Jurnal Teknik Komputer AMIK BSI
Volume 7, No.1, Januari 2021
P-ISSN 2442-2436, E-ISSN: 2550-0120
Akreditasi Ristekdikti, No: 36/E/KPT/2019 (Sinta 4)
DOI: 10.31294/jtk.v4i2

Komparasi Algoritma *Naive Bayes*, *Decision Tree* dan *Support Vector Machine* untuk Prediksi Penyakit Kanker Payudara

Lusa Indah Prahartiwi[1], Wulan Dari[2]

[1,2]STMIK Nusa Mandiri
[1]e-mail: lusa.lip@nusamandiri.ac.id
[2]e-mail: wulan.wld@nusamandiri.ac.id

Diterima	Direvisi	Disetujui
22-10-2020	02-11-2020	23-01-2021

Abstrak - Kanker payudara merupakan kanker paling umum pada wanita di seluruh dunia dengan menyumbang 25,4% dari total jumlah kasus baru yang didiagnosis pada tahun 2018. Kanker adalah sekelompok besar penyakit yang dapat dimulai di hampir semua organ atau jaringan tubuh ketika sel abnormal tumbuh tak terkendali, melampaui batas biasanya untuk menyerang bagian tubuh yang berdekatan dan/atau menyebar ke organ lain. Penyakit kanker payudara dapat diprediksi dengan pengetahuan data mining. Data mining dapat menemukan korelasi, pola, dan tren baru yang bermakna dengan memilah-milah data dalam jumlah besar yang disimpan dalam repositori, menggunakan teknologi pengenalan pola serta teknik statistik dan matematika. Penelitian ini membandingkan performa Algoritma *Naive Bayes*, *Decision Tree* dan *Support Vector Machine* untuk memprediksi penyakit kanker payudara. Dataset yang digunakan adalah data sekunder Breast Cancer Coimbra yang diambil dari UCI Repository. Hasil dari penelitian ini menunjukan bahwa Algoritma *Support Vector Machine* menghasilkan tingkat Accuracy tertinggi yaitu sebesar 74,29% dibandingkan dengan Algoritma *Naive Bayes* dan *Decision Tree*.

Kata Kunci: Prediksi, Algoritma *Naive Bayes*, *Decision Tree*, *Support Vector Machine*

Abstract *- Breast cancer was the most common cancer in women worldwide, contributing 25.4% of the total number of new cases diagnosed in 2018. Cancer is a large group of diseases that can start in almost any organ or tissue of the body when abnormal cells grow uncontrollably, go beyond their usual boundaries to invade adjoining parts of the body and/or spread to other organs. Breast cancer can be predicted with data mining knowledge. Data mining can discover meaningful new correlations, patterns and trends by sorting through the large amounts of data stored in repositories, using pattern recognition technology as well as statistical and mathematical techniques. This study compares the performance of the Naive Bayes Algorithm, Decision Tree and Support Vector Machine to predict breast cancer. The dataset used is secondary data from Breast Cancer Coimbra taken from the UCI Repository. The results of this study indicate that the Support Vector Machine algorithm produces the highest level of accuracy, which is 74.29% compared to the Naive Bayes algorithm and the Decision Tree.*

Keywords: Prediction, Algoritma Naive Bayes, Decision Tree, Support Vector Machine

PENDAHULUAN

World Cancer Research Fund mencatat Kanker payudara merupakan kanker paling umum pada wanita di seluruh dunia dengan menyumbang 25,4% dari total jumlah kasus baru yang didiagnosis pada tahun 2018.

World Health Organization menyatakan bahwa kanker adalah sekelompok besar penyakit yang dapat dimulai di hampir semua organ atau jaringan tubuh ketika sel abnormal tumbuh tak terkendali, melampaui batas biasanya untuk menyerang bagian tubuh yang berdekatan dan/atau menyebar ke organ lain. Proses terakhir ini disebut metastasis dan merupakan penyebab utama kematian akibat kanker. Neoplasma dan tumor ganas adalah nama umum lainnya untuk kanker. Beban kanker terus meningkat secara global, memberikan tekanan fisik, emosional dan finansial yang luar biasa pada individu, keluarga, komunitas dan sistem kesehatan. Banyak sistem kesehatan di negara berpenghasilan rendah dan

menengah paling tidak siap untuk menangani beban ini, dan sejumlah besar pasien kanker secara global tidak memiliki akses ke diagnosis dan pengobatan berkualitas tepat waktu. Di negara-negara di mana sistem kesehatannya kuat, tingkat kelangsungan hidup berbagai jenis kanker meningkat berkat deteksi dini yang dapat diakses, pengobatan berkualitas, dan perawatan penyintas.

