UJM 6 (1) (2017)
UNNES Journal of Mathematics
http://journal.unnes.ac.id/sju/index.php/ujm

KLASIFIKASI TINGKAT KESEJAHTERAAN KELUARGA JAWA TENGAH TAHUN 2015 MENGGUNAKAN METODE REGRESI LOGISTIK ORDINAL DAN *SUPPORT VECTOR MACHINE* (SVM)

Sely Agustina✉, Arief Agoestanto, Putriaji Hendikawati

Jurusan Matematika, FMIPA, Universitas Negeri Semarang, Indonesia
Gedung D7 Lt. 1, Kampus Sekaran Gunungpati, Semarang 50229

Info Artikel

Sejarah Artikel:
Diterima Januari 2017
Disetujui Maret 2017
Dipublikasikan Mei 2017

Keywords:
Klasifikasi, Regresi Logistik Ordinal, SVM

Abstrak

Tujuan dari penelitian ini adalah untuk mengetahui metode yang memberikan ketepatan hasil klasifikasi yang lebih baik antara Regresi Logistik Ordinal dan *Support Vector Machine* (SVM). Sampel data yang digunakan adalah data tingkat kesejahteraan keluarga Jawa Tengah yang diperoleh dari hasil survei Pendataan Keluarga (PK) tahun 2015, sejumlah 322 data keluarga dan dibagi menjadi data *training* 80% sejumlah 259 dan data *testing* 20% sejumlah 64. Metode Regresi Logistik Ordinal dilakukan dengan estimasi data *training* untuk menentukan model logit awal, uji signifikansi menggunakan uji rasio Likelihood dan uji Wald, model logit yang signifikan digunakan untuk mengklasifikasi data *testing*. Metode SVM dilakukan dengan memodelkan data *training* menggunakan fungsi kernel *Linear, Polynomial*, dan *Gaussian* RBF, fungsi kernel terbaik digunakan untuk mengklasifikasi data *testing*. Metode Regresi Logistik Ordinal menghasilkan nilai ketepatan klasifikasi sebesar 81,25%. Metode SVM dengan kernel Linear sebagai fungsi kernel terbaik menghasilkan nilai ketepatan klasifikasi sebesar 95,31%.

Abstract

The purpose of this study was to determine the accuracy of the method provides a better classification between Ordinal Logistic Regression and Support Vector Machine (SVM). Sample data used is the data rate of the family welfare in Central Java were obtained from the survey of Family Data Collection (PK) 2015 by BKKBN of Central Java province, some 322 family data and divided into training data 80% number 259 and data testing 20% by 64. Ordinal Logistic Regression method performed by the estimated training data to determine initial logit model, the significance test using the likelihood ratio test and Wald test, significant logit models were used to classify the data testing. SVM method is done by modeling the training data using Linear kernel function, polynomial, and Gaussian RBF, the kernel function is best used to classify the data testing. Ordinal Logistic Regression method produces a value of classification accuracy of 81.25%. Linear SVM method with the kernel as the kernel function best yield value of classification accuracy of 95.31%.

How to Cite

Agustina S., Agoestanto A., & Hendikawati, P. (2017). Klasifikasi Tingkat Kesejahteraan Keluarga Jawa Tengah Tahun 2015 Menggunakan Metode Regresi Logistik Ordinal dan Support Vector Machine (SVM). *Unnes Journal of Mathematics*, 6(1): 59-69.

✉Alamat korespondensi:
E-mail: selyagustina@gmail.com

p- ISSN 2252-6943
e- ISSN 2460-5859

PENDAHULUAN

Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) membagi keluarga dalam masyarakat menjadi tiga kategori keluarga yakni keluarga prasejahtera, keluarga sejahtera 1, dan keluarga sejahtera. Pembagian kategori keluarga tersebut dilakukan oleh petugas-petugas setempat dengan cara manual berdasarkan indikator-indikator yang didapat dari survei pendataan keluarga (BKKBN, 2015). Hasil pembagian jenis keluarga ini akan digunakan sebagai bahan bagi penyusunan kebijakan dan program pembangunan keluarga oleh pemerintah daerah. Serta menjadi dokumentasi kehidupan keluarga Indonesia menurut dimensi waktu (Sunarti, 2006).

Proses pendataan keluarga dan pembagian kategori keluarga harus dilakukan dengan tepat. Untuk membantu petugas dalam menentukan status tahapan keluarga sejahtera, maka dilakukan penelitian terhadap penentuan status tahapan keluarga sejahtera dengan menggunakan metode klasifikasi.

Menurut Prasetyo (2012), klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target *f* yang memetakan setiap set atribut (fitur) *x* ke satu dari sejumlah label kelas *y* yang tersedia. Algoritma klasifikasi menggunakan data *training* untuk membuat sebuah model. Model yang sudah dibangun tersebut kemudian digunakan untuk memprediksi label kelas data baru yang belum diketahui. Klasifikasi dapat membantu petugas menentukan kategori yang cocok dari suatu data yang kombinasinya rumit.

Ada banyak metode yang dapat digunakan untuk menyelesaikan kasus klasifikasi, diantaranya adalah metode regresi logistik ordinal dan metode Support Vector Machine. Kelebihan regresi logistik adalah memiliki *odds ratio* yang menunjukkan seberapa besar pengaruh variabel prediktor suatu kategori referensi pada suatu variabel respon (Webb dan Yohannes, 1999). Metode *Support Vector Machine* juga memiliki kelebihan selain dapat menghasilkan tingkat akurasi yang tinggi, juga memberikan error generalisasi yang lebih kecil daripada metode lain.

Secara umum regresi logistik ordinal merupakan salah satu metode statistika untuk menganalisis variabel respon yang mempunyai skala data ordinal yang memiliki 3 kategorik atau lebih. Pada regresi logistik ordinal model berupa kumulatif logit model. Sedangkan untuk variabel prediktor yang digunakan berupa data kategorik dan atau kuantitatif. Sifat ordinal dari respon Y pada model logit ini dituangkan dalam peluang kumulatif sehingga kumulatif logit model merupakan model yang didapat dengan membandingkan peluang kumulatif yaitu peluang kurang dari atau sama dengan kategori respon ke-j.

