JUSTIFY: Jurnal Sistem Informasi Ibrahimy
Vol. 2, No. 1, Juli 2023, hal. 74~80
E-ISSN: 2962-3642 | P-ISSN: 2962-3871
DOI: 10.35316/justify.v2i1.3370

Klasifikasi Kanker Payudara Menggunakan Algoritma SVM dengan Kernel RBF, Linier, dan Sigmoid

Ginanjar Abdurraman[1]

[1] Teknik Informatika, Teknik, Universitas Muhammadiyah Jember, Indonesia

Info Artikel

Riwayat Artikel:
Diterima : **24-April-2023**
Direvisi : **16-Juni-2023**
Disetujui : **18-Juli-2023**

Kata Kunci:

Kanker,
Support Vector Machine,
Kernel Linier,
Kernel RBF,
Kernel Sigmoid

ABSTRAK

Kanker payudara menjadi peringkat pertama baik dari kategori jenis kelamin maupun tingkat kematian. Penanganan yang terlambat sering ditemukan pada kasus kanker payudara yang menyebabkan meningkatnya faktor resiko kanker ini. Untuk itulah, diperlukan deteksi dini kanker payudara, sehingga penanganan dapat dilakukan tepat waktu, sehingga tangkat kematian karena kanker payudara dapat ditekan. Untuk itulah, dalam artikel ini ditawarkan deteksi dini kanker payudara menggunakan klasifikasi. Dataset pada penelitian ini menggunakan dataset kanker payudara wisconsin yang diambil dari Kaggle. Pada awalnya dataset memiliki missing value, selain itu data kategorikal belum dalam bentuk numerik, sehiingga perlu dilakukan preprocessing dengan teknik imputing missing value dan encoding untuk mengubah data kategorikal menjadi data numerik. Dataset dibagi menjadi dua proporsi, yakni 80% sebagai data training dan 20% sebagai data testing. Pada proses klasifikasi, dataset yang telah dilakukan preprocessing dilakukan klasifikasi menggunakan SVM dengan tiga kernel yang berbeda, yakni kernel linier, kernel RBF, dan kernel Sigmoid. Berdasarkan hasil penelitian yang telah diperoleh, kernel linier menunjukkan hasil kasifikasi terbaik jika diterapkan pada klasifikasi SVM, yakni dengan nilai akurasi mencapai 99%, dilanjutkan dengan performa kernel RBF dengan tingkat akurasi sebesar 92%, dan yang terakhir adalah kernel sigmoid dengan nilai akurasi 41%.

***Keywords*:**

Cancer,
Support Vector Machine,
Linear Kernel,
RBF Kernel,
Sigmoid Kernel

ABSTRACT

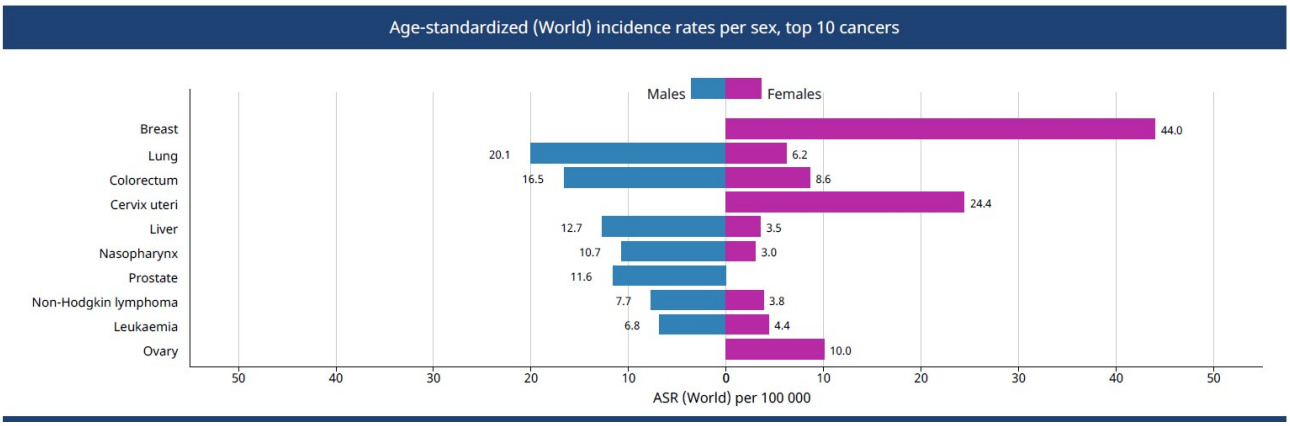
Breast cancer ranks first in both the gender category and the death rate. Late treatment is often found in cases of breast cancer which causes an increase in the risk factors for this cancer. For this reason, early detection of breast cancer is needed, so that treatment can be done in a timely manner, so that the death rate due to breast cancer can be reduced. For this reason, this article offers early detection of breast cancer using classification. The dataset in this study used the Wisconsin breast cancer dataset taken from Kaggle. Initially the dataset has a missing value, besides that the categorical data is not yet in numerical form, so it is necessary to do preprocessing with the missing value imputing technique and encoding to convert categorical data into numeric data. The dataset is divided into two proportions, namely 80% as training data and 20% as testing data. In the classification process, datasets that have been preprocessed are classified using SVM with three different kernels, namely the linear kernel, the RBF kernel, and the Sigmoid kernel. Based on the research results that have been obtained, the linear kernel shows the best classification results when applied to the SVM classification, namely with an accuracy value of up to 99%, followed by RBF kernel performance with an accuracy rate of 92%, and finally the sigmoid kernel with an accuracy value of 41%.

