SELEKSI FITUR DAN PARAMETER PADA METODE SUPPORT VECTOR MACHINE BERBASIS ALGORITMA GENETIKA PADA NASABAH TELEMARKETING BANK

TESIS

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

ERENE GERNARIA SIHOMBING
14000368

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
NUSA MANDIRI
JAKARTA
2015

SURAT PERNYATAAN ORISINALITAS

Yang bertandatangan di bawah ini :

Nama : Erene Gernaria Sihombing
NIM : 14000368
Program Studi : Magister IlmuKomputer
Jenjang : Strata Dua (S2)
Konsentrasi : *Management Information System*

Dengan ini menyatakan bahwa tesis yang telah saya buat dengan judul: "**Seleksi Fitur Dan Parameter Pada Metode Support Vector Machine Berbasis Algoritma Genetika Pada Nasabah Telemarketing Bank"** adalah hasil karya sendiri, dan semua sumber baik yang kutipan maupun yang dirujuk telah saya nyatakan dengan benar dan tesis ini 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, Agustus 2015
Yang Menyatakan

Erene Gernaria Sihombing

LEMBAR PENGESAHAN

LEMBAR PENGESAHAN

Tesis ini diajukan oleh:

Nama : Erene Gernaria Sihombing
NIM : 14000368
Program Studi : Magister Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : *Management Information System*
JudulTesis : **"Seleksi Fitur Dan Parameter Pada Metode Support Vector Machine Berbasis Algoritma Genetika Pada Nasabah Telemarketing Bank**

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, 28 Agustus 2015
Pascasarjana Magister Ilmu Komputer
STMIK Nusa Mandiri
Direktur

Prof. Dr. Ir. Kaman Nainggolan, MS

DEWAN PENGUJI

Penguji I : Dr. Sularso Budilaksono, M.Kom .....

Penguji II : Dr. Sfenrianto, M.Kom .....

Penguji III/ Pembimbing : Dr. Mochamad Wahyudi, MM, M.Kom, M.Pd.....

DAFTAR ISI

LEMBAR JUDUL SKRIPSI i
LEMBAR PERNYATAAN ORISINALITAS ii
LEMBAR PENGESAHAN iii
DAFTAR ISI iv

BAB I PENDAHULUAN
1.1 Latar Belakang 1
1.2 Identifikasi Masalah 4
1.3 Rumusan Masalah 4
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 4
1.6 Ruang Lingkup Penelitian 4
1.5 Sistematika Penulisan 5

BAB II LANDASAN TEORI
2.1 Tinjauan Pustaka 6
2.1.1. Nasabah Potensial Telemarketing 6
2.1.2. Data Mining 7
2.1.2.1 *Support Vector Machine* 11
2.1.2.2 *Parameter Selction* 13
2.1.3. Feature Selection 13
2.1.4 Genetic Algorithm 15
2.15 Pengujian Evaluasi dan Performa Prediksi 16
2.2 Tinjauan Studi 16
2.3 Kerangka Pemikirian 18

BAB III METODE PENELITIAN
3.1 Desain Penelitian 19
3.2 Pengumpulan Data 20
3.3 Metode Yang Disusulkan 22
3.4 Eksperimen dan Pengujian Metode 23
3.5 Evaluasi dan Validasi Hasil 24

BAB IV HASIL PENELITIAN DAN PEMBAHASAN
4.1 Tahapan Pengujian 25
4.2. Hasil Eksperimen dan Pengujian Metode 26
4.2.1 Metode Support Vector Machine 26
4.2.2 Parameter Support Vector Machines berbasis Genetic Algorithm 27
4.2.3 Seleksi Fitur Support Vector Machines berbasis Genetic Algorithm 28
4.2.4 Seleksi Fitur dan Parameter Support Vector Machines berbasis Genetic Algorithm 30

BAB V PENUTUP
5.1 Kesimpulan 32
5.2 Saran 32

Daftar Pustaka

BAB 1
PENDAHULUAN

3.3.Latar Belakang Penulisan

Penilaian nasabah perlu diprediksi dengan akurat, untuk mengidentifikasi sejumlah nasabah yang mempunyai tingkat respon yang relatif lebih tinggi sebagai nasabah yang potensial(Liao, Chen, dan Hsieh 2011)

