Klasifikasi Kabupaten di Provinsi Jawa Timur Berdasarkan Indikator Daerah Tertinggal dengan metode *Support Vector Machine* (SVM) dan *Entropy Based Fuzzy Support Vector Machine* (EFSVM)

Jefry Pranata Maulana dan Irhamah
Departemen Statistika, Fakultas Matematika, Komputasi, dan Sains Data
Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: irhamahn@gmail.com

Abstrak- ***Pemerintah menetapkan 4 Kabupaten dari 29 kabupaten di Provinsi Jawa Timur masuk dalam kategori daerah tertinggal pada tahun 2015. Penelitian ini akan digunakan metode Entropy Based Fuzzy Support Vector Machine (EFSVM) dan Support Vector Machine (SVM) untuk mengklasifikasikan kabupaten di Provinsi Jawa Timur dengan dan tanpa seleksi variabel. Terdapatnya imbalance pada data deerah tertinggal dimana kabupaten tertinggal jauh lebih sedikit dibandingkan kabupaten tidak tertinggal memerlukan metode klasifikasi untuk data imbalance, Salah satunya adalah EFSVM. Hasil menunjukan EFSVM memiliki Kinerja yang lebih baik pada AUC dibandingkan dengan SVM.. Seleksi variabel mampu meningkatkan AUC pada EFSVM namun tidak meningkatkan AUC pada SVM.***

Kata Kunci- Daerah Tertinggal, Data Imbalance, Entropy Fuzzy, Support Vector Machine (SVM)

I. PENDAHULUAN

Daerah tertinggal adalah daerah kabupaten yang wilayah serta masyarakatnya kurang berkembang dibandingkan dengan daerah lain dalam skala nasional [1]. Pemerintah menetapkan 4 Kabupaten pada Provinsi Jawa Timur masuk dalam kategori daerah yang tertinggal. Kabupaten tersebut adalah Kabupaten Situbondo, Kabupaten Bondowoso, Kabupaten Bangkalan, dan Kabupaten Sampang.

Pemerintah menetapkan daerah yang relatif kurang berkembang dibandingkan daerah lain dalam skala nasional sebagai daerah tertinggal. Dalam penetapan daerah tertinggal tahun 2015-2019 berdasarkan pada 6 (enam) indikator yaitu perekonomian masyarakat, kualitas dari sumberdaya manusia, prasarana, kemampuan keuangan lokal, aksesibilitas, dan karakteristik dari daerah. Masing-masing indikator terdapat sub indikator, sehingga terdapat 27 subindikator penetapan daerah tetinggal. Pemerintah menetapkan daerah menjadi tertinggal menggunakan beberapa pendekatan. Pendekatan dilakukan melalui analisa data seluruh kabupaten yang telah ditetapkan menjadi daerah tertinggal berdasarkan ketersediaan pada data-data terakhir daerah tersebut [1].

Penelitian-penelitian sebelumnya mengenai klasifikasi daerah tertinggal pernah dilakukan oleh Purwandari dan Hidayat [2] yang menggunakan Regresi Logistik Biner dengan hasil variabel yang signifikan adalah persentase penduduk miskin dan angka harapan hidup. Sebagian kekurangan dari metode-metode diatas adalah tidak mempertimbangkan terdapatnya *imbalance* pada data tersebut jumlah respon daerah tertinggal cenderung lebih sedikit dibandingkan dengan daerah tidak tertinggal. Salah satu metode klasifikasi untuk data *imbalance* adalah *Entropy Based Fuzzy Support Vector Machine* (EFSVM). EFSVM merupakan salah satu bentuk perluasan dari *Support Vector Machine* (SVM) yang menerapkan keanggotaan *fuzzy* berdasarkan *entropy* [3]. Penelitian ini bertujuan untuk klasifikasi data *imbalance* kabupaten – kabupaten di Provinsi Jawa Timur berdasarkan indikator daerah tertinggal dengan *Support Vector Machine* dan *Entropy based fuzzy support vector machine* dengan membandingkan nilai akurasi, *sensitivity, specificity*, dan AUC .

II. TINJAUAN PUSTAKA

A. *Fast Correlation Based Filter* (FCBF)

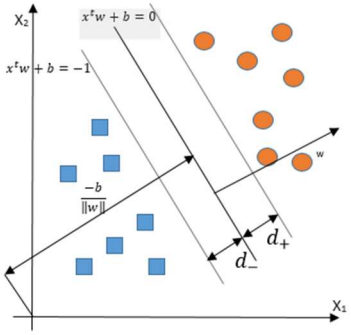
Fast Correlation Based Filter (FCBF) merupakan salah satu algoritma *feature selection* yang bersifat *multivariate* dan mengukur kelas variabel dan korelasi antara variabel-variabel berdasarkan nilai *entropy* [4]. Entropy dari variabel X dengan observasi sebanyak n didefinisikan pada Persamaan berikut.

$$H(X) = -\sum_{i=1}^{n} P(x_i)^2 \log\left(P(x_i)\right), i = 1, 2, ..., n \quad (1)$$

B. *Support vector machine* (SVM)

Support vector machine (SVM) pertama kali dikenalkan oleh Vapnik pada tahun 1992 . Prinsip dasar SVM adalah *linier classifier*, yaitu kasus klasifikasi yang secara linier dapat dipisahkan dan memaksimalkan batas- batas *hyperplane* [5].. Metode *Support Vector Machine* (SVM) melakukan klasifikasi himpunan vektor *training* berupa set data data berpasangan dari dua kelas [5].

