PREDIKSI KESEHATAN PARU-PARU PADA PASIEN KANKER PARU-PARU MENGGUNAKAN METODE RANDOM FOREST

Anggelina Karolina Teti[1], Febriani Astuti[2*]
[1,2] Universitas AKPRIND Indonesia, [*]Penulis Koresponden
e-mail: [1]karolinateti24@gmail.com, [2]febriani@akprind.ac.id

ABSTRACT

Lung health is an important aspect in the management of lung cancer patients. Healthy lung conditions can influence the prognosis and quality of life of patients. This study aims to classify lung health predictions in lung cancer patients using the Random Forest method. The analysis results indicate that the optimal model is obtained with parameters m=2, k=50, and k=100, using an 80:20 data proportion. The variable that most influences the prediction results is Hemoptysis. Model evaluation using the confusion matrix yields an accuracy rate of 95%. All independent variables significantly affect lung health predictions, including age, gender, air pollution, alcohol use, obesity, smoking, passive smoking, and hemoptysis.

Keywords: *Confusion Matrix, Lung Cancer Prediction, Random Forest, Variable Importance.*

INTISARI

Kesehatan paru-paru merupakan salah satu aspek penting dalam penanganan pasien kanker paru-paru. Kondisi paru-paru yang sehat dapat mempengaruhi prognosis dan kualitas hidup pasien. Penelitian ini bertujuan untuk melakukan prediksi klasifikasi kesehatan paru-paru pada pasien kanker paru-paru menggunakan metode Random Forest. Hasil analisis menunjukkan bahwa model optimal diperoleh dengan menggunakan parameter $\mathrm{m} = 2, \mathrm{k} = 50,$ dan $\mathrm{k} = 100$ dengan proporsi data 80:20. Variabel yang paling mempengaruhi hasil prediksi adalah Batuk Darah. Evaluasi model menggunakan *confusion matrix* menghasilkan tingkat akurasi sebesar 95%. Semua variabel bebas memiliki pengaruh signifikan terhadap prediksi kesehatan paru-paru antara lain usia, jenis kelamin, polusi udara, penggunaan alkohol, obesitas, merokok, perokok pasif, dan batuk darah.

Kata kunci: *Confusion Matrix*, Prediksi Kanker Paru-Paru, Random Forest, *Variabel Importance.*

1. PENDAHULUAN

Kanker paru-paru merupakan penyakit dengan tingkat kematian yang tinggi di seluruh dunia. Pada tahun 2020, jumlah kasus kanker paru-paru mencapai angka yang mencengangkan, menyebabkan ribuan kematian setiap tahunnya. Penyakit ini terjadi ketika zat karsinogen memicu pertumbuhan sel-sel yang tidak terkendali di paru-paru. Upaya pencegahan dan deteksi dini menjadi fokus utama untuk mengurangi angka kematian akibat kanker paru-paru. Gejala kanker paru-paru sering kali tidak spesifik, sehingga banyak penderita menganggap gejala awalnya sebagai gangguan pernapasan umum atau penyakit ringan lainnya. Hal ini dapat menyebabkan keterlambatan dalam diagnosis dan pengobatan yang tepat. Selain itu, rendahnya kesadaran di antara petugas layanan kesehatan untuk melakukan pemeriksaan lanjutan terhadap pasien dengan gejala mencurigakan memperburuk situasi ini. Kesulitan dalam mendeteksi kanker paru-paru mengakibatkan penyakit ini menjadi serius dan meningkatkan angka kematian yang tinggi. Kanker paru-paru umumnya terjadi pada individu dengan kebiasaan merokok dan pola hidup tidak sehat, menjadikannya salah satu jenis kanker yang paling umum di dunia, serta menempati posisi ketiga sebagai jenis kanker tersering di Indonesia. Sebagian besar kematian akibat kanker paru-paru terkait dengan kebiasaan merokok dan paparan asap rokok. Merokok merupakan salah satu faktor utama yang merusak paru-paru, dan kebiasaan ini sangat sulit dihentikan. Gangguan pada paru-paru dapat mengurangi efisiensi dan fungsi paru-paru dalam menyerap oksigen dari udara. Gejala dari penyakit kanker paru-paru menurut Mayo Clinic meliputi kelelahan, adanya benjolan, perubahan berat badan, perubahan kulit seperti menguning, batuk terus-menerus, nyeri otot, suara serak, dan kesulitan menelan. Kanker paru-paru sering terdiagnosis pada stadium lanjut, di mana sekitar 60-85% pasien tidak mengetahui penyakitnya. Hal ini disebabkan oleh anggapan bahwa batuk dan sesak yang diderita adalah hal biasa, sehingga mereka tidak memeriksakan diri ke layanan kesehatan. Kebanyakan petugas kesehatan hanya mengatasi gejala tanpa melakukan pemeriksaan lanjutan, sehingga diperlukan

metode yang dapat secara akurat mendeteksi kanker paru-paru melalui prediksi dari berbagai faktor dan gejala yang timbul.

