ISSN: 2548-9704

KOMPARASI ALGORITMA KLASIFIKASI DATA MINING UNTUK MEMPREDIKSI TINGKAT KEMATIAN DINI KANKER DENGAN DATASET *EARLY DEATH CANCER*

Panny Agustia Rahayuningsih[1]

Program Studi Sistem Informasi Akuntansi Kampus Kota Pontianak
Fakultas Teknologi Informasi, Universitas Bina Sarana Informatika
Jl. Abdurrahman Saleh No.18A, Pontianak
E-mail:panny.par@bsi.ac.id[1]

ABSTRACT

Cancer is something big in the world. Cancer is a malignant disease that is difficult to cure if the spread is too wide. However, detecting cancer cells as early as possible can reduce the risk of death. This study aims to predict the level of early detection of disasters in European countries using 5 classification algorithms, namely: Desecion Tree, Naïve Bayes, k-Nearset Neighbor, Random Forest and Neural Network of which algorithm is the best for this study. Tests carried out with several stages of research include: dataset (data contains), initial data processing, proposed method, credit method using 10 times cross validation, test results and t-test different tests. The alpha value is 0.05. if the probability is> 0.05 then H0 is accepted. If the probability is <0.05 then Ho is rejected. The results of the research that obtained performance with an accuracy value of 98.35% were the Neural Network algorithm. Whereas, the results of the research using the algirtic t-test with the best models are: Random Forest algorithm and Neural Network, the relatively good Naïve Bayes algorithm, the Desecion Tree algorithm is quite good and the poor algorithm is the K-Nearset Neighbor (K-NN) algorithm.

Keywords: Cancer, Algorithms, Classification, Probability

ABSTRAK

Penyakit Kanker merupakan sepuluh besar penyakit pembunuh di dunia. Kanker merupakan penyakit yang ganas dan sulit disembuhkan jika penyebarannya sudah terlalu luas. Akan tetapi, pendeteksian sel kanker sedini mungkin dapat mengurangi resiko kematian. Penelitian ini bertujuan untuk memprediksikan tingkat kematian dini kanker pada penduduk Eropa dengan menggunakan 5algoritma klasifikasi yaitu: *Desecion Tree, Naïve Bayes, k-Nearset Neighbour, Random Forest* dan *Neural Network* dari algoritma tersebut algoritma mana yang dianggap paling baik untuk penelitian ini. Pengujian dilakukan dengan beberapa tahapan penelitian antara lain: dataset (pengumpulan data), pengolahan data awal, metode yang diusulkan, pengujian metode menggunakan *10-fold cross validation,* evaluasi hasil dan uji beda t-test. Nilai alpha yang digunakan adalah 0.05. jika probabilitasnya >0.05 maka H_0 diterima. Sedangkan jika probabilitasnya <0.05 maka H_o ditolak.Hasil dari penelitian yang mendapatkan *performe* terbaik dengan nilai akurasi sebesar 98,35% adalah algoritma *Neural Network.* Sedangkan, hasil penelitian menggunakan uji *t-test* algoritma dengan model terbaik yaitu: algoritma *Random Forest* dan *Neural Network*, algoritma *Naïve Bayes* lumanyan baik, algoritma *Desecion Tree* cukup baik dan algoritma yang kurang baik adalah algoritma *K-Nearset Neighbour (K-NN).*

Kata kunci: Kanker, Algoritma, Klasifikasi, Probability

ISSN: 2548-9704

1. PENDAHULUAN

Menurut catatan WHO, Kanker merupakan sepuluh besar penyakit pembunuh di dunia[1]. Kanker adalah istilah yang digunakan untuk penyakit dimana sel-sel abnormal membelah tanpa kontrol dan mampu menyerang jaringan sel lain. Sel-sel kanker dapat menyebar ke bagian lain tubuh melalui pembuluh darah dan sistem limfe[2]. Kanker bukan hanya satu penyakit tapi banyak penyakit, ada lebih dari 100 berbagai jenis penyakit kanker[2]. Jaringan yang telah diserang oleh sel-sel kanker tidak akan berfungsi normal lagi dan akhirnya berujung pada kegagalan fungsi organ yang mengakibatkan kematian[2].

