J. Ris. & Ap. Mat. Vol. 4 No. 1 (2020) pp. 17-27
Jurnal Riset dan Aplikasi Matematika
e-ISSN: 2581-0154
URL: journal.unesa.ac.id/index.php/jram

KLASIFIKASI AKSES INTERNET OLEH ANAK-ANAK DAN REMAJA DEWASA DI JAWA TIMUR MENGGUNAKAN *SUPPORT VECTOR MACHINE*

MUKTI RATNA DEWI

Departemen Statistika Bisnis, Fakultas Vokasi, Institut Teknologi Sepuluh Nopember

mukti_ratna@its.ac.id

ABSTRAK

Pada tahun 2018 penetrasi pengguna internet di Indonesia naik sebesar 10,12% dibanding tahun sebelumnyan dengan pengguna terbanyak berada di Pulau Jawa yang mencapai 55%. Kelompok usia pengguna internet terbanyak berturut-turut berada pada usia 15 hingga 19 tahun, umur 20 hingga 24 tahun, dan anak-anak yang berumur 5 hingga 9 tahun. Berdasarkan penelitian yang dilakukan oleh Nurrahman (2017), faktor- faktor yang secara signifikan mempengaruhi penggunaan internet di Jawa Timur antara lain adalah umur, jenjang pendidikan, perbedaan tempat tinggal antara perkotaan dan pedesaan, status pekerjaan, perangkat yang digunakan dalam mengakses internet dalam tiga bulan terakhir serta kepemilikan bangunan. Untuk mengetahui besar tingkat pengguna internet maka perlu dilakukan pengelompokkan berdasarkan faktor-faktor yang mempengaruhinya. Oleh karena itu, penelitian ini melakukan klasifikasi akses internet oleh anak-anak dan remaja usia 6 hingga 21 tahun di Jawa Timur berdasarkan faktor-faktor yang signifikan menggunakan *Support Vector Machine* (SVM) dengan fungsi kernel *Radial Basis Fuction* (RBF). Berdasarkan nilai AUC sebesar 0,92, kinerja model SVM yang terbentuk tergolong sangat bagus (*excellent*) dengan nilai akurasi, sensitivitas, dan spesifisitas berturut-turut sebesar 86,45%; 84,64% dan 88,63%.

Kata Kunci: akses internet, klasifikasi, *Radial Basis Function, Support Vector Machine*

ABSTRACT

Internet user penetration in Indonesia rose by 10.12% in 2018 compared to the previous year with the most users in Java reaching 55%. The age groups of most internet users are respectively aged 15 to 19 years, ages 20 to 24 years, and children aged 5 to 9 years. Based on research conducted by Nurrahman (2017), factors that significantly influence internet use in East Java include age, education level, differences in residence between urban and rural areas, employment status, devices used in accessing the internet in the past three months and building ownership. To find out the level of internet users, it is necessary to do grouping based on the factors that influence it. Therefore, this study classifies internet access by children and adolescents aged 6 to 21 years in East Java based on significant factors using Support Vector Machine (SVM) with Radial Basis Fuction (RBF) as kernel function. Based on the AUC value of 0.92, the performance of the SVM model that was formed was classified as excellent with a value of accuracy, sensitivity, and specificity of 86.45%; 84.64% and 88.63%, respectively.

Keywords: internet access, classification, Radial Basis Function, Support Vector Machine

2010 *Mathematics Subject Classification*: 65D12
Tanggal Masuk: 29-02-20; direvisi: 20-04-20; diterima: 26-04-20

1 Pendahuluan

Sebagai upaya dalam menghadapi Revolusi Industri 4.0, pemerataan kualitas infrastruktur internet di seluruh tanah air menjadi penting. Meratanya akses internet diharapkan dapat mendorong perkembangan teknologi di daerah-daerah. Oleh sebab itu, penyelesaian infrastruktur digital harus digiatkan sehingga antar daerah di Indonesia dapat saling terhubung. Dengan demikian, jumlah penduduk yang bisa menikmati akses internet dengan lancar akan terus meningkat.

Pada tahun 2018 penetrasi pengguna internet di Indonesia adalah 64,8% dari total populasi 264,16 jiwa atau naik sebesar 10,12% dibanding tahun sebelumnyan dengan pengguna terbanyak berada di Pulau Jawa yang mencapai 55% [1]. Sementara itu, segmentasi pengguna internet terbanyak berturut-turut berada pada usia 15 hingga 19 tahun dan umur 20 hingga 24 tahun. Anak-anak berumur 5 hingga 9 tahun juga menggunakan internet dengan persentase sebesar 25,2% dari keseluruh sampel yang berada pada umur tersebut [2].