Data mining dapat diimplementasikan di banyak bidang, termasuk kesehatan. Penggunaan data mining untuk mendeteksi penyakit, khususnya kanker, semakin meningkat. Penelitian tentang penerapan data mining dalam analisis kanker paru pernah dilakukan oleh Sari et al. (2023). Data mining adalah proses meringkas pengetahuan menggunakan algoritma untuk mendeteksi pola, kecenderungan dalam data, dan hubungan yang tidak terlihat sebelumnya. Zai (2022) melakukan implementasi data mining dalam pengolahan data. Salah satu teknik yang sering digunakan dalam data mining adalah klasifikasi. Klasifikasi merupakan analisis data di mana terdapat model atau classifier yang digunakan untuk memprediksi class labels (Han et al., 2012). Dalam klasifikasi, terdapat dua proses yaitu training dan testing. Banyak metode yang dapat digunakan untuk klasifikasi, seperti Decision Tree, Random Forest, Naïve Bayes, dan K-Nearest Neighbor. Random Forest adalah salah satu metode populer yang merupakan pengembangan dari CART (Classification and Regression Trees). Metode ini menggunakan teknik bootstrap selection aggregating (bagging), di mana beberapa pohon keputusan dibangun secara independen berdasarkan sampel acak dari data pelatihan. Setiap pohon memberikan prediksi kelas, dan hasil akhirnya adalah agregat dari prediksi tersebut, sering kali menggunakan mode atau rata-rata.

Keunggulan data mining, seperti kemampuan untuk mengidentifikasi pola kompleks dan hubungan antarvariabel yang tidak terlihat secara langsung, sangat penting dalam proses ini. Dengan menerapkan teknik-teknik seperti Random Forest, peneliti dapat menghasilkan model prediktif yang dapat membantu dalam deteksi dini dan penanganan kanker paru-paru berdasarkan data klinis dan patologis yang tersedia. Hal ini menunjukkan bahwa penggunaan data mining dalam klasifikasi sangat bermanfaat untuk meningkatkan akurasi dalam memprediksi kondisi medis yang kompleks seperti kanker paru-paru. Dalam riset di bidang kesehatan, terdapat beberapa model klasifikasi yang sering digunakan.

Fadlilah et al. (2019) melakukan penelitian dengan klasifikasi objek fungsi kognitif pasien stroke menggunakan metode Random Forest Decision Trees. Hasil penelitian tersebut menunjukkan rata-rata akurasi sebesar 53,094% dengan jumlah fitur optimal 13 dan jumlah optimal trees 100, menggunakan K-Fold Cross Validation. Selanjutnya penelitian oleh Yang dan Chen (2015) dengan memprediksi stadium kanker paru-paru menggunakan informasi klinis dan patologi seperti rontgen, CT Scan, dan biopsi, dengan hasil rata-rata akurasi sebesar 81,97%. Selain itu, penelitian oleh Wulandari dan Perdana (2022) menggunakan algoritma Naïve Bayes untuk memprediksi apakah seseorang menderita kanker paru (yes/no). Hasilnya menunjukkan tingkat *recall* untuk kelas positif sebesar 98,77% dan kelas negatif sebesar 66,67%, dengan tingkat akurasi keseluruhan sebesar 94,62%. Selain itu Speiser et al. (2019) menggunakan Random Forest untuk melakukan model klasifikasi dan prediksi. Random Forest secara lengkap juga dibahas oleh Rigatti (2017) dalam bukunya berjudul "Random Forest". Penelitian tentang klasifikasi dengan Random Forest juga dibahas oleh Paul et al. (2018). Berdasarkan penelitian-penelitain terdahulu yang sudah dipaparkan semakin memperkuat peneliti untuk melakukan penelitian tentang "Prediksi Kesehatan Paru-Paru pada Pasien Kanker Paru-Paru Menggunakan Metode Random Forest".

