan membagi data menjadi *fold* yang sama rata sebanyak *k*. *Training* dan *testing* akan dilakukan sejumlah *k*. Pengunaan K-Fold Cross Validation pada *dataset* yang kecil sangat disarankan karena *dataset* bidsa digunakan semua data untuk *training* dan validasi. Dengan menggunakan semua data untuk validasi, estimasi error yang diberikan akan lebih merepresentasikan populasi data. (Vabalas, Gowen, Poliakoff, & Casson, 2019).

2.4 Faktor Risiko Kanker Paru-Paru

Pada penelitian yang dilakukan (Sholih, et al., 2019) pasien kanker paru-paru diberikan pertanyaan mengenai apakah mereka terekspos pada faktor-faktor risiko kanker paru-paru. Berikut merupakan hasil dari pertanyaan tersebut berurut dari yang paling memengaruhi terhadap kanker paru-paru dan yang paling tidak berpengaruhi kanker paru-paru:

1. Merokok,
2. Polusi Udara,

3. Makanan dan Minuman,
4. Zat Kimia,
5. Pekerjaan,
6. Penyakit Paru-Paru atau Kanker Turunan,
7. Kekurangan Berolahraga,
8. Konsumsi Alkohol,
9. Riwayat Penyakit Kanker Paru-Paru

2.5 Gejala Awal Kanker Paru-Paru

Pada penelitian yang dilakukan oleh (Krishnaiah, Narsimha, & Chandra, 2013) disebutkan bahwa ada 10 gejala awal yang bisa mengindikasi kanker paru-paru. Gejala-gejala tersebut meliputi:

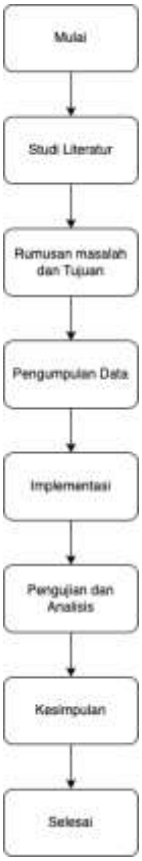
1. Dyspnea (Kesusahan bernafas),
2. Hemoptysis (Batuk berdarah),
3. Batuk tak kunjung sembuh,
4. Sakit di area dada atau perut
5. Cachexia (Berat badan menurun),
6. Dysphonia (Suara serak),
7. Dysphasia (Susah menelan),
8. Rasa sakit di bahu, dada, lengan,
9. Bronkitis
10. Pneumonia

3. METODOLOGI PENELITIAN

3.1 Diagram Alir Penelitian

Dalam penelitian ini, penulis telah melakukan studi literatur dari penelitian sebelumnya baik dari segi metode ataupun topik. Berdasarkan hasill studi literatur, penulis mengidentifikasi dan membuat rumusan masalah yang hendak dibawakan pada penelitian ini. Selanjutnya, penulis menetapkan tujuan dari penelitian ini berdasarkan rumusan masalah yang telah dirumuskan. Setelah studi literatur, penetapan tujuan, dan perumusan masalah selesai, penulis mulai mengumpulkan data untuk penelitian yang hendak dilakukan.

Gambar 1 di bawah ini merupakan langkah-langkah sistem:



Gambar 1 Diagram Alir Penelitian

Pada penelitian ini, data yang digunakan adalah data pasien yang telah didiagnosa kanker paru-paru pada 3 tahun terakhir. Setelah data dikumpulkan, tahap selanjutnya adalah implementasi metode. Implementasi metode akan melalui dua tahap, yaitu pemrosesan data melalui *preprocessing* dan pembuatan serta pelatihan model menggunakan metode Random Forest Decision Trees. Setelah implementasi selesai, maka dilaksanakan pengujian dan analisis. Hasil dari pengujian dan analisis diharapkan bisa menemukan kesimpulan.

