TUGAS AKHIR – SS141501

KLASIFIKASI KABUPATEN DI PROVINSI JAWA TIMUR BERDASARKAN INDIKATOR DAERAH TERTINGGAL DENGAN METODE *SUPPORT VECTOR MACHINE* DAN *ENTROPY BASED FUZZY SUPPORT VECTOR MACHINE*

JEFRY PRANATA M
NRP 06211645000026

Dosen Pembimbing
Irhamah, M.Si.,PhD

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2018

TUGAS AKHIR – SS141501

KLASIFIKASI KABUPATEN DI PROVINSI JAWA TIMUR BERDASARKAN INDIKATOR DAERAH TERTINGGAL DENGAN METODE *SUPPORT VECTOR MACHINE* DAN *ENTROPY BASED FUZZY SUPPORT VECTOR MACHINE*

JEFRY PRANATA M
NRP 06211645000026

Dosen Pembimbing
Irhamah, M.Si.,PhD

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2018

FINAL PROJECT – SS141501

CLASSIFICATION OF REGENCIES IN EAST JAVA BASED ON UNDERDEVELOPED REGION INDICATORS USING SUPPORT VECTOR MACHINE AND ENTROPY BASED FUZZY SUPPORT VECTOR MACHINE

JEFRY PRANATA M
SN 06211645000026

Supervisor
Irhamah, M.Si.,PhD

UNDERGRADUATE PROGRAMME
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS, COMPUTING, AND DATA SCIENCE
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2018

LEMBAR PENGESAHAN

KLASIFIKASI KABUPATEN DI PROVINSI JAWA TIMUR BERDASARKAN INDIKATOR DAERAH TERTINGGAL DENGAN METODE *SUPPORT VECTOR MACHINE* DAN *ENTROPY BASED FUZZY SUPPORT VECTOR MACHINE*

TUGAS AKHIR
Diajukan Untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Sains
Pada
Program Studi Sarjana Departemen Statistika
Fakultas Matematika, Komputasi, dan Sains Data
Institut Teknologi Sepuluh Nopember

Oleh :
Jefry Pranata Maulana
NRP. 062116 4500 0026

Disetujui oleh Pembimbing:

Irhamah, M.Si, Ph.D ( )
NIP. 19780406 200112 2 002

Mengetahui,
Kepala Departemen

Dr. Suhartono
NIP. 19710929 199512 1 001

SURABAYA, JULI 2018

KLASIFIKASI KABUPATEN DI PROVINSI JAWA TIMUR BERDASARKAN INDIKATOR DAERAH TERTINGGAL DENGAN METODE *SUPPORT VECTOR MACHINE* DAN *ENTROPY BASED FUZZY SUPPORT VECTOR MACHINE*

Nama Mahasiswa : Jefry Pranata M
NRP : 06211645000026
Departemen : Statistika-FMKSD-ITS
Dosen Pembimbing : Irhamah M.Si, PhD

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 Based Fuzzy, Support Vector Machine (SVM)

**********Halaman ini sengaja dikosongkan**********

CLASSIFICATION REGENCIES OF EAST JAVA BASED ON UNDERDEVELOPED REGION INDICATORS USING SUPPORT VECTOR MACHINE AND ENTROPY BASED FUZZY SUPPORT VECTOR MACHINE

Name : **Jefry Pranata M**
Student Number : **06211645000026**
Department : **Statistics**
Supervisors : **Irhamah, M.Si, Ph.D**

Abstract

The government set four regions in East Java to fall into the category of underdeveloped regions. There is an imbalance in the left behind data where there are far fewer regions than the non-disadvantaged regions. Therefore, a classification method is required which takes into account the imbalance in the data. One method of classification is the Support Vector Machine (SVM). Suppor Vector Manchine does not have high accuracy if applied to imbalance data. Therefore, a classification method for imbalance data is required. One of the classification methods for imbalance data is Entropy Based Fuzzy Support Vector Machine. This research will use EFSVM and SVM method to classify the regency in East Java province with and without variable selection. The result shows that EFSVM has better classification performance with AUC of 92.26%

Keyword*: Underdeveloped region, Imbalanced data, *Entropy Based Fuzzy, Support Vector Machine

**********This Page is intentionally left blank **********

KATA PENGANTAR

Puji Syukur kehadirat allah SWT, atas limpahan rahmat yang tidak pernah berhenti sehingga penulis dapat menyelesaikan Tugas Akhir yang berjudul **"KLASIFIKASI KABUPATEN DI PROVINSI JAWA TIMUR BERDASARKAN INDIKATOR DAERAH TERTINGGAL DENGAN METODE *SUPPORT VECTOR MACHINE DAN ENTROPY BASED FUZZY SUPPORT VECTOR MACHINE*"** dengan baik. Semua ini dari-Mu, karena-Mu, dan untuk-Mu. Penulis menyadari bahwa dalam penyusunan Tugas Akhir ini tidak terlepas dari bantuan dan dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Bapak Dr. Suhartono, S.Si.,M.sc selaku Kepala Departemen Statistika Fakultas Matematika, Komputasi, dan Sains Data Institut Teknologi Sepuluh Nopember Surabaya.
2. Bapak Dr. Sutikno, S.Si. M.Si, selaku Ketua Program Studi S1 Departemen Statistika Fakultas Matematika, Komputasi, dan Sains Data Institut Teknologi Sepuluh Nopember Surabaya.
3. Ibu Dr. Irhamah, M.Si selaku dosen pembimbing sekaligus dosen yang telah sabar dalam memberikan bimbingan dan saran.
4. Ibu Dr. Santi Wulan Purnami, S.Si selaku dosen wali atas dukungan dan semangat yang diberikan sewaktu perwalian.
5. Ibu dan bapak saya atas segala doa, kasih sayang dan dukungan yang tidak pernah habisnya.
6. Erin Relani yang selalu ada, memberikan semangat, cinta, kasih sayang, dan doa. Terimakasih atas segalanya, semoga sukses selalu. Bisa mencapai cita-cita dan bisa membanggakan orang tua serta orang-orang di sekitarnya.
7. Diana Sari dan Hendra Setiawan yang telah membantu mengerjakan Tugas Akhir. Terima kasih telah menjadi teman pembimbing yang baik.
8. Bapak Ali dan Keluarga yang selalu memberi tempat ketenangan untuk mengerjakan tugas akhir.

9. Teman-teman S1 Lintas Jalur Angkatan 2016, teman yang ada di kost yang selalu seru dan banyak berbagi, sahabat komunitas android di DTC dan yang lainnya.

Penulis sangat berharap hasil Tugas Akhir ini dapat bermanfaat bagi kita semua, serta kritik dan saran yang bersifat membangun guna perbaikan di masa mendatang.

Surabaya, Juli 2018

Penulis

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

**********Halaman ini sengaja dikosongkan**********

DAFTAR TABEL

Halaman

**********Halaman ini sengaja dikosongkan**********

DAFTAR LAMPIRAN

Halaman

**********Halaman ini sengaja dikosongkan**********

BAB I
PENDAHULUAN

1.1 Latar Belakang

Jawa Timur merupakan salah satu provinsi dengan sumbangan pada Produk Dromestik Bruto (PDRB) terbesar kedua di Indonesia setelah DKI Jakarta. Sumbangan terhadap kontribusi PDRB pada Provinsi Jawa Timur adalah sebesar 24,8 % terhadap PDRB total nasional atas harga berlaku pada tahun 2015 (Suryowati, 2016). Kontribusi nilai PDRB yang cukup besar menunjukan aktivitas ekonomi yakni perdagangan pertambangan, perindustrian, pertani--an dan lain-lain pada wilayah Jawa Timur yang cukup besar. Namun pada tahun 2015, Pemerintah menetapkan 4 Kabupaten pada Provinsi Jawa Timur masuk dalam kategori daerah yang tertinggal. Pemerintah menetapkan daerah yang relatif kurang berkembang dibandingkan daerah lain dalam skala nasional sebagai daerah tertinggal. Pada tahun 2015, pemerintah menetapkan 122 kabupaten tertinggal di Indonesia. Pemerintah menetapkan daerah tertinggal setiap lima tahun sekali. Persentase daerah tertinggal berdasarkan pulau di Indonesia adalah sebanyak 10,66% di Pulau Sumatera, Pulau Jawa & Pulau Bali sebanyak 4,92%, Pulau Kalimantan sebanyak 9,84% pulau Sulawesi sebanyak 14,75% pulau Maluku sebanyak 11,48%, Kepulauan Nusa Tenggara Barat dan Kepulauan Nusa Tenggara Timur sebanyak 21,31% Pulau Papua sebanyak 27,05%. Terdapat 6 kabupaten tertinggal di Pulau Jawa yang terdiri dari 4 kabupaten di Provinsi Jawa Timur dan 2 kabupaten di Provinsi Banten.

Daerah tertinggal di Provinsi Jawa Timur adalah Kabupaten Sampang, Kabupaten Bangkalan, Kabupaten Situbondo dan Kabupaten Bondowoso. Hal ini menunjukan terdapat ketimpangan pembangunan ekonomi pada kabupaten/kota pada pada Provinsi Jawa Timur karena memiliki PDRB terbesar kedua di Indonesia namun empat kabupaten masuk kategori daerah tertinggal. Sebagai implementasi otonomi daerah, Jawa Timur terdiri dari 29 Kabupaten dan 9 Kota. Secara umum, kabupaten memiliki jumlah infrastruktur yang lebih rendah diban dingkan dengan jumlah infrastruktur kota. Pada dasarnya pembangunan daerah merupakan kewenangan dari

pemerintah daerah baik provinsi maupun kabupaten, sedangkan pemerintah berfungsi sebagai motivator dan fasilitator dalam percepatan pembangunan pada daerah tertinggal. Pembangunan daerah tertinggal tidak mungkin berhasil tanpa dukungan dan kerja keras para pemangku kepentingan (*stakeholders*). Pemerintah menetapkan daerah tertinggal menggunakan beberapa pendekatan. Pendekatan dilakukan melalui analisa data seluruh kabupaten yang telah ditetapkan menjadi daerah tertinggal berdasarkan ketersediaan pada data-data terakhir.

Dalam penetapan daerah tertinggal tahun 2015-2019 berdasarkan pada 6 (enam) indikator yaitu perekonomian masyarakat, kualitas sumberdaya manusia, prasarana, kemampuan keuangan lokal, aksesibilitas, dan karakteristik daerah. Masing-masing indikator terdapat sub indikator, sehingga terdapat 27 subindikator penetapan daerah tetinggal. Subindikator tersebut adalah persentase penduduk miskin, pengeluaran konsumsi per kapita, angka harapan hidup, rata-rata lama sekolah, angka melek huruf, jumlah desa dengan jenis permukaan jalan terluas aspal/beton, jumlah desa dengan jenis permukaan jalan terluas diperkeras, jumlah desa dengan jenis permukaan jalan terluas tanah, jumlah desa dengan jenis permukaan jalan terluas lainnya, persentase rumah tangga pengguna listrik, persentase rumah tangga pengguna telepon, persentase rumah tangga pengguna air bersih, jumlah desa yang memiliki pasar tanpa bangunan permanen, jumlah sarana dan prasarana kesehatan per 1000 penduduk, jumlah dokter per 1000 penduduk, jumlah SD dan SMP per 1000 penduduk, rata-rata jarak dari kantor desa/kelurahan ke kantor kabupaten yang membawahi, jumlah desa dengan akses ke pelayanan kesehatan lebih dari 5 km, jarak desa ke pelayanan pendidikan dasar, persentase desa gempa bumi, persentase desa tanah longsor, persentase desa banjir, persentase desa bencana lainnya, persentase desa di kawasan hutan lindung, persentase desa berlahan kritis, persentase desa konflik satu tahun terakhir dan kemampuan keuangan daerah.

Indeks kapasitas fiskal dapat digunakan variabel kemampuan keuangan daerah karena terdiri dari komponen-komponen Pendapatan Asli Daerah, Dana Bagi Hasil, Dana Alokasi Umum dan pendapatan daerah lainnya, dimana dana bagi hasil dan dana alokasi

umum termasuk dalam dana transfer atau dana perimbangan (Lisna, Sinaga, Firdaus, & Sutomo, 2013). Kabupaten-kabupaten tertinggal yang dikatagorikan maju merupakan kabupaten yang terentaskan dari ketertinggalan, namun untuk menjaga agar status ketertinggalan stabil dan tidak turun lagi, maka kabupaten yang dikategorikan maju masih akan tetap dilakukan pembinaan oleh Kementerian Desa, Pembangunan Daerah Tertinggal dan Transmigrasi dan juga Kementerian /Lembaga lainnya.

Penelitian-penelitian sebelumnya mengenai daerah tertinggal pernah dilakukan oleh Oktora, Darwis dan Setyawanto (2015) dengan menggunakan metode Multivariate Adaptive Regression Splines (MARS) dengan hasil variabel yang paling berpengaruh membedakan daerah tertinggal dengan tidak tertinggal adalah pengeluaran konsumsi per kapita. Purwandari dan Hidayat (2017) menggunakan Regresi Logistik Biner dengan hasil variabel yang signifikan adalah persentase penduduk miskin dan angka harapan hidup. Handayani menggunakan metode regresi logistik biner dengan hasil variabel yang signifikan adalah persentase penduduk miskin, angka harapan hidup, dan angka melek huruf . Sulasih (2016) menggunakan metode *Rare Event Weighted Logistic Regression* dan *Truncated Regularied Iteratively Reweighted Least Square* pada data imbal*ance* desa tertinggal di Provinsi Jawa Timur dengan hasil *Rare Event Weighted Logistic Regression* dapat memprediksi desa tertinggal dengan baik.

Sebagian kekurangan dari metode-metode diatas adalah tidak mempertimbangkan terdapatnya adanya *imbalance* pada data tersebut jumlah respon daerah tertinggal selalu cenderung jauh lebih sedikit dibandingkan dengan daerah tidak tertinggal. *Imbalance* pada data menjadi salah satu tantangan pada komunitas data *mining* (Sun, 2009). Kelas *imbalance* terjadi ketika sebuah data didominasi oleh kelas mayoritas atau kelas yang secara signifikan lebih banyak kejadian daripada kelas yang langka atau kelas minoritas (Canedo, 2014). Umumnya, kelas dengan jumlah yang banyak ditandai sebagai kelas negatif, sedangkan kelas dengan jumlah sampel yang sedikit dinamakan kelas positif. *Support Vector Machine* (SVM) merupakan bagian dari metode pembelajaran yang digunakan untuk

klasifikasi. SVM memetakan vektor input ke sebuah ruang dimensi yang lebih tinggi dimana *hyperplane* pemisah dibangun.

