View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by E-Jurnal UIN (Universitas Islam Negeri) Alauddin Makassar

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
P-ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

KLASIFIKASI KANKER MENGGUNAKAN ALGORITMA NNGE, RANDOM FOREST, DAN RANDOM COMMITEE

MUHAMMAD NUR AKBAR

Program Studi Teknik Informatika Fakultas Sains dan Teknologi
Universitas Islam Negeri Alauddin Makassar

Email: muhammad.akbar@uin-alauddin.ac.id

ABSTRAK

Permasalahan yang dibahas dalam penelitian ini adalah seputar data pasien kanker pada sebuah klinik. Data yang digunakan yaitu data pasien dimana setiap pasien menjalani 4 tipe tes laboratorium. Dari data tes tersebut, dilakukan pemrosesan yang menghasilkan suatu pola atau model. Selanjutnya, pola tersebut digunakan untuk mendiagnosa pasien yang lain apakah menderita penyakit kanker atau tidak. Dalam masalah ini pemprosesan dilakukan dengan algoritma NNGE, Random Forest, Random Committee. Penggunaan ketiga algoritma tersebut diharapkan dapat menghasilkan klasifikasi dengan tingkat ketidaktepatan minimum. Sebelumnya data training dibagi menjadi 2 bagian, dimana 75% diambil sebagai data training dan 25% sisanya digunakan sebagai data validation. Hasil klasifikasi terhadap data 100 data uji yaitu sebanyak 37 pasien dinyatakan *malignant* dan sebanyak 63 pasien dinyatakan *benign*.

Kata Kunci*:* Klasifikasi, NNGE, *Random Forest, Random Committee, preprocessing*

I. PENDAHULUAN

Kanker merupakan salah satu penyakit penyebab kematian utama di dunia. Pada tahun 2012 diperkirakan terdapat 14 juta kasus baru kanker dan 8,2 juta kematian akibat kanker di dunia. Health Organization (WHO) melaporkan limba besar jenis kanker, yaitu kanker paru, prostat, kolorektum, kanker perut, dan kanker hati. Sedangkan pada perempuan kasus terbanyak berupa kanker payudara, kolerektum, paru-paru, serviks, serta kanker perut.

Angka kematian akibat kanker lebih rendah di negara maju dibandingkan negara berkembang. Perbedaan ini mencerminkan perbedaan faktor resiko dan keberhasilan penanganan deteksi, serta ketersediaan pengobatan. Berdasarkan data Riskesdas, prevalensi tumor/kanker di Indonesia menunjukkan adanya kenaikan

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 5 Nomor. 2, Oktober 2020
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

dari 1,4 per 1000 penduduk di tahun 2013 menjadi 1,79 per 1000 penduduk pada tahun 2018.

Untuk pencegahan dan pengendalian kanker di Indonesia, pemerintah telah melakukan berbagai upaya antara lain dengan melakukan deteksi dini pada kelompok masyarakat berpotensi. Pada umumnya pendeteksian tingkat keganasan kanker adalah dengan cara prognosis. Prognosis merupakan "tebakan terbaik" tim medis dalam menentukan sembuh atau tidaknya pasien dari penyakit kanker (F. Rachman dan S. W. Purnami, 2012). Selain dengan progonosis, cara lain dalam melakukan deteksi yaitu dengan memanfaatkan teknologi bioinformatika dengan menggunakan teknik *data mining* (G. I. Salama, dkk., 2012), karena telah terbukti dalam mendeteksi tingkat keganasan kanker (A. Bellaachia dan E. Guven, 2006).

Data mining adalah sebuah proses menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* dalam mengidentifikasi informasi dan pengetahuan terkait dari berbagai basis data yang besar (Turban, dkk., 2005). Pengetahuan yang dimaksud di sini berupa pola tersembunyi yang belum ditemukan sebelumnya, yaitu pola dari data hasil pemeriksaan pasien yang mengidap kanker atau kanker jinak.