Nasabah/pelanggan potensial merupakan individu yang harus dicari untuk mengidentifikasi pelanggan potensial yang dapat merespon dalam pemasaran produk (Chen, Hsu, dan Hsu 2011). Pemasaran berdasarkan pada mekanisme penyaringan pelanggan secara signifikan dapat mengurangi biaya integral dari pemasaran, dan juga dapat memperoleh lebih nasabah yang potensial untuk diterima (Chen, Hsu, dan Hsu 2011). Dalam *telemarketing* fokus utama adalah pada kualitas data prospek, oleh karena itu dalam memprediksi pelanggan yang memiliki probabilitas tinggi, untuk layanan tersebut, dapat dicapai dengan menggunakan teknik data mining(Vajiramedhin 2014).

Menganalisis *database* bank untuk manajemen perilaku pelanggan adalah sulit karena database bank yang multi-dimensi, terdiri dari catatan rekening bulanan dan catatan transaksi harian (Hsieh, 2004). Studi ini menunjukkan bahwa mengidentifikasi pelanggan dengan model penilaian perilaku karakter pelanggan membantu dan memfasilitasi pengembangan strategi pemasaran

Penelitian yang dilakukan oleh Romdhane et al. (2010) tujuan utama mendapatkan atau menilai profil pelanggan adalah untuk membangun model pelanggan yang dapat diandalkan dalam target promosi pemasaran dan akibatnya mendapatkan keuntungan yang lebih baik. Penelitian (Romdhane, Fadhel, dan Ayeb 2010) yaitu menggunakan *fuzzy clustering* untuk mengembangkan model pelanggan (profil literatur) untuk pemasaran bertarget. Langkah pertama yang dilakukan yaitu dengan mengkluster data menggunakan algoritma FCM untuk mengambil kelompok pelanggan. Lalu mengurangi jumlah atribut pada masing-masing kelompok, hasil darilangkah ini memperolehsatu set kelompok masing-masing yang digambarkan dengan satu kelompok yang berbeda dari atribut(atau karakteristik). Pada langkah terakhir darimodel, membangun satu setprofil pelanggan masing-masing dimodelkan oleh *Backpropagation Neural Network* dan dilakukan training data pada kelompok pelangga n yang sesuai.

Penelitian berikutnya dilakukan oleh (Chen, Hsu, and Hsu 2011) Model respon pelanggan mengacu pada estimasi probabilitas berdasarkan informasi prediksi individual pelanggan (misalnya, usia, lokasi tempat tinggal, pendapatan, dan tingkat pendidikan). Informasi ini kemudian digunakan untuk menentukan representasi fungsional untuk kemungkinan respon pelanggan yang digunakan khusus untuk kampanye promosi masa depan. Penelitian ini mengusulkan sebuah metode prediksi yang mengintegrasikan *union sequential pattern* dengan algoritma klasifikasi *Support Vector Machine* dan *Logistic Regression* untuk membangun model respon pelanggan

Berdasarkan penggunaan pola *union sequential*, ukuran pelanggan potensial didirikan dengan mengidentifikasi atribut dengan asosiasi tingkat tinggi. Model prediksi ini kemudian dibangun menggunakan algoritma klasifikasi tersebut. Model ini diusulkan lebih akurat memungkinkan kita untuk mengidentifikasi sekumpulan pelanggan dengan relatif lebih tinggi tingkat responnya, yaitu pelanggan yang potensial.

Penelitian berikutnya dilakukan oleh (Moro and Laureano 2011) yaitu meneliti tentang pelanggan deposito bank pada bank portugis, yang bertujuan untuk menemukan model yang mampu menjelaskan kesuksesan suatu kontak pelanggan. Model tersebut dapat meningkatkan efisiensi promosi dengan mengidentifikasikan karakteristik utama yang mempengaruhi kesuksesan, membantu dalam manajemen yang lebih baik dari sumber daya yang tersedia (misalnya usaha manusia, panggilan telepon, waktu) dan seleksi yang berkualitas tinggi dan terjangkau pada sekumpulan nasabah potensial. Pada penelitian ini menggunakan tiga model untuk dijadikan perbandingan yaitu Naivese Bayes, Decision Tree dan SVM. Setelah dilakukan pengujian didapat nilai AUC pada NB sebesar 0.870, DT sebesar 0.868, dan SVM sebesar 0.938.

