JUSTIFY: Jurnal Sistem Informasi Ibrahimy
Vol. 3, No. 1, Juli 2024, hal. 16-20
E-ISSN: 2962-3642 | P-ISSN: 2962-3871
DOI: 10.35316/justify.v3i1.5028

KLASIFIKASI TEKS MINING UNTUK DETEKSI KANKER MENGGUNAKAN SUPPORT VECTOR MACHINE

Ginanjar Abdurrahman[1], Hardian Oktavianto[2]

[1,2] Teknik Informatika,Taknik, Universitas Muhammadiyah Jember, Indonesia

Info Artikel

Riwayat Artikel:
Diterima : **24-Januari-2024**
Direvisi : **18-Maret-2024**
Disetujui : **08I-Juli-2024**

Kata Kunci:

Kanker,
Tiroid,
Paru-paru,
Usus besar

ABSTRAK

Banyak kasus kematian akibat kanker disebabkan oleh pasien yang terlambat memeriksakan diri. Pasien memeriksakan diri ke dokter jika kankernya sudah pada stadium tinggi (akut). Kanker merupakan penyebab kematian utama di dunia. Pada tahun 2020, tercatat hampir 10 juta kematian akibat kanker. Pembelajaran mesin adalah teknik yang meniru cara mesin (komputer) belajar dari data. Klasifikasi merupakan algoritma pembelajaran mesin untuk mencari pola dengan tujuan memisahkan kelas data. Salah satu algoritma klasifikasi untuk mengklasifikasikan data kanker adalah algoritma support vector machine (SVM).. Pada penelitian ini kanker (tiroid, paru-paru dan usus besar) akan diklasifikasi menggunakan algoritma SVM. Pasien nantinya akan diklasifikasikan menderita kanker tiroid, kanker paru-paru, atau kanker usus besar. Dataset ini merupakan data teks yang merupakan dataset publik yang diambil dari Kaggle. Setelah data melalui preprocessing dan dilakukan klasifikasi menggunakan algoritma SVM dengan proporsi pembagian train-test sebesar 70% sebagai data latih dan 30% sebagai data uji. Hasil penelitian manunjukkan bahwa matriks kinerja algoritma sebagai berikut: nilai akurasi sebesar 93,83%, nilai presisi sebesar 94,23%, nilai recall sebesar 94,35%, dan nilai F1-measure sebesar 94,27%

***Keywords*:**

Cancer,
Thyroid,
Lung,
Colon
Support Vector Machine

ABSTRACT

Many cases of death due to cancer are caused by patients being late in checking themselves. Patients go to the doctor if their cancer is already at a high stage (acute). Cancer is the main cause of death in the world. In 2020, almost 10 million deaths were recorded due to cancer. Machine learning is a technique that imitates the way machines (computers) learn from data. Classification is a machine learning algorithm to look for patterns with the aim of separating data classes. One classification algorithm to classify cancer data is the support vector machine (SVM) algorithm.. In this research, cancer (thyroid, lung and colon) will be classified using the SVM algorithm. Patients will later be classified as suffering from thyroid cancer, lung cancer or colon cancer. This dataset is text data which is a public dataset, taken from Kaggle. After the data goes through the preprocessing and classification is carried out using the SVM algorithm with a train-test split proportion of 70% as training data and 30% as testing data. The results show that the algorithm performance matrix was obtained as follows: accuracy value of 93.83%, precision value of 94.23%, recall value of 94.35%, and F1-measure value of 94.27%

Penulis Korespondensi:

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

1. PENDAHULUAN

Keterlambatan deteksi penyakit kanker akibat kurangnya kesadaran pasien menyebabkan tertundanya pengobatan oleh dokter. Saat pasien dirawat oleh dokter, penyakit kanker yang dideritanya sudah berada pada stadium tinggi (akut). Hal ini kemudian membuat angka prevalensi kematian akibat kanker menjadi tinggi. Menurut kajian penelitian yang dilakukan oleh [1] disebutkan bahwa kanker merupakan penyebab kematian utama di dunia. Pada tahun 2020, tercatat hampir 10 juta kematian akibat kanker.

Secara umum kanker dapat didefinisikan sebagai pertumbuhan abnormal sel pada tubuh manusia yang tentunya berpengaruh terhadap kesehatan manusia [2]. Menurut [3] kanker di Indonesia ranking 8 di Asia Tenggara dan ranking 23 di Asia, dengan kasus tertinggi untuk laki-laki (per 100.000 penduduk) adalah kanker paru mencapai 19,4 dengan angka kematian 10,9, kanker hati mencapai 12,4 dan angka kematian mencapai 7.6. Kemudian, untuk kasus tertinggi pada perempuan (per 100.000 penduduk) adalah kanker payudara mencapai 42,1 dengan rata-rata kematian 17, kemudian kanker rahim mencapai 23,4 dengan rata-rata kematian mencapai 13,9.

