PREDIKSI SPAM EMAIL MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE* DAN *PARTICLE SWARM OPTIMIZATION*

Eni Pudjiarti
Program Studi Teknik Informatika
STMIK Nusa Mandiri Jakarta
Jl. Damai No. 8 Warung Jati Barat Margasatwa, Jakarta Selatan. Telp. (021) 78839513
enipudjiarti@nusamandiri.ac.id

Abstract — *Spam* email *is the term used to describe the messages sent in mass emails or incoming email received without approval. Spam Email Filtering is a program used to detect unwanted email and prevent unsolicited email and to get into the inbox of email users. Many of the techniques used to create spam email filtering , one of them by using classification techniques. Support Vector Machine classifier is a supervised learning method is used to classify the data. But the Support Vector Machine has a weakness on the difficulty of selecting the appropriate features and optimal attribute weights are used to cause a degree of prediction accuracy is low. Therefore, in this study created a model algorithm and Support Vector Machine Support Vector Machine algorithm model based Particle Swarm Optimization to get email spam rule in predicting accuracy and provide a more accurate value. After testing the two models, namely Support Vector Machine Algorithm and Support Vector Machine -based Particle Swarm Optimization, the results obtained are thus obtained testing the algorithm using Support Vector Machine which is the value obtained 85.75 % accuracy and AUC value was 0.901, whereas the test using Support Vector Machine-based Particle Swarm Optimization value obtained 89.24 % accuracy and AUC value was 0.935 with a good level of diagnostic classification. So that both methods have different levels of accuracy that is equal to 3:49 % and AUC values of 0.034 difference*

Intisari — *Spam* email adalah istilah yang digunakan untuk menggambarkan pesan yang dikirim dalam email massal atau email yang masuk diterima tanpa persetujuan. *Spam Email Filtering* adalah program yang digunakan untuk mendeteksi email yang tidak diinginkan dan mencegah email yang tidak diminta dan untuk masuk ke inbox pengguna email. Banyak teknik yang digunakan untuk membuat email penyaringan *spam*, salah satunya dengan menggunakan teknik klasifikasi. *Support Vector Machine classifier* adalah metode pembelajaran terawasi digunakan untuk mengklasifikasikan data. Tapi *Support Vector Machine* memiliki kelemahan pada kesulitan memilih fitur yang tepat dan bobot atribut optimal digunakan untuk menyebabkan tingkat akurasi prediksi rendah. Oleh karena itu, dalam penelitian ini menciptakan algoritma Model dan *Support Vector Machine Support Vector Machine* Model algoritma berdasarkan *Particle Swarm Optimization* untuk mendapatkan *spam email* aturan dalam memprediksi akurasi dan memberikan nilai yang lebih akurat. Setelah menguji dua model, yaitu *Support Vector Algoritma* Mesin dan *Support Vector Machine* berbasis *Particle Swarm Optimization*, hasil yang diperoleh dengan demikian diperoleh menguji algoritma menggunakan *Support Vector Machine* yang merupakan nilai yang diperoleh akurasi 85,75% dan nilai AUC adalah 0,901, sedangkan Tes menggunakan *Support Vector Machine* berbasis *Particle Swarm Optimization* nilai yang diperoleh akurasi 89,24% dan nilai AUC adalah 0,935 dengan tingkat yang baik klasifikasi diagnostik. Sehingga kedua metode memiliki berbagai tingkat akurasi yaitu sebesar 03:49% dan nilai-nilai AUC 0,034 perbedaan.

Kata Kunci: *Feature Selection, Particle Swarm Optimization*, *Spam Email, Support Vector Machine Algorithm.*

PENDAHULUAN

Spam adalah email yang tidak diminta oleh pengguna (*unsolicitied* email) yang dikirim ke banyak orang. *Spam* email adalah istilah yang digunakan untuk menggambarkan pesan terkirim dalam email massal atau email masuk yang diterima tanpa persetujuan. *Spam* email biasanya datang dalam bentuk penawaran obat, proposal bisnis haram, pesan hoax atau iklan produk. Oleh karena itu email perlu diprediksi dengan akurat, karena hasil prediksi yang akurat sangat penting untuk mengurangi *spam*.

Klasifikasi adalah salah satu teknik dalam *data mining* yang digunakan untuk memprediksi kelompok keanggotaan (*class*) dari setiap *instance* data. Teknik klasifikasi yang biasa digunakan untuk membangun *e-mail spam filtering* diantaranya *Naïve Bayes, Support Vector Machine (SVM)* dan *k-nearest neighbor (kNN)*. Dari beberapa teknik tersebut yang paling sering

digunakan untuk klasifikasi data adalah *Support Vector Machine (SVM)*.

Support Vector Machine (SVM) adalah suatu metode *supervised learning* yang digunakan untuk melakukan klasifikasi data. *Support Vector Machine* (SVM) adalah kasus khusus dari keluarga algoritma yang kita sebut sebagai *regularized* metode klasifikasi *linier* dan metode yang kuat untuk minimalisasi resiko (Weiss, Indurkhya & Zhang, 2010). Dan kelebihan SVM lainnya adalah dapat meminimalkan kesalahan melalui memaksimalkan margin dengan memisahkan antara *hyper-lane* dan satu set data bahkan dengan jumlah *sample* yang kecil (Chunjiang & Yan, 2009). Namun demikian masalah aplikasi tertentu, tidak semua fitur ini sama-sama penting dan kinerja yang lebih baik dapat dicapai dengan membuang beberapa fitur dengan begitu fitur dalam SVM memiliki pengaruh penting dalam akurasi klasifikasi (Zhao, Fu, Ji, Tang & Zhou, 2011). Dataset yang tidak penting, fitur yang banyak atau sangat berhubungan secara signifikan akan mengurangi tingkat akurasi klasifikasi dengan menghapus fitur ini, dengan begitu tingkat akurasi efisiensi dan klasifikasi dapat diperoleh (Lin a, Shiue b & Chen, 2009).

Particle Swarm Optimization (PSO) sangat menarik untuk pemilihan fitur dimana kawanan partikel akan menemukan kombinasi fitur terbaik pada saat pencarian ruang masalah dana *Particle Swarm Optimization* (PSO) dapat menemukan solusi yang optimal dengan cepat (Parimala & Nallaswamy, 2012). *Particle Swarm Optimization* (PSO) banyak digunakan untuk memecahkan masalah optimasi, serta sebagai masalah seleksi fitur (Liu, Wang, Chen, Dong, Zhu & Wang, 2011). Dalam teknik *Particle Swarm Optimization* (PSO) terdapat beberapa cara untuk melakukan pengoptimasian diantaranya: meningkatkan bobot atribut (*attribute weight*) terhadap semua atribut atau variabel yang dipakai, menseleksi atribut (*attribute selection*) dan *feature selection*.