Kanker atau sering disebut dengan tumor secara umum dibagi menjadi dua macam, yaitu tumor jinak (*benign*) dan ganas (*malignant*). Tumor jinak memiliki kondisi dan perkembangan yang tidak bersifat kanker dimna dapt terdeteksi namun tidak menyebar dan merusak jaringan lain di sekitarnya. Pada level ganas atau kanker, tumor akan menyebar dan merusak jaringan dan organ sekitarnya (Rashmi, dkk., 2015). Kanker ganas jika tidak ditangani dengan cepat akan berdampak buruk bagi penderita. Dengan mengetahui jinak atau ganasnya kanker yang diderita maka dapat segera ditangani degan sesuai.

Pada penelitian ini digunakan data hasil pemeriksaan lab pasien sebagai data latih untuk mendiagnosa apakah seorang pasien menderita penyakit kanker atau tumor jinak. Data tersebut memiliki atribut : ID, umur, beberapa tipe uji lab, dan target sebagai indikasi apakah pasien menderita tumor jinak (*benign*) atau ganas (*malignant*).

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 5 Nomor. 2, Oktober 2020
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

II. METODE PENELITIAN

A. DATA

Digunakan data dari klinik 'X' yang merupakan data hasil pemeriksaan pasien penderita penyakit kanker. Data tersebut memiliki atribut : ID, umur, beberapa tipe uji lab, dan target sebagai indikasi apakah pasien menderita penyakit kanker atau tidak.

Terdapat dua buah data, yaitu data latih dan data uji dengan atribut yang sama. Data training latih dari 861 baris data yang digunakan untuk menemukan pola pasien yang menderita penyakit kanker. Sedangkan data uji memuat 100 baris data pasien yang harus diprediksi.

Tabel 1. Data latih

ID Pasien	Usia	Hasil Uji Lab				Malignant (1) / Benign (0)
		Tipe 1	Tipe 2	Tipe 3	Tipe 4	
ID0001	67	5	3	5	3	1
ID0002	58	5	4	5	3	1
…	…	…	…	…	…	…
ID0960	55	5	4	4	3	1
ID0961	55	5	4	3	3	1

Tabel 2. Data uji

ID Pasien	Usia	Hasil Uji Lab				Malignant (1) / Benign (0)
		Tipe 1	Tipe 2	Tipe 3	Tipe 4	
ID0017	45	5	4	5	3	?
…	…	…	…	…	…	…
ID0823	56	4	2	4	3	?

B. PREPROCESSING

Dari data latih yang digunakan ditemukan dua permasalahan, yaitu masalah ambiguitas dan pencilan. Masalah ambiguitas ditemukan pada beberapa baris data dengan hasil pemeriksaan yang sama tetapi memiliki diagnosa akhir yang berbeda. Sedangkan pencilan ditemukan pada pasien dengan ID ID0258.

Dari permasalahan tersebut dilakukan beberapa alternatif penanganan, diantaranya :

- Penanganan pertama : data yang ambigu diubah ke data yang paling banyak muncul, yang 50:50 diganti 0, dan data pencilan diganti dengan 4, dengan

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 5 Nomor. 2, Oktober 2020
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

pertimbangan data tersebut adalah data yang paling banyak muncul menurut tipe 1.

- Penanganan kedua : data yang ambigu diubah ke data yang paling banyak muncul, yang 50:50 diganti 1, dan data pencilan di ganti dengan 4, dengan pertimbangan data tersebut adalah data yang paling banyak muncul menurut tipe 1.
- Penanganan ketiga : data yang ambigu diubah ke data yang paling banyak muncul, yang 50:50 diganti 0, dan data pencilan di ganti dengan 5 dengan asumsi salah memasukkan data.
- Penanganan keempat : data yang ambigu diubah ke data yang paling banyak muncul, yang 50:50 diganti 1, dan data pencilan di ganti dengan 5 dengan asumsi salah memasukkan data.

Dari keempat alternatif penanganan tersebut, masing-masing data dibagi menjadi 2 bagian, dengan perbandingan 75 : 25, dimana 75% data digunakan sebagai data training dan 25% sisanya sebagai data validation. Proses pembagian data training dan data validation tersebut dilakukan secara random.

Pengambilan secara random dilakukan dengan mengambil 25% data tiap umur sebagai data validation dan 75% sisanya digunakan sebagai data training.

Misalnya data dengan umur 29 berjumlah 10 orang, maka 25% dari 10, yaitu 3 data, diambil sebagai data validation, dan sisanya digunakan sebagai data training.