Penulis Korespondensi:
Ginanjar Abdurrahman
Program Studi Teknik Informatika
Universitas Muhammadiyah Jember
Email: abdurrahmanginanjar@unmuhjember.ac.id

1. PENDAHULUAN

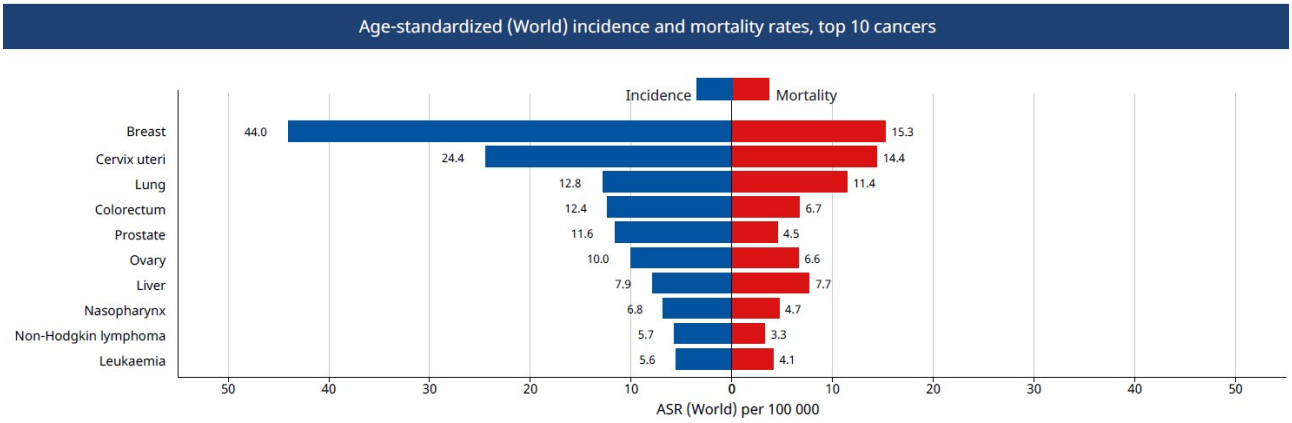
Berdasarkan [1] jumlah kasus baru untuk kanker di Indonesia pada tahun 2020 untuk semua jenis kelamin, semua umur mencapai total 396.914 kasus dengan rincian 25.943 kasus (14.1%) merupakan kanker paru-paru, 21.764 kasus (11,9%) kanker kolorektum, 16.412 kasus (9%) kanker liver, 15.427 kasus (8,4%) merupakan kanker nasopharynx, 13.563 kasus (7,4%) kanker prostat, dan sebanyak 90.259 (49,2%) merupakan kanker jenis lain. Sedangkan untuk laki-laki semua umur, jumlah kasus baru kanker mencapai 183.368, dengan rincian kanker paru-paru sebanyak 25.943 (14,1%), kanker kolorektum sebanyak 21.764 (11,9%), kanker liver sebanyak 16.412 (9%), kanker nasopharynx sebanyak 15.427 (8,4%), kanker prostat sebanyak 13.563 (8,4%), dan kanker jenis lain sebanyak 90.259 (49,2%). Adapun pada wanita semua umur, kanker payudara mencapai 65.858 (30,8%), kanker serviks sebanyak 36.63 (17,2%), kanker ovarium sebanyak 14.896 (7%), kanker kolorektum sebanyak 12.425 (5,8%), kanker tiroid sebanyak 9.053 (4,2%), dan kanker jenis lain sebanyak 74.681 (35%).