Pada penelitian ini *Particle Swarm Optimization* (PSO) akan diterapkan untuk pemilihan parameter *Support Vector Machine* (SVM) yang sesuai dan optimal, sehingga hasil prediksi lebih akurat.

BAHAN DAN METODE

a. *Spam* Email

Spam email atau bisa juga berbentuk junk mail adalah penyalahgunaan sistem pesan elektronik untuk mengirim berita iklan dan keperluan lainnya secara massal. Umumnya, *spam* menampilkan berita secara bertubi-tubi tanpa diminta dan sering kali tidak dikehendaki oleh penerimanya. Pada akhirnya, *spam* dapat menimbulkan ketidaknyamanan bagi para pengguna situs web. Orang yang menciptakan *spam* elektronik disebut *spammers*.

1) *Spam* yang masuk ke inbox *user*
   Spam yang masuk ke inbox biasanya berjenis:
   a. *Phising* email ini akan berpura-pura sebagai or
   b. *Hoax* : email peringatan atau nasehat palsu yang biasanya diakhiri dengan himbauan agar menyebarkannya seluas-luasnya.
2) *Account user* yang disalahgunakan *spammer* untuk mengirim *spam*
   Dalam hal ini *spammer* biasanya akan berusaha menebak *user* dan *password* email anda untuk bisa mengirim *spam* dari akun anda. Bila hal ini terjadi, maka efeknya sangat buruk untuk akun anda maupun reputasi network secara keseluruhan. Bila akun anda telah terbukti mengirim *spam*, maka *mail* adm akan mem-*blacklist* alamat email anda dan mengirimkan notifikasi ke *mail administrator* domain anda.
3) Penyebab email kita kebanjiran *spam*
   Biasanya *spam* email yang masuk menggunakan layanan *autoresponder* yang memang memiliki kemampuan untuk mengirimkan email secara massal dan otomatis, mulai ratusan hingga puluhan ribu email terkirim sekaligus.

b. *Support Vector Machine*

Support Vector Machine adalah sebuah metode seleksi yang membandingkan parameter standar seperangkat nilai diskrit yang disebut kandidat set, dan mengambil salah satu yang memiliki akurasi klasifikasi terbaik (Dong, Xia, Tu, & Xing, 2007). *Support Vector Machine* adalah salah satu alat yang paling berpengaruh dan kuat untuk memecahkan klasifikasi (Burges, 1998). *Support Vector Machines* (SVM) adalah seperangkat metode yang terkait untuk suatu metode pembelajaran, untuk kedua masalah klasifikasi dan *regresi* (Maimon, 2010). Dengan berorientasi pada tugas, kuat, sifat komputasi mudah dikerjakan, *Support Vector Machine* telah mencapai sukses besar dan dianggap sebagai *state-of-the-art classifier* saat ini (Huang, Yang, King, & Lyu, 2008).

Pada *Conference on Learning Theory* (COLT), Boser, Bernhard, Guyon, dan Vapnik tahun 1992, memperkenalkan *Support Vector Machine* (Premanode, Tzoumazou 2012) yaitu sebuah teknik *supervised learning* dari bidang *machine learning* yang dapat di aplikasikan kedalam *clasifikasi* dan *regresi* (Sewell, Taylor 2012). *Support Vector Machine* merupakan salah satu teknik yang relatif baru untuk melakukan

prediksi, *Support Vector Machine* berada dalam satu kelas dengan ANN dalam hal fungsi dan kondisi permasalahan yang bisa diselesaikan (Santosa, 2007). Dalam banyak implementasi *Supoort Vector Machine* memberikan hasil yang lebih baik dari ANN, dalam hal solusi yang dicapai. ANN menemukan solusi berupa local optimal dimana ANN akan selalu memberikan solusi yang berbeda dari setiap training, berbeda dengan *Support Vector Machine* yang menemukan solusi global optimal, dimana solusi akan memberikan hasil yang sama setiap dijalankan.

Ide dasar dari *Support Vector Machine* (SVM) adalah (Sewell, Taylor 2012) :

1. Memetakan *non-linear* input ke dimensi ruang fitur yang sangat tinggi (*kernel trick*).
2. Untuk kasus klasifikasi, membangun sebuah *hyperplane* pemisah (sebuah margin *pengklasifikasian* yang maksimal, atau dalam kasus *regresi*, menampilkan *linear regresi*, tetapi tanpa memperhatikan kesalahan-kesalahan kecil.

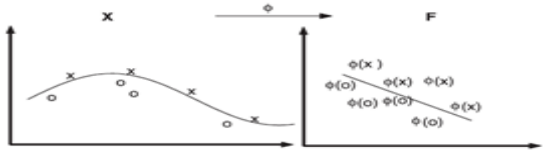
Yang menjadi karakteristik dari *Support Vector Machine* (SVM) adalah sebagai berikut :

1. Secara prinsip *Support Vector Machine* (SVM) adalah *linear classifier*.
2. *Pattern recognition* dilakukan dengan mentransformasikan data pada *input space* ke ruang yang berdimensi lebih tinggi, dan optimisasi dilakukan pada ruang *vector* yang baru tersebut. Hal ini membedakan *Support Vector Machine* (SVM) dari solusi *pattern recognition* pada umumnya, yang melakukan optimisasi parameter pada ruang hasil *transformasi* yang berdimensi lebih rendah daripada dimensi *input space*.
3. Menerapkan strategi *Structural Risk Minimization* (SRM).
4. Prinsip kerja *Support Vector Machine* (SVM) pada dasarnya hanya mampu menangani klasifikasi dua *class*.

Banyak teknik *data mining* atau *machine learning* yang dikembangkan dengan asumsi *kelinearan*, sehingga algoritma yang dihasilkan terbatas untuk kasus-kasus yang *linear*, maka untuk mengatasinya kita bisa menggunakan metode *kernel*.

Fungsi *kernel* yang biasa digunakan dalam *Support Vector Machine* (SVM) :

a. *Linear* : $x^T x$,
b. *Polinomial* : $(x^T x_i + 1)^P$,
c. *Radial basis function*(RBF) :exp $(-\frac{1}{2\sigma}||x-x_i||^2)$,
d. *Tangent hyperbolic*(*sigmoid*) : tanh $(\beta x^T x_i + \beta_1)$, dimana $\beta, \beta_1 \in R$?