Model terbaik diwujudkan oleh Support Vector Machine(SVM), mencapai prestasi prediksi yang tinggi. Menggunakan analisis sensitivitas, dengan mengukur pentingnya masukan dalam model SVM dan pengetahuan tersebut dapat digunakan oleh para manajer untuk meningkatkan pemasaran.

Naive Bayes(NB) mampu menghasilkan akurasi klasifikasi yang baik, terutama untuk data dimensi tinggi. Tetapi memiliki kelemahan dalam pemilihan atribut yang berhubungan dan pembobotan atribut sehingga dapat menurunkan kinerja klasifikasi(Wu et al. 2015)

Support Vector Machine dapat memecahkan masalah dengan sampel yang kecil, *non-linier* dan masalah dimensi yang tinggi dengan menggunakan struktur minimalisasi resiko (*structural risk minimization*) bukan minimalisasi resiko empiris (*empirical risk minimization*) (Yuxia dan Hongtao 2012). SVM digunakan untuk klasifikasi pola, dan ide

dasarnya adalah: pemetaan data dalam ruang *input* dengan *non-linier* mengubah keruang fitur dimensi tinggi, dimana masalah linear klasifikasi *hyper-plane* menjadi secara optimal(Wang dan Meng 2011).

Bila menggunakan SVM, dua masalah dihadapkan yaitu bagaimana memilih fitur yang optimal untuk SVM dan mengatur parameter terbaik . Kedua masalah sangat penting, karena pilihan fitur mempengaruhi kesesuaian parameter dan sebaliknya (Huang dan Wang 2006). Fitur yang banyak atau sangat berhubungan, secara signifikan akan mengurangi tingkat akurasi klasifikasi, dengan menghapus beberapa fitur, tingkat efisiensi akurasi dan klasifikasi dapat diperoleh (Lin et al. 2009). Selain pemilihan fitur, pengaturan parameter yang tepat dapat meningkatkan akurasi klasifikasi SVM (Huang dan Wang 2006). Kunci parameter dalam SVM sangat penting, keakuratan klasifikasi atau regresi ditentukan oleh sekelompok parameter yang sesuai (Yuxia dan Hongtao 2012).

Beberapa algoritmapun banyak direkomendasikan oleh peneliti dunia untuk mengoptimasi parameter pada machine learning, seperti: particle swarm optimization(Yukun Bao, Zhongyi Hu 2013), simulated annealing(Dehuai et al. 2012), genetic algorithm(Ilhan dan Tezel 2013)

Simulated annealing (SA) efektif pada pemuatan masalah optimasi pola, namun SA memiliki kecenderungan untuk terjebak dalam minimum lokal ketika suhu anil rendah (tingkat anil cepat) dan semakin tidak konvergen ketika suhu anil tinggi (tingkat anil lambat)(Zameer, Mirza, dan Mirza 2014).

Particle swarm optimization (PSO) memiliki kemampuan pencarian global yang kuat, juga dapat membantu mencari parameter yang optimum secara cepat (Wang et al., 2014), namun kinerja PSO diyakini memiliki ketergantungan yang sensitif pada parameter, dan cenderung terjebak dalam minimum lokal (Zameer et al., 2014), selain itu PSO juga sulit mendapatkan nilai yang optimum dalam mengoptimasi lebih dari sepuluh parameter.

Genetic algorithm atau algoritma genetika dapat mengatasi masalah yang nonlinier dengan diskontinuitas dan minimal lokal secara efisien serta lebih efisien dalam mengoptimasi lebih dari sepuluh parameter(Machairas, Tsangrassoulis, dan Axarli 2014).

Dari uraian tersebut diatas, maka dalam penelitian ini akan digunakan metode support vector machine yang dipadu dengan algoritma genetika yang akan digunakan untuk melakukan optimasi parameter support vector machine.

3.4.Identifikasi Masalah

Support Vector Machine dapat menyelesaikan masalah pada sampel data yang kecil yang ada pada konsumsi energi, tetapi *Support Vector Machine* memiliki kelemahan pada sulitnya pemilihan fitur yang sesuai dan parameter yang optimal sehingga menyebabkan tingkat akurasi menjadi rendah.

