E-Jurnal Matematika Vol. 12(2), Mei 2023, pp. 87-91 ISSN: 2303-1751
DOI: https://doi.org/10.24843/MTK.2023.v12.i02.p404

KLASIFIKASI PENYAKIT SIROSIS MENGGUNAKAN *SUPPORT VECTOR MACHINE*

Vania Riskasari YR[1§], I Putu Eka Nila Kencana[2], I Komang Gde Sukarsa[3]

[1]Program Studi Matematika, Fakultas MIPA – Universitas Udayana [vania.riskasari@gmail.com]
[2]Program Studi Matematika, Fakultas MIPA – Universitas Udayana [i.putu.enk@unud.ac.id]
[3]Program Studi Matematika, Fakultas MIPA – Universitas Udayana [gedesukarsa@unud.ac.id]
[§]*Corresponding Author*

ABSTRACT

Cirrhosis is one type of liver disease and is caused by forming fibrosis so that changes the liver structure become abnormal. Based on the presence of ascites, varicose veins, and bleeding, cirrhosis is divided into four clinical stages. This study aims to find the best classification model of cirrhosis using the support vector machine (SVM). SVM is a supervised learning method that aims to find the hyperplane with the maximum margin. In this study, the resulted model useful for determining the cirrhosis' stage from patients. The variables to classify are age, gender, ascites status, hepatomegaly status, spiders status, edema status, total bilirubin, total cholesterol, amount of albumin, amount of copper, alkaline phosphatase level test results, SGOT test results, amount of tryglycerides, amount of platelets, and prothrombin time. By applying radial basis function kernel, combination of parameter C and γ that gives the best accuracy is determined. The final model using SVM with parameters C = 1 and γ = 0,6 is the best model with the accuracy value of 67,86 percent.

Keywords: *Cirrhosis, Classification, Support Vector Machine*

1. PENDAHULUAN

Klasifikasi adalah teknik *data mining* yang bertujuan untuk menemukan model yang dapat mengklasifikasikan atau membedakan objek ke dalam sebuah kelas sehingga dapat dilakukan prediksi pada objek baru sesuai karakteristik kelas. Pada pembelajaran mesin, klasifikasi terdiri dari dua tahap yaitu pembelajaran untuk membangun model, dan menguji model dengan melakukan prediksi terhadap objek atau data baru (Han *et al*., 2012). Tujuan utama teknik klasifikasi pada pembelajaran mesin adalah mendapatkan model yang memaksimumkan kinerja pada data latih (Cervantes *et al*., 2020). Salah satu contoh aplikasi dari permasalahan klasifikasi adalah klasifikasi stadium penyakit sirosis.

Sirosis adalah penyakit proses perbaikan pada organ hati dengan membentuk bekas luka atau *fibrosis* setelah terjadi peradangan yang mengakibatkan hati sulit melakukan fungsinya. Penyebab utama penyakit ini di negara maju adalah virus hepatitis C dan konsumsi alkohol. Virus hepatitis B adalah penyebab umum di sub-sahara Afrika dan sebagian Asia. Sirosis berada pada urutan ke-14 penyebab kematian dengan prediksi total korban 1,03 juta kematian di dunia. Sirosis merupakan penyebab utama adanya transplantasi hati, sebanyak 5500 setiap tahunnya di Eropa (Tsochatzis *et al.*, 2014). Berdasarkan ada tidaknya varises, asites, dan pendarahan, penyakit sirosis dibagi menjadi empat stadium klinis (Amico, 2004).

Pencegahan kematian akibat penyakit sirosis dapat dilakukan dengan mendeteksi stadiumnya. Pada ranah pembelajaran mesin, deteksi dapat dilakukan dengan membangun model klasifikasi yang dapat digunakan untuk memprediksi stadium penyakit pasien. Hasil prediksi yang diperoleh selanjutnya digunakan sebagai acuan dalam menentukan penanganan dan pengobatan yang tepat sehingga dapat mencegah kematian. Salah satu metode yang dapat digunakan dalam membangun model klasifikasi penyakit sirosis adalah *support vector machine* (SVM).