Gambar 1. *Kernel* map mengubah *problem* yang tidak *linier* menjadi *linier* dalam *space* baru

c. *Particle Swarm Optimization*

Particle Swarm Optimization (PSO) merupakan algoritma pencarian berbasis populasi dan diinisialisasi dengan populasi solusi acak dan digunakan untuk memecahkan masalah optimasi (Abraham, Grosan & Ramos, 2006). *Particle Swarm Optimization* (PSO) teknik yang terinspirasi oleh proses alami burung yang berkelompok, dan juga dikenal sebagai segerombolan intelijen dengan mempelajari perilaku sosial atau kelompok hewan (Shukla, Tiwari & Kala, 2010). Untuk menemukan solusi yang optimal, masing-masing partikel bergerak ke arah posisi sebelumnya terbaik (pbest) dan terbaik posisi global (gbest). Kecepatan dan posisi partikel dapat diperbarui sebagai berikut persamaan:

$v_{ij}(t+1) = w * v_{i,j}(t) + c_1 * rand_1 * (pbest_{ij}(t) - p_{ij}(t)) + c_2 * rand_2 * (gbest_{ij}(t) - p_{ij}(t))$ (2.9)

$p_{ij}(t+1) = p_{ij}(t) + \beta * v_{ij}(t+1)$ (2.10)

Dimana :

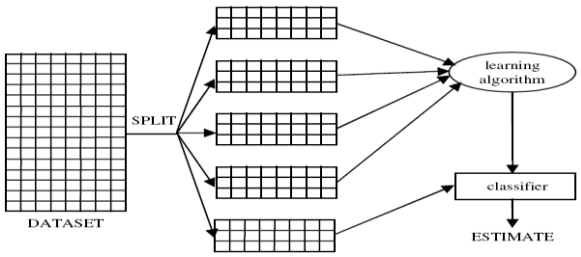
t : menunjukkan counter iterasi
V_{ij} : kecepatan partikel i pada dimensi ke-j (nilainya terbatas antara $[-v_{max}, v_{max}]$),
p_{ij} : posisi partikel i pada j dimensi (nilainya terbatas $[-p_{max}, p_{max}]$)
$pbest_{ij}$: posisi pbest partikel i pada dimensi ke-j
$gbest_{ij}$: posisi gbest dari dimensi ke-j
w : berat inersia (menyeimbangkan eksplorasi global dan local eksploitasi)
β : faktor kendala untuk mengontrol kecepatan berat (nilainya ke 1)
$rand_1$, $rand_2$: fungsi acak di rentang [0, 1]
c_1, c_2 : faktor pembelajaran pribadi dan sosial (nilainya ke 2)

d. Pengujian *K-Fold Cross Validation*

Salah satu pendekatan alternatif untuk "*train* dan *test*" yang sering di*adopsi* dalam beberapa kasus (dan beberapa lainnya terlepas dari ukurannya) yang disebut dengan *K-Fold Cross Validation* (Bramer, 2007) dengan cara

menguji besarnya *error* pada data *test* (santosa, 2007).

Kita gunakan k-1 sampel untuk *training* dan 1 sampel sisanya untuk *testing*. Misalnya ada 10 *subset* data, kita menggunakan 9 *subset* untuk *training* dan 1 *subset* sisanya untuk *testing*. Ada 10 kali *training* dimana pada masing-masing *training* ada 9 *subset* data untuk *training* dan 1 *subset* digunakan untuk *testing*. Setelah itu kemudian dihitung rata-rata *error* dan standar deviasi *error* (Santosa, 2007). Setiap bagian k pada gilirannya digunakan sebagai ujian menetapkan dan k lainnya -1 bagian digunakan sebagai *training set* (Bramer, 2007).



Gambar 2. *K-Fold Cross-validation*

e. Confusion Matrix

Confusion matrix merupakan *dataset* hanya memiliki dua kelas, kelas yang satu sebagai positif dan kelas yang lain sebagai negatif (Bramer, 2007). Metode ini menggunakan tabel *matrix*.

Tabel 1. Model *Confusion Matrix* (Vercellis, 2009)

		Predictions		Total
		-1 (negatif)	-1 (positif)	
Examples	-1 (negatif)	p	q	p+q
	+1 (positif)	u	v	u+v
	Total	p+u	q+v	m

Keterangan :

a. p adalah jumlah prediksi yang tepat bahwa *instance* bersifat negatif
b. q adalah jumlah prediksi yang salah bahwa *instance* bersifat positif
c. u adalah jumlah prediksi yang salah bahwa *instance* bersifat negatif
d. v adalah jumlah prediksi yang tepat bahwa *instance* bersifat positif.

Berikut adalah persamaan model *confusion matrix* :

a. nilai akurasi (acc) adalah proporsi jumlah prediksi yang benar. Dapat dihitung dengan menggunakan persamaan :

$$acc = \frac{p+v}{(p+q+u+v)} = \frac{p+v}{m} \quad \text{.....(1)}$$

b. tingkat negatif benar (tn) didefinisikan sebagai proporsi kasus negatif yang diklasifikasikan dengan benar, yang dihitung dengan menggunakan persamaan :

$$tn = \frac{p}{P+q} \quad \text{.....(2)}$$

c. tingkat negatif palsu (fn) adalah proporsi kasus positif yang salah diklasifikasikan sebagai negatif, yang dihitung dengan menggunakan persamaan :

$$fn = \frac{u}{u+v} \quad \text{.....(3)}$$

d. tingkat negatif palsu (fp) adalah proporsi kasus negatif yang salah diklasifikasikan sebagai positif, yang dihitung dengan menggunakan persamaan :

$$fp = \frac{q}{p+q} \quad \text{.....(4)}$$

e. penarikan kembali (*recall*) atau tingkat positif benar (tp) adalah proporsi kasus positif yang diklasifikasikan dengan benar, yang dihitung dengan menggunakan persamaan :

$$tp = \frac{u}{u+v} \quad \text{.....(5)}$$

f. presisi (p) adalah proporsi prediksi kasus positif yang benar, yang dihitung dengan menggunakan persamaan :

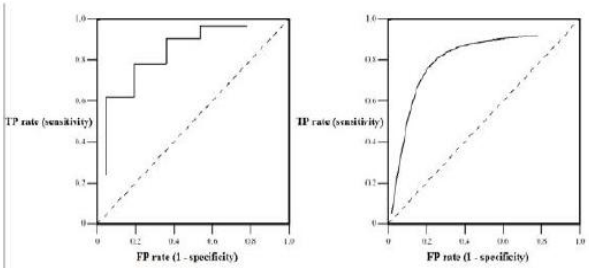
$$prc = \frac{v}{q+v} \quad \text{.....(6)}$$

f. Kurva ROC (*Receiver Operating Characteristic*)

