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

TESIS

ENI PUDJIARTI
14000531

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
NUSA MANDIRI
JAKARTA
2014

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

TESIS

Diajukan sebagai salah satu syarat untuk memperoleh gelar
Magister Ilmu Komputer (M.Kom)

ENI PUDJIARTI
14000531

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
NUSA MANDIRI
JAKARTA
2014

SURAT PERNYATAAN ORISINALITAS

Yang bertanda tangan di bawah ini :

Nama	: Eni Pudjiarti
NIM	14000531
Program Studi	: Magister Ilmu Komputer
Jenjang	: Strata Dua (S2)
Konsentrasi	: Management Information System

Dengan ini menyatakan bahwa tesis yang telah saya buat dengan judul: "Prediksi *Spam* Email Menggunakan Metode *Support Vector Machine* Dan *Particle Swarm Optimization*" adalah hasil karya sendiri, dan semua sumber baik yang kutip maupun yang dirujuk telah saya nyatakan dengan benar dan tesis belum pernah diterbitkan atau dipublikasikan dimanapun dan dalam bentuk apapun.

Demikianlah surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila dikemudian hari ternyata saya memberikan keterangan palsu dan atau ada pihak lain yang mengklaim bahwa tesis yang telah saya buat adalah hasil karya milik seseorang atau badan tertentu, saya bersedia diproses baik secara pidana maupun perdata dan kelulusan saya dari Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri dicabut/dibatalkan.

Jakarta, 8 Juli 2014
Yang menyatakan,

Materai Rp. 6.000,-

Eni Pudjiarti

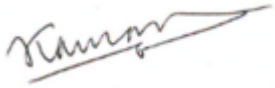
HALAMAN PENGESAHAN

Tesis ini diajukan oleh :

Nama	: Eni Pudjiarti
NIM	14000531
Program Studi	: Magister Ilmu Komputer
Jenjang	: Strata Dua (S2)
Konsentrasi	: *Management Information System*
Judul Tesis	: "Prediksi *Spam* Email Menggunakan Metode *Support Vector Machine* dan *Particle Swarm Optimization*"

Telah berhasil dipertahankan dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Magister Ilmu Komputer (M.Kom) pada Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri).

Jakarta, 8 Juli 2014
Pascasarjana Magister Ilmu Komputer
STMIK Nusa Mandiri
Direktur

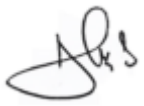


Prof. Dr. Ir. Kaman Nainggolan, MS

D E W A N P E N G U J I

Penguji I : Dr. Sularso Budilaksono

Penguji II : Windu Gata, M.Kom



Penguji III / Pembimbing : Dr. Rudi Lumanto

KATA PENGANTAR

Puji syukur alhamdulillah, penulis panjatkan kehadirat Allah SWT, yang telah melimpahkan rahmat dan karunia-Nya, sehingga pada akhirnya penulis dapat menyelesaikan tesis ini tepat pada waktunya. Dimana tesis ini penulis sajikan dalam bentuk buku yang sederhana. Adapun judul tesis, yang penulis ambil sebagai berikut "Prediksi *Spam* Email Menggunakan Metode *Support Vector Machine* dan *Particle Swarm Optimization*".

Tujuan penulisan tesis ini dibuat sebagai salah satu untuk mendapatkan gelar Magister Ilmu Komputer (M.Kom) pada Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (PPs MIK STMIK Nusa Mandiri).

Tesis ini diambil berdasarkan hasil penelitian atau riset mengenai prediksi *Spam* Email Menggunakan *Support Vector Machine* yang penulis lakukan menggunakan data publik yang sudah banyak diteliti oleh peneliti sebelumnya. Penulis juga lakukan mencari dan menganalisa berbagai macam sumber referensi, baik dalam bentuk jurnal ilmiah, buku-buku literatur, *internet*, dan lain-lain yang terkait dengan pembahasan pada tesis ini.

Penulis menyadari bahwa tanpa bimbingan dan dukungan dari semua pihak dalam pembuatan tesis ini, maka penulis tidak dapat menyelesaikan tesis ini tepat pada waktunya. Untuk itu ijinkanlah penulis kesempatan ini untuk mengucapkan ucapan terima kasih yang sebesar-besarnya kepada :

1. Allah SWT yang selalu mencurahkan nikmat dan rahmatNya pada saya sehingga saya dapat menyelesaikan tesis ini tepat pada waktunya.
2. Bapak Dr. Rudi Lumanto selaku pembimbing tesis yang telah menyediakan waktu, pikiran dan tenaga dalam membimbing penulis dalam menyelesaikan tesis ini.
3. Suami tercinta Zulfikar Indra Saputra, S.Kom yang telah memberikan dukungan dan semangat dalam menyelesaikan tesis.
4. Adinarto dan Suparti (Orang Tua), Bakri dan Andang Wiyati (Mertua), Rinda Sulastin (Adik), Faisal Rahman (Adik), Habibah Tri Rahayu (Adik) yang selalu memberikan dukungan moral dan spiritual kepada penulis.

5. Bapak Heri Kuswara M.Kom (Kepala Bagian BSI Career Center) dan Bapak Muhammad Islahudin (Kepala Bagian BSI Entrepreneur Center) yang selalu memberikan semangat dan dukungan.
6. Seluruh Staf BSI Career Center dan BSI Entrepreneur Center yang telah memberikan dukungan dan semangat.
7. Dinda Ayu Muthia, M.Kom yang telah memberikan bantuan dan dukungannya.
8. Seluruh staf pengajar (dosen) PPs MIK STMIK Nusa Mandiri yang telah memberikan pelajaran yang berarti bagi penulis selama menempuh studi.
9. Seluruh staf dan karyawan PPs MIK STMIK Nusa Mandiri yang telah melayani penulis dengan baik selama kuliah.
10. Dll

Serta semua pihak yang terlalu banyak untuk penulis sebutkan satu persatu sehingga terwujudnya penulisan tesis ini. Penulis menyadari bahwa penulisan tesis ini masih jauh sekali dari sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan karya ilmiah yang penulis hasilkan untuk yang akan datang.

Akhir kata semoga tesis ini dapat bermanfaat bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Jakarta, 8 Juli 2014

Eni Pudjiarti

Penulis

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan di bawah ini, saya :

Nama	: Eni Pudjiarti
NIM	14000531
Program Studi	: Magister Ilmu Komputer
Jenjang	: Strata Dua (S2)
Konsentrasi	: *Management Information System* (MIS)
Jenis Karya	: Tesis

Demi pengembangan ilmu pengetahuan, dengan ini menyetujui untuk memberikan ijin kepada pihak Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri) **Hak Bebas Royalti Non-Eksklusif (*Non-exclusive Royalti-Free Right*)** atas karya ilmiah kami yang berjudul : "Prediksi *Spam* Email Menggunakan Metode *Support Vector Machine* dan *Particle Swarm Optimization*" beserta perangkat yang diperlukan (apabila ada).

