PENERAPAN MODEL KLASIFIKASI REGRESI LOGISTIK, *SUPPORT VECTOR MACHINE*, *CLASSIFICATION AND REGRESSION TREE* TERHADAP DATA KEJADIAN DIFTERI DI PROVINSI JAWA BARAT

Hilman Dwi Anggana

Teknik Industri, Universitas Telkom, Jalan Telekomunikasi No. 01 Bandung; hilmandwianggana@telkomuniversity@ac.id

Abstrak

Salah satu permasalahan yang dihadapi Jawa Barat selama beberapa waktu terakhir adalah adanya kejadian luar biasa (KLB) penyakit Difteri. Upaya preventif untuk mengurangi merebaknya wabah suatu penyakit harus terus dilakukan seperti program sosialisasi, vaksinasi dan karantina. Selain dengan program yang telah disebutkan, kajian suatu penyakit dengan menggunakan pemodelan klasifikasi secara statistika menjadi salah satu alternatif dalam mendukung *early warning system* (EWS) suatu kejadian penyakit. Pada penelitian ini dilakukan penerapan model klasifikasi regresi logistik, *support vector machine* (SVM) dan *classification and regression tree* (CART) terhadap data kejadian Difteri di provinsi Jawa Barat. Hasil analisis menunjukkan bahwa model regresi logistik merupakan model yang kurang tepat diterapkan diantara tiga pilihan model ini karena memiliki nilai AUC terendah (nilai AUC sekitar 50%), didukung oleh tingkat akurasi dan tingkat ketepatan model mengklasifikasikan kelas positif (*sensitivity*) yang rendah. Sementara itu model yang paling tepat diterapkan adalah model SVM karena memiliki nilai AUC tertinggi (nilai AUC jauh diatas 50%), didukung oleh tingkat akurasi dan tingkat *sensitivity* yang tinggi.

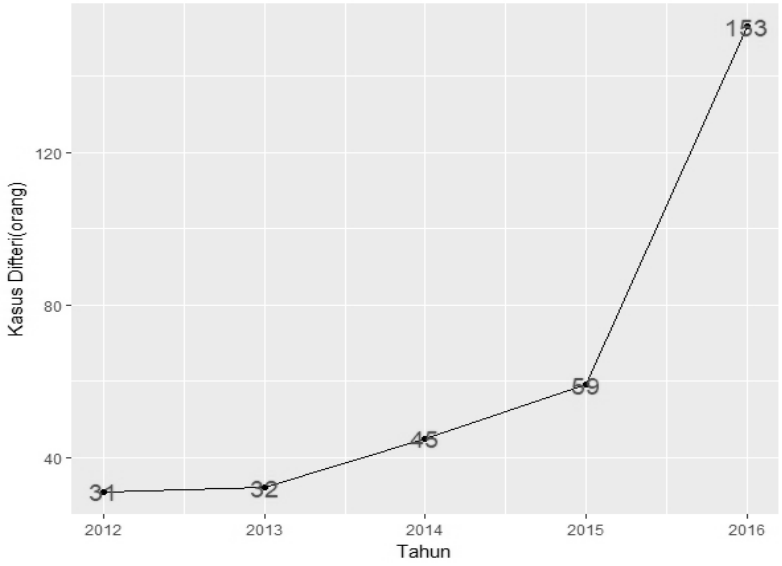
Kata Kunci. Penyakit Difteri, Model Klasifikasi, Regresi Logistik, SVM, CART

1. Pendahuluan

Jawa Barat merupakan provinsi di Indonesia yang memiliki program pembangunan dengan indikator kesehatan masyarakat. Hal ini tercermin dalam visi pembangunan kesehatan Jawa Barat yaitu "Tercapainya masyarakat Jawa Barat yang mandiri untuk hidup sehat". Dinas Kesehatan Jawa Barat sebagai pendorong, penggerak, fasilitator, dan advokator untuk terjadinya akselerasi pembangunan kesehatan di

Jawa Barat memiliki beberapa misi, salah satunya adalah meningkatkan sistem surveilans dalam upaya pencegahan dan pengendalian penyakit (http://www.diskes.jabarprov.go.id/index.php/pages/detail/2014/7/Visi-dan-Misi).

