Cyberspace: Jurnal Pendidikan Teknologi Informasi
Volume 5, Nomor 1, Maret 2021, 22-30

ARISON OF SUPPORT VECTOR MACHINE AND DECISION TREE METHODS IN THE CLASSIFICATION OF BREAST CANCER

Helmi Imaduddin[1], Brian Aditya Hermansyah[2] and Frischa Aura Salsabilla B[3]

[1]Informatika, Fakultas Komunikasi dan Infromatika, Universitas Muhammadiyah Surakarta
[2]Informatika, Fakultas Komunikasi dan Infromatika, Universitas Muhammadiyah Surakarta
[3]Informatika, Fakultas Komunikasi dan Infromatika, Universitas Muhammadiyah Surakarta
E-mail: helmi.imaduddin@ums.ac.id, brian.adityaherman@gmail.com, frischaaura10@gmail.com

Abstract

One of the most dangerous cancers in the world is breast cancer. This cancer occurs in many women, in some cases this cancer can also affect men, but it is very rare. The effects of this cancer are very dangerous for humans, in the worst case it can lead to death. So that serious prevention is needed against this cancer. One prevention can be done by early detection.

This study aims to implement machine learning methods to detect breast cancer in women. The algorithms used are Support Vector Machine (SVM) and Decision Tree (DT). After classifying the data provided, a comparison is made to find out which machine learning method has the best performance. The data used comes from the Gynecology Department of the University Hospital Center of Coimbra (CHUC), and can be downloaded for free on the UCI repository website. The results of this study indicate that the SVM algorithm with feature selection obtains the best classification results by obtaining an accuracy of 87.5%, a sensitivity of 90%, and a specificity of 85%. Thus this research obtains good results to be able to help provide solutions to detect breast cancer.

Keywords: *Breast Cancer, SVM, Decision Tree, Machine Learning*

Abstrak

Salah satu kanker yang paling berbahaya di dunia adalah kanker payudara. Kanker ini banyak terjadi pada wanita, dalam beberapa kasus kanker ini juga bisa menyerang pria namun sangat jarang. Efek dari kanker ini sangat berbahaya bagi manusia, dengan kasus terburuk bisa menyebabkan kematian. Sehingga diperlukan pencegahan secara serius terhadap kanker ini. Salah satu pencegahan dapat dilakukan dengan pendeteksian secara dini.

Penelitian ini bertujuan untuk mengimplementasian metode machine learning untuk mendeteksi kanker payudara pada wanita, Adapun algoritma yang dipakai yaitu Support Vector Machine (SVM) dan Decision Tree (DT). Setelah melakukan klasifikasi pada data yang diberikan selanjutnya dilakukan perbandingan untuk mengetahui metode machine learning yang memiliki performa terbaik. Data yang digunakan berasal dari Gynaecology Department of the University Hospital Centre of Coimbra (CHUC), dan dapat diunduh secara gratis di website UCI repository.

Hasil dari penelitian ini menunjukan bahwa algoritma SVM dengan seleksi fitur memperoleh hasil klasifikasi terbaik dengan memperoleh akurasi sebesar 87,5%, sensitivitas 90%, dan spesitifitas 85%. Dengan demikian penelitian ini memperoleh hasil yang baik untuk dapat membantu memberikan solusi untuk mendeteksi penyakit kanker payudara.

Kata Kunci: *Kanker Payudara, SVM, Decision Tree, Machine Learning*

1. PENDAHULUAN

Perkembangan teknologi *machine learning* sekarang sangat pesat, banyak bidang sudah menggunakan *machine learning* untuk meringankan pekerjaan ataupun memberikan solusi. Penggunaan *machine learning* dapat kita rasakan pada keseharian kita seperti pada video rekomendasi youtube, analisis sentimen pada sosial media dan lain sebagainya. Penggunaan teknologi *machine learning* dapat diterapkan dibanyak bidang, termasuk juga di bidang kesehatan. Dengan *machine learning* kita dapat menentukan karakteristik penyakit kanker berdasarkan gambar yang diperoleh dari hasil scan, ataupun kita dapat menentukan seberapa parah sebuah penyakit kanker dilihat dari bentuk dan kondisinya.