Masih menurut [1] juga menyebutkan bahwa kanker 10 besar berdasarkan jenis kelamin, yakni kanker payudara, kanker paru-paru, kanker kolorektum, kanker serviks uteri, kanker liver, kanker nasopharynx, kanker prostat, kanker non-hodgkin lymphoma, leukaemia, dan kanker ovarium. Untuk lebih jelasnya, sebaran kanker berdasarkan jenis kelamin dapat dilihat pada Gambar 1.



Gambar 1. data kanker 10 besar teratas berdasarkan jenis kelamin

Adapun data terkait data 10 besar kejadian kanker berdasarkan tingkat kematiannya [1], yakni: kanker payudara, kanker serviks uteri, kanker paru-paru, kanker kolorektum, kanker prostat, kanker ovarium, kanker liver, kanker nasopharynx, kanker non-hodgkin lymphoma, serta leukaemia. Selengkapnya, data sebaran kanker berdasarkan kejadian dan tingkat kematiannya dapat dilihat pada Gambar 2.



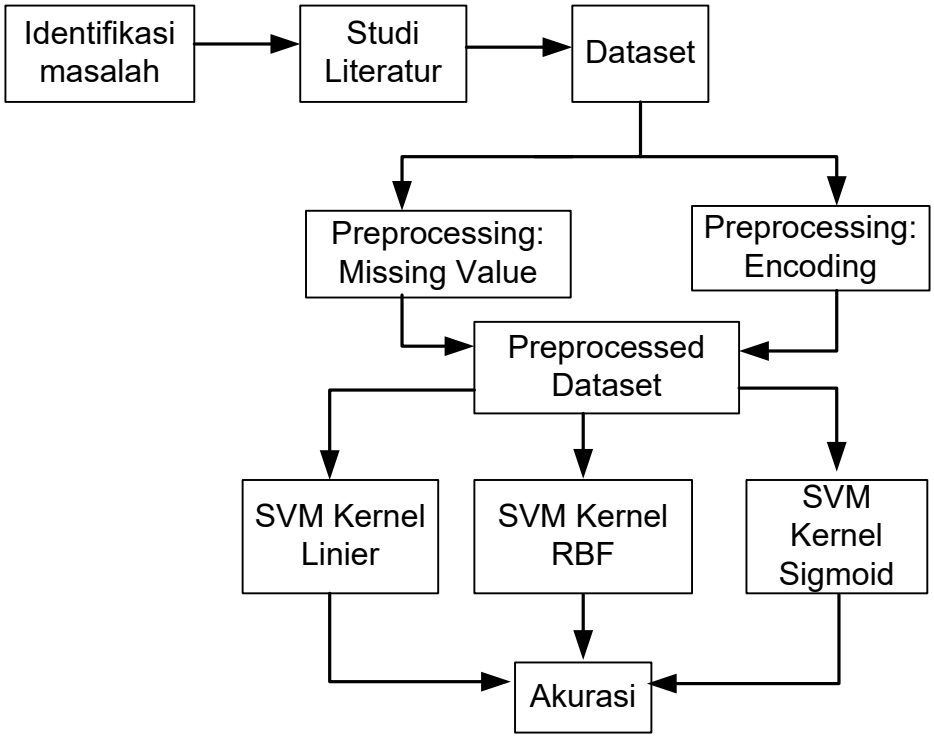
Gambar 2. Data 10 besar kejadian kanker berdasarkan tingkat kematian.