Ide dari dasar SVM adalah memaksimalkan batas *hyperplane*. *Hyperplane* dengan margin yang maksimal akan memberikan generalisasi yang lebih baik pada metode klasifikasi. SVM bekerja dengan baik pada himpunan data berdimensi tinggi (Pamuji , 2015). Suppor Vector Manchine tidak memiliki akurasi tinggi jika diterapkan pada data *imbalance* karena akan berakibat pada model yang terbentuk lebih cenderung merepresentasikan data pada kelompok negatif (Akbani, Kwek, & Japkowicz, 2005). Metode *Entropy based Fuzzy Support Vector Machine* adalah metode klasifikasi yang dapat diterapkan untuk data *imbalance* (Fan, Wang, Li, Gao, & Zha, 2017). Dalam teori informasi, *entropy* adalah ukuran efektif untuk kepastian. Shannon (2001) mendefinisikan *entropy* sebagai fungsi logaritmik negatif probabilitas terjadinya suatu peristiwa. Pada penelitian ini akan digunakan metode *Entropy Fuzzy Support Vector Machine* untuk mengklasifikasikan Kabupaten/Kota di Provinsi Jawa Timur dengan 27 indikator daerah tertinggal. *Support Vector Machine* (SVM) digunakan sebagai perbandingan peformansi klasifikasi. Seleksi variabel yang digunakan adalah *Fast Correlation Based Filter* (FCBF) dengan performansi klasifikasi meliputi akurasi, *sensitivity*, *specificity* dan *Area Under ROC Curve* AUC.

1.2 Rumusan Masalah

Penentuan daerah tertinggal pada dasarnya adalah bagaimana cara mengelompokkan atau mengklasifikasikan sejumlah daerah ke dalam kelompok tersebut dengan memperhatikan indikator yang ada. Data daerah tertinggal umumnya memiliki kondisi yang imbalanced. Hal ini berakibat hasil dari akurasi dan prediksi yang baik terhadap kelas negatif (mayoritas) sedangkan untuk kelas positif (minoritas) akan dihasilkan akurasi prediksi yang kurang baik. Dari uraian diatas,maka permasalahan dalam penelitian ini adalah menerapkan *Entropy based Fuzzy Support Vector Machine* (EFSVM) untuk klasifikasi imbalanced data dan menggunakan *Fast Correlation Based Filter* (FCBF) sebagai seleksi variabel

1.3 Tujuan Penelitian

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* dan membandingkan nilai akurasi, *sensitivity, specificity*, dan AUC.

1.4 Manfaat Penelitian

Manfaat dari penelitian ini adalah dapat memberikan alternatif metode klasifikasi khususnya untuk data *imbalance*.

1.5 Batasan Masalah

Data yang digunakan adalah data daerah tertinggal tahun 2015. variabel yang digunakan hanya 23 indikator daerah terteinggal dari Kementerian Desa dan variabel kemampuan keuangan daerah yang digunakan hanya Indeks Kapasitas Fiskal. Kernel yang digunakan adalah *radial basis function*, linier, dan sigmoid

**********Halaman ini sengaja dikosongkan**********

BAB II
TINJAUAN PUSTAKA

2.1 *Fast Correlation Based Filter* (FCBF)

Fast Correlation Based Filter (FCBF) merupakan salah satu algoritma variabel *selection* yang bersifat *multivariate* dan mengukur kelas variabel dan korelasi antara variabel-variabel (Alonso dkk 2015). Secara umum, variabel dikatakan bagus jika variabel tersebut relevan dengan konsep kelas namun tidak redundan pada variabel yang lain. Jika diterapkan korelasi antara dua variabel sebagai ukuran kebaikan, maka sebuah variabel dikatakan bagus untuk klasifikasi jika berkorelasi sangat tinggi dengan kelas namun tidak berkorelasi dengan variabel lainnya. Namun pengukuran dengan korelasi tidak mampu menangkap korelasi yang tidak linier selain itu korelasi mengharuskan semua variabel dan kelas mengandung nilai numerik. Penyelesaian untuk mengatasi kekurangan ini, Yu dan Liu (2009), menerapkan pendekatan lain yaitu memilih ukuran korelasi berdasarkan konsep *information theoretical entropy*. *Entropy* dari pengamatan X dengan variabel sebanyak *n* didefinisikan pada Persamaan berikut.

$$H(X) = -\sum_{i=1}^{n} P(x_i)\left(^{2}\log\left(P(x_i)\right)\right), i = 1, 2, ..., n \tag{2.1}$$

Entropy dari Pengamatan X jika diketahui pengamatan Y adalah sebagai berikut

$$H(X \mid Y) = -\sum_{i=1}^{n} P(y_i) \sum_{i=1}^{n} P(x_i \mid y_i)\ ^{2}\log\left(P(x_i \mid y_i)\right),\ i = 1, 2, ..., n \tag{2.2}$$

Dari *entropy* tersebut dapat diperoleh *Information Gain* sebagai berikut

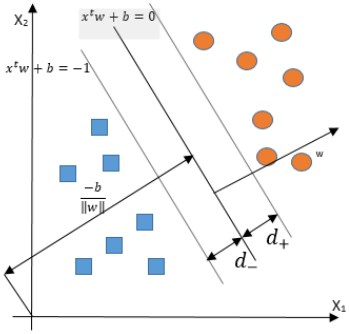
$$IG(X \mid Y) = H(X) - H(X \mid Y) \tag{2.3}$$

Untuk mengukur korelasi antar variabel, maka digunakan *symmetrical uncertainty*. Nilai *symmetrical uncertainty* berkisar pada rentang 0 sampai dengan 1. *Symmetrical uncertainty* dirumuskan sebagai berikut

$$SU(X, Y) = 2\left[\frac{IG(X \mid Y)}{H(X) + H(Y)}\right] \tag{2.4}$$

2.2 *Support vector machine* (SVM)

Support vector machine (SVM) pertama kali dikenalkan oleh Vapnik pada tahun 1992 pada saat dipresentasikan di *Annual Workshop on Computational Learning Theory* . Prinsip dasar SVM adalah linier *classifier*, yaitu kasus klasifikasi yang secara linier dapat dipisahkan. Misalkan diberikan himpunan $\mathbf{X}=\{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_n\}$ dinyatakan sebagai kelas positif jika $f(\mathbf{x}) \geq 0$ dan yang lainnya termasuk ke dalam kelas negatif.



Gambar 2.1 *Hyperplane* pada kasus linier

SVM melakukan klasifikasi himpunan vektor *training* berupa set data data berpasangan dari dua kelas (Gunn, 1998).

$(\mathbf{x}_i, y_i), \mathbf{x}_i \in R^n, y_i \in \{1,-1\}, i = 1,...,n$; $\mathbf{x}_i'\mathbf{w} + b = 0$ adalah *hyperplane* pemisah. d_- dan d_+ akan menjadi jarak terpendek dari objek paling dekat dari kelas -1 dan +1. Semua observasi harus memenuhi *constraint*

$$\mathbf{x}_i'\mathbf{w} + b \geq +1 \text{ untuk } y_i = +1 \quad (2.5)$$

$$\mathbf{x}_i'\mathbf{w} + b \geq -1 \text{ untuk } y_i = -1 \quad (2.6)$$

Kedua *constraint* dapat disederhanakan sebagai berikut

$$y_i\left[(\mathbf{w}'\mathbf{x}_i) + b\right] \geq 1, i = 1, 2, ...n \quad (2.7)$$

Garis pembatas pertama $\mathbf{x}_i'\mathbf{w} + b = 1$ mempunyai bobot dan jarak tegak lurus dari titik asal sebesar $\frac{|1-b|}{\|\mathbf{w}\|}$, sedangkan garis pembatas

kedua $\mathbf{x}_i'\mathbf{w} + b = -1$ mempunyai bobot dan jarak tegak lurus dari titik asal sebesar $\frac{|-1-b|}{\|\mathbf{w}\|}$. Nilai maksimum margin atau nilai jarak antar bidang pembatas adalah $\frac{1-b-(-1-b)}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$.

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

fungsi objektif

$$\min \frac{1}{2}\mathbf{w}'\mathbf{w} \tag{2.8}$$

constraint

$$y_i\left[\left(\mathbf{w}'\mathbf{x}_i\right) + b\right] \geq 1, i = 1, 2, ...n \tag{2.9}$$

Permasalahan optimasi minimum diatas dapat dibentuk sebagai fungsi *lagrange* multiplier sebagai berikut.

$$L\left(\mathbf{w}, b, \alpha\right) = \frac{1}{2}\mathbf{w}'\mathbf{w} - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\mathbf{w}'\mathbf{x}_i + b\right) - 1\right] \tag{2.10}$$

constraint

$$y_i\left[\left(\mathbf{w}'\mathbf{x}_i\right) + b\right] \geq 1, i = 1, 2, ...n \tag{2.11}$$

Nilai α_i adalah pengganda fungsi Lagrange. Persamaan diatas merupakan *primal space* sehingga perlu ditransformasi menjadi *dual space* agar lebih mudah dan efisien untuk diselesaikan (Gunn, 1998). Mengubah bentuk *primal* ke dual akan mendapatkan constraint yang lebih sederhana. Turunan pertama fungsi *lagrange primal space* terhadap **w** dan b adalah sebagai berikut

$$\frac{\partial \mathrm{L_p}\left(\mathbf{w}, \mathrm{b}, \alpha\right)}{\partial \mathbf{w}} = 0 \ : \ \mathbf{w} - \sum_{\mathrm{j}=1}^{\mathrm{n}}\alpha_\mathrm{i}\mathrm{y}_\mathrm{i}\mathbf{x}_\mathbf{i} = 0 \tag{2.12}$$

$$\frac{\partial \mathrm{L_p}\left(\mathbf{w}, \mathrm{b}, \alpha\right)}{\partial \mathrm{b}} = 0 \ : \ \sum_{\mathrm{j}=1}^{\mathrm{n}} \alpha_{\mathrm{i}} \mathrm{y}_{\mathrm{i}} = 0 \qquad (2.13)$$

Subtitusi persamaan 2.8 dan 2.9 kedalam *primal space* diperoleh *dual space* sebagai berikut

$$L_d\left(\alpha\right) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_{\mathbf{j}} \qquad (2.14)$$

constraint

$$\alpha_{\mathrm{i}} \geq 0 \qquad (2.15)$$

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}} \mathrm{y}_{\mathrm{i}} = 0 \qquad (2.16)$$

Nilai optimasi untuk didapatkan dengan menyelesaikan Persamaan (2.14) dengan mendapatkan nilai α

$$\begin{aligned} L_d\left(\alpha\right) &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_{\mathbf{j}} \\ &= \alpha_1 + \alpha_2 + \alpha_3 + \ldots + \alpha_n - \frac{1}{2} \alpha_1 \alpha_2 y_1 y_2 \mathbf{x}_1' \mathbf{x}_2 \\ &\quad - \frac{1}{2} \alpha_1 \alpha_3 y_1 y_3 \mathbf{x}_1' \mathbf{x}_3 - \ldots - \frac{1}{2} \alpha_{n-1} \alpha_n y_{n-1} y_n \mathbf{x}_{n-1}' \mathbf{x}_n \end{aligned} \qquad (2.17)$$

Selanjutnya L_d diturunkan terhadap α_1, $\alpha_{2,}$ $\alpha_{3,\ldots}$ α_n

$$\frac{\partial L_d}{\partial \alpha_1} = 1 - \frac{1}{2} \alpha_2 y_1 y_2 \mathbf{x}_1' \mathbf{x}_2 - \frac{1}{2} \alpha_3 y_1 y_3 \mathbf{x}_1' \mathbf{x}_3 - \ldots - \frac{1}{2} \alpha_n y_1 y_n \mathbf{x}_1' \mathbf{x}_n$$

$$\frac{\partial L_d}{\partial \alpha_2} = 1 - \frac{1}{2} \alpha_1 y_1 y_2 \mathbf{x}_2' \mathbf{x}_1 - \frac{1}{2} \alpha_3 y_2 y_3 \mathbf{x}_2' \mathbf{x}_3 - \ldots - \frac{1}{2} \alpha_n y_2 y_n \mathbf{x}_2' \mathbf{x}_n$$

.

.