Berdasarkan penjelasan diatas, dipilih alternatif yang pertama karena dari data hasil uji coba (akan dijelaskan pada Bab 3) dihasilkan akurasi data yang paling besar dibandingkan dengan ketiga alternative penanganan lainnya.

Selain itu, dalam preprocessing ini juga dilakukan feature selection, yaitu dengan menghilangkan atribut ID Pasien dengan pertimbangan atribut tersebut tidak memiliki pengaruh dalam permrosesan data.

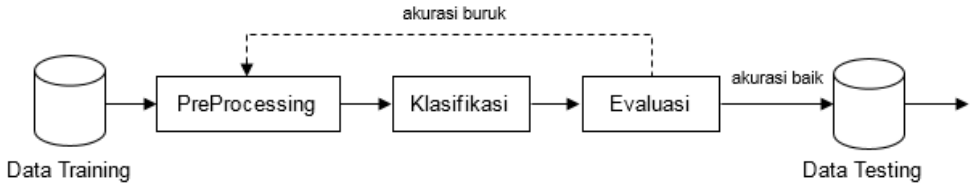
Sebelum memulai preprocessing, dilakukan penyederhanaan nama atribut untuk memudahkan dan mencegah terjadinya keambiguan atribut. Dalam kasus ini, dicontohkan sebagai berikut:

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 5 Nomor. 2, Oktober 2020
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

Tabel 3. Nama atribut setelah *preprocessing*

ID Pasien	Usia	Tipe 1	Tipe 2	Tipe 3	Tipe 4	Kelas

C. METODE KLASIFIKASI



Gambar 1. Proses klasifikasi

Penelitian ini menggunakan tools WEKA 3.8.3 dalam melakukan klasifikasi. WEKA adalah sebuah perangkat lunak yang memiliki banyak algoritma machine learning untuk keperluan data mining. WEKA juga memiliki banyak tools open source untuk pengolahan data, mulai dari pre-processing, classification, regression, clustering, association rules, dan visualization.

a. NNGE

K-Nearest Neighbor sangat sering digunakan dalam klasifikasi dengan tujuan dari algoritme ini adalah untuk mengklasifikasi objek baru berdasarkan atribut dan training samples (Larose, 2004).

Algoritma k-nearest neighbor (k-NN atau KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data training yang jaraknya paling dekat dengan objek tersebut. Teknik ini sangat sederhana dan mudah diimplementasikan. Mirip dengan teknik klastering, pengelompokkan suatu data baru berdasarkan jarak data baru itu ke beberapa data/tetangga (neighbor) terdekat. Dalam hal ini jumlah data/tetangga terdekat ditentukan oleh user yang dinyatakan dengan k. Misalkan ditentukan k=5, maka setiap data testing dihitung jaraknya terhadap data training dan dipilih 5 data training yang jaraknya paling dekat

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 5 Nomor. 2, Oktober 2020
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

ke data testing. Lalu periksa output atau labelnya masing-masing, kemudian tentukan output mana yang frekuensinya paling banyak. Lalu masukkan suatu data testing ke kelompok dengan output paling banyak. Misalkan dalam kasus klasifikasi dengan 3 kelas, lima data tadi terbagi atas tiga data dengan output kelas 1, satu data dengan output kelas 2 dan satu data dengan output kelas 3, maka dapat disimpulkan bahwa output dengan label kelas 1 adalah yang paling banyak. Maka data baru tadi dapat dikelompokkan ke dalam kelas 1. Prosedur ini dilakukan untuk semua data testing (Santosa, 2007).

Algoritma NNGE (Nearest Neighbor with Generalized Exemplar) merupakan pengembangan lebih lanjut dari algoritma k-NN yang menerapkan konsep exemplar. Konsep ini digunakan untuk mempercepat proses klasifikasi tanpa mengurangi akurasi dari hasil prediksi. Dalam kasus ini, algoritma ini cocok diterapkan untuk mengklasifikasikan 4 type hasil uji laboratorium untuk keperluan pendiagnosaan kanker.