Jika diasumsikan terdapat peubah respon Y berskala ordinal dengan J kategori dan $X^T = (x_1, x_2, \dots, x_p)$ adalah vektor peubah penjelas, maka peluang dari peubah respon kategori ke-j pada peubah penjelas X tertentu dapat dinyatakan dengan $P[Y = j|x] = \pi_j(x)$ dan peluang kumulatifnya adalah

$$P[Y \le j|x] = \frac{\exp(\alpha_j + X^T\beta)}{1+\exp(\alpha_j + X^T\beta)} \quad (1)$$

Dimana $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ merupakan nilai pengamatan ke-i (i=1,2,…,n) dari setiap variabel p variabel predictor. Pendugaan parameter regresi dilakukan dengan cara menguraikannya menggunakan transformasi logit dari $P[Y \le j|x]$.(Hosmer & Lemeshow, 2000)

Jika terdapat tiga kategori respon dimana j=0,1,2 maka model regresi logistik ordinal yang terbentuk adalah

$$Logit\ P[Y \le 0|x_i] = \ln\left(\frac{P[Y\le 1|x_i]}{1-P[Y\le 1|x_i]}\right) = \alpha_0 + X^T\beta$$

$$Logit\ P[Y \le 1|x_i] = \ln\left(\frac{P[Y\le 2|x_i]}{1-P[Y\le 2|x_i]}\right) = \alpha_1 + X^T\beta \quad (2)$$

Berdasarkan kedua peluang kumulatif pada persamaan (2) diperoleh peluang untuk masing-masing kategori respon sebagai berikut.

$$P(Y_j = 0) = \pi_0(x) = \frac{\exp(\alpha_0 + X^T\beta)}{1 + \exp(\alpha_0 + X^T\beta)}$$

$$P(Y_j = 1) = \pi_1(x) = \frac{\exp(\alpha_1 + X^T\beta)}{1 + \exp(\alpha_1 + X^T\beta)} - \frac{\exp(\alpha_0 + X^T\beta)}{1 + \exp(\alpha_0 + X^T\beta)}$$

$$P(Y_j = 2) = \pi_2(x) = 1 - \frac{\exp(\alpha_1 + X^T\beta)}{1+\exp(\alpha_1 + X^T\beta)} \quad (3)$$

Menurut Agresti (2002) untuk menentukan estimasi parameter digunakan metode maksimum Likelihood yang membutuhkan turunan pertama dan turunan kedua dari fungsi Likelihood. Maka log Likelihoodnya adalah

$$K = n\sum_{j=1}^{k-1}\left[Z_j\emptyset_j - Z_{j+1}g(\emptyset_j)\right] \quad (4)$$

Model nonlinier umum dapat ditulis

$$Y_j = logit\ (c_j) = \beta^T X_j \quad (5)$$

dengan

$\beta^T = (\theta_1\theta_2 \dots \theta_{k-1}\beta_1\beta_2 \dots \beta_p)$ adalah vector parameter

$X_j = (0 \dots 1 \dots .0\ x_1\ x_2 \dots x_p)^T$ dimana nilainya 1 pada kategori/klasifikasi j

Persamaan log Likelihood tersebut bukan merupakan fungsi linier β sehingga taksiran β dicari dengan menggunakan metode numerik. Metode yang dipakai untuk memecahkan masalah ini adalah metode Newton-Raphson. Prosedur Newton-Raphson untuk mencari taksiran β^T adalah :

1. Pilih taksiran awal $\beta_m, m = 1,2, \ldots$, misal diambil $\beta_1 = 0$
2. Pada setiap iterasi ke (m+1) hitung taksiran baru
$$\beta_{m+1} = \beta_m + A_{rs}^{-1} \frac{\partial K}{\partial \beta}$$
3. Iterasi berlanjut hingga diperoleh $\beta_{m+1} \approx \beta_m$

Untuk mengetahui pengaruh dari variabel independen dilakukan uji signifikansi secara keseluruhan dan secara individu. Dalam pengujian secara keseluruhan digunakan uji Rasio Likelihood. Pengujian ini digunakan untuk menguji kelayakan model yang diperoleh dari estimasi parameter, bertujuan untuk mengetahui apakah variabel independen yang terdapat dalam model berpengaruh nyata atau tidak secara keseluruhan (Hosmer and Lemeshow, 2000).

Hipotesis : H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1: paling sedikit ada satu $\beta_r \neq 0$ dengan r = 1,2, … , p

Statistik uji rasio Likelihood adalah

$$X^2_{hitung} = -2 \ln \left(\frac{Likelihood\ tanpa\ variabel\ bebas}{Likelihood\ dengan\ variabel\ bebas} \right) \quad (6)$$

Kriteria uji H_0 ditolak jika $X^2_{hitung} > X^2(\alpha, p)$. Penolakan H_0 memberi arti bahwa satu atau lebih parameter β yang ada pada model tidak sama dengan nol. Oleh karena itu, dengan mengetahui signifikan/ tidaknya parameter dapat diketahui signifikan/ tidaknya model.

Uji signifikansi secara individu dilakukan dengan menggunakan uji Wald yang diperoleh dengan cara mengkuadratkan rasio estimasi parameter dengan estimasi standar *error* nya, uji Wald dilakukan untuk mengetahui signifikansi parameter terhadap variabel dependen (Hosmer and Lemeshow, 2000).

Hipotesis : H_0: $\beta_r = 0$

H_1: $\beta_r \neq 0$ dengan r = 1,2, … , p

Statistik uji Wald

$$W_r = \left[\frac{\hat{\theta}_r}{SE(\hat{\theta}_r)} \right]^2 \quad (7)$$

Kriteria uji H_0 ditolak jika $W_r > x^2(\alpha, 1)$

Selanjutnya dilakukan uji kesesuaian model untuk mengetahui apakah terdapat perbedaan yang nyata antara hasil observasi dengan prediksi model. Pengujian ini menggunakan uji Hosmer dan Lemeshow (Hosmer and Lemeshow, 2000).