2. METODE PENELITIAN

Penelitian ini menggunakan metode deskriptif kuantitatif yang bertujuan untuk memberikan gambaran objektif mengenai keadaan tertentu dengan memanfaatkan data numerik. Dalam penelitian ini, data sekunder digunakan sebagai sumber informasi, yang diambil dari Kaggle, sebuah platform yang menyediakan berbagai dataset terkait machine learning dan data science (https://www.kaggle.com/).
Variabel bebas yang dianalisis mencakup usia, jenis kelamin, polusi udara, penggunaan alkohol, obesitas, merokok, perokok pasif, dan batuk darah, dengan fokus utama pada data penyakit paru-paru. Adapun variabel yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Variabel Penelitian

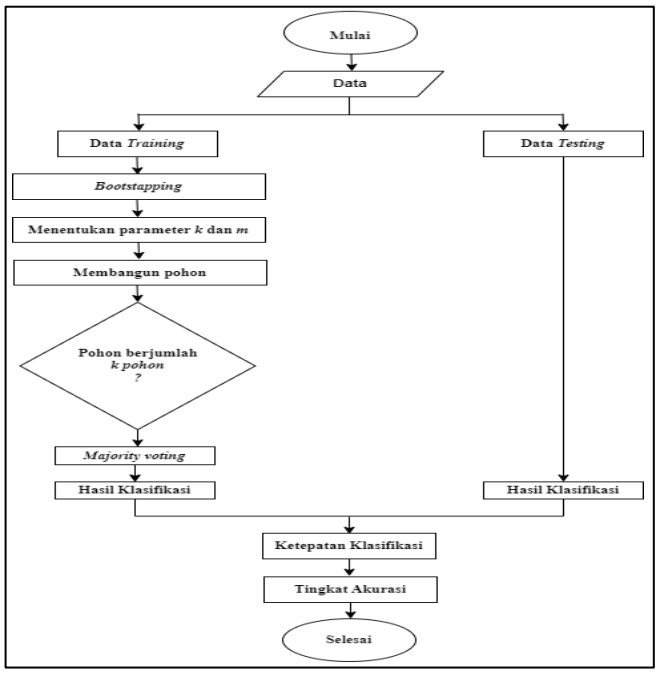
Variabel	Jenis	Skala Data
Usia (Tahun)	Independen	Interval
Jenis Kelamin (Kategori: Pria/Wanita)	Independen	Nominal
Polusi Udara	Independen	Nominal

(Kategori: 1: Sangat rendah, 2: Rendah, 3: Agak rendah, 4: Sedang, 5: Agak tinggi, 6: Tinggi, 7: Sangat tinggi)		
Penggunaan Alkohol (Kategori: 1: Tidak pernah, 2: Jarang sekali, 3: Jarang, 4: Kadang-kadang, 5: Cukup sering, 6: Sering, 7: Sangat sering, 8: Selalu)	Independen	Nominal
Obesitas (Kategori: 1: Tidak obesitas, 2: Sangat sedikit obesitas, 3: Sedikit obesitas, 4: Sedang, 5: Cukup obesitas, 6: Sangat Obesitas, 7: Obesitas tinggi)	Independen	Nominal
Merokok (Kategori: 1: Tidak pernah, 2: Sangat jarang merokok, 3: Jarang merokok, 4: Kadang-kadang merokok, 5: Cukup sering merokok, 6: Sering merokok, 7: Sangat sering merokok, 8: Selalu merokok)	Independen	Nominal
Perokok Pasif (Kategori: 1: Tidak pernah terpapar, 2: Sangat jarang terpapar, 3: Jarang terpapar, 4: Kadang-kadang terpapar, 5: Cukup sering terpapar, 6: Sering terpapar, 7: Sangat sering terpapar, 8: Selalu terpapar)	Independen	Nominal
Batuk Darah (Kategori: 1: Tidak pernah, 2: Sangat jarang, 3: Jarang, 4: Kadang-kadang, 5: Cukup sering, 6: Sering, 7: Sangat sering, 8: Setiap hari, 9: Kronis)	Independen	Nominal
Levels (Katgori: Low, High, Medium)	Dependen	Ordinal

Tahapan-tahapan analisis data dalam penelitian ini antara lain:

a. Menentukan banyaknya pohon (*k*).
b. Membuat suatu *bootstrap* sample dari gugus *data training* dengan pengembalian.
c. Membangun pohon klasifikasi dengan perhitungan *entropy* dan *information gain* menggunakan gugus *data training* baru yang terbentuk dari proses bootstrap.
d. Mengulangi langkah (b) dan (c) sebanyak *k* kali sehingga diperoleh k buah pohon acak. Setiap pohon klasifikasi akan menghasilkan satu keputusan sehingga didapatkan *k* buah keputusan. Penentuan klasifikasi didasarkan pada keputusan terbanyak (*majority vote*).
e. Interpretasi hasil.

Tahapan-tahapan penelitian ditampilkan dalam Gambar 1 berikut.