.

$$\frac{\partial L_d}{\partial \alpha_n} = 1 - \frac{1}{2} \alpha_1 y_1 y_n \mathbf{x}_n' \mathbf{x}_1 - \frac{1}{2} \alpha_2 y_2 y_n \mathbf{x}_n' \mathbf{x}_2 - \ldots \frac{1}{2} \alpha_{n-1} y_{n-1} y_n \mathbf{x}_n' \mathbf{x}_{n-1}$$

Nilai α_1, α_2, α_3,... α_n diperoleh dengan menghitung sistem persamaan diatas dan didapatkan nilai **w** dan b sebagai berikut

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i} \tag{2.18}$$

$$b = y_i - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i' \mathbf{x_j} \tag{2.19}$$

Pada kasus *non-separabel* beberapa data mungkin tidak bisa dikelompokkan secara benar atau terjadi *misclassification.* Sehingga fungsi objektif maupun kendala (constraint) dimodifikasi dengan mengikutsertakan variabel *slack* $\xi > 0$

fungsi objektif

$$\min \ \frac{1}{2}\mathbf{w}'\mathbf{w} + \mathrm{C}\sum_{\mathrm{i=1}}^{\mathrm{n}} \xi_{\mathrm{i}} \tag{2.20}$$

constraint

$$y_i \left(\mathbf{w}'\mathbf{x_i} + b\right) + \xi_i \geq 1, \ \xi_i \geq 0, \ \mathrm{i} = 1,2,\ldots,\mathrm{n} \tag{2.21}$$

Parameter C menggolongkan bobot yang diberikan kepada kesalahan klasifikasi. Minimalisasi fungsi tujuan dengan kendala menghasilkan kemungkinan margin tertinggi dari *hyperplane* pemisah . Fungsi *lagrange* untuk *primal space* adalah

$$\mathrm{L_p}\left(\mathbf{w}, \mathrm{b}, \alpha, \mu, \xi\right) = \frac{1}{2}\mathbf{w}'\mathbf{w} + \mathrm{C}\sum_{\mathrm{i=1}}^{\mathrm{n}} \xi_{\mathrm{i}} - \sum_{\mathrm{i=1}}^{\mathrm{n}} \alpha_{\mathrm{i}} \left\{ \mathrm{y_i} \left(\mathbf{x_i}'\mathbf{w} + \mathrm{b} - 1 + \xi_{\mathrm{i}} \right) \right\}$$

$$- \sum_{\mathrm{i=1}}^{\mathrm{n}} \mu_{\mathrm{i}} \xi_{\mathrm{i}}$$

constraint

$$y_i \left(\mathbf{w}'\mathbf{x_i} + b\right) + \xi_i \geq 1, \ \xi_i \geq 0, \ \mathrm{i} = 1,2,\ldots,\mathrm{n} \tag{2.22}$$

Bentuk dual dapat diperolah dari turunan fungsi *lagrange* untuk *primal* terhadap **w**, b, dan ξ sebagai berikut

$$\frac{\partial L_p(\mathbf{w}, b, \alpha, \mu, \xi)}{\partial \mathbf{w}} = 0 \;:\; \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \qquad (2.23)$$

$$\frac{\partial L_p(\mathbf{w}, b, \alpha, \mu, \xi)}{\partial b} = 0 \;:\; \sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (2.24)$$

$$\frac{\partial L_p(\mathbf{w}, b, \alpha, \mu, \xi)}{\partial \xi} = 0 \;:\; C - \alpha_i - \mu_i = 0 \qquad (2.25)$$

constraint

$$\alpha_i \geq 0 \qquad (2.26)$$

$$\mu_i \geq 0 \qquad (2.27)$$

$$\{y_i(\mathbf{x}_i'\mathbf{w} + b - 1 + \xi_i)\} = 0 \qquad (2.28)$$

$$\mu_i \xi_i = 0 \qquad (2.29)$$

Dengan mensubtitusikan kedalam bentuk dari *primal* space didapatkan bentuk *dual space* sebagai berikut

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_j \qquad (2.30)$$

constraint

$$0 \leq \alpha_i \leq C \qquad (2.31)$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (2.32)$$

Pada kasus separabel dan kasus - kasus non-separabel perbedaan keduanya hanya terletak dengan adanya penambahan *constraint* $0 \leq \alpha_i \leq C$ pada masalah non-separabel. Pada umumnya masalah dalam domain dunia nyata (*real world problem*) jarang yang bersifat linear separable. Kebanyakan bersifat non linear. SVM dimodifikasi dengan memasukkan fungsi *Kernel* untuk menyelesaikan masalah non linier. Dalam non linear SVM, pertamatama data $\mathbf{x}_i$ dipetakan oleh fungsi $\varphi(\mathbf{x}_i)$ ke ruang vektor yang berdimensi lebih tinggi. Pada ruang vektor yang baru ini, hyperplane yang memisahkan kedua class tersebut dapat dikonstruksikan

$$K\left(\mathbf{x_i},\mathbf{x_j}\right)=\varphi\left(\mathbf{x_i}'\right)\varphi\left(\mathbf{x}_j\right) \qquad (2.33)$$

$K\left(\mathbf{x_i},\mathbf{x_j}\right)$ adalah fungsi pembentuk matriks kernel yang akan digunakan untuk kasus non-linier. $\mathbf{x_i}$ adalah nilai variabel **x** pada pengamatan ke *i* dan $\mathbf{x_j}$ adalah nilai variabel **x** pada pengamatan ke *j*. Beberapa fungsi dari pembentuk matriks *kernel* yang umum digunakan pada SVM adalah

1. Kernel linier
   $K\left(\mathbf{x_i},\mathbf{x_j}\right)=\mathbf{x}_i'\mathbf{x_j}$
2. Kernel Radial Basis function
   $K\left(\mathbf{x_i},\mathbf{x_j}\right)=\exp\left(-\gamma\left\|\mathbf{x_i}-\mathbf{x_j}\right\|^2\right),\gamma>0$
3. Kernel Polinomial
   $K\left(\mathbf{x_i},\mathbf{x_j}\right)=\left(\gamma\mathbf{x}_i'\mathbf{x_j}+r\right)^p,\gamma>0$
4. Kernel Sigmoid
   $K\left(\mathbf{x_i},\mathbf{x_j}\right)=\tanh\left(\gamma\mathbf{x}_i'\mathbf{x_j}+r\right)$

Pemilihan fungsi kernel yang tepat merupakan hal yang sangat penting karena akan menentukan *feature* space dimana fungsi *classifier* akan dicari. Sepanjang fungsi kernelnya sesuai, SVM akan beroperasi secara benar meskipun tidak tahu pemetaan yang digunakan (Santosa, 2007). Menurut Hsu, Chang dan Lin (2004), fungsi kernel yang direkomendasikan untuk diuji pertama kali adalah fungsi kernel RBF karena dapat memetakan hubungan tidak *linear*. Berdasarkan langkah langkah yang telah dijelaskan dalam kasus linier, diperoleh fungsi *hyperplane* sebagai berikut.

$$\begin{aligned} f\left(x\right) &= sign\left(\mathbf{w}'\varphi\left(\mathbf{x}\right)+b\right) \\ &= sign\left(\left(\sum_{i=1}^{n}\alpha_i y_i\varphi\left(\mathbf{x}\right)\right)'\varphi\left(\mathbf{x}\right)+b\right) \\ &= sign\left(\sum_{i=1}^{n}\alpha_i y_i\varphi\left(\mathbf{x}\right)'\varphi\left(\mathbf{x}\right)+b\right) \end{aligned} \qquad (2.34)$$

$$f\left(x\right) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K\left(\mathbf{x}_i, \mathbf{x}_j\right) + b\right) \tag{2.35}$$

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 $f\left(x\right) = -1$ (Gunn, 1998).

2.3 *Entropy Based Fuzzy Membership*

Logika *Fuzzy* merupakan salah satu komponen pembentuk *soft computing*. Dasar logika *fuzzy* adalah teori himpunan *fuzzy*. Pada teori himpunan *fuzzy*, peranan derajat keanggotaan sebagai penentu keberadaan elemen dalam suatu himpunan sangatlah penting. Pada himpunan *fuzzy* nilai keanggotaan bernilai antara 0 sampai 1. Apabila x memiliki nilai keanggotaan *fuzzy* $\mu_A\left(x\right) = 0$, berarti x tidak menjadi anggota himpunan A, demikian pula apabila x memiliki nilai keanggotaan *fuzzy* $\mu_A\left(x\right) = 1$ berarti x menjadi anggota penuh pada himpunan A (Kusumadewi, 2010). Dalam teori informasi, *entropy* adalah rata-rata banyaknya informasi yang terkandung pada masing-masing pesan yang diterima (Shannon, 2001). *Entropy* menggambarkan kepastian tentang sumber informasi. Misalnya, *entropy* yang lebih kecil mengindikasikan bahwa informasi tersebut lebih meyakinkan. Dengan menggunakan *entropy*, kita dapat mengevaluasi kepentingan kelas dari sampel training. Maka kita menetapkan keanggotaan *fuzzy* pada sampel training berdasarkan kepastian kelasnya. Misalkan terdapat sampel pada data training $\left\{\mathbf{x}_i, y_i\right\}_{i=1}^{N}, y_i \in \left\{+1, -1\right\}$. Peluang $\mathbf{x}_i$ masuk kedala kelas positif dan kelas negatif adalah p_{+i} dan p_{-i}. Nilai *Entropy* $\mathbf{x}_i$ adalah sebagai berikut.

$$H = -p_{+i} \ln\left(p_{+i}\right) - p_{-i} \ln\left(p_{-i}\right) \tag{2.36}$$

Sebagai informasi dari sampel dapat direpresentasikan oleh neighbornya, evaluasi probabilitas berdasarkan pada *k nearest neighbors*. Misalnya untuk $\mathbf{x}_i$ pertama pilih *k nearest neighbors*

nya $\left\{\mathbf{x}_{i1,...,}\mathbf{x}_{ik}\right\}$. Kemudian dihitung jumlah kelas positif dan negatif pada k sampel yang terpilih. Jumlah dari kelas positif dan negatif masing-masing dinyatakan dalam num_{+i} dan num_{-i}. Peluang $\mathbf{x}_i$ masuk pada kelas positif dan negatif adalah

$$p_{+i} = \frac{num_{+i}}{k} \tag{2.37}$$

$$p_{-i} = \frac{num_{-i}}{k} \tag{2.38}$$

Entropy untuk kelas negatif adalah $H = \left\{H_{-1}, H_{-2}, ..., H_{-n_-}\right\}$. p_{+i} adalah peluang $\mathbf{x}_i$ masuk pada kelas positif, p_{-i} adalah peluang $\mathbf{x}_i$ masuk pada kelas negatif . Keanggotaan *entropy based fuzzy* untuk sampel negatif dievaluasi sebagai berikut. Pertama bagi negatif sampel kedalam *m* subset misalkan $\left\{Sub_l\right\}_{l=1}^{m}$. Urutkan subset-subset tersebut berdasarakan nilai – nilai entropy. $H_{Sub_1} < H_{Sub_2} < ... < H_{Sub_m}$ Kemudian keanggotaan *fuzzy* untuk sampel pada masing-masing subset dinyatakan sebagai berikut

$$FM_l = 1.0 - \beta * (l-1), l = 1, 2, ..., m \tag{2.39}$$

dimana FM_l adalah keanggotaan *fuzzy* untuk sampel yang didistribusikan pada subset Sub_l $FM_l \in [0,1]$ dan *m* adalah banyaknya subset pada *entropy* yang telah didapatkan dari persamaan (2.32) . Jika nilai β lebih kecil dari 0 maka didapatkan nilai keanggotaan *fuzzy* hanya satu anggota yaitu satu. Jika $\beta * (m-1) \le 1$ akan didapatkan nilai keanggotaan fuzzy yang bernilai negatif sehingga diperoleh pertidaksamaan dari nilai β adalah

$$\begin{aligned} &\beta * (m-1) \le 1 \\ &\beta \ge 0 \end{aligned} \tag{2.40}$$

Pertidaksamaan nilai β dapat disederhanakan $0 < \beta \le \frac{1}{m-1}$ (Fan, Wang, Li, Gao, & Zha, 2017). Nilai Keanggotaan *fuzzy* untuk data training sampel x_i dinyatakan sebagai berikut.

$$S_i = \begin{cases} 1.0 & \textbf{if } y_i = +1 \\ FM_l & \textbf{if } y_i = -1 \end{cases} \tag{2.41}$$

Nilai Keanggotaan *fuzzy* bernilai 1 pada kelas positif atau minoritas sedangkan nilai keanggotaan *fuzzy* pada kelas negatif atau minoritas nilainya akan tergantung dari nilai FM_l dari subset-subset kelas negatif tersebut (Fan, Wang, Li, Gao, & Zha, 2017)

2.4 Klasifikasi dengan *Entropy Fuzzy Support Vector Machine* (EFSVM)

Diberikan training S, dimana $S = \{(x_i, y_i, s_i)\}_{i=1}^{n}$, x_i adalah sampel berukuran n, $y_i \in \{+1, -1\}$ adalah menyatakan kelas (+1 untuk kelas positif dan -1 untuk kelas negatif), dan s_i adalah keanggotaan *entropy based fuzzy* yang ditentukan oleh Persamaan (2.32) . EFSVM menemukan daerah keputusan optimal yang membagi kelas positif dan negatif dengan margin sebesar mungkin. Untuk menemukan keputusan yang optimal, maka perlu untuk menyederhanakan masalah optimasi kuadrat berikut.
fungsi objektif

$$\min\left\{\frac{1}{2}\mathbf{w}^{\mathrm{t}}\mathbf{w} + C\sum_{i=1}^{n} s_i\xi_i\right\} \tag{2.42}$$

constraint

$$y_i\left(\mathbf{w}'\varphi(\mathbf{x_i}) + b\right) \ge 1 - \xi_i, \xi_i \ge 0, i = 1, ..., \mathrm{n} \tag{2.43}$$

Nilai $\mathbf{w}$ adalah vektor pembobot pada daerah keputusan, *b* menyatakan bias, merupakan fungsi nonlinear yang memetakan kedalam ruang *feature high dimensional* di mana daerah keputusan yang lebih baik dapat ditemukan, C adalah parameter regularisasi

yang dipilih terlebih dahulu untuk mengontrol *trade-off* antara margin klasifikasi dan biaya kesalahan klasifikasi. Variabel non-negatif ξ menyatakan variabel *slack* dari pada SVM, sedangkan $s_i\xi$ adalah ukuran error dengan bobot yang berbeda sesuai dengan s_i. Untuk mengatasi optimasi kuadratik, Persamaan (2.42) dapat dinyatakan persamaan langrange sebagai berikut.

$$L_p\left(\mathbf{w},b,\xi,\alpha,\mu\right)=\frac{1}{2}\mathbf{w}'\mathbf{w}+C\sum_{i=1}^{n}s_i\xi_i$$
$$-\sum_{i=1}^{n}\alpha_i\left(y_i\left(\mathbf{w}'\varphi\left(\mathbf{x_i}\right)+b\right)-1+\xi_i\right)-\sum_{i=1}^{n}\mu_i\xi_i$$

constraint

$$y_i\left(\mathbf{w}'\varphi\left(\mathbf{x_i}\right)+b\right)\geq 1-\xi_i,\xi_i\geq 0,i=1,...,\mathrm{n} \qquad (2.44)$$

persamaan diatas dinyatakan ke dual *problem* sebagai berikut

$$L_d\left(\alpha\right)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\varphi\left(\mathbf{x_i}\right)\varphi\left(\mathbf{x_i}\right) \qquad (2.45)$$

constraint

$$\sum_{i=1}^{n}\alpha_i y_i=0 \qquad (2.46)$$

$$0\leq\alpha_i\leq s_i C \qquad (2.47)$$
$$i=1,...,\mathrm{n}$$

Vektor pembobot **w** dapat dihitung sebagai berikut.

$$\mathbf{w}=\sum_{i=1}^{n}\alpha_i y_i\varphi\left(\mathbf{x_i}\right) \qquad (2.48)$$

$$b=y_i-\sum_{i=1}^{n}\alpha_i y_i K\left(\mathbf{x_i},\mathbf{x_j}\right) \qquad (2.49)$$

fungsi keputusan untuk EFSVM dinyatakan sebagai berikut

$$f\left(x\right)=sign\left(\sum_{i=1}^{n}\alpha_i y_i \varphi\left(\mathbf{x}_\mathbf{i}'\right)\varphi\left(\mathbf{x}_j\right)+b\right) \tag{2.50}$$