Manusia dibanjiri dengan data di banyak bidang. Sayangnya, data berharga ini, yang menghabiskan jutaan perusahaan untuk mengumpulkan dan menyusun, mendekam di gudang dan repositori. Masalahnya adalah bahwa tidak tersedia cukup analis manusia terlatih yang terampil dalam menerjemahkan semua data menjadi pengetahuan (Larose, 2005).

Penyakit kanker payudara dapat diprediksi dengan pengetahuan data mining. Data mining adalah proses menemukan pola yang berwawasan, menarik, dan baru, serta model deskriptif, dapat dipahami, dan prediktif dari data yang berskala besar (Ahmed et al., 2015). Data mining dapat menemukan korelasi, pola, dan tren baru yang bermakna dengan memilah-milah data dalam jumlah besar yang disimpan dalam repositori, menggunakan teknologi pengenalan pola serta teknik statistik dan matematika (Larose, 2005). Data mining diperkirakan akan menjadi "salah satu perkembangan paling revolusioner dalam dekade mendatang" (Larose, 2006).

Seperti halnya dengan teknologi informasi baru, penambangan data yang dilakukan dengan sedikit pengetahuan sangat berbahaya ketika harus menerapkan model yang kuat berdasarkan kumpulan data yang besar. Misalnya, analisis yang dilakukan pada data yang tidak diproses dapat menyebabkan kesimpulan yang salah, atau analisis yang tidak tepat dapat diterapkan pada kumpulan data yang memerlukan pendekatan yang sama sekali berbeda, atau model dapat diturunkan yang dibangun di atas asumsi yang sepenuhnya spekulatif. Jika diterapkan, kesalahan dalam analisis ini dapat menyebabkan kegagalan yang sangat mahal (Larose, 2005).

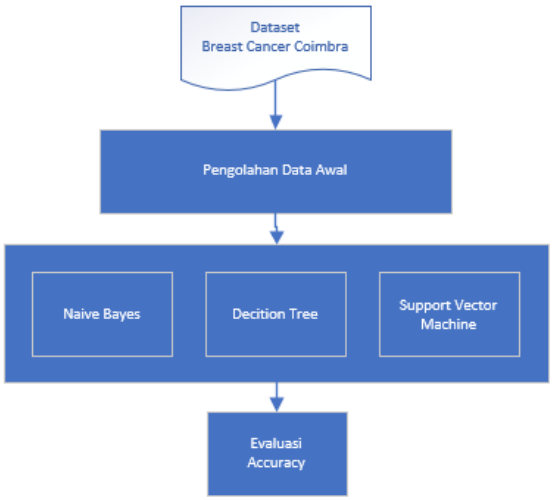
Salah satu peran data mining adalah prediksi. Prediksi mirip dengan klasifikasi dan estimasi, hanya saja untuk prediksi, hasilnya ada di masa depan (Larose, 2005).

Penelitian terdahulu telah dilakukan (Dumitru, 2009) untuk mendiagnosis penyakit kanker payudara menggunakan Algoritma *Naive Bayes*. Dataset yang digunakan adalah Wisconsin Prognostic Breast Cancer. Hasil penelitian menunjukan bahwa algortima *Naive Bayes* memiliki performa yang setara dengan teknik pembelajaran mesin lainnya dengan upaya komputasi yang rendah dan kecepatan tinggi.

Beberapa teknik pembelajaran mesin lain untuk memprediksi penyakit kanker payudara di antaranya Algoritma Gain Ratio (Aisyah & Sulistyo, 2016), Jaringan Syaraf Tiruan (Wibisono et al., 2019), dan *Support Vector Machine* (Chazar & Widhiaputra, 2020).