Penelitian yang dilakukan oleh [3] menunjukkan bahwa umur, jenjang pendidikan, perbedaan tempat tinggal antara perkotaan dan pedesaan, status pekerjaan, perangkat yang digunakan dalam mengakses internet dalam tiga bulan terakhir serta kepemilikan bangunan merupakan faktor-faktor yang secara signifikan mempengaruhi penggunaan internet di Jawa Timur. Untuk mengetahui besar tingkat pengguna internet maka perlu dilakukan pengelompokkan berdasarkan faktor-faktor yang mempengaruhinya.

Penelitian mengenai klasifikasi penggunaan internet pernah dilakukan oleh [4] menggunakan regresi logistik biner. Pada penelitian tersebut, variabel jenis kelamin juga dimasukkan sebagai prediktor dan terbukti signifikan mempengaruhi akses internet selama tiga bulan terakhir. Sementara itu, model regresi logistik biner yang terbentuk mampu mengelompokkan akses internet dengan akurasi keseluruhan sebesar 86,6%. Meskipun angka tersebut terbilang baik, namun akurasi model klasifikasi dapat ditingkatkan dengan menerapkan *machine learning*.

SVM (*Support Vektor Machine*) adalah salah satu metode yang akhir-akhir ini dikembangkan untuk meningkatkan performasi akurasi klasifikasi. SVM yang digagas oleh [5] merupakan salah satu *supervised learning model* di *machine learning* untuk penyelesaian permasalahan klasifikasi biner. Metode ini merupakan salah satu metode yang baik untuk mengatasi permasalahan klasifikasi berdimensi tinggi [6]. Konsep fundamental dari SVM adalah menemukan *hyperplane* dalam ruang dimensi-*N* yang dapat secara jelas mengklasifikasikan data [5]. Secara umum terdapat dua kondisi pada permasalahan klasifikasi biner, yaitu *linearly separable case* dan *non-linearly separable case*. Salah satu konsep yang dikenalkan untuk menangani kasus *non-linearly separable* adalah dengan menggunakan fungsi kernel yang dapat memproyeksikan data ke dimensi yang lebih tinggi sehingga kedua kelas dapat dipisahkan secara linier [5]. Penelitian yang dilakukan oleh [7] menganjurkan fungsi kernel *Radial Basis Function* (RBF) sebagai pilihan pertama dan utama dalam pemilihan *kernel trick* karena selain memiliki performa yang lebih baik, jumlah hiperparameter yang mempengaruhi kerumitan *model selection* lebih sedikit dibandingkan dengan fungsi kernel *polynomial* sehingga proses komputasinya lebih efisien. Hasil penelitian oleh [8] juga menunjukkan bahwa fungsi kernel RBF memiliki akurasi lebih baik dibandingkan dengan fungsi kernel lainnya.

Berdasarkan uraian di atas, penelitian ini akan menggunakan SVM dengan fungsi kernel RBF untuk klasifikasi akses internet oleh anak-anak dan remaja usia 6 hingga 21 tahun di Jawa Timur berdasarkan faktor-faktor yang terbukti berpengaruh signifikan pada penelitian terdahulu.

2 Tinjauan Pustaka

2.1 *Support Vector Machine* (SVM)

Semisal permasalahan klasifikasi sebanyak *m* poin dalam R^n direpresentasikan oleh matriks **A** yang berukuran *m* x *n* di mana setiap $\mathbf{A}_i$ merupakan matriks **D** berukuran *m* x *m* dengan nilai 1 atau -1 pada diagonal utamanya. Formulasi SVM untuk permasalahan linier adalah sebagai berikut:

$$\begin{aligned} &\min_{(\mathbf{w},\gamma,\xi)\in R^{n+1+m}} \nu\mathbf{e}^{\mathrm{T}}\boldsymbol{\xi} + \frac{1}{2}\|\mathbf{w}\|_2^2 \\ &s.t \quad \mathbf{D}\left(\mathbf{A}\mathbf{w} - \mathbf{e}\gamma\right) + \boldsymbol{\xi} \geq \mathbf{e},\ \boldsymbol{\xi} \geq \mathbf{0} \end{aligned} \tag{1}$$

di mana :

- $\boldsymbol{\xi}$: vektor *slack* non-negatif berukuran $m \times 1$.
- **w** : vektor normal berukuran $n \times 1$.
- **e** : vektor satu berukuran $m \times 1$.
- γ : parameter penentu lokasi bidang pemisah terhadap titik asal.
- *v* : parameter positif yang menyeimbangkan bobot dari *training error* $\mathbf{e}^{\mathrm{T}}\boldsymbol{\xi}$ dan *margin maximization term* $\frac{1}{2}\|\mathbf{w}\|_2^2$.