Berdasarkan data pada Gambar 1 dan Gambar 2, dapat dilihat bahwa kanker payudara menjadi peringkat pertama baik dari kategori jenis kelamin maupun tingkat kematian. Menurut [2] penanganan yang terlambat sering ditemukan pada kasus kanker payudara yang menyebabkan meningkatnya faktor resiko kanker ini. Untuk itulah, diperlukan deteksi dini kanker payudara, sehingga penanganan dapat dilakukan tepat waktu, sehingga tangkat kematian karena kanker payudara dapat ditekan. Untuk itulah, dalam artikel ini ditawarkan deteksi dini kanker payudara menggunakan klasifikasi.

Klasifikasi merupakan salah satu algoritma *supervised machine learning* untuk mengkategorikan kelas data [3]. Proses klasifiksi diartikan sebagai proses memperoleh model untuk mengidentifikasi kelas data, apabila model telah diperoleh, model tersebut dapat digunakan untuk klasifikasi kelas data baru [4]. *Support Vector Machine (SVM)* merupakan salah satu algoritma klasifikasi *(supervised learning)* yang dapat menangani data berdimensi tinggi, selain itu algoritma ini juga dapat menangani klasifikasi untuk data-data non-linier yang tidak dapat dipisahkan secara linier [5]. Algoritma SVM sangat baik digunakan untuk klasifikasi biner, yakni klasifikasi dengan dua kelas keputusan. Hal ini disampaikan oleh [6].

2. METODE PENELITIAN

Alur penelitian dapat dilihat pada Gambar 3.



Gambar 3. Bagan Penelitian

3. HASIL DAN ANALISIS

Pada bagian ini dijelaskan hasil penelitian berdasarkan alur penelitian yang telah digambarkan pada Gambar 3, yakni: *preprocessing* dan Klasifikasi SVM. Pada bagian *preprocessing*, dataset dilakukan imputing missing data dan encoding data kategorikal menjadi data numerik. Kemudian, pada bagian klasifikasi SVM, dataset yang sudah dilakukan preprocessing diklasiifikasikan menggunakan algoritma SVM dengan tiga kernel yang berbeda, yakni kernel linier, kernel RBF, serta kernel Sigmoid.

3.1 Studi Literatur

Ada beberapa penelitian yang relevan yang telah dilakukan sebagai dasar dalam penelitian ini, diantaranya, penelitian yang telah dilakukan oleh [7] yang berjudul Klasifikasi Kanker Menggunakan Algoritma NNGE, Random Forest, dan Random Committee. Penelitian ini menggunakan data pasien yang menjalani 4 jenis tes laboratorium. Pada tahap preprocessing dilakukan penanganan data ambigu dan data outlier. Selanjutnya data yang sudah dilakukan preprocessing dilanjutkan pada tahap klasifikasi menggunakan 3 metode, yakni NNGE, Random Forest, dan Random Committee, sehingga menghasilkan nilai akurasi untuk masing-masing metode, yakni akurasi untuk NNGE sebesar 100 %, akurasi untuk Random Forest sebesar 93,38%, dan akurasi untuk Random Committee sebesar 100%.

Penelitian selanjutnya dilakukan oleh [8] yang berjudul Komparasi Fungsi Kernel Metode Support Vector Machine untuk Analisis Sentimen Instagram dan Twitter (Studi kasus: Komisi Pemberantasan Korupsi). Dataset yang digunakan pada penelitian ini adalah data komentar dari twitter dan Instagram yang nantinya dipetakan menjadi sentiment positif, negatif atau netral. Penelitian ini bertujuan untuk membandingkan kinerja Support Vector Machine dalam klasifikasi sentiment berdasarkan nilai kernel SVM, yakni kernel linier, kernel polynomial, serta kernel sigmoid. Dari hasil penelitian, diketahui nilai akurasi dari implementasi kernel linier adalah 89.70%, sedangkan nilai akurasi dari implementasi kernel polynomial sebesar 81,45%, dan kinerja kernel sigmoid membuat nilai akurasi model sebesar 79,83%.