Gambar 1. Tahapan-Tahapan Penelitian

3. HASIL DAN PEMBAHASAN

3.1 Data *Training* dan Data *Testing*

Klasifikasi kesehatan paru-paru pada metode Random Forest menggunakan 8 variabel independen yang diduga berpengaruh terhadap prediksi kesehatan paru-paru. Diantaranya usia, jenis kelamin, polusi udara, penggunaan alkohol, obesitas, merokok, perokok pasif, batuk darah. Langkah awal pengklasifikasian menggunakan Random Forest adalah pembagian data training dan data testing.

Dalam melakukan klasifikasi, data perlu dibagi menjadi dua yaitu data training dan data testing. Model klasifikasi dibangun berdasarkan data training dan kinerjanya diukur berdasarkan data testing. Klasifikasi untuk memprediksi kesehatan paru-paru dilakukan dengan menggunakan metode Random Forest.

Pembagian data training dan data testing pada metode klasifikasi ini dilakukan dengan beberapa percobaan. Percobaan ini digunakan untuk memperoleh proporsi terbaik yaitu proporsi dengan akurasi klasifikasi tertinggi. Percobaan dilakukan dengan bantuan software R Studio. Percobaan proporsi yang digunakan adalah proporsi 80:20, yang mana proporsi ini dipilih karena memberikan error klasifikasi yang kecil (Fitriyaningsih dan Basani, 2019). Berdasarkan pembagian proporsinya maka perbandingan data training dan data testing adalah sebagai berikut:

Tabel 2. Pembagian Data Training dan Data Testing Proporsi 80:20

Keterangan	Data Training	Data Testing
Jumlah	80	20
Presentase	80%	20%

Berdasarkan Tabel 2 di atas, banyak data yang digunakan adalah 100 data. Diberikan proporsi 80:20, sehingga banyak data training adalah 80 data dan data testing adalah 20 data.

3.2 *Bootstrapping*

Langkah pengklasifikasian selanjutnya dari Random Forest adalah bootstrapping. Bootstrapping adalah pengambilan sampel Z dan N gugus data training secara acak dengan pengembalian. Kemudian pohon klasifikasi dibangun menggunakan gugus data training baru yang terbentuk dari proses bootstrapping. Pembangunan pohon klasifikasi dilakukan hingga diperoleh k buah pohon acak. Setiap pohon klasifikasi

akan menghasilkan satu keputusan sehingga didapatkan k buah keputusan. Penentuan klasifikasi didasarkan pada keputusan terbanyak (*majority vote*).
Menurut (Hastie, Tibshirani, & Freidman, 2008) jumlah variabel yang diambil secara acak dapat ditentukan melalui $\sqrt{p}$. Nilai *m* juga dapat diperoleh dari $\frac{1}{2}\sqrt{p}$ dan $2\sqrt{p}$ (Breiman, 2001). Berikut perhitungan jumlah variabel penjelas *m*.

$$m_2 = \frac{1}{2}\sqrt{p} = \frac{1}{2}\sqrt{8} \approx 1 \text{ ……..(1)}$$

$$m_1 = \sqrt{p} = \sqrt{8} \approx 2 \text{ ……..(2)}$$

$$m_3 = 2\sqrt{p} = 2\sqrt{8} \approx 5 \text{ ……..(3)}$$

Sehingga diperoleh m yang digunakan adalah 1, 2, dan 5. Pada penelitian ini, penulis melakukan percobaan dengan proporsi data training dan data testing 80:20. Proporsi dilakukan 3 percobaan m yaitu 1, 2, dan 5 serta percobaan nilai k yang digunakan yaitu 25, 50,100.

3.3 Mencari Model Terbaik Metode Random Forest

Pemilihan model terbaik pada klasifikasi dengan metode Random Forest dapat dilakukan dengan membandingkan besar akurasi pada setiap percobaan. Percobaan dibangun menggunakan data training dan kemudian akurasi atau kinerjanya diukur berdasarkan data testing. Berikut akurasi yang diperoleh dari percobaan menggunakan proporsi 80:20.

Tabel 3. Akurasi Percobaan Metode Random Forest dengan Proporsi 80:20

m	*k*	Akurasi
1	25	0,95
	50	0,90
	100	0,85
2	25	0,90
	50	**0,95**
	100	**0,95**
3	25	0,90
	50	0,90
	100	0,95

Berdasarkan Tabel 3 didapatkan akurasi terbesar diperoleh oleh model dengan rata-rata terbesar yakni $m = 2$ dan $k = 50, k = 100$ dengan nilai sebesar 0,95. Oleh karena itu, penerapan metode *Random Forest* dalam prediksi Kesehatan Paru-Paru pada Pasien Kanker Paru-Paru akan menggunakan model dengan $m = 2$ dan $k = 50, k = 100$.