3.2 Pengumpulan Data

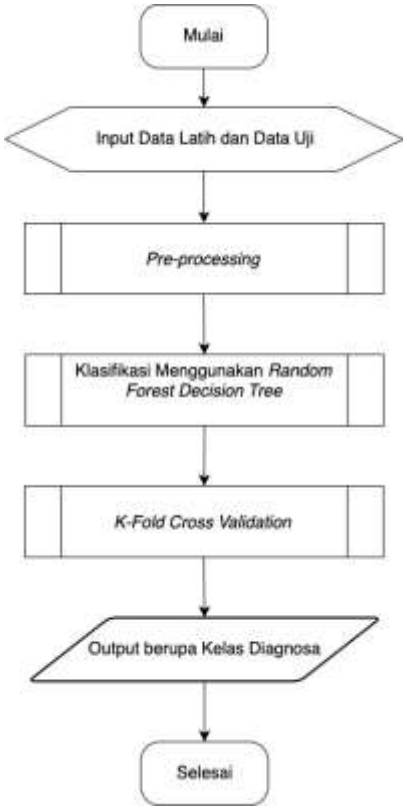
Tahapan pengambilan data dilakukan melalui Google Form kepada responden dengan kriteria telah didiagnosis dengan kanker paru-paru. Pertanyaan dari kuesioner merupakan fitur-fitur yang merupakan gejala-gejala dan faktor risiko berdasarkan penelitian oleh Sholih (2019). Fitur-fitur tersebut merupakan kesulitan bernafas, batuk berdarah, batuk yang tak kunjung sembuh, rasa sakit di area data atau perut, penurunan berat badan, suara serak, kesusahan menelan, rasa sakit di daerah tangan atau pundak, bronkitis, pneumonia, merokok, memakan makanan bergizi, mengkonsumsi

alkohol, kekurangan berolahraga, memiliki riwayat kanker turunan, dan tereskpos ke zat kimia. Responden menjawab survey yang telah diberikan berdasarkan saat mereka didiagnosis kanker paru-paru. Responden merupakan pasien dari dr. Andika Chandra Putra, Sp.P, Dokter spesialis paru dan kedokteran respirasi. Terdapat juga pengambilan data melalui kuesioner yang berbeda untuk yang tidak terdiagnosis atau non-pasien kanker paru-paru dengan pertanyaan yang sama. Data yang dikumpulkan terjumlah 43 dengan data terbagi menjadi 18 pasien kanker paru-paru dan 25 yang tidak terdiagnosis kanker paru-paru.

3.3 Implementasi

Tahap implementasi adalah tahapan penerapan klasifikasi menyesuaikan dengan rancangan yang telah dibuat terlebih dahulu. Implementasi pada penelitian ini dilakukan menggunakan bahasa pemrograman Pyhton. Pemilihan bahasa Pyhton karena penulis lebih familiar dengan bahasa Pyhton dan juga simplisitas penulisan kode tanpa harus menuliskan tipe data sebelumnya dan variabel yang dinamis.

Penulis menggunakan metode Random Forest Decision Trees untuk menyelesaikan masalah pada objek yang telah ditentukan. Tahapan yang akan dimulai adalah dengan menginputkan data yang akan digunakan untuk pelatihan dan pengujian, selanjutnya dilakukan pra-pemrosesan data untuk mem*filter* data, menghilangkan data *null*, *splitting* data, dan normalisasi. Setelah itu dilakukan proses klasifikasi menggunakan metode Random Forest Decision Tree. Setelah proses klasifikasi perlu dilakukan K-Fold Cross Validation untuk menentukan jumlah fitur dan jumlah *tree* dari klasifikasi. Tahapan-tahapan tersebut akan dijelaskan lebih dalam pada Gambar 2.



Gambar 2 Diagram Alir Program

3.4 Evaluasi

Pengujian model algoritme Random Forest Decision Trees menggunakan Confusion Matrix akan 10 uji coba dengan menggunakan *dataset* yang sama tetapi dengan Train Test Split sehingga pembagian data latih dan data uji akan berbeda setiap kali dilakukan *testing*. Hal ini dilakukan untuk mencari rata-rata akurasi dan Classification Report yang berisi *precision*, *recall*, dan *f1-score*.