Hipotesis: H_0 = Model sesuai

H_1 = Model tidak sesuai

Statistik Uji

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n\,\bar{\pi}_k)^2}{n_k \pi_k (1-\pi_k)} \quad (8)$$

dengan $O_k = \sum_{j=1}^{nk} y_j$; $\bar{\pi}_k = \sum_{j=1}^{nk} \frac{m_j \bar{\pi}_j}{nk}$

g = jumlah grup

nk = banyaknya subjek pada grup ke-k

O_k = jumlah nilai variabel respon grup ke-k

m_j = banyak observasi yang memiliki nilai $\bar{\pi}_j$

$\bar{\pi}_k$ = rata-rata taksiran probabilitas

Kriteria uji H_0 ditolak jika $\hat{C} > x^2(\alpha, g-2)$

Support Vector Machine (SVM) merupakan salah satu bagian dari Data Mining yang digunakan untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi (Santoso, 2007). Pada dasarnya SVM bekerja dengan prinsip *linier classifier*, kemudian dikembangkan untuk dapat bekerja pada kasus non linear dengan menggunakan konsep kernel pada ruang kerja berdimensi tinggi (Nugroho dkk, 2003). Pada klasifikasi linear SVM dibagi menjadi 2 jenis yaitu *separable* dan *nonseparable*.

Misalkan diberikan himpunan $X = \{x_1, x_2, \ldots, x_n\}$, dengan $x_i \in R^p$, dengan telah diketahui X memiliki pola tertentu, yaitu apabila x_i termasuk dalam suatu kelas maka diberi label $y_i = +1$, jika tidak diberi label $y_i = -1$ untuk itu label masing-masing dinotasikan $y_i \in \{-1, +1\}$ sehingga data berupa pasangan (x_1, y_1), $(x_2, y_2), \ldots, (x_i, y_i)$ dimana i=1,2,…,n yang mana n adalah banyak data. diasumsikan kedua kelas -1 dan +1 dapat terpisah secara sempurna oleh fungsi pemisah berdimensi p, yang didefinisikan : $w^T x + b = 0$. w dan b adalah parameter model.

Data x_i yang termasuk dalam kelas -1 dapat dirumuskan sebagai berikut.

$[w^T x_i + b] \leq -1$, untuk $y_i = -1$ (9)

Sedangkan data x_i yang termasuk dalam kelas +1 dapat dirumuskan sebagai berikut.

$[w^T x_i + b] \geq +1$, untuk $y_i = +1$ (10)

Klasifikasi kelas data pada persamaan (9) dan (10) dapat digabungkan dengan notasi sebagai berikut.

$y_i[w^T x_i + b] \geq 1$, i = 1, 2,…, n (11)

(Prasetyo, 2012)

Untuk mendapatkan fungsi pemisah terbaik adalah dengan mencari fungsi pemisah yang terletak ditengah-tengah antara dua bidang pembatas kelas, sama dengan memaksimalkan margin atau jarak antara dua set objek dari kelas yang berbeda (Santosa, 2007). Fungsi pemisah optimal dihitung dengan memaksimalkan

margin $\rho(w,b)$ untuk jarak x ke fungsi pemisah (w,b) adalah.

$$d(w,b;x) = \frac{|w^T x+b|}{||w||} \quad (12)$$

(Gunn, 1998)

Selanjutnya, diformulasikan kedalam persamaan *quadratic programming* (QP), dengan meminimalkan invers persamaan (2), seperti berikut.

$$\frac{1}{2}||w||^2, \text{ dimana } ||w||^2 = w^T w \quad (13)$$

dengan syarat

$$y_i[(w^T x)+b]-1 \geq 0\,, \quad i = 1,2,3,\dots,n$$

(Prasetyo, 2012)

Optimalisasi ini dapat diselesaikan dengan fungsi *Lagrange Multiplier*

$$L(w,b,\alpha) = \frac{1}{2}w^T w - \sum_{i=1}^{n} \alpha_i\{\, y_i[w^T x_i + b]-1\} \quad (14)$$

(Prasetyo, 2012)

Nilai α_i adalah fungsi *Lagrange Multiplier,* yang bernilai nol atau positif $(\alpha_i \geq 0)$. Dari hasil perhitungan ini diperoleh α_i kebanyakan bernilai positif. Data yang berkorelasi dengan α_i yang positif disebut *support vector* (Vapnik, 1995). Kelas dari data yang akan diprediksi atau data testing dapat ditentukan berdasarkan fungsi sebagai berikut.

$$f(x_t) = \sum_{i=1}^{ns} \alpha_i\, y_i x_i . x_t + b \quad (15)$$

Metode SVM juga dapat digunakan dalam kasus non-separable dengan memperluas formulasi yang terdapat pada kasus linier. Masalah optimasi sebelumnya baik pada fungsi obyektif maupun kendala dimodifikasi dengan mengikutsertakan variabel *Slack* $\xi > 0$ yaitu merupakan sebuah ukuran kesalahan klasifikasi. Formulasinya sebagai berikut.

$$y_i[(w^T x_i)+b] \geq 1-\xi_i, \text{ I = 1, 2, …, n} \quad (16)$$

Sehingga persamaan (14) menjadi sebagai berikut.

$$\Phi(w,\xi) = \frac{1}{2}w^T w + C\ \sum_{i=1}^{n} \xi_i \quad (18)$$

(Gunn, 1998)

Model optimasi (18) dapat diselesaikan dengan menggunakan fungsi *Lagrange,* yaitu.

$$L(w,b,\alpha,\xi,\beta) = \frac{1}{2}w^T w + C\ \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i\, y_i[(w^T x_i)+b]-1+\xi_i) - \sum_{1}^{n} \beta_i\, \xi_i \quad (19)$$

(Kecman, 2005)

Untuk menyederhanakan persamaaan (19) harus ditransformasi kedalam fungsi *Lagrange Multiplie* itu sendiri (dualitas masalah). Sehingga menjadi sebuah persamaan.

$$\max_a L_d = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j y_i y_j x_i^T x_j \quad (20)$$

dengan batas $0 \leq \alpha_i \leq C, i = 1,2,\dots,n$ dan $\sum_{i=1}^{n} \alpha_i\, y_i$

(Kecman, 2005)

Pada umumnya masalah dalam dunia nyata (*real world problem*) jarang yang bersifat linier separable (tidak terpisahkan secara linier), tetapi bersifat non-linear (Nugroho, dkk, 2003). Untuk menyelesaikan problem non-linear, SVM dimodifikasi dengan memasukkan fungsi kernel. Kernel dapat didefinisikan sebagai suatu fungsi yang memetakan fitur data dari dimensi awal (rendah) ke fitur yang lebih tinggi (bahkan jauh lebih tinggi).