Kanker merupakan penyakit yang ganas dan sulit disembuhkan jika penyebarannya sudah terlalu luas. Akan tetapi, pendeteksian sel kanker sedini mungkin dapat mengurangi resiko kematian[2]. Kanker adalah sekelompok penyakit yang dapat menyebabkan hampir semua tanda atau gejala. Tanda-tanda dan gejala akan tergantung pada di mana kanker, seberapa besar itu, dan berapa banyak mempengaruhi organ atau jaringan di dekatnya. Jika kanker telah menyebar (metastasis), gejala dapat muncul di berbagai bagian tubuh[3]. Penyakit kanker merupakan salah satu penyebab kematian utama di seluruh dunia. Pada tahun 2012,kanker menjadi penyebab kematian sekita 8,2 juta orang. Berdasarkan data GLOBOCAN, International Agency Of Research on Cancer (IARC) diketahui bahwa pada tahun 2012 terdapat 14.067.894 kasus baru kanker dan 8.201.575 kematian akibat kanker di seluruh dunia. Penyebab terbesar kematian akibat kanker setiap tahunnya antara lain disebabkan oleh kanker paru, hati, perut, kolorektal, dan kanker payudara[12].

Penyakit kanker kini merupakan penyakit yang paling berbahaya di Negara Eropa. Di benua Eropa yang terdiri dari 53 negara, WHO menyebutkan penyakit jantung mencabut nyawa lebih dari 4 juta orang pada 2016. Jumlah kematian ini setara dengan 45 persen dari semua kematian di negara-negara Eropa[5]. Sementara itu kanker bertanggung jawab atas kurang dari setengah angka kematian akibat penyakit jantung di Eropa secara keseluruhan. Kanker, seperti dipaparkan dalam European Heart Journal, kini membunuh lebih banyak lelaki daripada penyakit jantung di 12 negara Eropa ini: Belgia, Denmark, Prancis, Israel, Italia, Luxembourg, Belanda, Norwegia, Portugal, Slovenia, Spanyol dan Inggris Raya[12].

Klasifikasi menentukan tingkat kematian dini kanker merupakan hal yang sangat penting untuk melakukan diagnosis secara dini agar dapat mengetahui Negara mana yang lebih baik,lebih buruk dan tidak ada perubahan yang signifikan dari rata-rata Negara Inggris. Dalam proses klasifikasi terdapat banyak algoritma yang digunakan untuk menentukan mana algoritma yang dianggap paling baik pada proses klasifikasi. Penelitian ini dilakukan untuk mengetahui algoritma data mining mana yang baik dalam mengklasifikasi data Earlydeathcancer pada penduduk Eropa. Penulis membandingkan beberapa metode klasifikasi data mining,diantaranya yaitu Algoritma *C4.5*, *Neural Network*, *Naives Bayes*, *K-NN* dan *Random Forest*. Hasil klasifikasi dari masing-masing algoritma akan dibandingkan dan dilihat tingkat akurasinya.

2. METODOLOGI PENELITIAN

2.1 Penyakit Kanker

Penyakit kanker adalah sel tubuh yang mengalami mutasi (perubahan) dan tumbuh tidak kendali serta membelah lebih cepat dibandingkan dengan sel normal. Sel kanker tidak mati setelah usianya cukup., melainkan tumbuh terus dan bersifat invasif sehingga sel normal tubuh dapat terdesak atau malah mati[12]. Penyakit kanker adalah penyakit yang timbul akibat pertumbuhan tidak normal sel jaringan tubuh yang berubah menjadi sel kanker[5].

ISSN: 2548-9704

2.2 Data Mining

Data Mining adalah suatu disiplin ilmu yang bertujuan menemukan, menggali atau menambahkan pengetahuan dari data atau informasi yang kita miliki[14].