Profil kesehatan Jawa Barat merupakan output yang diperoleh dari laporan dan evaluasi program kesehatan selama satu tahun termasuk juga bagaimana proses surveilans dalam upaya pencegahan dan pengendalian penyakit dilakukan. Fakta menarik yang sering ditemukan pada laporan profil kesehatan Jawa Barat terkait upaya pencegahan dan pengendalian penyakit adalah adanya kejadian luar biasa (KLB) penyakit, salah satunya adalah penyakit Difteri. Informasi yang diperoleh dari kepala dinas kesehatan Jawa Barat, Suhendar, mengatakan bahwa hingga pertengahan Desember 2017 telah terjadi 153 kasus Difteri dengan korban meninggal 14 orang yang tersebar di beberapa kota dan kabupaten di Jawa Barat, seperti kabupaten Purwakarta, Karawang, Bekasi dan kota Depok (https://regional.kompas.com/read/2017/12/15/08124791/14-orang-meninggal-akibat-difteri-di-jawa-barat) . Hal ini diperkuat laporan profil kesehatan Jawa Barat tahun 2016 bahwa selama periode 2012 – 2016 telah terjadi peningkatan KLB Difteri, dengan distribusi kasus setiap tahun, 2012 (31 kasus), 2013 (32 kasus), 2014 (45 kasus), 2015 (59 kasus), dan 2016 (153 kasus). Dengan demikian, ini menjadi indikasi bahwa KLB Difteri di Jawa Barat bukanlah sesuatu yang baru bahkan fenomenanya semakin mengkhawatirkan pada tahun 2016 dengan kenaikan kasus hampir 300%



Gambar 1. Kasus Difteri Jawa Barat 2012 - 2016

Upaya preventif untuk mengurangi merebaknya wabah suatu penyakit sering dilakukan pemerintah dan instansi terkait, dalam hal ini dinas kesehatan, diantaranya adalah program vaksinasi dan karantina masyarakat. Selain dengan program vaksinasi dan karantina masyarakat, pemerintah dan instansi terkait juga dapat melakukan sosialisasi terkait jenis penyakit, karakteristik penyakit, faktor-faktor terjadinya penyakit dan pola penyebaran penyakit sebagai bagian dari *Early Warning System* (EWS) suatu penyakit menular di suatu daerah. EWS merupakan komponen penting dalam manajemen risiko penyakit sehingga kajian yang berkaitan dengan EWS suatu penyakit menarik untuk dilakukan. Salah satu teknik yang dapat dilakukan untuk mengkaji kejadian suatu penyakit adalah menggunakan model klasifikasi atau prediksi secara statistika.

Beberapa penelitian mengenai penyakit Difteri telah dilakukan diantaranya, Kartono (2005) meneliti hubungan lingkungan rumah dan kejadian Difteri di Kabupaten Tasikmalaya dan Garut, Kusuma (2012) meneliti faktor-faktor yang berhubungan dengan kejadian Difteri di Kabupaten Sidoarjo, Arifin dan Prasasti (2016) meneliti faktor-faktor yang berhubungan dengan kasus Difteri anak di Puskesmas Bangkalan, Puspita dkk. (2017) meneliti penyebaran penyakit Difteri dengan pengaruh karantina dan vaksinasi menggunakan model matematika. Berdasarkan beberapa penelitian yang telah dilakukan ini, belum ada yang mengkaji mengenai klasifikasi atau prediksi kasus Difteri berdasarkan faktor-faktor risiko yang ada khususnya di provinsi Jawa Barat.

Pada penelitian ini, terdapat dua pendekatan yang dilakukan dalam pemodelan, yaitu pendekatan statistika tradisional yang mengasumsikan data berdistribusi tertentu dan pendekatan statistika yang tidak mensyaratkan data berdistribusi tertentu. Untuk teknik klasifikasi statistika tradisional yang digunakan adalah model regresi logistik. Sedangkan pada pendekatan statistika yang tidak mensyaratkan data berdistribusi tertentu digunakan model *support vector machine* (SVM) dan *classification and regression tree (*CART*)*. Tujuan dari penelitian ini adalah menerapkan model klasifikasi regresi logistik, SVM, dan CART terhadap data kejadian Difteri di provinsi Jawa Barat kemudian membandingkan dan menganalisis hasil yang diperoleh sehingga menjadi informasi tambahan yang dapat digunakan sebagai EWS bagi instansi terkait.

