GENETIC ALGORITHM FOR FEATURE SELECTION IN NAÏVE BAYES IN LIFE RESISTANCE CLASSIFICATION ON BREAST CANCER PATIENT

Dhika Malita Puspita Arum*; Andri Triyono

ABSTRACT

Breast cancer is the most common cancer in women's suffering and is the second leading cause of death for women (after lung cancer). More than one million cases and nearly 600,000 breast cancer deaths occur worldwide each year. Survival is generally defined as surviving patients over a period of time after the diagnosis of the disease. Accurate predictions about the likelihood of survival of breast cancer patients can allow doctors and healthcare providers to make more informed decisions about patient care. To classify the survival of breast cancer patients can do the utilization of data mining techniques with Naive Bayes algorithm. Naive Bayes is very simple and efficient but very sensitive to the features so from it the selection of the appropriate features is in need because irrelevant features can reduce the level of accuracy. Naive Bayes will work more effectively when combined with some attribute selection procedures such as Genetic Algorithm. In this study the researchers proposed the Genetic Algorithm for Feature Selection on Naive Bayes so as to improve the accuracy of breast cancer survival classification results. In this study using a private dataset breast cancer patients. The results show that Naive Bayes Genetic Algorithm has a higher accuracy of 90% compared to Naive Bayes with 86% accuracy .

Keywords*; Breast Cancer, Survival, Classification, Feature Selection, Naive Bayes, Genetic Algorithm*

Correspondence:
Dhika Malita Puspita Arum
Universitas An Nuur, Email; dhika.malita11@gmail.com

PENDAHULUAN

Kanker payudara (*breast cancer*) adalah kanker yang paling umum di derita kaum wanita. Kanker Payudara merupakan kanker yang berkembang di jaringan payudara. Tanda-tanda kanker payudara di antaranya adanya benjolan pada payudara, perubahan bentuk payudara dan terdapat bercak atau sisik merah pada puting susu (Kabel & Baali, 2015) . Kanker payudara adalah Penyebab utama kematian kedua bagi wanita (setelah kanker paru-paru) (Asri et al., 2016) . Lebih dari satu juta kasus dan hampir 600.000 kematian kanker payudara terjadi di seluruh dunia setiap tahunnya. Ini mewakili 20,5% dari total kasus kanker umum di seluruh dunia dan persentase 36,2 % kanker payudara

diukur sebagai yang dapat diobati (Durgalakshmi & Vijayakumar, 2015). Banyak faktor resiko kanker payudara yang berada di luar kendali (tidak dapat dicegah) seperti usia, riwayat keluarga, dan riwayat kesehatan. Namun, ada beberapa faktor resiko yang dapat dikendalikan (dicegah), seperti berat badan, aktivitas fisik, dan konsumsi alkohol (Breast cancer risk factors: Preventable and non-preventable, 2012).

Beberapa aplikasi data mining telah digunakan dalam beberapa tahun terakhir untuk prediksi kanker dan prognosis (Kate & Nadig, 2017) Prognosis adalah bidang kedokteran yang berhubungan dengan analisis kelangsungan hidup pasien(Delen et al., 2005). Khusus untuk kanker payudara, peneliti telah menggunakan berbagai metode data mining untuk memprediksi kerentanan, diagnosis, kekambuhan dan ketahanan hidup(Kate & Nadig, 2017) . Ketahanan Hidup umumnya didefinisikan sebagai pasien yang masih hidup selama jangka waktu tertentu setelah diagnosis penyakit (Delen et al., 2005). Prediksi yang akurat tentang kemungkinan ketahanan hidup penderita kanker payudara dapat memungkinkan dokter dan penyedia layanan kesehatan membuat keputusan yang lebih tepat mengenai perawatan pasien (Kate & Nadig, 2017).

Beberapa peneliti telah memprediksi kanker payudara dengan teknik data mining di antaranya adalah yang dilakukan oleh K.M.Al Aidaroos, A.A.Bakar dan Z.Othman (Medical Data Classification with Naive Bayes Approach, 2012) dengan membandingkan metode klasifikasi *Naive Bayes* dengan metode klasifikasi lainnya yaitu *Logistic Regresion, Decision Tree dan Neural Network* menggunakan 15 dataset medical salah satunya adalah dataset kanker payudara, Naive Bayes mendapatkan akurasi tertinggi yaitu 97.30% dan AUC 0.99.

Penelitian mengenai prediksi kanker payudara juga dilakukan oleh Shweta, K Shika , A dan Sunita S (Kharya et al., 2014) prediksi kanker payudara dengan menggunakan *Naive Bayes Classifier*, hasil eksperimen mengungkapkan bahwa NBC adalah pendekatan yang efisien untuk ekstraksi yang signifikan dari kumpulan data kanker payudara dengan Akurasi maksimal mencapai 93% . Pada penelitian yang dilakukan oleh Oman Somantri, Mohammad Khambali dengan menerapkan algoritma genetika pada Naive Bayes untuk klasifikasi kategori cerpen online memeperlihatkan adanya peningkatan akurasi dari 78,59% menjadi 84,29%. Pada penelitian yang dilakukan oleh Dwi Cahya Putri Buani (STMIK Nusa Mandiri Jakarta, 2018)

memanfaatkan seleksi fitur menggunakan Algoritma Genetika untuk meningkatkan akurasi dari prediksi algoritma Naive Bayes. dari penerapan metode tersebut dapat dihasilkan predikisi untuk penyakit hepatitis memiliki akurasi sebesar 96, 77%, hasil prediksi ini meningkat dari penelitian sebelumnya dengan menggunakan algoritma yang sama yaitu algoritma naïve bayes tanpa dilakukan seleksi fitur hasil akurasinya adalah 83, 71%, selisih hasil penelitian sebelumnya dengan penelitian ini adalah 13.06%.