Kanker payudara merupakan satu diantara banyak kanker yang paling berbahaya. Kanker payudara bisa terjadi pada manusia, baik pria dan wanita. Pada pria jumlah kasus yang ditemukan sangat sedikit, berbeda pada wanita yang jumlah kasusnya cenderung lebih banyak [1]. Kanker payudara memiliki jumlah penderita paling banyak kedua di dunia setelah kanker kulit. Dari semua wanita yang menderita kanker, sebanyak 21% adalah penderita kanker payudara. Kanker ini menyebabkan lebih dari 60% kematian bagi penderitanya, Oleh sebab itu diperlukan metode khusus untuk mendeteksi kanker payudara [2].

Penelitian ini bertujuan untuk membantu memberikan solusi bagi masalah kanker payudara dengan melakukan implementasikan *machine learning* (ML) untuk mengklasifikasikan data dari kanker payudara. Penelitian ini menggunakan dua algoritma, yaitu *Support Vector Machine* (SVM) dan *Decision Tree* (DT) untuk klasifikasi data kanker payudara, kemudian membandingkan hasil dari keduanya untuk menentukan algoritma yang memiliki performa lebih baik. Adapaun kriteria penilaian yang digunakan untuk mengukur performa klasifikasinya adalah akurasi, spesifisitas, dan sensitivitas yang didapat dengan metode *confusion matrix*.

2. KAJIAN PUSTAKA

2.1. Algoritma *Support Vector Machine*

Algoritma *Support Vector Machine* (SVM) merupakan salah satu dari algoritma ML dengan pendekatan *supervised*. Metode ini melakukan klasifikasi dengan membagi data kedalam dua kelas menggunakan garis vektor, yang dinamakan *hyperplane* [3]. *Hyperplane* paling baik dapat dicari dengan menggunakan cara mengukur jarak kemudian mencari titik maksimalnya, proses pencarian *hyperplane* ini merupakan inti dari algoritma SVM. Algoritma SVM kini banyak diterapkan di berbagai bidang, karena algoritma ini mempunyai performa yang lebih bagus dibandingkan metode yang lainnya. Pada penerapan klasifikasi menggunakan SVM tidak semua data digunakan untuk proses klasifikasi [4].

Tingkat akurasi pada model SVM bergantung pada *kernel* dan parameter yang digunakan. Dilihat dari karakteristik yang dimiliki, SVM terbagi kedalam dua bagian,

yaitu SVM linier dan SVM non-linier. Pada SVM linier proses pemisahan data dilakukan secara linier, sedangakan SVM non-linier prosesnya menggunakan *kernel trick* pada ruang berdimensi tinggi [5].

2.2. Algoritma *Decision Tree*

Decision Tree (DT) adalah sebuah metode *unsupervised* yang berbentuk seperti pohon. Pohon tersebut dapat mengelompokkan data yang ukurannya besar menjadi data dengan ukuran yang kecil [6]. Dalam DT setiap simpul menyatakan pengujian sebuah atribut yang dimiliki, setiap cabang menunjukan hasil dari pegujian yang dilakukan, pada simpul daun menunjukan kelas targetnya. Simpul yang paling atas disebut simpul akar, dan pada simpul akar ini awal mula terbentuknya proses *Decision Tree*. Simpul daun memberikan informasi terkait keputusan akhir atau target kelas dari peroses DT. Alur dari metode DT dimulai simpul akar menuju kebawah sampai ke simpul daun yang menentukan prediksi kelas [7].