2.3 Algoritma Klasifikasi Data Mining

Klasifikasi Data Mining adalah suatu metode pembelajaran., untuk memprediksi nilai dari sekelompok attribut dalam menggambarkan dan membedakan kelas data atau konsep yang bertujuan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui[8].

2.4 Algoritma Tree C4.5

Algortima C4.5 merupakan bagian dari kelompok algoritma decision tree dan merupakan kategori 10 algoritma yang paling popular[9]. Algoritma C4.5 diperkenalkan oleh J.Ross Quinlan diakhir tahun 1970 hingga awal tahun 1980-an. J. Ross Quinlan seorang peneliti dibidang mesin pembelajaran yang merupakan pengembangan dari algoritma ID3 (*Interative Dichotomiser*), algoritma tersebut digunakan untuk membentuk pohon keputusan. Pohon keputusan dianggap sebagai salah satu pendekatan yang paling popular, dalam klasifikasi pohon keputusan terdiri dari sebuah node yang membentuk akar node akar tidak memiliki inputan. Node lain yang bukan sebagai akar tetapi memiliki tepat satu inputan disebut node lainnya dinamakan daun. Daun mewakili nilai target yang paling tepat dari salah satu *class*[10].

Konsep dari algoritma C4.5 adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (rule). C4.5 adalah algoritma yang cocok untuk masalah klasifikasi dan data mining. Memetakan nilai atribut menjadi *class* yang dapat diterapkan untuk klasifikasi baru[11].

Tahapan dalam membangun sebuah pohon keputusan dengan algoritma C4.5, yaitu[11]:

1. Menyiapkan *data training.* Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon. Akar akan diambil dari atribut yang dipilih, dengan cara mengitung nilai gain dari masing-masing atribut.,nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai gain dari atribut, hitung dahulu nilai entropy. Untuk menghitung nilai entropy digunakan rumus.

$$Entropy\ (S) = \sum_{i=1}^{n} -pi * Log_2 pi \quad (1)$$

Keterangan:

S : himpunan kasus

A : atribut

N : Jumlah partisi S

Pi : proporsi dari Si terhadap S

3. Kemudian hitung mulai *gain* yang menggunakan rumus:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|si|}{|s|} * Entropy(si) \quad (2)$$

Keterangan :
S : himpunan kasus
A : fitur
n : jumlah partisi atribut A
| Si | : proporsi Si terhdapa S
| S | : jumlah kasus dalam S

4. Ulangi langkah ke-2 hingga semua record terpartisi.

5. Proses partisi pohon keputusan akan berhenti saat:

ISSN: 2548-9704

a. Semua record dalam simpul N mendapat kelas yang sama.
b. Tidak ada atribut di dalam record yang dipartisi lagi.
c. Tidak ada record di dalam cabang yang kosong.

2.5 *Neural Network*

Neural Network adalah satu set unit input/output yang terhubung dimana tiap relasinya memiliki bobot, selama *fase* pembelajaran, *neural network* menyesuaikan bobot sehingga dapat memprediksikan *class* yang benar dari *tupple*. *Neural Network* dimaksudkan untuk mensimulasikan perilaku sistem biologi susunan syaraf manusia, yang terdiri dari sejumlah besar unit pemroses (Alpanyandi, 2010). Neuron mempunyai relasi dengan *synapse* yang mengelilingi neuron-neuron lainnya. Susunan sayaraf tersebut dipresentasikan dalam neural network berupa graf yang terdiri dari simpul (neuron) yang dihubungkan dengan busur, yang berkorespondensi dengan *synapse*. Sejak tahun 1950-an, *neural network* telah digunakan untuk tujuan prediksi, bukan hanya klasifikasi tapi juga regresi dengan atribut target *continu*.

