Jurnal EduMatSains, 1 (2) Januari 2017, 119-136

Studi Perbandingan Pemilihan Fitur untuk *Support Vector Machine* pada Klasifikasi Penilaian Risiko Kredit

Desri Kristina Silalahi[1*], Hendri Murfi[2], Yudi Satria[3]

[1]Jurusan Pendidikan Matematika FIP UPH, Jln. Boulevard Mh. Thamrin 1100
[2,3]Departemen Matematika Fakultas MIPAUniversitas Indonesia, Kampus UI Depok 16424

*e-mail: desri.kristina@uph.edu

Abstract

Credit scoring is a system or method used by banks or other financial institutions to determine the debtor feasible or not get a loan. One of credit scoring method is used to classify the characteristics of debtor is Support Vector Machine (SVM). SVM has an excellent generalization ability to solve classification problems in a large amount of data and can generate an optimal separator function to separate two groups of data from two different classes. One of the success using SVM method is dependent on features selection process that will affect the level of classification accuracy. Various methods have done to features selection, because not all the features are able to give best classification results. Features selection that used this study is Variance Threshold, Univariate Chi - Square, Recursive Feature Elimination (RFE) and Extra Trees Classifier (ETC). Data in this study use secondary data from the database in UCI machine learning responsitory. Based on simulations to compare the accuracy of using feature selection method on SVM in classification ofcredit riskscoring, obtained that Variance Threshold and Univariate Chi – Square method can decrease accuracy while RFE and ETC method can increase accuracy. RFE method gives better accuracy.

Keywords: *Credit scoring, Credit risk, Feature selection, Support vector machine*

PENDAHULUAN

Dalam kehidupan perekonomian suatu negara, bank memiliki peranan penting dalam perekonomian. Menurut UU Perbankan No. 10 Tahun 1998, bank adalah badan usaha yang menghimpun dana dari masyarakat dalam bentuk simpanan dan menyalurkannya kepada masyarakat dalam bentuk kredit dan / atau bentuk – bentuk lainnya dalam rangka meningkatkan taraf hidup orang banyak. Dengan demikian hampir seluruh keperluan setiap orang dan segenap lapisan masyarakat dalam kegiatan perekonomian terkait dengan perbankan.

Salah satu pelayanan dalam dunia perbankan adalah memberi pinjaman kredit kepada nasabah yang memenuhi syarat perbankan. Kredit merupakan suatu fasilitas keuangan yang memungkinkan seseorang atau badan usaha meminjam uang untuk membeli produk dan membayarnya kembali dalam jangka waktu yang ditentukan. Sedangkan dalam UU No. 10 tahun 1998 menyebutkan bahwa kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam

meminjam antara bank dengan pihak lain. yang mewajibkan pihak peminjam untuk melunasi utangnya setelah jangka watu tertentu dengan pemberian bunga. Kredit merupakan sumber utama penghasilan bagi sebuah bank dan juga sekaligus sumber risiko operasi bisnis terbesar, karena sebagian besar dana operasional bank diputar dalam bentuk kredit.

Risiko kredit ditimbulkan oleh debitur yang secara kredit tidak dapat membayar utang dan memenuhi kewajiban. Risiko kredit merupakan sesuatu yang tidak dapat dihindari. Tindakan yang diperlukan adalah bagaimana mendeteksi, mengukur, dan mengelola risiko agar tidak menimbulkan kerugian yang besar. Dalam menekan atau meminimalisasi risiko kredit, bank perlu melakukan proses analisis data nasabah, data jaminan hingga proses pengambilan keputusan untuk melihat calon debitur yang layak diberi pinjaman. Proses tersebut merupakan proses penilaian atau *scoring* dengan menggunakan data – data historis calon peminjam untuk diklasifikasikan layak diberi pinjaman atau tidak. Calon debitur yang lolos seleksi disebut *good debitur* sehingga permohonan kreditnya akan dikabulkan. Sebaliknya, calon debitur yang tidak lolos seleksi disebut *bad debitur* sehingga permohonan kreditnya akan ditolak. Proses tersebut dinamakan *credit scoring*.

Credit scoring atau penilaian kredit merupakan sistem atau cara yang digunakan oleh bank atau lembaga keuangan lainnya untuk mengambil keputusan mengenai kelayakan seorang calon debitur untuk diterima menjadi nasabah kredit (Thomas,L.C, 2002). Banyak analisis yang digunakan untuk mengklasifikasikan karakteristik calon debitur dengan sifat *good* dan *bad.* Metode klasifikasi yang paling populer dalam model *credit scoring* adalah analisis diskriminan dan regresi logistik. Kedua metode ini merupakan teknik klasifikasi yang sederhana dan mudah, tetapi kedua metode ini memiliki keterbatasan. Metode ini tidak efesien untuk jumlah data yang besar karena membutuhkan waktu komputasi yang lebih lama dan beberapa asumsi harus terpenuhi seperti data dipisahkan secara linier serta data harus mengikuti distribusi tertentu (Abdou, H. & Pointon, J., 2011). Metode yang juga sering digunakan untuk klasifikasi adalah *Support Vector Machine (SVM)*, dapat digunakan untuk jumlah data yang besar dan dapat menghasilkan fungsi pemisah *(classifier)* yang optimal untuk memisahkan dua kelompok data dari dua kelas yang berbeda. SVM telah mempunyai kemampuan generalisasi yang baik dalam

berbagai bidang, termasuk dalam penilaian risiko kredit (Gestel, T.V & Baesens, B, 2009).