Dengan **Hak Bebas Royalti Non-Eksklusif** ini pihak STMIK Nusa Mandiri berhak menyimpan, mengalih-media atau *bentuk*-kan, mengelolaannya dalam pangkalan data (*database*), mendistribusikannya dan menampilkan atau mempublikasikannya di *internet* atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari kami selama tetap mencantumkan nama kami sebagai penulis/pencipta karya ilmiah tersebut.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak STMIK Nusa Mandiri, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 8 Juli 2014
Yang menyatakan,

Materai Rp. 6.000,-

Eni Pudjiarti

ABSTRAK

Nama : Eni Pudjiarti
NIM 14000531
Program Studi : Magister Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : *Management Information System* (MIS)

Judul : "Prediksi *Spam* Email Menggunakan Metode *Support Vector Machine* dan *Particle Swarm Optimization"*

Spam Email adalah istilah yang digunakan untuk menggambarkan pesan terkirim dalam email massal atau email masuk yang diterima tanpa persetujuan. *Spam* Email *Filtering* adalah suatu program yang digunakan untuk mendeteksi email yang tidak diinginkan dan tidak diminta dan mencegah email tersebut untuk masuk ke dalam kotak masuk pengguna email. Banyak teknik yang digunakan untuk membuat *spam* email *filtering*, salah satunya dengan menggunakan teknik klasifikasi. Pengklasifikasi *Support Vector Machine* adalah suatu metode *supervised learning* yang digunakan untuk melakukan klasifikasi data. Tetapi *Support Vector Machine* memiliki kelemahan pada sulitnya pemilihan fitur yang sesuai dan optimal pada bobot atribut yang digunakan sehingga menyebabkan tingkat akurasi prediksi menjadi rendah. Oleh karena itu, dalam penelitian ini dibuatkan model algoritma *Support Vector Machine* dan model algoritma *Support Vector Machine* berbasis *Particle Swarm Optimization* untuk mendapatkan rule dalam memprediksi *spam* email dan memberikan nilai akurasi yang lebih akurat. Setelah dilakukan pengujian dengan dua model yaitu Algoritma *Support Vector Machine* dan *Support Vector Machine* berbasis *Particle Swarm Optimization* maka hasil yang didapat adalah algoritma sehingga didapat pengujian dengan menggunakan *Support Vector Machine* dimana didapat nilai *accuracy* adalah 85.75% dan nilai AUC adalah 0.901, sedangkan pengujian dengan menggunakan *Support Vector Machine* berbasis *Particle Swarm Optimization* didapatkan nilai *accuracy* 89.24% dan nilai AUC adalah 0.935 dengan tingkat diagnosa *good classification*. Sehingga kedua metode tersebut memiliki perbedaan tingkat akurasi yaitu sebesar 3.49% dan perbedaan nilai AUC sebesar 0.034.

Kata kunci:
Spam email, Algoritma *Support Vector Machine*, seleksi atribut, *Particle Swarm Optimization*

ABSTRACT

Nama	: Eni Pudjiarti
NIM	14000531
Program Studi	: Magister Ilmu Komputer
Jenjang	: Strata Dua (S2)
Konsentrasi	: *Management Information System* (MIS)
Judul	: "*Email Spam Prediction Method Using Support Vector Machine and Particle Swarm Optimizatio"*

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.

Keywords:
Spam Email, Support Vector Machine Algorithm, Feature Selection, Particle Swarm Optimization

DAFTAR ISI

Halaman

HALAMAN SAMPUL i
HALAMAN JUDUL ii
HALAMAN PERNYATAAN ORISINALITAS iii
HALAMAN PENGESAHAN iv
KATA PENGANTAR v
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS vii
ABSTRAK viii
ABSTACT ix
DAFTAR ISI x
DAFTAR TABEL xii
DAFTAR GAMBAR xiii
DAFTAR LAMPIRAN xiv
BAB I PENDAHULUAN 1
1.1 Latar Belakang Penulisan 1
1.2 Identifikasi Masalah 2
1.3 Rumusan Masalah 2
1.4 Tujuan Penelitian 2
1.5 Manfaat Penelitian 3
1.6 Ruang Lingkup Penelitian 3
1.7 Sistematika Penulisan 3
BAB II LANDASAN TEORI 5
2.1 Tinjauan Studi 5
2.1.1 Model Penelitian Mahsus 5
2.1.2 Model Penelitian Hsu 6
2.1.3 Model Penelitian Ren 8
2.2 Tinjauan Pustaka 10
2.2.1 *Spam* Email 10
2.2.2 *Support Vector Machine* 13

2.2.3 *Particle Swarm Optimization* 18
2.2.4 Pengujian K-Fold Cross Validation 20
2.2.5 *Confution Matrix* 20
2.2.6 Kurva ROC 22
2.3 Kerangka Pemikiran 23
BAB III METODOLOGI PENELITIAN 25
3.1 Desain Penelitian 25
3.2 Pengumpulan Data 26
3.3 Pengolahan Data Awal 27
3.4 Metode yang Diusulkan 28
3.5 Eksperimen dan Pengujian Metode 30
3.6 Evaluasi dan Validasi Hasil 30
BAB IV HASIL PENELITIAN DAN PEMBAHASAN 31
4.1 Hasil Eksperimen dan Pengujian Metode 31
4.1.1 Support Vector Machine 31
4.1.2 Support Vector Machine berbasis Berbasis Particle Swarm Optimization 31
4.2 Evaluasi dan Validasi Hasil 32
4.2.1 Hasil Pengujian Support Vector Machine 32
4.2.2 Hasil Pengujian Support Vector Machine Machine Berbasis Particle Swarm Optimization 35
4.2.3 Analisis Hasil Pengujian Support Vector Machine 38
4.3 Pembahasan 39
4.3 Implikasi Penelitian 40
BAB V PENUTUP 41
5.1 Kesimpulan 41
5.2 Saran 41
DAFTAR REFERENSI 42
LAMPIRAN 47