Langkah pembelajaran algoritma *backpropagation* adalah sebagai berikut:

1. Inisialisasikan bobot jaringan secara acak (biasanya antara-1.0 s/d 1.0)
2. Untuk setiap data pada data training, hitung input untuk simpul berdasarkan nilai input dan bobot jaringan saat itu. Dengan menggunakan rumus:

$$Input_j = \sum_{i=1}^{n} O\, i\, w_{ij} + \square_j \quad (3)$$

Keterangan:
O_i = output simpul I dari layer sebelumnya.
W_{ij} =bobot relasidari simpul I pada layer sebelumnya ke simpul j.
$\square_j$ = bias (sebagai pembatas)

3. Berdasarkan input dari langkah kedua, selanjutnya membangkitkan output untuk simpul menggunakan fungsi aktifitas sigmoid:

$$ouput = \frac{1}{1+c^{-input}} \quad (4)$$

4. Hitung nilai error antara nilai yang diprediksi dengan nilai yang sesungguhnya menggunakan rumus:

$$Error_j = ouput_j * (1\text{-}\ Output_j) * (Target_j\text{-}\ Output_j) \quad (5)$$

Keterangan:
$Output_j$ = Output actual dari simpul j
$Target_j$ = nilai target yang sudah diketahui pada data training.

5. Setelah nilai error dihitung, selanjutnya dibalik ke layer sebelumnya (*backpropagatioan*). Untuk menghitung nilai error pada hidden layer, menggunakan rumus:

$$ErrorJ = ouput_j * (1\text{-}\ ouput_j) * \sum_{k=1}^{n} errork\, w_{jk} \quad (6)$$

Keterangan:
$Output_j$ = Output actual dari simpul j
$Error_j$ = Error dari simpul k
W_{jk} = Bobot relasi dari simpul j ke simpul k pada layer berikutnya.

6. Nilai error yang dihasilkan dari langkah sebelumnya digunakan untuk memperbarui bobot relasi, dengan menggunakan rumus:

$$W_{ij} = W_{ij} + i * Error_j * Output_i \quad (7)$$

Keterangan:
W_{ij} = bobot relasi dari unit i pada layer sebelumnya ke unit j.
l = *learning* rate (konstantan, nilainya 0 sampai dengan 1)
Error = Error pada output layer simpul j
Output = Output dari simpul i.

ISSN: 2548-9704

2.6 *Naive Bayes*

Naïve Bayes merupakan metode yang tidak memiliki aturan[15]. *Naïve Bayes* menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training. *Naïve Bayes* merupakan metode klasifikasi popular dan masuk dalam sepuluh algoritma terbaik dalam data mining [18]. Bentuk umum dari teorema bayes dapat dilihat pada persamaan berikut:

$$P(H|X)\ \frac{P(H|X)P(H)}{P(X)} \quad (8)$$

Dimana:
X : Data dengan class yang belum diketahui.
H : Hipotesis data X merupakan suatu class spesifik
P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X (posterior probability)
P(H) : Probabilitas hipotesis H (prior probability)
P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) : Probabilitas dari X

2.7 *K-Nearest Neighbour*

Algoritma *K-Nearest Neighbour* adalah suatu metode yang menggunakan algoritma supervised [18].Tujuan dari algoritma *K-Nearest Neighbour* adalah untuk mengklasifikasi objek baru berdasarkan atribut dan training samples[15]. Pada proses klasifikasi algoritma ini tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori algoritma *K-Nearest Neighbour* menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari sampel uji yang baru.

Euclidean Distance

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(X_k - Y_k)^2} \quad (9)$$

Manhattan Distance

$$(x,y) = \sqrt{\sum_{k=1}^{n}|X_k - Y_k|} \quad (10)$$

Untuk mengukur jarak dari atribut yang mempunyai nilai besar, seperti atribut pendapatan, maka dilakukan normalisasi. Normalisasi bisa dilakukan dengan *min-max normalization* atau Z-*score* standardization[15]. Jika data training terdiri dari atribut campuran antara numerik dan kategori, lebih baik gunakan *min-max normalization* [15].