b. Random Forest

Random forest adalah sebuah metode pengelompokan ensemble yang di dalamnya terdapat banyak decision trees dan output kelas yang modus output kelas oleh pohon individu. Algoritma ini dikembangkan oleh Leo Breiman dan Adele Cutler, dan "Random Forest" adalah merek dagang mereka. Istilah ini berasal dari Random Decision Tree yang pertama kali diusulkan oleh Tin Kam Ho dari Bell Labs pada tahun 1995. Metode ini menggabungkan ide "Bagging" dari Breiman dan Random feature selection, diperkenalkan secara terpisah oleh Ho dan Amit dan Geman untuk membangun koleksi Decision Tree dengan variasi yang terkontrol.

c. Random Commitee

Random committee merupakan salah satu dari sekian banyak random tree classifier. Dalam kasus klasifikasi, Random committee algorithm menghasilkan prediksi dari rata-rata perkiraan kemungkinan tentang

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 5 Nomor. 2, Oktober 2020
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

klasifikasi tersebut. Random committee algorithm merupakan sebuah jalan yang efektif untuk meningkatkan akurasi klasifikasi yang tidak begitu baik ketika dibandingkan dengan rata-rata yang sederhana dan individual model.

D. METODE EVALUASI

a. Holdout

Dalam *holdout method*, data awal yang memiliki label kelas akan dipartisi menjadi 2 data yang dinamakan *training set* dan *test set*. Model klasifikasi kemudian dibangun menggunakan data yang ada pada *training set*, kemudian model tersebut akan diujikan ke *test set*, lalu hasil prediksi akan dibandingkan dengan kelas aktual untuk mendapatkan nilai akurasi. Proporsi data yang digunakan untuk training set dan test set tergantung pada kebutuhan analisis misalnya 50%-50%, 75%-25%, 90%-10% dimana umumnya jumlah data *training* sama atau lebih banyak dari data *testing*.

b. Accuracy

Umumnya evaluasi kinerja dari suatu algoritma klasifikasi dapat dilakukan dengan menggunakan *confusion matrix*, yang akan menghasilkan nilai akurasi. Akurasi dalam klasifikasi merupakan persentase ketepatan record data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi (Han dan Kember, 2006).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Keterangan :

True Positive (TP) : jumlah record positif yang diklasifikasikan sebagai positif

False Positive (FP) : jumlah record negatif yang diklasifikasikan sebagai positif

True Negative (TN) : jumlah record negatif yang diklasifikasikan sebagai negatif

False Negative (FN) : jumlah record negatif yang diklasifikasikan sebagai positif

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 5 Nomor. 2, Oktober 2020
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

c. Root Mean Square Error

Root Mean Square Error (RMSE) merupakan besarnya tingkat kesalahan hasil klasifikasi/prediksi, dimana semakin kecil (mendekati 0) nilai RMSE maka hasil prediksi akan semakin akurat. Nilai RMSE dapat dihitung dengan persamaan sebagai berikut.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Xi-Yi)\text{^}2}{n}} \quad (2)$$

Keterangan :

Xi : nilai data aktual

Yi : nilai data prediksi

n : jumlah data

III. HASIL DAN PEMBAHASAN

A. HASIL KLASIFIKASI DATA LATIH

Dari percobaan yang dilakukan sebelumnyadiperoleh hasil sebagai berikut.

a. Penanganan pertama : data yang ambigu diubah ke data yang paling banyak muncul, yang 50:50 di ganti 0, dan data pencilan di ganti dengan 4, dengan pertimbangan data tersebut adalah data yang paling banyak muncul menurut tipe 1.

Tabel 4. Hasil pengujian penanganan pertama

Metode	Data Training		Data Validation	
	Akurasi	RMSE	Akurasi	RMSE
NNGE	100%	0	93.02%	0.2641
RandomForest	99.38%	0.1027	92.56%	0.2559
Random Committe	100%	0	91.63%	0.2725

b. Penanganan kedua : data yang ambigu diubah ke data yang paling banyak muncul, yang 50:50 di ganti 1, dan data pencilan di ganti dengan 4, dengan pertimbangan data tersebut adalah data yang paling banyak muncul menurut tipe 1.