Untuk meminimalisir tingkat kematian, perlu deteksi dini dengan cara mengklasifikasikan kanker agar dapat segera ditangani oleh dokter. Kanker dalam penelitian ini, selanjutnya diklasifikasikan menjadi: kanker tiroid, kanker paru, dan kanker usus. Menurut [4] kanker tiroid merupakan tumor ganas yang menyerang sel parenkim pada kelenjar tiroid, sedangkan menurut [5] kanker tiroid memiliki perilaku klinis bervariasi dengan prevalensi kematian yang tinggi. Menurut [6] kanker paru memiliki proporsi 11,4% dari 19,3 juta kasus baru keseluruhan jenis kanker, yang menjadikannya berada pada urutan terbanyak kedua setelah kanker payudara. Kanker usus [7] merupakan tumor ganas yang muncul pada bagian usus besar dan menempati urutan keempat pada semus jenis kanker di Indonesia.

Dengan data kanker yang telah disebutkan sebelumnya, perlu adanya deteksi dini kanker, agar penanganan dapat dilakukan secara tepat dan cepat, sehingga kanker belum berada pada stadium yang membahayakan. Sebagai alternatif solusi untuk deteksi dini kanker, dalam penelitian ini ditawarkan pembelajaran mesin.

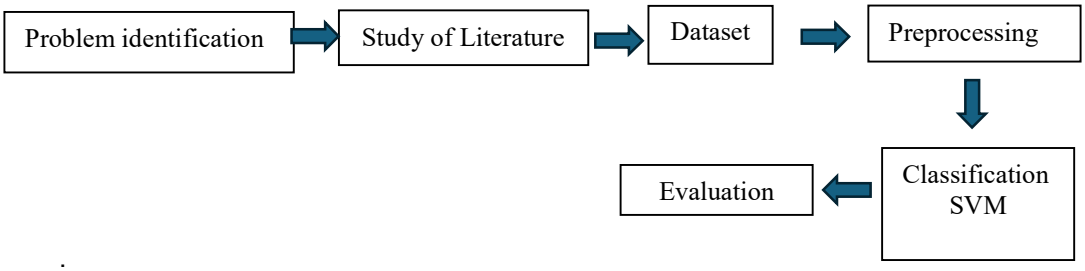
Pembelajaran mesin adalah teknik yang meniru cara mesin (komputer) belajar dari data. Klasifikasi merupakan algoritma pembelajaran mesin dalam konteks pembelajaran terawasi untuk mencari pola dengan tujuan memisahkan kelas data. Algoritma klasifikasi biasanya digunakan untuk memprediksi data yang belum mempunyai kelas data tertentu [8]. Salah satu algoritma klasifikasi yang dapat digunakan untuk mengklasifikasikan data kanker adalah algoritma support vector machine (SVM).

Algoritma SVM banyak digunakan untuk klasifikasi teks. Seperti halnya penelitian yang dilakukan oleh [9] Penelitian ini bertujuan untuk mengklasifikasikan berita menggunakan Support Vector Machine menggunakan 510 sampel berita dengan batas klasifikasi 3 kategori berita. Hasil kinerja algoritma SVM menunjukkan akurasi tertinggi yang diperoleh sebesar 88% untuk nilai parameter C=1, kernel linier dengan pembagian data uji dan data latih sebesar 90% dan 10%. Selain itu, juga melakukan klasifikasi teks menggunakan SVM. Penelitian ini bertujuan untuk melakukan klasifikasi Buku dengan SVM pada Digital Library. Dataset yang digunakan berasal dari https://opac.unesa.ac.id . Dataset ini terdiri dari sepuluh kategori buku yang disesuaikan dengan Dewey Decimal Classification (DDC). Preprocessing yang dilakukan adalah penghapusan terhadap daa ganda, selain itu pada preprocessing juga dilakukan casefolding, tokenizing, dan stopwords. Dataset yang digunakan juga terdiri dari 1000 record sebagai uji data. Pengambilan sampel dilakukan secara acak untuk menyeimbangkan data. Tahap ekstraksi fitur dilakukan menggunakan Term-Frequency – Inverse Document Frekuensi (TF-IDF) untuk mengolah teks menjadi numerik. Pada implementasi algoritma SVM, digunakan empat fungsi kernel, yaitu kernel linier, RBF, Polynomial, dan Sigmoid. Pengujian algoritma dilakukan pada dataset dengan tiga proporsi uji, yakni 60:40, 70:30, dan 80:20. Nilai akurasi klasifikasi dengan metode SVM, diperoleh nilai akurasi tertinggi sebesar 69,24% pada kernel linier. Sedangkan nilai precission 71%, recall 61%, dan F1-Score 64%.

