Statistika, Vol. 21 No. 1, 1 – 4
Mei 2021

Diagnosis Penderita Penyakit Kanker Paru Menggunakan *Support Vector Machine* dan *Naïve Bayes*

MUHAMMAD IQBAL YUNAN HELMI[1], DIAN ANGGRAENI[2], ALFIAN FUTUHUL HADI[3]

[1,2,3]Program Studi Matematika Fakultas MIPA Universitas Jember, Indonesia
Email: [1]iqbalyunan1111@gmail.com

ABSTRAK

Menurut data jenis kanker yang menjadi penyebab kematian terbanyak adalah kanker paru, mencapai 1,7 juta kematian pertahun. Penyakit ini disebabkan oleh banyak faktor salah satunya genetika. Dalam penelitian ini akan dilakukan diagnosis kanker paru menggunakan metode *Support Vector Machine* (SVM) dan *Naïve Bayes*. *Naïve Bayes* merupakan teknik prediksi berbasis probabilitas sederhana yang berdasarkan pada model fitur independent, sedangkan klasifikasi menggunakan SVM dapat dijelaskan secara sederhana yaitu usaha untuk mendapatkan *hyperplane* sebagai fungsi pemisah terbaik yang dapat memisahkan dua kelas yang berbeda pada ruang input. Pada penelitian ini akan dibandingkan metode SVM dan *Naive Bayes* untuk didapatkan mana metode yang mempunyai akurasi terbaik. Data *microarray* yang digunakan pada penelitian ini berupa 80 individu dengan masing-masing jumlah ekspresi genetiknya 2408. Sebanyak 60 individu tergolong ke dalam kelas kanker, dan 20 individu termasuk ke dalam kelas normal. Hasil dari penelitian ini adalah SVM mempunyai nilai akurasi sebesar 90% dan *Naïve Bayes* mempunyai nilai akurasi sebesar 75%.

Kata Kunci: *Support Vector Machine*, *Naïve Bayes*, *Hyperplane*, *Microarray*, Ekspresi genetik.

ABSTRACT

According to data, the type of cancer that is the most common cause of death is lung cancer, reaching 1.7 million deaths in one year. This disease is caused by many factors, one of which is genetics. In this study, lung cancer will be diagnosed using the Support Vector Machine (SVM) and Naïve Bayes methods. Naïve Bayes is a simple probability-based prediction technique based on an independent feature model, while the classification using SVM can be explained simply, namely the effort to get a hyperplane as the best separator function that can separate two different classes in the input space. In this study, the SVM and Naive Bayes methods will be compared to obtain which method has the best accuracy. The microarray data used in this study were 80 individuals, with each genetic expressions 2408. A total of 60 individuals belonged to the cancer class, and 20 individuals belonged to the normal class. The results of this study are SVM has an accuracy value of 90% and Naïve Bayes has an accuracy value of 75%.

Keywords: Support Vector Machine, Naïve Bayes, Hyperplane, Microarray, Genetic expression.

1. PENDAHULUAN

Kanker paru adalah salah satu jenis penyakit paru yang memerlukan penanganan dan tindakan yang cepat dan terarah. Menurut data jenis kanker WHO, yang menjadi penyebab kematian terbanyak adalah kanker paru, mencapai 1,7 juta kematian pertahun. Ada beberapa faktor yang dapat mempengaruhi terjangkitnya kanker paru, salah satu penyebabnya yaitu mutasi ekspresi genetika sel paru (Bakhtiar dan Soeprijanto, 2006). Jumlah ekspresi genetika kanker paru sangat banyak sehingga dibutuhkan sebuah sistem klasifikasi kanker paru yang dapat mengklasifikasikan antara sel yang beresiko kanker paru dan sel sehat. Beberapa metode yang dapat digunakan untuk pengklasifikasi kanker paru dalam ilmu satatistika antara lain *Naïve Bayes* dan *Support Vector Machine (SVM).*

Metode *Naive Bayes* merupakan metode yang hanya membutuhkan jumlah data *training* kecil untuk menentukan estimasi paremeter dalam proses pengklasifikasian. *Naïve Bayes* merupakan teknik prediksi berbasis probabilitas sederhana yang berdasarkan pada model fitur *independent* (Ridwan dkk, 2013). sedangkan klasifikasi menggunakan SVM dapat dijelaskan secara sederhana yaitu usaha untuk mendapatkan *hyperplane* sebagai fungsi pemisah terbaik yang dapat memisahkan dua kelas yang berbeda pada ruang input (Nugroho dkk). Ayu, Foriana dan Santi di tahun 2012 melakukan penelitian diagnosis kanker payudara dengan menggunakan SVM dengan hasil penelitian tersebut menunjukkan ketepatan klasifikasi metode SVM mencapai 94%. Penelitian Hidayatul dan Yunita di tahun 2018 membahas mengenai diagnosis penyakit jantung menggunakan metode *Naïve Bayes* dengan ketepatan klasifikasi sebesar 92,31%. Beberapa penelitian yang sebelumnya telah dilakukan, SVM dan *Naïve Bayes* diketahui memiliki akurasai yang tinggi. Penelitian ini dilakukan dengan tujuan untuk melakukan perbandingan antara algoritma *Naive Bayes* dengan SVM dalam memprediksi keberhasilan metode pengobatan kanker paru, karena pengobatan penyakit ini sangat bergantung pada diagnosis pasti.