3.5.Rumusan Masalah

Rumusan masalah yang pada penelitian ini adalah seberapa besar akurasi metode *Support Vector Machine* yang dipadukan dengan algoritma genetika dengan cara optimasi parameter *Support Vector Machine?*

3.6.Tujuan Penelitian

Tujuan dari penelitian ini adalah menerapkan metode *Support Vector Machine* untuk peningkatan optimasi parameter potensial yang dipadukan dengan algoritma genetika guna meningkatkan akurasi prediksi nasabah telemarketing.

3.7.Manfaat Penelitian

a. Pada penelitian ini menerapkan teori permodelan yang diharapkan dapat memberikan masukan terhadap metode Support Vector Machine yang dapat menemukan parameter yang paling baik dan ditingkatkan dengan Algoritma genetika untuk optimasi parameter.
b. Penelitian ini dapat dijadikan acuan untuk para telemarketing bank sebagai strategi pemasaran untuk memilih nasabah potensial dengan melihat parameter-parameter yang sesuai.

3.8.Ruang Lingkup Penelitian

Ruang lingkup pembahasan dalam penelitian ini dibatasi pada metode *Support Vector Machine* dengan *Support Vector Machine* berbasis *Genetic Algorithm* dengan cara optimasi parameter dalam penentuan potensial nasabah menggunakan *Bank Marketing* data.

3.9.Sistematika Penulisan

Disajikan dalam lima bab dan masing-masing bab terdiri dai 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 dan hipotesis.

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 2
LANDASAN TEORI

3.10. Tinjauan Pustaka

Tinjauan pustaka dalam penulisan tesis ini dilakukan dengan menggunakan buku dan jurnal yang berhubungan dengan tema yang dipilih. Secara lebih detail tinjauan dalam penulisan tesis ini dapat dijelaskan sebagai berikut;

2.1.1 Nasabah Potensial Telemarketing

Nasabah merupakan sumber utama keuntungan perusahaan/bank dan menduduki posisi yang sangat penting dalam strategi pengembangan bisnis modern. Namun, ada perbedaan besar antara pelanggan karena sebagian besar keuntungan berasal dari pelanggan berkualitas, sehingga perusahaan harus dapat mengidentifikasi yang mana pelanggan yang berkualitas/potensial (Xing dan Xin-feng 2010)

Nilai Potensial pelanggan dapat diukur dengan metode RFM. Model RFM adalah model yang digunakan untuk mengidentifikasi perilaku pelanggan(Weiwen et al. 2008). Model ini menggunakan pendekatan tiga dimensi dari data transaksi pelanggan yaitu:

1) Recency(R): Transaksi terakhir
2) Frequency(F): Jumlah total pembelian
3) Monetary(M): Nilai transaksi

Telemarketing adalah teknik pemasaran interaktif telepon yang dilakukan oleh seorang marketer untuk mengumpulkan sejumlah calon pelanggan melalui telepon untuk memasarkan barang dagangan atau jasa(Vajiramedhin 2014). Pemasaran langsung adalah pemasaran menemukan prospek pinpoint untuk layanan tambahan berdasarkan data pelanggan yang dikumpulkan dalam database yang dikenal sebagai database pemasaran. Sebuah database pelanggan potensial bisa mendapatkan keuntungan besar dari pemasaran langsung seperti komunikasi, iklan dan analisis..

Yang paling sukses dalam *telemarketing* adalah fokus pada kualitas data prospek, memprediksi pelanggan diharapkan memiliki probabilitas yang lebih tinggi untuk menggunakan layanan ini dengan menggunakan teknik data mining. Untuk memahami perilaku pelanggan, banyak bank telah mengadopsi teknik prediktif didasarkan pada data mining untuk memprediksi data pelanggan untuk mengklasifikasikan pelanggan sebelum

menawarkan layanan khusus. Prediksi atau klasifikasi adalah tugas yang paling penting dalam data mining yang biasanya diterapkan untuk mengklasifikasikan kelompok data (Vajiramedhin, 2014)

Atribut yang menggambarkan karakteristik nasabah yang baik adalah(Moro dan Laureano 2011).