Adapun penelitian yang telah dilakukan oleh [9] yang berjudul Penerapan Metode Support Vector Machine (SVM) untuk mendeteksi Penyalahgunaan Narkoba, metode SVM digunakan untuk deteksi jenis narkotika pemakai, yang didasarkaan pada gejala yang dialami. Dataset pada penelitian ini adalah pasien rawat jalan BNN Provinsi Maluku yang berjumlah 101 pasien, dengan 23 macam gejala serta jenis narkoba yang digunakan. Jenis narkotika yang digunakan adalah Sabu, Ganja, Lem dan Sintesis. Splitting dataset yang digunakan untuk pencarian nilai akurasi adalah 60%:40%, 70%:30%, serta 80%:20%. Penelitian ini menggunakan Pada penerapannya, ada 2 metode yang digunakan, yakni SVM linier dan SVM Non-Linier. Untuk SVM Linier, nilai akurasi untuk dataset 60%:40% adalah 77,5%, untuk dataset 70%:30% adalah 83,3%, serta untuk dataset 80%:20% adalah 80%. Sedangkan untuk implementasi SVM Non-Linier, pada setiap splitting dataset, dibagi menjadi 8 parameter kernel untuk dua jenis kernel yang berbeda. Dalam hal ini kernel polynomial dan kernel RBF. Hasil dari kernel polynomial, pada data splitting 60%:40% diperoleh nilai akurasi terbaik sebesar 77,5%, pada data splitting 70%:30% diperoleh nilai akurasi terbaik sebesar 83,3%, sedangkan pada data splitting 80%:20% diperoleh akurasi terbaik sebesar 95%. Selanjutnya dari kernel RBF, pada data splitting 60%:40% diperoleh nilai akurasi terbaik sebesar 80%, pada data splitting 70%:30% diperoleh nilai akurasi terbaik sebesar 83,3%, dan pada data splitting 80%:20% diperoleh nilai akurasi terbaik sebesar 90%.

Penelitian yang dilakukan oleh [10], melakukan klasifikasi terhadap 22.335 data tweet mengenai kebijakan PSBB menggunakan algioritma Support Vector Machine untuk analisis sentiment. Pada penelitian ini, digunakan 4 model SVM berdasarkan kernel Linier, RBF, Polinomoial, serta Sigmoid. Kinerja algoritma SVM diuji menggunakan k-fold cross validation untuk memperoleh nilai akurasi model. Hasil klasifikasi model menggunakan kernel RBF merupakan model terbaik yang diperoleh, dengan nilai akurasi sebesar 95,94%.

3.2 Dataset

Dataset yang digunakan merupakan dataset public kanker payudara Wisconsin yang diambil dari Kaggle. Dataset ini terdiri dari 570 sel kanker dengan 30 fitur untuk menentukan apakah sel kanker jinak (Malignant) atau kanker ganas (Benign).

3.3 Preprocessing

Pada tahap ini, dilakukan dua proses, yakni *imputing missing data* dan *encoding* data kategorikal. Hal ini dikarenakan, pada dataset masih terdapat beberapa missing value dan data kategorikal pada kelas keputusan belum dalam bentuk numerik, sehingga untuk penanganan data pada *preprocessing* perlu dilakukan *imputing* dan *encoding*.

3.3.1 *Imputing Missing Data*

Dalam penanganan missing data, missing data perlu diidentifikasi terlebih dahulu pada setiap fiturnya. Dalam dataset, missing data dituliskan dengan nilai 0, hanya saja python menganggap nilai 0 sebagai nilai data, bukan sebagai keberadaan missing data. Sehingga perlu diubah sebagai entitas NaN terlebih dahulu dengan menggunakan perintah replace nol menjadi NaN. Setelah diubah menjadi NaN, banyaknya missing data dari setiap fitur, teridentifikasi seperti terlihat pada Tabel 1.

Setelah missing data teridentifikasi, Langkah selanutnya adalah dengan mengubah nilai-nilai missing data tersebut menggunakan Teknik imputing. Teknik imputing yang dipilih adalah imputing mengunakan nilai rata-rata (mean) dengan perintah fillna pada python.