2. METODE PENELITIAN

Jenis data yang digunakan adalah data sekunder yang didapat dari situs https://www.ncbi.nlm.nih.gov. Data yang digunakan sebanyak 80 individu dengan 2408 variabel gen kanker paru yang diklasifikasikan kedalam kelas Normal yang didefinisikan dengan 0 dan kelas Kanker yang didefinisikan dengan 1. Metode analisis data yang digunakan dalam penelitian ini adalah SVM dan *Naïve Bayes* dengan menggunakan *software R studio*. Langkah-langkah yang dilakukan dalam penelitian adalah sebagai berikut:

1. Data yang diperoleh disesuaikan menjadi bentuk matriks terlebih dahulu menggunakan *Microsoft Excel*.
2. Melakukan spliting data 75:25 dengan proporsi sama setiap kelas. Kemudian membuat data frame dari data *training* dan data *testing*.
3. Melakukan Klasifikasi data menggunakan Support Vector Machine.
4. Melakukan Klasifikasi data menggunakan Naïve Bayes.
5. Membandingkan hasil klasifikasi model SVM dan model Naïve Bayes berdasarkan *confusion matrix*.

Tabel 1 Confussion Matrix

Kelas Asli	Kelas Prediksi	
	Negatif	Positif
Negatif	*True Negative* (TN)	*False Positive* (FP)
Positif	*False Negative* (FN)	*True Positive* (TP)

3. HASIL DAN PEMBAHASAN

Gambaran secara umum mengenai data penelitian pada ekspresi genetika kanker paru-paru akan disajikan menggunakan *microarray* data. Data *microarray* yang digunakan berupa 80 individu dengan masing-masing jumlah genetiknya 2408. Sebanyak 60 individu tergolong ke dalam kelas kanker, dan 20 individu termasuk ke dalam kelas normal. Data yang telah diinputkan ke program R kemudian dibagi menjadi proses *training* dan data *testing*. Data ekspresi genetika dibagi menjadi 75% proses *training* dan 25% data *testing*. Data dibagi secara random dengan proporsi yang sama, jumlah data *training* sebanyak 60 data dan data *testing* sebanyak 20 data.

Support Vector Machine

Proses pengolahan data *training* pada SVM menggunakan fungsi kernel *linear, polynomial, radial*, dan *sigmoid*. Jumlah data pada proses *training* sebanyak 60 data dengan perbandingan 18 data kelas normal dan 42 data kelas kanker, sedangkan data *testing* yang digunakan sebanyak 20 data dengan 5 kelas normal dan 15 kelas kanker. Dalam pengujian data *training* dilakukan metode *k-fold cross validation* dengan nilai *k* adalah 5 dan 10 pada masing-masing

kernel. Hasil yang didapat adalah nilai *error* terkecil dari masing-masing fungsi kernel seperti pada Tabel 2.

Tabel 2 Nilai *error* klasifikasi

Fungsi Kernel	*5fold*		*10fold*		Hasil *Training*
	Cost	*Error*	*Cost*	*Error*	
Linear	**0,001**	**0,1166667**	0,001	0,1333333	**100%**
Polynomial	0,001	0,3	1	0,2666667	85%
Radial	1	0,2666667	10	0,25	100%
Sigmoid	10	0,2	10	0,1333333	100%

Berdasarkan nilai *error* yang didapat, dapat diketahui bahwa fungsi kernel terbaik adalah kernel *linear* dengan 0,1166 sebagai nilai *error* terkecil. Sehingga untuk pemodelan klasifikasi pada SVM akan dilakukan menggunakan fungsi kernel *linear*. Fungsi kernel terbaik digunakan untuk memprediksi klasifikasi kelas dalam data *testing*.

Tabel 3 Akurasi kernel *linear*

Fungsi Kernel	*5fold*		*10fold*		Hasil *Testing*
	Cost	Akurasi	*Cost*	Akurasi	
Linear	**0,001**	**0,9008**	0,001	0,8590	**100%**

Dapat dilihat pada Tabel 3 pengujian menggunakan *k-fold cross validation* dengan nilai *k* sebesar 5 dan 10 menghasilkan rata-rata akurasi yang berbeda. Hasil pada *5-fold cross validation* memberikan hasil yang lebih baik dengan nilai rata-rata akurasi 0,9008. Nilai rata-rata akurasi dari *10-fold cross validation* sebesar 0,859. Untuk mengetahui hasil klasifikasi dengan data pengujian menggunakan kernel *linear* dapat dilihat dengan menggunakan *confusion matrix* yang disajikan pada Tabel 5 berikut ini.