Pada metode DT pemilihan atribut untuk menjadi simpul akar, didasarkan pada atribut yang memiliki nilai *Information Gain* (IG) tertinggi. IG adalah suatu ukuran korelasi yang dapat menggambarkan hubungan ketergantungan antara dua peubah. Sebelum mencari nilai *IG* Langkah yang pertama dilakukan terlebih dahulu adalah menentukan nilai *entropy*. Cara untuk menghitung nilai *entropy* dapat dilakukan dengan rumus 1 berikut.

$$Entropy(s) = \sum_{i=1}^{n} -pi * log_2\, pi \quad (1)$$

Keterangan:
S = Himpunan kasus
n = Jumlah partisi S
pi = Proporsi dari Si terhadap S

Setelah mendapatkan nilai *entropy*, selanjutnya pemilihan atribut dilakukan dengan nilai IG terbesar. Untuk menghitung IG bisa menggunakan rumus 2 berikut.

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

Keterangan:
S = Himpunan kasus
A = Atribut
n = Jumlah partisi atribut A
|Si| = Jumlah kasus pada partisi ke-i
|S| = Jumlah kasus dalam S

2.3. Machine learning

Machine learning (ML) adalah sebuah cabang dari kecerdasan buatan (*Artificial Intelligence*), ML dapat membuat mesin belajar tanpa diprogram secara eksplisit. Dengan menggunakan ML kita dapat membuat sebuah model yang dapat melakukan pembelajaran dari sebuah dataset. Pada perkembangannya ML sudah sering kali

diterapkan pada kehidupan nyata untuk membantu meringankan pekerjaan manusia, seperti ketika kita menonton video youtube, maka akan ada video rekomendasi yang diberikan kepada kita, itu merupakan salah satu contoh dari penggunaan ML. Selain itu ML juga bisa diterapkan dibidang lain seperti kedokteran untuk menentukan jenis penyakit, marketplace untuk rekomendasi produk, media sosial untuk mengukur sentimen, dan lain sebagainya [8].

Cara kerja ML diadopsi dari cara manusia berfikir, yaitu dengan cara menggunakan contoh yang diberikan kemudian menentukan pola dari contoh tersebut. Dengan cara tersebut mesin dapat menjawab atau menentukan solusi untuk masalah yang diinginkan. Pada proses perkembangannya, ML memiliki 4 jenis metode, yaitu:

a. *Supervised*

Supervised adalah algoritma ML yang dapat melakukan pembelajaran pada data yang sudah diberikan label, metode ini biasanya digunakan untuk melakukan proses klasifikasi.

b. Unsupervised

Unsupervised merupakan metode ML yang dapat belajar dengan data tanpa label, artinya data unstuktur dapat digunakan pada metode ini. Metode *unsupervised* biasanya digunakan untuk melakukan *clustering*.

c. Semi-*supervised*

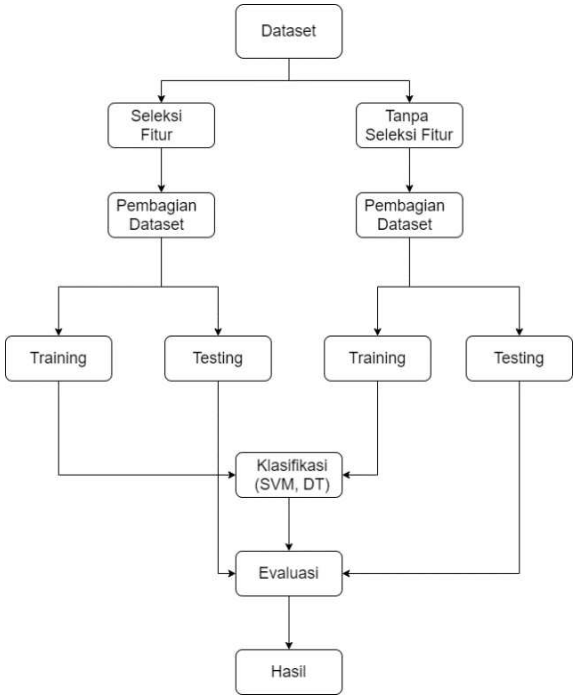
Semi-*supervised* adalah metode yang digunakan untuk melakukan pembelajaran dengan cara menggabungkan metode *supervised* dan un*supervised*. Pada prosesnya metode ini dapat bekerja dengan data yang berlabel dan data tanpa label. Biasanya metode ini digunakan ketika data berlabel jumlahnya sedikit atau susah untuk menemukan data berlabel.

d. Reinforcement

Reinforcement merupakan algoritma yang memiliki kelebihan dalam berinteraksi dengan proses belajar yang dilaksanakan, algoritma ini menambahkan poin disaat performa model terus menjadi baik, ataupun kurangi poin (*error*) disaat model yang dihasilkan mengalami penurunan performa. Salah satu implementasi yang ditemukan ialah pada mesin pencari.