Dalam SVM non-linear, data x dipetakan oleh fungsi $\Phi(x)$ ke ruang vektor yang berdimensi lebih tinggi. Dimisalkan untuk n sampel data

$$((\Phi(x_1), y_1); (\Phi(x_2), y_2); \dots; (\Phi(x_n), y_n)) \quad (21)$$

Proses pemetaan memerlukan perhitungan *dot product* dua buah data pada ruang fitur baru. *Dot product* dua buah vector (x_i) dan (x_j) dinotasikan sebagai $\Phi(x_1).\Phi(x_t)$. Nilai *dot product* tersebut dapat dihitung tanpa mengetahui fungsi transformasi Φ dengan memakai komponen kedua buah vector tersebut di ruang dimensi asal, seperti berikut.

$$\mathrm{K}(x_i, x_t) = \Phi(x_1).\Phi(x_t) \quad (22)$$

Nilai $\mathrm{K}(x_i, x_t)$ merupakan fungsi kernel yang menunjukkan pemetaan non-linear pada *feature space*. Prediksi himpunan data dengan dimensi fitur yang baru diformulasikan dengan.

$$\mathrm{f}\big(\Phi(x)\big) = sign\ (\sum_{i=1}^{ns} \alpha_i\, y_i \mathrm{K}(x_i, x_t) + b) \quad (23)$$

dengan

ns : jumlah data yang menjadi *support vector*

x_i : *support vector*

x_t : data testing yang akan diprediksi

(Prasetyo, 2012)

Fungsi kernel yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Kernel linear
   $$\mathrm{K}(x_i, x_t) = x_i^T x_t \quad (24)$$
2. Kernel *Radial Basis Function* (RBF) atau kernel Gaussian.
   $$\mathrm{K}(x_i, x_t) = \exp(-\frac{1}{2\sigma^2}||\, x_i - x_t||^2) \quad (25)$$
3. Kernel Polynominal
   $$\mathrm{K}(x_i, x_t) = (x_i^T x_t + 1)^d \quad (26)$$

(Kecman, 2005)

Penelitian tentang perbandingan klasifikasi menggunakan regresi logistik ordinal dan *support vector machine* (SVM) pernah dilakukan oleh Santi Wulan Purnami (2012) untuk klasifikasi tingkat keganasan *breast cancer.* Hasil yang diperoleh yaitu 56,60% dengan metode regresi logistik ordinal dan 98,11% menggunakan *Support Vector Machine*,

sehingga metode SVM memiliki ketepatan klasifikasi lebih baik dibandingkan dengan regresi logistik ordinal untuk klasifikasi tingkat keganasan *breast cancer.*

METODE

Metode yang digunakan pada penelitian ini adalah perumusan masalah, pengumpulan data, pengolahan dan analisis data, dan penarikan kesimpulan. Perumusan masalah yang dimaksudkan adalah suatu usaha untuk membatasi permasalahan, sehingga diperoleh bahan kajian yang jelas.

Tahapan pengumpulan data, dalam penelitian ini data yang digunakan adalah data pendataan keluarga di Provinsi Jawa Tengah pada tahun 2015. Data tersebut merupakan data sekunder yang diperoleh dari (BKKBN) Provinsi Jawa Tengah. Pengambilan sampel dilakukan dengan teknik sampling *proporsional random sampling* yaitu dari populasi sejumlah 9 juta kepala keluarga di Provinsi Jawa Tengah, sampel yang diambil untuk penelitian ini adalah 322 data kepala keluarga.

Variabel yang digunakan dalam penelitian ini terdiri atas variabel respon (Y) dan variabel predictor (X) yang diuraikan dalam tabel 1 sebagai berikut.