Tabel 4 Confusion matrix SVM

Kelas Asli	Kelas Prediksi	
	Normal	Kanker
Normal	4	1
Kanker	1	14

Berdasarkan hasil *confusion matrix* dapat dijelaskan bahwa terdapat kesalahan klasifikasi pada setiap kelas. Tingkat akurasi yang dihasilkan dari fungsi kernel *linear* dengan 18 data terklasifikasikan secara benar dari total 20 data adalah sebesar 90%.

Naïve Bayes

Proses pengolahan data pada *Naïve Bayes* menggunakan metode *k-fold cross validation*, dengan nilai *k* sebesar 5 dan 10. Jumlah data pada proses *training* sebanyak 60 data dengan perbandingan 18 data kelas normal dan 42 data kelas kanker, sedangkan data *testing* yang digunakan sebanyak 20 data dengan 5 kelas normal dan 15 kelas kanker. Hasil yang didapat adalah nilai akurasi terbaik seperti pada Tabel 5.

Tabel 5 Akurasi Testing Naïve Bayes

Fungsi Kernel	*5fold*		*10fold*		Hasil *Testing*
	Cost	Akurasi	*Cost*	Akurasi	
Linear	**0,001**	**0,77**	0,001	0,8114	**75%**

Berdasarkan nilai rata-rata akurasi yang didapat, diketahui bahwa metode *10-fold cross validation* dengan 0,8114 sebagai nilai rata-rata terbesar dan metode *5-fold cross validation* memiliki akurasi 0,77. Model *Naïve Bayes* terbaik digunakan untuk memprediksi klasifikasi

kelas dalam data *testing*. Untuk mengetahui hasil klasifikasi dapat dilihat dengan menggunakan *confusion matrix* yang disajikan pada Tabel berikut ini.

Tabel 6 *confusion matrix* metode *Naïve Bayes*

Kelas Asli	Kelas Prediksi	
	Normal	Kanker
Normal	3	2
Kanker	3	12

Berdasarkan hasil *confusion matrix* dapat dijelaskan bahwa terdapat kesalahan klasifikasi pada setiap kelas. Tingkat akurasi yang dihasilkan dari *Naïve Bayes* dengan 15 data terklasifikasikan secara benar dari total 20 data adalah sebesar 75%.

Perbandingan

Setelah melakukan pengujian menggunakan metode SVM dan *Naive Bayes*, diketahui bahwa metode SVM mampu mengklasifikasikan data ekspresi genteika kanker paru kedalam kelas normal dan kelas kanker dengan akurasi sebesar 90%, sedangkan hasil pengujian menggunakan metode *Naïve Bayes* memiliki akurasi sebesar 75%. Berdasarkan hasil akurasi klasifikasi yang diperoleh, dapat dinyatakan bahwa metode SVM lebih baik jika dibandingkan dengan metode *Naive Bayes*. Pada penelitian ini SVM memiliki akurasi yang tinggi karena data yang digunakan adalah data dengan dua *output class*, sehingga sesuai dengan SVM yang merupakan metode klasifikasi *linear* dengan konsep menemukan garis pemisah terbaik antara dua jenis *class*, sedangkan *Naïve Bayes* merupakan metode klasifikasi yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class* berdasarkan variabelnya.

4. KESIMPULAN

Berdasarkan hasil pengujian metode *Naive bayes* dan SVM terhadap data ekspresi genetika kanker paru menggunakan program R, dapat diambil kesimpulan bahwa metode SVM memiliki hasil klasifikasi yang lebih baik dari metode *Naïve Bayes*. Pengujian metode SVM menggunakan kernel *linear* dengan metode *5-fold cross validation* dan *10-fold cross validation* memiliki tingkat ketepatan klasifikasi sebesar 90%, sedangkan untuk metode *Naïve Byaes* menggunakan metode *5-fold cross validation* dan *10-fold cross validation* memiliki ketepatan klasifikasi sebesar 75%.

DAFTAR PUSTAKA

Ayu, Foriana dan Santi Wulan. 2012. Analisis Diagnosis Kanker Payudara Menggunakan Regresi Logistik dan Support Vector Machine (SVM) Berdasar Hasil Mamografi. Surabaya: Institut Teknologi Sepuluh Nopember.

Bakhtiar, A dan Bambang. (2006). Kanker Paru dan Penatalaksanaanya. Surabaya: Universitas Airlangga.

Hidayatul, S. H dan Yuita A. S. 2018. Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes. Malang: Universita Brawijaya.

Nugroho, A.S., Arief B. Witarto dan Dwi Handoko. 2003. Suport Vector Machines : Teori Aplikasinya dalam Bioinformatika. Kuliah Umum Ilmu Komputer.com.

Ridwan, M., Suyono, H., & Sarosa, M. (2013). Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. Jurnal EECCIS, 59-64.

World Health Organization. 2018. WHO Media Centre di http://www.who.int/mediacentre/factsheets/fs297/en/. (di akses 20 November 2020).