2. Tinjauan Pustaka

2.1. Model Klasifikasi

Model klasifikasi adalah model atau fungsi M yang memprediksi label kelas $\hat{y}$ berdasarkan variabel masukan $\boldsymbol{x}$ sehingga diperoleh $\hat{y} = M(\boldsymbol{x})$ dimana $\hat{y} \in \{c_1, c_2, \dots, c_k\}$ adalah label kelas prediksi (Zaki dan Meira, 2014). Jika variabel label kelas $\hat{y}$ adalah variabel numerik maka model klasifikasi adalah model prediksi, contohnya regresi linier ganda. Untuk membangun sebuah model klasifikasi dapat dilakukan dengan dua tahap yaitu *learning step* dan *classification step* (Han dkk., 2012). *Learning step* adalah sebuah tahapan untuk membangun model klasifikasi berdasarkan *training data*, sedangkan *classification step* adalah tahapan untuk memprediksi label kelas dengan model yang telah dibangun berdasarkan data yang diberikan atau *testing data*, tetapi pada prakteknya *classification step* dapat juga diterapkan pada *training data*. Beberapa model klasifikasi yang sering digunakan diantaranya adalah regresi logistik, *support vector machine*, *classification and regression tree*, *neural network*, *discriminant analysis*, *naïve bayes*, *k-nearest neighbor*, tetapi pada penelitian ini hanya akan membahas tiga model klasifikasi pertama.

2.2. Regresi Logistik

Model regresi logistik adalah salah satu model tradisional yang dapat digunakan untuk klasifikasi data. Penerapan regresi logistik yang sering dilakukan diantaranya adalah klasifikasi calon nasabah baru kartu kredit berdasarkan variabel karakteristik calon nasabah, klasifikasi calon mahasiswa pascasarjana berdasarkan variabel karakteristik calon mahasiswa, klasifikasi status kota/kabupaten berdasarkan variabel ekonomimaikro. Jika Y adalah variabel dependen/label kelas biner maka :

$$Y = \begin{cases} 1 \text{ ; jika ada} \\ 0 \text{ ; jika tidak ada} \end{cases} \quad (1)$$

dengan $\boldsymbol{x} = [x_1, x_2, \dots, x_p]$ adalah vektor variabel independen dan $\pi(\boldsymbol{x}) = E(Y|x_1, x_2, \dots, x_p)$ adalah peluang sebuah unit observasi diklasifikasikan menjadi anggota dari salah satu kelas biner. Model regresi logistik menurut Hosmer dan Lemeshow (2000) :

$$\pi(\boldsymbol{x}) = \frac{exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\}}{1 + exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\}} \quad (2)$$

Model peluang $\pi(\boldsymbol{x})$ diatas tidak linear terhadap parameter model sehingga dilakukan tranformasi logit dimana $logit(\pi(\boldsymbol{x})) = log(\pi(\boldsymbol{x})/1 - \pi(\boldsymbol{x}))$. Jika sebuah unit observasi baru $\boldsymbol{x}^* = [x_1^*, x_2^*, \dots, x_p^*]$ maka peluang unit observasi baru masuk kedalam salah satu kelas biner adalah :

$$\hat{\pi}(\boldsymbol{x}^*) = \frac{exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \cdots + \hat{\beta}_p x_p^*\}}{1 + exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \cdots + \hat{\beta}_p x_p^*\}} \quad (3)$$

`

2.3 *Support Vector Machine*

Permasalahan klasifikasi pada umumnya adalah jarang menemukan fungsi yang *linear separable*. Metode yang dapat digunakan untuk permasalahan klasifikasi yang *nonlinear* adalah menggunakan pendekatan statistika nontradisional yang tidak mengasumsikan data berdistribusi tertentu, salah satu model klasifikasinya adalah model *support vector machine* (SVM). Andaikan Φ adalah fungsi *kernel* dan $\boldsymbol{x}_i = [x_{1i}, x_{2i}, \dots, x_{pi}]$ adalah fitur atau variabel independen observasi ke-*i*, $\boldsymbol{x}_i$ dipetakan oleh fungsi *kernel* menjadi $\boldsymbol{x}_i \rightarrow \Phi(\boldsymbol{x}_i)$ sehingga fungsi *hyperlane* yang *linear separable* dapat dinyatakan

$$\boldsymbol{w}\Phi(\boldsymbol{x}_i) + b = 0 \quad (4)$$

dengan $\boldsymbol{w}$ dan b adalah parameter model. Jika kelas biner pertama $y_i = -1$ dan kelas biner kedua $y_i = 1$ maka

$$\boldsymbol{w}\Phi(\boldsymbol{x}) + b \begin{cases} \geq 1, \text{ jika } y_i = 1 \\ \leq -1, \text{ jika } y_i = -1 \end{cases} \quad (5)$$