$$\left(\mathbf{x}_i, y_i\right), \mathbf{x}_i \in R^n, y_i \in \{1, -1\}, i = 1, ..., n$$



Gambar 1 *Hyperplane*

$\mathbf{x_i'w} + b = 0$ adalah *hyperplane* pemisah. d_- dan d_+ akan menjadi jarak terpendek dari objek paling dekat dari kelas -1 dan +1. Semua observasi harus memenuhi *constraint*

$$\mathbf{x_i'w} + b \geq +1 \text{ untuk } y_i = +1 \quad (2)$$

$$\mathbf{x_i'w} + b \geq -1 \text{ untuk } y_i = -1 \quad (3)$$

Kedua *constraint* dapat disederhanakan sebagai berikut

$$y_i\left[\left(\mathbf{w'x_i}\right) + b\right] \geq 1, i = 1, 2, \ldots n \quad (4)$$

Garis pembatas pertama $\mathbf{x_i'w} + b = 1$ mempunyai bobot dan jarak tegak lurus dari titik asal sebesar

$$\frac{|1-b|}{\|\mathbf{w}\|} \quad (5)$$

garis pembatas kedua $\mathbf{x_i'w} + b = -1$ mempunyai bobot dan jarak tegak lurus dari titik asal sebesar

$$\frac{|-1-b|}{\|\mathbf{w}\|} . \quad (6)$$

Nilai maksimum margin atau nilai jarak antar bidang pembatas adalah $\frac{1-b-(-1-b)}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$. (7)

Hyperplane yang paling optimal diperoleh dengan meminimumkan $\frac{1}{2}\|\mathbf{w}\|^2$ atau meminimumkan fungsi objektif

$$\min \frac{1}{2}\mathbf{w'w} \quad (8)$$

constraint

$$y_i\left[\left(\mathbf{w'x_i}\right) + b\right] \geq 1, i = 1, 2, \ldots n \quad (9)$$

Misalkan terdapat *Hyperplane* yang optimal diperoleh denga meminimumkan $\frac{1}{2}\mathbf{w'w}$ dengan batasan $y_i\left[\left(\mathbf{w'x_i}\right) + b\right] \geq 1, i = 1, 2, \ldots n$ [1]. Permasalahan optimasi minimum diatas dapat dibentuk sebagai fungsi *lagrange* multiplier primal sebagai berikut.

$$L_p\left(\mathbf{w}, b, \alpha\right) = \frac{1}{2}\mathbf{w'w} - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\mathbf{w'x_i} + b\right) - 1\right] \quad (10)$$

Bentuk dual dari persamaan adalah sebagai berikut

$$L_d\left(\alpha\right) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x_i'x_j} \quad (11)$$

Optimasi didapatkan dengan menghitung Persamaan (11) dengan mendapatkan nilai α

$$L_d\left(\alpha\right) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x_i'x_j}$$

$$= \alpha_1 + \alpha_2 + \alpha_3 + \ldots + \alpha_n - \frac{1}{2}\alpha_1\alpha_2 y_1 y_2 \mathbf{x_1'x_2} - \frac{1}{2}\alpha_1\alpha_3 y_1 y_3 \mathbf{x_1'x_3} - \ldots - \frac{1}{2}\alpha_{n-1}\alpha_n y_{n-1} y_n \mathbf{x_{n-1}'x_n}$$

Selanjutnya L_d diturunkan terhadap α_1, α_2, $\alpha_{3,\ldots}$ α_n sebagai berikut

$$\frac{\partial L_d}{\partial \alpha_1} = 1 - \frac{1}{2}\alpha_2 y_1 y_2 \mathbf{x_1'x_2} - \frac{1}{2}\alpha_3 y_1 y_3 \mathbf{x_1'x_3} - \ldots - \frac{1}{2}\alpha_n y_1 y_n \mathbf{x_1'x_n}$$

$$\frac{\partial L_d}{\partial \alpha_2} = 1 - \frac{1}{2}\alpha_1 y_1 y_2 \mathbf{x_2'x_1} - \frac{1}{2}\alpha_3 y_2 y_3 \mathbf{x_2'x_3} - \ldots - \frac{1}{2}\alpha_n y_2 y_n \mathbf{x_2'x_n}$$

.

.

.

$$\frac{\partial L_d}{\partial \alpha_n} = 1 - \frac{1}{2}\alpha_1 y_1 y_n \mathbf{x_n'x_1} - \frac{1}{2}\alpha_2 y_2 y_n \mathbf{x_n'x_2} - \ldots \frac{1}{2}\alpha_{n-1} y_{n-1} y_n \mathbf{x_n'x_{n-1}}$$

Nilai α_1, α_2, $\alpha_{3,\ldots}$ α_n diperoleh dengan menghitung sistem persamaan diatas dan didapatkan nilai **w** dan b sebagai berikut

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x_i} \quad (12)$$

$$b = y_i - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x_i'x_j} \quad (13)$$

Fungsi *hyperplane* yang terbentuk adalah

$$f\left(x\right) = sign\left(\mathbf{w'x} + b\right) \quad (14)$$

Nilai *sign* menunjukan jika nilai $f\left(x\right) \geq 0$ maka $f\left(x\right) = +1$, jika meiliki nila $f\left(x\right) < 0$ maka $f\left(x\right) = -1$