2.5 Evaluasi Kinerja Metode Klasifikasi

Kinerja klasifikasi menunjukkan kemampuan metode klasifi--kasi untuk memprediksikan kelas suatu data. Hasil dari klasifikasi dapat disusun dalam sebuah *confusion matrix* berikut

Tabel 2.1 *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
Dari *confusion matrix* tersebut dapat dihitung nilai akurasi, sensiti--vity, dan specificity. Selain itu, performa klasifikasi juga diukur melalui ukuran kinerja klasifikasi yang relevan pada data *imbalan--ce*, yaitu G-*mean.*

1. Akurasi

Akurasi menilai keseluruhan efektivitas algoritma dengan memperkirakan probabilitas nilai benar dari label kelas. Nilai Akurasi dinyatakan sebagai berikut.

$$Akurasi=\frac{TN+TP}{TN+TP+FN+FP} \tag{2.51}$$

2. Sensitivity

Sensitivity adalah ukuran keakuratan dari sampel positif. Nilai sensitivity menyatakan berapa banyak sampel kelas positif yang diberi label dengan benar. Nilai *sensitivity* dinyatakan sebagai berikut.

$$Sensitivity=\frac{TP}{TP+FN} \tag{2.52}$$

3. Specificity

Specificity menilai keefektifan algoritma pada kelas negatif. Nilai *specificity* menyatakan berapa banyak sampel kelas negatif yang diberi label dengan benar. Nilai *specificity* dinyatakan sebagai berikut

$$Specificity = \frac{TN}{TN + FP} \tag{2.53}$$

4. AUC (Area Under ROC Curve)

Area dibawah kurva Receiver Operating Characteristic (AUC) adalah indikator ringkasan dari kinerja kurva ROC yang dapat merangkum kinerja pada suatu classifier kedalam metrik tunggal. Tidak seperti kesulitan yang ditemui pada perbandingan dari kurva ROC yang berbeda khususnya pada kasus kurva ROC dengan in--seksi, AUC dapat mengurutkan model berdasarkan kinerja keselu--ruhan, sebagai hasilnya, AUC lebih dipertimbangkan pada penila--ian model (Batista, Ronaldo, & Monard, 2004).

$$AUC = \frac{1}{2}\left(sensitivity + specificity\right) \times 100\% \tag{2.54}$$

2.6 *Stratified* K-*Fold Cross Validation*

Cross validation adalah metode statistik untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua bagian yaitu data training yang digunakan untuk training dan data testing yang digunakan untuk memvalidasi model (Refaeilzadeh dkk, 2008). K-fold *cross validation* akan membagi data ke dalam k subset yang saling bebas yaitu S_1, S_2, ..., S_k dengan jumlah data tiap subset hampir sama, selanjutnya jika satu subset menjadi data testing maka *k-1* subset yang akan menjadi data training (Han dkk, 2006). Data biasanya distratifikasi sebelum dipecah kedalam k-*fold*. Stratifikasi adalah proses penyusunan ulang data untuk memastikan setiap fold merupakan representasi yang baik dari keseluruhan data. Misalnya dalam masalah klasifikasi biner dimana masing-masing kelas terdiri dari 50% data, cara yang terbaik adalah dengan mengatur data sedemikian rupa sehingga dalam setiap fold, setiap kelasnya terdapat sekitar setengah sampel (Refaeilzadeh dkk, 2008).

2.7 Daerah Tertinggal

Daerah tertinggal adalah daerah kabupaten yang wilayah serta masyarakatnya kurang berkembang dibandingkan dengan daerah lain dalam skala nasional (Kementerian Desa, 2015). 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). Nilai Garis Kemiskinan pada suatu daerah tertentu akan berbeda-beda. Nilai Garis Kemiskinan juga akan berbeda pada setiap tahun (Badan Pusat Statistik, 2014).

b. Pengeluaran per kapita per tahun yang 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 metode penghitungan menggunakan metode Rao (Badan Pusat Statistik, 2014)

c. Angka Harapan Hidup

Angka Harapan Hidup secara konsepsi diartikan sebagai rata-rata jumlah tahun hidup yang dapat dijalani oleh seseorang hingga akhir hayatnya. (Badan Pusat Statistik, 2014). Angka Harapan Hidup (AHH) merupakan salah satu indikator yang digunakan untuk menilai derajat kesehatan penduduk, artinya jika angka harapan hidup meningkat, maka derajat kesehatan penduduk juga meningkat serta memperpanjang usia harapan hidupnya. 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), *Mean Years of Schooling (MYS)* didefinisikan sebagai jumlah tahun yang digunakan oleh penduduk dalam menjalani pendidikan formal. Cakupan 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. Penghitungan RLS pada usia 25 tahun ke atas juga mengikuti standar internasional yang digunakan oleh UNDP (Badan Pusat Statistik, 2014).

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. Hasilnya jika semakin besar akan semakin baik atau menggambarkan kondisi/tingkat kesejahteraan yang lebih baik (Badan Pusat Statistik, 2014).

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 (Badan Pusat Statistik, 2014). Jenis permukaan jalan merupakan salah satu dari 27 indikator yang ditetapkan oleh kementerian desa.

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. Kementerian Desa menetapkan batas IDM kurang dari sama dengan 0,599 sebagai desa tertinggal (Kementerian Desa, 2015)

h. Indeks Kapasitas Fiskal

Indeks Kapasitas Fiskal indeks yang menunjukan kemampu--an 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 lain yang penggunaannya dibatasi untuk membiayai pengeluaran tertentu) untuk membiayai tugas pemerin--tahan setelah dikurangi belanja pegawai dan dikaitkan dengan jumlah penduduk miskin. Potensi fiskal dapat diukur dengan kom--ponen-komponennya terdiri dari Pendapatan Asli Daerah, Dana Bagi Hasil, Dana Alokasi Umum dan pendapatan daerah lainnya yang sah, dimana dana bagi hasil dan dana alokasi umum termasuk dalam dana transfer atau dana perimbangan (Direktorat Jenderal Perimbangan Keuangan Kementerian Keuangan, 2013). Semakin tinggi nilai Indeks Kapasitas Fiskal maka semakin banyak nilai hibah yang ditransfer dari pemerintahan pusat ke pemerintahan daerah tersebut.

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan dalam penelitian ini diperoleh dari Statistik Potensi Daerah 2014, Profil Kesehatan Provinsi Jawa Timur 2014 serta Realisasi Kemampuan Keuangan Daerah Tahun 2014. Unit penelitian ini adalah kabupaten

3.2 Variabel Penelitian

Variabel yang digunakan pada penelitian ini adalah 23 Indikator yang telah ditetapkan pemerintah untuk menilai daerah tertinggal dapat dilihat pada Tabel 1 berikut ini

Tabel 3.1 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

Tabel 3.2 Variabel Penelitian *(Lanjutan*)

Variabel	Keterangan	Skala	Satuan
X_{18}	persentase desa banjir	Rasio	persentase
X_{19}	persentase desa bencana lainnya	Rasio	persentase
X_{20}	persentase desa di kawasan hutan lindung	Rasio	persentase
X_{21}	persentase desa konflik satu tahun terakhir	Rasio	persentase
X_{22}	Indeks Kapasitas Fiskal	Rasio	Indeks
X_{23}	Indeks Desa Membangun	Rasio	Indeks

3.3 Struktur Data

Berikut adalah struktur data dari penelitian ini.

Tabel 3.3 Struktur Data Penelitian

Kabupaten	X_1	X_2	…	X_{27}	Y
Kabupaten 1	X_{11}	X_{12}	…	X_{127}	Tdk Tertinggal
Kabupaten 2	X_{21}	X_{22}	…	X_{227}	Tdk Tertinggal
…	…	…	…	…	…
Kabupaten 25	X_{251}	X_{252}	…	X_{2527}	Tdk Tertinggal
kabupaten 26	X_{261}	X_{262}	…	X_{2627}	Tertinggal
…	…	…	…	…	…
Kabupaten 29	X_{291}	X_{292}	…	X_{2927}	Tertinggal

3.4 Metode Analisis Data

Setelah diperoleh data dari data sekunder selanjutnya akan dilakukan analisis data. 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 (2.1)
    b. Menghitung nilai *Information Gain* menggunakan persamaan (2.3)
    c. Menghitung nilai *Symmetrical uncertainty* mengguakan persamaan (2.4)
3. Membagi data menjadi data *testing* dan data *training*.
4. Klasifikasi Data dengan *Support Vector Machine* (SVM).

a. Menentukan fungsi kernel yang digunakan pemodelan menggunakan persamaan
b. Menentukan nilai *initial value* parameter kernel dan parameter *cost* untuk optimasi
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, dipilih k =7 (ladayya, 2018)
   b. Menghitung jumlah sampel positif dan negatif pada (k-NN) .
   c. Menghitung peluang dari sampel positif dan negatif menggunakan persamaan (2.37) dan (2.38)
   d. Menghitung nilai entropy dari dari masing-masing data *training* menggunakan persamaan (2.36)
   e. Pada sampe training $\mathbf{x_i}$ dari kelas $y_i = -1$ dilakukan pengelompokan kedalam *m* subset. Menghitung nilai pada batas atas (*thrUp)* dan nilai pada batas bawah (*thrLow)* dengan rumus sebagai berikut

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

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

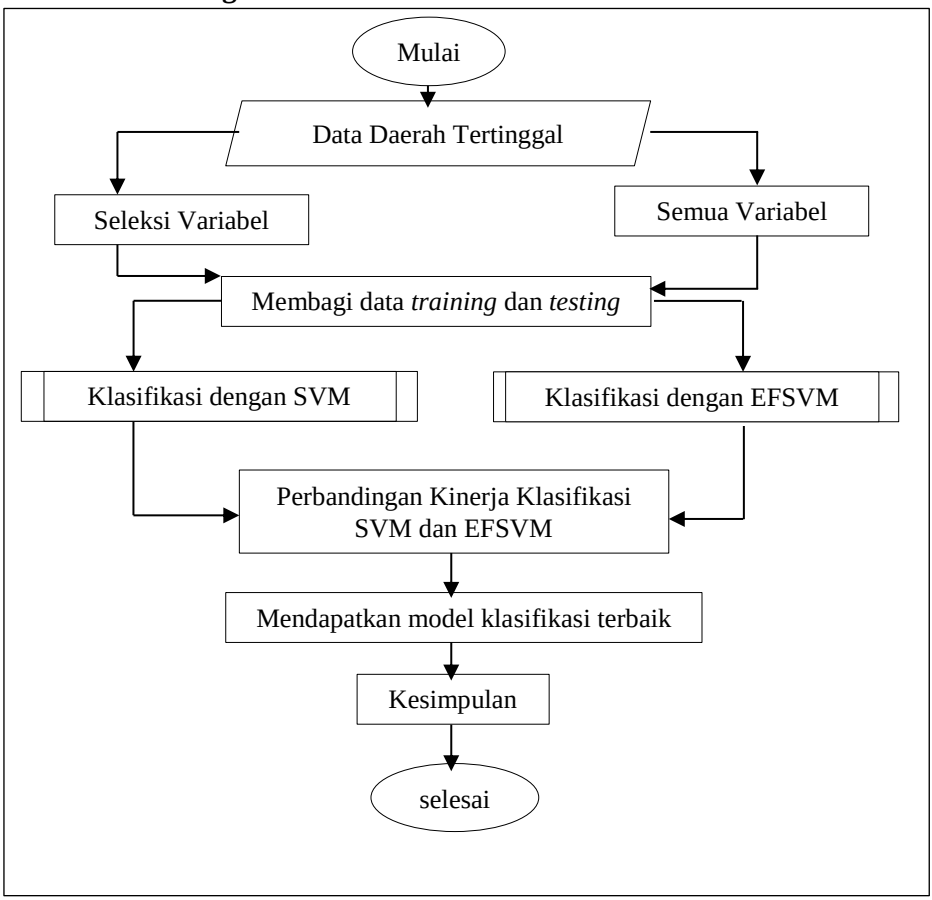
   f. Setelah didapatkan nilai *thrUp* dan *thrLow* pada masing-masing subset, sample *training* dari kelas $y_i = -1$ dikelompokan berdasarkan nilai *entropy*
   g. Menghitung nilai *fuzzy membership* $\left(FM_l\right)$ untuk masing-masing subset menggunakan persamaan (2.39)
   h. Menetapkan entropy based fuzzy membership untuk semua *training* sampel x_i menggunakan persamaan

(2.41). Sampel dari kelas $y_i = +1$ diberi niai $s_i = 1$, sedangkan untuk sampel dari kelas $y_i = +1$ memiliki nilai s_i yang sesuai dengan telah dihitung pada langkah f.

i. Optimasi Parameter kernel dan parameter cost terbaik berdasarkan nilai AUC

6. Membandingkan Kinerja hasil klasifikasi dengan metode SVM dan EFSVM

3.5 Diagram Alir



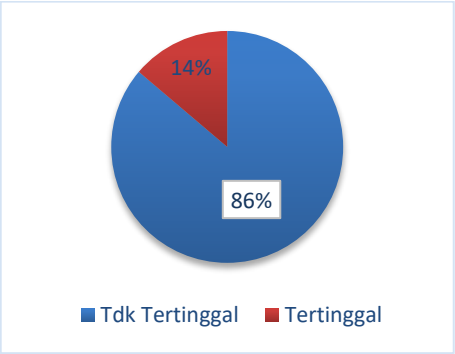
Gambar 3.1 Diagram Alir

BAB IV
ANALISIS DAN PEMBAHASAN

Pada subbab ini membahas mengenai hasil analisis data untuk menjawab permasalahan dari rumusan masalah yang diambil. Analisis yang dilakukan pertama untuk mengetahui karakteristik kabupaten-kabupaten di Provinsi Jawa Timur berdasarkan indikator-indikator daerah tertinggal

4.1 Karakteristik Data Daerah Tertinggal Berdasarkan Indi--kator Daerah tertinggal

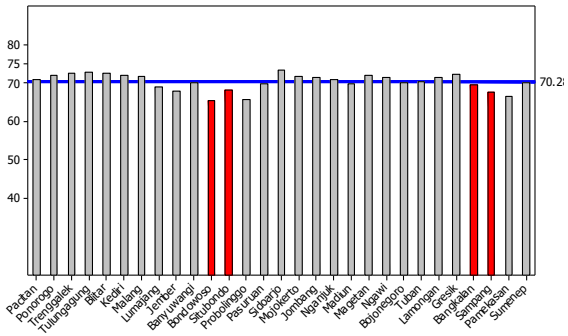
Terdapat 29 Kabupaten di Provinsi Jawa Timur. Sebanyak 4 kabupaten atau sekitar 14% dari total kabupaten di Provinsi Jawa Timur yang berstatus sebagai daerah tertinggal.