Untuk menaksir parameter model $\boldsymbol{w}$ dan b dilakukan dengan cara memaksimumkan *margin* (jarak), yang mana *margin* antara observasi terdekat setiap kelas terhadap *hyperplane* adalah $1/\|\boldsymbol{w}\|$ dan *margin* antar kelas adalah $2/\|\boldsymbol{w}\|$. Permasalahan ini dapat diformulasikan menggunakan metode *quadratic programming* sehingga meminimumkan $\|\boldsymbol{w}\|^2$ dengan kendala $y_i(\boldsymbol{w}\Phi(\boldsymbol{x}_i) + b) \geq 1$, $i = 1,2, \dots, n$ (Prasetyo, 2014).

Beberapa fungsi kernel menurut Prasetyo (2014) adalah sebagai berikut :

1. Kernel *linear* : $K(x_i, x_j) = x_i \,.\, x_j$
2. Kernel *polynomial* : $K(x_i, x_j) = (x_i \,.\, x_j + c)^2$
3. Kernel *radial basis function* : $K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$ (6)
4. Kernel *sigmod* : $K(x_i, x_j) = \tanh(\sigma(x_i \,.\, x_j) + c)$
5. Kernel *invers multiquadratic* : $K(x_i, x_j) = \frac{1}{\sqrt{\|x_i - x_j\|^2 - c^2}}$

2.4 *Classification and Regression Tree*

Model *classification and regression tree* (CART) merupakan teknik klasifikasi statistika non tradisional lainnya yang mensyaratkan data tidak berdistribusi tertentu. Pada model CART dikembangkan sebuah pohon keputusan untuk menyelesaikan permasalahan klasifikasi atau regresi berdasarkan variabel karakteristik/independen yang menyertainya. Jika variabel dependen yang digunakan adalah variabel kategori, model CART akan menghasilkan sebuah pohon klasifikasi dan jika variabel dependennya numerik akan dihasilkan sebuah

pohon regresi. Prinsip dari metode CART adalah melakukan penyekatan terhadap data secara rekursif

Secara umum tahapan dalam membangun sebuah model CART ada tiga (Deconinck dkk., 2005). Tahap pertama, membangun pohon secara maksimal menggunakan prosedur penyekatan biner. Selanjutnya, tahap *prunning* untuk menghasilkan pohon keputusan yang tidak *overfitting*. Tahap terakhir, memilih pohon keputusan optimum berdasarkan evaluasi terhadap *predictive error*. Untuk lebih lengkapnya, penjelasan mengenai metode CART dapat ditemukan pada Deconick dkk. (2005) dan Johra (2017).

2.5 *Confusion Matrix*

Model klasifikasi yang telah dibangun berdasarkan *training data* perlu diukur performanya. Salah satu cara yang dapat dilakukan untuk mengukur performansi model klasifikasi adalah dengan membuat *confusion matrix*. Prinsip dasar *confusion matrix* adalah membuat tabel perbandingan hasil klasifikasi yang diperoleh berdasarkan model dan hasil klasifikasi yang aktual. Pada jenis klasifikasi untuk label kelas biner, *confusion matrix* dapat disajikan pada Tabel 1dibawah ini
`

Tabel 1. *Confusion Matrix* untuk Label Kelas Biner

Aktual	Hasil Klasifikasi	
	Kelas Positif	Kelas Negatif
Kelas Positif	*True Positive* (TP)	*False Negative* (FN)
Kelas Negatif	*False Positive* (FP)	*True Negative* (TN)

Keterangan :
TP : Banyaknya anggota yang terklasifikasikan benar dari kelas positif
TN : Banyaknya anggota yang terklasifikasikan benar dari kelas negatif
FP : Banyaknya anggota yang terklasifikasikan salah dari kelas positif
FN : Banyaknya anggota yang terklasifikasikan salah dari kelas negatif

Ukuran yang sering digunakan untuk mengukur performa sebuah model klasifikasi adalah tingkat akurasi, *sensitivity*, dan *specifity*. Tingkat akurasi merupakan rasio antara banyaknya anggota yang terklasifikasikan benar (TP+TN) dengan banyaknya anggota keseluruhan (TP+FN+FP+TN). *Sensitivity* atau *true positive rate* (TPR) merupakan rasio antara banyaknya anggota kelas positif yang terklasifikasikan benar (TN) dengan banyaknya anggota kelas positif keseluruhan (TP+FN).