.Pada kasus non-linier SVM, pertama-tama data $\mathbf{x}_i$ dipetakan oleh fungsi $\varphi\left(\mathbf{x_i}\right)$ ke ruang vektor yang berdimensi lebih tinggi dan membutuhkan fungsi *kernel* $K\left(\mathbf{x_i}, \mathbf{x_j}\right)$. $\mathbf{x_i}$ adalah nilai variabel **x** pada pengamatan ke *i* dan $\mathbf{x_j}$ adalah nilai variabel **x** pada pengamatan ke *j*. Fungi kernel akan menghasilkan matriks kernel yang berukuran n x n dengan n adalah banyaknya pengamatan/observasi. Beberapa fungsi dari *kernel* yang umum digunakan pada *Support Vector Machine* (SVM) adalah kernel linier, kernel *Radial Basis Function*, kernel sigmoid, dan kernel polinomial seperti pada persamaan berikut ini.

1. Kernel linier

$$K\left(\mathbf{x_i}, \mathbf{x_j}\right) = \mathbf{x_i'x_j} \quad (15)$$

2. Kernel *Radial Basis function*

$$K\left(\mathbf{x_i}, \mathbf{x_j}\right) = \exp\left(-\gamma\left\|\mathbf{x_i} - \mathbf{x_j}\right\|^2\right), \gamma > 0 \quad (16)$$

3. Kernel Polinomial

$$K\left(\mathbf{x_i}, \mathbf{x_j}\right) = \left(\gamma\mathbf{x_i'x_j} + r\right)^p, \gamma > 0 \quad (17)$$

4. Kernel Sigmoid

$$K\left(\mathbf{x_i}, \mathbf{x_j}\right) = \tanh\left(\gamma \mathbf{x_i}' \mathbf{x_j} + r\right) \quad (18)$$

Pemilihan jenis kernel akan berpengaruh kepada nilai akurasi, *sensitivity, specificity,* dan AUC. fungsi kernel yang direkomendasikan untuk diuji pertama kali adalah fungsi kernel RBF karena dapat memetakan hubungan tidak *linear* [1] . Nilai **w**, **b**, dan fungsi *hyperplane* adalah sebagai berikut

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \varphi\left(\mathbf{x_i}\right);\ b = y_i - \sum_{i=1}^{n} \alpha_i y_i K\left(\mathbf{x_i}, \mathbf{x_j}\right)$$

$$f\left(x\right) = sign\left(\sum_{i=1}^{n} \alpha_i y_i \varphi\left(\mathbf{x}\right)' \varphi\left(\mathbf{x}\right) + b\right)$$

$$= sign\left(\sum_{i=1}^{n} \alpha_i y_i K\left(\mathbf{x_i}, \mathbf{x_j}\right) + b\right) \quad (19)$$

C. Klasifikasi dengan *Entropy Fuzzy Support Vector Machine* (EFSVM)

Perbedaan antara metode SVM dan EFSVM terdapatnya pada nilai pembobot s_i pada parameter *cost* [4]. Perbedaan antara SVM dan SVM dengan penambahan keanggotaan fuzzy terletaak pada penerapan fuzzy membership (s_i). Nilai *si* yang lebih kecil mengurangi efek dari parameter cost pada persamaan sehingga sampel $\mathbf{x_i}$ dikurangi kepentingannya [4].

D Evaluasi Kinerja Klasifikasi

Pengukuran kinerja klasifikasi dilakukan untuk mengetahui kinerja klasifikasi untuk memprediksikan kelas suatu data. Hasil dari jumlah obervasi yang benar atau salah dari model *Support Vector Machine* (SVM) diklasifikasikan oleh model dapat disusun dalam sebuah *confusion matrix* berikut.

Tabel 1. *confusion matrix*

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Keterangan:
TP : Banyaknya prediksi benar pada kelas positif
FP : Banyaknya prediksi salah pada kelas positif
TN : Banyaknya prediksi benar pada kelas negatif
FN : Banyaknya prediksi salah pada kelas negatif

Berdaarkan tabel *confusion matrix* dapat dihitung Nilai Akurasi, *Sensitivity*, *Specificity* dan AUC adalah sebagai berikut

$$Akurasi = \frac{TN + TP}{TN + TP + FN + FP} \quad (20)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (21)$$

$$Specificity = \frac{TN}{TN + FP} \quad (22)$$

$$AUC = \frac{1}{2}\left(sensitivity + specificity\right) \times 100\% \quad (23)$$

E Daerah Tertinggal

Daerah tertinggal adalah daerah kabupaten yang wilayah serta masyarakatnya kurang berkembang dibandingkan dengan daerah lain dalam skala nasional [6]. Penentuan sebuah kabupaten tertinggal ditentukan beberapa indikator sebagai berikut.

a. Persentase Penduduk Miskin

Penduduk yang memiliki rata-rata pengeluaran perkapita per bulan dibawah Garis Kemiskinan dikategorikan sebagai penduduk miskin. Garis Kemiskinan (GK) merupakan penjumlahan dari Garis Kemiskinan Makanan (GKM) dan Garis Kemiskinan Non Makanan (GKNM) [7].

b. Pengeluaran per kapita per tahun disesuaikan

Pengeluaran per kapita per tahun yang disesuaikan ditentukan dari nilai pengeluaran per kapita dan paritas daya beli. Paritas daya beli pada metode baru menggunakan 96 komoditas dimana 66 komoditas merupakan makanan dan sisanya adalah komoditas nonmakanan [7].