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (11)$$

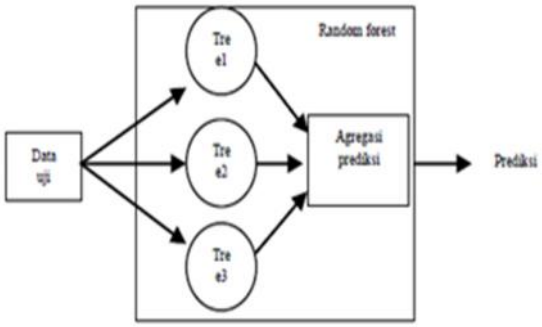
x^* = Standarisasi min-max

Dimana:

Min (x) = Nilai minimum data sampel
Max (x) =Nilai maximum data sampel

2.8 *Random Forest*

Random forest adalah algoritma klasifikasi den regresi yang merupakan bagian dari kelompok ensemble learning. Model base classifier yang dipakai adalah decision tree sehingga membentuk ‘forest’ atau hutan. Random forest dipopulerkan oleh Leo Breiman[7].



Gambar 1 Proses Prediksi Random Forest

ISSN: 2548-9704

Dalam random forest, pemilihan atribut dalam setiap kali sebuah node akan dipecah diambil secara acak,. Pertama-tama setiap tree diberi sample data training dengan menggunakan metode bagging, dan tiap tree dibangun menggunakan metode yang sama untuk membangun CART. Perbedaan yang mencolok terdapat pada proses pemilihan splitting criterion. Alih-alih mempertimbangkan seluruh atribut yang ada pada data, random forest hanya mempertimbangkan subset dari keseluruhan atribut (biasanya jumlah atribut yang diseleksi ditentukan oleh user). Atribut yang akan diseleksi ini didapatkan secara acak. Seperti CART, pembangunan tree akan berhenti ketika data sudah homogeny atau batas jumlah data minimum terlewati. Namun terdapat variasi Random Forest yang menentukan besar kedalaman tree maksimum[8].

Random Forest dapat menangani data dengan dimensi tinggi dengan baik dibanding model classifier lainnya. Berbeda dengan Decision Tree biasa, overfitting diatasi dengan menjaga variansi model tree dalam forest. Dalam Random umumnya tidak terdapat leaf pruning (penghilangan node leaf), tetapi variasi Random Forest seperti Hough Forest dalam implementasinya menerapakan leaf pruning untuk menghilangkan leaf node dengan probabilitas rendah[8]. Random Forest juga memiliki nama Decission Tree dalam pengaplikasiannya memiliki beberapa variasi, anatar lain Extremely Randomized Tree, dan Hough Forest. Random Forest dapat melakukan klasifikasi, regresi dan bahkan clustering. Hal ini menjadi daya tarik random forest dan turut mendorong aplikasi dan penelitian random forest dalam bidang computer vision[8].

2.9 Tahap Penelitian

Dalam penelitian ini penulis menggunakan sebuah *software* yaitu rapid miner. Penelitian ini bertujuan melakukan perbandingan untuk menentukan algoritma mana yang memiliki akurasi paling baik dengan menggunakan dataset *Early Death Cancer*. Pada penelitian ini dilakukan beberapa langkah-langkah atau tahapan penelitian antara lain: dataset (pengumpulan data), pengolahan data awal, metode yang diusulkan, pengujian metode, evaluasi hasil dan uji beda t-test.

2.10 Dataset

Penelitian ini, penulis menggunakan dataset dari *Public Health England, Health Profile Datasheets* 2013. Penulis peroleh data tersebut dari http://www.apho.org.uk/resource/browse.aspx?RID=126684. *Public Health England* merupakan suatu lembaga yang yang bergerak dibidang kesehatan. PHE adalah lembaga eksekutif, yang disponsori oleh Departemen kesehatan. Tujuannya untuk melindungi dan meningkatkan kesehatan dan kesejahteraan bangsa, dan mengurangi kesenjangan kesehatan. Dataset yang digunakan adalah *Early Death Cancer*. Pada dataset *early death cancer* terdapat 7 atribut dan 1 class seperti pada tabel 1 dibawah ini.