Sedangkan *specifity* atau *true negative rate* (TNR) merupakan rasio antara banyaknya anggota kelas negatif yang terklasifikasikan benar (TN) dengan banyaknya anggota kelas negatif keseluruhan (FP+TN).

Selain menggunakan nilai akurasi, *sensitivity*, dan *specifity*, untuk mengukur performa model klasifikasi dapat dilakuan secara visual menggunakan kurva *receiver operating characteristic* (ROC). Untuk mempermudah intrepetasi kurva ROC diubah kedalam bentuk skalar melalui AUC. Jika nilai AUC mendekati satu maka akurasi model klasifikasi semakin tinggi sehingga model semakin baik (Fawcett, 2006).

3. Metodologi Penelitian

3.1. Data

Data yang digunakan pada penelitian ini adalah data kejadian Difteri untuk setiap kota/kabupaten di Provinsi Jawa Barat tahun 2012 – 2016 dan faktor-faktor risiko yang menyertainya yang diperoleh dari Profil Dinas Kesehatan Provinsi Jawa Barat (http://www.diskes.jabarprov.go.id/index.php/arsip/categories/MTEz/profile-kesehatan). Variabel dependen adalah variabel kelas biner dengan nilai $y = 0$ dan $y = 1$. Jika kota/kabupaten mengalami kejadian Difteri maka $y = 1$ dan sebaliknya jika tidak mengalami kejadian Difteri maka $y = 0$. Variabel independen adalah faktor-faktor risiko diantaranya Persentase Rumah Sehat (X_1), Persentase Akses Sanitasi Jamban Layak (X_2), dan Persentase Tempat Umum dan Pengelolaan Makanan Sehat (X_3). Setelah dilakukan verifikasi dan *cleaning data* diperoleh sebanyak 130 unit observasi selama periode 2012 – 2016.

3.2. Metode Analisis

Secara umum, langkah-langkah dalam tahapan analisis data adalah sebagai berikut

1. Melakukan pengambilan data untuk *training data* dan *testing data* secara acak tanpa pengembalian dengan proporsi 80% untuk *training data* dan 20% untuk *testing data*.
2. Membangun model klasifikasi Regresi Logistik, SVM, dan CART menggunakan *training data*.
3. Melakukan validasi model klasifikasi yang telah dibangun pada langkah 3) terhadap *training data* dan *testing data.*

4. Menentukan *confusion matrix* kemudian hitung nilai akurasi, *sensitivity*, dan *specifity*.
5. Melakukan validasi model klasifikasi menggunakan model yang dibangun dengan semua data dengan proporsi *testing data* yang berlainan jika hasil validasi terhadap *training data* dan *testing data* berbeda.
6. Memilih model terbaik berdasarkan nilai AUC

4. Hasil dan Pembahasan

4.1. Pendahuluan

Model regresi logistik, SVM, dan CART dibangun berdasarkan *training data* yang sama. Jumlah kota/kabupaten yang digunakan sebagai unit observasi pada *training data* adalah 105 unit observasi yang diambil secara acak tanpa pengembalian dari data awal. Tahap estimasi parameter model-model klasifikasi dilakukan dengan menggunakan *software R-Studio*. Penentuan kelas biner untuk kelas positif dan kelas negatif dilakukan berdasarkan keperluan bahwa model-model klasifikasi yang dibangun lebih diutamakan untuk memprediksi kota-kabupaten dengan status terdapat kejadian Difteri sehingga kota/kabupaten yang teridentifikasi terdapat kejadian Difteri dikategorikan sebagai kelas positif.

4.2. Model Regresi Logistik

Estimasi parameter model regresi logistik pada software *R-Studio* dilakukan dengan menggunakan metode *iteratively reweighted least square*. Hasil klasifikasi menggunakan model regresi logistik pada *training data* dan *testing data* disajikan pada Tabel 2

Tabel 2. *Confusion Matrix* Model Regresi Logistik

Aktual	Hasil Klasifikasi			
	Training Data		*Testing Data*	
	Y=1	*Y=0*	*Y=1*	*Y=0*
Y=1(Terjadi Difteri)	24	28	8	4
Y=0(Tidak terjadi Difteri)	24	29	4	9

Berdasarkan Tabel 2 diperoleh informasi sebanyak 53 unit observasi pada *training data* terklasifikasikan benar dan 52 unit observasi terklasifikasikan salah. Sedangkan pada *testing data*, sebanyak 17 unit observasi terklasifikasikan benar dan 8 unit observasi terklasifikasikan salah. Hasil ini menjadi indikasi bahwa klasifikasi menggunakan model regresi logistik terhadap *training data* menghasilkan tingkat kesalahan klasifikasi yang relatif tinggi, sedangkan pada testing data menghasilkan tingkat kesalahan klasifikasi relatif rendah.