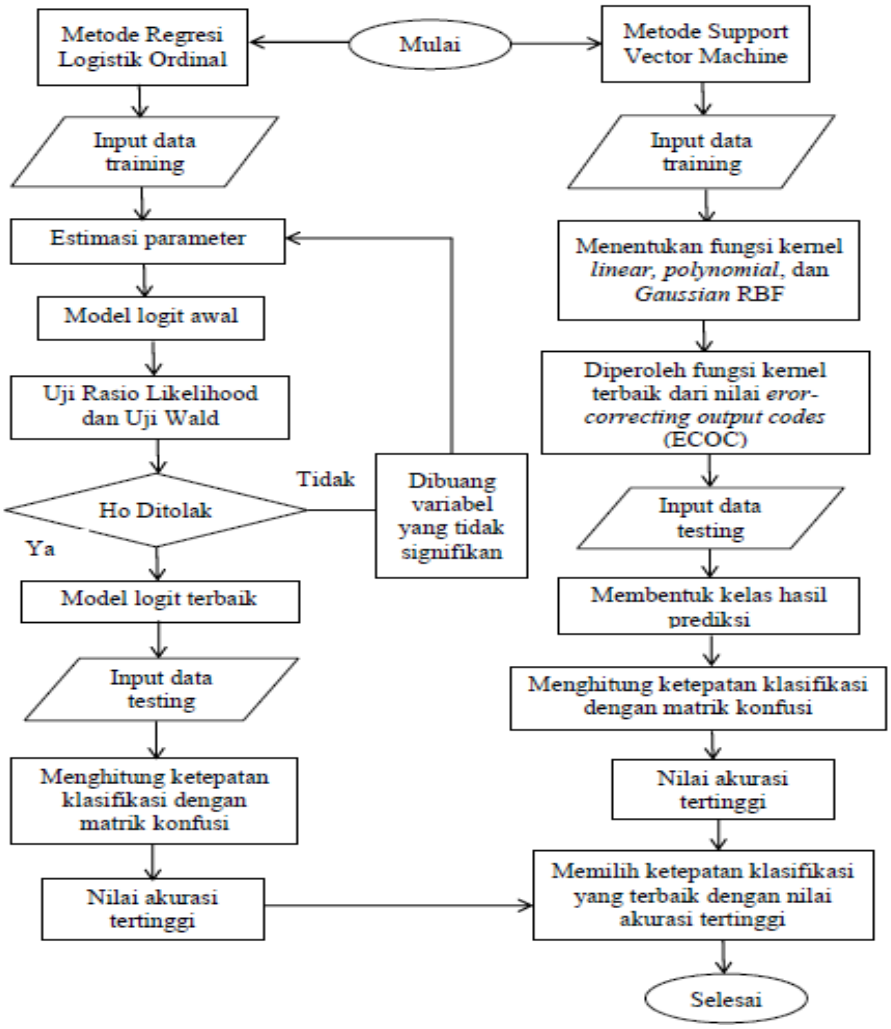
Tabel 1 Deskripsi Variabel

Var	Label	Kategori
Y	Tingkat kesejahteraan keluarga	0=Keluarga pra-sejahtera 1=Keluarga sejahtera 1 2=Keluarga sejahtera
X1	Keluarga membeli satu stel pakaian	0=Tidak 1=Ya
X2	Keluarga makan minimal 2 kali sehari.	0=Tidak 1=Ya 2=Tidak berlaku
X3	Keluarga berobat ke fasilitas kesehatan	0=Tidak 1=Ya
X4	Keluarga memiliki pakaian yang berbeda	0=Tidak 1=Ya
X5	Keluarga makan daging/ikan/telur	0=Tidak 1=Ya
X6	Keluarga menjalankan ibadah agama	0=Tidak 1=Ya
X7	Pasangan usia subur menjadi peserta KB	0=Tidak 1=Ya 2=Tidak berlaku
X8	Keluarga memiliki tabungan	0=Tidak 1=Ya
X9	Keluarga berkomunikasi dengan seluruh anggota keluarga	0=Tidak 1=Ya
X10	Keluarga ikut kegiatan sosial	0=Tidak 1=Ya
X11	Keluarga memiliki akses informasi	0=Tidak 1=Ya
X12	Keluarga menjadi pengurus kegiatan sosial	0=Tidak 1=Ya
X13	Keluarga mempunyai balita ikut kegiatan posyandu	0=Tidak 1=Ya 2=Tidak berlaku
X14	Keluarga mempunyai balita ikut kegiatan BKB	0=Tidak 1=Ya 2=Tidak berlaku
X15	Keluarga mempunyai remaja ikut kegiatan BKR	0=Tidak 1=Ya 2=Tidak berlaku
X16	Ada anggota keluarga masih remaja ikut PIK-R/M	0=Tidak 1=Ya 2=Tidak berlaku
X17	Keluarga lansia ikut kegiatan BKL	0=Tidak 1=Ya 2=Tidak berlaku
X18	Keluarga mengikuti kegiatan UPPKS	0=Tidak 1=Ya 2=Tidak berlaku
X19	Apakah jenis atap rumah terluas	0=Daun/ Rumbia 1=Seng/ Asbes 2=Genteng/ Sirap 3=Lainnya
X20	Apakah jenis dinding rumah terluas	0=Tembok 1=Kayu/ Seng 2=Bambu 3=Lainnya
X21	Apakah jenis lantai rumah terluas	0=Ubin/ Keramik/ Marmer 1=Semen/ Papan 2=Tanah 3=Lainnya
X22	Apakah sumber penerangan utama	0=Listrik 1=Genset/ Diesel 2=Lampu Minyak 3=Lainnya

Var	Label	Kategori
X23	Apakah sumber air minum	0=Ledeng/ Kemasan 1=Sumur Terlindung/ Pompa 2=Air hujan/ Air sungai 3=Lainnya
X24	Apakah bahan bakar utama untuk memasak	0=Listrik/ Gas 1=Minyak Tanah 2=Arang/ Kayu 3=Lainnya
X25	Apakah fasilitas tempat buang air besar	0=Jamban sendiri 1=Jamban bersama 2=Jamban umum 3=Lainnya
X26	Status kepemilikan rumah/bangunan tempat tinggal	0=Milik sendiri 1=Sewa/ kontrak 2=Menumpang 3=Lainnya
X27	Berapa luas rumah/bangunan keseluruhan (m2)	(kontinu)
X28	Berapa orang yang tinggal dan menetap di rumah/ bangunan ini (orang)	(kontinu)

Tahapan analisis data dapat dilihat pada gambar 1. Penjelsan tahapan analisis data tersebut adalah sebagai berikut.

1. Menerjemahkan variabel dari bahasa menjadi variabel kategori pada tabel 1
2. Membagi data menjadi dua bagian yaitu data training 80% sejumlah 258 data dan data testing 20% sejumlah 64 data.
3. Melakukan klasifikasi menggunakan metode Regresi Logistik Ordinal dengan bantuan program SPSS v16.0
   a. Melakukan estimasi parameter.
   b. Menentukan model logit awal.
   c. Melakukan uji signifikansi secara keseluruhan menggunakan Uji Rasio Likelihood dan uji signifikansi secara individu menggunakan Uji Wald untuk mengetahui variabel yang berpengaruh dalam model.
   d. Menentukan model logit akhir.
   e. Melakukan uji kesesuaian model menggunakan uji Hosmer dan Lemeshow.
   f. Menghitung ketepatan klasifikasi.
4. Melakukan klasifikasi menggunakan metode Support Vector Machine (SVM) dengan bantuan program MATLAB R2015b. Software ini memiliki tools-tools yang dapat memudahkan dalam proses pembuatan program (Hartono, 2012).
   a. Melakukan transformasi data sesuai dengan metode SVM *multiclass.*
   b. Menentukan fungsi kernel untuk pemodelan yaitu kernel *linear, polynomial,* dan *Gaussian* RBF.
   c. Diperoleh fungsi kernel terbaik dari nilai *eror-correcting output codes* (ECOC).
   d. Membentuk kelas hasil prediksi menggunakan kernel terbaik.
   e. Menghitung nilai ketepatan klasifikasi dengan matrik konfusi
5. Membandingkan ketepatan klasifikasi yang diperoleh dari Regresi Logistik Ordinal dengan SVM.
6. Membuat kesimpulan



Gambar 1. Diagram alir (*flowchart*) Teknik Analisis Data

HASIL DAN PEMBAHASAN

Analisis Deskriptif

Analisis deskriptif digunakan untuk memperoleh gambaran data secara umum. Data yang digunakan pada penelitian ini adalah sebanyak 322 data hasil pendataan keluarga 2015 dengan persentase 59 % diantaranya adalah keluarga sejahtera, 21% adalah keluarga sejahtera 1, sedangkan sisanya yaitu 20% merupakan keluarga prasejahtera.

Persentase keluarga sejahtera pada data training yaitu 49% , keluarga sejahtera 1 sebesar 23%, dan keluarga prasejahtera sebesar 28%. Sedangkan persentase data testing untuk keluarga sejahtera sebesar 9%, keluarga sejahtera 1 sebesar 27%, dan keluaga prasejahtera sebesar 64%.

