Infotekmesin
Vol.14, No.01, Januari 2023
p-ISSN: 2087-1627, e-ISSN: 2685-9858
DOI: 10.35970/infotekmesin.v14i1.1751, pp.155-162

Penerapan Data *Mining* dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma *Random Forest*

Laura Sari[1*], Annisa Romadloni[2], Rostika Listiyaningrum[3]
[1, 3]Program Studi Teknik Informatika, Politeknik Negeri Cilacap
[1,2]Jln. Dr. Soetomo No.1 Karangcengis Sidakaya, Kabupaten Cilacap, 53212, Indonesia
E-mail: laurasari@pnc.ac.id[1], annisa.romadloni@pnc.ac.id[2], nadhifa007@gmail.com[3]

Info Naskah:
Naskah masuk: 18 Januari 2023
Direvisi: 19 Januari 2023
Diterima: 27 Januari 2023

Abstrak

Kanker merupakan penyebab kematian tertinggi kedua di dunia. Di Indonesia termasuk penyakit dengan tingkat kematian yang tinggi. Sebagian besar penderita tidak mengetahui bahwa dirinya terkena kanker paru sehingga penanganan menjadi terlambat. Metode prediksi dengan tingkat akurasi yang tinggi diperlukan untuk mendeteksi secara dini kanker paru. Penelitian sebelumnya menggunakan metode kalsifikasi data *mining* dengan algoritma *Naïve Bayes* untuk meprediksi terjadinya kanker paru. Penelitian tersebut menghasilkan nilai *recall* yang tinggi untuk kelas positif (kelas *Yes*) namun rendah untuk kelas negatif (kelas *No*). Penelitian ini dibuat dengan algoritma *Random Forest* yang dikenal memilki performa yang baik. Pemodelan dioptimalkan dengan menerapkan teknik *K-fold Cross Validation*. Algoritma *Random Forest* menghasilkan nilai *Accuracy* yang lebih tinggi daripada algoritma Naïve Bayes, yaitu sebesar 98,4%. Bahkan algoritma ini menghasilkan *Recall* 100% untuk kelas positif dan 80% untuk kelas negatif serta memberikan prediksi 100% benar terlihat dari nilai AUC sebesar 1. Meskipun uji statistik dengan tingkat signifikansi 5% menunjukkan hasil dari kedua algoritma tersebut tidak berbeda secara signifikan.

Keywords:
data mining;
random forest;
prediction;
naïve bayes.

Abstract

Cancer is the second highest cause of death in the world. In Indonesia, it is a disease with a high mortality rate. Most patients do not realize that they have lung cancer thus the treatment is sometimes too late. A prediction method with a high degree of accuracy is needed to detect lung cancer earlier. Previous research used data mining calcification methods with the Naïve Bayes algorithm to predict lung cancer. This research resulted in high recall values for the positive class (Yes class) but low for the negative class (No class). This research was made using the Random Forest algorithm which is known to have good performance. The modeling is optimized by applying the K-fold Cross Validation technique. The Random Forest algorithm produces a higher Accuracy value than the Naïve Bayes algorithm, which is 98.4%. This algorithm produces 100% Recall for the positive class, 80% for the negative class and provides a 100% correct prediction as can be seen from the AUC value of 1. Although a statistical test with a significance level of 5% shows the results of the two algorithms are not significantly different.

***Penulis korespondensi:**
Laura Sari
E-mail: laurasari@pnc.ac.id

1. Pendahuluan

Kanker, menurut definisi dari National Cancer Institute, adalah penyakit genetik yang yang disebabkan oleh perubahan gen yang mengontrol fungsi sel, terutama fungsi untuk tumbuh dan membelah. Kanker merupakan sel-sel baru yang tumbuh secara abnormal kemudian menyerang bagian tubuh kontralateral dan menyebar ke organ lain [1]. Kanker merupakan penyebab kematian tertinggi kedua di dunia. Menurut *The Global Cancer Burden* memperkirakan kasus baru kanker meningkat menjadi 19,3 juta dan sekitar 10 juta kematian pada tahun 2020. Satu dari 5 orang di seluruh dunia mengidap kanker selama hidup mereka. Satu dari 8 pria dan satu dari 11 wanita meninggal karena penyakit ini [2].

Data dari Global Cancer Observatory dari WHO menyatakan 10 jenis kanker paling mematikan di dunia. Kanker paru menduduki posisi pertama dengan 1.796.144 kematian disusul kanker kolokretal (935.173 kematian) dan terakhir kanker prostat (375.304 kematian). Di Indonesia kanker paru menjadi jenis kanker dengan angka kejadian tertinggi sekitar 34.783 kasus baru dan 30.843 kematian selama tahun 2020 [3]. Faktor utama kanker di Indonesia adalah merokok, selain merokok secara langsung, asap rokok yang terhirup juga maningkatkan resiko kanker paru. Faktor lainnya yaitu genetik, ada Riwayat kanker pada keluarga, minum kopi lebih dari 6 gelas/hari, penyakit paru kronik, konsumsi alkohol, konsumsi daging yang digoreng atau dipanggang, polusi udara, dan terpapar zat kimia [4].

Gejala dari penyakit menurut Mayo Clinic secara umum diantaranya kelelahan, adanya benjolan, perubahan berat badan, perubahan kulit seperti menguning, batuk terus menerus, nyeri otot, suara serak, dan kesulitan menelan [5]. Kanker paru sering terdiagnosis pada stadium lanjut. Bahkan sekitar 60-85% pasien kanker paru tidak mengetahui penyakitnya. Hal ini dikarenakan pasien menganggap batuk dan sesak yang diderita merupakan hal biasa dan tidak memeriksakan diri ke layanan kesehatan. Selain itu, kebanyakan petugas kesehatan hanya mengatasi gejala tanpa melakukan pemeriksaan lanjutan [4]. Oleh sebab itu diperlukan suatu metode yang dapat secara akurat mendeteksi kanker paru melalui prediksi dari beberapa faktor dan gejala yang timbul.