Dalam bidang *machine learning*, dimensi data *input* yang digunakan disebut sebagai fitur, sehingga fitur merupakan hal yang penting dalam menentukan akurasi klasifikasi. Dengan demikian, pemilihan fitur merupakan sebuah tahapan penting karena fitur yang terseleksi, sangat mempengaruhi tingkat akurasi dari klasifikasi. Berbagai metode dilakukan untuk melakukan pemilihan fitur, karena tidak semua fitur mampu memberikan hasil klasifikasi baik. Salah satu keberhasilan menggunakan metode SVM adalah proses pemilihan fitur dan parameter (Zhao, M, 2011). Metode pemihan fitur dikelompokkan menjadi tiga yaitu metode *Filter,* metode *Wrapper* dan metode *Embedded* (Guyon, I & Elisseeff, A.2003). Metode *Filter* merupakan pemilihan fitur yang mengasumsikan bahwa setiap fitur saling bebas sehingga mengabaikan adanya fitur yang saling bergantung. Yang termasuk dalam metode Filter yaitu *Variance Threshold* dan *Univariate Chi – Square*. Metode Wrapper dan metode Embedded memilih fitur yang saling bergantung terhadap pengklasifikasian. Metode Wrapper cenderung mengalami *overfitting* yang menyebabkan klasifikasi tidak efisien, sedangkan metode *Embedded* mengurangi kecenderungan terhadap *overfitting* serta kompleksitas komputasionalnya lebih baik dari pada metode Wrapper. Salah satu metode Embedded adalah *Recursive Feature Elimination (RFE)*. Metode RFE ini telah diaplikasikan dalam beberapa masalah klasifikasi, obat dan non obat serta menghasilkan tingkat akurasi yang lebih tinggi (Korkmaz S, 2014). Selain ketiga kelompok pemilihan fitur tersebut, ada metode yang lain yaitu *Extra Trees Clasifier (ETC)*. Metode ETC merupakan metode pemilihan fitur berbasis *decision tree*. Metode ETC ini telah diaplikasikan dalam pemilihan fitur untuk SVM pada masalah analisis sentimen dan menghasilkan akurasi lebih tinggi dibandingkan dengan metode pemilihan fitur lainnya (Prawira, A, 2014).

METODE PENELITIAN

A. Pengolahan Data

Pengolahan data menggunakan metode *Support Vector Machine (SVM).* SVM dikembangkan oleh Boser, Guyon & Vapnik yang pertama kalinya dipresentasikan pada Tahun 1992 di *Annual Workshop on Computational Learning Theory*. SVM merupakan salah satu mesin pembelajaran alternatif yang digunakan

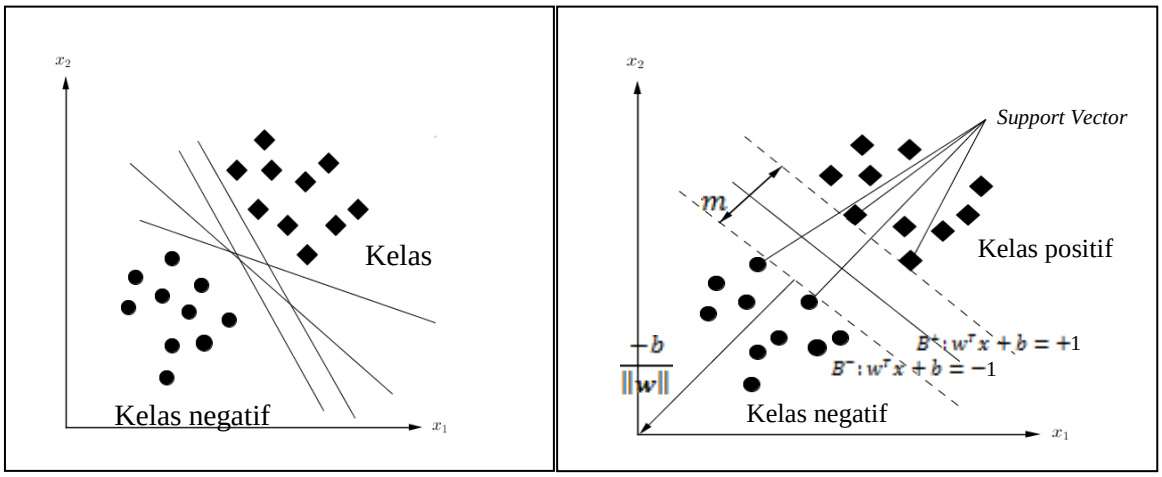
dalam menyelesaikan persoalan klasifikasi dengan menggunakan konsep *maximum margin.*

SVM merupakan salah satu metode *supervised learning*, karena himpunan data pelatihan (*trainingdata*) berupa vektor *input* yang diberikan target sebagai *output*. Tujuan pembelajaran ini adalah membangun model yang dapat menghasilkan *output* yang benar jika diberikan *input* yang baru. Model umum yang digunakan adalah model linear. Model linear tersebut digunakan untuk memisahkan data pembelajaran ke dalam dua kelas yang berbeda, yaitu kelas positif dan kelas negatif.

Maximal margin classifier

Misalkan $\{\boldsymbol{x_i}, y_i\}, \ i = 1, 2, \dots, N$ adalah himpunan pasangan data sebanyak N, dengan $\boldsymbol{x_i} = [x_1, x_2, \dots, x_m]$ adalah vektor baris berdimensi m (banyaknya fitur) dan $y_i \in \{+1, -1\}$ adalah target (kelas) pada setiap vektor baris, persamaan $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = 0$ dinamakan *hyperplane* yang memisahkan data menjadi dua kelas, dengan $\boldsymbol{w}$ adalah vektor yang mempresentasikan parameter bobot, $\boldsymbol{x}$ adalah vektor input dan *b* adalah bias atau error yang berupa skalar. Pada Gambar 1 dapat dilihat bahwa *hyperplane* yang digunakan untuk mengklasifikasikan data tidaklah unik maka dengan SVM, *hyperplane* optimal tidak hanya memisahkan data tetapi juga memiliki margin yang maksimum.