Tabel 1. Identifikasi banyaknya missing data pada dataset

Fitur	Banyaknya missing data
radius_mean	0

texture_mean	0
perimeter_mean	0
area_mean	0
smoothness_mean	0
compactness_mean	0
concavity_mean	13
concave points_mean	13
symmetry_mean	0
fractal_dimension_mean	0
radius_se	0
texture_se	0
perimeter_se	0
area_se	0
smoothness_se	0
compactness_se	0
concavity_se	13
concave points_se	13
symmetry_se	0
fractal_dimension_se	0
radius_worst	0
texture_worst	0
perimeter_worst	0
area_worst	0
smoothness_worst	0
compactness_worst	0
concavity_worst	13
concave points_worst	13
symmetry_worst	0
fractal_dimension_worst	0

3.3.2 *Encoding* Data Kategorikal

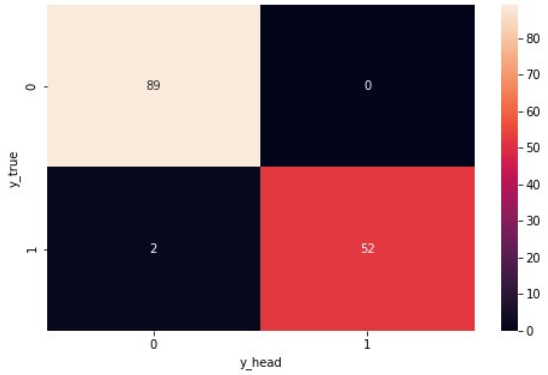
Data-data kategorikal terdapat pada kelas keputusan, dalam hal ini kelas keputusan *Benign* (Kanker Jinak) dan kelas keputusan *Malignant* (Kanker Ganas). Data kategorikal ini perlu diubah terlebih dahulu menjadi data numerik, sehingga dapat dibaca oleh python. Metode yang digunakan untuk mengubah data kategorikal menjadi data numerik, digunakan teknik encoding menggunakan label encoder. Sehingga diperoleh untuk kanker jinak dikategorikan dengan 0, dan kanker ganas dikategorikan dengan 1.

3.4 Klasifikasi SVM

Data yang telah dilakukan *preprocessing*, selanjutnya diklasifikasikan menggunakan algoritma SVM dengan diujicoba menggunakan 3 kernel yang berbeda, yakni kernel Linier, Kernel RBF, dan Kernel Sigmoid.

3.4.1 Klasifikasi SVM Menggunakan Kernel Linier

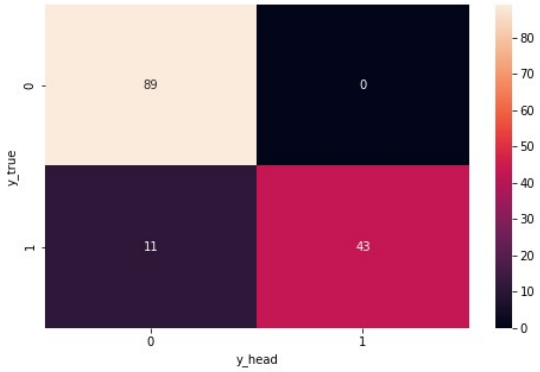
Hasil klasifikasi menggunakan SVM dengan kernel linier, diperoleh nilai True Positif sebesar 89, True Negatif sebesar 52, False Positif sebesar 0 dan False Negatif sebesar 2 Sehingga diperoleh nilai Akurasi Model sebesar 99 %.Visualisasi *confusion matriks* dapat dilihat pada Gambar 4.



Gambar 4 Visualisasi confusion matriks untuk SVM kernel linier

3.4.2 Klasifikasi SVM Menggunakan Kernel RBF

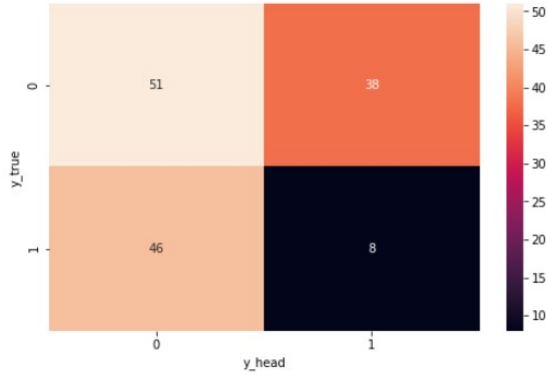
Hasil klasifikasi menggunakan SVM dengan kernel linier, diperoleh nilai True Positif sebesar 89, True Negatif sebesar 43, False Positif sebesar 0 dan False Negatif sebesar 11. Sehingga diperoleh nilai Akurasi Model sebesar 92 %.Visualisasi *confusion matriks* dapat dilihat pada Gambar 5.