DAFTAR TABEL

Tabel 2.1 Perbandingan Penelitian Terkait 9
Tabel 2.2 AND Problem 18
Tabel 2.3 Model *Confusion Matrix* 21
Tabel 3.1 Tabel atribut yang digunakan 28
Tabel 3.2 Spesifikasi *hardware* dan *software* 30
Tabel 4.1 Eksperiment penentuan nilai *training cycle* SVM 31
Tabel 4.2 Eksperiment penentuan nilai *training cycle* SVM berbasis PSO 31
Tabel 4.3 Model *Confusion Matrix* untuk Metode *Support Vector Machine* 33
Tabel 4.4 Nilai *accuracy, sensitivity, specificity, ppv*, dan *npv* untuk Metode *Support Vector Machine* 34
Tabel 4.5 Model *Confusion Matrix* untuk Metode *Support Vector Machine* berbasis *Particle Swarm Optimization* 36
Tabel 4.6 Nilai *accuracy, sensitivity, specificity, ppv*, dan *npv* untuk Metode *Support Vector Machine* berbasis *Particle Swarm Optimization* 37
Tabel 4.7 Pengujian algoritma SVM dan SVM berbasis PSO 38

DAFTAR GAMBAR

Gambar 2.1 Bagan GSVM-DC 6
Gambar 2.2 Data *Preprocessing* 7
Gambar 2.3 Feature-Fusion framework for spam filtering 8
Gambar 2.4 Konsep SVM untuk mencari *hyperplane* terbaik 15
Gambar 2.5 Kernel map 17
Gambar 2.6 K-*fold cross-validation* 20
Gambar 2.7 Grafik ROC (discrete dan continous) 22
Gambar 2.8 Kerangka Pemikiran 24
Gambar 3.1 Metode yang diusulkan 29
Gambar 4.1 Desain Model Validasi 32
Gambar 4.2 Model Pengujian validasi *Support Vector Machine* 33
Gambar 4.3 Kurva ROC dengan Metode *Support Vector Machines* 35
Gambar 4.4 Model pengujian validasi *support vector machine* berbasis algoritma *Particle Swarm Optimization* 36
Gambar 4.5 Kurva ROC dengan *support vector machine* berbasis *Particle Swarm Optimization* 38
Gambar 4.6 Kurva ROC dengan *support vector machine* dan *support vector machine* berbasis *Particle Swarm Optimization* 39

DATA LAMPIRAN

Lampiran 1 Data *Spam* Email.....47
Lampiran 2 Data non *Spam* Email 53

BAB I
PENDAHULUAN

1.1. Latar Belakang Masalah

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.

1.2. Identifikasi Masalah

Support Vector Machine dapat menyelesaikan masalah *decision tree* khususnya sampel data yang kecil yang ada di email, tetapi *Support Vector Machine* memiliki kelemahan pada sulitnya pemilihan fitur yang sesuai dan optimal pada bobot atribut yang digunakan sehingga menyebabkan tingkat akurasi prediksi menjadi rendah.

1.3. Rumusan Masalah

Rumusan masalah yang diangkat pada penelitian ini adalah seberapa tinggi akurasi metode *Support Vector Machine* apabila *Particle Swarm Optimization* (PSO) diterapkan pada proses pemilihan parameter yang optimal dengan cara melakukan seleksi atribut untuk prediksi *spam* email?

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah menerapkan *Particle Swarm Optimization* (PSO) untuk pemilihan parameter yang sesuai (C, gamma dan epsilon) pada *Support Vector Machine* (SVM), sehingga hasil prediksinya lebih akurat.

1.5. Manfaat Penelitian

1. Manfaat penelitian ini adalah membantu para user untuk mengetahui apakah email yang diterima tersebut adalah *spam* atau bukan.
2. Manfaat teoritis dari penelitian ini yaitu diharapkan dapat memberikan sumbangan untuk pengembangan teori yang berkaitan dengan penerapan *Particle Swarm Optimization* pada *Support Vector Machine* untuk meningkatkan akurasi prediksi *spam* email.

1.6. Ruang Lingkup Penelitian

Ruang lingkup pembahasan dalam penelitian ini dibatasi pada perbandingan metode *Support Vector Machine* dengan *Support Vector Machine* berbasis *Particle swarm optimization* (PSO) dalam memprediksi *spam* email.

1.7. Sistematika Penulisan

Disajikan dalam lima bab dan masing-masing bab terdiri dari beberapa sub bab yaitu sebagai berikut:

Bab I Pendahuluan

Bab ini membahas tentang latar belakang penulisan, identifikasi permasalahan, rumusan masalah, tujuan penelitian, ruang lingkup penelitian

Bab II Landasan Teori

Bab ini membahas tentang landasan teori yang melandasi penelitian.

Bab III Metode Penelitian

Bab ini berisi tentang metode penelitian yang membahas tentang perancangan penelitian, tahap *computing approach* dan pengembangan sistem.

Bab IV Hasil dan Pembahasan

Bab ini berisi tentang hasil dan pembahasan yang menguraikan tentang implementasi sistem, pengukuran serta implikasi penelitian.

Bab V Kesimpulan dan Saran

Bab ini membahas kesimpulan dari penelitian dan saran untuk penelitian selanjutnya.

BAB II
LANDASAN/KERANGKA PEMIKIRAN

2.1. Tinjauan Studi

Ada beberapa penelitian yang menggunakan *Support Vector Machine* sebagai pengklasifikasi dalam klasifikasi *spam* email, diantaranya :

2.1.1. Model Penelitian Mahsus

Penelitian yang dilakukan oleh Moh. Mahsus. ZK, Abdurahman Baizal, Ssi, M.Kom, Shaufiah, ST, MT mengenai *spam* email. *Spam* email adalah pembanjiran internet dengan banyak salinan pesan yang sama, dan memaksa pengguna internet untuk menerimanya walaupun itu tidak diinginkan. Pengguna email membutuhkan waktu yang lebih lama untuk membaca dan memutuskan apakah email yang diterima tersebut adalah *spam* atau bukan.berdasarkan motivasi tersebut email *spam* filtering banyak dikembangkan. Banyak teknik yang digunakan untuk pembangunan sebuah *spam* email filtering, dan salah satunya dengan menggunakan metode *Support Vector Machine* karena terbukti memiliki kemampuan generalisasi yang baik. Untuk meningkatkan akurasi dan efisiensi pemrosesan, SVM dimodifikasi dengan paradigma *granular computing* dengan memecah data menjadi bagian-bagian kecil informasi serta metode *data cleaning* yang digunakan untuk mengurangi jumlah data yang akan dilatih. *Dataset* yang digunakan adalah *dataset* EMCL PKDD 2006 *Discovery Challenge Task A* dengan rincian *data training* sebanyak 4000 data dan data *testing* sebanyak 1000 data. *Dataset* yang digunakan sebelumnya telah melalui proses *preprocessing* untuk dilakukan ekstraksi dan normalisasi. Penelitian ini bertujuan untuk merancang dan membangun sistem *spam* email *filtering* menggunakan *Granular Support Vector Machines with Data Cleaning*, serta melakukan analisa akurasi dan efisiensi waktu pada sistem yang dibangun dengan membandingkan ukuran *F-Measure* dan waktu pemrosesan dengan SVM yang tidak dimodifikasi. Selain itu juga akan dilakukan analisa jumlah *support vector* yang ditemukan dengan akurasi pengujian terhadap waktu *training* pada GSVM-DC.