Memaksimumkan *margin* berarti memaksimumkan nilai $\frac{1}{\|\boldsymbol{w}\|}$ yang setara dengan meminimumkan nilai $\|\boldsymbol{w}\|^2$, maka pencarian *hyperplane* terbaik dengan memaksimumkan *margin* dapat dirumuskan menjadi masalah optimisasi berikut (Scholkopf, B & Smola, A., 2002) :



Gambar 1. H*yperplane – hyperplane* yang memisahkan data pelatihan (kiri) dan *hyperplane* dengan margin maksimum (kanan)

$$\arg\min_{\boldsymbol{w}, b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$s.t\ \ y_i(\boldsymbol{w^T x}_i + b) \geq 1$$

Soft margin dengan metode kernel

Dalam SVM, untuk mengatasi beberapa data yang tidak dapat diklasifikasikan secara benar *(misclassification error)*, maka dilakukan margin lunak *(soft margin)*. *Soft margin* tersebut memungkinkan beberapa data yang berada pada sisi *decision boundary* yang salah atau memberikan kelunakan untuk beberapa data yang salah dalam pengklasifikasian. Untuk merealisasikan *soft margin* ini, diperkenalkan variabel *slack* $\xi_i \geq 0$ dimana $i = 1, 2, \dots, n$ dengan satu variabel *slack* untuk setiap data pelatihan (Scholkopf & Smola, 2002). Variabel *slack* bernilai $\xi_i = |y_i - f(\boldsymbol{x_i})|$, sehingga:

- Jika $\xi_i = 0$ maka data berada pada batas margin dan pada sisi *decision boundary* yang benar.
- Jika $0 < \xi_i < 1$ maka data berada di dalam margin dan pada sisi *decision boundary* yang benar.
- Jika $\xi_i > 1$ maka data berada pada sisi *decision boundary* yang salah atau data salah diklasifikasikan

Namun, dalam aplikasinya tidak semua data dapat dipisahkan secara linear oleh *hyperplane* walaupun telah menambahkan variabel *slack*. Untuk mengatasi hal ini digunakan metode Kernel dengan memetakan data ke dimensi lebih tinggi, sehingga diharapkan data dapat dipisahkan secara linear atau disebut juga bersifat *linearly separable* (Scholkopf, B & Smola, A, 2002). Adapun fungsi kernel yang biasa digunakan dalam SVM yaitu: (Bishop, Christopher M. 2006).

1. Fungsi Kernel Linear:$K(\boldsymbol{x},\ \boldsymbol{x'}) = \boldsymbol{x}^T\boldsymbol{x}$
2. Fungsi Kernel Polinomial:
   $K(\boldsymbol{x},\ \boldsymbol{x'}) = (\boldsymbol{x}^T\boldsymbol{x'} + 1)^p$
3. Fungsi Kernel *Radial Basis Function (RBF):*$K(\boldsymbol{x}, \boldsymbol{x'}) = \exp(-\gamma\|\boldsymbol{x} - \boldsymbol{x'}\|^2)$

Dalam penelitian ini akan digunakan fungsi kernel RBF karena dapat memetakan data input secara nonlinear ke dimensi yang lebih tinggi sehingga diharapkan dapat menangani kasus ketika hubungan antara label kelas dan fitur – fitur yang tidak linear (Hsu, C, 2010) dan menghasilkan akurasi lebih baik dibandingkan dengan fungsi kernel lainnya. Sehingga, diperoleh masalah *soft margin* dengan metode kernel sebagai berikut :

$$\max_{a} \quad \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i y_i a_j y_j \left(\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)\right)$$

$$s.t \quad \sum_{i=1}^{n} a_i y_i = 0$$

$$0 \leq a_i \leq C \qquad , i = 1,\ 2,\ \dots,\ n$$

Maka, *hyperplane*metode *soft margin* dengan metode kernel adalah

$$f(\boldsymbol{x}) = \sum_{i\in S}^{n} a_i^* y_i \left(\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)\right) + b(12)$$

dengan S adalah himpunan indeks *support vector*.

B. Proses *Learning*

Pada proses *learning* bertujuan untuk penentuan parameter dari metode pada data pelatihan yang diberikan. Dalam SVM dengan fungsi kernel RBF dibutuhkan dua parameter yaitu parameter C dan γ. Parameter C merupakan parameter yang digunakan untuk mengukur *trade off* dari kesalahan pengklasifikasian data pelatihan dan nilai parameter γ yang semakin besar berarti sebagian data pelatihan akan semakin berpengaruh terhadap data pelatihan yang lainnya. Metode pendekatan yang digunakan untuk mendapatkan kombinasi yang optimal dari C dan γ yaitu pendekatan *Grid Search (*Hsu, C, 2010). Proses *Grid Search* secara lengkap membutuhkan waktu proses yang lama, sehingga direkomendasikan untuk menggunakan metode pembagian *Grid Search* menjadi dua tahap, yaitu:

1) Loose Grid Search, dilakukan pemilihan parameter $C = 2^x$, dengan $x = -5, -3, \dots, 15$ dan $\gamma = 2^y$, dengan $y = -15, -13, \dots, 3$. Dalam proses pemilihan parameter C dan γ yang optimal dilakukan secara iterative untuk setiap pemasangan C dan γ yang berbeda.