c. Angka Harapan Hidup

Angka Harapan Hidup secara konsepsi diartikan sebagai rata-rata jumlah tahun hidup yang dapat dijalani oleh seseorang hingga akhir hayatnya [7]. Angka Harapan Hidup dihitung dari penjumahan usia orang meninggal pada tahun tersebut dibagi dengan banyaknya orang meninggal Rumus dari angka harapan hidup adalah sebagai berikut.

$$\text{Angka Harapan Hidup} = \frac{Usia_1 + Usia_2 + ... + Usia_k}{n_k}$$

n_k = banyaknya orang meninggal per tahun

d. Rata-rata Lama Sekolah

Rata-rata Lama Sekolah (RLS),didefinisikan sebagai jumlah tahun yang digunakan oleh penduduk dalam menjalani pendidikan formal. Cakupan dari penduduk yang dihitung RLS adalah penduduk berusia 25 tahun ke atas. RLS dihitung untuk usia 25 tahun ke atas dengan asumsi pada umur 25 tahun proses pendidikan telah berakhir [7].

e. Angka Melek Huruf

Angka Melek Huruf adalah perbandingan jumlah penduduk usia 15 tahun keatas yang dapat membaca dan menulis huruf latin maupun huruf lainnya terhadap seluruh penduduk usia 15 tahun keatas dikali dengan seratus [1]. Hasilnya jika semakin besar akan semakin baik atau menggambarkan kondisi/tingkat kesejahteraan yang lebih baik

f. Jumlah desa dengan jenis permukaan jalan terluas

Jenis Permukaan jalan terluas adalah jenis permukaan jalan aspal/beton, diperkeras (dengan kerikil atau batu), tanah, dan lainnya yaitu terbuat dari kayu /papan yang biasanya digunakan di daera rawa, termasuk jalan setapak, jalan di hutan dan sejenisnya [1].

g. Indeks Desa Membangun

Indeks Desa Membangun (IDM) merupakan ukuran untuk tingkat kemajuan desa pada suatu kabupaten. Semakin rendah nilai IDM maka semakin banyak desa yang berstatus sebagai desa tertinggal di kabupaten tersebut [1].

h. Indeks Kapasitas Fiskal

Indeks Kapasitas Fiskal indeks yang menunjukan kemampuan keuangan masing-masing daerah yang dicerminkan melalui penerimaan dana umum Anggaran Pendapatan dan Belanja Daerah (tidak termasuk dana alokasi khusus, dana darurat, dana pinjaman lama, dan penerimaan-penerimaan lain yang penggunaannya dibatasi untuk membiayai pengeluaran tertentu) [3].

III. METODOLOGI PENELITIAN

A. Sumber Data

Data yang digunakan dalam penelitian ini diperoleh dari Statistik Potensi Daerah 2014 [7], Profil Kesehatan Provinsi Jawa Timur 2014 [8-9] serta Realisasi Kemampuan Keuangan Daerah Tahun 2014 [10-11]. Unit penelitian ini adalah kabupaten.

B. Variabel Penelitian

Variabel-variabel yang digunakan pada penelitian ini disajikan pada Tabel 2 berikut.

Tabel 2 Variabel penelitian

Variabel	Keterangan	Skala	Satuan
Y	Y= -1 (Daerah Tidak tertinggal) Y= +1 (Daerah tertinggal)	Nominal	Desa
X_1	Persentase penduduk miskin	Rasio	Persentase
X_2	Pengeluaran per kapita yang disesuaikan	Rasio	Rupiah
X_3	Angka Harapan hidup	Rasio	Tahun
X_4	Rata-rata lama Sekolah	Rasio	Tahun
X_5	Angka Melek Huruf	Rasio	Persentase
X_6	Jumlah desa dengan jenis permukaan jalan terluas aspal/beton	Rasio	Desa
X_7	Jumlah desa dengan jenis permukaan jalan terluas diperkeras	Rasio	Desa
X_8	Jumlah desa dengan jenis permukaan jalan terluas tanah	Rasio	Desa
X_9	Persentase rumah tangga pengguna listrik	Rasio	Persentase
X_{10}	Persentase rumah tangga pengguna telepon	Rasio	Persentase
X_{11}	Persentase rumah tangga pengguna air bersih	Rasio	Persentase
X_{12}	Jumlah prasarana kesehatan per 1000 penduduk	Rasio	Prasarana
X_{13}	Jumlah desa yang memiliki pasar tanpa bangunan permanen	Rasio	Pasar
X_{14}	Jumlah dokter per 1000 penduduk	Rasio	Orang
X_{15}	Jumlah SD dan SMP per 1000 penduduk	Rasio	Sekolah
X_{16}	persentase desa gempa bumi	Rasio	Persentase
X_{17}	persentase desa tanah longsor	Rasio	Persentase
X18	persentase desa banjir	Rasio	persentase
X19	persentase desa bencana lainnya	Rasio	persentase
X20	persentase desa di kawasan hutan lindung	Rasio	persentase
X21	persentase desa konflik satu tahun terakhir	Rasio	persentase
X22	Indeks Kapasitas Fiskal	Rasio	Indeks
X23	Indeks Desa Membangun	Rasio	Indeks

C Langkah Analisis

Langkah-langkah analisis pada penelitian ini adalah sebagai berikut.