Berdasarkan klasifikasi prediksi Kesehatan Paru-Paru pada Pasien Kanker Paru-Paru dengan *m* = 2 dan *k* = 50 dan 100 dilakukan pengujian ketepatan klasifikasi dengan menggunakan *data testing*. Berikut hasil ketepatan klasifikasi *data testing* menggunakan *m* = 2 dan *k* = 50 dan 100.

3.4 Ketepatan Klasifikasi Metode Random Forest

Berdasarkan klasifikasi prediksi kesehatan paru-paru pada pasien kanker paru-paru dengan *m* = 2 dan *k* = 50 dan 100 dilakukan pengujian ketepatan klasifikasi dengan menggunakan data testing. Berikut hasil ketepatan klasifikasi data testing menggunakan *m* = 2 dan *k* = 50 dan 100.

Tabel 4. Ketepatan Klasifikasi Metode Random Forest

X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**Y**	**Prediksi_ RF50.2_ 100.2._ Test**
33	1	2	4	4	3	2	4	Low	Low
17	1	3	1	2	2	4	3	Medium	Medium

X1	X2	X3	X4	X5	X6	X7	X8	Y	Prediksi_ RF50.2_ 100.2._ Test
46	1	2	3	3	2	3	4	Medium	Medium
44	1	6	7	7	7	8	7	High	High
29	2	6	7	7	7	7	7	High	High
19	1	3	2	3	2	2	3	Medium	Medium
25	2	3	1	3	1	4	1	Low	Low
27	2	3	1	3	2	2	2	Low	Low
17	2	1	2	1	3	2	2	Low	Low
22	1	2	1	2	6	1	2	Low	Low
35	1	1	3	2	2	1	4	Low	Low
23	2	4	2	4	2	4	4	Low	Medium
38	2	5	2	2	2	5	3	Low	Low
47	2	2	3	1	2	1	5	Low	Low
44	2	2	3	2	7	6	2	Low	Low
33	2	1	6	7	3	4	7	Medium	Medium
33	1	3	2	3	2	2	3	Medium	Medium
52	2	1	2	3	3	2	4	Medium	Medium
44	1	6	7	7	7	8	7	High	High
28	1	6	7	7	7	8	7	High	High

Berdasarkan hasil klasifikasi *data testing* pada Tabel 4 diperoleh bahwa pada data ke-23 memiliki data aktual low, sedangkan hasil prediksinya adalah medium.

3.5 *Confusion Matrix* Metode Random Forest

Berdasarkan hasil perbandingan data aktual dengan hasil prediksi di atas, langkah selanjutnya adalah membuat tabel *confusion matrix*. Tabel *confusion matrix* adalah perbandingan dari data aktual dengan hasil prediksi. Berikut tabel *confusion matrix* metode *Random Forest*.

Tabel 5. Confusion Matrix Metode Random Forest ($m = 2, k = 50$, dan $k = 100$).

Data Testing	Prediksi		
	High	Low	Medium
High	4	0	0
Low	0	9	1
Medium	0	0	6

Berdasarkan Tabel 5 didapatkan bahwa:

a. Pada bagian *confusion matrix*, data yang dikategorikan *high* setelah diprediksi akan tetap *high* dengan nilai sebanyak 4, data yang dikategorikan *high* setelah diprediksi akan berubah menjadi *low* dengan nilai sebanyak 0, dan data yang dikategorikan *high* setelah diprediksi akan berubah menjadi *medium* dengan nilai sebanyak 0.
b. Sedangkan pada bagian *low*, data yang dikategorikan *low* setelah diprediksi akan berubah menjadi *high* dengan nilai sebanyak 0, data yang dikategorikan *low* setelah diprediksi akan tetap menjadi *low* dengan nilai sebanyak 9, dan data yang dikategorikan *low* setelah diprediksi akan berubah menjadi *medium* dengan nilai sebanyak 1.
c. Selanjutnya pada bagian *medium*, data yang dikategorikan *medium* setelah diprediksi akan berubah menjadi *high* dengan nilai sebanyak 0, data yang dikategorikan *medium* setelah diprediksi akan

berubah menjadi *low* dengan nilai sebanyak 0, dan data yang dikategorikan *medium* setelah diprediksi akan tetap menjadi *medium* dengan nilai sebanyak 6.