SVM merupakan metode pembelajaran terawasi pada pembelajaran mesin di mana pengembangan dan implementasi algoritma

SVM menarik bagi penelitian teoretis maupun terapan dalam *machine learning*, *data mining*, dan bioinformatika (Izenman, 2008). Menurut Bishop (2006), pendekatan SVM dalam menyelesaikan masalah didukung oleh konsep margin. Tujuan dari metode SVM adalah untuk memisahkan dataset ke dalam dua kelas yang berbeda oleh suatu *hyperplane* optimal dengan margin yang maksimum (Vapnik, 1995). SVM dapat digunakan untuk melakukan klasifikasi pada data yang dapat dipisahkan secara linear dan nonlinear. Untuk mengatasi masalah yang dipisahkan secara linear, dapat digunakan *hard margin classifier*. Sedangkan, *soft margin classifier* digunakan untuk mengatasi masalah data yang tidak dapat dipisahkan secara linear (nonlinear).

Karena tidak semua data dapat dipisahkan secara linear maka klasifikasi yang diperoleh tidak memiliki kemampuan generalisasi yang tinggi walaupun telah ditentukan *hyperplane* optimal. Untuk mengatasi hal tersebut, dikenal konsep kernel pada SVM. Terdapat empat contoh kernel yaitu kernel linear, *polynomial*, *radial basis function* (RBF), dan *multilayer perceptron*.

Standar dari metode SVM adalah mengatasi masalah klasifikasi dua kelas, namun untuk mengatasi masalah klasifikasi yang lebih dari dua kelas dapat digunakan *multi-class* SVM yang mendekomposisikan masalah klasifikasi *multi-class* menjadi serangkaian masalah klasifikasi biner sehingga standar dari metode SVM tetap dapat digunakan. Salah satu skema dalam *multi-class* SVM adalah *one-versus-rest*. Pada skema ini, dari masalah klasifikasi K kelas, dibangun sebanyak K fungsi keputusan yang membagi kelas ke-k dan bukan kelas ke-k untuk $k = 1,2, \dots, K$. Kemudian, label kelas ditentukan oleh model klasifikasi yang memberikan output terbesar.

Darsyah (2014) mengklasifikasikan penyakit tuberkulosis menggunakan pendekatan SVM dengan kernel RBF dan regresi logistik. Hasil penelitian tersebut menunjukkan bahwa metode SVM menghasilkan nilai akurasi lebih tinggi dibandingkan dengan regresi logistik yaitu sebesar 98%. Puspitasari *et al.* (2018) mengklasifikasikan penyakit gigi dan mulut ke dalam empat kelas menggunakan *multi-class* SVM. Pada penelitian tersebut, metode SVM memperoleh hasil optimal dengan nilai rata-rata akurasi sebesar 93,329% menggunakan kernel *radial basis function* (RBF).

Mencermati beberapa hal yang telah dipaparkan di atas, penelitian ini bertujuan untuk menemukan nilai akurasi terbaik dari model klasifikasi dari penyakit sirosis menggunakan *multi-class* SVM skema *one-versus-rest* dengan kernel RBF

1.1 *Support Vector Machine* (SVM)

Diasumsikan terdapat himpunan data training T yang berisi pasangan (x_i, y_i) dengan $x_i \in R^m$ dan $y_i \in \{-1, +1\}$ untuk $i = 1,2, \dots, N$. Sehingga, T dapat didefinisikan sebagai $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$. Jika terdapat $w \in R^m$, $b \in R$, dan suatu bilangan positif ε sedemikian sehingga untuk setiap $y_i = +1$ berlaku $(w \cdot x_i + b) \geq \varepsilon$ dan untuk setiap $y_i = -1$ berlaku $(w \cdot x_i + b) \leq -\varepsilon$, maka himpunan data training T dan permasalahan klasifikasinya dapat dipisahkan secara linear (Deng *et al.*, 2013).Misalkan fungsi keputusan:

$$f(x) = w^{\mathrm{T}} \cdot x + b \quad (1)$$

dengan w adalah vektor berdimensi m dan b adalah bias (Abe, 2005). Jika $f(x)$ pada persamaan (1) sama dengan nol, maka persamaan ini mewakili *hyperplane* yang memisahkan kedua kelas data. *Margin* dari *hyperplane* didefinisikan sebagai berikut:

$$d = \frac{2}{\|w\|} \quad (2)$$

Untuk menemukan *hyperplane* yang optimal, maka harus ditemukan *hyperplane* yang memiliki *margin* maksimum pada persamaan (2). *Margin* akan bernilai maksimum apabila nilai $\|w\|$ minimum. Meminimumkan $\|w\|$ ekuivalen dengan meminimumkan $\|w\|^2$, sehingga diperoleh permasalahan optimal kuadratik sebagai berikut:

$$\min \frac{1}{2}\|w\|^2 \quad (3)$$

dengan fungsi kendala $y_i\big((w \cdot x) + b\big) - 1 \geq 0$ untuk $i = 1,2, \dots, N$. Permasalahan optimisasi pada persamaan (3) dapat diselesaikan dengan membentuk fungsi Langrange berikut:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i \{y_i(w^{\mathrm{T}} x_i + b) - 1\} \quad (4)$$

dengan $\alpha_i = (\alpha_1, \alpha_2, \dots, \alpha_N)^{\mathrm{T}}$ adalah vektor pengali Lagrange. Permasalahan optimisasi pada persamaan (4) merupakan permasalahan data yang dipisahkan secara linear dengan *hard margin* classifier. Sedangkan, permasalahan optimasi dari *soft margin* dapat diperoleh dengan menambahan total dari variabel *slack* ξ_i yang dikalikan dengan parameter regularisasi C terhadap persamaan (3), yang dapat dituliskan sebagai berikut (Abe, 2005):

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad (5)$$

dengan fungsi kendala $y_i\big((w \cdot x) + b\big) \geq 1 - \xi_i$ dan $\xi_i \geq 0$ untuk $i = 1,2,\dots,N$ dan ξ_i adalah variable *slack* yang menyatakan jarak suatu data yang berada di luar *support lines*. Untuk itu, didefinisikan fungsi Lagrange pada kasus *soft margin* untuk menyelesaikan persamaan (5) yang memuat variabel Lagrange α_i dan β_i sebagai berikut (Abe, 2005):

$$L(w,b,\xi,\alpha,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\big(y_i\big(w^{\mathrm{T}}x_i + b\big) - 1 + \xi_i\big) - \sum_{i=1}^{N}\beta_i\xi_i \quad (6)$$

dengan $\alpha_i = (\alpha_1, \alpha_2, \dots, \alpha_N)^{\mathrm{T}}$ dan $\beta_i = (\beta_1, \beta_2, \dots, \beta_N)^{\mathrm{T}}$.

1.2 *Kernel Tricks*

Untuk meningkatkan keterpisahan linear, didefinisikan fungsi kernel yang memetakan ruang input asli ke ruang *dot product* yang berdimensi tinggi. Terdapat empat contoh kernel, yaitu:

a. Kernel Linear (Abe, 2005):

$$K(x_i, x_j) = x_i^{\mathrm{T}}x_j$$

b. Kernel *Polynomial* (Abe, 2005):

$$K(x_i, x_j) = \big(x_i^{\mathrm{T}}x_i + 1\big)^d$$

c. Kernel RBF (Abe, 2005):

$$K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right)$$

d. Kernel *Multilayer Perceptron* (Kecman, 2001):