1. Statistika deskriptif untuk karakteristik data
2. *Feature selection* dengan metode FCBF
   a. Menghitung Nilai *entropy* menggunakan persamaan
   b. Menghitung nilai dari *Information Gain* dengan rumus sebagai berikut

$$IG(X \mid Y) = H(X) - H(X \mid Y)$$

3. Membagi data menjadi data *testing* dan data *training*.
4. Klasifikasi Data dengan *Support Vector Machine* (SVM).
   a. Menentukan jenis dari fungsi kernel yang digunakan pemodelan menggunakan persamaan
   b. Menentukan nilai *initial value* parameter kernel dan parameter *cost* untuk optimasi. *Initial value* berupa *range* dari parameter *cost* dan parameter dari fungsi kernel
   c. Optimasi Parameter kernel dan parameter cost terbaik berdasarkan nilai AUC
5. Klasifikasi Data dengan *Fuzzy Entropy Support Vector Machine*
   a. Menentukan k-Nearest Neighbors (k-NN) untuk masing-masing sampel
   b. Menghitung jumlah sampel positif dan negatif pada (k-NN) .
   c. Menghitung peluang dari sampel positif dan negatif

$$p_{+i} = \frac{num_{+i}}{k} \; ; p_{-i} = \frac{num_{-i}}{k}$$

   d. Menghitung nilai entropy dari dari masing-masing data *training*

$$H = -p_{+i} \ln(p_{+i}) - p_{-i} \ln(p_{-i})$$

   e. Pada sampel training $\mathbf{x_i}$ dari kelas negatif dikelompokan kedalam *m* subset kemudian menghitung nilai pada batas atas (*thrUp)* dan nilai pada batas bawah (*thrLow)* dengan rumus sebagai berikut

$$thrUp = H_{\min} + \frac{1}{m}(H_{\max} - H_{\min});$$

$$thrLow = H_{\min} + \frac{l-1}{m}(H_{\max} - H_{\min}).$$

   f. Setelah didapatkan nilai batas atas dan batas bawah pada masing-masing subset, sampel data *training* dari kelas negatif dikelompokan berdasarkan nilai *entropy*.
   g. Menghitung nilai *fuzzy membership* untuk masing-masing subset

$$FM_l = 1.0 - \beta * (l-1), l = 1, 2, ..., m$$

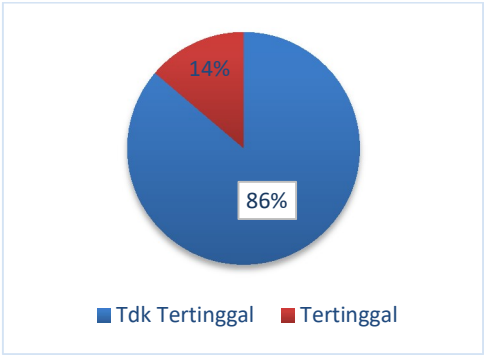
   h. Menetapkan entropy based fuzzy membership untuk semua *training* sampel x_i menggunakan Sampel dari kelas positif diberi niai $s_i = 1$, sedangkan untuk sampel dari kelas negatif memiliki nilai s_i yang sesuai dengan telah dihitung pada langkah sebelumnya. Semakin tinggi nilai *entropy,* maka semakin mendekati satu nilai dari *fuzzy membership.*

i. Optimasi Parameter kernel dan parameter *cost* terbaik berdasarkan nilai AUC. N
j. Membandingkan Kinerja hasil klasifikasi dengan metode SVM dan EFSVM berdasarkan nilai AUC. Dipilih perbandingan nilai AUC karena data yang dianalisis adalah data *imbalance*.

IV. ANALISIS DAN PEMBAHASAN

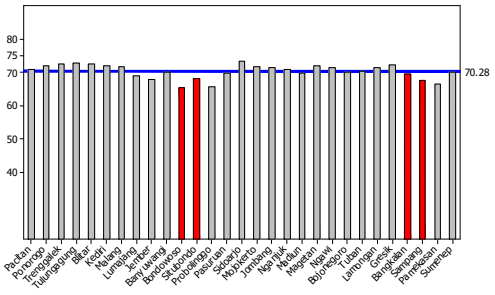
A. Karakteristik Data Daerah Tertinggal

Terdapat 29 Kabupaten di Provinsi Jawa Timur. Sebanyak 4 kabupaten atau sekitar 14% dari total kabupaten di Provinsi Jawa Timur yang berstatus sebagai daerah tertinggal.



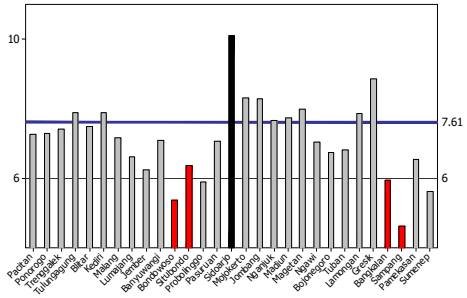
Gambar 2 Proporsi Kabupaten Tertinggal

Grafik diatas proporsi kelas kabupaten tertinggal lebih sedikit dibandingkan kelas kabupaten tidak tertinggal dengan rasio kelas mayoritas dibandingkan dengan kelas minoritas adalah 25:4 yang menunjukkan kategori *low imbalance.*



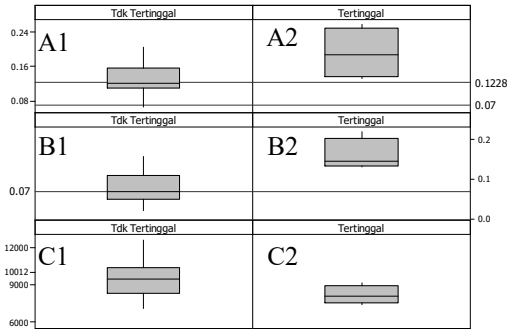
Gambar 3 Angka Harapan Hidup di Jawa Timur

Rata-rata angka harapan hidup di Provinsi Jawa Timur adalah 70,28 tahun . Grafik batang yang bewarna merah adalah kabupaten tertinggal. Semua kabupaten tertinggal memiliki angka harapan hidup di bawah rata-rata Jawa Timur.