Penelitian ini akan membandingkan performa Algoritma *Naive Bayes*, *Decision Tree* dan *Support Vector Machine* untuk memprediksi penyakit kanker payudara

METODOLOGI PENELITIAN



Sumber: Hasil Penelitian (2020)

Gambar 1. Metodologi Penelitian

Berdasarkan Gambar 1, Tahapan metode penelitian yang dilakukan yaitu:

1. Pengumpulan dataset.
   Dataset berupa data sekunder Breast Cancer Coimbra yang diambil dari UCI Repository. Dataset terdiri dari 10 atribut yaitu Age, B MI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, dan Clasification. Dataset Breast Cancer Coimbra terdiri dari 116 record.
2. Pengolahan Data Awal.
   Tidak ada atribut yang perlu dihapus pada dataset Breast Cancer Coimbra, dikarenakan semua atribut diperlukan dalam pengolahan data.
3. Metode yang diusulkan.
   Pada penelitian ini akan diusulkan Algoritma *Naive Bayes*, Algoritma *Decision Tree* dan Algoritma *Support Vector Machine*.
4. Eksperimen.
   Eksperimen dilakukan dengan mengolah data Breast Cancer Coimbra menggunakan Algoritma *Naive Bayes*, *Decision Tree* dan *Support Vector Machine*. Kemudian hasil

eksperimen dari ketiga Algoritma tersebut akan dikomparasi untuk mengetahui Algoritma mana yang memiliki performance terbaik

5. Evaluasi.
   Metode yang diusulkan akan diuji tingkat akurasinya dengan Confusion Matrix dan Receiver Operating Characteristic Curve (ROC Curve

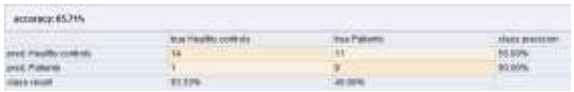
HASIL DAN PEMBAHASAN

1. Eksperimen

Untuk mendapatkan hasil prediksi yang lebih akurat, penulis melakukan perbandingan dengan beberapa eksperimen sebagai berikut:

a. Algoritma *Naive Bayes*

Eksperimen pertama dilakukan dengan menggunakan Algoritma *Naive Bayes*. Hasil eksperimen diperoleh Accuracy sebesar 65,71% yang dapat dilihat pada Gambar 2.



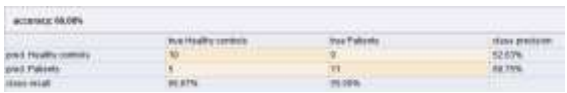
Sumber: Hasil Penelitian (2020)

Gambar 2. Hasil Eksperimen dengan Algoritma *Naive Bayes*

Berdasarkan Gambar 1, diketahui bahwa jumlah True Positive (TP) sebanyak 14 record yang diklasifikasikan sebagai Healthy Controls dan True Negative (TN) sebanyak 9 record yang diklasifikasikan sebagai Patients. Jumlah False Positive (FP) sebanyak 11 record yang diprediksi sebagai Healthy controls tetapi ternyata Patients dan 1 record False Negative(FN) yang diprediksi Patient tetapi ternyata Healthy controls.

b. Algoritma *Decision Tree*

Eksperimen kedua dilakukan dengan Algoritma *Decision Tree*. Accuracy sebesar 60% didapat dari hasil eksperimen melalui Algoritma *Decision Tree* yang dapat dilihat pada Gambar 3.



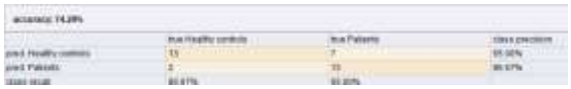
Sumber: Hasil Penelitian (2020)

Gambar 3. Hasil Eksperimen dengan Algoritma Decition Tree

Berdasarkan Gambar 3, diketahui bahwa jumlah True Positive (TP) sebanyak 10 record yang diklasifikasikan sebagai Healthy Controls dan True Negative (TN) sebanyak 11 record yang diklasifikasikan sebagai Patients. Jumlah False Positive (FP) sebanyak 9 record yang diprediksi sebagai Healthy controls tetapi ternyata Patients dan 5 record False Negative(FN) yang diprediksi Patient tetapi ternyata Healthy controls

c. Algoritma *Support Vector Machine* (SVM)

Eksperimen ketiga dilakukan melalui Algortima *Support Vector Machine*(SVM). Hasil ekperimen dapat dilihat pada Gambar 4.



Sumber: Hasil Penelitian (2020)

Gambar 4. Hasil Eksperimen dengan Algoritma *Support Vector Machine* (SVM)

Berdasarkan Gambar 4, diketahui bahwa nilai Accuracy yang dihasilkan sebesar 74,29%. Jumlah True Positive (TP) sebanyak 13 record yang diklasifikasikan sebagai Healthy Controls dan True Negative (TN) sebanyak 13 record yang diklasifikasikan sebagai Patients. Jumlah False Positive (FP) sebanyak 7 record yang diprediksi sebagai Healthy controls tetapi ternyata Patients dan 2 record False Negative(FN) yang diprediksi Patient tetapi ternyata Healthy controls

2. Evaluasi

Hasil eksperimen Algoritma *Naive Bayes*, Algoritma *Decision Tree* dan Algoritma *Support Vector Machine* (SVM) akan dievaluasi dengan menggunakan Confusion Matrix dan Receiver Operating Characteristic Curve (ROC Curve).

a. Confusion Matrix

Analisa hasil komparasi dilakukan dengan membandingkan hasil pengujian Confusion Matrix dari ketiga Algoritma.