Data mining telah banyak digunakan untuk mendeteksi suatu penyakit. *Data mining* merupakan suatu proses meringkas suatu pengetahuan menggunakan algoritma untuk mendeteksi pola spesifik, kecenderungan dalam data, dan aturan mekanis. Algoritma ini digunakan untuk mengetahui hubungan antar data yang sebelumnya tidak terlihat [6].

Penelitian mengenai kanker paru telah dilakukan oleh Wulandari dan Perdana (2022). Penelitian tersebut menggunkaan algoritma Naïve Bayes dan bertujuan untuk memprediksi apakah seseorang menderita kanker paru (yes/no). Berdasarkan pengukuran tingkat akurasi algoritma tersebut, diperoleh tingkat persentasi *recall* untuk kelas positif 98,77% dan kelas negatif sebesar 66,67%. Nilai *precision* sebesar 95,24% dan tingkat akurasi sebesar 94,62%. Penelitian tersebut menghasilkan model dengan *recall* kelas positif (kelas Yes) yang tinggi namun rendah untuk kelas negatif (kelas No).

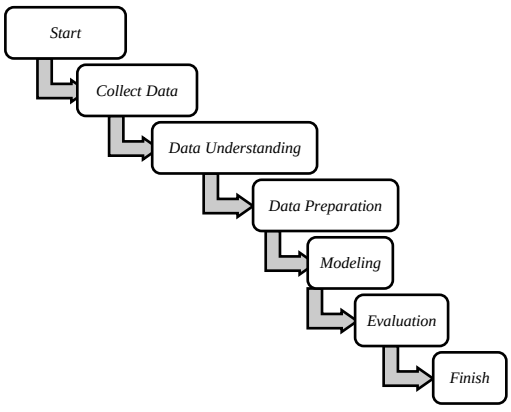
Naïve Bayes merupakan algoritma *data mining* yang sering digunakan. Algoritma ini melakukan komputasi yang cepat karena algoritmanya cukup sedrhana. Namun, naïve bayes rentan terhadap bias karena kelangkaan data (*scarcity data*). Selain itu juga membutuhkan asumsi seperti semua prediktor data sama penting dan *independent* [7].

Penelitian terpisah mengenai data *mining* dilakukan oleh Depari, dkk (2022) dengan berjudul Perbandingan Model *Decision Tree*, *Naïve Bayes* dan *Random Forest* untuk Prediksi Klasifikasi Penyakit Jantung. Penelitian tersebut menghasilkan nilai akurasi Random Forest lebih baik dari Naïve Bayes. Hasil yang sama diperoleh dari penelitian yang dilakukan oleh [8], [9], [7], [10], [11], [12], [13], [14], dan [15]. *Random Forest* merupakan algoritma berbasis esemble yang dibangun berdasarkan algoritma *Decision Tree* dan dikenal memiliki performa yang baik. Algoritma berbasis esemble adalah gabungan dari beberapa teknik pembelajaran mesin (*machine learning*) yang digabungkan menjadi satu model prediktif. Algoritma tersebut dibuat untuk mengurangi kesalahan, bias, dan meningkatkan ketepatan prediksi.

Perbedaan antara penelitian ini dengan penelitian sebelumnya meliputi dua hal. Penelitian ini tidak hanya menerapkan penggabungan Algoritma *Random Forest* dan *K-Fold Cross Validation*, akan tetapi juga menggunakan teknik SMOTE pada tahap *construct* data. Kedua hal ini dilakukan sebagai upaya untuk mengoptimalkan performa model prediksi. Penelitian ini bertujuan untuk mengetahui tingkat akurasi algoritma *data mining* yaitu *Random Forest* dalam memprediksi kanker paru. Selanjutnya hasilnya akan dibandingkan dengan hasil dari algoritma *Naïve Bayes*.

2. Metode

Tahapan penelitian yang dilakukan dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1 *Collect Data*

Data yang digunakan dalam penelitian ini berupa data sekunder yaitu dataset *Survey Lung Cancer* yang diperoleh melalui web *www.kaggle.com*. Data berbentuk file cvs yang terdiri dari 16 kolom dan 309 baris. Dimana kolom menyatakan variabel dan baris menyatakan banyak responden. Lebih jelas, sampel data pada dataset ditunjukkan pada Tabel 1.

Adapun keterangan dari setiap kolom dijelaskan dalam Tabel 2. Dari 16 kolom, hanya terdapat satu data bertipe numerik yaitu *AGE* nilainya berkisar antara 21 hingga 87 tahun. Sedangkan 15 kolom lain bertipe kategorik dan masing-masing memiliki 2 kategori yaitu 1 / 2 dan YES / NO.

Tabel 1. Sampel Dataset

survey_lung_cancer

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY
1	M	69	1	2	2	1	1	2	1
2	M	74	2	1	1	1	2	2	2
3	F	59	1	1	1	2	1	2	1
4	M	63	2	2	2	1	1	1	1
5	F	63	1	2	1	1	1	1	1
6	F	75	1	2	1	1	2	2	2
7	M	52	2	1	1	1	1	2	1
8	F	51	2	2	2	2	1	2	2
9	F	68	2	1	2	1	1	2	1