Tabel 5. Hasil pengujian penanganan kedua

Metode	Data Training		Data Validation	
	Akurasi	RMSE	Akurasi	RMSE

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 5 Nomor. 2, Oktober 2020
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

NNGE	100%	0	92.09%	0.2812
RandomForest	99.54%	0.1041	92.09%	0.2506
Random Committe	100%	0	91.63%	0.2697

c. Penanganan ketiga : data yang ambigu diubah ke data yang paling banyak muncul, yang 50:50 diganti 0, dan data pencilan di ganti dengan 5 dengan asumsi salah memasukkan data.

Tabel 6. Hasil pengujian penanganan ketiga

Metode	Data Training		Data Validation	
	Akurasi	RMSE	Akurasi	RMSE
NNGE	100%	0	92.56%	0.273
RandomForest	100%	0	91.63%	0.259
Random Committe	99.23%	0.107	91.63%	0.248

d. Penanganan keempat : data yang ambigu diubah ke data yang paling banyak muncul, yang 50:50 diganti 1, dan data pencilan di ganti dengan 5 dengan asumsi salah memasukkan data.

Tabel 7. Hasil pengujian penanganan keempat

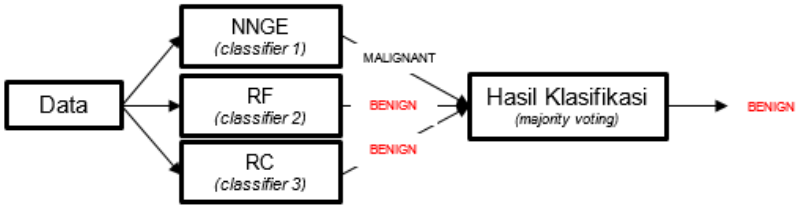
Metode	Data Training		Data Validation	
	Akurasi	RMSE	Akurasi	RMSE
NNGE	91.64%	0	92.56%	0.273
RandomForest	100%	0	91.63%	0.289
Random Committe	99.38%	0	91.63%	0.257

Dari keempat penanganan tersebut, data diurutkan berdasarkan akurasi dari data validation. Berdasarkan hasil yang diperoleh, maka dipilih alternatif yang pertama karena akurasi data yang didapatkan paling besar dibandingkan dengan ketiga alternatif penanganan lainnya.

B. HASIL KLASIFIKASI DATA UJI

Dalam melakukan klasifikasi pada data uji digunakan *ensemble classifier* dengan menggunakan *majority voting* dalam menentukan kelas.

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 5 Nomor. 2, Oktober 2020
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id



Gambar 2. Proses Ensemble Classifier

Hasil klasifikasi terhadap data 100 data uji yaitu sebanyak 37 pasien dinyatakan *malignant* dan sebanyak 63 pasien dinyatakan *benign*.

IV. KESIMPULAN

Dari uji coba yang telah dilakukan dapat disimpulkan bahwa penanganan kasus klasifikasi kanker dengan algoritma NNGE, Random Forest dan Random Committee efektif memberikan hasil yang akurat. Berdasarkan penggunaan ketiga algoritma ini pada data validation, didapatkan nilai akurasi yang tinggi.

DAFTAR PUSTAKA

F. Rachman dan S. W. Purnami, "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM)," JURNAL SAINS DAN SENI ITS Vol. 1, No. 1, (Sept. 2012) ISSN: 2301-928X, pp. D-130, 2012.

G. I. Salama, M. B. Abdelhalim dan M. A.-e. Zeid, "Experimental Comparison of Classifiers for Breast Cancer Diagnosis," dalam International Conference Computer Engineering and Systems (ICCES), Cairo, 2012.

A. Bellaachia dan E. Guven, "Predict ing Breast Cancer Survivability Using Data Mining Techniques," dalam SIAM Conference on Data Mining, Washington DC, 2006.

Turban, Efraim., et al. 2005. "Decision Support Systems and Intelligent Systems"

Rashmi, G. D., Lekha, A., & Bawane, N. (2015). Analysis of Efficiency of Classification and Prediction Algorithms (Naïve Bayes) for Breast Cancer Dataset. In 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)

Larose, Daniel T. 2004. K-Nearest Neighbor Algorithm.

Santosa, Budi. 2007. Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta : Graha Ilmu.

Han, Jiawei., Kember, Micheline. 2006. Data Mining : Concept and Technique. San Francisco : Morgan Kaufmann Publishers. 2006.