Gambar 4.1 Proporsi Kabupaten Tertinggal

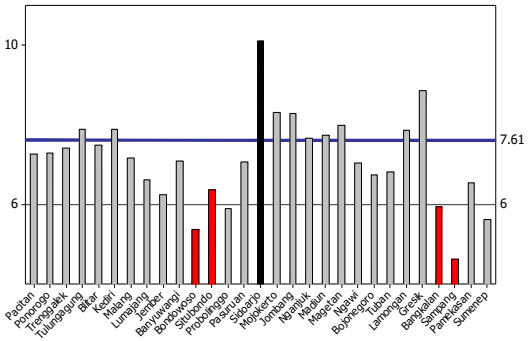
Gambar 4.1 menunjukan bahwa dapat disimpulkan data kabupaten tertinggal di Provinsi Jawa Timur adalah data *imbalance.* Rasio kelas mayoritas dibandingkan dengan kelas minoritas adalah 25:4 yang menunjukkan kategori *low imbalance*. Berdasarkan gambar 4.2 dibawah menunjukkan kabupaten-kabupaten yang tertinggal di Provinsi Jawa Timur memiliki angka harapan hidup dibawah angka harapan hidup Provinsi Jawa Timur angka harapan hidup nasional dan yaitu 70,2 tahun dan 70,9 tahun. Rata-rata angka harapan hidup di Provinsi Jawa Timur adalah 70,28 tahun . Hal ini berarti

bahwa setiap bayi yang lahir hidup di Jawa Timur mempunyai harapan untuk bertahan hidup sampai usia 70,28 tahun Banyak hal yang melatarbelakangi angka harapan hidup di suatu daerah pada posisi tinggi atau rendah. Salah satu diantaranya adalah keberhasilan program kesehatan pemerintah dan gaya hidup sehat penduduk pada wilayah tersebut



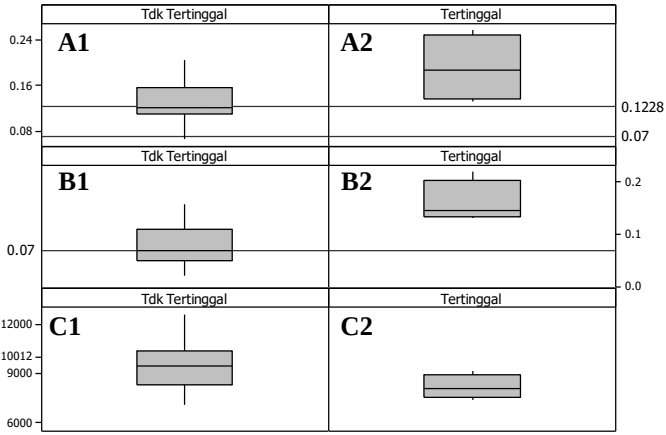
Gambar 4.2 Angka harapan hidup Provinsi Jawa Timur

Rata-rata lama sekolah di Provinsi Jawa Timur pada masing-masing kabupaten dapat dilihat pada gambar 4.3 dengan warna merah adalah kabupaten tertinggal. Semua kabupaten– kabupaten tertinggal di Provinsi Jawa Timur memiliki rata-rata lama sekolah kurang dari rata-rata provinsi yaitu sebesar 7,61 tahun. 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 sepanjang kurang dari 6 tahun atau tidak tamat sekolah dasar. Kabupaten Sampang memiliki angka rata-rata lama sekolah paling rendah diantara kabupaten tertinggal yang lain. Kabupaten Sodoarjo memiliki rata-rata lama sekolah sebesar 10,12 tahun yang merupakan kabupaten dengan nilai rata-rata lama sekolah tertinggi. Kabupaten Sidoarjo juga merupakan satu satunya kabupaten yang rata-rata penduduknya bersekolah hingga tamat SMP. Rata-rata lama sekolah di Provinsi Jawa Timur terlihat sebesar 7,61 tahun.



Gambar 4.3 Rata-rata lama sekolah Provinsi Jawa Timur

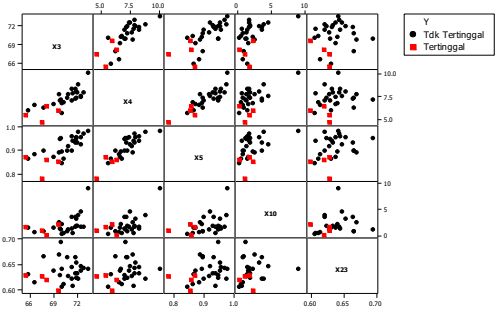
Gambar 4.4 dibawah adalah bentuk boxplot berturut-turut A, B, dan C adalah boxplot dari persentase penduduk miskin, angka melek huruf, dan konsumsi per kapita yang disesuaikan dari kabupaten tertinggal dan tidak tertinggal.



Gambar 4.4 Boxplot persentase penduduk miskin, angka tidak melek huruf, dan konsumsi per kapita

Nilai persentase dari penduduk miskin di Jawa Timur adalah sebesar 0,1228 atau 12,28%. Pada Boxplot angka tidak melek huruf, 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. Boxplot konsumsi perkapita 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. Hal ini menunjukan rata-rata daya beli penduduk di kabupaten tertinggal cenderung untuk konsumsi komoditas makanan dan non makanan kurang dari Rp 10.012. Penyebaran data dari beberapa variabel terdapat pada gambar 4.5. Terdapat pola data yang menyebar sehingga sehingga sulit dipsahkan fungsi pemisah linier sehingga diperlukan bantuan kernel untuk mendapatkan fungsi pemisah yang optimal



Gambar 4.5 Pola persebaran data beberapa variabel

4.2 *Fast Correlation Based Filter untuk Seleksi Variabel*

Feature *Correlation Based Filter (FCBF)* adalah seleksi fitur digunakan pada penelitian ini. Seleksi fitur adalah salah satu tahapan praproses klasifikasi. Seleksi fitur dilakukan dengan cara memilih fitur-fitur yang relevan yang mempengaruhi hasil klasifikasi. Seleksi fitur digunakan untuk mengurangi dimensi data dan fitur-fitur yang tidak relevan. Hasil dari seleksi fitur dengan metode FCBF adalah sebagi berikut. *Variabel yang terseleki Feature Correlation based filter (FCBF)* adalah variabel yang memiliki *information Gain* paling besar terhadap kategori ketertinggalan kabupaten

Tabel 4.1 Variabel terpilih hasil FCBF

No	Variabel	Nama Variabel	*Information Gain*
1	X_5	Angka melek huruf	0,341
2	X_4	Rata-rata lama sekolah	0,341
3	X_9	Persentase rumah tangga pengguna listrik	0,341
4	X_3	Angka Harapan Hidup	0,303

Berdasarkan tabel 4.1, variabel yang terpilih dari 23 variabel adalah variabel angka harapan hidup, rata-rata lama sekolah, angka melek huruf. Ketiga variabel memiliki nilai *Information Gain* yang sama. Nilai *Information Gain* diperoleh dari nilai *entropy* sebelum pemisahan dikurangi dengan nilai *entropy* setelah pemisahan.

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

Sebanyak 29 data dibagi menjadi data *training* dan data *testing* menggunakan 4 fold dengan perbandingan. Masing-masing anggota Fold adalah sebagai berikut.

Tabel 4.2 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

Pada tabel 4.2 diatas, anggota yang bercetak tebal adalah kabupaten tertinggal. Masing masing fold akan memiliki 1 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. Pembagian data menjadi 4 fold karena jumlah kelas minoritas yaitu kabupaten-kabupaten tertinggal berjumlah sebanyak empat kabupaten. Pembagian menjadi empat fold maka otomatis perbandingan data *training* dan *testing* adalah 75:25.

4.4 Klasifikasi dengan *Support Vector Machine*

Dalam penelitian ini, Klasikasi dengan SVM dilakukan pada semua variabel dan variabel yang telah terseleksi pada *Feature Correlation Based Filter* yaitu empat variabel terseleksi. Nilai *cost* yang dipilih pada *range* 2^{-6}, 2^{-5}, 2^{-4} ,…2^{1}, 2^{2},… 2^{5}, 2^{6} (Fan dkk, 2017) dan nilai *gamma* yang dipilih pada *range* yang sama sehingga ada terdapat kombinasi 121 kombinasi nilai *cost* dan *gamma*. Hasil dari AUC dan akurasi dari klasifikasi dengan *Support Vector Machine* dapat dilihat pada tabel 4.4

Tabel 4.3 Hasil AUC dan Akurasi data *training* kernel RBF (%)

Cost	γ	Variabel terseleksi		Semua Variabel	
		AUC	Akurasi	AUC	Akurasi
2^{-6}	2^{-6}	50	86,20	50	86,20
	2^{-5}	50	86,20	50	86,20
	2^{-4}	50	86,20	50	86,20
	⋮	⋮	⋮	⋮	⋮
	2^{6}	50	86,20	50	86,20
2^{-5}	2^{-6}	50	86,20	50	86,20
	2^{-5}	50	86,20	50	86,20
	2^{-4}	50	86,20	50	86,20
	⋮	⋮	⋮	⋮	⋮
	2^{6}	50	86,20	50	86,20
⋮	⋮	⋮	⋮	⋮	⋮
2^{6}	2^{-6}	82,67	94,23	**100**	**100**
	2^{-5}	83,33	95,40	100	100
	2^{-4}	95,83	98,86	100	100
	⋮	⋮	⋮	⋮	⋮
	2^{3}	**100**	**100**	100	100
	⋮	⋮	⋮	⋮	⋮
	2^{6}	100	100	100	100

Tabel 4.3 menunjukkan nilai AUC yang hampir sama pada nilai *cost* sebesar 2^{-6} dan 2^{-5} yaitu sebesar 50% . Nilai AUC tertinggi yang didapatkan dari data *training* SVM variabel terseleksi adalah sebesar 100% dan akurasi tertinggi adalah sebesar 100% yang terdapat pada nilai *cost* sebesar 2^{6} dan nilai *gamma* sebesar 2^{3} . Nilai

AUC tertinggi SVM semua variabel pada data *training* adalah sebesar 100% berada pada nilai *cost* sebesar 2^6 dengan nilai *gamma* sebesar 2^{-6}, 2^{-5}, 2^{-4},… 2^5, 2^6. *Cost* dan *gamma* yang terbaik adalah 2^6 dan 2^{-3}. Hasil dari nilai AUC tertinggi dan akurasi data *training* dengan menggunakan kernel linier dan sigmoid adalah sebagai berikut.

Tabel 4.4 Hasil rata-rata AUC dan Akurasi dari data *training* kernel linier dan kernel sigmoid (%)

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

Pemilihan nilai *cost* dan gamma optimum berdasarkan nilai AUC karena data yang digunakan adalah data *imbalance*. Hasil nilai *cost* dan *gamma* dari data *training* digunakan untuk model data *testing*. Hasil ketepatan klasifikasi pada data *testing* adalah sebagai berikut.

Tabel 4.5 Hasil ketepatan klasifikasi pada 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

Berdasarkan tabel 4.5, Metode SVM dengan menggunakan variabel terseleksi dan semua variabel yang terdapat pada kernel RBF menghasilkan nilai AUC data *testing* sebesar 50% dan nilai akurasi sebesar 86,16% , nilai *specificity* yang tinggi yaitu sebesar 100% namun nilai *sensitivity* yang dihasilkan sebesar 0% . Hal ini menunjukan model hanya mampu mengklasifikasikan semua kabupaten yang tidak tertinggal dengan benar namun tidak satu pun kabupaten tertinggal diklasifikasikan dengan benar. Metode SVM

dengan menggunakan kernel linier menghasilkan nilai AUC yang lebih rendah dari kernel RBF sedangkan kernel sigmoid menghasilkan nilai AUC yang lebih tinggi jika menggunakan semua variabel. Model yang didapatkan adalah sebagai berikut

Tabel 4.6 Model SVM dari tiga jenis kernel

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

4.5 Penentuan Nilai *Entropy* berdasarkan *fuzzzy*

Langkah pertama dalam klasifikasi dengan EFSVM adalah menentukan keanggotaan *entropy* based *fuzzy* Keanggotaan *entropy* based *fuzzy* ini nantinya akan digunakan sebagai input dalam klasifikasi dengan EFSVM yang berguna untuk menjamin kepentingan dari kelas positif (minoritas) dan mengurangi adanya bias yang disebabkan oleh kelas negatif (mayoritas). Pada keanggotaan *entropy* based *fuzzy* terlebih dahulu didapatkan nilai *entropy* dari masing-masing sampel. Nilai *entropy* ditentukan berdasarkan nilai kedekatan satu sampel dengan sampel yang lain menggunakan k-nearest neighbors . K-nearest neighbor yang dengan dipilih adalah nilai k sebesar 7. Jarak yang digunakan dalam penelitian ini adalah jarak *Euclidean* .

Tabel 4.7 Matriks jarak sampel x_i

	1	2	3	…	29
1	0	7,863725	6,693651	…	9,470636
2	7,863725	0	5,321984	…	6,587308
3	6,693651	5,321984	0	…	8,64671
…	…	…	…		…
29	9,470636	6,587308	8,64671	…	0

Setelah didapatkan matriks jarak, kemudian didapatkan 7 sampel dengan jarak terdekat dari masing-masing sampel. Kemudian dihitung peluang sampel tersebut masuk pada kelas positif dan negatif sehingga dari hasil tersebut dapat digunakan untuk menghitung nilai *entropy* untuk masing-masing sampel. Tabel 4.8 menunjukkan nilai *entropy* untuk masing-masing sampel

Tabel 4.8 nilai *entropy* untuk masing-masing sampel

Data ke	Entropy
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0,410116
10	0
11	0,682908
12	0,59827
13	0,682908
,,	…
28	0,59827
29	0,59827

Nilai *entropy* yang didapatkan pada Tabel 4,8 digunakan untuk mengelompokkan data pada kelas negatif atau kabupaten tidak tertinggal kedalam subset-subset, Sampel dalam satu subset akan memiliki nilia keanggotaan *fuzzy* yang sama.