Analisis menggunakan Metode Regresi Logistik Ordinal

Analisis menggunakan metode Regresi Logistik Ordinal dilakukan dengan bantuan program SPSSv16. Langkah pertama, dari data *training* yang diinputkan, dilakukan estimasi parameter untuk memperoleh model awal tahap pertama. Hasil estimasi seperti pada tabel 2 berikut.

Tabel 2 Estimasi Parameter

Variabel	Estimsi Parameter
Y=0	-9,321
Y=1	-5,503
X_1	-5,158
X_2	-12,410
X_3	-7,020
X_5	-6,983
X_7	-2,541
X_8	-2,674
X_9	-3,751
X_{13}	-6,102
X_{14}	6,015
X_{27}	0,019
X_{28}	-0,466

Model awal dari hasil estimasi tersebut diuji parameter secara keseluruhan menggunakan uji *Rasio Likelihood* dengan Hipotesis :

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_5 = \beta_7 = \beta_8 = \beta_9 = \beta_{13} = \beta_{14} = \beta_{27} = \beta_{28} = 0$ (Model Tidak Signifikan)

H_1 : Paling sedikit salah satu dari $\beta_r \neq 0$, dimana r = 1, 2, 3, 5, 7, 8, 9, 13, 14, 27, 28 (Model Signifikan)

Daerah Kritis :

H_0 ditolak jika signifikansi $< 5\%(\alpha)$ atau nilai $\chi^2_{\text{hit}} > \chi^2_{(0,05;11)}$ dimana nilai $\chi^2_{(0,05;11)}$ = 19,68

Tabel 3 Uji *Rasio Likelihood*

Model	-2 Log Likelihood	Chi-Square	df	Sig,
Intercept Only	538,005			
Final	184,473	353,532	15	,000

Link function: Logit,

Berdasarkan tabel 3 diperoleh nilai signifikansi 0 dan nilai Chi-Square adalah 353,532. Karena nilai signifikansi = $0 < 5\%(\alpha)$ atau 353,532 $(\chi^2_{hit}) > 19{,}68$ $(\chi^2_{(0,05;11)})$ maka H_0 ditolak. Penolakan H_0 memberi arti bahwa satu atau lebih parameter β yang ada pada model tidak sama dengan nol.

Kemudian dilakukan uji parameter secara individu menggunakan uji *Wald* untuk mengetahui apakah dalam model tersebut terdapat variabel yang tidak signifikan, dengan hipotesis:

H_0 : $\beta_r = 0$ (parameter tidak signifikan atau variabel bebas tidak memiliki hubungan yang kuat dengan variabel respon)

H_1 : $\beta_r \neq 0$ dimana r = 1, 2, 3, 5, 7, 8, 9, 13, 14, 27, 28 (parameter signifikan atau variabel bebas memiliki hubungan yang kuat dengan variabel respon)

Daerah kritis: H_0 ditolak jika sig. $< 5\%$ (α) atau $W_r \chi^2_{(0,05;11)}$

Hasil uji Wald dapat dilihat dalam tabel 4 berikut.

Tabel 4 Uji Wald

Variabel Bebas	Nilai β	W_r	Sig.	$\chi^2_{(0,05;11)}$	Keputusan
X_1	-5,158	1,075	0,001	19,68	Ditolak
X_2	-12,410	5,103	0,000	19,68	Ditolak
X_3	-7,020	,259	0,031	19,68	Ditolak
X_5	-6,983	,549	0,000	19,68	Ditolak
X_7	-2,541	9,542	0,000	19,68	Ditolak
X_8	-2,674	0,277	0,000	19,68	Ditolak
X_9	-3,751	4,211	0,000	19,68	Ditolak
X_{13}	-6,102	,976	0,026	19,68	Ditolak
X_{14}	6,015	,570	0,033	19,68	Ditolak
X_{27}	0,019	0,466	0,000	19,68	Ditolak
X_{28}	-0,466	0,995	0,000	19,68	Ditolak

Hasil uji *Wald* nilai sig. seluruh variabel kurang dari 5% (α) artinya seluruh variabel signifikan, sehingga diperoleh keputusan bahwa model signifikan dan variabel yang tetap dimasukkan ke dalam model adalah X_1, X_2, X_3, X_5, X_7, X_8, X_9, X_{13}, X_{14}, X_{27}, dan X_{28}

Berikut adalah model yang diperoleh dari hasil analisis.

$$Logit\ 1 = -9{,}321 - 5{,}158X_1 - 12{,}410X_2 - 7{,}020X_3 - 6{,}983X_5 - 2{,}541X_7 - 2{,}674X_8 - 3{,}751X_9 - 6{,}102X_{10} + 6{,}015X_{14} + 0{,}019X_{27} - 0{,}466X_{28}$$

$$Logit\ 2 = -5{,}503 - 5{,}158X_1 - 12{,}410X_2 - 7{,}020X_3 - 6{,}983X_5 - 2{,}541X_7 - 2{,}674X_8 - 3{,}751X_9 - 6{,}102X_{10} + 6{,}015X_{14} + 0{,}019X_{27} - 0{,}466X_{28}$$

Model logit 1 merupakan model peluang kategori pertama atau kategori keluarga prasejahtera dan model logit 2 merupakan model peluang kategori kedua atau kategori keluarga sejahtera 1 atau kategori kelurga sejahtera. Kemudian dengan model akhir yang telah signifikan tersebut dilakukan prediksi kategori kelas dari data testing sebanyak 64 data, dengan cara menghitung nilai C_j dan nilai peluang untuk masing-masing kelas seperti dalam tabel 5 berikut.