Fungsi objektif pada persamaan (1) merupakan permasalahan non-linier yang kurang efisien dari segi komputasi. Oleh karena itu, persamaan (1) ditransformasi ke dalam *dual space* agar waktu komputasi lebih efisien [9]. Transformasi ini disebut juga *Wolfe dual problem* yang memiliki formulasi sebagai berikut:

$$\begin{aligned} &\min_{\mathbf{u}\in R^m} \ \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{D}\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{D}\mathbf{u} - \mathbf{e}^{\mathrm{T}}\mathbf{u} \\ &\text{s.t.} \quad \mathbf{e}^{\mathrm{T}}\mathbf{D}\mathbf{u} = 0,\ 0 \leq \mathbf{u} \leq \nu\mathbf{e} \end{aligned} \tag{2}$$

dengan variabel primal dari **w** memiliki formulasi

$$\mathbf{w} = \mathbf{A}^{\mathrm{T}}\mathbf{D}\mathbf{u} = \sum_{\{i|u_i>0\}}^{m} u_i D_{ii}\mathbf{A}_i^{\mathrm{T}} \tag{3}$$

dan γ diberikan oleh

$$\gamma = \mathbf{A}_i\mathbf{w} - D_{ii} = \mathbf{A}_i\mathbf{A}^{\mathrm{T}}\mathbf{D}\mathbf{u} - D_{ii}. \tag{4}$$

Pada kasus non-linier, sebuah fungsi kernel $K\left(\mathbf{A}, \mathbf{A}^{\mathrm{T}}\right)$ digunakan untuk memetakan data ke dimensi yang lebih tinggi dan kemudian dicari *hyperplane* menggunakan SVM linier untuk memisahkan dua kelas secara optimal. Hasil dari formulasi dual dari SVM non-linier adalah sebagai berikut:

$$\begin{aligned} &\min_{\mathbf{u}\in R^m} \ \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{D}K\left(\mathbf{A}, \mathbf{A}^{\mathrm{T}}\right)\mathbf{D}\mathbf{u} - \mathbf{e}^{\mathrm{T}}\mathbf{u} \\ &\text{s.t.} \quad \mathbf{e}^{\mathrm{T}}\mathbf{D}\mathbf{u} = 0,\ 0 \leq \mathbf{u} \leq \nu\mathbf{e} \end{aligned} \tag{5}$$

Bidang pemisah non-linier yang dihasilkan adalah:

$$K\left(\mathbf{x}^{\mathrm{T}}, \mathbf{A}^{\mathrm{T}}\right)\mathbf{D}\mathbf{u} = \gamma \tag{6}$$

di mana

$$\gamma = K\left(\mathbf{A}_i, \mathbf{A}^{\mathrm{T}}\right)\mathbf{D}\mathbf{u} - D_{ii},\ i \in \mathbf{I} := \left\{ j \mid 0 < u_j < v \right\} \tag{7}$$

Classifier yang dihasilkan adalah sebagai berikut:

$$f\left(\mathbf{x}\right) = sign\left(g\left(\mathbf{x}\right)\right) \tag{8}$$

dengan *g* (**x**) adalah persamaan (3).

2.2 Fungsi Kernel

Terdapat beberapa alternatif fungsi kernel yang digunakan dalam membangun model SVM, di antaranya adalah linier, polinomial, RBF, dan sigmoid. Namun, secara umum *Radial Basis Function* (RBF) adalah pilihan pertama dalam pemilihan fungsi kernel pada SVM [7]. Kernel ini secara nonlinier memetakan sampel ke dalam ruang dimensi yang lebih tinggi sehingga, tidak seperti kernel linier, dapat menangani kasus ketika hubungan antara label kelas dan atribut nonlinier. Selain itu, kernel linier sebenarnya adalah kasus khusus dari RBF karena kernel linier dengan parameter *v* memiliki performa klasifikasi yang sama dengan kernel RBF dengan beberapa nilai parameter μ dan *v* [10]. Sebagai tambahan, kernel sigmoid memiliki persamaan dengan kernel RBF untuk nilai parameter tertentu [11]. Terakhir, jumlah hiperparameter kernel RBF lebih sedikit dibandingkan dengan kernel polinomial. Jumlah hiperparameter ini mempengaruhi kompleksitas pemilihan model sehingga dan waktu komputasi yang dibutuhkan. Formulasi kernel RBF adalah sebagai berikut:

$$K\left(\mathbf{A},\mathbf{A}\right)_{ij} = \exp\left(-\mu\left\|\mathbf{A}_i^T - \mathbf{A}_j\right\|^2\right) \tag{9}$$

di mana μ adalah parameter kernel dengan dengan $i, j = 1, 2, \ldots, m$ dan $\ell = m$.