Grafik kurva ROC (*Receiver Operating Characteristic*) digunakan untuk mengevaluasi akurasi *classifier* dan untuk membandingkan klasifikasi yang berbeda model (Vercellis, 2009). Sebuah grafik ROC adalah grafik dua dimensi dengan proporsi negatif pada sumbu *horisontal* dan proporsi positif yang benar di sumbu *vertikal* (Vercellis, 2009). Kegunaan kurva ROC (*Receiver Operating Characteristic*) adalah (Gorunescu, 2011) adalah untuk radar selama Perang Dunia II untuk mendeteksi benda-benda musuh di medan perang, teori deteksi sinyal, dalam psikologi ke rekening untuk deteksi sinyal persepsi, penelitian medis dan dalam mesin pembelajaran dan penelitian data mining, serta masalah klasifikasi.

Dalam masalah klasifikasi menggunakan kelas keputusan dua (klasifikasi biner), masing-masing objek dikelompokkan dalam (P, N), yaitu positif atau negatif. Sementara model klasifikasi beberapa (misalnya, pohon keputusan)

menghasilkan label kelas diskrit (menunjukkan hanya kelas diprediksi objek), pengklasifikasi lainnya (misalnya, *naive bayes*, jaringan saraf) menghasilkan output yang berkesinambungan, yang ambang batas yang berbeda mungkin diterapkan untuk memprediksi keanggotaan kelas, secara teknis, ROC kurva, juga dikenal sebagai grafik ROC, dua-dimensi grafik dimana tingkat TP diplot pada sumbu Y- dan tingkat FP diplot pada X- sumbu (Gorunescu, 2011).



Gambar 3. Grafik ROC (*discrete/continuous case*)

Pada gambar 3 garis diagonal membagi ruang ROC, yaitu :

1. (a) poin diatas garis diagonal merupakan hasil klasifikasi yang baik.
2. (b) point dibawah garis diagonal merupakan hasil klasifikasi yang buruk

Dapat disimpulkan bahwa, satu point pada kurva ROC adalah lebih baik dari pada yang lainnya jika arah garis melintang dari kiri bawah ke kanan atas didalam grafik.

Untuk keakurasian nilai AUC dalam klasifikasi data mining dibagi menjadi lima kelompok (Gorunescu, 2011), yaitu :

a. 0.90 – 1.00 = klasifikasi sangat baik (*excellent classification*)
b. 0.80 – 0.90 = klasifikasi baik (*good classification*)
c. 0.70 – 0.80 = klasifikasi cukup (*fair classification*)
d. 0.60 – 0.70 = klasifikasi buruk (*poor classification*)
e. 0.50 – 0.60 = klasifikasi salah (*failure*)

g. Kerangka Pemikiran

Model kerangka pemikiran yang digunakan adalah *method improvement* (perbaikan metode), yang sering digunakan pada penelitian di bidang sains dan teknik, termasuk bidang computing di dalamnya. Komponen dari model kerangka pemikiran perbaikan metode (*methode improvement*) adalah ***indicators, proposed method, objectives*** dan ***measurements*** (Polancic, 2007). Kerangka pemikiran pada penelitian ini dimulai dari prediksi hasil pemilihan umum. Maka dengan ini penulis mencoba membuat sebuah *soft computing* dengan menggunakan *Support Vector Machine* berbasis *Particle Swarm Optimization* (PSO).

Metode Penelitian

Pengertian penelitian dalam akademik yaitu digunakan untuk mengacu pada aktivitas yang rajin dan penyelidikan sistematis atau investigasi di suatu daerah, dengan tujuan menemukan atau merevisi fakta, teori, aplikasi dan tujuannya adalah untuk menemukan dan menyebarkan pengetahuan baru (Berndtsson, Olsson, & Lundell, 2008).

Menurut (Dawson, 2009) ada empat metode penelitian yang umum digunakan yaitu tindakan penelitian, eksperimen, studi kasus dan survey. Dalam konteks penelitian, metode yang dilakukan mengacu kepada pemecahan masalah yang meliputi mengumpulkan data, merumuskan hipotesis atau proposisi, pengujian hipotesis, menafsirkan hasil, dan kesimpulan (Berndtssom, Hansson, Olsson, & Lundell, 2008).

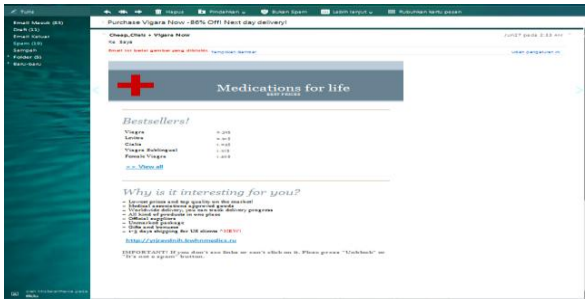
Dalam penelitian ini dilakukan beberapa langkah yang dilakukan dalam proses penelitian.

1. Pengumpulan Data
   Pada tahap ini ditentukan data yang akan diproses. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua data ke dalam data set, termasuk variabel yang diperlukan dalam proses. Data untuk eksperimen ini dikumpulkan, kemudian diseleksi dari data yang tidak sesuai.
2. Pengolahan Awal Data
   Ditahap ini dilakukan penyeleksian data, ditransformasikan kebentuk yang diinginkan sehingga dapat dilakukan persiapan dalam pembuatan model. Model dipilih berdasarkan kesesuaian data dengan metode yang paling baik dari beberapa metode pengklasifikasian yang sudah digunakan. Model yang digunakan adalah algoritma *Support Vector Machine*.
3. Metode Yang Diusulkan
   Untuk meningkatkan akurasi dari Algoritma *Support Vector Machine*, maka dilakukan penambahan dengan menambahkan metode *Particle Swarm Optimization*.
4. Eksperimen dan Pengujian Metode
   Untuk eksperimen data penelitian, penulis menggunakan *RapidMiner* 5.3 untuk mengolah data.
5. Evaluasi dan Validasi Hasil
   Evaluasi dilakukan untuk mengetahui akurasi model Algoritma *Support Vector Machine*. Validasi digunakan untuk melihat perbandingan hasil akurasi dari model yang digunakan dengan hasil yang telah ada sebelumnya. Teknik validasi yang digunakan adalah *cross validation*.