Gambar 5. Visualisasi confusion matriks untuk SVM Kernel RBF

3.4.3 Klasifikasi SVM Menggunakan Kernel Sigmoid

Hasil klasifikasi menggunakan SVM dengan kernel linier, diperoleh nilai True Positif sebesar 51, True Negatif sebesar 8 False Positif sebesar 38 dan False Negatif sebesar 46. Sehingga diperoleh nilai Akurasi Model sebesar 41 %.Visualisasi *confusion matriks* dapat dilihat pada Gambar 6.



Gambar 6. Visualisasi confusion matriks untuk SVM Kernel Sigmoid

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah diperoleh, kernel liner menunjukkan hasil kasifikasi terbaik jika diterapkan pada klasifikasi SVM, yakni dengan nilai akurasi mencapai 99%, dilanjutkan dengan performa kernel RBF dengan tingkat akurasi sebesar 92%, dan yang terakhir adalah kernel sigmoid dengan nilai akurasi 41%.

REFERENSI

[1] A. I. Sutnick and S. Gunawan, “Cancer in Indonesia,” *JAMA J. Am. Med. Assoc.*, vol. 247, no. 22, pp. 3087–3088, 2021, doi: 10.1001/jama.247.22.3087.

[2] A. ; Nurrohmah, A. Aprianti, and S. Hartutik, “Risk factors of breast cancer in burma,” *Gaster J. Heal. Sci.*, vol. 21, no. 4, pp. 432–437, 2022, doi: https://doi.org/10.30787/gaster.v20i1.777.

[3] P. Bimo, N. Setio, D. Retno, S. Saputro, and B. Winarno, “Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C4.5,” *Prism. Pros. Semin. Nas. Mat.*, vol. 3, pp. 64–71, 2020.

[4] R. Nanda, E. Haerani, S. K. Gusti, and S. Ramadhani, “Klasifikasi Berita Menggunakan Metode Support Vector Machine,” vol. 5, no. 2, pp. 269–278, 2022.

[5] I. Mahendro and D. Abimanto, “Analisa Kepuasan Mahasiswa Terhadap E-Learning Menggunakan Algoritma Support Vector Machine,” *J. Sains Dan Teknol. Marit.*, vol. 23, no. 1, p. 97, 2022, doi: 10.33556/jstm.v23i1.333.

[6] G. N. Kurniawati, “Algoritma Machine Learning yang Harus Kamu Pelajari di Tahun 2021,” 2021. https://www.dqlab.id/algoritma-machine-learning-yang-perlu-dipelajari (accessed Feb. 05, 2022).

[7] M. N. U. R. Akbar, “KLASIFIKASI KANKER MENGGUNAKAN ALGORITMA NNGE , RANDOM FOREST , DAN RANDOM COMMITEE,” vol. 5, pp. 289–298, 2020.

[8] A. Zaiem and N. Charibaldi, “Komparasi Fungsi Kernel Metode Support Vector Machine untuk Analisis Sentimen Instagram dan Twitter (Studi Kasus : Komisi Pemberantasan Korupsi),” vol. 9, no. 2, pp. 33–42, 2021.

[9] P. Metode, S. Vector, and M. Svm, “PENERAPAN METODE SUPPORT VECTOR MACHINE (SVM) UNTUK MENDETEKSI PENYALAHGUNAAN NARKOBA Application of Support Vector Machine (SVM) Method to Detect Drug Abuse,” vol. 01, no. 02, pp. 111–122, 2022.

[10] A. Fahrurozi and H. Parasian, “IMPLEMENTASI ALGORITMA KLASIFIKASI SUPPORT VECTOR MACHINE UNTUK ANALISA SENTIMEN PENGGUNA,” pp. 149–162, 2021.