2.3 *Cross Validation* (CV) dan *Grid Search*

Hiperparameter adalah karakteristik model yang nilainya tidak dapat ditaksir dari data. Nilai tersebut harus ditetapkan sebelum proses pembelajaran dimulai. Salah satu teknik dalam pemilihan hiperparameter fungsi kernel adalah metode *grid search* yang digabung dengan *cross validation*.

Grid search adalah proses pemindaian data untuk menemukan parameter yang optimal untuk model yang diberikan. Metode ini akan membangun dan mengevaluasi model untuk setiap kombinasi hiperparameter yang telah ditentukan. Oleh karena itu, *grid search* tidak efisien secara komputasi dan mungkin membutuhkan waktu lama sampai di dapatkan kombinasi parameter yang optimal. Meskipun demikian, metode ini juga memiliki kelebihan yaitu proses algoritmanya dapat diparalelkan karena setiap pasangan hiperparameter saling bebas [12].

Cross validation (CV) sendiri adalah metode yang sering digunakan untuk evaluasi model. Jenis yang paling sederhana dari CV adalah metode *holdout* yang memisahkan data menjadi dua set, yaitu data *training* dan data *testing*. Data *training* digunakan untuk membangun model, sementara data *testing* digunakan untuk menguji performasi model yang dibentuk oleh data *training*. Namun, metode ini memiliki kekurangan karena hasil evaluasi mungkin akan jauh berbeda tergantung pada pembagian data ke dalam set *training* maupun *testing*.

K-fold cross validation adalah pengembangan metode *holdout* agar performasinya lebih baik. Kumpulan data dibagi menjadi *k subset* dan metode *holdout* diulang sebanyak *k* kali. Setiap pengulangan, salah satu dari *k subset* digunakan untuk data *testing* dan sisanya, *k-1 subset*, digunakan untuk membentuk model. Performasi dihitung dari rata-rata akurasi dari semua *k* percobaan. Metode ini memerlukan waktu komputasi yang lebih lama karena algoritma *cross validation* harus diulang dari awal sebanyak *k* kali. Meskipun demikian, pada penelitian ini *k-fold cross validation* tetap menjadi algoritma pilihan untuk seleksi *hyperparatemer* karena bias dan variansi yang dihasilkan oleh teknik ini lebih kecil dibandingkan *cross validation*.

2.4 Evaluasi Performasi Klasifikasi

Ketepatan klasifikasi suatu *classifier* secara keseluruhan dapat diukur melalui akurasi dan secara khusus melalui sensitivitas untuk kelas positif serta spesifisitas untuk kelas negatif. Pada kasus klasifikasi biner, ukuran ketepatan klasifikasi dapat dilihat melalui tabel klasifikasi yang ditunjukkan oleh Tabel 1.

Tabel 1: Tabel Klasifikasi

Kondisi Sebenarnya	**Hasil Prediksi**	
	Positif	**Negatif**
Positif	TP	FN
Negatif	FP	TN

Keterangan dari masing-masing istilah pada Tabel 1 adalah sebagai berikut [13]:
- TP: *True Positive*, yatu hasil prediksi dan kondisi sebenarnya sama-sama positif.
- TN: *True Negative*, yaitu hasil prediksi dan kondisi yang sebenarnya sama-sama negatif.
- FP: *False Positive*, yaitu hasil prediksi positif namun sebenarnya negatif.
- FN: *False Negative*, hasil prediksi negatif namun sebenarnya positif.

Akurasi, sensitivitas dan spesifisitas dapat dihitung dengan rumus berikut:

$$\text{a) Akurasi} = \frac{TN+TP}{TN+TP+FN+FP} \tag{10}$$

$$\text{b) Sensitivitas} = \frac{TP}{TP+FN} \tag{11}$$

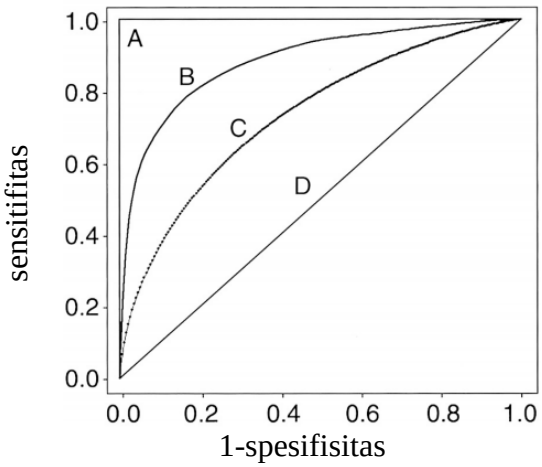
$$\text{c) Spesifisitas} = \frac{TN}{TN+FP} \tag{12}$$