3. METODE PENELITIAN

Penelitian ini menggunakan bahasa pemrograman *python* untuk menjalankan programnya, serta menggunakan beberapa library yang sudah disediakan oleh *python* untuk pemrosesan *machine learning*. Alur penelitian dimulai dari mencari dataset, kemudian membagi dataset tersebut kedalam dua bagian yang berbeda, yaitu data *training* untuk proses pembuatan model dan data *testing* untuk proses evaluasi model, hingga nanti melakukan evaluasi dengan metode *confusion matrix*. Adapun alur penelitian yang lengkap terdapat pada Gambar 1.



Gambar 1. Alur Penelitian

3.1 Data

Dataset penelitian ini berasal dari *Gynaecology Department of the University Hospital Centre of Coimbra (CHUC)*, yang dapat diakses pada UCI *machine learning repository* secara gratis. Dataset yang dipakai merupakan hasil dari tes wanita yang didiagnosis mengalami kanker payudara, direkam sejak tahun 2009 sampai 2013. Keseluruhan dataset terdiri 116 data yang terdiri atas 64 orang mengalami kanker payudara dan 52 yang sehat [9].

Dalam dataset ada 9 atribut yang digunakan dalam pegukuran, yaitu umur, glukosa, *resistin*, HOMA, *insulin, leptin, body mass index* (BMI), *adiponectin*, dan MCP-1. Dalam proses klasifikasi nantinya akan menggunakan 2 skema, pertama menggunakan keseluruhan atribut yang ada dalam dataset, kedua dengan melakukan seleksi fitur untuk menentukan atribut yang memiliki pengaruh paling besar dalam proses klasifikasi, sehingga dapat diperoleh hasil klasifikasi yang lebih baik.

3.2 Seleksi Fitur

Proses Seleksi fitur berguna untuk memilih fitur apa saja yang mempunyai pengaruh yang lebih besar dalam proses klasifikasi. Pada awalnya dataset yang dipakai memiliki 9 fitur (atribut) yaitu *Age*, BMI, *Glucose*, *Insulin*, HOMA, *Leptin*, *Adiponectin*, *Resistin*, dan MCP-1. Setelah proses seleksi fitur didapat 4 fitur yang paling berpengaruh dalam proses klasifikasi, yaitu *Glucose*, *Resistin*, *Age*, dan BMI. Semua fitur dari dataset yang digunakan dideskripsikan pada Tabel 1.

Tabel 2. Deskripsi Fitur

No	Fitur	Deskripsi
1	Age	Umur Responden
2	Glucose	Level Glukosa
3	BMI	Body Mass Index
4	Insulin	Level Insulin
5	HOMA	Homeostatis Model Assessment
6	Leptin	Level Leptin
7	Adiponectin	Level Adiponectin
8	Resistin	Level Resistin
9	MCP-1	Monocyte Chemoattractan Protein-1

Penelitian ini menggunakan 4 skema dengan mengkombinasikan 2 algoritma dan penggunaan seleksi fitur. Hal ini sengaja dilakukan untuk membuktikan seberapa besar penggunaan seleksi fitur dapat meningkatkan klasifikasi. Skema penelitian yang dilakukan bisa dilihat dalam Tabel 2.