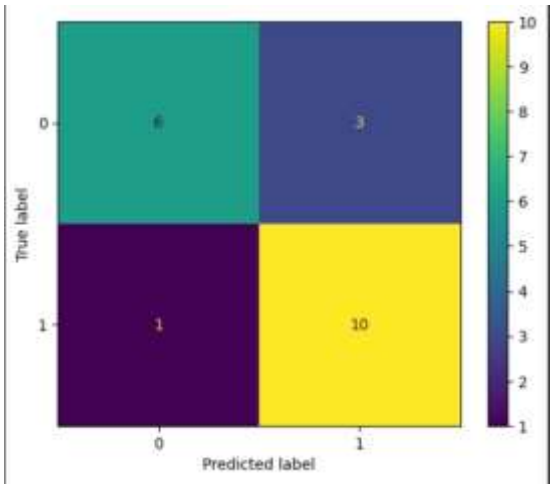
Pengujian model algoritme Random Forest Decision Trees menggunakan K-Fold Cross Validation akan dijalankan ujicoba dengan 2, 5, dan 8-*fold*. Dalam pengujian K-Fold Cross Validation *dataset* akan dibagi menjadi *fold* dengan besar yang sama rata sebanyak *k*. *Training* dan *testing* akan dilakukan sejumlah *k* dimana setiap percobaan menggunakan data yang dipartisi ke-K sebagai data uji dan sisa data akan digunakan sebagai data latih.

4. HASIL DAN PEMBAHASAN

4.1 Hasil Pengujian *Confusion Matrix*

Berikut akan ditampilkan hasil salah satu ujicoba Confusion Matrix yang ditunjukkan pada Gambar 3. Pada Gambar 3, label 0 menunjukkan terindikasi kanker paru-paru dan

label 1 menunjukkan tidak terindikasi kanker paru-paru. Menggunakan Gambar 4 dapat diketahui juga nilai *accuracy*, *precision*, *recall*, dan *f1-score*.



Gambar 3 Hasil *Confusion Matrix*

Hasil dari pengujian pada data uji merupakan nilai-nilai berikut, yaitu, nilai dari *True Positive* (TP) sejumlah 6, nilai dari *True Negative* (TN) sejumlah 10, nilai dari *False Positive* (FP) sejumlah 3, dan nilai dari *False Negative* (FN) sejumlah 1. Dari keempat nilai tersebut bisa ditemukan nilai *accuracy*, *precision*, *recall*, dan *f1-score*. Perhitungan manual dari nilai *accuracy*, *precision*, *recall*, dan *f1-score* dapat dilihat di bawah ini.

1. Perhitungan pada semua kelas

   a. *Accuracy*

   $$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

   $$Accuracy = \frac{6+10}{6+10+3+1} = 0.8$$

2. Perhitungan pada kelas kanker

a. *Precision*

$$Precision = \frac{TP}{TP+FP}$$

$$Precision = \frac{6}{7} = 0.86$$

b. *Recall*

$$Recall = \frac{TP}{TP+FN}$$

$$Recall = \frac{6}{9} = 0.67$$

c. *f1-score*

$$f1\text{-}score = \frac{2 * precision * recall}{precision+recall}$$

$$f1\text{-}score = \frac{2 * 0.86 * 0.67}{0.86 + 0.67} = 0.75$$

3. Perhitungan pada kelas tidak kanker

a. *Precision*

$$Precision = \frac{TN}{TN+FN}$$

$$Precision = \frac{10}{13} = 0.77$$

b. *Recall*

$$Recall = \frac{TN}{TN+FP}$$

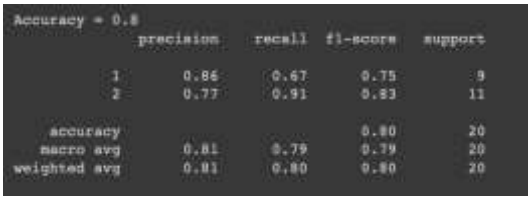
$$Recall = \frac{10}{11} = 0.91$$

c. *f1-score*

$$f1\text{-}score = \frac{2 * precision * recall}{precision+recall}$$

$$f1\text{-}score = \frac{2 * 0.77 * 0.91}{0.77 + 0.91} = 0.83$$

Untuk melakukan validasi terhadap perhitungan manual, maka akan dilakukan perhitungan *accuracy*, *precision*, *recall*, dan *f1-score* menggunakan *classification report* dari *scikit-learn*. Berikut hasil perhitungan menggunakan *classification report* pada Gambar 4.