2) Fine Grid Search, dilakukan pencarian yang lebih kecil pada titik yang ditemukan pada proses sebelumnya. Misalkan pada proses *Loose Grid Search* diperoleh C yang optimal adalah x^* dan γ optimal yang diperoleh adalah γ^*, maka pada proses *Fine Grid Search* akan mencari nilai parameter

$C = 2^{x^*+x'}$ dengan $x' = -2, -1.75, -1.5, \dots, 2$

Dan $\gamma = 2^{y^*+y'}$ dengan $y' = -2, -1.75, -1.5, \dots, 2.$

C. Proses Evaluasi Model

Evaluasi model bertujuan untuk mengetahui keakuratan model fungsi klasifikator dalam memprediksi data baru yang bukan termasuk dalam data pelatihan. *K – Fold Cross Validation* digunakan untuk menghitung akurasi model fungsi klasifikator terhadap data baru.

Dalam penelitian ini, yang digunakan *5 – Fold Cross Validation*. Untuk menyajikan hasil *K – Cross Validation*, digunakan *confusion matrix*.

Tabel 1. Skema *Confusion Matrix*

		Target Prediksi	
		-1	1
Target Sebenarnya	-1	*True Negative (TN)*	*False Positive (FP)*
	1	*False Negative (FN)*	*True Positive (TP)*

Berdasarkan skema *confusion matrix* pada tabel 1, menurut Gorunescu (2011) satuan kinerja model yaitu sebagai berikut:

- *Succes rate* atau tingkat kesuksesan adalah proporsi jumlah prediksi yang benar. Dapat dihitung dengan:

  $Succes\ rate = \frac{TP+TN}{TP+TN+FP+FN}$

- *True Positif Rate (TPR)* adalah membandingkan proporsi *True Positif (TP)* terhadap tupel yang positif. Dapat dihitung dengan:

  $TPR = \frac{TP}{TP+FN}$

- *False Positif Rate (FPR)* adalah membandingkan proporsi *False Positif (FP)* terhadap tupel yang negatif. Dapat dihitung dengan:

  $FPR = \frac{FP}{FP+TN}$

ROC Curve(Receiver Operating Charateristic) adalah alat visual yang berguna untuk membandingkan dua atau lebih model klasifikasi. *ROC Curve* adalah grafik dua dimensi dengan *False Positive* sebagai garis horizontal dan *True Positive* sebagai garis vertikal (Gorunescu, F, 2011). Tingkat akurasi nilai AUC *(Area Under The Curve)* dalam klasifikasi dibagi menjadi lima kelompok dinyatakan dalam tabel berikut :

Berdasarkan skema *confusion matrix* pada tabel 1, menurut Gorunescu (2011) satuan kinerja model yaitu sebagai berikut:

- *Succes rate* atau tingkat kesuksesan adalah proporsi jumlah prediksi yang benar. Dapat dihitung dengan:

  $Succes\ rate = \frac{TP+TN}{TP+TN+FP+FN}$

- *True Positif Rate (TPR)* adalah membandingkan proporsi *True Positif (TP)* terhadap tupel yang positif. Dapat dihitung dengan:

  $TPR = \frac{TP}{TP+FN}$

- *False Positif Rate (FPR)* adalah membandingkan proporsi *False Positif (FP)* terhadap tupel yang negatif. Dapat dihitung dengan:

  $FPR = \frac{FP}{FP+TN}$

Tabel 2. Kelompok Tingkat Akurasi Nilai AUC *(Area Under The Curve)* dalam Klasifikasi

Interval Nilai AUC	Akurasi Klasifikasi
0.90 – 1.00	Klasifikasi sangat baik *(excellent classification)*
0.80 – 0.90	Klasifikasi baik *(good classification)*
0.70 – 0.80	Klasifikasi cukup *(fair classification)*
0.60 – 0.70	Klasifikasi buruk *(poor classification)*
0.50 – 0.60	Klasifikasi salah *(failure classification)*

D. Proses Pemilihan Fitur

Dalam proses klasifikasi, dimensi data *input* yang digunakan disebut sebagai fitur**.** Dalam proses klasifikasi, pemilihan fitur merupakan sebuah tahapan penting karena fitur yang terseleksi sangat berpengaruh terhadap penentuan fungi klasifikator *(hyperplane).* Metode pemilihan fitur yang digunakan yaitu :

1. *Variance Threshold*

Metode *Variance Threshold* adalah metode yang mengeliminasi fitur yang memiliki variansi di bawah batas tertentu. Persamaan yang digunakan untuk mencari varians dari masing – masing fitur adalah:

$$Var(x_j) = \frac{\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2}{n}$$

dengan:

$x_j \quad , j = 1, 2, \dots, m$ merupakan fitur

$\overline{x}_j \quad , j = 1, 2, \dots, m$ merupakan rata – rata dari suatu fitur

n merupakan jumlah data.

2. *Univariate Chi – Square*

Nilai dari *Chi – Square Statistic* dapat dihitung menggunakan persamaan berikut:

$$\chi^2 = \sum_{i=1}^{b}\sum_{j=1}^{m}\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}$$

dengan:

o_{ij} merupakan frekuensi yang diperoleh atau diamati

e_{ij} merupakan frekuensi yang diharapkan.

Adapun langkah – langkah pengujian *Chi – Square Statistic* yaitu:

- Hipotesis

  H_0: tidak ada hubungan yang signifikan antara fitur data set dengan *credit scoring*

  H_a: ada hubungan yang signifikan antara fitur data set dengan *credit scoring*.
- Menentukan taraf nyata yaitu $\alpha = 0{,}05$
- Menghitung o_{ij} dan e_{ij} . Untuk menghitung frekuensi harapan e_{ij}, akan dibuat tabel kontingensi.