Tabel 1. Atribut Dataset *Early Death Cancer*

No	**Nama Atribut**	**Keterangan**
1	*Period*	2009-2011
2	*Aggregate Numerator*	Jumlah kematian akibat kanker yang terdaftar pada tahun 2009, 2010, 2011, orang yang berusia <75
3	*Aggregate Denominator*	Jumlah dari 2011 penduduk pertengahan tahun berdasarkan Sensus memperkirakan 2009, 2010, 2011, orang yang berusia

ISSN: 2548-9704

		<75
4	*Single Year Numerator*	Rata-rata jumlah tahunan kematian akibat kanker, 2009-2011, orang berusia <75
5	*Single Year Denominato*	Rata-rata penduduk pertengahan tahun berdasarkan 2011 Sensus tahunan memperkirakan 2009-2011, orang berusia <75
6	*Indicator Value*	Angka kematian dini (DSR) dari semua kanker (orang yang berusia <75)
7	*95% CI*	Lebih rendah 95% CI dan Atas 95% CI
8	*Significance* ***(Class)***	Menunjukkan jika suatu daerah secara signifikan lebih baik, buruk atau tidak berbeda secara signifikan dari rata-rata Inggris, berdasarkan apakah nilai rata-rata Inggris termasuk dalam 95%

2.11 Pengolahan Awal Data

Jumlah data yang digunakan adalah 363 data. Untuk mendapatkan data yang berkualitas, beberapa teknik yang dilakukan adalah sebagai berikut[17].

1. Data validation, untuk mengidentifikasi dan menghapus data yang ganjil *(outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (missing value). Missing data terlihat
2. Data integration and Transformation, untuk meningkatkan akurasi dan efisiensi algoritma. Data ditransformasikan ke dalam software RapidMiner.
3. Data size reduction and dicrtization, untuk memperoleh data set dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat informatif.

2.12 Metode dan Evaluasi

Dalam penelitian ini adapun metode dan evaluasi yang akan dilakukan dengan menggunakan metode eksperimen yaitu menggunakan beberapa metode klasifikasi data mining terhadap dataset *Earlideathcancer* yang dimana untuk menentukan tingkat kematian dini kanker pada penduduk Eropa. Data akan diolah dengan menggunakan algoritma *Decision Tree, Naïve Baye, K-NN, Random Forest, Neural Network* dan menghasilkan model, maka terhadap model yang dihasilkan tersebut dilakukan pengujian menggunakankan *k-fold cross validation*, kemudian dilakukan evaluasi dan validasi hasil dengan *confusion matrix* dan uji beda t-test.

N. Pengujian Model

Pengujian model dalam penelitian ini menggunakan *Cross Validation* adalah teknik validasi dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi (Han & Kamber, 2006). *Cross Validation* akan dilakukan sebanyak 10 kali untuk memperkirakan akurasinya. Pada penelitian ini, nilai k berjumlah 10 atau 10-*fold cross validation.* Berikut contoh 10-*fold cross validation* pada tabel 2 dibawah ini.

ISSN: 2548-9704

Tabel 2. *10-Fold Cross Validation*

n-Validasi	Dataset									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

O. Evaluasi Hasil

Pada penelitian algoritma klasifikasi data mining terdapat evaluasi untuk mengetahui tingkat akurasi dari algoritma klasifikasi tersebut. Dalam algoritma klasifikasi dibagi menjadi 2 data pada umunya yaitu data training dan data testing. Data training digunakan untuk membuat suatu pola dalam membentuk sebuah model klasifikasi. Sedangkan, data testing digunakan untuk mengukur akurasi dari algoritma klasifikasi apakah berhasil melakukan klasifikasi dengan benar.