Tabel 2. Skema Penelitian

No	Algoritma	Fitur
1	SVM	Fitur 9
2	SVM	Fitur 4
3	DT	Fitur 9
4	DT	Fitur 4

3.3 Pembagian Dataset

Pada proses ini dataset yang diperoleh dibagi menjadi 2 bagian, yang pertama untuk data *training* kemudian yang kedua untuk data *testing*. Data *training* akan digunakan dalam proses pembuatan model klasifikasi, kemudian data *testing* hanya dipakai untuk melakukan evaluasi dari model yang sudah dibuat. Adapun besarnya pembagian dataset yang dipakai adalah data *training* sebanyak 80% dan data *testing* sebanyak 20%.

3.4 Training

Proses *training* adalah proses untuk membuat model *machine learning* yang nantinya digunakan untuk mengklasifikasi data. Pada proses pembuatan model ini data yang dipakai adalah data *training*.

3.5 Testing

Data *testing* yang didapat dari hasil pembagian dataset akan digunakan untuk melakukan proses *testing* pada model yang dibuat, sehingga nanti dapat diketahui seberapa baik proses klasifikasi yang dilakukan. Data *testing* tidak boleh digunakan pada proses *training*, supaya pada saat *testing* model benar benar belajar dari data baru.

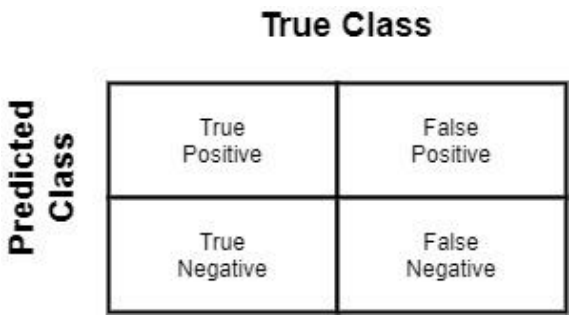
3.6 Klasifikasi

Klasifikasi merupakan proses analisis pada sebuah data yang dapat menghasilkan sebuah model klasifikasi, model yang sudah didapat bisa dipakai untuk menentukan kelas yang terkandung dalam sebuah dataset yang sudah diberi label. Model tersebut dinamakan *classifier*. Jadi, dengan *classifier* ini kita dapat menentukan kelas yang ada

dalam data (Sutoyo, 2018). Dalam penelitian ini model klasifikasi digunakan untuk mengelompokkan dataset kedalam dua macam polaritas, yaitu 1 dan 0. Angka 1 melambangkan terinfeksi kanker dan angka 0 melambangkan tidak terinfeksi kanker. Proses klasifikasi dilakukan dengan menggunakan dua algoritma yaitu SVM dan DT, setelah diperoleh hasilnya kemudian membandingkan performa yang terbaik diantara kedua algoritma tersebut.

3.7 Evaluasi

Evaluasi ini merupakan proses untuk mengukur performa klasifikasi dengan menggunakan data *testing*. Evaluasi dilakukan menggunakan metode *confusion matrix*, dimana metode tersebut akan menghasilkan data *True Positive*, *False Negative, True Negative*, dan *False Positive*. *Confusion matrix* bisa dilihat pada Gambar 2.



Gambar 2. *Confusion matrix*

Setelah mendapatkan *confusion matrix* selanjutnya adalah mengukur tingkat akurasi, spesifisitas, dan sensitivitas. Untuk mengukur akurasi dapat dilakukan dengan persamaan 3.1, untuk mengukur sensitivitas dapat menggunakan persamaan 3.2, dan untuk mengukur spesifisitas dapat dilakukan dengan persamaan 3.3.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Keterangan:
TP = *True Positive*
TN = *True Negative*
FP = *False Positive*
FN = *False Negative*

$$Sensitivity = \frac{TP}{TP+FN} \quad (4)$$

Keterangan:
TP = *True Positive*
FN = *False Negative*

$$Spesificity = \frac{TN}{TN+FP} \quad (5)$$

Keterangan:
TN = *True Negative*
FP = *False Positive*

4. HASIL DAN PEMBAHASAN

Dari penelitian yang dilakukan menunjukan bahwa skema metode SVM dengan menggunakan fitur seleksi memperoleh hasil terbaik, yaitu memperoleh hasil akurasi sebesar 87,5%, sensitivitas 90%, dan spesitifitas 85%. Sedangkan hasil paling buruk diperoleh oleh skema metode SVM tanpa fitur seleksi dengan memperoleh akurasi sebesar 70%, sensitivitas 64%, dan spesifistas 80%.