2.1.2. Model Penelitian Hsu

Penelitian yang dilakukan oleh Chih Hsu dan Ying Yu mengenai penyaringan *spam* email. *Support Vector Machine* adalah alat untuk membangun model penyaringan *spam* yang baik. Namun, kinerja dari model ini tergantung pada pemilihan parameter. Seleksi parameter SVM akan mempengaruhi klasifikasi kinerja selama proses pelatihan. Dalam studi ini, kami menggunakan metode Taguchi gabungan dan metode staelin untuk mengoptimalkan SVM dalam penyaringan *spam* email. Hasil menunjukkan bahwa metode yang diusulkan dapat menghasilkan nilai akurasi tinggi. Dalam pengujian ini enam *dataset* telah digunakan. Masing-masing berisi pesan non *spam* dan pesan *spam* diambil dari 500 pesan *spam* dan 500 pesan *non spam*.

2.1.3. Model Penelitian Ren

Penelitian yang dilakukan oleh Qinqing Ren mengusulkan kerangka novel berdasarkan support vector machine (SVM): kerangka fitur-fusion. SVM terbukti menjadi metode yang bagus untuk menyaring spam di koleksi teks berskala besar dengan state-of-the-art performance oleh banyak peneliti di dunia. Ekstraksi fitur adalah langkah kunci untuk klasifikasi SVM. Fitur-fusion kerangka berdasarkan SVM adalah metode tingkat fitur dan memiliki kelebihan untuk memecahkan beberapa masalah praktis penyaringan spam dan mengarah pada kinerja tinggi, terutama mengurangi biaya komputasi. Metode fitur-fusion diusulkan dalam makalah ini dapat memadukan informasi dari kata-kata dalam email dengan menggunakan rumus TFIDF dan kata-kata dalam kolom header subyek email dengan meningkatkan berat, dan meningkatkan kinerja penyaringan spam. Hasil penelitian menunjukkan bahwa kerangka fitur-fusion melebihi SVM murni dan metode SVM berbasis lainnya spam skala besar penyaringan dalam hal kurva ROC.

2.1.4. Rangkuman Penelitian Terkait

Ketiga penelitian terkait memiliki model yang sama, pengklasifikasi *support vector machine* memiliki akurasi yang paling tinggi diantara pengklasifikasi lainnya. *Particle Swarm Optimization* dapat digunakan sebagai

metode pemilihan fitur agar menghasilkan akurasi *support vector machine* yang lebih tinggi. Dari tinjauan studi di atas, dapat diketahui bahwa *support vector machine* merupakan pengklasifikasi terbaik untuk memecahkan masalah prediksi *spam* email. Pada tesis ini, penulis menggunakan pengklasifikasi *support vector machine* dan *Particle Swarm Optimization* untuk meningkatkan tingkat akurasi dari pengklasifikasi.

2.2. Tinjauan Pustaka

Dalam tesis ini, penulis menggunakan referensi dari buku untuk menjelaskan pengklasifikasi *Support Vector Machine* dan metode pemilihan fitur.

2.2.1. *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*.

2.2.2. *Support Vector Machine*

Pada *Conference on Learning Theory* (COLT), Boser, Bernhard, Guyon, dan Vapnik tahun 1992, memperkenalkan SVM (Premanode, Tzoumazou 2012) yaitu sebuah teknik *supervised learning* dari bidang *machine learning* yang dapat di aplikasikan kedalam clasifikasi dan regresi (Sewell, Taylor 2012). SVM merupakan salah satu teknik yang relatif baru untuk melakukan prediksi, SVM berada dalam satu kelas dengan ANN dalam hal fungsi dan kondisi permasalahan yang bisa di selesaikan(Santosa, 2007). Dalam banyak implementasi SVM memberikan hasil yang lebih baik dari ANN, dalam hal solusi yang di capai. ANN menemukan solusi berupa local optimal dimana ANN akan selalu memberikan solusi yang berbeda dari setiap training, berbeda dengan SVM yang menemukan solusi global optimal, dimana solusi akan memberikan hasil yang sama setip di jalankan.

SVM 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). SVM 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, SVM telah mencapai sukses besar dan dianggap sebagai *state-of-the-art classifier* saat ini (Huang, Yang, King, & Lyu, 2008).

Dua kelas data yang digambarkan sebagai lingkaran dan padat titik-titik yang disajikan di angka ini. Secara intuitif diamati, ada banyak keputusan *hyperplanes* yang dapat digunakan untuk memisahkan kedua kelompok data. Namun, yang digambarkan dengan angka ini dipilih sebagai yang menguntungkan memisahkan bidang, karena mengandung maksimal margin antara dua kelas. Karena itu, dalam tujuan fungsi SVM, sebuah istilah *regularization* mewakili margin muncul. Apalagi seperti yang terlihat di angka ini, hanya mereka yang penuh poin disebut mendukung vector terutama menentukan memisahkan bidang, sementara poin lain tidak memberi kontribusi untuk di semua margin. Dalam kata lain, hanya sejumlah titik penting untuk klasifikasi tujuan dalam kerangka SVM dan dengan demikian harus diambil (Huang, Yang, King, & Lyu, 2008).

Hyperplane terbaik adalah *hyperplane* yang terletak ditengah-tengah antara dua set obyek dari dua *class*. *Hyperplane* pemisah terbaik antara kedua *class* dapat ditemukan dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya. *Margin* adalah jarak antara *hyperplane* tersebut dengan *pattern* terdekat dari masing-masing *class. Pattern* yang paling dekat ini disebut sebagai *support vector* (Aydin, Karakose & Akin, 2011).

2.2.3. *Particle Swarm Optimization*

Optimasi adalah proses menyesuaikan kepada masukan atau karakteristik perangkat, proses matematis, atau percobaan untuk menemukan output minimum atau maksimum hasil. Input terdiri dari variabel, proses atau fungsi dikenal sebagai fungsi biaya, fungsi tujuan atau kemampuan fungsi dan output adalah biaya atau tujuan, jika proses adalah sebuah percobaan, kemudian variabel adalah masukan fisik untuk percobaan (Haupt & Haupt, 2004).