Tabel 1. Hasil Uji Komparasi Algoritma *Naive Bayes*, *Decision Tree* dan SVM

Algoritma	Accuracy	Sensitivity	Specificity	PPV	NPV
Naive Bayes	65,71%	93,33%	45,00%	56,00%	90,00%
Decision Tree	60,00%	66,67%	55,00%	52,63%	68,75%
SVM	74,29%	86,67%	65,00%	65,00%	86,67%

Sumber: Hasil Penelitian (2020)

Berdasarkan Tabel 1, Algoritma *Support Vector Machine* menghasilkan tingkat Accuracy tertinggi yaitu sebesar 74,29% dibandingkan dengan Algoritma *Naive Bayes* yang memperoleh Accuracy sebesar 65,71%. Sedangkan Algoritma Decition Tree mengahasilkan Accuracy terendah bila dibandingkan dengan kedua Algoritma lainnya yaitu hanya sebesar 60%. Algoritma *Naive Bayes* menghasilkan nilai Sensitivity dan NPV tertinggi yaitu sebesar 93,33% dan 90%.

b. Kurva ROC

Kurva ROC dari ketiga Algoritma disajikan pada Gambar 5.

Sumber: Hasil Penelitian (2020)

Gambar 5. Kurva ROC Algoritma *Naive Bayes*, *Decision Tree* dan *Support Vector Machine* (SVM)

KESIMPULAN

Penyakit kanker payudara dapat diprediksi dengan menerapkan pengetahuan data mining. Eksperimen dilakukan dengan membandingkan Algoritma *Naive Bayes*, *Decision Tree* dan *Support Vector Machine* untuk mengetahui performa dari ketiga Algoritma tersebut. Dari hasil eksperimen diperoleh Algoritma *Support Vector Machine* menghasilkan tingkat Accuracy tertinggi yaitu sebesar 74,29%. Hasil uji Accuracy, Sensitivity, PPV dan NPV pada Algoritma *Decision Tree* menunjukan bahwa Algoritma *Decision Tree* memiliki performa terburuk. Dari rata-rata hasil uji Accuracy, Sensitivity, Specificity, PPV, dan NPV Algoritma *Support Vector Machine* menghasilkan nilai tertinggi sehingga Algoritma *Support Vector Machine*(SVM) memiliki performa terbaik dibandingkan dengan algortima *Naive Bayes* dan *Decision Tree*

REFERENSI

Ahmed, A. M., Rizaner, A., Ulusoy, A. H., Silva, L. L. A., Silva, L. L. A., Kaur, P., Singh, M., Josan, G. S., Zaki, M. J., Carmona, C., Castillo, G., Millán, E., Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., Erven, G. Van, Marbouti, F., … Magalhães, J. (2015). DATA MINING Fundamental Concepts and Algorithms. In *Procedia Computer Science* (Vol. 35, Issue 6). https://doi.org/10.1016/j.sbspro.2013.10.240

Aisyah, B., & Sulistyo, Y. (2016). *Klasifikasi Kanker Payudara Menggunakan Algoritma Gain Ratio*. *8*(2), 2–5.

Chazar, C., & Widhiaputra, B. E. (2020). *INFORMASI (Jurnal Informatika dan Sistem Informasi) Volume 12 No.1 / Mei/ 2020*. *12*(1), 67–80.

Dumitru, D. (2009). *Prediction of recurrent events in breast cancer using the Naive Bayesian classification*. *36*(2), 92–96.

Larose, D. T. (2005). *An Introduction to Data Mining*.

Larose, D. T. (2006). Data Mining Methods and Models. In *Data Mining Methods and Models*. https://doi.org/10.1002/0471756482

Wibisono, G., Hermawan, A., Studi, P., Teknologi, M., Yogyakarta, U. T., Gejala, F. P., Kanker, P., Dengan, P., & Jaringan, P. (2019). *FAKTOR-FAKTOR PENENTU GEJALA PENYAKIT*. *1*(1), 1–6.