  Berdasarkan Tabel 3 tersebut bahwa $A_1, A_2, \dots, A_m$ adalah fitur – fitur yang ada pada satu fitur.
- Mencari nilai frekuensi yang diharapkan $\left(e_{ij}\right)$

  $e_{ij} = \frac{(n_{i.})(n_{.j})}{n}$; $i = 1, 2, \dots, b$;

  $j = 1, 2, \dots, m$
- Menghitung nilai *Chi – Square Statistic* menggunakan dengan berdistribusi *chi – square* dengan derajat kebebasan yaitu $(b-1)(m-1)$
- Menarik kesimpulan

  Dengan kriteria keputusan:Tolak H_0 jika $\chi^2 \geq \chi^2_{tabel(\alpha,v)}$.

Tabel 3. Bentuk Umum Tabel Kontingensi

	A_1	A_2	...	A_m	Jumlah
Label kelas (L_1) = 1	o_{11}	o_{12}	...	o_{1m}	$n_{1.}$
Label kelas (L_2)= -1	o_{21}	o_{22}	...	o_{2m}	$n_{2.}$
Jumlah	$n_{.1}$	$n_{.2}$	...	$n_{.m}$	N

3. *Recursive Feature Elemination (RFE)*

Algoritma dari SVM RFE pertama kali diusulkan oleh Guyon, Weston, Barnhill & Vapnik (2002), yaitu sebagai berikut:

1. Diberikan $\{\boldsymbol{x}_i, y_i\}$, $i = 1, 2, \dots, n$ adalah himpunan pasangan data sebanyak *n*, dengan $\boldsymbol{x}_i \in \mathbb{R}^m$ dan $y_i \in \{+1, -1\}$ adalah target.
2. Pelatihan klasifikasi dengan menggunakan SVM untuk mendapatkan vektor $\boldsymbol{w}$ yang mempresentasikan parameter bobot.
$$\boldsymbol{w} = \sum_{i=1}^{n} a_i y_i \boldsymbol{x}_i$$
3. Hitung vektor $\boldsymbol{w}$ yang mempresentasikan parameter bobot
4. Hitung skor ranking untuk semua fitur $c_j = (w_j)^2$ dengan w_j , $j = 1, 2, \dots, m$ adalah elemen pada vektor $\boldsymbol{w}$
5. Pencarian fitur dengan skor ranking terendah yaitu $f = \arg\min(c_j)$
6. Mengeliminasi fitur dengan $c = f$ (fitur yang dieliminasi bisa lebih dari satu fitur).

4. *Extra Trees Classifier (ETC)*

ETC merupakan pemilihan fitur yang berbasis *decision tree*. Algoritma ETC berbeda dari metode *decision tree* lainnya, yaitu menggunakan parameter, membagi *node* dengan memilih titik potong seluruhnya secara acak dan membangun setiap *tree* dengan menggunakan sampel data pelatihan asli. Parameter yang digunakan dalam algoritma ETC yaitu:

- *K* merupakan parameter yang menentukan jumlah fitur yang diambil secara acak pada setiap *node*.
- n_{min} merupakan parameter yang menentukan banyaknya sampel minumum yang digunakan untuk melakukan pemisahan *node*.
- *M* merupakan parameter yang menentukan jumlah pohon di *ensemble*.

Metode ETC mengeliminasi fitur – fitur yang berada di bawah bobot yang ditentuka sesuai dengan algoritma Pseudo – code dari algoritma *Extra Trees* (Geurts, 2006)

HASIL DAN PEMBAHASAN

Tiga *dataset* yaitu *German Credit*, *Australian Credit* dan *Japanese Credit*. Ketiga *dataset* ini merupakan data yang berisi karakteristik penilaian dari calon debitur yang mengajukan kredit.

Pada proses awal akan dilakukan pemilihan fitur. Metode pemilihan fitur yang digunakan yaitu *Variance Threshold*, *Univariate Chi Square*, *Recursive Feature Elimination (RFE)* dan *Extra Trees Classifier (ETC)*. Berikut hasil pemilihan fitur pada ketiga *dataset* tersebut.

Setelah dilakukan pemilihan fitur, selanjutnya data pelatihan ini akan diproses menggunakan metode SVM dengan fungsi kernel RBF. Karena penelitian ini menggunakan SVM dengan fungsi kernel RBF maka dibutuhkan parameter C dan γ. Setelah parameter C dan γ yang optimal terpilih, maka selanjutnya melakukan evaluasi model. Evaluasi model yang digunakan adalah *5 – Fold Cross Validation* untuk menghitung kinerja yang dihasilkan SVM. Satuan kinerja yang digunakan adalah akurasi atau *succes rate*.

Tabel 4. Gambaran Data Penelitian

Dataset	**Jumlah Calon Debitur**	**Jumlah Fitur Kategorik**	**Jumlah Fitur Numerik**	**Total Fitur**	**Jumlah Kelas**	**Jumlah *Good Debitur***	**Jumlah *Bad Debitur***
German Credit	1000	13	7	20	2	700	300
Australian Credit	690	8	6	14	2	307	383
Japanese Credit	690	9	6	15	2	307	383

[Sumber data: http://archive.ics.uci.edu/ml/datasets/]

Tabel 5. Jumlah Fitur yang Dihasilkan oleh Metode Pemilihan Fitur pada Ketiga *Dataset*

	Variance Threshold	***Univariate Chi Square***	***Recursive Feature Elimination (RFE)***	***Extra Trees Classifier (ETC)***
German Credit	11	10	15	11
Australian Credit	8	10	1	3
Japanese Credit	8	10	1	3