Tabel 2.2
Tabel Atribut karakteristik nasabah

Nama Atribut	Deskripsi
Age	Umur nasabah
Job	Pekerjaan nasabah
Education	Pendidikan terakhir
Marital Status	Status Perkawinan
Annual balance	Saldo tahunan
Housing	Kepemilikan rumah
Loans in delay ?	Hutang pinjaman
Contact	Jenis Kontak yang dapat dihubungi
Day	Tanggal marketing
Month	Bulan marketing
Duration	Lamanya dihubungi
Campaign	Promosi
Pdays	Promosi perhari
Previous	Promosi sebelumnya
Poutcome	Hasil sebelumnya

Sumber : (moro dan laureano, 2011)

2.1.2 Data Mining

Data mining, sering juga disebut *knowledge discovery in database* (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Santosa, 2007). Menurut Gartner *Data Mining* adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan

dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Larose, 2005).

Istilah *data mining* dan *knowledge discovery in database* (KDD) seringkali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah *data mining*. Proses KDD secara garis besar dapat dijelaskan sebagai berikut (Larose, 2005):

1. *Data Selection*

   Pemilihan (seleksi) data dari sekumpulan data operasional perlu di lakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Pre-processing/ Cleaning*

   Sebelum proses *data mining* dapat di laksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (*tipografi*). Juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. *Transformation*

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Data mining*

   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation/ Evaluation*

   Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan dari bagian dari proses KDD yang di sebut *interpretation*. Tahap ini mencakup pemeriksaan

apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

Enam fase CRSIP-DM (Larose, 2005):

1. Fase Pemahaman Bisnis (*Business Understanding Phase*)
    a. Penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan.
    b. Menerjemahkan tujuan dan batasan menjadi formula dari permasalahan *data mining*.
    c. Menyiapkan strategi awal untuk mencapai tujuan.
2. Fase Pemahaman Data (*Data Understanding Phase*)
    a. Mengumpulkan data.
    b. Menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal.
    c. Mengevaluasi kualitas data.
    d. Jika diinginkan, pilih sebagian kecil group data yang mungkin mengandung pola dari permasalahan.
3. Fase Pengolahan Data (*Data Preparation Phase*)
    a. Siapkan dari data awal, kumpulan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu di laksanakan secara intensif.
    b. Pilih kasus dan variabel yang ingin dianalisis dan yang sesuai analisis yang akan dilakukan.
    c. Lakukan perubahan pada beberapa variable jika di butuhkan.
    d. Siapkan data awal sehingga siap untuk perangkat permodelan.
4. Fase Pemodelan (*Modeling Phase*)
    a. Pilih dan aplikasikan teknik permodelan yang sesuai.
    b. Kalibrasi aturan model untuk mengoptimalkan hasil.
    c. Perlu diperhatikan bahwa beberapa teknik mungkin untuk digunakan pada permasalahan *data mining* yang sama.
    d. Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk menjadikan data ke dalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik *data mining* tertentu.

5. Fase Evaluasi (*Evaluation Phase*)
    a. Mengevaluasi satu atau lebih model yang di gunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektifitas sebelum disebarkan untuk digunakan.
    b. Menetapkan apakah terdapat model yang memenuhi tujuan pada fase awal.
    c. Menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik.
    d. Mengambil keputusan berkaitan dengan penggunaan hasil dari *data mining.*
6. Fase Penyebaran (*Deployment Phase*)
    a. Menggunakan model yang dihasilkan. Terbentuknya model tidak menandakan telah terselesaikannya proyek.
    b. Contoh sederhana penyebaran: Pembuatan laporan.
    c. Contoh kompleks penyebaran: Penerapan proses *data mining* secara paralel pada departemen lain.

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (Larose, 2005):

1. Deskripsi

    Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup professional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

2. Estimasi

    Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh, akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, indeks berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik dari nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya.

Contoh lai yaitu estimasi nilai indeks prestasi kumulatif mahasiswa program pascasarjana dengan melihat nilai indeks prestasi mahasiswa tersebut pada saat mengikuti program sarjana.

3. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi menghasilkan nilai dari hasil di masa mendatang.

Contoh prediksi dalam bisnis dan penelitian adalah :

- Prediksi harga beras dalam tiga bulan yang akan dating.
- Prediksi presentase kenaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikkan.

Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

4. Klasifikasi

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, pengolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, pendapatan rendah.