Particle Swarm Optimization (PSO) adalah metode pencarian penduduk, yang berasal dari penelitian untuk pergerakan organisme dari kelompok burung atau ikan, seperti algoritma genetika, *Particle Swarm Optimization* (PSO) melakukan pencarian menggunakan populasi (*swarm*) dari individu (partikel) yang diperbaharui dari iterasi untuk iterasi (Fei, Miao & Liu, 2009).

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).

2.2.4. 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 di sebut 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. Dari situ lalu di hitung 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).

2.2.5. *Confusion Matrix*

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

2.2.6. Kurva ROC (*Receiver Operating Characteristic*)

Grafik kurva ROC (Receiver operasi karakteristik) 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 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, naif 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 di mana tingkat TP diplot pada sumbu Y-dan tingkat FP diplot pada X-sumbu (Gorunescu, 2011).

2.3. 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).

BAB III
METODOLOGI PENELITIAN

3.1. Desain 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 kedalam data set, termasuk variabel yang diperlukan dalam proses. Data untuk eksperimen ini dikumpulkan, lalu 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 oleh beberapa peneliti sebelumnya. 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. Sedangkan untuk pengujian metode, penulis mencoba mengusulkan memasang *software* anti *spamming* pada PC atau Laptop.

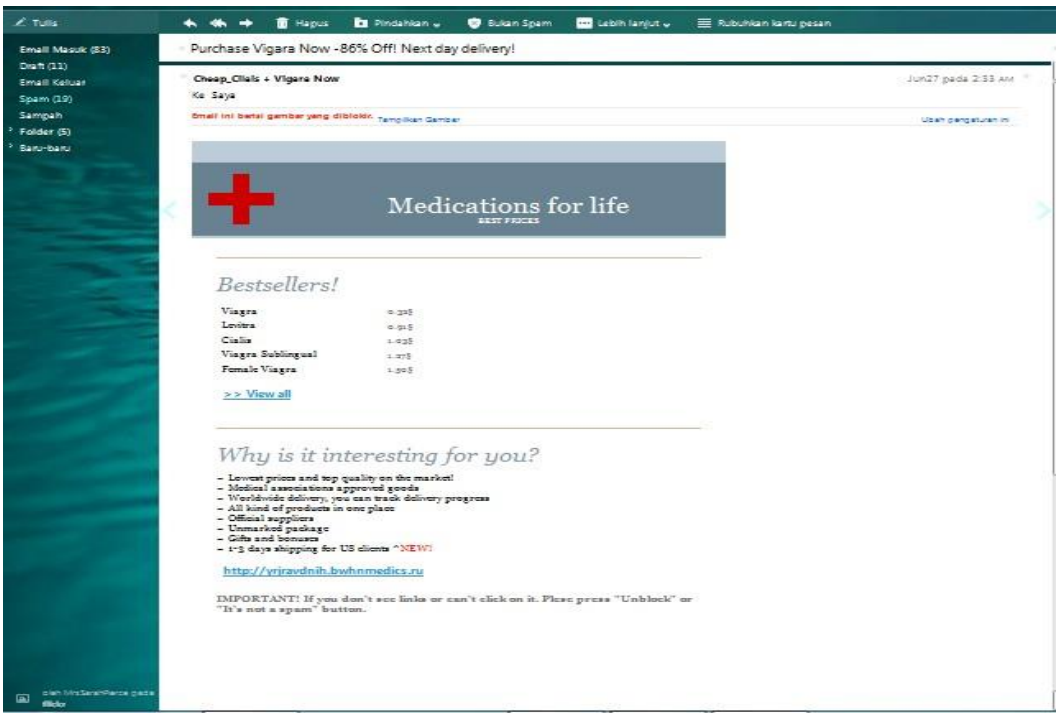
5. Evaluasi dan Validasi Hasil

   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*.

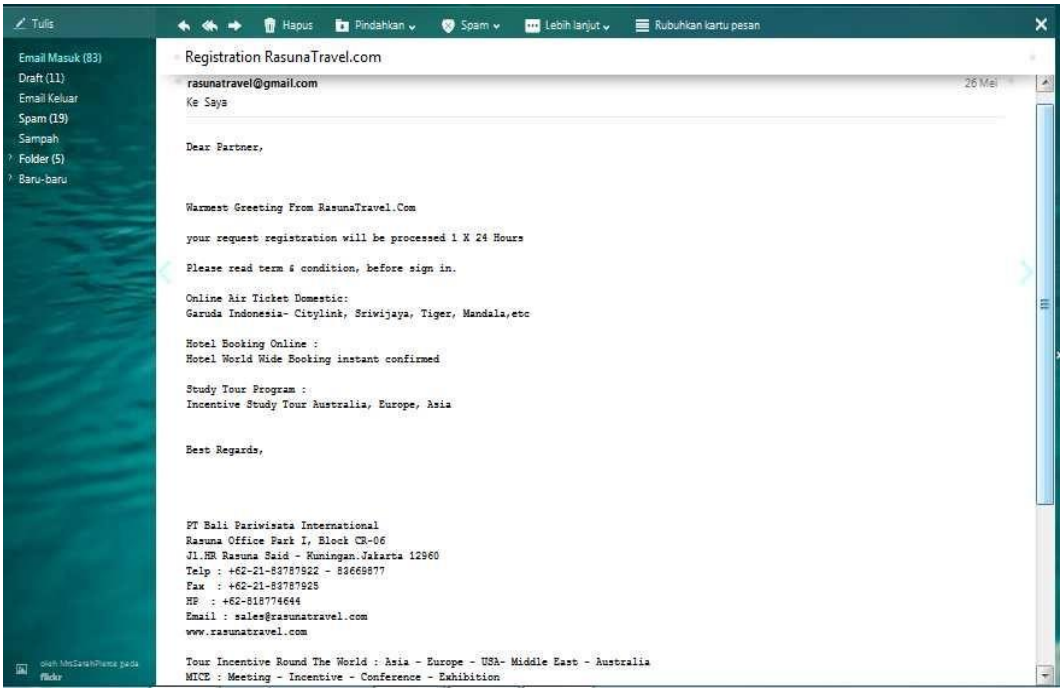
3.2. Pengumpulan Data

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:



Sedangkan contoh data non spam sebagai berikut:



3.3. 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):