Pengumpulan Data

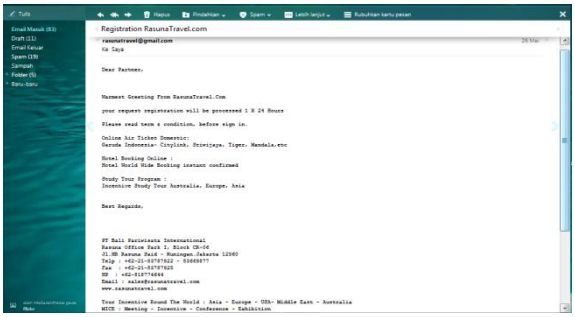
Teknik pengumpulan data ialah teknik atau cara-cara yang dapat digunakan untuk menggunakan data (Riduwan, 2008). Data yang didapat sebanyak 50 dataset. Penulis menggunakan data tersebut dari *spam* email yahoo terbaru.

Contoh data *spam* sebagai berikut :



Gambar 4. Data *Spam* Email

Sedangkan contoh data non *spam* sebagai berikut



Gambar 5. Data *Non Spam* Email

Pengolahan Data Awal

Untuk mengurangi lamanya waktu pengolahan data, penulis hanya menggunakan 25 data *spam* dan 25 data *non spam* sebagai data *training*. *Dataset* ini dalam tahap *preprocessing* harus melalui 3 proses, yaitu:

1. *Tokenization*
   Yaitu mengumpulkan semua kata yang muncul dan menghilangkan tanda baca maupun simbol apapun yang bukan huruf.
2. *Stopwords Removal*
   Yaitu penghapusan kata-kata yang tidak relevan seperti *the, of, for, with* dan sebagainya.
3. *Stemming*
   Yaitu mengelompokkan kata ke dalam beberapa kelompok yang memiliki kata dasar yang sama, seperti *drug*, *drugged* dan *drugs* di mana kata dasar dari semuanya adalah kata *drug*.

Untuk mendapatkan data yang berkualitas, beberapa teknik yang dilakukan sebagai berikut (Vecellis, 2009) :

a. Data *validation*, untuk mengidentifikasikan dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*).
b. Data *integration and transformation*, untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal. Data ditransformasikan ke dalam *software RapidMiner*. Tabel kategorikal atribut terlihat pada tabel 3.1
c. Data *size reduction and discritization*, untuk memperoleh *dataset* dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat *informative*.

Tabel 1. Tabel Atribut Yang Digunakan

No	Atribut	Nilai
1.	Data *Spam*	Berapa banyak *spam* dalam sehari yang masuk email
2.	Data *Non Spam*	Berapa banyak *non spam* dalam sehari yang masuk email

Metode yang diusulkan

Pada tahap *modeling* ini dilakukan pemprosesan data *training* sehingga akan membahas metode Algoritma yang diuji dengan memasukkan data *spam* email kemudian dianalisa dan dikomparasi.

Eksperimen dan Pengujian Metode

Tahap modeling untuk menyelesaikan prediksi *spam* email dengan menggunakan dua metode yaitu Algoritma *Support Vector Machine* dan Algoritma *Particle Swarm Optimization*.

1. *Support Vector Machine* yaitu suatu sebuah metode seleksi fitur, dan mengambil salah satu yang memiliki akurasi klasifikasi terbaik.
2. *Particle Swarm Optimization* (PSO) yaitu metode optimasi yang melakukan pencarian menggunakan populasi (*swarm*) dari individu (partikel) yang diperbaharui dari iterasi untuk iterasi dengan menyeleksi atribut yang ada.

Pada penelitian ini yang digunakan adalah penelitian *Eksperimen*. Penulis melakukan proses *Eksperimen* menggunakan aplikasi *RapidMiner* 5.3.

Evaluasi dan Validasi Hasil

Model yang diusulkan pada penelitian tentang prediksi *Spam* email adalah dengan menerapkan *Support Vector Machine* dan *Support Vector Machine* berbasis *Particle Swarm Optimization*. Penerapan algoritma *Support*

Vector Machine dengan menentukan nilai *weight* terlebih dahulu. Setelah didapatkan nilai akurasi dan AUC terbesar, nilai *weight* tersebut akan dijadikan nilai yang akan digunakan untuk mencari nilai akurasi dan AUC tertinggi. Sedangkan penerapan algoritma *Support Vector Machine* berbasis *Particle Swarm Optimization* beracuan pada nilai *weight* pada algoritma tersebut. Setelah ditemukan nilai akurasi yang paling ideal dari parameter tersebut langkah selanjutnya adalah menentukan nilai *weight*. Setelah ditemukan nilai akurasi yang paling ideal dari parameter tersebut langkah selanjutnya adalah menentukan *weight* sehingga terbentuk struktur algoritma yang ideal untuk pemecahan masalah tersebut.

HASIL DAN PEMBAHASAN

Support Vector Machine

Nilai *Training cycles* dalam penelitian ini ditentukan dengan cara melakukan uji coba memasukkan C, epsilon. Berikut ini adalah hasil dari percobaan yang telah dilakukan untuk penentuan nilai *Training cycles*:

Tabel 2. *Eksperiment* Penentuan Nilai *Training Cycles SVM*

C	Epsilon	SVM	
		accuracy	AUC
0.0	0.0	84.25%	0.915
0.6	0.6	85.25%	0.892
0.7	0.7	84.00%	0.870
0.8	0.8	83.75%	0.844
0.9	0.9	81.00%	0.823
1.0	0.0	85.75%	0.901

Hasil terbaik pada *eksperiment* SVM diatas adalah dengan C=1.0 dan *Epsilon*=0.0 dihasilkan *accuracy* 85.75% dan AUCnya 0.901 untuk SVM dengan C=1.0 dan *Epsilon*=0.0 dihasilkan *accuracy* 89.24% dan AUCnya 0.935.