2.5 Kurva *Receiving Operating Characteristics* (ROC)

Kurva ROC (*Receiver Operating Characteristics*) adalah representasi grafis dari hubungan timbal balik antara sensitivitas dan spesifisitas [14]. Kurva ini sering digunakan untuk mengevaluasi model klasifikasi karena mempunyai kemampuan secara menyeluruh dan cukup baik [15]. Beberapa informasi yang didapatkan dari kurva ROC adalah sebagai berikut:

i. Menunjukkan *tradeoff* antara sensitivitas dan spesifisitas dimana peningkatan sensitivitas akan diikuti dengan penurunan spesifisitas.
ii. Semakin dekat kurva ke sudut kiri atas, semakin tinggi akurasi keseluruhan model klasifikasi [16].
iii. Semakin dekat kurva membentuk sudut 45 derajat, semakin rendah akurasi model klasifikasi.
iv. Area di bawah kurva (AUC) adalah ukuran akurasi model. Semakin tinggi nilai AUC maka semakin baik model dalam memprediksi kelas negatif sebagai kelas negatif dan kelas positif sebagai kelas positif.

Kurva ROC memplot *True Positive Rate* (TPR= sensitivitas) terhadap *False Positive Rate* (FPR= 1-spesifisitas) di mana TPR berada pada sumbu *y* dan FPR pada sumbu *x*. Pada Gambar 1, kurva A memiliki nilai AUC sebesar 1 yang menunjukkan bahwa model dapat membedakan antara kelas negatif dan positif dengan sempurna. Sebaliknya, kurva D yang memiliki nilai AUC sebesar 0,5 menunjukkan performasi model paling buruk karena model mengklasifikasikan data secara acak. Umumnya, kurva ROC yang terbentuk dari model klasifikasi adalah kurva B dan C. Dalam hal ini, kurva B memiliki luas area di bawah kurva ROC lebih besar dari kurva C sehingga model klasifikasi yang diwakili oleh kurva B memiliki performasi klasifikasi lebih baik dari model klasifikasi yang diwakili oleh kurva C. Kategori pengklasifikasian model berdasarkan nilai AUC diberikan oleh Tabel 2 [17].



Gambar 1: Kurva AUC dengan nilai AUC yang berbeda-beda

Tabel 2: Kategori Pengklasifikasian Model Berdasarkan Nilai AUC

Nilai AUC	Model Diklasifikasikan Sebagai
0,90-1,00	*Excellent*
0,80-0,90	*Good*
0,70-0,80	*Fair*
0,60-0,70	*Poor*
0,50-0,60	*Fail*

3 Metode Penelitian

3.1 Data

Data akses internet di Jawa Timur diperoleh dari Survei Sosial Ekonomi Nasional (SUSENAS) 2017 oleh Badan Pusat Statistik (BPS) yang melibatkan 300.000 responden se-Indonesia. Dari total responden tersebut, 25.248 responden di antaranya merupakan anak-anak dan remaja dewasa berumur 6 sampai 21 tahun di Jawa Timur dengan persentase sebanyak 48% mengakses internet dalam tiga bulan terakhir. Data tersebut yang kemudian digunakan dalam penelitian ini. Setiap responden memiliki satu respon dengan 6 variabel prediktor seperti yang ditampilkan pada Tabel 3. Variabel prediktor ini selanjutnya akan disebut *features* dalam SVM.

Tabel 3: Variabel Penelitian

Variabel	Deskripsi Variabel	Kategori	Skala
Y	Mengakses internet dalam tiga bulan terakhir	0 : Tidak 1 : Ya	Nominal
X_1	Menggunakan komputer dalam tiga bulan terakhir	0 : Tidak 1 : Ya	Nominal
X_2	Menggunakan HP dalam tiga bulan terakhir	0 : Tidak 1 : Ya	Nominal
X_3	Jenis Kelamin	0 : Perempuan 1 : Laki – laki	Nominal
X_4	Umur	-	Rasio
X_5	Jenjang pendidikan tertinggi yang sedang/ pernah diikuti	0 : ≤ SD / Sederajat 1 : SMP / Sederajat 2 : ≥ SMA/ Sederajat	Ordinal
X_6	Tempat tinggal	0 : Pedesaan 1 : Perkotaan	Nominal

3.2 Prosedur Analisis

Prosedur analisis secara garis besar dibagi menjadi tiga bagian:

1. Pra-pemrosesan data (*data preprocessing*)
   - Membuat variabel *dummy* untuk semua variabel kategori.
   - Menskala ulang (*rescaling*) variabel kontinyu.
   - Membagi data menjadi data *training* dan *testing* dengan rasio 9:1.
2. Membangun model SVM. Langkah-langkahnya adalah sebagai berikut:
   - Melakukan seleksi parameter terbaik dari fungsi kernel RBF menggunakan data *training*.
   - Membangun model SVM dengan parameter terbaik dari fungsi kernel RBF.
3. Melakukan evaluasi performasi model SVM dengan data *testing*.
   - Menghitung akurasi, sensitivitas, dan spesifisitas.
   - Membuat ROC plot dan menghitung AUC.