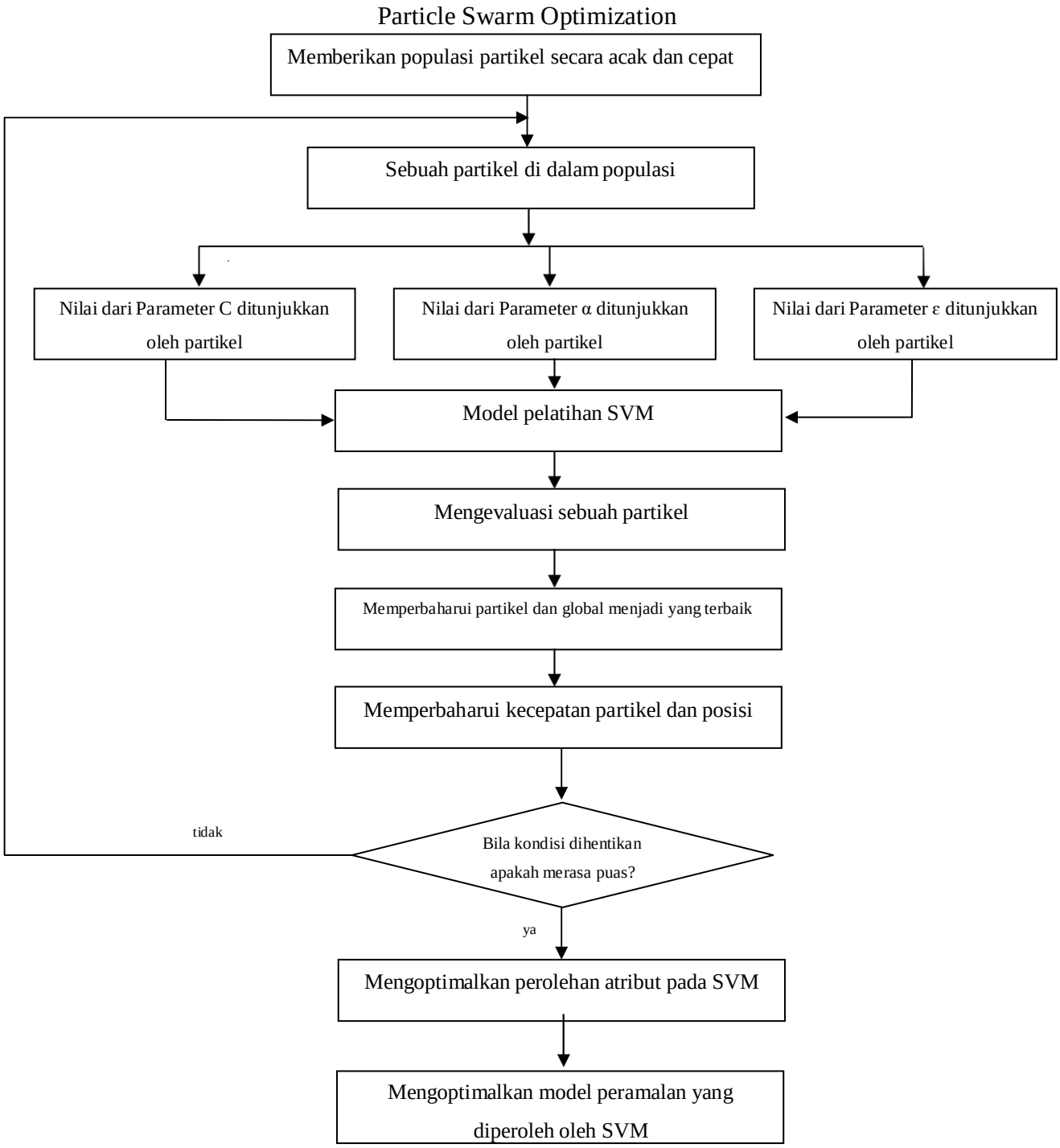
a. Data validation, untuk mengidentifikasikan dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*).
b. Data *integration and transformationi,* untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan 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 3.1. 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

3.4. Metode yang diusulkan

Pada tahap modeling ini dilakukan pemprosesan data training sehingga akan membahas metode algoritma yang diuji dengan memasukkan data spam email kemudian di analisa dan dikomparasi. Berikut ini bentuk gambaran metode algoritma yang diuji.



Gambar 3.1. model yang diusulkan

3.5. 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 kali ini yang digunakan adalah penelitian Experiment. Penulis melakukan proses eksperimen menggunakan aplikasi RapidMiner 5.3. Sedangkan spesifikasi komputer yang penulis gunakan dalam penelitian ini dapat dilihat pada tabel 3.2.

Tabel 3.2 spesifikasi *hardware* dan *software*

Software	*hardware*
Sistem Operasi: Windows 7	CPU: Intel Pentium Dual Core
Data Mining: Rapid Miner versi 5.3	Memory: 4 GB
	Harddisk: 160 GB

3.6. Evaluasi dan Validasi Hasil

Model yang diusulkan pada penelitian tentang prediksi penyakit diabetes 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.

BAB IV
HASIL DAN PEMBAHASAN

4.1. Hasil Eksperimen dan Pengujian Metode

4.1.1. 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 4.1. 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.

4.1.2. 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*:

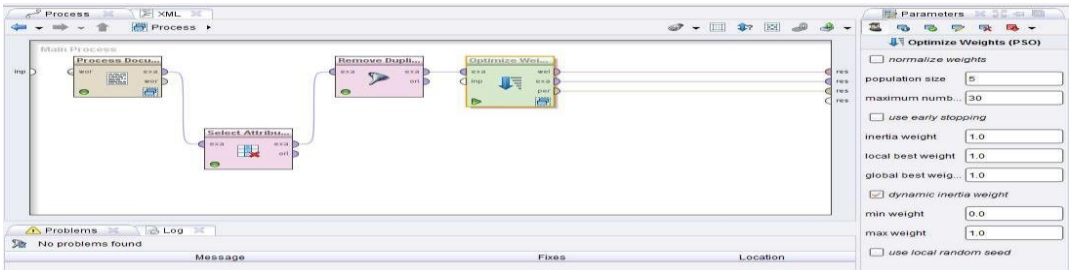
Tabel 4.2 Eksperiment penentuan nilai *training cycle* SVM berbasis PSO