Tabel 6. Perbandingan Hasil Akurasi pada Ketiga *Dataset*

Dataset	**Tanpa Pemilihan Fitur**	**Dengan Pemilihan Fitur**			
		Variance Threshold	***Univariate Chi – Square***	***RFE***	***ETC***
German Credit	(77,6±0)%	(76,9±0)%	(74,5±0)%	(81,3±2)%	(79,0±1)%
Australian Credit	(87,1±1)%	(80,1±0)%	(86,8±1)%	(95,7±1)%	(90,6±2)%
Japanese Credit	(85,8±7)%	(78,6±2)%	(85,7±7)%	(95,1±4)%	(91,6±7)%

Pada Tabel 6 terlihat bahwa dengan metode pemilihan fitur menggunakan metode *Variance Threshold* dan *Univariate Chi – Square* mengurangi akurasi klasifikasi sedangkan metode ETC dan RFE dapat meningkatkan akurasi sebelumnya dan metode RFE menghasilkan akurasi lebih tinggi dibandingkan dengan metode pemilihan fitur lainnya.

Untuk menyajikan hasil dari *5 – Fold Cross Validation* digunakan *confusion matrix*. *Confusion matrix* berisi informasi sebenarnya dan prediksi pada klasifikasi seperti skema *confusion matrix* pada Tabel 1.

Tabel 7. *Confusion Matrix* pada Data *German Credit*

Tanpa Pemilihan Fitur			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	142	158
		Good	66	634
Variance Threshold			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	138	162
		Good	69	631
Univariate Chi – Square			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	145	155
		Good	81	619
Recursive Feature Elimination (RFE)			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	140	133
		Good	54	673
Extra Trees Classifier (ETC)			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	149	140
		Good	70	641

Pada Tabel 7, pada data *German Credit* memiliki 1000 calon debitur, klasifikasi SVM dengan fungsi kernel RBF pada penilaian risiko kredit diperoleh bahwa 142 diprediksi *'bad'* sesuai dengan hasil sebenarnya *'bad*, 66 diprediksi *'bad'* tetapi hasil sebenarnya *'good*, 158 diprediksi *'good '*tetapi hasil sebenarnya *'bad'* dan 634 diprediksi *'good'* sesuai dengan hasil sebenarnya *'good'*. Sedangkan setelah melakukan pemilihan fitur menggunakan metode *Variance Threshold*, diperoleh bahwa 138 diprediksi *'bad'* sesuai dengan hasil sebenarnya *'bad'*, 69 diprediksi *'bad '*tetapi hasil sebenarnya *'good'*, 162 diprediksi *'good '*tetapi hasil sebenarnya *'bad'* dan 631 diprediksi *'good'*sesuai dengan dengan hasil sebenarnya *'good'*. Pemilihan fitur menggunakan metode *Univariate Chi – Square,* diperoleh bahwa 145 diprediksi *'bad '*sesuai dengan hasil sebenarnya *'bad'*, 81 diprediksi *'bad '*tetapi hasil sebenarnya *'good'*, 155 diprediksi *'good'* tetapi hasil sebenarnya *'bad'* dan 619 diprediksi *'good'* sesuai dengan dengan hasil sebenarnya *'good'*.

Tabel 8. *Confusion matrix* pada data *Australian Credit*

Tanpa Pemilihan Fitur			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	320	63
		Good	26	281
Variance Threshold			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	352	31
		Good	106	201
Unvariate Chi – Square			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	320	63
		Good	28	279
Recursive Feature Elimination (RFE)			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	325	26
		Good	4	335
Extra Trees Classifier (ETC)			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	312	48
		Good	17	313

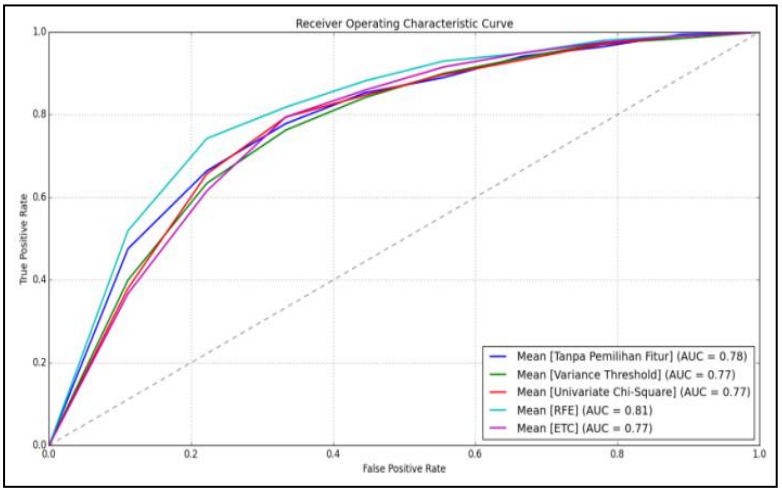
Tabel 9. C*onfusion matrix* pada data *Japanese Credit*

Tanpa Pemilihan Fitur			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	309	74
		Good	24	283
Variance Threshold			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	*Bad*	351	32
		Good	116	191
Unvariate Chi – Square			Klasifikasi Prediksi	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	306	77
		Good	22	285
Recursive Feature Elimination (RFE)			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	323	28
		Good	6	333
Extra Trees Classifier (ETC)			**Klasifikasi Prediksi**	
			Bad	***Good***
	Klasifikasi Sebenarnya	***Bad***	318	47
		Good	11	314