Tabel 5 Hasil Prediksi Menggunakan Metode Regresi Logistik Ordinal

No	C_1	C_2	π_1	π_2	π_3	P	A
1	0,874	0,759	0,874	0,115	0,885	2	2
2	0,278	0,872	0,278	0,594	0,128	1	1
3	0,302	0,808	0,302	0,505	0,192	1	1
4	0,864	0,743	0,864	-0,121	0,879	2	1
5	0,661	0,7	0,661	0,039	0,3	0	0
6	0,566	0,856	0,566	0,29	0,71	2	2
7	0,529	0,836	0,529	0,307	0,164	0	0
8	0,626	0,777	0,626	0,150	0,223	0	0
9	0,874	0,759	0,874	0,115	0,885	2	2
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	
60	0,737	0,664	0,737	-0,073	0,336	0	0
61	0,225	0,807	0,225	0,582	0,192	1	1
62	0,796	0,695	0,796	-0,101	0,304	0	0
63	0,693	0,507	0,693	0,186	0,814	2	2
64	0,879	0,774	0,879	-0,105	0,225	0	1

Hasilnya 36 keluarga prasejahtera, 8 keluarga sejahtera 1, dan 8 keluarga sejahtera yang diprediksi tepat. Tetapi 2 keluarga prasejahtera menjadi keluarga sejahtera 1 dan 1 keluarga prasejahtera menjadi keluarga sejahtera. 5 keluarga sejahtera 1 diprediksi menjadi keluarga prasejahtera dan 1 keluarga sejahtera 1 menjadi keluarga sejahtera. 2

keluarga sejahtera diprediksi menjadi keluarga prasejahtera dan 1 keluarga sejahtera diprediksi menjadi keluarga sejahtera 1. Hasil prediksi tersebut dihitung ketepatan klasifikasinya menggunakan matrik konfusi dalam tabel 6.

Tabel 6 Matriks Konfusi

Kelas Hasil Observasi	Kelas Hasil Prediksi		
	Pra Sejahtera	Sejahtera 1	Sejahtera
Pra Sejahtera	36	2	1
Sejahtera 1	5	8	1
Sejahtera	2	1	8

$$Akurasi = \frac{jumlah\ prediksi\ \ benar}{jumlah\ prediksi} \times 100\%$$

$$Akurasi = \frac{36+8+8}{36+2+1+5+8+1+2+1+8} \times 100\%$$

$Akurasi = 81{,}25\ \%$

Jadi akurasi hasil klasifikasi metode Regresi Logistik Ordinal adalah 81,25%

Analisis menggunakan Metode SVM

Analisis menggunakan metode *Support Vector Machine* (SVM) dilakukan menggunakan bantuan program Matlab R2015b.

Langkah pertama *load* data *training* dan data *testing* dalam *workspace* matlab. Kemudian menentukan fungsi kernel yang akan digunakan untuk pemodelan yaitu kernel *linear, polynomial,* dan *Gaussian* RBF menggunakan fungsi *fitcecoc.* Fungsi *fitcecoc* adalah kode untuk klasifikasi multikelas dengan bekerja dengan mereduksi menjadi klasifikasi biner.

Fungsi untuk memanggil kernel Gaussian

```
>>
g=templateSVM('Standardize',1,'KernelFunction','G
aussian')
>>
Mdl=fitcecoc(TrainX,TrainY,'Learners',g,'ClassNam
es',{'0','1','2'});
>> CVMdl=crossval(Mdl);
>> oosLoss=kfoldLoss(CVMdl)
```

Diperoleh output

```
    oosLoss =
 0.4884
```

Fungsi untuk memanggil kernel Polynomial

```
>>
p=templateSVM('Standardize',1,'KernelFunction','P
olynomial')

>>
Mdl=fitcecoc(TrainX,TrainY,'Learners',g,'ClassNam
es',{'0','1','2'});
>> CVMdl=crossval(Mdl);
>> oosLoss=kfoldLoss(CVMdl)
```

Diperoleh output

```
    oosLoss =
 0.2713
```

Fungsi untuk memanggil kernel *Linear*

```
>>
l=templateSVM('Standardize',1,'KernelFunction','L
inear')
>>
Mdl=fitcecoc(TrainX,TrainY,'Learners',g,'ClassNam
es',{'0','1','2'});
>> CVMdl=crossval(Mdl);
>> oosLoss=kfoldLoss(CVMdl)
```

Diperoleh output

```
    oosLoss =
 0.1667
```

Output yang diperoleh berupa nilai *eror-correcting output codes* (ECOC) dari masing-masing fungsi kernel seperti dalam tabel 7 berikut.

Tabel 7 Perbandingan nilai ECOC pada fungsi kernel *Linear, Polynomial*, dan *Gaussian* RBF

	Kernel		
	Linear	*Polynomial*	*Gaussian*RBF
Nilai ECOC	0,1667	0,2713	0,4884

Fungsi kernel terbaik dari nilai *eror-correcting output codes* (ECOC) terkecil 0,1667 yaitu kernel *linear.* Fungsi kernel terbaik tersebut digunakan untuk memprediksi kelas baru menggunakan data *testing.*

```
>>
Mdl=fitcecoc(TestX,TestY,'Learners',l,'FitPoste
rior',1,'ClassNames',{'0','1','2'},'Verbose',2;
```

Training binary learner 1 (SVM) out of 3 with 17 negative and 41 positive observations.
Negative class indices: 2
Positive class indices: 1

Fitting posterior probabilities for learner 1 (SVM).
Training binary learner 2 (SVM) out of 3 with 6 negative and 41 positive observations.
Negative class indices: 3
Positive class indices: 1

Fitting posterior probabilities for learner 2 (SVM).
Training binary learner 3 (SVM) out of 3 with 6 negative and 17 positive observations.
Negative class indices: 3
Positive class indices: 2

Fitting posterior probabilities for learner 3 (SVM).

```
>>
[label,~,~,Posterior]=resubPredict(Mdl,'Verbose',1);
```

Predictions from all learners have been computed.

Loss for all observations has been computed.

Computing posterior probabilities...

```
>> Mdl.BinaryLoss

    ans =
    quadratic

>> idx=randsample(size(TestX,1),64,1);
    Mdl.ClassNames

    ans =
      '0'
      '1'
      '2'

>>
table(TestY(idx),label(idx),Posterior(idx,:),'VariableNames',{'TrueLabel','PredLabel','Posterior'})
```

Output yang diperoleh berupa tabel perbandingan kelas asli dan kelas hasil prediksi serta nilai *probabilitas posterior*, nilai *probabilitas posterior* tersebut merupakan nilai peluang estimasi prediksi seperti dalam tabel 8.