Support Vector Machine berbasis *Particle Swarm Optimization*

Nilai *training cycles* dalam penelitian ini ditentukan dengan cara melakukan uji coba memasukkan C, Epsilon dan *population size*. Berikut ini adalah hasil dari percobaan yang telah dilakukan untuk penentuan nilai *training cycles* :

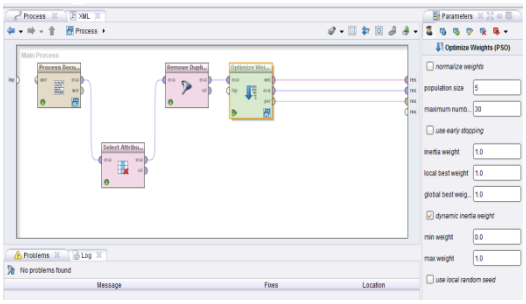
Tabel 3. *Eksperiment* penentuan nilai *training cycle* SVM berbasis PSO

C	Epsilon	SVM		Population size	SVM-PSO	
		Accuracy	AUC		Accuracy	AUC
0.0	0.0	84.25%	0.915	5	86.47%	0.916
0.6	0.6	85.25%	0.892	5	89.24%	0.935
0.7	0.7	84.00%	0.870	5	87.72%	0.920
0.8	0.8	83.75%	0.844	5	86.69%	0.907
0.9	0.9	81.00%	0.823	5	85.73%	0.891
1.0	0.0	85.75%	0.901	5	88.47%	0.928

Hasil terbaik pada *eksperiment* SVM berbasis PSO diatas adalah dengan C=1.0 dan *Epsilon*=0.0 serta *population size*=5 yang dihasilkan *accuracy* 88.47% dan AUCnya 0.928 dan dengan C=0.6 dan *Epsilon*=0.6 serta *population size*=5 untuk SVM berbasis PSO dihasilkan *accuracy* 89.24% dan AUCnya 0.935.

Evaluasi dan Validasi Hasil

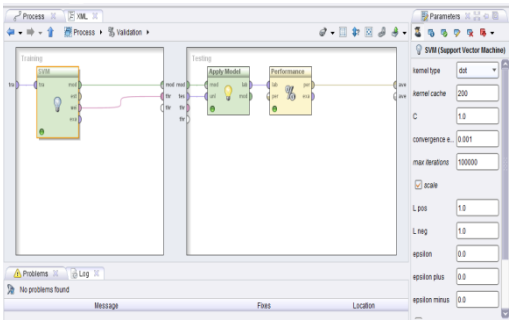
Hasil dari pengujian model yang dilakukan adalah memprediksi *spam* email dengan *Support Vector Machine* dan *Support Vector Machine* berbasis *Particle Swarm Optimization* untuk menentukan nilai *accuracy* dan AUC. Dalam menentukan nilai tingkat keakurasian dalam model *Support Vector Machine* dan Algoritma *neural network* berbasis *Particle Swarm Optimization*. Metode pengujiannya menggunakan *cross validation* dengan desain modelnya sebagai berikut :



Gambar 6. *Desain* Model *Validasi*

Hasil Pengujian Model *Support Vector Machine*

Pada penelitian penentuan hasil *spam* email menggunakan Algoritma *Support Vector Machine* berbasis pada *framework RapidMiner* sebagai berikut :



Gambar 7. Model Pengujian validasi *Support Vector Machine*

Nilai *accuracy*, *precision*, dan *recall* dari data *training* dapat dihitung dengan menggunakan *RapidMiner*. Hasil pengujian dengan menggunakan model *Support Vector Machine* didapatkan hasil pada tabel 3

1. *Confusion Matrix*

Tabel 4 data *training* yang digunakan terdiri dari 200 data *spam* dan 200 data *non spam*. Untuk data *spam*, 161 diklasifikasikan , 161 diklasifikasikan *ya* sesuai dengan prediksi yang dilakukan dengan metode SVM, dan 39 data diprediksi *ya* tetapi ternyata hasil prediksinya tidak. Untuk data *non spam*, 182 diklasifikasikan *ya* sesuai dengan prediksi yang dilakukan dengan metode SVM, dan 18 data diprediksi tidak ternyata hasil prediksinya *ya*.

Tabel 4. Model *Confusion Matrix* untuk Metode *Support Vector Machine*

Accuracy: 85.75% +/-5.01% (mikro: 85.75%)			
	True Ya	*True* Tidak	*Class precission*
Pred. Ya	161	18	89.94%
Pred. Tidak	39	182	82.35%
Class recall	80.50%	91.00%	

2. Kurva ROC

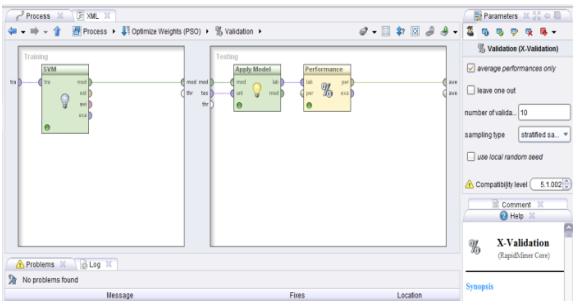
Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua metode komparasi bisa dilihat pada gambar 4.2 yang merupakan kurva ROC untuk Algoritma *Support Vector Machine*. Kurva ROC pada gambar 8 mengekspresikan *confusion matrix* dari Tabel 4. Garis horizontal adalah *false positives* dan garis vertikal *true positives*.

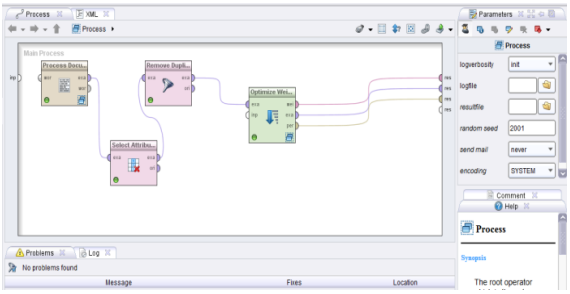
Gambar 8. Kurva ROC dengan metode *Support Vector Machine*

Dari gambar 8 terdapat grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.901 dimana diagnosa hasilnya *fair classification*.

Hasil Pengujian Model *Support Vector Machine* berbasis Algoritma *Particle Swarm Optimization* (PSO).

Pada penelitian penentuan hasil *spam* email menggunakan Algoritma *Support Vector Machine* berbasis *Particle Swarm Optimization* (PSO) pada *framework RapidMiner* sebagai berikut :





Gambar 9. Model Pengujian validasi *Support Vector Machine* berbasis *Particle Swarm Optimization* (PSO)

Nilai *accuracy*, *precision*, dan *recall* dari data *training* dapat dihitung dengan menggunakan *RapidMiner*. Hasil pengujian dengan menggunakan model *Support Vector Machine* didapatkan hasil pada tabel 5

1. *Confusion matrix*

Tabel 4.5 data *training* yang digunakan terdiri dari 200 data *spam* dan 200 data *non spam*. Untuk data spam, 171 diklasifikasikan *ya* sesuai dengan prediksi yang dilakukan dengan metode SVM, dan 29 data diprediksi *ya* tetapi ternyata hasil prediksinya tidak. Untuk data non spam, 185 diklasifikasikan *ya* sesuai dengan prediksi yang dilakukan dengan metode SVM, dan 14 data diprediksi tidak ternyata hasil prediksinya *ya*.