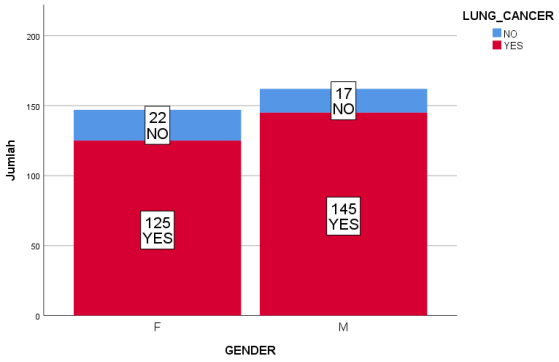
Tabel 2. Penjelasan Kolom

No	Kolom	Keterangan
1	*Gender*	Jenis Kelamin (M = Pria, F = Wanita)
2	*Age*	Umur Pasien (21 – 87)
3	*Smoking*	Status merokok (1 = No, 2 = Yes)
4	*Yellow Finger*	Jari menguning (1 = No, 2 = Yes)
5	*Anxiety*	Gangguan kecemasan (1 = No, 2 = Yes)
6	*Peer Pressure*	Tekanan lingkungan (1 = No, 2 = Yes)
7	*Chronic Disease*	Penyakit kronis (1 = No, 2 = Yes)
8	*Fatigue*	Kelelahan (1 = No, 2 = Yes)
9	*Allergy*	Memiliki alergi (1 = No, 2 = Yes)
10	*Wheezing*	Napas berbunyi (1 = No, 2 = Yes)
11	*Alcohol*	Konsumsi alkohol (1 = No, 2 = Yes)
12	*Coughing*	Batuk (1 = No, 2 = Yes)
13	*Shortness of Breath*	Napas pendek/tersengal (1 = No, 2 = Yes)
14	*Swallowing Difficulty*	Kesulitan menelan (1 = No, 2 = Yes)
15	*Chest Pain*	Sakit dada (1 = No, 2 = Yes)
16	*Lung Cancer*	Mengidap kanker paru (Yes, No)

2.2 Data Understanding

Dataset *Survey Lung Cancer* terdiri dari sekitar 52% pria dan 48% wanita dengan rata-rata usia 62 tahun. Gambar 2 memperlihatkan bahwa pria dengan kanker paru jumlahnya lebih banyak dari wanita dengan kanker paru. Pada Gambar 3 merupakan perintah untuk mendeskripsikan statistic data.

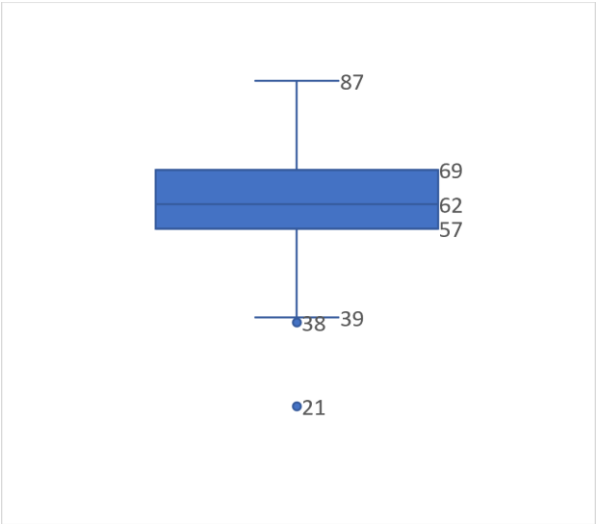
Gambar 4 memperlihatkan distribusi data Usia. Usia responden paling banyak berkisar antara 57 hingga 69 tahun. Dari gambar tersebut juga terlihat adanya pencilan (outlier). Namun hal ini akan diperiksa pada tahap *Data Preparation*. Sebaran pasien kanker paru berdasarkan usia ditunjukkan oleh Gambar 5.



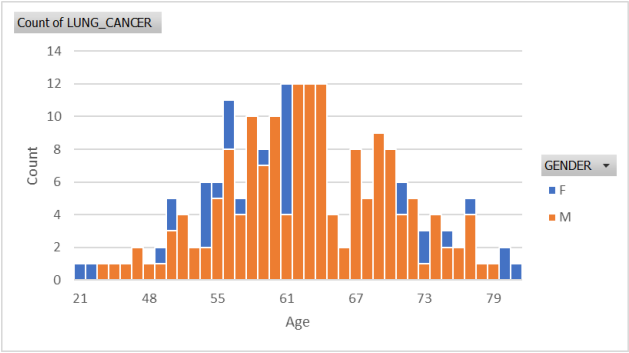
Gambar 2. Diagram Batang Jenis Kelamin

```
> summary(survey_lung_cancer$AGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   57.00   62.00   62.67   69.00   87.00
```

Gambar 3. Statistik Deskriptif Data Usia

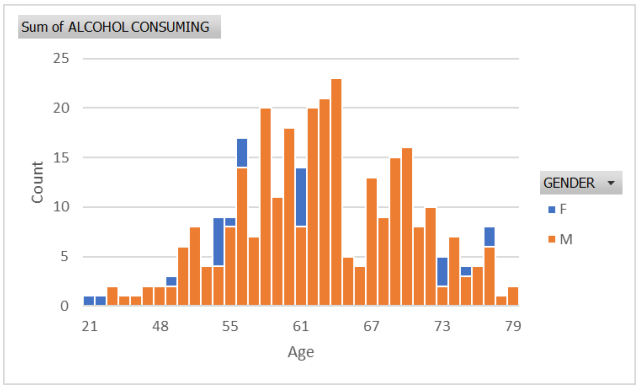


Gambar 4. Boxplot Data Usia

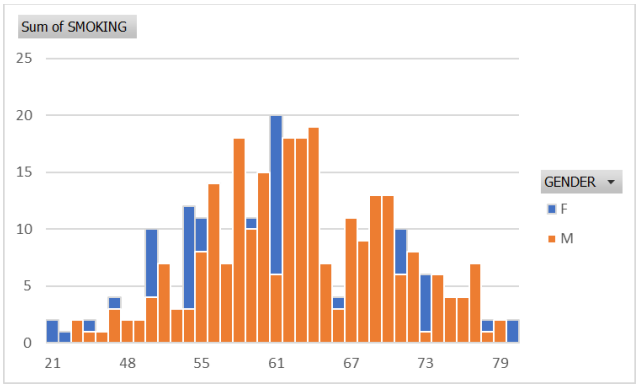


Gambar 5. Sebaran Kanker Paru Berdasarkan Usia

Dari Gambar 2 dan Gambar 5 terlihat bahwa Kanker Paru sebagian besar diderita oleh pria. Hal ini dapat diketahui penyebabnya dengan memperhatikan Gambar 6. Gambar tersebut memperlihatkan bahwa responden yang mengkonsumsi rokok dan alkohol lebih banyak pria di sebagaian besar usia.