Berdasarkan Tabel 5. diketahui bahwa hasil prediksi benar untuk kategori *high* sebanyak 4, hasil prediksi benar untuk kategori *low* sebanyak 9, dan hasil prediksi benar untuk kategori *medium* sebanyak 6. Berikut perhitungan akurasinya:

$$Akurasi = \frac{\Sigma Data\ uji\ benar\ klasifikasi}{\Sigma\ Data\ uji} \times 100\% \quad \text{.....(4)}$$

$$Akurasi = \frac{19}{20} \times 100\% = 0{,}95 \times 100\% = 95\% \quad \text{.....(5)}$$

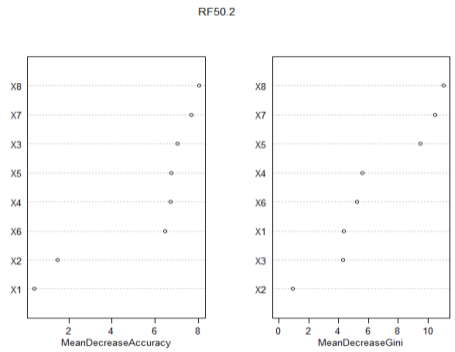
Berdasarkan perhitungan akurasi di atas, diperoleh akurasi klasifikasi sebesar 0,95 atau 95%, sehingga klasifikasi prediksi kesehatan paru-paru pada pasien kanker paru-paru dapat dikatakan sudah sangat baik, karena memiliki nilai akurasi 0,90 – 1,00.

3.6 Variable Importance

Klasifikasi prediksi kesehatan paru-paru pada pasien kanker paru-paru menghasilkan model terbaik pada proporsi 80:20 dengan *m* = 2 dan *k* = 50, dan 100. Berikut *variable importance* yang diperoleh.

	High	Low	Medium	MeanDecreaseAccuracy	MeanDecreaseGini
X1	0.36813424	-0.5686023	1.2947933	0.400208	4.3320012
X2	0.03179847	0.9300470	0.6555934	1.473756	0.9085409
X3	3.54350788	3.5199428	3.8066877	7.034223	4.2959198
X4	3.85520929	6.0280175	2.9838189	6.693983	5.6208595
X5	4.04225833	5.9184174	5.2113138	6.736822	9.5067633
X6	4.53413720	1.0725620	5.7489853	6.443800	5.2537589
X7	6.62003268	5.0475411	6.3422399	7.673189	10.4648707
X8	4.80840626	6.7679597	5.4047288	8.020803	11.0172858

Gambar 2. Gambar Output *Importance* untuk $m = 2$ dan $k = 50$ dari Software R



Gambar 3. *Variable Importance* untuk $m = 2$ dan $k = 50$

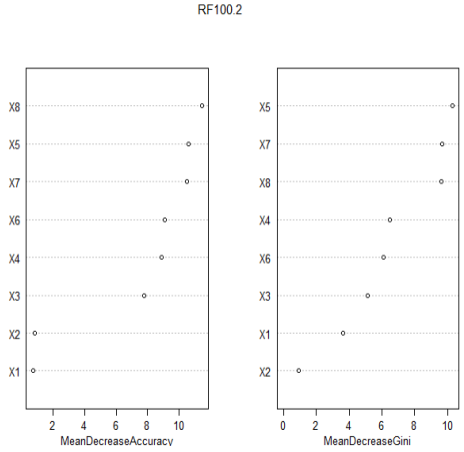
Gambar 2 dan Gambar 3 menunjukkan bahwa semakin besar angka *Mean Decrease Accuracy*, maka semakin besar juga peran variabel tersebut dalam mempengaruhi hasil analisis. Berdasarkan Gambar 2 dan 3, terlihat bahwa semua variabel berpengaruh terhadap hasil klasifikasi prediksi kesehatan paru-paru pada pasien kanker paru-paru.

Variabel yang berpengaruh secara berurutan yaitu X_8 (Batuk darah) dengan *mean decrease accuracy* sebesar 8,020803, X_7 (Perokok pasif) dengan *mean decrease accuracy* sebesar 7, 673189, X_3 (Polusi udara) dengan *mean decrease accuracy* sebesar 7,034223, X_5 (Obesitas) dengan *mean decrease accuracy* sebesar 6,736822, X_4 (Penggunaan alkohol) dengan *mean decrease accuracy* sebesar 6,693983, X_6 (Merokok) dengan *mean decrease accuracy* sebesar 6,443800, X_2 (Jenis kelamin) dengan *mean decrease accuracy* sebesar 1,473756, dan X_1 (Usia) dengan *mean decrease accuracy* sebesar 0,400208.

```
> ##importance
> importance(RF100.2)
          High       Low    Medium MeanDecreaseAccuracy MeanDecreaseGini
X1 -1.2065984 -1.4833181 2.6243489           0.7530792        3.6261125
X2  0.8729495  0.4590306 0.1711314           0.8745762        0.9125805
X3  5.3856413  2.5560779 6.0987524           7.7962731        5.1192647
X4  5.8844773  7.5301266 4.2040817           8.8867780        6.5055643
X5  7.0763402  7.2922719 7.8095519          10.5978311       10.2843005
X6  6.4871011  2.4334253 7.3957490           9.0650606        6.1100183
X7  8.4169866  5.7748678 9.8593780          10.5006238        9.6888573
X8  7.0619901  8.0073844 7.2684575          11.4161960        9.6460360
```