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

4.2. Evaluasi dan Validasi Hasil

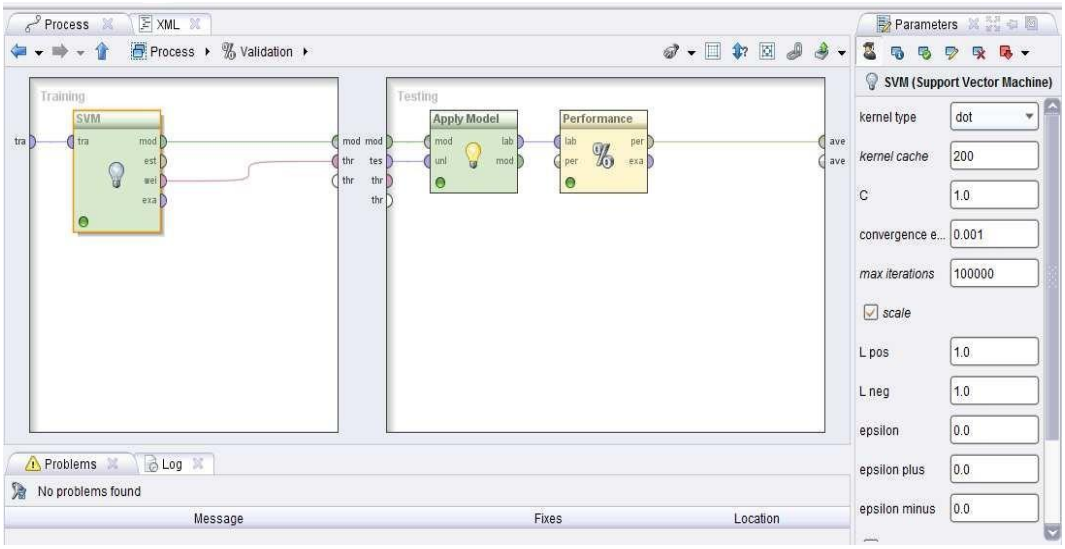
Hasil dari pengujian model yang dilakukan adalah memprediksi spam email dengan *Support Vector Machine* dan *Support Vector Machine* berbasis *particle swarm optimization* untuk menentukan nilai *accuracy* dan AUC. Dalam menentukan nilai tingkat keakurasian dalam model *support vector machine* dan algoritma *neural network* berbasis *particle swarm optimization*. Metode pengujiannya menggunakan *cross validation* dengan desain modelnya sebagai berikut:



Gambar 4.1. Desain Model Validasi

4.2.1 Hasil Pengujian Model Support Vector Machine

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



Gambar 4.2. 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 4.1.

1. Confusion Matrix

Tabel 4.3 data training yang digunakan terdiri dari 200 data spam dan 200 data non spam. Untuk data spam, 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 4.3 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%	

Berdasarkan tabel 4.1 tersebut menunjukkan bahwa, tingkat akurasi dengan menggunakan algoritma SVM adalah sebesar 85.75%, dan dapat dihitung untuk mencari nilai *accuracy*, *sensitivity*, *specificity*, *ppv*, dan *npv* hasilnya dan dapat dihitung untuk mencari nilai *accuracy*, *sensitivity*, *specificity*, *ppv*, dan *npv* pada persamaan dibawah ini:

$$\textit{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{161 + 182}{161 + 182 + 39 + 18} = 0.8575$$

$$\textit{Sensitivity} = \frac{TP}{TP + FN} = \frac{161}{161 + 18} = 0.8994$$

$$\textit{Specificity} = \frac{TN}{TN + TP} = \frac{182}{182 + 161} = 0.5306$$

$$PPV = \frac{TP}{TP + FP} = \frac{161}{161 + 39} = 0.805$$

$$NPV = \frac{TN}{TN + FN} = \frac{182}{182 + 18} = 0.91$$

Hasil perhitungan terlihat pada tabel 4.4 dibawah ini:

Tabel 4.4 Nilai *accuracy*, *sensitivity*, *specificity*, *ppv*, dan *npv* untuk Metode *Support Vector Machine*

	Nilai (%)
Accuracy	85.75
Sensitivity	89.94
Specificity	53.06
PPV	81.5
NPV	9.1

2. Kurva ROC

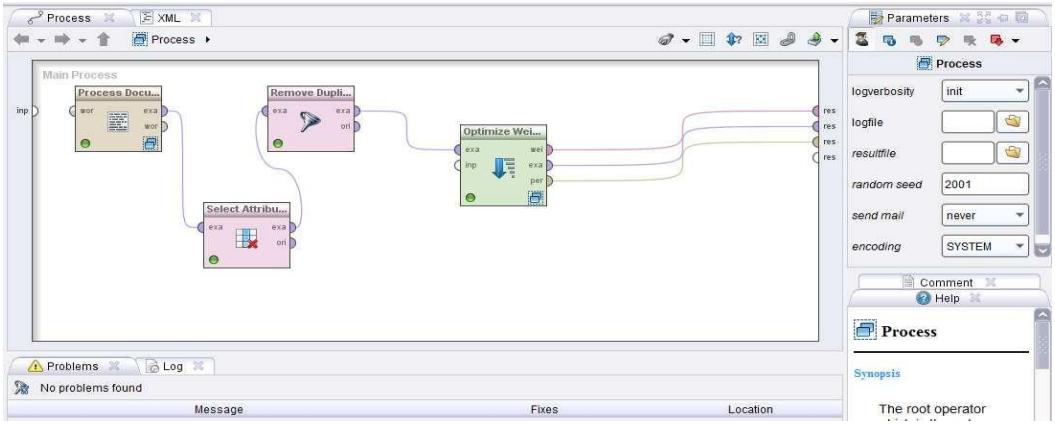
Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua metode komparasi bisa dilihat pada gambar 4.2 yang merupakan kurva ROC untuk algoritma *Support Vector Machine*. Kurva ROC pada gambar 4.2 mengekspresikan *confusion matrix* dari Tabel 4.1. Garis horizontal adalah *false positives* dan garis vertikal *true positives*.