(a)



(b)

Gambar 6. (a),(b) Sebaran Pasien Positif yang Mengkonsumsi Alkohol dan Rokok

2.3 Data Preparation

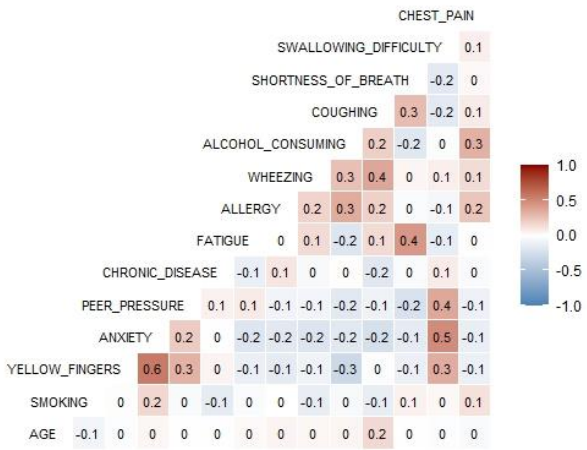
Tahap selanjutnya adalah memeriksa kualitas data. Tahap ini penting agar tujuan dari penelitian dapat tercapai, misalnya hasil prediksi yang dihasilkan lebih akurat. Tahap pertama adalah melakukan identifikasi data pencilan. Data pencilan atau disebut *outlier* adalah data yang sangat berbeda dengan data lainnya. Nilainya terlalu rendah atau terlalu tinggi. Melalui pengamatan sekilas pada Gambar 4, terdapat pencilan pada variabel AGE, yaitu 21 dan 38. Selanjutnya melalui pemeriksaan menggunakan RStudio diperoleh terdapat satu data pencilan yaitu 21.

Tahap kedua melakukan identifikasi data kosong (*missing data*). Data kosong adalah data yang kosong atau hilang. Menurut identifikasi menggunakan RStudio didapatkan dataset yang digunakan sudah cukup baik karena tidak mengandung data kosong, seperti yang diperlihatkan oleh Gambar 7. Setelah tahap identifkasi, selanjutnya dilakukan pembersihan data dengan menghapus satu pencilan yang terdapat dalam dataset. Sehingga data yang digunakan sekarang tersisa 308 data responden.

Selanjutnya dilakukan pengujian korelasi untuk melihat adakah hubungan yang kuat antar variabel (multikolinearitas). Pengujian ini memberikan peringatan dini bahwa variabel tersebut mungkin tidak sesuai untuk beberapa model yang membutuhkan asumsi independent kuat seperti *Naïve Bayes*. Uji korelasi dilakukan menggunakan RStudio dengan algoritma *Pearson Correlation*. Pengujian tersebut menghasilkan matriks korelasi seperti yang terlihat dalam Gambar 8.

```
                GENDER                     AGE
                     0                       0
             SMOKING                YELLOW_FINGERS
                     0                       0
             ANXIETY                PEER_PRESSURE
                     0                       0
     CHRONIC DISEASE                     FATIGUE
                     0                       0
             ALLERGY                     WHEEZING
                     0                       0
   ALCOHOL CONSUMING                     COUGHING
                     0                       0
SHORTNESS OF BREATH SWALLOWING DIFFICULTY
                     0                       0
          CHEST PAIN                  LUNG_CANCER
                     0                       0
```

Gambar 7. Hasil Identifikasi Data Pencilan dan Data Kosong



Gambar 8. Matriks Korelasi

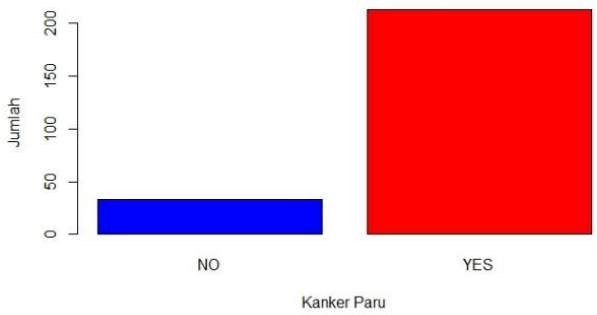
Berdasarkan matriks korelasi di atas, terdapat korelasi/hubungan yang cukup kuat pada beberapa variabel. Variabel-variabel tersebut adalah YELLOW_FINGER dan ANXIETY sehingga masih cukup sesuai jika menggunakan algoritma dengan asumsi independensi kuat seperti *Naïve Bayes*.

Langkah selanjutnya yaitu melakukan *construct* data. Langkah ini dimulai dengan membagi data menjadi 2 bagian yaitu data *train* dan data *test*. Data *train* digunakan untuk membangun model, sedangkan data *test* digunakan untuk menguji model yang telah dibuat sekaligus mengetahui kinerja model. Rasio yang digunakan adalah 80% data *train* dan 20% data *test*. Ukuran ini paling banyak dipilih terutama untuk data berukuran kecil. Hasil dari proses pembagian ini tercipta 246 data *train* dan 15 data *testing*.