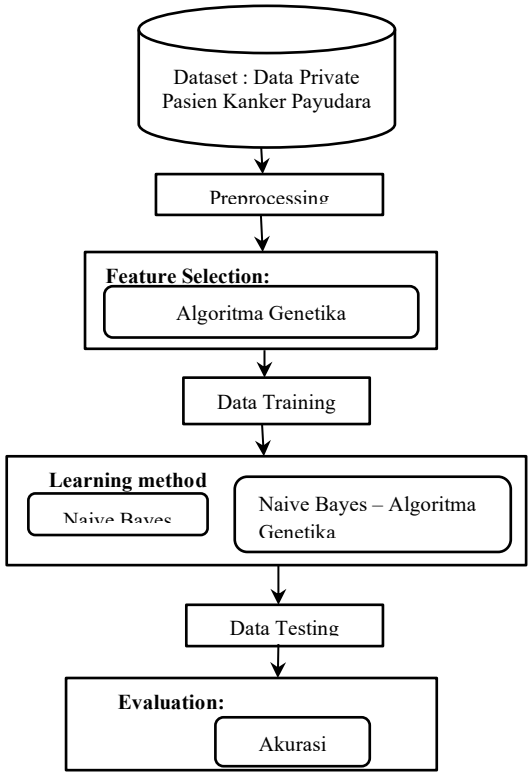
Metode Data Mining yang di gunakan untuk klasifikasi diantaranya adalah *Artificial Neural Network* (ANN), *Support Vectore Machine*, *Decision Tree, Naive Bayes* dan *Logistic Regression*(Wu et al., 2008). *Naive Bayes Classifier* adalah Pengklasifikasi statistik yang didasarkan pada teorema bayes yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas (Jiawei, et al., 2012) . Naive Bayes sangat sederhana dan efisien namun sangat sensitif terhadap fitur maka dari itu pemilihan fitur yang sesuai sangat di perlukan karena fitur-fitur yang tidak relevan dapat mengurangi tingkat akurasi(Chen et al., 2009). *Naive Bayes Classifier* akan bekerja lebih efektif jika dikombinasikan dengan beberapa prosedur pemilihan atribut (Witten, et al., 2011). Tujuan dari pemilihan atribut ini untuk mengurangi jumlah fitur serta menghilangkan fitur yang tidak relevan, pemilihan atribut dapat meningkatkan akurasi dan mempercepat algoritma data mining (Harb & Desuky, 2014).

Algoritma Genetika adalah algoritma pencarian stokastik, paralel, heuristik yang terinspirasi oleh prinsip dasar seleksi alam dimana seleksi alam merupakan proses biologis di mana individu yang lebih kuat kemungkinan besar menjadi pemenang dalam lingkungan yang bersaing(Prasetio, 2020). Dalam penelitian ini Algoritma Genetika digunakan untuk Seleksi Fitur pada Algoritma Naive *Bayes* sehingga dapat meningkatkan akurasi klasifikasi ketahanan hidup penderita kanker payudara.

METODE PENELITIAN

Data

Penelitian ini menggunakan data private pasien kanker payudara. Dataset ini berjumlah 100 record dan 11 atribut. Variabel yang ada dalam data ini diantaranya adalah Umur, Status Pernikahan, Ras, Askes, Stadium, Grade, Diagnosa Utama, Behavior, Riwayat Keluarga, Tindakan, Komplikasi dan Ketahanan Hidup yang akan dijadikan sebagai label.



Gambar 1: Metode yang diusulkan

Tahap Preprocessing

Dilakukan penghapusan data yang kosong atau missing value sebanyak 38 data, data yang terkumpul sebanyak 100 record. Dari Variabel yang ada Kemudian di tentukan variabel- variabel yang akan di gunakan sebagai variabel prediktor dan variabel target. Variabel Umur, Status Pernikahan, Ras, Askes, Stadium, Grade, Diagnosa Utama, Behavior, Riwayat Keluarga, Tindakan dan Komplikasi merupakan variabel prediktor , sedangkan untuk Ketahanan Hidup akan di jadikan Variabel Target.