4 Hasil dan Pembahasan

4.1 Variabel *Dummy* dan *Feature Scaling*

Variabel kategori tidak dapat digunakan secara langsung di SVM karena metode ini didasarkan pada jarak Euclidean. Oleh karena itu, perlu dibuat variabel *dummy* yang mewakili variabel kategori untuk mendefinisikan matriks jarak. Bila satu variabel kategori memiliki *m* level maka perlu dibentuk *m* variabel *dummy*. Sebagai contoh, variabel X_1 yang memiliki dua level 0: Tidak dan 1: Ya akan digantikan oleh variabel *dummy* X_1.Tidak dan X_1.Ya. Jika pada suatu baris X_1= "Tidak" maka pada variabel *dummy* X_1.Tidak= 1 dengan X_1.Ya= 0. Begitu pula bila suatu baris X_1= "Ya", nilai pada variabel *dummy* X_1.Tidak= 0 dengan X_1.Ya= 1. Tabel 4 menampilkan variabel *dummy* yang terbentuk untuk setiap variabel kategori. Pembentukan variabel *dummy* ini meningkatkan dimensi data secara signifikan. Namun, selama jumlah level dalam variabel kategori tidak terlalu banyak, proses pembentukan model dengan variabel *dummy* lebih stabil dibandingkan dengan menggunakan variabel asli.

Tabel 4: Variabel Dummy untuk Setiap Variabel Kategori

Variabel	Deskripsi Variabel	Kategori	Variabel *Dummy*
X_1	Menggunakan komputer dalam tiga bulan terakhir	0 : Tidak 1 : Ya	$X_{1.}$Tidak $X_{1.}$Ya
X_2	Menggunakan HP dalam tiga bulan terakhir	0 : Tidak 1 : Ya	$X_{2.}$Tidak $X_{2.}$Ya
X_3	Jenis Kelamin	0 : Perempuan 1 : Laki – laki	$X_{3.}$Perempuan $X_{3.}$Laki-laki
X_5	Jenjang pendidikan tertinggi yang sedang/ pernah diikuti	0 : ≤ SD/ Sederajat 1 : SMP/ Sederajat 2 : ≥ SMA/ Sederajat	$X_{5.}$≤ SD/ Sederajat $X_{5.}$SMP/ Sederajat $X_{5.}$≥ SMA/ Sederajat
X_6	Tempat tinggal	0 : Pedesaan 1 : Perkotaan	$X_{6.}$Pedesaan $X_{6.}$Perkotaan

Setelah mengubah variabel kategori ke variabel *dummy*, langkah selanjutnya adalah melakukan *scaling* terhadap variabel kontinyu. Fungsi kernel RBF menggunakan jarak euclidean untuk membandingkan dua sampel. Bila setiap *feature* memiliki *range* skala yang berbeda-beda maka jarak euclidean hanya akan memperhitungkan *feature* dengan *range* skala paling tinggi. Pada penelitian ini variabel kontinyu X_4 akan diskala ulang (*rescaling*) sehingga data berada pada skala 0 sampai 1 yang menyesuaikan *range* data pada variabel *dummy*.

Selanjutnya, pembentukan model SVM akan menggunakan variabel-variabel *dummy* pada Tabel 4 ditambah dengan satu variabel kontinyu X_4 yang telah diskala ulang sehingga terdapat total 12 *features*.

4.2 Seleksi Parameter

Seleksi parameter melibatkan parameter regulasi SVM (v) dan parameter fungsi kernel; dalam hal ini μ yang merupakan parameter fungsi kernel RBF. Parameter v mengontrol *trade-off* antara menemukan garis yang memaksimalkan margin dan meminimalkan kesalahan klasifikasi. Sementara itu, parameter μ pada kernel RBF berfungsi untuk mengontrol pengaruh titik baru terhadap *decision boundary* yang memisahkan kedua kelas. Ketika nilai μ terlalu kecil, model SVM tidak bisa menangkap kompleksitas dan bentuk data. Sebaliknya, semakin besar nilai μ, semakin besar kemungkinan terjadinya kasus *overfitting*. Pada kasus *overfitting*, *classifier* memiliki kinerja yang tinggi saat memodelkan data *training* namun rendah saat memodelkan data baru.

Saat melakukan seleksi paramater, perlu diberikan *range* nilai μ dan v untuk kemudian dicari kombinasi terbaiknya melalui metode *grid search* dan 10-*fold cross validation.* Pada penelitian ini, proses seleksi parameter melibatkan 22.723 data dengan nilai μ = 0,5; 1; 1,5; 2 dan v = 0,25; 0,5; 0,75; 1. Hasil dari seleksi parameter disajikan dalam Tabel 5.