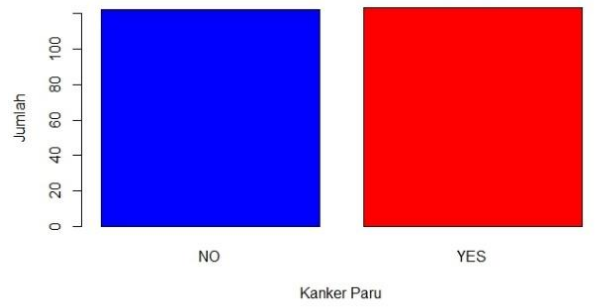
Sebelum dilakukan pemodelan, perlu diperiksa proporsi data *train* pada variabel tujuan (*LUNG_CANCER*).

Hasilnya ditunjukkan Gambar 9. Hasil pemeriksaan memberikan proporsi variabel tujuan (*LUNG_CANCER*) dalam data *train* sebesar 13:87 (*NO*:*YES)*. Proporsi semacam ini belum bisa dikatakan seimbang. Data *train* yang tidak seimbang akan mempengaruhi proses pengklasifikasian, yang menyebabkan bias terhadap kelas mayoritas. Hal ini dapat mengakibatkan *error* yang tinggi bahkan penghilangan sama sekali kelas minoritas [16]. Oleh karena itu, perlu dilakukan penyeimbangan data. Teknik yang dilakukan adalah teknik *Synthetic Minority Oversampling Technique* (SMOTE).

Teknik SMOTE merupakan teknik yang paling terkenal dalam mengatasi data tidak seimbang. Teknik ini mirip dengan teknik *oversampling* yaitu menduplikasi data dari kelas minoritas sehingga jumlahnya sama dengan jumlah data kelas mayoritas. Namun SMOTE tidak hanya menduplikasi data yang sama, akan tetapi SMOTE akan membuat sampel baru yang menyerupai data asli dari kelas minoritas sehingga kelas minoritas menjadi jauh lebih beragam [16]–[21]. Setelah proses penyeimbangan, proporsi variabel tujuan di data *train* menjadi seimbang yaitu 49,8 : 50,2 (*NO : YES*) seperti pada Gambar 10.



Gambar 9. Proporsi Data Train Sebelum Proses Penyeimbangan



Gambar 10. Proporsi Data Train Setelah Proses Penyeimbangan

2.4 *Modelling*

Penelitian ini akan dilakukan penerapan data *mining* dalam memprediksi terjadinya penyakit kanker paru. Metode data *mining* yang digunakan adalah *Random Forest* dan *Naïve Bayes*.

2.4.1 Random Forest

Random Forest merupakan algoritma berbasis *ensemble* yang dibangun berdasarkan *Decision Tree*. Algoritma berbasis *ensemble* adalah gabungan dari beberapa teknik pembelajaran mesin (*machine learning*) yang digabungkan menjadi satu model prediktif. *Random Forest* bekerja dengan cara membuat banyak *Decision Tree*. Kemudian dari seluruh hasil prediksi yang dibuat, dilakukan *majority voting* untuk menentukan hasil prediksi akhir. Hal ini secara langsung dapat mengatasi masalah ketika melakukan klasifikasi menggunakan satu pohon keputusan saja sering kali tidak optimal [22]–[24].

Pembentukan *Decision Tree* pada algoritma *Random Forest* sama dengan proses pada *Classification and Regression Tree* (CART), hanya saja pada *Random Forest* tidak dilakukan *pruning* (pemangkasan). *Indeks Gini* digunakan untuk memilih fitur di setiap simpul internal dari *Decision Tree.* Nilai *Indeks Gini* dapat dihitung menggunakan persamaan (1).

$$Gini(S_i) = 1 - \sum_{i=0}^{c-1} p_i^2 \quad (1)$$

dengan p_i merupakan frekuensi relative kelas C_i di dalam data set. C_i merupakan kelas untuk $i = 1, \dots, c-1$, dan c adalah jumlah kelas yang telah ditentukan.

Kualitas *split* pada fitur k ke dalam subset S_i merupakan jumlah sampel milik kelas C_i, kemudian dihitung sebagai jumlah pertimbangan indikasi Gini dari subset yang dihasilkan. Data dapat dihitung dengan rumus persamaan (2).

$$Gini_{split} = \sum_{i=0}^{k-1} \left(\frac{n_i}{n}\right) Gini(S_i) \quad (2)$$

dengan n_i merupakan jumlah sampel dalam subset S_i setelah di split dan n merupakan jumlah sampel di node yang diberikan.

Misalkan $\{h(x, \Theta_k), k = 1, \dots\}$ dengan $\{\Theta_k\}$ merupakan vector random yang independent identically distributed (iid) dan tiap pohon memilih kelas yang paling banyak dari rata-rata (majority vote). Fungsi margin untuk Random Forest dihitung dengan rumus persamaan (3).

$$mr(X,Y) = P_\Theta(h(X,\Theta) = Y) - \max_{j \neq Y} P_\Theta(h(X,\Theta) = j) \quad (3)$$

dan kekuatan himpunan pengklasifikasi $\{h(X, \Theta)\}$ adalah seperti pada persamaan (4).

$$s = E_{X,Y} mr(X,Y) \quad (4)$$

dengan asumsi $s \geq 0$, ketidaksamaan *Chebychev* serta penurunan variansi mr dari fungsi margin untuk metode *Random Forest*, akan didapatkan persamaan batas atas kesalahan generalisasi seperti persamaan (5).