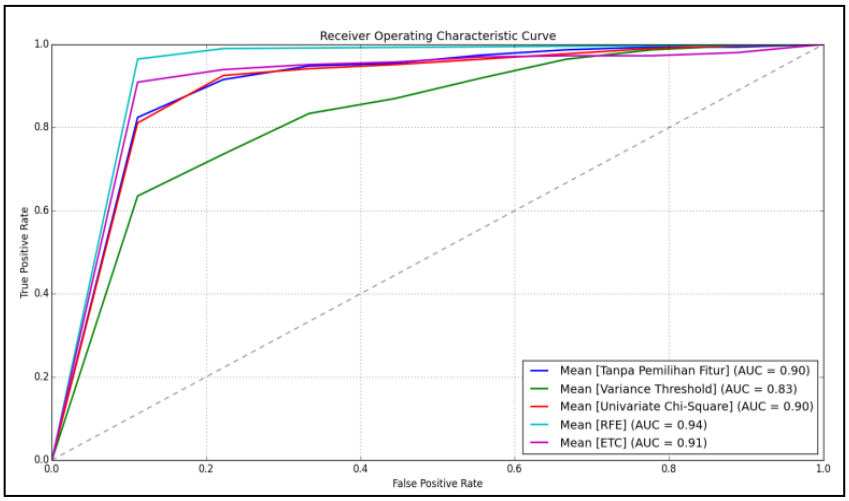
Pada Tabel 8, pada data *Australian Credit* memiliki 690 calon debitur, klasifikasi SVM dengan fungsi kernel RBF penilaian risiko kredit diperoleh diperoleh bahwa 320 diprediksi *'bad'* sesuai dengan hasil sebenarnya *'bad*, 63 diprediksi *'bad'* tetapi hasil sebenarnya *'good,* 26 diprediksi *'good'* tetapi hasil sebenarnya *'bad'* dan 281 diprediksi *'good'* sesuai dengan hasil sebenarnya *'good'*. Sedangkan setelah melakukan pemilihan fitur menggunakan metode *Variance Threshold,* diperoleh bahwa 352 diprediksi *'bad'* sesuai dengan hasil sebenarnya *'bad'*, 106 diprediksi *'bad'* tetapi hasil sebenarnya *'good'*, 31 diprediksi *'good'* tetapi hasil sebenarnya *'bad'* dan 201 diprediksi *'good'* sesuai dengan dengan hasil sebenarnya *'good'*. Pemilihan fitur menggunakan metode *Univariate Chi – Square*, diperoleh bahwa 320 diprediksi *'bad'* sesuai dengan hasil sebenarnya*'bad'*, 28 diprediksi *'bad'* tetapi hasil sebenarnya *'good'*, 63 diprediksi *'good'* tetapi hasil sebenarnya *'bad'* dan 279 diprediksi *'good'* sesuai dengan dengan hasil sebenarnya *'good'*. Pemilihan fitur menggunakan metode RFE, diperoleh bahwa 325 diprediksi *'bad'* sesuai dengan hasil sebenarnya *'bad'*, 4 diprediksi *'bad'* tetapi hasil sebenarnya *'good'*, 26 diprediksi *'good '*tetapi hasil sebenarnya *'bad'* dan 335 diprediksi *'good'* sesuai

dengan dengan hasil sebenarnya *'good'*. Pemilihan fitur menggunakan metode ETC, diperoleh bahwa 312 diprediksi *'bad '*sesuai dengan hasil sebenarnya *'bad'*, 17 diprediksi *'bad '*tetapi hasil sebenarnya *'good'*, 48 diprediksi *'good'* tetapi hasil sebenarnya *'bad'* dan 313 diprediksi *'good'* sesuai dengan dengan hasil sebenarnya *'good'*.

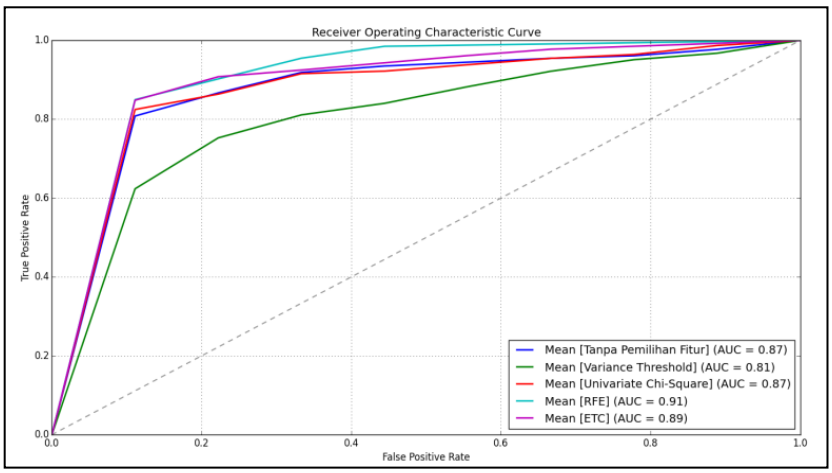
Pada Tabel 9, pada data *Japanese Credit* memiliki 690 debitur, klasifikasi SVM dengan fungsi kernel RBF penilaian risiko kredit diperoleh diperoleh bahwa 309 diprediksi *'bad'* sesuai dengan hasil sebenarnya *'bad*, 24 diprediksi *'bad'* tetapi hasil sebenarnya *'good*, 74 diprediksi *'good'* tetapi hasil sebenarnya *'bad'* dan 283 diprediksi *'good'* sesuai dengan hasil sebenarnya *'good'*. Sedangkan setelah melakukan pemilihan fitur menggunakan metode *Variance Threshold*, diperoleh bahwa 351 diprediksi *'bad'* sesuai dengan hasil sebenarnya *'bad'*, 116 diprediksi *'bad'* tetapi hasil sebenarnya *'good'*, 32 diprediksi *'good'* tetapi hasil sebenarnya *'bad'* dan 191 diprediksi *'good'* sesuai dengan dengan hasil sebenarnya *'good'*. Pemilihan fitur menggunakan metode *Univariate Chi – Square*, diperoleh bahwa 306 diprediksi *'bad'* sesuai dengan hasil sebenarnya *'bad'*, 22 diprediksi *'bad'* tetapi hasil sebenarnya *'good'*, 77 diprediksi *'good'* tetapi hasil sebenarnya *'bad'* dan 285 diprediksi *'good'* sesuai dengan dengan hasil sebenarnya *'good'*. Pemilihan fitur menggunakan metode RFE, diperoleh bahwa 323 diprediksi *'bad'* sesuai dengan hasil sebenarnya *'bad'*, 6 diprediksi *'bad'*tetapi hasil sebenarnya *'good'*, 28 diprediksi *'good '*tetapi hasil sebenarnya *'bad'* dan 333 diprediksi *'good'* sesuai dengan dengan hasil sebenarnya *'good'*.