4.2. Model SVM

Estimasi parameter model SVM pada software *R-Studio* dilakukan dengan menggunakan fungsi kernel *radial basis function*. Penggunaan fungsi kernel *radial basis function* berdasarkan *trial and error* terhadap *training data*, yang mana fungsi kernel yang memberikan hasil klasifikasi terbaik adalah fungsi kernel *radial basis function*. Hasil klasifikasi menggunakan model SVM pada *training data* dan *testing data* disajikan pada Tabel 3

Tabel 3. *Confusion Matrix* Model SVM

Aktual	Hasil Klasifikasi			
	Training Data		*Testing Data*	
	Y=1	*Y=0*	*Y=1*	*Y=0*
Y=1(Terjadi Difteri)	40	12	5	7
Y=0(Tidak terjadi Difteri)	20	33	6	7

Berdasarkan Tabel 3 diperoleh informasi sebanyak 73 unit observasi pada *training data* terklasifikasikan benar dan 32 unit observasi terklasifikasikan salah. Sedangkan pada *testing data*, sebanyak 12 unit observasi terklasifikasikan benar dan 13 unit observasi terklasifikasikan salah. Hasil ini menjadi indikasi bahwa klasifikasi menggunakan model SVM terhadap *training data* menghasilkan tingkat kesalahan klasifikasi yang relatif rendah, sedangkan pada testing data menghasilkan tingkat kesalahan klasifikasi relatif tinggi.

4.3. Model CART

Estimasi parameter model CART pada software *R-Studio* dilakukan dengan banyak pemisah optimum sebesar 6 berdasarkan *predictive error* terkecil. Hasil klasifikasi menggunakan model CART pada *training data* dan *testing data* disajikan pada Tabel 4.

Tabel 4. *Confusion Matrix* Model CART

Aktual	Hasil Klasifikasi			
	Training Data		*Testing Data*	
	Y=1	*Y=0*	*Y=1*	*Y=0*
Y=1(Terjadi Difteri)	37	15	5	7
Y=0(Tidak terjadi Difteri)	11	42	5	8

Berdasarkan Tabel 4 diperoleh informasi sebanyak 79 unit observasi pada *training data* terklasifikasikan benar dan 26 unit observasi terklasifikasikan salah. Sedangkan pada *testing data*, sebanyak 13 unit observasi terklasifikasikan benar dan 12 unit observasi terklasifikasikan salah. Hasil ini menjadi indikasi bahwa klasifikasi menggunakan model CART terhadap *training data* menghasilkan tingkat kesalahan klasifikasi yang relatif rendah, sedangkan pada testing data menghasilkan tingkat kesalahan klasifikasi relatif tinggi.

4.4. Validasi *Training Data* dan *Testing Data*

Validasi model dilakukan pada *training data* dan *testing data* berdasarkan hasil klasifikasi/*confusion matrix* untuk setiap model klasifikasi. Hasil validasi untuk setiap model klasifikasi disajikan pada Tabel 5

Tabel 5. Validasi *Training Data*

Model Klasifikasi	Ukuran Performansi			
	Akurasi	*Sensitivity*	*Specitifity*	AUC
Regresi Logistik	50.48%	46.15%	54.72%	50.44%
SVM	69.52%	76.92%	62.26%	78.07%
CART	75.24%	71.15%	79.25%	77.83%

Berdasarkan Tabel 5 diperoleh informasi bahwa tingkat akurasi model klasifikasi pada *training data* bervariasi, model CART merupakan model klasifikasi dengan tingkat akurasi tertinggi (75.24%), sebaliknya model regresi logistik merupakan model klasifikasi dengan tingkat akurasi terendah (50.48%). Tingkat *sensitivity* tertinggi diperoleh untuk model SVM (76.92%) dan tingkat *sensitivity* terendah untuk model regresi logistik (46.15%). Hal ini menjadi indikasi bahwa model SVM memiliki kecenderungan untuk mengklasifikasikan *training data* ke kelas biner positif

(terjadinya Difteri). Model SVM memiliki nilai AUC tertinggi (78.07%) sedangkan model regresi logistik memiliki nilai AUC terendah (77.83%) sehingga berdasarkan kriteria nilai AUC, SVM merupakan model terbaik dan regresi logistik merupakan model terburuk. Dengan demikian, berdasarkan validasi terhadap *training data*, cukup bukti untuk mengatakan bahwa SVM merupakan model terbaik dan regresi logistik merupakan model terburuk dalam mengklasifikasikan kejadian Difteri.