Selain itu, [10] juga melakukan klasifikasi teks menggunakan SVM. Penelitian ini bertujuan untuk melakukan klasifikasi Buku dengan SVM pada Digital Library. Dataset ini terdiri dari sepuluh kategori buku yang disesuaikan dengan *Dewey Decimal Classification (DDC). Preprocessing* yang dilakukan adalah penghapusan terhadap daa ganda, selain itu pada *preprocessing* juga dilakukan *case folding, tokenizing, dan stopwords*.Dataset yang digunakan juga terdiri dari 1000 records sebagai uji data. Pengambilan sampel dilakukan secara acak untuk menyeimbangkan data. Tahap ekstraksi fitur dilakukan menggunakan *Term-Frequency – Inverse Document Frequency (TF-IDF)* untuk mengolah teks menjadi numerik. Pada implementasi algoritma SVM, digunakan empat fungsi kernel, yakni kernel linier, RBF, Polynomial, dan Sigmoid. Pengujian algoritma dilakukan pada dataset dengan tiga proporsi uji, yakni 60:40, 70:30, dan 80:20. Nilai akurasi klasifikasi dengan metode SVM, diperoleh nilai akurasi tertinggi sebesar 69,24% pada kernel linier. Sedangkan nilai precission 71%, recall 61%, dan F1-Score 64%.

Pada penelitian ini kanker (tiroid, paru-paru dan usus besar) akan diklasifikasi menggunakan algoritma Support Vector Machine (SVM). Pasien nantinya akan diklasifikasikan menderita kanker tiroid, kanker paru-paru, atau kanker usus besar. Dataset ini merupakan data teks yang merupakan dataset publik yang diambil dari Kaggle.

2. METODE PENELITIAN



.

3. HASIL DAN ANALISIS

3.1. Data Preprocessing

Secara umum teknik preprocessing data dilakukan dengan menggunakan library Natural Language Toolkit (NLTK) dengan Python. Proses yang dilakukan pada text preprocessing pada penelitian ini adalah sebagai berikut

3.1.1 Case folding

Pada tahap ini, semua huruf dalam dokumen diubah menjadi huruf kecil. Hal ini bertujuan agar kata yang sama tidak terdeteksi berbeda hanya karena perbedaan penggunaan huruf kapital pada kata tertentu. Dalam penelitian ini dilakukan dengan menggunakan fungsi doc.lower() dengan Python

3.1.2 Tokenization

Proses ini memecah dokumen teks menjadi bagian-bagian yang lebih sederhana. Bagian yang lebih sederhana kemudian disebut sebagai token. Token dapat berupa kata, frasa, atau berupa kalimat

3.1.3 Stemming

Pada tahap ini, kata-kata yang mempunyai awalan dan akhiran dihilangkan dari awalan dan akhiran tersebut, sehingga terbentuk kata dasar. Proses ini bertujuan untuk menyederhanakan dan membakukan kata, sehingga pemrosesan bahasa alami (NLP) lebih efektif. Pada penelitian ini Porter Stemmer digunakan untuk tahap stemming.

3.1.4 Filtering/Stopword removal

Pada tahap ini, kata-kata yang dianggap tidak penting pada kategori kelas keputusan akan dihilangkan. Dalam penelitian ini, stopwords bahasa Inggris digunakan dari perpustakaan NLTK Python dengan mengambil kata-kata populer dari bahasa Inggris..

3.1.5 Punctuation Removal

Pada tahap ini, karakter tanda baca seperti titik (.), koma (,), tanda tanya (?), dan sebagainya dihilangkan.

3.1.6 Numeric Removal

Dari dataset yang ada, masih terdapat karakter numerik di beberapa dokumen. Proses penghapusan numerik adalah proses menghilangkan karakter numerik pada dokumen.

3.1.7 Lemmatization

Proses ini mengekstrak akar kata dari kata yang tersemat dengan tujuan untuk mereduksi variasi kata pada akar kata. Lemmatisasi ini mempertimbangkan analisis morfologi kata, yaitu mengelompokkan berbagai bentuk infleksi kata agar dapat dianalisis menjadi satu item. Dalam bahasa Inggris misalnya, kata run, running, ran adalah semua bentuk dari kata run, oleh karena itu, run adalah inti dari semua kata tersebut dan lemmatization mengembalikan kata sebenarnya dari bahasa tersebut

3.2. Klasifikasi menggunakan SVM

Dalam teknik pengambilan sampel data, diambil 2000 dokumen dari total data 7569 dokumen dengan menggunakan teknik random sampling dengan penggantian. Teknik pengambilan sampel ini melibatkan total 133317 fitur/istilah/kata.