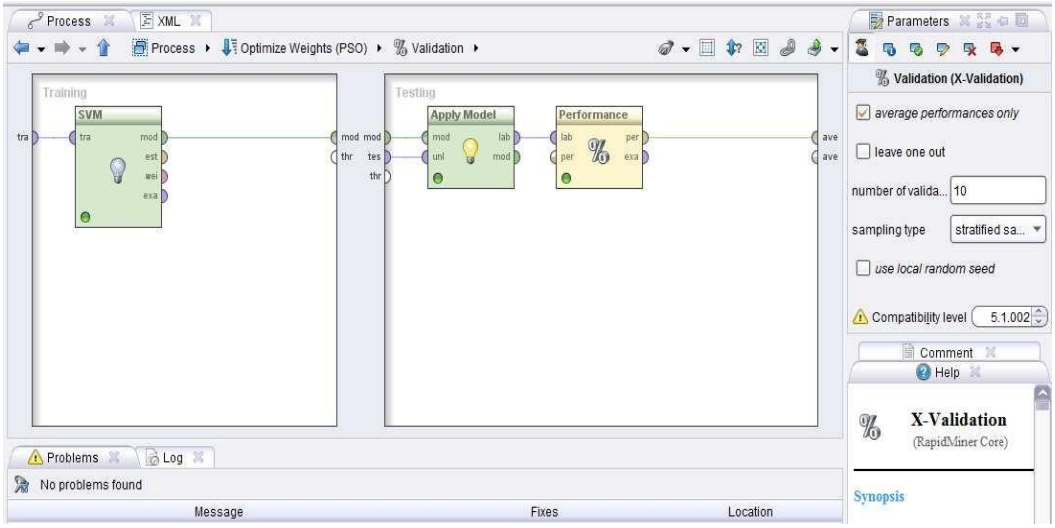
Gambar 4.3. Kurva ROC dengan metode *Support Vector Machine*

Dari gambar 4.3 terdapat grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.901 dimana diagnosa hasilnya *fair classification*.

4.2.2 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 4.4 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 4.2

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 4.5 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%	

Berdasarkan tabel 4.5 tersebut menunjukkan bahwa, tingkat akurasi dengan menggunakan algoritma SVM berbasis *Particle Swarm Optimization* (PSO) adalah sebesar 89.24% dan dapat dihitung untuk mencari nilai *accuracy*, *sensitivity*, *specificity*, *ppv*, dan *npv* hasilnya dan dapat dihitung untuk mencari nilai *accuracy*, *sensitivity*, *specificity*, *ppv*, dan *npv* pada persamaan di bawah ini:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{171 + 185}{171 + 185 + 29 + 14} = 0.8922$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{171}{171 + 14} = 0.9243$$

$$Specificity = \frac{TN}{TN + TP} = \frac{185}{185 + 171} = 0.5196$$

$$PPV = \frac{TP}{TP + FP} = \frac{171}{171 + 29} = 0.855$$

$$NPV = \frac{TN}{TN + FN} = \frac{185}{185 + 14} = 0.9296$$

Hasil perhitungan terlihat pada tabel 4.6 dibawah ini:

Tabel 4.6 Nilai *accuracy, sensitivity, specificity, ppv,* dan *npv* Metode *Support Vector Machine* berbasis *Particle Swarm Optimization*

	Nilai (%)
Accuracy	89.24
Sensitivity	92.43
Specificity	51.96
PPV	8.6
NPV	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 4.4 mengekspresikan *confusion matrix* dari Tabel 4.3. Garis horizontal adalah *false positives* dan garis vertikal *true positives*.

Gambar 4.5. Kurva ROC dengan Metode *Support Vector Machine* berbasis *Particle Swarm Optimization* (PSO)

Dari Gambar 4.5 terdapat grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.928 dimana diagnosa hasilnya *Fair classification*.

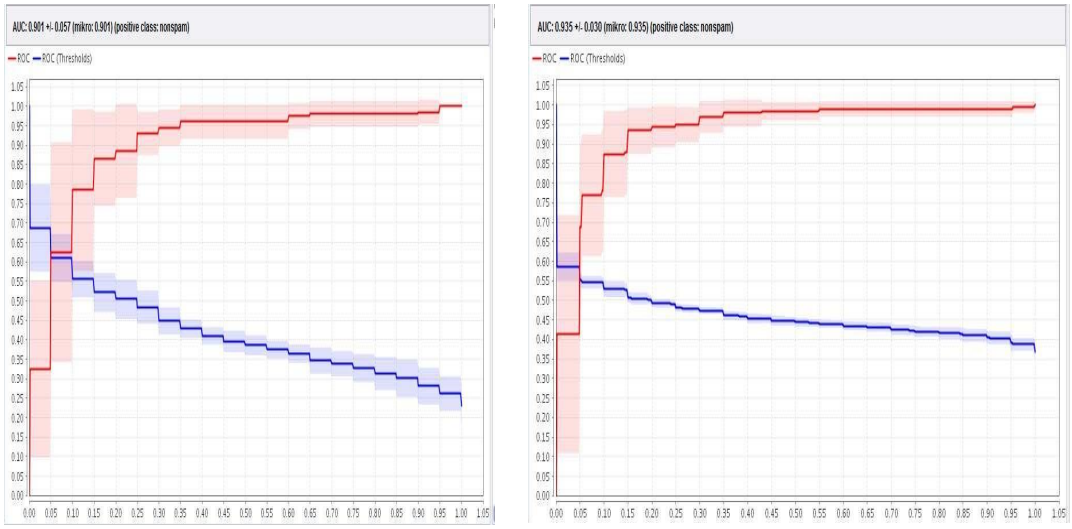
4.2.3 Analisis Evaluasi dan Validasi Model

Dari hasil pengujian diatas, baik evaluasi menggunakan *confusion matrix* mampu ROC *curve* terbukti bahwa hasil pengujian algoritma SVM berbasis PSO memiliki nilai akurasi yang lebih tinggi dibandingkan dengan algoritma SVM nilai akurasi untuk model algoritma SVM sebesar 85.75% dan nilai akurasi untuk model algoritma SVM berbasis PSO sebesar 89.24% dengan selisih akurasi 3.49% dapat dilihat pada tabel 4.7 dibawah ini:

Tabel 4.7 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 SVM berbasis PSO (*Particle Swarm Optimization*) menghasilkan nilai 0.935 dengan nilai diagnosa *Fair Classification*, dan selisih nilai keduanya sebesar 0.034. dapat dilihat pada Gambar 4.5 dibawah ini.



Gambar 4.6 Kurva ROC dengan *Support Vector Machine* dan *Support Vector Machine* berbasis *Particle Swarm Optimization*

Dengan demikian algoritma SVM berbasis PSO dapat memberikan solusi untuk permasalahan dalam prediksi spam email. Untuk rinciannya dapat dilihat pada tabel 4.5 dan Gambar 4.6.

4.3. Pembahasan

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.

4.4. Implikasi Penelitian

Implikasi penelitian pada tesis ini peneliti arahkan pada tiga aspek, yaitu:

1. Aspek Sistem

   Penerapan kebijakan dalam penentuan prediksi spam email akan membawa pengaruh pada sistem, di mana 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*"**, di mana *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*.

BAB V
PENUTUP

5.1. 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 email *spam* 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 reviewnya.

5.2. Saran

Walaupun pengklasifikasi *Support Vector Machine* sudah sering digunakan dan mempunyai performa yang baik dalam mengklasifikasi 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.

DAFTAR 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.