Evaluasi menggunakan *Confusion matrix* untuk memberikan keputusan yang diperoleh dalam *training* dan *testing*. *Confusion matrix,* memberikan penilaian *performace* klasifikasi berdasarkan objek dengan benar atau salah. Untuk mendapatkan hasil akurasi yang lebih baik maka dilakukan percobaan. Dari percobaan yang dilakukan dalam penelitian ini adalah menghitung nilai rata-rata keseluruhan.

P. Uji Beda T-Test

Selain menggunakan *Confusion matrix* , pada penelitian ini juga mencoba melakukan pengujian dengan menggunakan metode uji T-Test untuk mendapatkan model terbaik. T-Test adalah metode pengujian untuk menentukan dua sampel yang tidak berhubungan yang memiliki nilai rata-rata berbeda. Caranya dengan membandingkan perbedaan dari ke dua sampel tersebut.

3. HASIL DAN PEMBAHASAN

Dari hasil penelitian dengan menggunakan alat bantu rapid miner ini menunjukan algritma Neural Network dengan performa terbaik yang mendapat nilai akurasi sebesar 98.35%. Algoritma Random Forest dengan nilai sebesar akurasi 96.42%, algoritma Naïve Bayes dengan nilai sebesar akurasi 86.23%. algoritma C4.5 dengan nilai akurasi sebesar 69.97%. sedangkan untuk algoritma K-NN mendapatkan nilai akurasi sebesar 43.80%. Nilai akurasi tersebut didapat dengan menggunakan data EarlyDeathCancer. Keseluruhan hasil dari penelitian ini dapat dilihat pada tabel III untuk melihat hasil perbandingan akurasi pada masing-masing algoritma.

Tabel 3. Hasil Nilai Akurasi

Metode	Akurasi
C4.5	69,97%
Naive Bayes	86,23%
K-NN	43,80%
Random Forest	96,42%
Neural Network	98,35%

ISSN: 2548-9704

Setelah mendapatkan nilai akurasi dari 5 algortima tersebut, penelitian selanjutnya adalah melakukan pengujian dengan uji beda t-test. Pada pengujian ini, uji t-test beda memiliki nilai signifikan yang berbeda pada setiap algoritmanya. Nilai alpha yang digunakan adalah 0.05. jika probabilitasnya >0.05 maka H_0 diterima (tidak ada perbedaan yang signifikan) sedangkan jika probabilitasnya <0.05 maka H_o ditolak (ada perbedaan yang signifikan). Berikut ini tabel dari hasil uji T-test dari keseluruhan algoritma klasifikasi.

Tabel 4: Hasil Uji *T-Test*

	C4.5	Naïve Bayes	K-NN	Random Forest	Neural Network
C4.5	-	0.022	0.001	0.000	0.000
Naïve Bayes	-	-	0.000	0.001	0.000
K-NN	-	-	-	0.000	0.000
Random Forest	-	-	-	-	0.189
Neural Network	-	-	-	-	-

Dari tabel diatas, algortima yang lebih kecil dari nilai alpa 0.05 menunjukkan adanya perbedaan yang signifikan yaitu: algoritma C4.5 dengan Naïve Bayes, K-NN, Random Forest dan Neural Network. Algortima Naïve Bayes dengan K-NN, Random Forest dan Neural Network. Kemudian, algoritma K-NN dengan Random Forest dan Neural Network. Sedangkan untuk algoritma Random Forest dengan Neural Network menunjukkan hasil lebih dari nilai alpa sehingga tidak ada perbedaan yang singnifikan. Jadi, model terbaik untuk dataset *Early Death Cancer* ini yaitu: algoritma *Random Forest* dan *Neural Network*.