$$K(x_i, x_j) = \tanh\big(x_i^{\mathrm{T}}x_j + b\big)$$

2. METODE PENELITIAN

2.1 Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini adalah data stadium penyakit sirosis dari pasien yang diperoleh dari Kaggle.com dengan judul *Cirrhosis Prediction Dataset* dengan total sebanyak 412 data. Data pasien diklasifikasikan ke dalam empat target/kelas (Y) yaitu stadium 1, stadium 2, stadium 3, dan stadium 4 yang dilengkapi dengan 15 atribut/*feature* (X) yaitu umur, jenis kelamin, status asites, status hepatomegali, status spiders, status edema, jumlah bilirubin dalam tubuh, jumlah kolesterol dalam tubuh, jumlah albumin dalam tubuh, jumlah copper pada urine dalam tubuh, hasil tes alkaline phosphatase, hasil tes SGOT, jumlah trigliserida dalam tubuh, jumlah trombosit dalam tubuh, dan masa protrombin. Data pada kelas stadium 1 berjumlah 21 data, pada kelas stadium 2 berjumlah 92 data, pada kelas stadium 3 berjumlah 155 data, dan pada kelas stadium 4 berjumlah 144 data. Data yang diperoleh dibagi menjadi dua, yaitu data *training* untuk melatih model dan data *testing* untuk menguji model.

2.2 Teknik Analisis Data

Langkah-langkah yang dilakukan dalam penelitian ini adalah sebagai berikut:

1. *Data Preprocessing*
   Data preprocessing yang dilakukan adalah mengatasi data yang hilang atau *missing value* pada atribut data dengan cara menghapus baris data yang mengandung *missing value*. Selain itu, dilakukan normalisasi data karena terdapat rentang yang terlalu jauh pada data menggunakan *min-max normalization* yang bertujuan untuk mendapatkan rentang yang sama pada data yaitu dalam rentang 0 sampai 1.
2. *Data Splitting*
   Data dibagi menjadi data *training* dan data *testing*. Dari keseluruhan data, sebanyak 90% digunakan sebagai data *training* dan 10% digunakan sebagai data *testing*. Pembagian/*splitting* data dilakukan secara acak pada masing-masing kelas.
3. Analisis menggunakan *multi-class* SVM dengan skema *one-versus-rest*, dengan membangun empat fungsi keputusan (*hyperplane*) untuk setiap model klasifikasi. Adapun langkah analisis adalah sebagai berikut:
   i. Melatih model klasifikasi dengan menetapkan nilai parameter C dan nilai parameter kernel RBF (γ) yang digunakan untuk membangun model SVM dengan $C = 1, 2, 4, 8, 16$ dan $\gamma =$ 0,01; 0,02; 0,03; 0,04; 0,05; 0,06; 0,07; 0,08; 0,09; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 1.
   ii. Melakukan 10-*fold cross validation* pada masing-masing model klasifikasi yang dibangun menggunakan data *training*.
   iii. Menghitung nilai akurasi dari masing-masing model klasifikasi menggunakan *confusion matrix*.
   iv. Melakukan pengujian model pada model yang dibangun menggunakan data *testing* dan menghitung nilai akurasi dari masing-masing model klasifikasi menggunakan *confusion matrix*.

3. Memilih model SVM dengan parameter-parameter pada model yang menghasilkan akurasi tertinggi pada data *training* dan rata-rata nilai akurasi tertinggi pada 10-*fold cross validation* dan tidak mengalami *overfitting* (kinerja baik pada *training* namun buruk pada *testing*).

3. HASIL DAN PEMBAHASAN

Setelah dilakukan *data preprocessing*, diperoleh dataset baru dengan jumlah 276 data yang terdiri dari 12 data dari kelas stadium 1, 59 data dari kelas stadium 2, 111 data dari kelas stadium 3, dan 94 data dari kelas stadium 4. Data set baru yang diperoleh akan digunakan dalam proses analisis menggunakan *multi-class* SVM dengan skema *one-versus-rest*.