Tabel 8. Hasil Prediksi Menggunakan Metode SVM

No	Kelas Asli	Kelas Prediksi
1	2	2
2	1	1
3	1	1
4	1	1
5	0	0
6	2	2
7	0	0
8	0	0
9	2	2
10	0	0
11	1	1
12	0	0
13	2	0
14	1	1
15	2	0
.	.	.
.	.	.
.	.	.
60	0	0
61	1	1
62	0	0
63	2	0
64	1	1

Dari tabel prediksi diperoleh hasil bahwa terdapat 3 keluarga kategori kelas sejahtera yang diprediksi sebagai keluarga kelas prasejahtera. Karena terdapat 3 data yang diprediksi salah.

Hasil prediksi yang telah diperoleh dihitung akurasi ketepatan klasifikasinya menggunakan matrik konfusi, dapat dilihat dalam tabel 9 sebagai berikut.

Tabel 9 Matrik Konfusi untuk Menghitung Akurasi Metode SVM

Kelas asli	Kelas hasil prediksi		
	Kelas = 0	Kelas = 1	Kelas = 2
Kelas = 0	32	0	0
Kelas = 1	0	20	0
Kelas = 2	3	0	9

$$Akurasi = \frac{jumlah\ prediksi\ \ benar}{jumlah\ prediksi} \times 100\%$$

$$Akurasi = \frac{32 + 20 + 9}{32 + 0 + 0 + 0 + 20 + 0 + 3 + 0 + 9} \times 100\%$$

$$Akurasi = 95{,}31\ \%$$

Jadi akurasi hasil klasifikasi metode SVM adalah 95,31%.

Setelah hasil klasifikasi dengan kedua metode tersebut dibandingkan untuk mengetahui metode mana yang memberikan nilai ketepatan klasifikasi yang lebih baik.

SIMPULAN

Simpulan yang diperoleh dari penelitian ini adalah (1) Model terbaik yang diperoleh dari hasil klasifikasi menggunakan metode Regresi Logistik Ordinal adalah

$$\begin{aligned} Logit\ 1 = {} & -9{,}321 - 5{,}158X_1 - 12{,}410X_2 \\ & - 7{,}020X_3 - 6{,}983X_5 \\ & - 2{,}541X_7 - 2{,}674X_8 \\ & - 3{,}751X_9 - 6{,}102X_{10} \\ & + 6{,}015X_{14} + 0{,}019X_{27} \\ & - 0{,}466X_{28} \end{aligned}$$

$$\begin{aligned} Logit\ 2 = {} & -5{,}503 - 5{,}158X_1 - 12{,}410X_2 \\ & - 7{,}020X_3 - 6{,}983X_5 \\ & - 2{,}541X_7 - 2{,}674X_8 \\ & - 3{,}751X_9 - 6{,}102X_{10} \\ & + 6{,}015X_{14} + 0{,}019X_{27} \\ & - 0{,}466X_{28} \end{aligned}$$

Dengan ketepatan klasifikasi adalah 81,25%

(2) Ketepatan klasifikasi dengan menggunakan metode SVM dengan fungsi kernel *Linear* sebesar 95,31%. (3) Metode SVM memberikan ketepatan hasil klasifikasi yang lebih baik jika dibandingkan dengan metode Regresi Logistik Ordinal pada tingkat kesejahteraan keluarga

karena nilai ketepatan klasifikasi yang dihasilkan lebih tinggi. (4) Perbandingan antara klasifikasi manual oleh petugas Pendataan Keluarga dan klasifikasi menggunakan Regresi Logistk Ordinal atau *Support Vector Machine* (SVM) adalah metode *Support Vector Machine* (SVM) dapat melakukan klasifikasi dalam data yang jumlahnya besar dalam waktu yang singkat. Sementara dengan cara manual data hanya dapat diklasifikasi satu-satu, sehingga membutuhkan waktu yang lebih lama dan jumlah petugas yang lebih banyak.

SARAN

Berdasarkan hasil penelitian yang telah dilakukan, peneliti memberikan beberapa saran (1) Penentuan indikator status tahapan keluarga sejahtera hendaknya diperhatikan kembali, karena dalam penelitian ini dari 28 indikator (variabel) hanya 11 yang signifikan. (2) Perbandingan metode klasifikasi Regresi Logistik Ordinal dan SVM pada penelitian ini menggunakan bantuan program SPSSv16 dan Matlab R2015, untuk penelitian selanjutnya sebaiknya menggunakan bantuan program yang sama. (3) Penelitian mengenai klasifiksi yang lain dapat menggunakan metode *Support Vector Machine* karena telah terbukti menghasilkan nilai ketepatan yang tinggi untuk klasifikasi kesejahteraan keluarga.

DAFTAR PUSTAKA

Agresti, A. 2002. *Categorical Data Analysis Second Edition*. Florida: Jhon Wiley & Sons, Inc.

BKKBN. 2015. *Panduan Tata Cara Pencatatan dan Pelaporan Pendataan Keluarga Tahun 2015*. Jakarta: BKKBN.

Gunn, S. R. 1998. *Support Vector Machines for Classification and Regression*. Technical *Report*. Southampon: University of Southampton.

Hartono, A. F. 2012. Implementasi Jaringan Syaraf Tiruan Backpropagtion Sebagai Sistem Pengenalan Citra Daging Babi dan Citra Daging Sapi. *UNNES Journal of Mathematics*. Vol 1. No 2. 2012.

Hosmer, D. W. and Lemeshow, S. 2000. *Applied Logistic Regression Second Edition*. Florida: Jhon Wiley & Sons, Inc.

Kecman, V. 2005. *Support Vector Machine – An Introduction*. Netherlands: Springer -Verlag Berlin Heidelberg.

Nugroho, A. S., Witarto, A. B., Handoko, D. 2003. *Support Vector Machine – Teori dan Aplikasinya dalam Bioinformatika*. [online]. [diunduh pada 18 November 2015]. Tersedia pada: http://asnugroho.net/papers/ikcsvm.pdf

Prasetyo, E. 2012. *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: Andi.

Santosa, B. 2007. *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.

Sunarti, E. 2006. *Indikator Keluarga Sejahtera: Sejarah Pengembangan, Evaluasi, dan Keberlanjutannya*. Bogor: Fakultas Ekologi Manusia IPB

Vapnik, V. N. 1999. *The Nature of Statistical Learning Theory Second Edition.* New York: Springer.

Webb, P., and Yohannes, Y. 1999. *Classification And Regression Trees, CART*. Washington D. C: International Food Policy Research Institute.