Gambar 4 Rata-rata lama sekolah

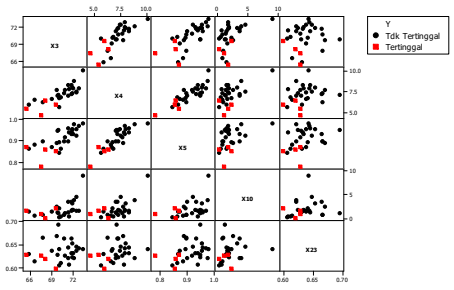
Rata-rata lama sekolah di Provinsi Jawa Timur pada masing-masing kabupaten dapat dilihat pada Gambar 3 dengan bewarna merah adalah kabupaten tertinggal. Kabupaten Sodoarjo memiliki rata-rata lama sekolah sebesar 10,12 tahun yang merupakan kabupaten nilai rata-rata lama sekolah tertinggi. Terdapat tiga Kabupaten tertinggal yaitu Kabupaten Bondowoso, Kabupaten Bangkalan dan Kabupaten Sampang bahkan memiliki rata-rata lama sekolah kurang dari 6 tahun sehingga dapat disimpulkan rata-rata penduduk yang berumur 25 tahun keatas hanya bersekolah kurang dari 6 tahun atau tidak tamat pendidikan sekolah dasar. Secara umum, rata-rata lama sekolah di Provinsi Jawa Timur terlihat sebesar 7,61 tahun.



Gambar 5 Boxplot persentase penduduk miskin (A) , angka tidak melek huruf (B), dan konsumsi per kapita (C)

Pada Boxplot C menunjukkan konsumsi per kapita di semua kabupaten tertinggal memiliki nilai yang lebih rendah dari angka konsumsi per kapita Jawa Timur yang sebesar Rp 10.012, Pada Boxplot B, kabupaten tertinggal memiliki angka tidak melek huruf yang jauh lebih tinggi dari angka nasional Jawa Timur yaitu sebesar 0,07. Semakin besar angka tidak melek huruf, maka semakin banyak penduduk yang berusia 15 tahun keatas yang tidak bisa membaca dan menulis huruf latin di kabupaten tersebut. Persentase penduduk miskin pada kabupaten yang tertinggal pada Boxplot A memiliki nilai yang lebih tinggi dari kabupaten-kabupaten lain yang tidak tertinggal.

Penyebaran data dari beberapa variabel terdapat pada gambar 6. Terdapat pola data yang menyebar sehingga sehingga sulit dipsahkan fungsi pemisah linier sehingga diperlukan bantuan kernel untuk mendapatkan fungsi pemisah yang optimal



Gambar 6 Pola persebaran data beberapa variabel

B Seleksi Variabel dengan *Fast Correlation Based Filter (FCBF)*

Seleksi fitur dengan *Fast Correlation Based Filter (FCBF)* dilakukan dengan cara memilih fitur-fitur yang relevan yang mempengaruhi hasil klasifikasi. Hasil dari seleksi fitur dengan metode FCBF terdapat pada tabel 3. Variabel yang terpilih adalah Angka

melek huruf, Rata rata lama sekolah, persentase pengguna listrik dan angka harapan hidup

Tabel 3 Variabel terpilih hasil FCBF

No	Nama Variabel	*Information Gain*
1	Angka melek huruf	0,341
2	Rata-rata lama sekolah	0,341
3	Persentase rumah tangga pengguna listrik	0,341
4	Angka Harapan Hidup	0,303

Information Gain pada variabel angka harapan hidup adalah yang terendah yaitu sebesar 0,303.

C Pembagian data *training* dan *testing* dengan Stratified K-*Fold Cross Validation*

K-*fold* yang digunakan dalam penelitian ini adalah sebanyak 4 fold. Berikut adalah hasil pembagian 4 fold.

Tabel 4 Hasil *Kfold Cross Validation*

Fold	Anggota
Fold 1	10, 5, 22, 15, 23, **12**,17
Fold 2	28, 14, 2 , 6, 29, **11**, 19
Fold 3	16, 4, 9, 24 ,7, **27**, 20
Fold 4	1, 21, 18, 25, **26** , 3 , 8, 13

Berdasarkan Tabel 4, angka yang bercetak tebal adalah kabupaten tertinggal sehingga masing-masing fold memiliki satu kabupaten tertinggal. Kabupaten dengan data ke 12, 11, 27, dan 26 berturut urut adalah Kabupaten Situbondo, Kabupaten Bondowoso, Kabupaten Sampang, dan Kabupaten Bangkalan sedangkan sisanya adalah kabupaten-kabupaten lain yang tidak tertinggal. Penentuan k-fold menggunakan *Stratified* dimana pada masing-masing fold terdapat satu kabupaten tertinggal.