Kemudian, *data splitting* dilakukan dan diperoleh jumlah data *training* dan data *testing* pada masing-masing kelas yang dituangkan ke dalam Tabel 1.

Tabel 1. Jumlah Data *Training* dan *Testing*

Kelas	Jumlah data *training*	Jumlah data *testing*
Stadium 1	10	2
Stadium 2	53	6
Stadium 3	100	11
Stadium 4	85	9
Jumlah	248	28

(Sumber: Diolah, 2022)

Dari Tabel 1. diperoleh jumlah data *training* adalah sebanyak 248 data yang terdiri dari 10 data dari kelas stadium 1, 53 data dari kelas stadium 2, 100 data dari kelas stadium 3, dan 85 data dari kelas stadium 4. Sedangkan jumlah data *testing* sebanyak 28 data yang terdiri dari 2 data dari kelas stadium 1, 6 data dari kelas stadium 2, 11 data dari kelas stadium 3, dan 9 data dari kelas stadium 4. Setelah dataset dibagi menjadi data *training* dan data *testing*, dilakukan pembangunan model klasifikasi menggunakan *multi-class* SVM dengan skema *one-versus*-rest dengan nilai parameter $C = 1, 2, 4, 8,$ dan 16 dan menggunakan kernel RBF dengan parameter $\gamma =$ 0,01; 0,02; 0,03; 0,04; 0,05; 0,06; 0,07; 0,08; 0,09; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; dan 1. Tahap pelatihan atau *training* dilakukan dengan membangun model klasifikasi yang menggunakan kombinasi parameter C dan γ. Sehingga diperoleh 95 model klasifikasi. Masing-masing model klasifikasi memiliki nilai akurasi atau ketepatan model dalam melakukan klasifikasi.

Setelah model klasifikasi *multi-class* SVM dengan skema *one-versus-rest* dibangun, dilakukan 10-*fold cross validation* menggunakan data *training* untuk mengukur kekonsistenan model dalam melakukan klasifikasi. Dalam melakukan 10-*fold cross validation*, dibentuk 10 partisi data. Terdapat 8 partisi yang berisi 25 data dan 2 partisi yang berisi 24 data. Kemudian, dilakukan iterasi sebanyak 10 kali. Pada iterasi ke-1, partisi ke-1 berperan sebagai data *testing* dan partisi lainnya berperan sebagai data *training*, dan seterusnya. Kemudian, dihitung rata-rata dari nilai akurasi pada setiap iterasi.

Kemudian, dilakukan pengujian pada model dengan melakukan prediksi terhadap label kelas pada data *testing*. Label kelas dari hasil prediksi dan label kelas aktual dituangkan ke dalam *confusion matrix* yang akan memberikan nilai ketepatan klasifikasi atau akurasi dari model yang telah dibangun. Berikut adalah *confusion matrix* untuk model klasifikasi dengan parameter $C = 1$ dan $\gamma = 0{,}6$ yang dituangkan dalam Tabel 2. berikut:

Tabel 2. *Confusion Matrix* untuk Model dengan parameter $C = 1$ dan $\gamma = 0{,}6$

Kelas Aktual	Kelas Prediksi			
	Stadium			
	1	2	3	4
Stadium 1	0	0	2	0
Stadium 2	0	1	3	2
Stadium 3	0	0	10	1
Stadium 4	0	0	1	8

(Sumber: Diolah, 2022)

Dari Tabel 2. diperoleh untuk kelas stadium 1 dan stadium 2 tidak ada data yang terklasifikasi dengan benar, untuk kelas stadium 3 terdapat 6 data yang terklasifikasi dengan benar, dan untuk kelas stadium 4 terdapat 10 data yang terklasifikasi dengan benar. Dari *confusion* matrix yang terbentuk dihitung nilai akurasi model pada data *testing*. Tabel 3 menunjukkan nilai akurasi dari model dengan parameter $C = 1$ dan $\gamma = 0{,}6$.