Accuracy = 0.8

	precision	recall	f1-score	support
1	0.86	0.67	0.75	9
2	0.77	0.91	0.83	11
accuracy			0.80	20
macro avg	0.81	0.79	0.79	20
weighted avg	0.81	0.80	0.80	20

Gambar 4 *Classification Report*

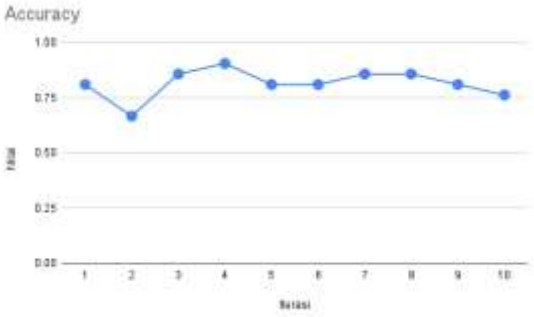
Hasil pengujian menggunakan *confusion matrix* untuk uji coba lainnya bisa dilihat pada Tabel 1 berikut.

Tabel 1 Hasil Pengujian Train Test Split

Iterasi	Akurasi	Kelas	*Precision*	*Recall*	*F1-score*
1	0.8095238095	1	0.75	0.75	0.75
		2	0.85	0.85	0.85
2	0.6666666667	1	0.55	0.75	0.63
		2	0.8	0.62	0.7
3	0.8571428571	1	1	0.62	0.7
		2	0.8	1	0.9
4	0.9047619048	1	0.9	0.9	0.9
		2	0.91	0.91	0.91
5	0.8095238095	1	0.71	0.71	0.71
		2	0.86	0.86	0.86
6	0.8095238095	1	0.86	0.67	0.75
		2	0.79	0.92	0.85
7		1	0.86	0.75	0.8

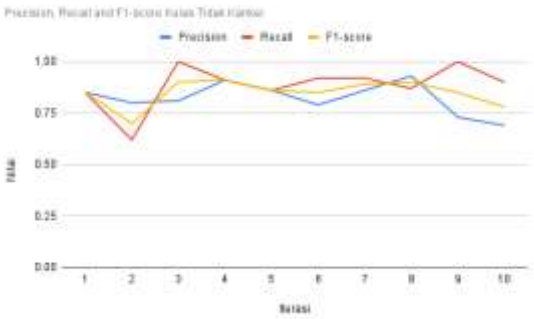
	0.857142857143	2	0.86	0.92	0.89
8	0.857142857143	1	0.71	0.83	0.77
		2	0.93	0.87	0.9
9	0.809523809095	1	1	0.6	0.75
		2	0.73	1	0.85
10	0.761904761619	1	0.88	0.64	0.74
		2	0.69	0.9	0.78
Rata-rata	0.813285714143	1	0.822	0.722	0.757
		2	0.823	0.885	0.849

Representasi grafik pada Tabel 4.14 hasil pengujian K-Fold Cross Validation dapat dilihat pada Gambar 5, 6, dan 7.



Gambar 5 Hasil Akurasi Confusion Matrix

Gambar 6 Hasil *precision*, *recall*, dan *f1-score* Kelas Kanker



Gambar 7 Hasil *precision*, *recall*, dan *f1-score* Kelas Tidak Kanker

4.2 Hasil Pengujian Confusion Matrix

Pada pengujian dengan menggunakan Confusion Matrix dengan melakukan 10 kali uji coba ditemukan bahwa nilai akurasi tertinggi adalah sebesar 0,9047619048 yang didapatkan pada uji coba ke-4. Ditemukan juga nilai akurasi terendah adalah sebesar 0,6666666667 yang didapatkan pada uji coba ke-2. Rata-rata nilai akurasi dari semua uji coba adalah 0,8132857143.