Tabel 4.9 Nilai Batas atas dan Batas Bawah Subset

Subset ke	Batas Bawah	Batas Atas
1	0	0,136581621
2	0,136581621	0,273163242
3	0,273163242	0,409744863
4	0,409744863	0,546326484
5	0,546326484	0,682908105

Sampel pada kelas negatif dengan nilai *entropy* antara 0 sampai kurang dari 0,136581621 akan masuk pada subset 1, selanjutnya data dengan nilai *entropy* sebesar 0,136581621 sampai kurang dari 0,273163242 akan masuk pada subset 2 dan seterusnya sampai subset ke-5, Distribusi dari sampel kelas negatif untuk masing-masing subset ditunjukkan pada Tabel 4,10

Tabel 4.10 Distribusi Sampel Kelas Negatif Pada tiap Subset

Subset	Anggota
1	1 , 2 , 3 , 4 , 5 , 6 , 7 , 8, 10, 14, 15, 16, 17 18, 19, 20, 21, 22, 23, 24, 25
2	…
3	…
4	9
5	13, 28, 29

Berdasarkan Tabel 4.10 dapat dilihat bahwa untuk subset ke 2 dan 3 tidak terdapat sampel dari kelas negatif didalamnya sedangkan untuk subset ke-1 terdapat 21 sampel negatif, subset ke-4 terdapat 1 sampel dari kelas negatif, dan pada subset ke-5 terdapat 3 sampel dari kelas negatif, Nilai keanggotaan *fuzzy* untuk sampel (FM_l) pada masing-masing subset subset ditunjukkan pada Tabel 4.11

Tabel 4.11 Nilai FM_l untuk Masing-masing Subset

Subset ke	FM
1	1
2	0,95

Tabel 4.11 Nilai FM_l untuk Masing-masing Subset

Subset ke	**FM**
3	0,9
4	0,85
5	0,8

Sehingga akan didapatkan keanggotaan *entropy based fuzzy* untuk masing-masing sampel pada Tabel 4.12

Tabel 4.12 Keanggotaan *Entropy Based Fuzzy*

Data ke	Si
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	0,85
10	1
11	**1**
12	**1**
13	0,8
…	…
26	**1**
27	**1**
28	0,8
29	0,8

Nilai s_i pada Tabel 4.12 selanjutnya digunakan sebagai input dalam klasifikasi menggunakan *Entropy based Fuzzy Support Vector Machine* (EFSVM). Nilai dari keanggotaan pada kelas minoritas/kelas positif yaitu pada data ke 11,12,26, dan 27, bernilai 1. Nilai keanggotaan *fuzzy* pada kelas negatif atau kelas mayoritas bernilai sesuai dengan subset anggota yang telah ditentukan pada tabel 4.11

4.6 Klasifikasi dengan *Entropy Fuzzy Support Vector Machine* (EFSVM)

Pada klasifikasi dengan EFSVM keanggotaan *entropy* based *fuzzy* pada persamaan didapatkan terlebih dahulu. Keanggotaan *entropy* based *fuzzy* tersebut digunakan sebagai input pada klasifikasi EFSVM dimana nilainya dikalikan dengan parameter cost (C). Parameter *cost* yang dipilih pada *range* 2^{-6}, 2^{-5}, 2^{-4} ,…2^{-1}, 2^{1}, 2^{2},… 2^{6} , Parameter *gamma* yang dipilih pada *range* 2^{-6}, 2^{-5}, 2^{-4} ,…2^{-1}, 2^{1}, 2^{2},… 2^{6}

Tabel 4.13 AUC dan Akurasi EFSVM data *training* kernel RBF (%)

Cost	γ	Variabel terseleksi		Semua Variabel	
		AUC	Akurasi	AUC	Akurasi
2^{-6}	2^{-6}	58,33	88,58	70,83	92,05
	2^{-5}	58,33	88,58	66,67	90,91
	2^{-4}	50	86,20	70,83	92,05
	⋮	⋮	⋮	⋮	⋮
	2^{6}	50	97,72	95,18	97,72
2^{-5}	2^{-6}	69,48	89,67	95,18	97,72
	2^{-5}	64,58	87,34	94,48	96,54
	2^{-4}	73,64	90,08	**97,98**	**96,54**
	⋮	⋮	⋮	⋮	⋮
	2^{6}	77,15	90,08	94,66	90,08
	⋮	⋮	⋮	⋮	⋮
2^{-4}	2^{-5}	**93,31**	**88,47**	95,97	93,07
	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
2^{6}	2^{-6}	82,67	94,23	86	75,86
	2^{-5}	83,33	95,40	86,69	77,07
	2^{-4}	95,83	98,86	88,01	79,33
	⋮	⋮	⋮	⋮	⋮
	2^{6}	100	100	86,03	75,92

Tabel 4.13 menunjukan nilai *cost* dan nilai *gamma* optimal dari EFSVM kernel RBF dengan menggunakan seleksi variabel adalah 2^{-4} dan 2^{-5} yang menghasilkan nilai AUC tertinggi yaitu sebesar 93,31% sedangkan EFSVM dengan menggunakan seleksi variabel kernel RBF nilai cost dan *gamma* yang optimal adalah 2^{-5} dan 2^{-4}. Hasil nilai AUC dan akurasi dari data *training* dari kernel linier dan sigmoid adalah sebagai berikut.

Tabel 4.14 AUC dan Akurasi EFSVM data *training* kernel linier dan sigmoid (%)

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

Nilai AUC tertinggi pada data *training* adalah EFSVM dengan semua variabel kernel sigmoid yaitu sebesar 98% dan nilai AUC terendah terdapat pada EFSVM seleksi variabel kernel linier yaitu sebesar 92,65%. Hasil nilai AUC dan akurasi pada data *testing* dari kernel RBF, linier, dan sigmoid adalah sebagai berikut,

Tabel 4.15 Hasil ketepatan klasifikasi EFSVM pada data *testing* (%)

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 4.15 menunjukan nilai AUC sebesar 92,26% yang lebih tinggi dari hasil AUC pada data testing dengan menggunakan SVM, EFSVM pada kernel linier dan RBF menghasilkan nilai AUC yang sama . Nilai AUC terendah diperoleh oleh EFSVM linier dan sigmoid. Nilai *Sensitivity* sebesar 100% menunjukan model mampu mengklasifikasikan semua kabupaten tertinggal dengan benar. Model yang diperoleh adalah sebagai berikut.

Tabel 4.16 Model ESVM dari tiga jenis kernel

Kernel	Variabel	Model
RBF	Semua	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \exp\left(-2^{4} \left\|x_i - x_j\right\|^2 - 0,71429\right)$
Linier	Semua	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \left(x_i' x_j - 0,27273\right)$
Sigmoid	Semua	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \left(\tanh\left(2^{-3} \mathbf{x}_i' \mathbf{x}_j + 0\right) - 0,2\right)$
RBF	Seleksi	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \exp\left(-2^{-5} \left\|x_i - x_j\right\|^2 - 0,27273\right)$
Linier	Seleksi	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \left(x_i' x_j - 0,11111\right)$
Sigmoid	Seleksi	$f(\mathbf{x}) = \sum_{i=1}^{29}\sum_{i=1}^{29} \alpha_i y_i \left(\tanh\left(2^{6} \mathbf{x}_i' \mathbf{x}_j + 0\right) - 0,2\right)$

4.7 Evaluasi Kinerja Metode Klasifikasi

Metode *Support Vecttor Machine* dan *Entropy Fuzzy Support Vector Machine* (EFSVM) dengan *selected* variabel dan seluruh variabel menghasilkan akurasi, *sensitivity*, *specificity*, dan AUC dari parameter *cost* dan *gamma* yang optimum dan model terbaik dari masing-masing metode adalah sebagai berikut.

Tabel 4.17 Evaluasi Kinerja Metode Klasifikasi (%)

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

Tabel 4.17 menunjukan Nilai AUC pada metode ESVM lebih tinggi dari SVM. Metode EFSVM yang terbaik adalah dengan menggunakan seleksi variabel. Hal ini sesuai dengan penelitian sebelumnya yang dilakukan oleh Fan dkk pada tahun 2017. Pemilihan jenis kernel sigmoid pada metode SVM mampu meningkatkan *sensitivity* menjadi 50% . Nilai *Sensitivity* sebesar 100 menunjukan model EFSVM mampu mengklasifikasikan semua kabupaten tertinggal dengan benar.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil analisis dan pembahasan dapat diambil kesimpulan sebagai berikut,

1. Hasil model terbaik klasifikasi data *imbalance* kabupaten tertinggal di Provinsi Jawa Timur adalah dengan metode *Entropy Based Fuzzy Support Vector Machine* dari variabel yang telah terseleksi dengan kernel *Radial Basis Function* dan kernel *linier* ,
2. Seleksi variabel mampu meningkatkan nilai AUC pada metode *Entropy Based Fuzzy Support Vector Machine* namun tidak meningkatkan AUC pada metode *Support Vector Machine*

5.2 Saran

Berdasarkan hasil penelitian diatas, saran yang diberikan adalah

1. Perlu dilakukan perbandingan dengan metode EFSVM dengan nilai β, m, k dan posisi kabupaten tertinggal pada masing-masing fold yang berbeda-beda
2. Pada penelitian selanjutnya, diharapkan menggunakan data dengan jumlah data yang lebih banyak yang memiliki perbandingan kelas minoritas dan mayoritas yang lebih besar

**********Halaman ini sengaja dikosongkan**********

DAFTAR PUSTAKA

Akbani, R., Kwek, S., & Japkowicz, N. (2005). *Support Vector Machines to Imbalanced.* San Antonio: 1Department of Computer Science, University of Texas .

Alonso, A. N. (2015). *Feature Selection for High Dimensional Data. Artificial Intelligence: Fondations, Theory, and Algorthms.* Springel International Publishing Switzerland.

Badan Pusat Statistik. (2014). *Statistik Potensi Desa Jawa Timur.* Jakarta: Badan Pusat Statistik.

Batista, G., Ronaldo, C., & Monard, M. (2004). A Study of the Behavior of Several Methods fo Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter*, 20-29.

Canedo, V. B. (2014). A Review of Microarray Datasets and Applied Feature Selection Methods. *information Science*, 111-135.

Dinas Kesehatan Provinsi Jawa Timur. (2014). *Profil Kesehatan Provinsi Jawa Timur.* Surabaya: Dinas Kesehatan Provinsi Jawa Timur.

Direktorat Perimbangan Keuangan dari Kementerian Keuangan. (2013). *Affirmative Policy pada Percepatan Pembangunan Infrastruktur Daerah Untuk Meningkatkan Kesejahteraan Rakyat.* Jakarta: Kementerian Keuangan.

Fan, Q., Wang, Z., Li, D., Gao, D., & Zha, H. (2017). Entropy Based Fuzzy Support Vector Machine for Imbalance Datasets. *Knowledge Based System*, 87-99.

Gunn, S. (1998). *Support Vector Machine for Clasification and Regression* . Southamton: University of Southamton.

Han, J., & M, J. P. (2006). *Data Mining: Concept and Techniques.* San Fransisco: Morgan Kaufmaan.

Hsu, W. C., Chang, C. C., & Lin, C. J. (2004). *A Practical Guide to AnalySupport Vector Machine.* Taipei: Departement of Computer Science National Taiwan University.

Kementerian Desa, P. D. (2015). *Rencana Strategis Direktorat Jenderal Pembangunan Daerah Tertinggal 2015-209.*

Jakarta: Kementerian Desa, Pembangunan Daerah Tertinggal dan Transmigrasi.

Kusumadewi , S., & Purnomo, H. (2010). *Aplikasi Logika Fuzzy untuk Pendukung Keputusan.* Yogyakarta: Graha Ilmu.

Lisna, V., Sinaga, B., Firdaus, M., & Sutomo, S. (2013). Dampak Kapasitas Fiskal terhadap Penurunan Kemiskinan: Suatu Analisis Simulasi Kebijakan. *Jurnal Ekonomi dan Pembangunan Indonesia*, 1-26.

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

Nash, M., & Bradford, D. (2001). *Parametric and Non Parametric Logistic Regression for Prediction of Precense/ Absence of an Amphibian.* Las Vegas: US Environmental Protection Agency Office of Research and Development National Exposure Research Laboratory Environmental Sciences.

Oktora, S., Darwis, S., & Setyawanto, G. R. (2015). *Pemodelan dan Pengklasifikasian Kabupaten Tertinggal di Indonesia dengan Pendekatan Multivariate Adaptive Regression Splines (MARS).* Bandung: Jurusan Statistika, Universitas Padjadjaran.

Pamuji, Y. S. (2015). *Klasifikasi Penerima Program Beras Miskin.* Semarang: Universitas Diponegoro.

Purwandari, T., & Hidayat, Y. (2017). *Pemodelan Ketertinggalan Suatu Daerah* . Bandung: Jurusan Statistika, Universitas Padjadjaran.

Refaeilzadeh, P., Tang, L., & Liu, H. (2007). *Cross Validation.* Arizona: Arizona State University.

Santosa, B. (2007). *Teknik Pemanfaatan Data untuk Keperluan bisnis.* Surabaya: Graha Ilmu .

Shannon, C. (2001). A Mathematical Theory of Communication. *The Bell System Technical*, 379-423.

Sulasih, D. (2016). *Rare Event Weighted Logistic Regression untuk Klasifikasi Imbalance Data.* Surabaya: Jurusan Statistika Institut Teknologi Sepuluh Nopember.

Sun, Y. W. (2009). Clasification Of Imbalance Data. *International Journal of Pattern Recognition and Artificial Intelligence*, 687-719.