Selain itu penelitian ini memberikan informasi bahwa dengan melakukan fitur seleksi pada dataset dapat meningkatkan akurasi dalam proses klasifikasi secara signifikan. Hasil dari penelitian dapat dilihat pada tabel 4.1.

Tabel 4.1 Hasil Klasifikasi

Algoritma	Fitur	Akurasi	Sensitivitas	Spesifisitas
SVM	Fitur 9	87,5%	90%	85%
SVM	Fitur 4	70%	64%	80%
DT	Fitur 9	83,3%	88%	80%
DT	Fitur 4	75%	69%	81%

5. KESIMPULAN DAN SARAN

Dari hasil penelitian yang diperoleh, bisa dikesimpulkan bahwa metode klasifikasi menggunakan seleksi fitur mempunyai performa yang lebih baik dari pada yang tidak menggunakan seleksi fitur. Kemudian Algoritma SVM memiliki performa yang sangat baik untuk digunakan dalam klasifikasi data kanker payudara, sehingga dapat membantu mendiagnosis orang yang terinfeksi kanker payudara dengan lebih baik.

Saran untuk pengembangan bisa dilakukan dengan menggunakan algoritma *machine learning* yang lain seperti *random forest*, *XGBoost*, dan lain sebagainya. Ketika nanti data yang digunakan cukup besar bisa juga menggunakan *deep learning* untuk memperoleh hasil yang lebih baik. Disamping itu penelitian ini juga bisa dijadikan referensi untuk penelitian selanjutnya.

DAFTAR PUSTAKA

Journal Article

[1] F.Y. A'la, A.E. Permanasari, & N.A. Setiawan, *A Comparative Analysis of Tree-based Machine Learning Algorithms for Breast Cancer Detection*. l2th International Conference on Information & Communication Technology and System (lCTS). 2019.

[2] B.A. Farahdiba, Y.S. Nugroho, *Klasifikasi Kanker Payudara Menggunakan Algoritma Gain Ratio*. Jurnal Teknik Elektro Vol. 8 No.2. 2016.

[3] A. Perdana, M.T. Furqon, & Indriati, *Penerapan Algoritma Support Vector Machine (SVM) Pada Pengklasifikasian Penyakit Kejiwaan Skizofrenia (Studi Kasus: RSJ. Radjiman Wediodiningrat, Lawang)*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol.2, hlm. 3162-3167. 2018.

[4] Neneng, K. Adi, R.R. Isnanto, *Support Vector Machine Untuk Klasifikasi Citra Jenis Daging Berdasarkan Tekstur Menggunakan Ekstraksi Ciri Gray Level Co-Occurrence Matrices (GLCM)*. Jurnal Sistem Informasi Bisnis. pp. 1-10. 2016.

[5] A.M. Puspitasari, D.E. Ratnawati,& A.W. Widodo, *Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol.2, No.2. 2018.

[6] P.B.N. Setio, D.R.S. Saputro, & B. Winarno, *Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C4.5.* PRISMA, Prosiding Seminar Nasional Matematika 3, pp. 64-71. 2020.

[7] Rismayanti, *Decision Tree Penentuan Masa Studi Mahasiswa Prodi Teknik Informatika (Studi Kasus: Fakultas Teknik dan Komputer Universitas Harapan Medan)*. Jurnal Sistem Informasi Vol.2. 2018.

[8] R.Herbrich, & T.Graepel, *Introduction To Machine Learning With Applications In Information Security*. California: CRC Press. 2018.

[9] M. Patricio, et al, *Using Resistin, glucose, age and BMI to predict the presence of breast cancer.* BMC Cancer. Vol. 18, no.1. 2018.

[10] I. Sutoyo, *Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik*. Jurnal PILAR Nusa Mandiri Vol. 14, No.2. 2018.