.

$$PE \leq \frac{\bar{\rho}(1-s^2)}{s^2} \quad (5)$$

dengan $\bar{\rho}$ adalah nilai rata-rata korelasi, dengan persamaan (6)

$$\bar{\rho} = \frac{E_{\Theta,\Theta'}(\rho(\Theta,\Theta')sd(\Theta)sd(\Theta'))}{E_{\Theta,\Theta'}(sd(\Theta)sd(\Theta'))} \quad (6)$$

Kelebihan dari *Random Forest* antara lain menghasilkan *error* yang kecil, memberikan akurasi yang baik dalam klasifikasi, dan efektif untuk mengatasi data yang tidak lengkap [12]. Pemodelan dengan *Random Forest* dapat lebih dioptimalkan dengan penerapan teknik evaluasi model yang disebut *K-fold Cross Validation*. Teknik ini membagi data menjadi *k* bagian. Setiap bagian akan menjadi data test secara bergantian, sehingga semua data memiliki kesempatan menjadi data test maupun data train. Tidak seperti Cross Validation yang hanya membagi data menjadi data train dan data test secara tetap [25]–[27].

2.4.2 Naïve Bayes

Algoritma *Naïve Bayes* dapat dijadikan sebagai pengklasifikasi probabilistik sederhana yang akan memperkirakan himpunan peluang dengan memperhitungkan kemunculan serta kombinasi nilai dalam himpunan data tertentu. Algoritma menggunakan teorema *Bayes* yang menganggap bahwa semua atribut bersifat berdiri sendiri atau tidak berkaitan satu sama lain berdasarkan nilai variabel kelas. Secara matematis, algoritma ini dapat dituliskan berdasarkan persamaan (7).

$$P(H|E) = \frac{P(E|H)\times P(H)}{P(E)} \quad (7)$$

Dimana $P(H|E)$ adalah kemungkinan atau peluang hipotesis berdasarkan kondisi (*posterior probability*), $P(E|H)$ adalah peluang parameter E berdasarkan kondisi pada hipotesis H, kemudian $P(H)$ adalah peluang hipotesis H (*prior probability*) dan $P(E)$ adalah peluang parameter E (*prior probability*).

Kelebihan algoritma *Naïve Bayes* adalah tidak memerlukan data *train* yang banyak, efisien dari segi ruang dan waktu karena algoritma yang sederhana, serta dapat menangani atribut yang tidak relevan dengan baik [8].

2.4.3 Confusion Matriks

Confusion Matriks adalah tabel yang digunakan untuk melihat akurasi serta performa algoritma yang dihasilkan dari klasifikasi, baik dibuat untuk mengklasifikasi maupun memprediksi atribut dari data *test*. Metode ini dikembangkan sebagai penilaian algoritma *machine learning* yang diterapkan dalam menyelesaikan masalah klasifikasi. Dalam *Confusion Matriks* terdapat *False Negative* (FN), *False Positive* (FP), *True Negative* (TN), dan *True Positive* (TP). Asumsi dalam *Confusion Matriks* diperllihatkan dalam Tabel 3.

Tabel 3. Asumsi dalam *Confusion Matriks*

Kelas Prediksi	Kelas Aktual	
	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

TP adalah kondisi dimana baik prediksi maupun nilai aktualnya benar; FN adalah kasus dimana nilai prediksi tidak benar tetapi nilai aktualnya benar; FP adalah kasus dimana nilai prediksi benar tapi nilai aktualnya tidak benar. Terdapat beberapa macam ukuran diantaranya akurasi, *recall*, dan presisi. Persamaan untuk menghitung nilai akurasi, *recall*, dan presisi ditunjukkan dalam Tabel 4.

Tabel 4. Rumus Evaluasi Performa Model

Matriks Performa	Rumus
Akurasi	$\frac{TP+TN}{TP+TN+FP+FN}\times 100\%$
Recall	$\frac{TP}{TP+FN}\times 100\%$
Presisi	$\frac{TP}{TP+FP}\times 100\%$

Akurasi merupakan rasio prediksi benar dengan keseluruhan data. *Recall* merupakan rasio prediksi positif benar dibandingkan dengan keseluruhan data aktual positif. Presisi merupakan rasio prediksi positif benar dibandingkan dengan seluruh data aktual positif. Ukuran kinerja lain yang bisa digunakan adalah AUC. AUC merepresentasikan derajat atau ukuran keterpisahan. Semakin tinggi nilai AUC semakin baik model memisahkan kelas target [7].

3 Hasil dan Pembahasan

Pemodelan dan analisis menggunakan RStudio. Model *Random Forest* yang dihasilkan melalui *K-fold Cross Validation* menggunakan data *train*. Dimana ringkasan pada model diperlihatkan pada Gambar 11. Berdasarkan *summary* model, diketahui bahwa jumlah variabel optimal yang dipertimbangkan untuk dipecah pada setiap simpul pohon adalah 2. Model optimal dari *Random Forest* dapat ditunjukkan pada Gambar 12. Evaluasi model *Random Forest* diperoleh hasil seperti Gambar 13.

```
Random Forest

246 samples
 15 predictor
  2 classes: 'NO', 'YES'

No pre-processing
Resampling: Cross-Validated (2 fold, repeated 1 times)
Summary of sample sizes: 124, 122
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.8740085  0.3254170
   8    0.8537811  0.2991952
  15    0.8578794  0.3303983

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

Gambar 11. Hasil Model Random Forest

```
Call:
 randomForest(x = x, y = y, mtry = param$mtry)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 10.57%
Confusion matrix:
    NO YES class.error
NO  14  19  0.57575758
YES  7 206  0.03286385
```

Gambar 12. Model Optinal Random Forest

```
# A tibble: 1 × 5
  Accuracy Recall Specificity Precision   AUC
     <dbl>  <dbl>       <dbl>     <dbl> <dbl>
1    0.984      1         0.8     0.983     1
```

Gambar 13. Hasil *Confusion Matriks* Randon Forest