Suryowati, E. (2016, Februari 5). *Perekonomian Indonesia masih dominan di Pulau Jawa dan Sumatera*. Retrieved from KOMPAS:ekonomi.kompas.com/read/2016/02/05/51449126/Perekonomian.RI.Masih.Dominan.di.Jawa.dan.Sumatera

**********Halaman ini sengaja dikosongkan**********

LAMPIRAN

Lampiran 1
Data Indikator-indikator daerah tertinggal

No	Kabupaten	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	Pacitan	-1	0,1618	7656	70,74683	7,27	0,1043	160	9
2	Ponorogo	-1	0,1153	8383	71,8828	7,28	0,0498	276	31
3	Trenggalek	-1	0,131	8417	72,51494	7,41	0,046	137	17
4	Tulungagung	-1	0,0875	9505	72,87546	7,89	0,0303	263	4
5	Blitar	-1	0,1022	9245	72,49518	7,49	0,0779	243	5
6	Kediri	-1	0,1277	9633	72,04074	7,88	0,0684	338	6
7	Malang	-1	0,1107	8817	71,77889	7,17	0,0673	379	9
8	Lumajang	-1	0,1175	7895	69,07239	6,62	0,1297	189	15
9	Jember	-1	0,1128	8227	67,79799	6,24	0,1023	181	60
10	Banyuwangi	-1	0,0929	10379	69,92687	7,1	0,0501	202	12
11	Bondowoso	1	0,1476	9176	65,42726	5,36	0,1309	160	59
12	Situbondo	1	0,1315	8383	68,08448	6,36	0,1423	132	4
13	Probolinggo	-1	0,2044	9877	65,74814	5,9	0,1359	317	13
14	Pasuruan	-1	0,1086	8293	69,82783	7,06	0,052	325	39
15	Sidoarjo	-1	0,064	12632	73,42845	10,11	0,0195	347	3
16	Mojokerto	-1	0,1056	11208	71,76401	8,32	0,0591	286	17
17	Jombang	-1	0,108	9709	71,37397	8,28	0,0434	269	33
18	Nganjuk	-1	0,1314	10754	70,86526	7,67	0,0829	264	20
19	Madiun	-1	0,1204	10667	69,75904	7,74	0,1121	190	16
20	Magetan	-1	0,118	10539	71,90756	7,99	0,0443	226	9
21	Ngawi	-1	0,1488	10143	71,32969	7,04	0,1011	100	108
22	Bojonegoro	-1	0,1548	8964	70,11457	6,73	0,1076	397	33
23	Tuban	-1	0,1664	8906	70,25275	6,81	0,137	322	6
24	Lamongan	-1	0,1568	9545	71,47288	7,86	0,0608	433	41
25	Gresik	-1	0,1341	11514	72,19857	8,87	0,0246	341	15
26	Bangkalan	1	0,2238	7459	69,61537	5,94	0,1473	259	21
27	Sampang	1	0,258	7798	67,4765	4,62	0,2207	145	41
28	Pamekasan	-1	0,1774	7478	66,55786	6,55	0,1173	167	22
29	Sumenep	-1	0,2049	7143	70,02066	5,62	0,1563	283	42

Lampiran 2
Data Indikator-indikator daerah tertinggal (*Lanjutan*)

No	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
1	2	0,9961	0,79	,5871	1,674307	19	0,083715	1,172015
2	0	1	1,84	0,7954	1,539601	16	0,068144	0,987516
3	3	0,998	1,78	0,515	1,453156	33	0,090276	0,949356
4	3	1	1,73	0,6665	1,438029	23	0,081695	0,875022
5	0	0,9976	4,52	0,6839	1,313122	13	0,055225	0,945833
6	0	0,9989	1,87	0,6682	1,300905	13	0,06433	0,712833
7	2	0,9986	3,59	0,8121	1,298333	24	0,032448	0,779158
8	1	1	1,55	0,6741	1,420529	12	0,085738	0,919739
9	7	0,989	0,83	0,7295	1,254485	23	0,045519	0,789273
10	3	0,9972	1,17	0,7212	1,530777	107	0,068007	0,817968
11	0	0,9868	1,74	0,5645	1,431989	7	0,100398	1,179674
12	0	0,996	0,2	0,4643	1,496968	9	0,075074	1,031511
13	0	0,9924	1,49	0,6544	1,192736	6	0,059151	1,20951
14	1	0,9966	1,37	0,5963	1,350743	12	0,05798	0,822551
15	0	1	9,04	0,9161	0,941493	51	0,07054	0,471706
16	1	1	1,75	0,7655	1,248965	19	0,043905	0,820188
17	4	1	1,38	0,7629	1,47833	12	0,072904	0,865127
18	0	0,9981	1,26	0,7773	1,293216	8	0,06071	0,884629
19	0	0,9989	2,57	0,7454	1,467385	17	0,096441	0,870935
20	0	0,999	3,17	0,8975	1,76185	31	0,08139	1,075622
21	9	0,999	3,59	0,7983	1,671843	41	0,077311	0,972423
22	0	0,9991	2,26	0,7782	1,631794	25	0,079521	1,000498
23	0	1	0,44	0,7672	1,523847	31	0,050562	0,842998
24	0	0,9988	0,56	0,7958	1,715127	23	0,06318	1,275394
25	0	1	3,91	0,8923	1,277371	16	0,086178	0,825539
26	1	0,9892	2,21	0,618	1,420988	39	0,049692	1,200016
27	0	0,996	1,1	0,7227	1,361902	16	0,044281	1,583306
28	0	1	0,72	0,901	1,437414	30	0,062184	1,309458
29	6	0,9895	0,27	0,6863	1,641676	38	0,055285	1,513303

Lampiran 3
Data Indikator-indikator daerah tertinggal (*Lanjutan*)

No	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}
1	0,795322	0,567251	0,128655	0,204678	0,631579	0,005848	0,0942	0,6652
2	0	0,153094	0,13355	0,107492	0,358306	0	0,1159	0,6231
3	0,019108	0,375796	0,343949	0,191083	0,713376	0	0,0891	0,6377
4	0,02952	0,088561	0,059041	0,118081	0,254613	0,04059	0,129	0,6435
5	0,016129	0,100806	0,044355	0,080645	0,322581	0,012097	0,103	0,6474
6	0	0,026163	0,063953	0,113372	0,119186	0,017442	0,0781	0,6328
7	0,017949	0,130769	0,064103	0,161538	0,282051	0,005128	0,1104	0,6645
8	0	0,063415	0,073171	0,121951	0,17561	0,004878	0,1232	0,637
9	0,012097	0,068548	0,21371	0,112903	0,137097	0,016129	0,1134	0,6654
10	0,147465	0,018433	0,078341	0,193548	0,184332	0,013825	0,1556	0,6952
11	0	0,100457	0,118721	0,228311	0,255708	0	0,132	0,6281
12	0,014706	0,125	0,279412	0,588235	0,213235	0,007353	0,1289	0,6193
13	0,006061	0,039394	0,163636	0,312121	0,29697	0	0,0826	0,6303
14	0	0,060274	0,194521	0,191781	0,186301	0,010959	0,1433	0,6292
15	0	0	0,122857	0,08	0,002857	0,017143	0,3413	0,6413
16	0	0,075658	0,194079	0,144737	0,180921	0,049342	0,1876	0,6322
17	0	0,022876	0,179739	0,107843	0,104575	0,03268	0,132	0,6443
18	0	0,059859	0,165493	0,140845	0,25	0,056338	0,1063	0,6274
19	0	0,067961	0,07767	0,101942	0,26699	0,019417	0,1226	0,6702
20	0	0,06383	0,021277	0,093617	0,093617	0	0,0939	0,6535
21	0	0,02765	0,225806	0,317972	0,373272	0	0,0566	0,621
22	0	0,111628	0,360465	0,255814	0,325581	0,016279	0,2245	0,6405
23	0	0,003049	0,106707	0,085366	0,259146	0,036585	0,114	0,6108
24	0	0	0,130802	0,033755	0,181435	0,025316	0,0876	0,6068
25	0	0,016854	0,308989	0,08427	0,053371	0,019663	0,1962	0,6208
26	0	0,021352	0,039146	0,135231	0,010676	0,010676	0,0766	0,5987
27	0	0,177419	0,107527	0,193548	0	0,010753	0,07	0,6276
28	0	0,126984	0,074074	0,301587	0	0,015873	0,1175	0,6136
29	0	0,063253	0,042169	0,213855	0,039157	0,012048	0,132	0,6054

Lampiran 4
Nilai α yang didapatkan dari kernel RBF, linier, dan sigmoid

RBF[1]	RBF[2]	Linier[1]	Linier[2]	Sigmoid[1]	Sigmoid[2]
0,0068006559	0,0000019229	0,0037664386	0,0000000062	0,0122346900	0,0312499977
0,0000009892	0,0000001557	0,0000000004	0,0000000005	0,0000000385	0,0000000788
0,0070829026	0,0000002307	0,0060400177	0,0000000006	0,0000001822	0,0000000510
0,0000298604	0,0000001830	0,0000000002	0,0000000004	0,0000000460	0,0000000511
0,0004301597	0,0000003274	0,0000000006	0,0000000009	0,0000000262	0,0000000070
0,0000064793	0,0000001833	0,0000000005	0,0000000005	0,0000000422	0,0000000074
0,0002948744	0,0000002141	0,0000000002	0,0000000008	0,0000000102	0,0000000566
0,0085700164	0,0483197675	0,0312499540	0,0531249701	0,0057826977	0,0265624989
0,0064300485	0,0499999656	0,0089428743	0,0499999999	0,0132804694	0,0249999926
0,0052970030	0,0000001871	0,0000000002	0,0000000008	0,0000000121	0,0000000181
0,0312499994	0,0624999995	0,0312500000	0,0624999999	0,0156249995	0,0312499999
0,0312499994	0,0624999994	0,0312500000	0,0624999999	0,0156249999	0,0312499998
0,0240165727	0,0499999854	0,0249999999	0,0499999996	0,0124998386	0,0156233871
0,0017991244	0,0000002253	0,0000006916	0,0000000010	0,0000002702	0,0000010767
0,0066499441	0,0000004447	0,0000000001	0,0000000002	0,0000000072	0,0000000078
0,0012105638	0,0000001454	0,0000000003	0,0000000004	0,0000000416	0,0000000501
0,0000014084	0,0000001329	0,0000000002	0,0000000003	0,0000000423	0,0000000501
0,0000872136	0,0000001953	0,0000000012	0,0000000008	0,0000000484	0,0000000076
0,0000356468	0,0000002053	0,0000000006	0,0000000010	0,0000000387	0,0000000053
0,0015909533	0,0000001610	0,0000000002	0,0000000004	0,0000000216	0,0000000501
0,0060605810	0,0000003286	0,0000000077	0,0000000015	0,0000005134	0,0000000020
0,0033295814	0,0000004385	0,0000000091	0,0000000028	0,0000000437	0,0000000025
0,0025054454	0,0133218585	0,0000000018	0,0000000338	0,0000000329	0,0000000020
0,0021890457	0,0000001590	0,0000000007	0,0000000005	0,0000000170	0,0000000501
0,0027273571	0,0000001745	0,0000000002	0,0000000003	0,0000000394	0,0000000501
0,0312499994	0,0624999994	0,0312500000	0,0624999999	0,0156249995	0,0312499998
0,0312499994	0,0624999995	0,0312500000	0,0624999999	0,0156249994	0,0312499997
0,0136565996	0,0383524134	0,0249999997	0,0468749769	0,0124998128	0,0265624978
0,0241969710	0,0499999924	0,0249999998	0,0499999996	0,0062010163	0,0000000009

Keterangan
RBF[1] = nilai alfa dari EFSVM kernel RBF semua variabel
RBF[2] = nilai alfa dari EFSVM kernel RBF seleksi variabel
Linier[1] = nilai alfa dari EFSVM kernel linier semua variabel
Linier[2] = nilai alfa dari EFSVM kernel linier seleksi variabel
Sigmoid[1] = nilai alfa dari EFSVM kernel sigmoid semua variabel
Sigmoid[2] = nilai alfa dari EFSVM kernel sigmoid seleksi variabel

Lampiran 5
Syntax SVM

```
library(kernlab)
library(e1071)
Kfold = function(y, k) {
y = as,vector(as,matrix(y))
datax = data,frame(y = y, idx = c(1:length(y)))
n = length(y)
nk = ceiling(n/k)
datak = vector("list", k)
levely = as,numeric(levels(as,factor(y)))
for (j in 1:length(levely)) {
datai = datax[which(datax[,1]==levely[j]), ]
nc = dim(datai)[1]
nck = ceiling(nc/k)
# acak data setiap setiap kelas
set,seed(2222)
sam = sample(nc, replace = F)
sam_datai = datai[sam,]
for (i in 1:k) {
if (i==k) {
datak[[i]] = c(datak[[i]],sam_datai[(((i-1)*nck)+1):nc,2])
} else {
datak[[i]] = c(datak[[i]],sam_datai[(((i-1)*nck)+1):(i*nck),2])
}
}
}
}
# grid search
gridSearchx = function(y, x, C, Sigma, fold_sample) {
#-----------------#
```

Lampiran 6
Syntax SVM (Lanjutan[1]**)**

```
# ACCURACY
#------------------#
pred<-function(x,lable){
C=NULL
n = length(x)
for (i in 1:n){
if(x[i]>0){C[i]=1}
else {C[i]=-1}
}
TP = 0
for (i in 1:n) {if (C[i]==1 & lable[i]==1) (TP=TP+1)}
FN = 0
for (i in 1:n) {if (C[i]!=1 & lable[i]==1) (FN=FN+1)}
FP = 0
for (i in 1:n) {if (C[i]==1 & lable[i]!=1) (FP=FP+1)}
TN = 0
for (i in 1:n) {if (C[i]!=1 & lable[i]!=1) (TN=TN+1)}
Sensi = TP/(TP+FN)
Spesi = TN/(TN+FP)
Gmean = sqrt(Sensi*Spesi)
AUC = (1+(TP/(TP+FN))-(FP/(FP+TN)))/2
accuracy=mean(C==lable)
result=list(C,accuracy)
conmat=table(C,lable) #confusion matrix
list(accuracy=accuracy, conmat=conmat,
Sensi = Sensi, Spesi = Spesi,
Gmean = Gmean, AUC = AUC)
}
y = as,numeric(as,matrix(y))
x = x
```

Lampiran 7
Syntax SVM (Lanjutan[2]**)**

```
C = C
Sigma = Sigma
fold_sample = fold_sample
nf = length(fold_sample)
nC = length(C)
nS = length(Sigma)
result = vector("list", nC)
for (i in 1:nC){
# Matrix Akurasi
mat,has = matrix(0, nrow = nf, ncol = nS)
colnames(mat,has) = Sigma
rownames(mat,has) = c(1:nf)
# Matrix Spesi
Spesi = matrix(0, nrow = nf, ncol = nS)
colnames(Spesi) = Sigma
rownames(Spesi) = c(1:nf)
# Matrix Sensi
Sensi = matrix(0, nrow = nf, ncol = nS)
colnames(Sensi) = Sigma
rownames(Sensi) = c(1:nf)
# Matrix Gmean
Gmean = matrix(0, nrow = nf, ncol = nS)
colnames(Gmean) = Sigma
rownames(Gmean) = c(1:nf)
# Matrix AUC
AUCx = matrix(0, nrow = nf, ncol = nS)
colnames(AUCx) = Sigma
rownames(AUCx) = c(1:nf)
# Matrix CV
CV = matrix(0, nrow = 1, ncol = nC)
colnames(CV) = Sigma
#n = 0
conmat = vector("list", nS)
for (k in 1:nS) {
```