Tabel 5: Hasil Seleksi Parameter Menggunakan *Grid Search* dan 10-*fold Cross Validation*

No.	μ	v	Akurasi
1	0,5	0,25	86,66%
2	1,0	0,25	86,72%
3	1,5	0,25	86,78%
4	2,0	0,25	86,85%
5	0,5	0,50	86,69%
6	1,0	0,50	86,80%
7	1,5	0,50	86,87%
8	2,0	0,50	86,97%
9	0,5	0,75	86,72%
10	1,0	0,75	86,83%
11	1,5	0,75	86,98%
12	2,0	0,75	86,94%
13	0,5	1,00	86,70%
14	1,0	1,00	86,85%
15	1,5	1,00	86,96%
16	2,0	1,00	86,91%

Berdasarkan Tabel 5 kombinasi parameter (μ, v) terbaik adalah (1,5; 0,75) dengan tingkat akurasi 86,98%. Parameter terpilih ini kemudian dimasukkan ke dalam *classifier* SVM.

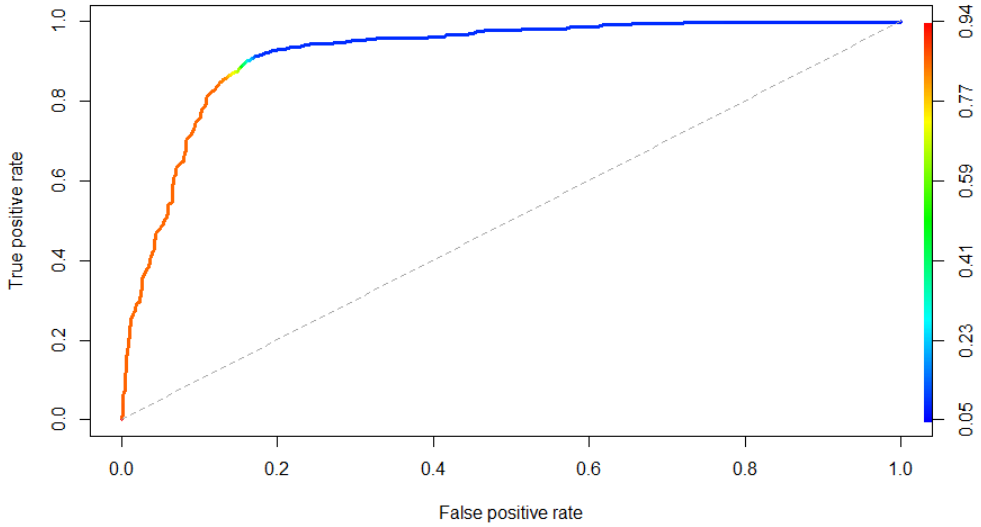
4.3 Evaluasi Kinerja Klasifikasi

Algoritma *classifier* SVM yang terbentuk dijalankan pada data *testing* penggunaan internet untuk mengevaluasi kinerja model klasifikasi. Penerapan model ini akan menghasilkan kelas prediksi yang kemudian dibandingkan dengan kelas asli dari data *testing*. Performasi klasifikasi metode SVM dapat dihitung melalui tabel klasifikasi yang disajikan oleh Tabel 6.

Tabel 6: Tabel klasifikasi dari model SVM yang terbentuk

Kondisi Sebenarnya	Hasil Prediksi	Kondisi Sebenarnya
	Positif (0: Tidak)	Negatif (1: Ya)
Positif (0: Tidak)	3351	411
Negatif (1: Ya)	608	3203

Berdasarkan Tabel 6, diperoleh akurasi model sebesar 86,54% serta sensitivitas dan spesifisitas berturut-turut adalah 84,64% dan 88,63%. Artinya, model SVM yang terbentuk secara keseluruhan mampu mengklasifikasikan responden baru sebagai pengguna atau bukan pengguna internet dengan tepat sebesar 86,54%. Sementara itu, ketepatan model dalam mengklasifikasikan responden baru yang tidak menggunakan internet dengan tepat adalah sebesar 84,64% dan 88,63% untuk responden yang menggunakan internet.



Gambar 2: Kurva ROC dari evaluasi model SVM

Kurva ROC yang terbentuk pada Gambar 2 merepresentasikan rata-rata nilai sensitivitas untuk semua kemungkinan nilai spesifisitas dengan luasan di bawah kurva (AUC) sebesar 0,92. Meskipun akurasi model hanya 86,54%, namun berdasarkan nilai AUC model SVM yang terbentuk masih termasuk model yang kinerjanya sangat bagus (*excellent*).