```
# A tibble: 1 × 5
  Accuracy Recall Specificity Precision   AUC
     <dbl>  <dbl>       <dbl>     <dbl> <dbl>
1    0.952  0.947           1         1 0.989
```

Gambar 14. *Confusion Matriks* Algoritma*Naïve Bayes*

Table 5. Perbandingan Performa Algoritma *Naive Bayes* dan *Random Forest*

	Accuracy	Recall	Specificity	Precision	AUC
Naive_Bayes	0,951613	0,947368	1	1	0,98947
Random_Forest	0,983871	1	0,8	0,982759	1

```
        Friedman rank sum test

data:  Pengujian_beda$Nilai, Pengujian_beda$Ukuran and Pengujian_beda$Algoritma
Friedman chi-squared = 0.84211, df = 4, p-value = 0.9327
```

Gambar 15. Uji Beda Signifikansi

Berdasarkan hasil pada Gambar 13, diperoleh informasi bahwa kemampuan model dalam memprediksi variabel tujuan sebesar 98,4%. Sedangkan dari keseluruhan data aktual pada responden dengan kanker paru (Kelas *YES*), model mampu menebak dengan benar sebesar 100%. Jika dibandingkan dengan keseluruhan data aktual pasien yang tidak mengidap kanker paru (Kelas *NO*), model mampu memprediksi dengan benar sebanyak 80%. Sedangkan dari keseluruhan hasil prediksi yang mampu ditebak oleh model, model dari algoritma *Random Forest* mampu menebak dengan benar kelas positif (*YES*) sebesar 98,3%. Sedangkan hasil RStudio, diperoleh nilai AUC sebesar 1. Hasil model prediksi menggunakan algoritma *Naïve Bayes* menghasilkan *confusion matriks* seperti yang ditampilkan dalam Gambar 14.

Berdasarkan *output* pada Gambar 14, diperoleh informasi bahwa kemampuan model dalam memprediksi variabel tujuan sebesar 95,2%. Sedangkan dari keseluruhan data aktual pada responden dengan kanker paru (Kelas *YES*), model mampu menebak dengan benar sebesar 94,7%. Jika dibandingkan dengan keseluruhan data aktual pasien yang tidak mengidap kanker paru (Kelas *NO*), model mampu memprediksi dengan benar sebesar 100%. Sedangkan dari keseluruhan hasil prediksi yang mampu ditebak oleh model, model *Naïve Bayes* mampu menebak dengan benar kelas *YES* sebesar 100%. Sedangkan hasil dari RStudio, diperoleh nilai AUC sebesar 0,9894737.

Ringkasan perbandingan performa Algoritma *Naïve Bayes* dan *Random Forest* ditampilkan dalam Tabel 5. Klasifikasi ini mengharapkan dapat sebanyak mungkin memprediksi dengan benar responden yang terkena kanker paru (kelas *YES*), sehingga matriks yang digunakan adalah *Recall*. Nilai *Recall* tertinggi dihasilkan saat menggunakan metode *Random Forest*. Begitupun dengan nilai *Accuracy* dan *AUC*. Namun memiliki nilai *Specificity* dan *Precision* yang lebih kecil dibandingkan *Naïve Bayes*.

Pengujian beda signifikansi hasil dari kedua metode ditunjukkan pada Gambar 15. Uji beda signifikansi dilakukan dengan dengan Uji Friedman. Uji Freadman merupakan uji statistik nonparametrik untuk data berpasangan tanpa asumsi normalitas dan varians populasi yang tidak diketahui. Berdasarkan hasil yang ditunjukkan pada Gambar 15, nilai *p-value* sebesar 0,9327 dimana lebih besar dari tingkat signifikasi 0,05. Hal ini berarti kedua algoritma tersebut memberikan hasil yang tidak berbeda secara signifikan. Artinya baik algoritma Naïve Bayes maupun Random Forest memiliki akurasi yang baik dalam memprediksi kanker paru.

4 Kesimpulan

Penelitian ini menggunakan dataset *Survey Lung Cancer*. *P*engukuran performa model dari algoritma *Random Forest* dan *Naïve Bayes* menujukkan bahwa secara umum nilai dari matriks *Accuracy*, *Recall*, *Spesificity*, *Precision* dan AUC pada kedua algoritma sangat tinggi. Algoritma Random Forest menghasilkan nilai *Accuracy* yang lebih tinggi daripada algoritma *Naïve Bayes*, yaitu sebesar 98,4%. Bahkan algoritma ini menghasilkan *Recall* 100% untuk kelas positif dan 80% untuk kelas negatif serta memberikan prediksi 100% benar terlihat dari nilai AUC sebesar 1. Meskipun uji statistik dengan tingkat signifikansi 5% menunjukkan hasil dari kedua algoritma tersebut tidak berbeda secara signifikan.

Daftar Pustaka

[1] J. Braithwaite, "What Is Cancer?," in *The Lancet*, vol. 131, no. 3383, 1888, pp. 1287–1289. doi: 10.1016/S0140-6736(02)16666-9.

[2] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[3] "The Global Cancer Observatory," 2020. https://gco.iarc.fr/today/data/factsheets/cancers/15-Lung-fact-sheet.pdf (accessed Jan. 13, 2023).

[4] S. Sugiharto, R. A. Putri, S. Simanjuntak, and O. Larissa, "Kanker Paru, Faktor Resiko Dan Pencegahannya," in *Seminar Nasional Hasil Penelitian dan Pengabdian Kepada Masyarakat (SENAPENMAS)*, 2021.

[5] S. R. Rahmadania, "Fakta-fakta Hari Kanker Sedunia 2022, Dirayakan Setiap Tanggal 4 Februari," *detikHealth*, 2022. https://health.detik.com/berita-detikhealth/d-5925795/fakta-fakta-hari-kanker-sedunia-2022-dirayakan-setiap-tanggal-4-februari (accessed Jan. 13, 2023).