Tabel 5. Model *Confusion Matrix* untuk metode *Support Vector Machine* berbasis *Particle Swarm Optimization*

Accuracy: 89.24% +/-4.76% (mikro: 89.24%)			
	True Ya	*True* Tidak	*Class Precission*
Pred. Ya	171	14	92.43%
Pred. Tidak	29	185	86.45%
Class Recall	85.50%	92.96%	

2. Kurva ROC

Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua metode *komparasi* bisa dilihat pada Gambar 4.4 yang merupakan kurva ROC untuk Algoritma *Support Vector Machine* berbasis *Particle Swarm Optimization* (PSO). Kurva ROC pada gambar 9 mengekspresikan *confusion matrix* dari Tabel 5 Garis *horizontal* adalah *false positives* dan garis vertikal *true positives.*

Gambar 10. Kurva ROC dengan Metode *Support Vector Machine* berbasis *Particle Swarm Optimization* (PSO)

Dari gambar 10 terdapat grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.928 dimana *diagnosa* hasilnya *fair classification*.

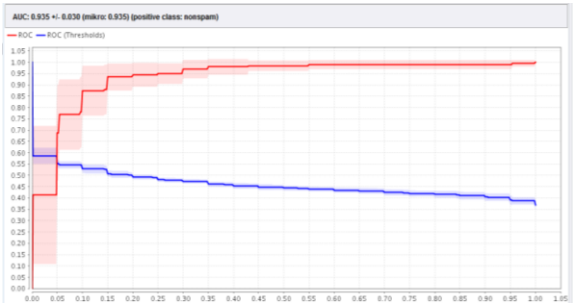
Analisis Evaluasi dan Validasi Model

Dari hasil pengujian di atas, baik evaluasi menggunakan *confusion matrix* mampu ROC *curve* terbukti bahwa hasil pengujian Algoritma *Support Vector Machine* berbasis *Particle Swarm Optimization* memiliki nilai akurasi yang lebih tinggi dibandingkan dengan Algoritma *Support Vector Machine* nilai akurasi untuk model Algoritma *Support Vector Machine* sebesar 85.75% dan nilai akurasi untuk model Algoritma *Support Vector Machine* berbasis *Particle Swarm O* sebesar 89.24% dengan selisih akurasi 3.49% dapat dilihat pada tabel 6 dibawah ini:

Tabel 6. Pengujian Algoritma SVM dan SVM berbasis PSO

	Accuracy	AUC
SVM	85.75%	0.901
SVM berbasis PSO	89.24%	0.935

Untuk evaluasi menggunakan ROC *curve* sehingga menghasilkan nilai AUC (*Area Under Curve*) untuk model Algoritma SVM menghasilkan nilai 0.901 dengan nilai *diagnosa Fair Classification*, sedangkan untuk Algoritma *Support Vector Machine* (SVM) berbasis *Particle Swarm Optimization* (PSO) menghasilkan nilai 0.935 dengan nilai diagnosa *Fair Classification*, dan selisih nilai keduanya sebesar 0.034 dapat dilihat pada Gambar 11 dibawah ini



Gambar 11. Kurva ROC dengan *Support Vector Machine* dan *Support Vector Machine* berbasis *Particle Swarm Optimization*

Dengan demikian Algoritma *Support Vector Machine* (SVM) berbasis *Particle Swarm Optimization* (PSO) dapat memberikan solusi untuk permasalahan dalam prediksi *spam* email. Untuk rinciannya dapat dilihat pada tabel 6. dan Gambar 11.

Berdasarkan hasil *eksperiment* yang dilakukan untuk memecahkan masalah prediksi *spam* email, dapat disimpulkan bahwa hasil *eksperiment* menggunakan metode *Support Vector Machine* mempunyai tingkat akurasi sebesar 84.25% dan mempunyai nilai AUC sebesar 0.915. setelah dilakukan penyesuaian pada parameter C dan *epsilon* didapat nilai akurasi terbaik untuk Algoritma *Support Vector Machine* yaitu mempunyai akurasi sebesar 85.75% dan nilai AUCnya sebesar 0.901. Sedangkan *eksperiment* kedua yang dilakukan dengan menggunakan metode *Support Vector Machine* berbasis *Particle Swarm Optimization* mempunyai nilai akurasi sebesar 88.47% dan nilai AUC sebesar 0.928. Setelah dilakukan penyesuaian pada parameter C dan *epsilon* dan *population* didapat nilai akurasi terbaik untuk Algoritma *Support Vector Machine* berbasis *Particle Swarm Optimization* yaitu mempunyai akurasi sebesar 89.24% dan nilai AUC sebesar 0.935.

Implikasi Penelitian

Implikasi penelitian diarahkan pada tiga aspek, yaitu:

1. Aspek Sistem

Penerapan kebijakan dalam penentuan prediksi *spam* email akan membawa pengaruh pada sistem, dimana melalui data-data yang ada di email akan mempermudah dalam menentukan *spam* dan *non spam* email. Dalam penerapannya dapat menggunakan *software* anti-*spam*: "*spam assasin*", dimana *software* ini merupakan *software* terbaik. Dengan memanfaatkan perangkat yang ada, seperti *hardware* pada jaringan, sistem akan berjalan dengan lebih baik dalam memecahkan permasalahan yang ada.

2. Aspek Manajerial

Secara manajerial kinerja *software* akan lebih membawa dampak yang signifikan dalam memprediksi *spam* email. Selain itu penerapannya dapat dilakukan yaitu melindungi semua akun email di PC, melindungi terhadap "*phishing*", pencurian identitas dan penipuan email lainnya, blacklist dan blok domains dan email serta pelaporan *spam* dengan satu kali klik.

3. Penelitian Lanjutan

Penelitian semacam ini dapat dikembangkan pada bidang informasi. Penelitian ini juga dapat dikembangkan dengan Algoritma aturan klasifikasi yang lain, seperti Algoritma *Naïve Bayes*, *kNN.*

KESIMPULAN

Untuk mengklasifikasikan *spam* dengan data berupa *spam* email, salah satu pengklasifikasi yang dapat digunakan adalah pengklasifikasi *Support Vector Machine*. Hal ini dikarenakan *Support Vector Machine* sangat sederhana dan efisien. Selain itu *Support Vector Machine* juga sangat populer digunakan untuk prediksi *spam* dan memiliki performa yang baik pada banyak domain.