Naive Bayes

Naive Bayes Classifier (Jiawei, et al., 2012) adalah Pengklasifikasian statistik yang didasarkan pada teorema bayes yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. Bayesian Classification terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database yang besar. Bentuk umum teorema bayes adalah sebagai berikut

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)} \quad .....(2.1)$$

Dimana :

X = data dengan kelas yang belum diketahui

H = Hipotesa data X merupakan suatu kelas spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X (Posterior Probability)

P(H) = probabilitas hipotesis H (prior probability)

Algoritma Genetika

Algoritma Genetika adalah algoritma pencarian stokastik, paralel, heuristik yang terinspirasi oleh prinsip dasar seleksi alam dimana seleksi alam merupakan proses biologis di mana individu yang lebih kuat kemungkinan besar menjadi pemenang dalam lingkungan yang bersaing. Algoritma genetika

menyediakan kerangka kerja untuk mempelajari efek dari faktor-faktor yang diilhami secara biologis seperti pemilihan pasangan, reproduksi,mutasi, dan persilangan informasi genetik. Tiga operator digunakan oleh algoritma genetika yaitu Selection, Crossover dan Mutation(Prasetio, 2020).

HASIL DAN PEMBAHASAN

Pre Processing

Dilakukan penghapusan data yang kosong atau missing value sebanyak 38 data, data yang terkumpul sebanyak 100 record. Dari Variabel yang ada Kemudian di tentukan variabel- variabel yang akan di gunakan sebagai variabel prediktor dan variabel target. Variabel Umur, Status Pernikahan, Ras, Askes, Stadium, Grade, Diagnosa Utama, Behavior, Riwayat Keluarga, Tindakan dan Komplikasi merupakan variabel prediktor , sedangkan untuk Ketahanan Hidup akan di jadikan Variabel Target.

Eksperimen pada Metode Naive Bayes

Eksperimen yang di lakukan pada Algoritma Naive Bayes tanpa fitur seleksi adalah dengan percobaan menggunakan number of validation 2-10 .

Tabel 1; Hasil Eksperimen Metode Naive Bayes - Algoritma Genetika

Number of validation	Akurasi
2	79%
3	76,05%
4	85%
5	86%
6	84,13%
7	83,27%
8	83,89%
9	81,82%
10	**86%**

Tabel 2; Hasil Eksperimen Metode Naive Bayes dan Naïve Bayes - Algoritma Genetika

Model	Akurasi
Naïve Bayes	**86%**
Naïve Bayes – Algoritma Genetika	**90%**

SIMPULAN DAN SARAN

Berdasarkan hasil percobaan dengan menggunakan Naive Bayes dan Naive Bayes Algoritm Genetika diperoleh akurasi Naive Bayes sebesar 86 % dan sedangkan akurasi Naive Bayes-Algoritma Genetika sebesar 90%. Hal tersebut membuktikan bahwa Algoritma Genetika yang digunakan untuk Pembobotan atribut dapat meningkatkan akurasi Naive Bayes. Kenaikan akurasi meningkat sebanyak 4 %.

DAFTAR PUSTAKA

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using Machine Learning Algorithms for

Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, *83*(Fams), 1064–1069. https://doi.org/10.1016/j.procs.2016.04.224

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, *36*(3 PART 1), 5432–5435. https://doi.org/10.1016/j.eswa.2008.06.054

Delen, D., Walker, G., & Kadam, A. (2005). *Predicting breast cancer* survivability *: a comparison of three data mining methods*. https://doi.org/10.1016/j.artmed.2004.07.002

Durgalakshmi, B., & Vijayakumar, V. (2015). Progonosis and modelling of breast cancer and its growth novel naive bayes. *Procedia* Computer *Science*, *50*, 551–553. https://doi.org/10.1016/j.procs.2015.04.102

Harb, H. M., & Desuky, A. S. (2014). Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization. *International Journal of Computer Applications*, *104*(5), 975–8887. https://doi.org/10.5120/18197-9118

Kabel, A. M., & Baali, F. H. (2015). Breast Cancer : Insights into Risk Factors , Pathogenesis , Diagnosis and Management. *Insights into Risk Factors , Pathogenesis , Diagnosis and* Managemen, *3*(2), 28–33. https://doi.org/10.12691/jcrt-3-2-3

Kate, R. J., & Nadig, R. (2017). Stage-specific predictive models for breast cancer survivability. *International Journal of Medical Informatics*, *97*, 304–311. https://doi.org/10.1016/j.ijmedinf.2016.11.001

Kharya, S., Agrawal, S., & Soni, S. (2014). Naive Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer. *International Journal of Computer Applications*, *92*(10), 26–31. https://doi.org/10.5120/16045-5206

Prasetio, R. T. (2020). Seleksi Fitur dan Optimasi Parameter k-NN Berbasis Algoritma Genetika pada Dataset Medis. *Jurnal Responsif*, *2*(2), 213–221.

STMIK Nusa Mandiri Jakarta, D. C. P. B. (2018). Prediksi Penyakit Hepatitis Menggunakan Algoritma Naïve Bayes Dengan Seleksi Fitur Algoritma Genetika. *Evolusi : Jurnal Sains Dan Manajemen*, *6*(2), 1–5. https://doi.org/10.31294/evolusi.v6i2.4381

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. In *Knowledge and Information Systems* (Vol. 14, Issue 1). https://doi.org/10.1007/s10115-007-0114-2