Gambar 4. Gambar Output *Importance* untuk $m = 2$ dan $k = 100$ dari Software R

Selanjutnya variabel importance diberikan pada Gambar 5.



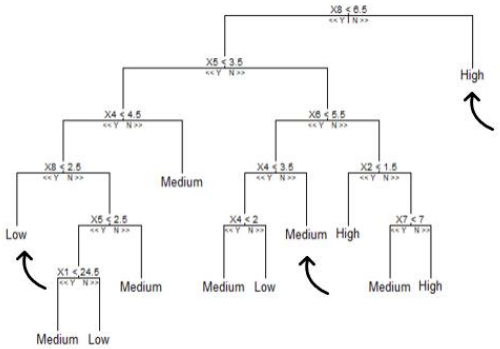
Gambar 5. *Variable Importance* untuk $m = 2$ dan $k = 100$

Gambar 4. dan 5. di atas menunjukkan bahwa semakin besar angka *Mean Decrease Accuracy*, maka semakin besar juga peran variabel tersebut dalam mempengaruhi hasil analisis. Berdasarkan Gambar 4. dan 5., pada penelitian ini semua variabel berpengaruh terhadap hasil klasifikasi prediksi kesehatan paru-paru pada pasien kanker paru-paru.

Variabel yang berpengaruh secara berurutan yaitu X_8 (Batuk darah) dengan *mean decrease accuracy* sebesar 11,4161960, X_5 (Obesitas) dengan *mean decrease accuracy* sebesar 10,5978311, X_7 (Perokok pasif) dengan *mean decrease accuracy* sebesar 10,5006238, X_6 (Merokok) dengan *mean decrease accuracy* sebesar 9,0650606, X_4 (Penggunaan alkohol) dengan *mean decrease accuracy* sebesar 8,8867780, X_3 (Polusi udara) dengan *mean decrease accuracy* sebesar 7,7962731, X_2 (Jenis kelamin) dengan *mean decrease accuracy* sebesar 0,8745762, dan X_1 (Usia) dengan *mean decrease accuracy* sebesar 0,7530792.

3.7 Pohon Klasifikasi Metode Random Forest

Berdasarkan klasifikasi Prediksi Kesehatan Paru-Paru pada Pasien Kanker Paru-Paru dengan $m = 2$, $k = 50$, dan 100 akan dilakukan pembentukan pohon klasifikasi. Berikut pohon klasifikasi dengan menggunakan $m = 2, k = 50$ dan 100. Gambar 6 merupakan pohon klasifikasi untuk $m = 2$ dan $k = 50$.

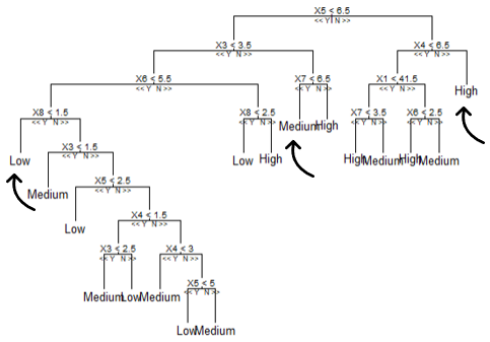


Gambar 6. Pohon Klasifikasi untuk $m = 2$ dan $k = 50$

Gambar 6. di atas menunjukkan bahwa:

a. Kesehatan paru-paru pada pasien yang menderita kanker paru-paru dikatakan rendah/*low* jika $X_8 < 2.5, X_4 < 4.5, X_5 < 3.5$, dan $X_8 < 6.5$.
b. Kesehatan paru-paru pada pasien yang menderita kanker paru-paru dikatakan sedang/*medium* jika $X_4 > 3.5, X_6 < 5.5, X_5 > 3.5$, dan $X_8 < 6.5$.
c. Kesehatan paru-paru pada pasien yang menderita kanker paru-paru dikatakan tinggi/*high* jika $X_8 > 6.5$.