D Klasifikasi Kabupaten dengan *Support Vector Machine* (SVM)

Klasikasi dengan SVM dilakukan pada semua variabel dan variabel yang telah terseleksi pada Feature Correlation Based Filter . Nilai *cost* yang dipilih pada *range* 2^{-6}, 2^{-5}, 2^{-4} ,…2^{1}, 2^{2},… 2^{5}, 2^{6} [4] dan nilai *gamma* yang dipilih pada *range* yang sama

Tabel 5 Hasil AUC dan Akurasi data *training* SVM (%)

Kernel	Variabel	AUC	Akurasi	*Cost*	*Gamma*
RBF	Semua	100	100	2^{6}	2^{-6}
RBF	Seleksi	100	100	2^{6}	2^{3}
Linier	Semua	50	86,20	2^{-6}	-
Linier	Seleksi	100	100	2^{2}	-
Sigmoid	Semua	60,12	73,48	2^{5}	2^{-4}
Sigmoid	Seleksi	95,83	73,48	2^{4}	2^{-6}

Berdasarkan Tabel 5, metode SVM dengan menggunakan kernel RBF memiliki nilai AUC yang tinggi yaitu sebesar 100% pada semua variabel dan seleksi variabel. Nilai AUC terendah didapatkan pada kernel linier dengan menggunakan semua variabel. Nilai AUC dihitung berdaarkan persamaan (23) Seleksi variabel menghasilkan nilai AUC yang lebih tinggi dari kernel linier dan kernel sigmoid. Hasil dari kinerja klasifikasi pada data *testing* ditampilkan pada Tabel 6.

Tabel 6 Hasil AUC dan Akurasi data *testing* (%)

Kernel	Variabel	AUC	Akurasi	*Sensitivity*	*Specificity*
RBF	Semua	50	86,16	0	100
RBF	Seleksi	50	86,16	0	100
Linier	Semua	50	86,16	0	100
Linier	Seleksi	34,52	56,25	0	65,48
Sigmoid	**Semua**	**66,96**	**79,46**	**50**	**83,93**
Sigmoid	Seleksi	48,81	65,63	25	72,62

Tabel 6 menunjukan nilai AUC tertinggi pada data *testing* terdapat pada SVM dengan menggunakan kernel sigmoid pada semua variabel. Nilai AUC daan *sensitivity* pada kernel RBF sebesar 50% dan 0%. Hal ini menunjukan SVM dengan kernel RBF hanya mampu mengklasifikasikan kabupaten yang tidak tertinggal dengan benar. Model yang terbentuk dapat dilihat pada Tabel 7. Pada data *testing,* metode SVM dengan menggunakan seleksi variabel tidak meningkatkan nilai AUC.

Tabel 7 Model SVM dari tiga jenis kernel

Kernel	Model
	Semua Variabel
RBF	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \exp\left(-2^{3}\left\|x_i - x_j\right\|^2 - 0.1190523\right)$
Linier	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \left(x_i' x_j + 0.05041696\right)$
Sigmoid	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \left(\tanh\left(2^{-4}\mathbf{x}_i'\mathbf{x}_j + 0\right) + 0.003388564\right)$
	Seleksi Variabel
RBF	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \exp\left(-2^{3}\left\|x_i - x_j\right\|^2 - 0.1839123\right)$
Linier	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \left(x_i' x_j + 0.3348949\right)$
Sigmoid	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \left(\tanh\left(2^{-6}\mathbf{x}_i'\mathbf{x}_j + 0\right) - 0.4186657\right)$

E Klasifikasi Kabupaten dengan *Entropy Based Fuzzy Support Vector Machine* (EFSVM)

Parameter *cost* yang dipilih pada EFSVM adalah dipilih pada *range* 2^{-6}, 2^{-5}, 2^{-4} ,…2^{-1}, 2^{1}, 2^{2},… 2^{6} [4]. Parameter *gamma* yang dipilih pada *range* 2^{-6}, 2^{-5}, 2^{-4} ,…2^{-1}, 2^{1}, 2^{2},… 2^{6}. Hasil nilai AUC dan akurasi pada data *training* adalah sebagai berikut

Tabel 8 Hasil AUC dan Akurasi data *training* EFSVM (%)

Kernel	Variabel	AUC	Akurasi	*Cost*	*Gamma*
RBF	Semua	97,98	96,54	2^{-5}	2^{-4}
RBF	Seleksi	93,31	88,47	2^{-4}	2^{-5}
Linier	Semua	97,37	95,46	2^{-5}	-
Linier	Seleksi	92,65	87,34	2^{-4}	-
Sigmoid	Semua	98,03	96,59	2^{-6}	2^{-3}
Sigmoid	Seleksi	93,31	88,47	2^{-5}	2^{6}

Tabel 8 menunjukan nilai AUC yang didapatkan pada data *training* dengan menggunakan variabel seleksi memiliki nilai AUC yang lebih rendah dari AUC dari semua variabel . Hasil ini menunjukan seleksi variabel tidak meningkatkan nilai AUC pada data *training*. Nilai AUC tertinggi pada data *training* terdapat pada kernel sigmoid dengan menggunakan semua variabel. Hasil nilai AUC dan akurasi pada *testing* dapat dilihat seperti Tabel 9

Tabel 9 Hasil kinerja data *testing* EFSVM (%)