Selanjutnya dilakukan validasi pada *testing data*. Hasil validasi untuk setiap model klasifikasi disajikan pada Tabel 6

Tabel 6. Validasi *Testing Data*

Model Klasifikasi	Ukuran Performansi Model Klasifikasi			
	Akurasi	*Sensitivity*	*Specitifity*	AUC
Regresi Logistik	68%	20.00%	69.23%	67.94%
SVM	48%	41.67%	53.85%	62.18%
CART	52%	41.67%	61.54%	56.73%

Berdasarkan Tabel 6 diperoleh informasi bahwa tingkat akurasi model klasifikasi pada *testing data* bervariasi, model regresi logistik merupakan model klasifikasi dengan tingkat akurasi tertinggi (68%), sebaliknya model SVM merupakan model klasifikasi dengan tingkat akurasi terendah (48%). Tingkat *sensitivity* tertinggi diperoleh untuk model SVM dan CART (41.67%) dan tingkat *sensitivity* terendah untuk model regresi logistik (20%). Hal ini menjadi indikasi bahwa model SVM dan CART memiliki kecenderungan untuk mengklasifikasikan *testing data* ke kelas biner positif (terjadinya Difteri). Model regresi logistik memiliki nilai AUC tertinggi (67.94%) sedangkan model CART memiliki nilai AUC terendah (56.73%). Dengan demikian, berdasarkan validasi terhadap *testing data*, cukup bukti untuk mengatakan bahwa regresi logistik merupakan model terbaik dan CART merupakan model terburuk dalam mengklasifikasikan kejadian Difteri.

4.5 Validasi *Testing Data* Dengan Proporsi Tertentu

Hasil validasi yang dilakukan pada *testing data* dan *training data* adalah berbeda. Salah satu penyebab hasilnya berbeda adalah karena sampel untuk *testing data* tidak mencukupi. Selanjutnya akan dilakukan validasi pada *testing data* dengan proporsi *testing data* yang berlainan terhadap model klasifikasi yang dibangun menggunakan

keseluruhan data. Hasil validasi model klasifikasi dengan proporsi *testing data* 20%, 40%, 60%, 80%, dan 100% disajikan pada Tabel 7

Tabel 7. Validasi *Testing Data* dengan Proporsi Tertentu

Proporsi Testing Data (p)	Model Klasifikasi	Ukuran Performansi Model Klasifikasi			
		Akurasi	*Sensitivity*	*Specitifity*	AUC
p=20%	Regresi Logistik	51.85%	30.77%	71.43%	51.10%
	SVM	70.37%	84.62%	57.14%	80.77%
	CART	81.48%	84.62%	78.57%	88.19%
p=40%	Regresi Logistik	50.94%	38.46%	62.96%	50.71%
	SVM	69.81%	80.77%	59.26%	77.21%
	CART	73.58%	73.08%	74.07%	76.99%
p=60%	Regresi Logistik	53.16%	33.33%	72.50%	52.92%
	SVM	72.15%	79.49%	65.00%	80.38%
	CART	77.22%	71.79%	82.50%	80.35%
p=80%	Regresi Logistik	46.67%	32.69%	60.38%	46.53%
	SVM	69.52%	76.92%	62.26%	78.07%
	CART	75.24%	71.15%	79.25%	77.83%
p=100%	Regresi Logistik	51.54%	37.50%	65.15%	51.33%
	SVM	65.38%	70.31%	60.61%	75.11%
	CART	70.77%	65.62%	75.76%	74.14%

Berdasarkan Tabel 7 diperoleh informasi bahwa untuk semua proporsi *testing data*, model CART merupakan model dengan tingkat akurasi tertinggi dan model regresi logistik dengan tingkat akurasi terendah. Tingkat *sensitivity* tertinggi diperoleh untuk model SVM dan CART untuk proporsi testing data 20% (84.62%), SVM untuk proporsi testing data 40% (80.77%), 60% (79.49%), 80% (76.92%), dan 100% (70.31%). Sedangkan model regresi logistik merupakan model dengan tingkat *sensitivity* terendah untuk semua proporsi *testing data*. Hal ini menjadi indikasi bahwa model SVM memiliki kecenderungan untuk mengklasifikasikan ke kelas biner positif (terjadinya Difteri). Untuk proporsi *testing data* 20%, model CART merupakan model terbaik karena memiliki nilai AUC tertinggi (88.19%), sedangkan untuk proporsi *testing data* lainnya, model SVM merupakan model terbaik sehingga dapat disimpulkan bahwa model terbaik untuk klasifikasi kejadian Difteri adalah model SVM.