Selanjutnya dengan menggunakan $m = 2, k = 100$ diberikan gambar pohon klasifikasi sebagai berikut.



Gambar 7. Pohon Klasifikasi untuk $m = 2$ dan $k = 100$

Gambar 7 di atas menunjukkan bahwa :

1. Kesehatan paru-paru pada pasien yang menderita kanker paru-paru dikatakan rendah/*low* jika X8 < 1.5, X6 < 5.5, X3 < 3.5, dan X5 < 6.5.
2. Kesehatan paru-paru pada pasien yang menderita kanker paru-paru dikatakan sedang/*medium* jika X7 < 6.5, X3 > 3.5, dan X5 < 6.5.
3. Kesehatan paru-paru pada pasien yang menderita kanker paru-paru dikatakan tinggi/*high* jika X4 > 6.5, X5 > 6.5.

Berdasarkan pohon klasifikasi dan *variable importance*, diketahui terdapat delapan variabel yang berpengaruh terhadap prediksi kesehatan paru-paru pada pasien kanker paru-paru, diantaranya usia, jenis kelamin, polusi udara, penggunaan alkohol, obesitas, merokok, perokok pasif, dan batuk darah. Variabel yang menjadi akar dalam pembentukan pohon klasifikasi di atas adalah variabel rata-rata obesitas dan batuk darah.

4. KESIMPULAN

Berdasarkan hasil analisis menggunakan metode Random Forest pada prediksi kesehatan paru-paru pada pasien kanker paru-paru diperoleh kesimpulan sebagai berikut:

1. Prediksi klasifikasi menggunakan metode Random Forest dilakukan dengan berbagai kombinasi m dan k. Berdasarkan kombinasi yang digunakan menghasilkan model optimal yaitu dengan $m = 2, k = 50$, dan $k = 100$ pada proporsi 80:20. Variabel yang paling mempengaruhi hasil analisis menggunakan metode Random Forest adalah batuk darah.
2. Dengan melihat tingkat akurasinya menggunakan *confusion matrix*, diperoleh bahwa prediksi kesehatan paru-paru pada pasien kanker paru-paru menggunakan metode Random Forest memperoleh nilai tingkat akurasi sebesar 0,95 atau 95%. Hal ini menunjukkan bahwa prediksi kesehatan paru-paru pada pasien kanker paru-paru dapat dikatakan sudah baik, karena memiliki nilai akurasi 0,95.
3. Berdasarkan *variabel importance*, semua variabel seperti usia, polusi udara, penggunaan alkohol, obesitas, merokok, perokok pasif, dan batuk darah memiliki pengaruh signifikan terhadap prediksi kesehatan paru-paru pada pasien dengan kanker paru-paru menggunakan metode random forest. namun, variabel yang paling berpengaruh adalah batuk darah.

DAFTAR PUSTAKA

Benaya, D. (2024). Implementasi Random Forest dalam klasifikasi kanker paru-paru. *JOINTER: Journal of Informatics Engineering*, *5*(01), 27-31.

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, 3(3), 3005-3013.

Fitriyaningsih, I., & Basani, Y. (2019). Flood prediction with ensemble machine learning using BP-NN and SVM. Jurnal Teknologi dan Sistem Komputer, 7(3), 93-97.

Han, J., Kamber, M., & Pei, J. (2012). Clustering analysis. Data Mining: Concept and Technique, MK imprint of Elsevier, New York, 478-490.

Marzuq, R. D., Wicaksono, S. A., & Setiawan, N. Y. (2023). Prediksi Kanker Paru-Paru menggunakan Algoritme Random Forest Decision Tree. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *7*(7), 3448-3456.

Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., & Kundu, S. (2018). Improved random forest for classification. *IEEE Transactions on Image Processing*, *27*(8), 4012-4024.

Permana, A. Y., Fazri, H. N., Athoilah, M. F. N., Robi, M., & Firmansyah, R. (2023). Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest. Jurnal Ilmiah Teknik Informatika dan Komunikasi, 3(2), 27-41.

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, *47*(1), 31-39.

Sari, L., Romadloni, A., & Listyaningrum, R. (2023). Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest. *Infotekmesin*, *14*(1), 155-162.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, *134*, 93-101.

Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications, 237, 121549.

Wulandari, E., & Perdana, A. (2022). Klasifikasi Kanker Paru-Paru Menggunakan Metode Naive Bayes. I-Robot Jurnal, 6(2), 1-10.

Zai, C. (2022). Implementasi Data Mining Sebagai Pengolahan Data. *Jurnal Portal Data*, *2*(3).