[6] I. W. Gamadarenda and I. Waspada, "Implementasi Data Mining Untuk Deteksi Penyakit Ginjal Kronis (Pgk) Menggunakan K-Nearest Neighbor (Knn) Dengan Backward Data Mining Implementation For Detection Of Chronic Kidney (Ckd) Using K-Nearest Neighbor (Knn) With Backward Elimination," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 7, no. 2, pp. 417–426, 2020, doi: 10.25126/jtiik.202071896.

[7] I. Amal, "Klasifikasi Menggunakan Naive Bayes, Decision Tree, dan Random Forest," 2021. https://rstudio-pubs-

static.s3.amazonaws.com/717459_5136236cf5064b8d973e4d8c1b863943.html#5_Cross_validation (accessed Jan. 13, 2023).

[8] A. I. Kusumarini, P. A. Hogantara, M. Fadhlurohman, and S. Kom. , M. K. Nurul Chamidah, "Perbandingan Algoritma Random Forest, Naive Bayes, Dan Decision Tree Dengan Oversampling Untuk Klasifikasi Bakteri E.Coli," *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya*, vol. 2, no. 1, pp. 792–799, 2021.

[9] B. Bawono and R. Wasono, "Perbandingan Metode Random Forest dan Naive Bayes," *Jurnal Sains dan Sistem Informasi*, vol. 3, no. 7, pp. 343–348, 2019, [Online]. Available: http://prosiding.unimus.ac.id

[10] G. M. Momole and E. Mailoa, "Perbandingan Naïve Bayes Dan Random Forest Dalam Klasifikasi Bahasa Daerah," vol. 9, no. 2, pp. 855–863, 2022.

[11] R. Leonardo, J. Pratama, and Chrisnatalis, "Perbandingan Metode Random Forest Dan Naïve Bayes Dalam Prediksi Keberhasilan Klien Telemarketing," vol. 3, pp. 455–459, 2020.

[12] S. Amaliah and M. Nusrang, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi Di Kedai Kopi Konijiwa Bantaeng," *Variansi: Journal of Statistic and Its Application on Teaching and Research*, vol. 4, no. 2, pp. 121–127, 2022, doi: 10.35580/variansiunm31.

[13] D. H. Depari *et al.*, "Perbandingan Model Decision Tree , Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," vol. 4221, pp. 239–248, 2022.

[14] P. Sejati et al., "Studi Komparasi Naive Bayes , K-Nearest Neighbor , Dan Random Forest Untuk Prediksi Calon Mahasiswa Yang Diterima Atau Comparative Study Of Naive Bayes , K-Nearest Neighbor , And Random Forest For The Prediction Of Prospective Students," Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), vol. 9, no. 7, pp. 1341–1348, 2022, doi: 10.25126/jtiik.202296737.

[15] Ramadani and B. H. Hayadi, "Perbandingan Metode Naive Bayes Dan Random Forest Untuk Menentukan Prestasi Belajar Siswa Pada Jurusan RPL (Studi Kasus SMK Swasta Siti Banun Sigambal)," Journal Computer Science and Information Technology(JCoInT) Program Studi Teknologi Informasi, no. 2, p. 2022, 2022, [Online]. Available: http://jurnal.ulb.ac.id/index.php/JCoInT/index

[16] D. Dablain, B. Krawczyk, and N. v. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans Neural Netw Learn Syst*, pp. 1–14, 2022, doi: 10.1109/TNNLS.2021.3136503.

[17] V. Nugraha, "Menghadapi Imblanced Target Variable dengan SMOTE," *RPubs*, 2021. https://rpubs.com/VicNP/UBL-SmoteClassif (accessed Jan. 16, 2023).

[18] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, "Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model," *IEEE Access*, vol. 9, pp. 78621–78634, 2021, doi: 10.1109/ACCESS.2021.3083638.

[19] D. Mualfah, W. Fadila, and R. Firdaus, "Teknik SMOTE untuk Mengatasi Imbalance Data pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 2, pp. 107–113, 2022, doi: 10.37859/coscitech.v3i2.3912.

[20] A. N. Kasanah, M. Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 2, pp. 196–201, 2019, doi: 10.29207/resti.v3i2.945.

[21] A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[22] F. Putri, Sanni Ucha; Irawan, Eka; Rizky, "Implementasi Data Mining Untuk Prediksi Penyakit Diabetes," *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, vol. 2, no. 1, pp. 39–46.

[23] D. Alita and A. Rahman, "Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier," *Jurnal Komputasi*, vol. 8, no. 2, pp. 50–58, 2020, doi: 10.23960/komputasi.v8i2.2615.

[24] Y. Yuliani, "Algoritma Random Forest Untuk Prediksi Kelangsungan Hidup Pasien Gagal Jantung Menggunakan Seleksi Fitur Bestfirst," vol. 5, no. 2, pp. 298–306, 2022.

[25] K. Pal and B. v. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 83–87. doi: 10.1109/ICCMC48092.2020.ICCMC-00016.

[26] K. Phinzi, D. Abriha, and S. Szabó, "Classification efficacy using k-fold cross-validation and bootstrapping resampling techniques on the example of mapping complex gully systems," *Remote Sens (Basel)*, vol. 13, no. 15, 2021, doi: 10.3390/rs13152980.

[27] M. Madanan, A. Venugopal, and N. C. Velayudhan, "Applying an optimal feature ranking and selection algorithm and random forest classifier algorithm along with k-fold cross validation for classification of blood cancer cells," *European Journal of Molecular and Clinical Medicine*, vol. 7, no. 11, pp. 774–789, 2020, [Online]. Available: https://www.embase.com/search/results?subaction=viewrecord&id=L2010514747&from=export