Gambar 2. Grafik perbandingan *ROC Curve* metode pemilihan fitur pada *data set German Credit*



Gambar 3. Grafik perbandingan *ROC Curve* metode pemilihan fitur pada *data set Australian Credit*



Gambar 4. Grafik perbandingan *ROC Curve* metode pemilihan fitur pada *data set Japanese Credit*

Pemilihan fitur menggunakan metode ETC, diperoleh bahwa 318 diprediksi *'bad'* sesuai dengan hasil sebenarnya *'bad'*, 11 diprediksi *'bad'* tetapi hasil sebenarnya *'good'*, 47 diprediksi *'good'* tetapi hasil sebenarnya *'bad'* dan 314 diprediksi *'good'* sesuai dengan dengan hasil sebenarnya *'good'*.

ROC Curve ini digunakan untuk melihat perbandingan *True Positif Rate (TPR)* dengan *False Positif Rate (FPR).* Berikut perbandingan *ROC Curve* metode pemilihan fitur pada masing – masing *data set.*

Berdasarkan Gambar 4 terlihat bahwa hasil AUC *(Area Under Curve)* dari metode pemilihan fitur RFE memiliki nilai AUC

yang lebih tinggi yaitu 0,91. Hal ini berarti dengan melakukan pemilihan fitur menggunakan metode RFE memiliki kemampuan prediksi lebih

tinggi dibandingkan metode pemilihan fitur yang lainnya. Sehingga berdasarkan kelompok tingkat akurasi nilai AUC, maka tingkat akurasi dari metode RFE adalah klasifikasi sangat baik *(excellent classification).*

KESIMPULAN

Berdasarkan simulasi untuk membandingkan nilai akurasi penggunaan metode pemilihan fitur *Variance Threshold*, *Univariate Chi – Square*, RFE dan ETC pada SVM dalam klasifikasi penilaian risiko kredit, maka diperoleh kesimpulan bahwa:

- Metode *Variance Threshold* dan *Univariate Chi – Square* dapat mengurangi akurasi klasifikasi setelah dilakukukan pemilihan fitur.
- Metode RFE dan ETC dapat meningkatkan akurasi klasifikasi setelah dilakukukan pemilihan fitur. Metode RFE memberikan akurasi yang lebih baik.

DAFTAR PUSTAKA

Abdou, H. & Pointon, J. 2011. Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of The Literature. *Intelligent Systems in Accounting, Finance & Management vol 18, 59 – 88.*

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Cambridge: Spinger.

Boser, B, Guyon, I & Vapnik, V. 1992. Annual Workshop on Computational Learning Pattern Recognition. *Kluwer Academy Publisher*. Boston.

Gestel, T.V & Baesens, B. 2009. *Credit Risk Management Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital*. New York: Oxford University Press.

Geurts, P, Ernst, D & Wehenkel, L. 2006. Extremely Randomized Trees. *Mach Learn vol 63, 3 – 42.*

Gorunescu, F. 2011. *Data Mining: Concepts, Models and Techniques*. Verlag Berlin Heidelberg: Springer.

Guyon, I, Weston, J, Barnhill, S & Vapnik, V. 2002. Gene Selection For Cancer Classification Using Support Vector Machines. *Machine Learning, vol. 46(1-3), 389 – 422.*

Guyon, I & Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research vol 3, 1157 – 1182* .

Hsu, C. W., Chang, C. C., & Lin, C. J. 2010. *A Practical Guide to Support Vector Classification*. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Korkmaz S, Zararsiz G & Goksuluk D. 2014. Drug / Nondrug Classification Using Support Vector Machine With Various Feature Selection Strategies. *Computer Methods And Programs In Biomedicine vol 117*. 51 – 60.

Prawira, A. 2014. *Analisis Kinerja Pemilihan Fitur untuk Support Vector Machine (SVM) pada Masalah Analisis Sentimen*. Skripsi, Universitas Indonesia.

Scholkopf, B & Smola, A. 2002. *Learning with Kernels – Support Vector Machines, Regularization, Optimizationand Beyond*. Cambridge: The MIT Press.

UCI *machine learning responsitory (*http://archive.ics.uci.edu/ml/datasets/). Diakses pada 23 Februari 2015.

Thomas, L. C., Edelman, D. B. & Crook, L. N. 2002. *Credit Scoring and Its Applications*. Philadelphia: Society for Industrial and Applied Mathematics.

Zhao, M., Fu, C., Ji, L., Tang, K. & Zhou, M. 2011. Feature Selection and Parameter Optimization For Support Vector Machines: A New Approach Based on Genetic Algorithm with Feature Chromosomes. *Expert System with Applications, vol 38(5), 5197-5204*