5. Simpulan dan Saran

Pada penelitian ini, teknik klasifikasi kota/kabupaten yang mengalami kejadian Difteri dapat dilakukan dengan menggunakan model klasifikasi kelas biner yaitu regresi logistik, SVM dan CART, yang dapat digunakan sebagai informasi tambahan dalam mendukung komponen EWS KLB Difteri di Jawa Barat. Hasil analisis menunjukkan bahwa model regresi logistik merupakan model yang kurang tepat (model terburuk) diterapkan diantara tiga pilihan model ini karena memiliki nilai AUC terendah, didukung oleh tingkat akurasi dan tingkat ketepatan model mengklasifikasikan kelas positif (*sensitivity*) yang rendah. Sementara itu model yang paling tepat (model terbaik) diterapkan adalah model SVM karena memiliki nilai AUC tertinggi, didukung oleh tingkat akurasi dan tingkat *sensitivity* yang tinggi. Sebagai saran, untuk memperbaiki hasil klasifikasi, selanjutnya teknik klasifikasi dapat dilakukan dengan menggunakan model klasifikasi berbasis *ensemble method*.

Daftar Pustaka

Arifin, I.F. & Prasasti, C.I. 2017. Faktor yang Berhubungan dengan Kasus Difteri Anak di Puskesmas Bangkalan Tahun 2016. Jurnal Berkala Epidemiologi Vol 5 No 1 hlm 26-36.

Deconinck, E., Hancock, T., Coomans, D., Massart, D.L. & Van der Heyden, Y. 2005. *Classification of Drugs in Absorption Classes Using The Classification and Regression Trees Methodology*. Journal of Pharmaceutical and Biomedical Analysis 29 91-103.

Fawcett, T. 2006. *An Introduction to ROC Analysis*. *Pattern Recognition Letters* 27 pp 861-874.

Han, J., M. Kamber., dan J. Pei. 2012. *Data Mining Concepts and Techniques, third edition*. Waltham: Morgan Kaufman.

Hosmer, D.W. & Lemeshow, S. 2000. *Applied Logistic Regression, second edition.* New York: John Wiley & Sons Inc.

Johra, M.B. 2018. Perbandingan Faktor Determinan Niat Kewirausahaan Dengan *Classification and Regression Tree* di Indonesia, Filipina, dan Malaysia. Jurnal Euclid Vol 5 No 1 pp 90.

Kartono, B. 2008. Lingkungan Rumah dan Kejadian Difteri di Kabupaten Tasikmalaya dan Kabupaten Garut. Jurnal Kesehatan Masyarakat Nasional Vol 2 No 5.

Lestari, K.S. 2012. Faktor-faktor yang Berhubungan Dengan Kejadian Difteri di Kabupaten Sidoarjo. Depok: Universitas Indonesia.

Prasetyo, E. 2014. Mengolah Data Menjadi Informasi Menggunakan Matlab. Yogyakarta: AndiPublisher.

Puspita, G., Kharis, M. & Supriyono. 2017. Pemodelan Matematika Pada Penyebaran Penyakit Difteri Dengan Menggunakan Karantina dan Vaksinasi. UNNES Journal of Mathematics vol 6 No 1.

Suhendar, D. 2017. 14 Orang Meninggal Akibat Difteri di Jawa Barat. https://regional.kompas.com/read/2017/12/15/08124791/14-orang-meninggal-akibat-difteri-di-jawa-barat.

Zaki, MJ. & Meira JR, W. 2014. *Data Mining and Analysis: Foundations and Algorithms*. New York : Cambridge University Press.

http://www.diskes.jabarprov.go.id/index.php/arsip/categories/MTEz/profile-kesehatan

http://www.diskes.jabarprov.go.id/index.php/pages/detail/2014/7/Visi-dan-Misi.