591	"Mesothelin; MTB: Mycobacterium tuberculosis; ...
7517	section all disclosure information for gie ed...
5395	"We also studied the oncogenic potential of SE...
243	"Interferon COVID19SARSCOV2IranIn this study e...
4652	"evaluate the clinicopathologic characteristic...
...	...
1841	" levels of physical activity change througho...
3309	the bacteroides fragilis b fragilis produ...
4946	sarscov2 has resulted in numerous cases of cor...
3018	subcutaneous hydration and medicationsinfusion...
3855	"A. Proliferation assay in Mero-14 cells. The ...

2000 rows × 1 columns

Data ini kemudian diuji dengan menggunakan train-test split dengan proporsi data 70% sebagai data latih dan 30% sebagai data uji. Dari hasil pengujian tersebut diperoleh matriks kinerja algoritma sebagai berikut: nilai akurasi sebesar 93,83%, nilai presisi sebesar 94,23%, nilai recall sebesar 94,35%, dan nilai F1-measure sebesar 94,27%.

4. KESIMPULAN

Setelah data melalui proses preprocessing yang telah disebutkan dan diklasifikasikan menggunakan algoritma Support Vector Machine (SVM), diperoleh kesimpulan sebagai berikut:

Dari total data sebanyak 7569 dokumen, diambil 2000 dokumen sebagai sampel data dengan menggunakan teknik random sampling dengan replacement yang melibatkan total 133317 fitur/istilah/kata. Data tersebut kemudian diuji menggunakan algoritma support vector machine (SVM) dengan proporsi pembagian train-test sebesar 70% sebagai data pelatihan dan 30% sebagai data pengujian. Dari hasil pengujian tersebut diperoleh matriks kinerja algoritma sebagai berikut: nilai akurasi sebesar 93,83%, nilai presisi sebesar 94,23%, nilai recall sebesar 94,35%, dan nilai F1-measure sebesar 94,27%.

REFERENSI

[1] I. Buana and D. A. Harahap, “Asbestos, Radon Dan Polusi Udara Sebagai Faktor Resiko Kanker Paru Pada Perempuan Bukan Perokok,” *AVERROUS J. Kedokt. dan Kesehat. Malikussaleh*, vol. 8, no. 1, p. 1, 2022, doi: 10.29103/averrous.v8i1.7088.

[2] T. Agustin, “Potensi Metabolit Aktif Dalam Sayuran Cruciferous Untuk Menghambat Pertumbuhan Sel Kanker,” *J. Penelit. Perawat Prof.*, vol. 1, no. November, pp. 89–94, 2019.

[3] Widyawati, “Hari Kanker Sedunia 2019,” 2019.

[4] A. W. Sindi Yulia Mustika, “MANAJEMEN ANESTESI PADA PAPILLARY THYROID CARCINOMA: SEBUAH LAPORAN KASUS,” vol. 4, no. November, pp. 1377–1386, 2022.

[5] A. Nur, A. Santosa, and A. Siti Komariyah, “Karakteristik Kanker Tiroid di Maluku Utara Tahun 2017-2018,” *J. Endur. Kaji. Ilm. Probl. Kesehat.*, vol. 8, no. 2, pp. 246–252, 2023, [Online]. Available: https://creativecommons.org/licenses/by/4.0/%0Ahttps://sinta.kemdikbud.go.id/journals/profile/1162.

[6] S. Sugiharto, R. A. P. Simanjuntak, and O. Larissa, “Kanker Paru, Faktor Risiko Dan Pencegahannya,” *Pros. SENAPENMAS*, p. 613, 2021, doi: 10.24912/psenapenmas.v0i0.15060.

[7] S. J. Zannah, I. S. Murti, and S. Sulistiawati, “Hubungan Usia dengan Stadium Saat Diagnosis Penderita Kanker Kolorektal di RSUD Abdul Wahab Sjahranie Samarinda,” *J. Sains dan Kesehat.*, vol. 3, no. 5, pp. 701–705, 2021, doi: 10.25026/jsk.v3i5.629.

[8] 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.

[9] 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.

[10] D. H. Amalia and W. Yustanti, “Klasifikasi Buku Menggunakan Metode Support Vector Machine pada

Digital Library," *J. Informatics Comput. Sci.*, vol. 3, no. 01, pp. 55–61, 2021, doi: 10.26740/jinacs.v3n01.p55-61.