Lampiran 8
Syntax SVM (Lanjutan[3]**)**

```
CVx = 0
Mconmat = matrix(0,nrow = nf, ncol = 4)
Lampiran Syntax SVM
colnames(Mconmat) = c("TN", "FN", "FP", "TP")
#n = n+1
for (j in 1:nf) {
xj = x[-fold_sample[[j]],]
yj = y[-fold_sample[[j]]]
hasil =try(svm(xj,yj,cost=C[i],gamma=Sigma[k]))
Xj = x[fold_sample[[j]],]
Yj = y[fold_sample[[j]]]
fx=predict(hasil,Xj)
akurasi=pred(fx,Yj)
conmatx = akurasi$conmat
for (con in 1:length(conmatx)) {
Mconmat[j,con] = conmatx[con]
}
mat,has[j,k] = akurasi$accuracy
Spesi[j,k] = akurasi$Spesi
Sensi[j,k] = akurasi$Sensi
Gmean[j,k] = akurasi$Gmean
AUCx[j,k] = akurasi$AUC
}
}
result[[i]] = list(Akurasi = mat,has,
Sensitivity = Sensi,
Specificity = Spesi,
Gmeans = Gmean,
AUC = AUCx)
}
names(result) = C
return(result)
}
```

Lampiran 9
Lampiran Syntax EFSVM

```
EFSVMrbf= function(y, x, C, Sigma, fold_sample,B,K,M) {
 quadb<-function(x,y,cost, Sigma){
  library(kernlab)
  x = as,matrix(x)
  y = as,matrix(y)
  m=dim(x)[1]
  rbf <- rbfdot(sigma = Sigma)
  ## create H matrix etc,
  H <- kernelPol(rbf,x,,y)
  c <- matrix(rep(-1,m))
  A <- t(y)
  b <- 0
  l <- matrix(rep(0,m))
  u <- cost
  r <- 0
  capture,output(sv <- ipop(c,H,A,b,l,u,r,verb=TRUE,max-
iter=400))
  ipopsol<-primal(sv)
  alpha<-matrix(ipopsol, nrow=m)
  #-------------------------------------------------#
  # Calculation of the normal vector W and bias term b
  #-------------------------------------------------#
  w=t(alpha*y)%*%(x) #W
  ff=as,matrix(matrix(rep(alpha*y,m),m,m))%*%H
  fout=matrix(t(apply(ff,2,sum)))
  pos=which(alpha>1e-6)
  b = mean(y[pos]-fout[pos]) #b
  list(W = w, b = b)
}
EbasedF = function(x, y, Beta, kNN, m) {
  y = as,numeric(y)
  ndata = dim(x)[1]
  rownames(x) = c(1:ndata)
```

Lampiran 10
Syntax ESVM (*Lanjutan*[1])

```
jarak1 = as,matrix(dist(x, method = "euclidean", diag = TRUE,
upper = TRUE, p = 2))
 coba=diag(ndata)
 jarak=coba+jarak1
 # H
 H = rep(0, ndata)
 LN = data,frame()
 for (i in 1:ndata){
  Nterdekat = as,numeric(names(sort(jarak[i,]))[1:kNN])
  # L=label
  LNi = y[Nterdekat]
  LN = rbind(LN,LNi)
  Hpi = table(LNi)/kNN
  lHpi = length(as,vector(Hpi))
  if (lHpi>1) {
   H[i] = -Hpi[2]*log(Hpi[2])-Hpi[1]*log(Hpi[1])
  } else {
   H[i] = 0
  }
 }
 colnames(LN) = paste0("NN",c(1:kNN))
 # Subl
 Sub = vector("list", m)
 thr = matrix(0, m, 2)
 for (l in 1:m) {
  thrUp = min(H) + (l/m)*(max(H) - min(H))
  thrLow = min(H) + ((l-1)/m)*(max(H) - min(H))
  for (i in 1:ndata) {
   if (y[i]==-1) {
    if (H[i]>=thrLow & H[i]<=thrUp) {
     Sub[[l]] = c(Sub[[l]], i)
    }
   }
  }
```

Lampiran 11
Syntax ESVM (*Lanjutan*^{2})

```
thr[l,1] = thrLow
   thr[l,2] = thrUp
  }
  # FM
  FM = c()
  for (l in 1:m){
    Fmx = 1-Beta*(l-1)
    FM = c(FM, Fmx)
  }
  # si
  Si = rep(1, ndata)
  for (l in 1:m) {
    Subl = Sub[[l]]
    Si[Subl] = FM[l]
  }
  list(dist = jarak, H = H, thr = thr,
       Subl = Sub, FM = FM, si = Si, LN = LN)
}
  pred<-function(x,lable){
    C=NULL
    n = length(x)
    for (i in 1:n){
      if(x[i]>0){C[i]=1}
      else {C[i]=-1}
    }
    TP = 0
    for (i in 1:n) {if (C[i]==1 & lable[i]==1) (TP=TP+1)}
    FN = 0
    for (i in 1:n) {if (C[i]!=1 & lable[i]==1) (FN=FN+1)}
    FP = 0
    for (i in 1:n) {if (C[i]==1 & lable[i]!=1) (FP=FP+1)}
    TN = 0
    for (i in 1:n) {if (C[i]!=1 & lable[i]!=1) (TN=TN+1)}
    Sensi = TP/(TP+FN)
```

Lampiran 12
Syntax ESVM (*Lanjutan*[3])

```
    Spesi = TN/(TN+FP)
Gmean = sqrt(Sensi*Spesi)
    AUC = (1+(TP/(TP+FN))-(FP/(FP+TN)))/2
    accuracy=mean(C==lable)
    result=list(C,accuracy)
    conmat=table(C,lable) #confusion matrix
    list(accuracy=accuracy, conmat=conmat,
         Sensi = Sensi, Spesi = Spesi,
         Gmean = Gmean, AUC = AUC)
  }
  x = x
  C = C
  Sigma = Sigma
  fold_sample = fold_sample
  nf = length(fold_sample)
  nC = length(C)
  nS = length(Sigma)
  result = vector("list", nC)
  for (i in 1:nC){
    # Matrix Akurasi
    mat,has = matrix(0, nrow = nf, ncol = nS)
    colnames(mat,has) = Sigma
    rownames(mat,has) = c(1:nf)
    # Matrix Spesi
    Spesi = matrix(0, nrow = nf, ncol = nS)
    colnames(Spesi) = Sigma
    rownames(Spesi) = c(1:nf)
    # Matrix Sensi
    Sensi = matrix(0, nrow = nf, ncol = nS)
    colnames(Sensi) = Sigma
    rownames(Sensi) = c(1:nf)
    # Matrix Gmean
    Gmean = matrix(0, nrow = nf, ncol = nS)
```

Lampiran 13
Syntax ESVM (*Lanjutan*[4])

```
colnames(Gmean) = Sigma
   rownames(Gmean) = c(1:nf)
   # Matrix AUC
   AUCx = matrix(0, nrow = nf, ncol = nS)
   colnames(AUCx) = Sigma
   rownames(AUCx) = c(1:nf)
   # Matrix CV
   CV = matrix(0, nrow = 1, ncol = nC)
   colnames(CV) = Sigma
   #n = 0
   conmat = vector("list", nS)
   for (k in 1:nS) {
    CVx = 0
    Mconmat = matrix(0,nrow = nf, ncol = 4)
    colnames(Mconmat) = c("TN", "FN", "FP", "TP")
    #n = n+1
    for (j in 1:nf) {
     xj = x[-fold_sample[[j]],]
     yj = y[-fold_sample[[j]],]
     Si = EbasedF(x = xj,y = yj,Beta = B,kNN = K, m = M)
     si = Si$si
     hasil = try(quadb(xj,yj,C[i]*si,Sigma[k]))
     if(inherits(hasil, "try-error")) {
      mat,has[j,k] = 0
     } else {
      Xj = x[fold_sample[[j]],]
      Yj = y[fold_sample[[j]],]
      fx=t(hasil$W %*% t(as,matrix(Xj))) + hasil$b
      akurasi=pred(fx,Yj)
      conmatx = akurasi$conmat
```

Lampiran 14
Syntax ESVM (*Lanjutan*[5])

```
for (con in 1:length(conmatx)) {
          Mconmat[j,con] = conmatx[con]
        }
        mat,has[j,k] = akurasi$accuracy
      Spesi[j,k] = akurasi$Spesi
        Sensi[j,k] = akurasi$Sensi
        Gmean[j,k] = akurasi$Gmean
        AUCx[j,k] = akurasi$AUC
       }
     }
     for (j in 1:nf) {
       CVx = CVx + (1-mat,has[j,k])*length(fold_sam-
ple[[j]])/length(y)
     }
     CV[,i] = CVx
     conmat[[k]] = Mconmat
   }
   result[[i]] = list(Akurasi = mat,has,
                      Sensitivity = Sensi,
                      Specificity = Spesi,
                      Gmeans = Gmean,
                      AUC = AUCx)
  }
  names(result) = C
  return(result)
}
```

Lampiran 15
Lampiran Syntax pemilihan jenis kernel EFSVM

```
quadl<-function(x,y,cost, Sigma){
  library(kernlab)
  x = as,matrix(x)
  y = as,matrix(y)
  m=dim(x)[1]
  p <- vanilladot()
  ## create H matrix etc,
  H <- kernelPol(p,x,,y)
  c <- matrix(rep(-1,m))
  A <- t(y)
  b <- 0
  l <- matrix(rep(0,m))
  u <- cost
  r <- 0
  capture,output(sv <- ipop(c,H,A,b,l,u,r,verb=TRUE,max-
iter=400))
  ipopsol<-primal(sv)
  alpha<-matrix(ipopsol, nrow=m)
  #-------------------------------------------------#
  # Calculation of the normal vector W and bias term b
  #-------------------------------------------------#
  w=t(alpha*y)%*%(x) #W
  ff=as,matrix(matrix(rep(alpha*y,m),m,m))%*%H
  fout=matrix(t(apply(ff,2,sum)))
  pos=which(alpha>1e-6)
  b = mean(y[pos]-fout[pos]) #b
```

Lampiran 16
Lampiran Syntax pemilihan jenis kernel EFSVM (*Lanjutan*)

```
quadtan<-function(x,y,cost, Sigma){
  library(kernlab)
  x = as,matrix(x)
  y = as,matrix(y)
  m=dim(x)[1]
  p <- tanhdot(scale = Sigma)
  ## create H matrix etc,
  H <- kernelPol(p,x,,y)
  c <- matrix(rep(-1,m))
  A <- t(y)
  b <- 0
  l <- matrix(rep(0,m))
  u <- cost
  r <- 0
  capture,output(sv <-
ipop(c,H,A,b,l,u,r,verb=TRUE,maxiter=400))
  ipopsol<-primal(sv)
  alpha<-matrix(ipopsol, nrow=m)
  #-------------------------------------------------#
  # Calculation of the normal vector W and bias term b
  #-------------------------------------------------#
  w=t(alpha*y)%*%(x) #W
  ff=as,matrix(matrix(rep(alpha*y,m),m,m))%*%H
  fout=matrix(t(apply(ff,2,sum)))
  pos=which(alpha>1e-6)
  b = mean(y[pos]-fout[pos]) #b
  list(W = w, b = b)
}
```

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini, mahasiswa Departemen Statistika FMKSD ITS:

Nama : Jefry Pranata Maulana
NRP : 06211645000026

menyatakan bahwa data yang digunakan dalam Tugas Akhir/~~Thesis~~ ini merupakan data sekunder yang diambil dari ~~Penelitian/ Buku/ Tugas Akhir/ Thesis~~/ Publikasi lainnya yaitu:

Sumber : a. Badan Pusat Statistik Provinsi Jawa Timur
b. Dinas Kesehatan Provinsi Jawa Timur
c. Kementerian Keuangan

Keterangan : a. Statistik Potensi Desa Provinsi Jawa Timur 2014
b. Profil Kesehatan Provinsi Jawa Timur 2014
c. Realisasi Kemampuan Keuangan Daerah 2014

Surat Pernyataan ini dibuat dengan sebenarnya. Apabila terdapat pemalsuan data maka saya siap menerima sanksi sesuai aturan yang berlaku.

Mengetahui
Pembimbing Tugas Akhir

Irhamah, PhD
NIP. 19780406 200112 2 002

Surabaya, 30 Juli 2018

Jefry Pranata Maulana
NRP. 06211645000026

BIODATA PENULIS

Penulis terlahir dengan nama Jefry Pranata Maulana, biasa dipanggil Jefry, Pranata, atau Gilang. Penulis dilahirkan di Jombang pada tanggal 30 Agustus 1993 dan merupakan anak kedua dari pasangan Bapak Jalil dan Ibu Sugiwati. Pendidikan formal yang ditempuh penulis adalah SDN Tunggorono 1 Jombang, SMPN 1 Jombang, dan SMAN 2 Jombang. Setelah lulus dari SMA dan Diploma 3 Statistika di ITS Surabaya, Penulis melanjutkan ke S1 Lintas Jalur Statistika pada tahun 2016. Aktifitas penulis semasa kuliah semester 1 hingga 4 adalah mengikuti komunitas android dan bekerja di DTC. Bagi Penulis, bekerja bermanfaat tidak hanya mendapatkan penghasilan, tetapi juga memberikan pengalaman unik. Penulis sangat hobi sekali dengan musik dan olahraga. Penulis juga mengikuti beberapa seminar di luar Surabaya, seperti di UGM dan UNY selama masa kuliah semester 1 dan 2. Penulis juga sangat senang dengan musik Bagi penulis “musik adalah warna untuk menghiasi hidup”. Segala kritik, saran dan pertanyaan untuk penulis dapat dikirimkan melalui alamat email jefri.88999@gmail.com atau jika kurang jelas bisa juga menghubungi di No. Hp 083811669111. Terimakasih.