Tabel 3. Akurasi Model dengan $C = 1$ dan $\gamma = 0{,}6$

Rata-rata Akurasi *10-fold cross validation*	Akurasi *Training*	Akurasi *Testing*
53,65%	66,94%	67,86%

(Sumber: Diolah, 2022)

Dari Tabel 3. diperoleh untuk model klasifikasi dengan parameter $C = 1$ dan $\gamma = 0{,}6$ menghasilkan rata-rata akurasi pada 10-*fold cross validation* sebesar 53,65%, nilai akurasi pada *training* sebesar 66,94% dan pada *testing* sebesar 67,68%, sehingga model tidak mengalami *overfitting*. Selain itu, model dengan parameter $C = 1$ dan $\gamma = 0{,}6$ memiliki nilai akurasi tertinggi pada 10-*fold cross validation* dan pada data *training* dibandingkan dengan model lainnya.

4. KESIMPULAN DAN SARAN

Model klasifikasi *multi-class* SVM dalam mengklasifikasikan penyakit sirosis dengan skema *one-versus-rest* menggunakan parameter $C = 1$ dan $\gamma = 0{,}6$ tidak mengalami *overfitting* dan memiliki nilai akurasi tertinggi pada data *training* yaitu sebesar 66,94% dengan rata-rata nilai akurasi pada 10-*fold cross validation* tertinggi sebesar 53,65%.

Pada penelitian ini, hasil yang diperoleh kurang memuaskan. Hal ini disebabkan oleh ketidakseimbangan pada data. Untuk itu, pada penelitian lebih lanjut, disarankan untuk mengatasi masalah data yang tidak seimbang menggunakan beberapa metode, salah satunya adalah *synthetic minority over-sampling technique* (SMOTE).

Untuk meningkatkan keterpisahan linear, penelitian ini hanya menggunakan kernel RBF. Untuk itu, penelitian lebih lanjut disarankan untuk menggunakan kernel yang lain seperti kernel linear, *polynomial*, dan *multilayer perceptron*.

DAFTAR PUSTAKA

Abe, Shigeo. 2005. *Support Vector Machines for Pattern Classification*. 2nd ed. New York: Springer.

Amico, Gennaro D. 2004. "Esophageal Varices: From Appearance to Rupture; Natural History and Prognostic Indicators." *Portal Hypertension in the 21st Century*: 147–54.

Bishop, Christopher M. 2006. *Pattern Recognition And Machine Learning*. New York: Springer.

Cervantes, Jair, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. 2020. "A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends." *Neurocomputing*.

Darsyah, Moh Yamin. 2014. "Klasifikasi Tuberkulosis Dengan Pendekatan Metode Supports Vector Machine (SVM)." *Statistika* 2(2): 37–41.

Deng, Naiyang, Yingjie Tian, and Chunhua Zhang. 2013. *Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions*. CRC Press.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. *Data Mining Concepts And Techniques*. 3rd ed. USA: Elsevier Inc.

Izenman, Alan Julian. 2008. *Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning*. New York: Springer.

Kaggle.com. 2020. "Cirrhosis Prediction Dataset." http://www.kaggle.com/datasets/fedesoriano/cirrhosis-prediction-dataset (April 5, 2022).

Kecman, Vojislav. 2001. *Learning And Soft Computing - Support Vector Machines, Neural Networks, And Fuzzy Logic Models*. London, England: The MIT Press.

Puspitasari, Ana Mariyam, Dian Eka Ratnawati, and Agus Wahyu Widodo. 2018. "Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine." *J-Ptiik* 2(2): 802–10. http://j-ptiik.ub.ac.id.

Tsochatzis, Emmanuel A., Jaime Bosch, and Andrew K. Burroughs. 2014. "Liver Cirrhosis." *The Lancet* 383: 1749–61.

Vapnik, Vladimir N. 1995. Theory *The Nature of Statistical Learning*. 2nd ed. New York: Springer.