Dari pengolahan data yang sudah dilakukan, penggabungan metode *Support Vector Machine* dan *Particle Swarm Optimization*, terbukti dapat meningkatkan akurasi pengklasifikasi *Support Vector Machine*. Data email dapat diklasifikasi dengan baik ke dalam bentuk *spam* dan *non spam*. Akurasi *Support Vector Machine* sebelum menggunakan metode *Particle Swarm Optimization* mencapai 85.75%. Sedangkan setelah menggunakan penggabungan metode *Particle Swarm Optimization* akurasinya meningkat hingga mencapai 89.24%. Peningkatan akurasi mencapai 4%. Untuk mendukung penelitian, penulis menggunakan *software* anti-spam "*spam assasin*" untuk memprediksi *spam* email dan non *spam*. Model yang terbentuk dapat diterapkan pada seluruh data email, sehingga dapat dilihat secara langsung hasilnya dalam bentuk *spam* dan *non spam*. Hal ini dapat membantu seseorang untuk menghemat waktu saat memprediksi email tanpa harus mengkhawatirkan adanya *spoiler* dan pemberian *rating* yang tidak sesuai dengan *review*nya.

Walaupun pengklasifikasi *Support Vector Machine* sudah sering digunakan dan mempunyai performa yang baik dalam pengklasifikasi teks, namun ada beberapa hal yang dapat ditambahkan untuk penelitian selanjutnya:

1. Menggunakan metode yang lain, seperti metode email *filtering* dan lain-lain agar hasilnya bisa dibandingkan dengan metode yang sudah umum digunakan. Baik penggunaan metode-metode ataupun digabung.
2. Menggunakan pengklasifikasi lain yang mungkin di luar *supervised learning*. Sehingga bisa dilakukan penelitian yang berbeda dari umumnya yang sudah ada.
3. Menggunakan data review dari domain yang berbeda, misalnya review SMS *spam* dan lain sebagainya.

REFERENSI

Abraham, A., Grosan, C., & Ramos, V. (2006). *Swarm Intelligence In Data Mining.* Verlag Berlin Heidelberg: Springer.

Alpaydın, E. (2010). *Introduction To Machine Learning.* London: Massachusetts Institute Of Technology.

Bellazzi, R., & Zupanb, B. (2008). Predictive Data Mining In Clinical Medicine: Current Issues And And Guidelines. *International Journal Of Medical Informatics 7 7* , 81–97.

Berndtssom, M., Hansson, J., Olsson, B., & Lundell, B. (2008). *A Guide For Students In Computer Science And Information Systems.* London: Springer.

Bramer, M. (2007). *Principles Of Data Mining.* Verlag London: Springer.

Burges, C. J. (1998). *A Tutorial On Support Vector Machines For Pattern Recognition.* Boston: Kluwer Academic Publishers.

Dawson, C. W. (2009). *Projects In Computing And Information System A Student's Guide.* England: Addison-Wesley.

Dong, Y., Xia, Z., Tu, M., & Xing, G. (2007). An Optimization Method For Selecting Parameters In Support Vector Machines. *Sixth International Conference On Machine Learning And Applications* , 1.

Fei, S. W., Miao, Y. B., & Liu, C. L. (2009). Chinese Grain Production Forecasting Method Based On Particle Swarm Optimization-Based Support Vector Machine. *Recent Patents On Engineering 2009 , 3*, 8-12.

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

G. Sudipto, M. Adam, M. Nina, M. Rajeev, O. Liadan, 2003, *Clustering Data Streams*: *Theory and Practice*, Radical Eye Software, pages 1-4.

Han, J., & Kamber, M. (2007). *Data Mining Concepts and Techniques*.

Haupt, R. L., & Haupt, S. E. (2004). *Practical Genetic Algorithms.* Untied States Of America: A John Wiley & Sons Inc Publication.

https://nic.itb.ac.id/mail/spam-dan-kiat-mengatasinya

http://sinaga17.wordpress.com/2014/02/26/software-anti-spamming-terbaik/

Huang, K., Yang, H., King, I., & Lyu, M. (2008). *Machine Learning Modeling Data Locally And Globally.* Berlin Heidelberg: Zhejiang University Press, Hangzhou And Springer-Verlag Gmbh.

Larose, D. T. (2007). *Data Mining Methods And Models.* New Jersey: A John Wiley & Sons.

Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, S. (2011). An Improved Particle Swarm Optimization for Feature Selection. *Journal of Bionic Engineering Vol 8* , 1-10.

Maimon, O. (2010). *Data Mining And Knowledge Discovery Handbook.* New York Dordrecht Heidelberg London: Springer.

Moertini, V. S. (2002). Data Mining Sebagai Solusi Bisnis. *Integral, Vol. 7 No. 1, April 2002* , 4.

Nugroho, A. S. (2008). Support Vector Machine: Paradigma Baru Dalam Softcomputing. *Konferensi Nasional Sistem Dan Informatika,* 92-99.

Parimala, R., & Nallaswamy, R. (2012). Feature Selection using a Novel Particle Swarm Optimization and It's Variants. *I.J. Information Technology and Computer Science,* , 16-24.

Ren, Qinqing. (2010). Feature-Fusion Framework for Spam Filtering Based on SVM.

Shukla, A., Tiwari, R., & Kala, R. (2010). *Real Life Applications Of Soft Computing.* New York: Taylor & Francis Group.

techterms@whatis.com,what is *spam filter*?, http://searchmidmarketsecurity.techtarget.com/sDefinition/0,,sid198_gci931766,00.html,midmarket IT security definition, diakses pada tanggal 20 oktober 2009.

Vercellis, C. (2009). *Business Intelligence Data Mining And Optimization For Decision Making* . United Kingdom: A John Wiley And Sons, Ltd., Publication.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals Of Predictive Text Mining.* London: Springer.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools And Techniques.* Burlington, Usa: Morgan Kaufmann Publishers.

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. *School Of Computer Science And Technology* , 5197–5204.

BIODATA PENULIS

Eni Pudjiarti, M.Kom. Lahir di Jakarta 4 Oktober 1985. Sebagai Dosen Tetap di AMIK BSI Jakarta dan STMIK Nusa Mandiri Jakarta. Tahun 2008 Lulus Program Diploma Tiga (DIII) dari AMIK BSI Jakarta Program Studi Komputerisasi Akuntansi. Tahun 2011 Lulus S1 dari STMIK Nusa Mandiri Jakarta Jurusan Sistem Informasi. Tahun 2014 Lulus S2 dari Pasca Sarjana Magister Ilmu Komputer STMIK Nusa Mandiri Jakarta Konsentrasi *Management Information System*.