4. KESIMPULAN DAN SARAN

Dari hasil penelitian yang telah dilakukan dengan menggunakan dataset *EarlyDeathCancer* dapat disimpulkan bahwa:

1. Penelitian menggunakan 5 algoritma yaitu: decision tree, naïve bayes, k-nn, random forest, dan neural network.
2. Pengujian dilakukan dengan menggunakan dataset dari *Public Health England, Health Profile Datasheets* 2013 yaitu dataset *EarlyDeathCancer*. Tahapan peneltian antara lain: dataset (pengumpulan data), pengolahan data awal, metode yang diusulkan, pengujian metode (10-*fold cross validatio*), evaluasi hasil menggunakan *Confusion matrix* dan uji beda t-test.
3. Hasil penelitian dari pengujian metode menggunakan *Confusion matrix* menghasilkan performe terbaik dengan nilai akurasi sebesar 98,35% adalah algortima *Neural Network.*
4. Urutan algoritma yang dihasilkan pada pengujian menggunakan uji beda t-test mendapatkan hasil model algoritma yang terbaik adalah Random Forest dan Neural Network, algoritma Naïve Bayes lumanyan baik, algoritma C4.5 cukup baik sedangkan untuk algoritma K-NN algoritma yang kurang baik digunakan pada dataset ini.

REFERENSI

[1] World Health organization, http://www.who.int/mediacentre/factsheets/fs310/en/. diakses tanggal 21 Agustus 2016

[2] Kementerian Kesehatan Republik Indonesia, http://www.dharmais.co.id/index.php/what-is-cancer-id.html. diakses tanggal 21 Agustus 2016

[3] Kementerian Kesehatan Republik Indonesia, http://www.dharmais.co.id/index.php/what-are-the-warning-signs.html. diakses tanggal 21 Agustus 2016

ISSN: 2548-9704

[4] https://beritagar.id/artikel/gaya-hidup/kanker-kini-pembunuh-nomor-satu-di-eropa-barat. diakses tanggal 22 Agustus 2016.

[5] http://www.depkes.go.id/resources/download/pusdatin/infodatin/infodatin-kanker.pdf. diakses tanggal 21 Agustus 2016

[6] Alfisahrin C4.5, Naive Bayes dan Neural Network, S. (2014). Komparasi Algoritma Untuk Memprediksi Penyakit Jantung. Jakarta: Pascasarjana Magister Ilmu Komputer STMIK Nusa Mandiri.

[7] Breiman, Leo. 2001. Random Forests. Kluwer-Academic Publishers

[8] Criminisi, A., Shotton, J. 2013. Decision Forest for Computer Vision and Medical Image Analysis. Springer-Verlag London.

[9] Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques. San Fransisco: Morgan kauffman.

[10] Han, & Kamber. (2006). Data Mining Concepts and technique. San Francisco: Diane Cerra.

[11] Kusrini, & Luthfi, E. T. (2009). Algoritma Data Mining.Yogyakarta: Andi Offset

[12] Mugi Wahidin, dkk. Situasi Penyakit Kanker. Pusat Data dan Informasi. Kemeterian Kesehatan RI. Diakses pada tanggal 21 Agustus 2016.

[13] Maimon & Rokach. (2010). Data Mining and knowledge Discovery Handbook. New York: Springer.

[14] Susanto, S., & Suryadi, D. (2010). Pengantar Data MiningMenggali Pengetahuan dari Bongkahan Data. Yogyakarta: C.V ANDI OFFSET

[15] Toni Arifin.2015.Metode Data Mining untuk Klasifikasi Data sel Nukleus dan Sel Radang Berdasarkan Analisa Tekstur. Informatika Vol II No 2 September 2015.Diambil dari: http://ejournal.bsi.ac.id/assets/files/Toni_-_Univ_BSI_-_OK_(425-433).pdf

[16] Vercellis, C. (2009). Data Mining and Optimization for Decision Making. Italy: WILEY.

[17] Vercellis. (2009). Business Intelligence: Data Mining and Optimization for Decision Making Decision Making.

[18] Wu & Kumar. (2009). The Top Ten Algorithms in Data Mining. USA: CRC Press.