Pengujian menggunakan Confusion Matrix juga menghasilkan *Classification Report* yang berisi *precision*, *recall*, dan *f1-score*. Nilai tertinggi *precision* pada kelas 1 ditemukan pada dua uji coba yaitu nilai sebesar 1.0 yang ditemukan pada uji coba ke-3 dan uji coba ke-8. Rata-rata nilai *precision* pada kelas 1 adalah 0,822. Nilai tertinggi *precision* pada kelas 2 ditemukan pada uji coba ke-8 dengan nilai sebesar 0,93. Rata-rata nilai *precision* pada kelas 2 adalah 0,823.

Nilai tertinggi *recall* pada kelas 1 adalah sebesar 0,9 yang ditemukan pada uji coba ke-4. Rata-rata nilai *recall* pada kelas 1 adalah 0,722. Nilai tertinggi *recall* pada kelas 2 ditemukan pada dua uji coba yaitu nilai sebesar 1.0 yang ditemukan pada uji coba ke-3 dan uji coba ke-9. Rata-rata nilai *recall* pada kelas 2 adalah 0,885.

Nilai tertinggi *f1-score* pada kelas 1 kelas 1 adalah sebesar 0,9 yang ditemukan pada uji coba ke-4. Rata-rata nilai *Precision* pada kelas 1 adalah 0,757. Nilai tertinggi *precision* pada kelas 2 ditemukan pada uji coba ke-4 dengan nilai sebesar 0,91. Rata-rata nilai *Precision* pada kelas 2 adalah 0,849.

4.3 Hasil Pengujian K-Fold Cross Validation

Pengujian sistem menggunakan K-Fold Cross Validation dilakukan 3 kali dengan menggunakan 2, 5, dan 8-*fold*. Dalam pengujian K-Fold Cross Validation *dataset* akan dibagi menjadi *fold* dengan besar yang sama rata sebanyak *k*. Hasil Pengujian K-Fold Cross Validation dapat dilihat pada Tabel 4.15.

Tabel 2 Hasil Pengujian K-Fold Cross Validation

	Jumlah Data	Akurasi

K-fold	Ujicoba ke-	Data Uji	Data Latih	
2	1	22	22	0.80952381
	2	22	22	0.76190476
5	1	10	34	0.88888889
	2	10	34	0.77777778
	3	10	34	0.875
	4	10	34	0.75
	5	10	34	0.75
8	1	6	38	0.83333333
	2	6	38	0.66666667
	3	6	38	0.8
	4	6	38	0.8
	5	6	38	1
	6	6	38	0.4
	7	6	38	0.8
	8	6	38	0.6

Representasi grafik pada Tabel 4.13 hasil pengujian K-Fold Cross Validation dapat dilihat pada Gambar 8.

Gambar 7 Hasil Akurasi K-Fold Cross Validation

4.4 Hasil Pengujian Confusion Matrix

Pengujian K-Fold Cross Validation yang menggunakan berbagai nilai *k* menghasilkan menggunakan nilai *k* sebesar 5 memiliki akurasi tertinggi dengan akurasi sebesar 0,808333334 dengan 10 data uji dan 34 data latih. Uji coba pertama menunjukkan akurasi tertinggi yang menghasilkan akurasi sebesar 0,88888889.