5 Kesimpulan dan Saran

Metode SVM memiliki performa yang *excellent* berdasarkan nilai AUC untuk klasifikasi akses internet oleh anak-anak dan remaja dewasa di Jawa Timur. Secara umum, model SVM yang terbentuk mampu mengklasifikan kasus dengan benar sebesar 86,45%. Sementara itu, sebesar 84,64% responden baru yang tidak menggunakan internet dalam tiga bulan terakhir dapat diklasifikasn dengan tepat dan sebesar 88,63% responden baru yang menggunakan internet diklasifikasikan dengan benar.

Meskipun kinerja model SVM yang dihasilkan oleh penelitian ini dikategorikan sangat bagus, namun nilai akurasi yang didapatkan tidak berbeda jauh dengan algoritma klasifikasi tradisional. Oleh karena itu pada penelitian selanjutnya, penambahan variabel numerik sebagai prediktor dapat dipertimbangkan untuk meningkatkan kinerja algoritma SVM. Selain itu, mengingat waktu komputasi dari SVM yang terbilang lama maka penggunaan metode RSVM (*Reduced Support Vector Machine*) dapat diimplementasikan.

Daftar Pustaka

[1] Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), "Survei APJII yang Ditunggu-tunggu, Penetrasi Internet Indonesia 2018," *Buletin APJII*, p. 1, Mei 2019.

[2] Asosiasi Penyelenggara Jasa Internet Indonesia, "Penetrasi dan Profil Perilaku Pengguna Internet di Indonesia," Asosiasi Penyelenggara Jasa Internet Indonesia, Jakarta, Infografis 2019.

[3] Ryan R. Nurrahman, "Analisis Faktor-Faktor yang Mempenharuhi Kepala Rumah Tangga Jawa Timur dalam Mengakses Internet Tahun 2017," Institut Teknologi Sepuluh Nopember, Surabaya, Tugas Akhir 2017.

[4] Octavia Dian Pratama Nia Anggraini, "Analisis Faktor-Faktor yang Mempengaruhi Anak-Anak dan Remaja di Jawa Timur Dalam Mengakses Internet Menggunakan Regresi Logistik Biner," Institut Teknologi Sepuluh Nopember, Surabaya, Tugas Akhir 2019.

[5] Corinna Cortes and Vladimir Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, September 1995. [Online]. http://link.springer.com/article/10.1007%2FBF00994018?LI=true

[6] T Verplancke et al., "Support Vector Machine Versus Logistic Regression Modeling for Prediction of Hospital Mortality in Critically Ill Patients with Haematological Malignancies," *BMC Medical Informatics and Decision Making*, vol. 8, p. 56, Desember 2008. [Online]. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2612652/

[7] Chih W. Hsu, Chih C. Chang, and Chih J. Lin, "A practical guide to Support Vector Classification," Information Engineering, Naional Taiwan University, Taipe, 2008.

[8] Noviyanti Santoso, "Comparative Study of Kernel Function for Support Vector Machine on Financial Dataset," *International Journal of Soft Computing*, vol. 13, no. 4, pp. 129-133, 2018.

[9] Budi Santosa, *Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis/Studi*, 1st ed. Yogyakarta: Graha Ilmu, 2007.

[10] S. Sathiya Keerthi and Chih-Jen Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667–1689, July 2003.

[11] Hsuan-Tien Lin and Chih-Jen Lin, "A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods," Department of Computer Science, National Taiwan University, Technical report 2003. [Online]. http://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf

[12] Chenghai Yang et al., "Evaluating unsupervised and supervised image classification methods for mapping cotton root rot.," *Precision Agriculture*, vol. 16, pp. 201-215, April 2015.

[13] Wen Zhu, Nancy Zeng, and Ning Wang, "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations," *Health Care and Life Sciences*, 2010.

[14] Arian R. V. Erke and Peter M. Th. Pattynama, "Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology," *European Journal of Radiology*, pp. 88-94, 1998.

[15] S.M. Chou, J.W. Shan, Y. Guo, and L. Zhang, "Automated Breast Cancer Detection and Classification Using Ultrasound Image : A Survey, Pattern Recognition," vol. 43, pp. 299-317, 2010.

[16] Mark H. Zweig and Gregory Campbell, "Receiver Operating Characteristic (ROC) Plots : A Fundamental Evaluation Clinical Medicine," *Clinical Chemistry*, pp. 561-577, 1993.

[17] Ertugrul Colak et al., "Comparison of Semiparametric, Parametric, and Nonparametric ROC Analysis for Continuous Diagnostic Tests Using a Simulation Study and Acute Coronary Syndrome Data," *Computational and Mathematical Methods in Medicine*, vol. 2012, p. 7, 2012. [Online]. http://downloads.hindawi.com/journals/cmmm/2012/698320.pdf