Kernel	Variabel	AUC	Akurasi	*Sensitivity*	*Specificity*
RBF	Semua	79,76	83,04	75	84,52
RBF	**Seleksi**	**92,26**	**86,61**	**100**	**84,52**
Linier	Semua	65,48	76,79	50	80,95
Linier	**Seleksi**	**92,26**	**86,61**	**100**	**84,52**
Sigmoid	Semua	65,48	76,79	50	80,95
Sigmoid	Seleksi	91,96	86,16	100	83,93

Tabel 9 menunjukan metode *Entropy Based Support Vector Machine* (EFSVM) dengan menggunakan seleksi variabel kernel RBF dan kernel linier menghasilkan nilai AUC yang sama yaitu sebesar 92,26 % yang merupakan AUC tertinggi . Hasil yang berbeda dari data *training* ternyata didapatkan seleksi variabel dengan *Fast Correlation Based Filter* (FCBF) ternyata mampu meningkatkan AUC pada data *testing*. Nilai *Sensitivity* sebesar 100 menunjukan model EFSVM mampu mengklasifikasikan semua kabupaten tertinggal dengan benar.

Hasil kinerja dari data *testing* terbaik dapat dilihat pada Tabel 10.

Tabel 10 Evaluasi Kinerja Metode Klasifikasi (%)

Metode	Kernel	AUC	Akurasi	*Sensitivity*	*Specificity*
SVM Semua	Sigmoid	66,96	79,46	50	83,93
EFSVM Seleksi	**RBF**	**92,26**	**86,61**	**100**	**84,52**
EFSVM Seleksi	**Linier**	**92,26**	**86,61**	**100**	**84,52**

Pada Tabel 10 menunjukan klasifikasi terbaik dari model SVM hanya mempu mendapatkan nilai AUC sebesar 66,9% dengan akurasi sebesar 79,46%. Metode EFSVM memiliki nilai *sensitivity* sebesar 100% yang berarti mampu mengklasifikasikan kelas minoritas dengan benar atau mampu mengklasifikasikan kabupaten tertinggal dengan benar.

V KESIMPULAN

Hasil analisis dan pembahasan yang telah dipaparkan dapat diambil kesimpulan sebagai berikut. Hasil model terbaik klasifikasi data *imbalance* kabupaten di Provinsi Jawa Timur dengan meggunakan Indikator Daerah tertinggal tertinggal adalah dengan metode EFSVM dari variabel yang telah terseleksi dengan kernel *Radial Basis Function* (RBF) dan kernel *linier.* Seleksi variabel dengan menggunakan FCBF mampu meningkatkan nilai dari AUC pada metode *Entropy Based Fuzzy Support Vector Machine* (EFSVM) namun tidak meningkatkan nilai AUC pada metode *Support Vector Machine.* (SVM) Saran dari hasil penelitian diatas, diharapkan menggunakan data dengan jumlah data yang lebih. banyak yang memiliki perbandingan kelas minoritas dan mayoritas lebih besar. Perlu dilakukan perbandingan dengan metode EFSVM dengan nilai β, *m, k* dan posisi kabupaten tertinggal pada masing-masing fold yang berbeda-beda.

DAFTAR PUSTAKA

[1] W. C. Hsu, C. C. Chang and C. J. Lin, A Practical Guide to Support Vector Machine., Taipei: Departement of Computer Science National Taiwan University, 2004.

[2] H. Liu and Y. Lei, "Feature Selection for High Dimensional Data : A Fast Correlation Based Filter Solution," *Proceeding of Twentieth International Conference on Machine Learning (ICML-2003),* 2003.

[3] N. Hasanah, Analisis Diskriminan untuk Evaluasi status Ketertinggalan Kabupaten, Bogor: Jurusan Statistika Institut Pertanian Bogor, 2009.

[4] V. Lisna, B. Sinaga, M. Firdaus and S. Sutomo, "Dampak Kapasitas Fiskal terhadap Penurunan Kemiskinan: Suatu Analisis Simulasi Kebijakan," *Jurnal Ekonomi dan Pembangunan Indonesia,* pp. 1-26, 2013.

[5] S. Gunn, Support Vector Machine for Clasification and Regression, Southamton: University of Southamton, 1998.

[6] Q. Fan, Z. Wang, D. Li, D. Gao and H. Zha, "Entropy Based Fuzzy Support Vector Machine for Imbalance Datasets," *Knowledge Based System,* pp. 87-99, 2017.

[7] Badan Pusat Statistik, Statistik Potensi Desa Jawa Timur, Jakarta: Badan Pusat Statistik., 2014.

[8] K. Desa, Rencana Strategis Direktorat Jenderal Pembangunan Daerah Tertinggal 2015-2019, Jakarta: Kementerian Desa, Pembangunan Daerah Tertinggal dan Transmigrasi, 2015.

[9] Dinas Kesehatan Provinsi Jawa Timur, Profil Kesehatan Provinsi Jawa Timur, Surabaya: Dinas Kesehatan Provinsi Jawa Timur, 2014 .

[10] Direktorat Jenderal Perimbangan Keuangan Kementerian Keuangan, Affirmative Policy Dalam Percepatan Pembangunan Daerah Untuk Meningkatkan Kesejahteraan Rakyat., Jakarta: Kementerian Keuangan, 2013.

[11] T. &. H. Y. Purwandari, Pemodelan Ketertinggalan Daerah, Bandung: Jurusan Statistika, Universitas Padjadjaran, 2017.