Uji coba K-Fold Cross Validation juga dilakukan dengan nilai *k* sebesar 2 yang dengan 22 data uji dan 22 data latih. Uji coba ke-1 menghasilkan akurasi sebesar 0,80952381 dan uji coba ke-2 menghasilkan akurasi sebesar 0,76190476. Pada pengujian menggunakan nilai k sebesar 8 dengan 6 data uji dan 38 data latih ditemukan terdapat perbedaan yang jauh antara akurasi uji coba. Uji coba kelima menghasilkan akurasi sebesar 1. Namun, uji coba ke-6 menghasilkan akurasi sebesar 0,4. Terdapat 3 uji coba yang menghasilkan akurasi yang sama sebesar 0,8, yaitu uji coba ketiga, uji coba keempat, dan uji coba ketujuh.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

1. Pembentukan rancangan model algoritme Random Forest Decision Trees membutuhkan 16 fitur yang berisi gejala-gejala kanker paru-paru dan faktor risiko kanker paru-paru. Gejala dan faktor risiko tersebut diambil dari penelitian yang telah ada.
2. Pada pengujian dengan menggunakan Confusion Matrix dengan melakukan 10 kali uji coba ditemukan bahwa nilai akurasi tertinggi adalah sebesar 0,9047619048 yang didapatkan pada uji coba ke-4. Ditemukan juga nilai akurasi terendah adalah sebesar 0,6666666667 yang didapatkan pada uji coba ke-2. Rata-rata nilai akurasi dari semua uji coba adalah 0,8132857143. Sementara pengujian K-fold Cross Validation yang menggunakan berbagai nilai k menghasilkan menggunakan nilai *k* sebesar 5 memiliki rata-rata akurasi tertinggi dengan akurasi sebesar 0,808333334 dengan 10 data uji dan 34 data latih.
3. Hasil dari penelitian prediksi diagnosis kanker paru-paru menggunakan *data mining* adalah model algoritme Random Forest Decision Trees dengan akurasi 0.813.

5.2 Saran

1. *Dataset* yang digunakan pada penelitian ini masih termasuk sedikit dengan jumlah data

44. Hal ini menyebabkan permasalahan saat pengujian menggunakan Train Test Split. Penulis menyarakankan untuk memperbanyak jumlah data dengan cara bekerja sama dengan rumah sakit kanker yang bisa membantu dalam pencarian data.

2. Diharapkan pada penelitan selanjutnya dilakukan pengujian untuk jumlah *tree* yang digunakan sehingga hasil lebih akurat. Pengujian tersebut bisa dilakukan menggunakan K-Fold Cross Validation.

3. Diharapkan pada penelitian selanjutnya pengimplementasian *website* dilakukan sampai *hosting* sehingga bisa diakses oleh masyarakat umum.

6. DAFTAR PUSTAKA

Sholih, M. G., Perwitasari, D. A., Hendriani, R., Sukandar, H., Barliana, M. I., Suwantika, A., Diantini, A. (2019). Risk factors of Lung Cancer in Indonesia: a qualitative study. *Journal of Advanced Pharmacy Education & Research*, 41-45.

WHO. (2020, Januari 1). Retrieved from https://www.who.int/publications/m/item/cancer-idn-2020

Krishnaiah, V., Narsimha, D., & Chandra, D. (2013). Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. *International Journal of Computer Science and Information Technologies, Vol. 4 (1)*, 39-45.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.* Waltham: Morgan Kaufmann Publishers.

Fadlilah, M. S., Wihandika, R. C., & Rahayudi, B. (2019). Klasifikasi Penurunan Fungsi Kognitif Pasien Stroke Menggunakan Metode Klasifikasi Random Forest. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol.3*, 3005-3013.

Yun, H. (2021). Prediction model of algal blooms using logistic regression and confusion matrix. *International Journal of Electrical and Computer Engineering (IJECE)*, 2407-2413.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. (2019, November 7). Machine learning algorithm validation with a limited sample size. *PLoS ONE 14*, p. 11.

WHO. (2021, December 13). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cancer-in-children

Pusat Data dan Informasi. (2020, Oktober 16). Retrieved from www.pdpersi.co.id

Yang, H., & Chen, Y.-P. P. (2015). Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information. *Expert Systems with Applications 42*, 6168-6176.

Denisko, D., & Hoffman, M. M. (2018, February 20). Classification and interaction in random forests. *PNAS*, pp. 1690-1692.