5. Pengklusteran

Pengklusteran merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidak miripan dengan *record-record* dalam kluster lain. Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan atau homogen, yang mana kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.

6. Asosiasi

Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

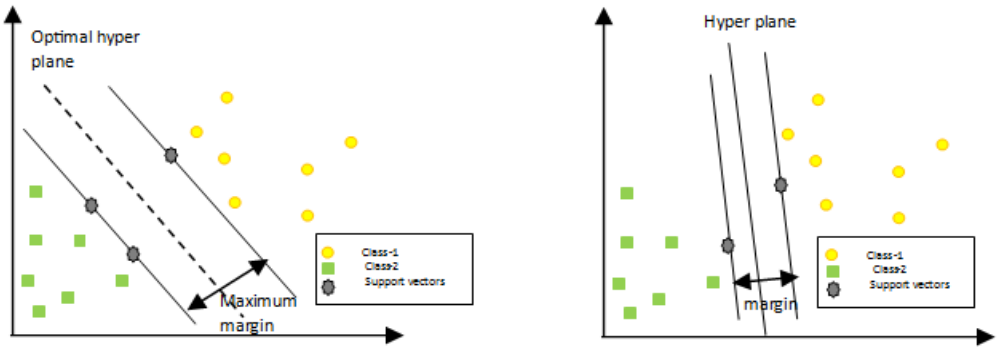
2.1.2.1 Support Vector Machine

Support Vector Machine (SVM) salah satu teknik klasifikasi data yang muncul, yang diperkenalkan oleh Vapnik(1995), dan baru-baru ini banyak digunakan di berbagai bidang termasuk masalah klasifikasi pengenalan pola, bio informatika dan keuangan (Zhao et al.

2011). Secara konseptual SVM adalah sebuah mesin linier, dilengkapi dengan fitur-fitur khusus, dan berdasarkan metode *structural risk minimization* (SRM) dan sebuah *statistical learning theory*.Akibatnya, SVM dapat memberi kinerja yang baik dalam masalah generalisasi pengenalan pola, tanpa memasukkan masalah, pengetahuan *domain* yang memberikan fitur yang unik diantara mesin-mesin belajar lainnya (Gorunescu, 2011).

SVM memecahkan permasalahan seperti sampel kecil, dimensi tinggi, nonlinier dan masalah lokal minimum. SVM telah menunjukkan kinerja yang baik diberbagai bidang seperti pengenalan pola dan regresi(Wang & Meng, 2011). SVM adalah seperangkat metode yang terkait untuk suatu metode pembelajaran, untuk kedua masalah klasifikasi dan regresi(Oded Maimon, 2010). Masalah klasifikasi dapat dibatasi untuk mempertimbangkan masalah dua-kelas tanpa mengurangi keadaan yang umum. Hal ini dapat digambarkan sebagai berikut: misalkan dua kelas objek yang diberikan, kita lalu dihadapkan objek baru, dan harus menetapkan kesalah satu dari dua kelas tersebut (Yong Shi, Yingjie Tian, Gang Kou, Yi Peng, 2011).

SVM adalah metode klasifikasi dua kelas dan teori ini didasarkan pada gagasan minimalisasi risiko struktural. SVM meminimalkan kesalahan generalisasi dan memaksimalkan margin geometris antara dua kelas, maka juga dikenal sebagai pengklasifikasi maksimum margin. SVM menggunakan fungsi kernel untuk memetakan data input ke dalam ruang fitur dimensi tinggi dan menemukan *hyper plane* optimal untuk memisahkan data dua kelas (Aydin, Karakose, & Akin, 2011).



Gambar 2.2 *Hyper plane* dan *support vectors*
(Aydin et al., 2011)

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 et al., 2011).

Untuk ruang *n*-dimensi, data masukan xi(i = 1...k) milik kelas1atau kelas 2 dan label terkait menjadi -1untuk kelas1dan +1untuk kelas 2.

Tujuan dari SVM adalah untuk memisahkan data kelas dengan cara maksimal margin *hyper plane*. Dengan demikian, SVM menjamin untuk memaksimalkan jarak antara data yang paling dekat dengan *hyper plane*. Jika input data dapat dipisahkan secara linear, pemisahan *hyper plane* dapat diberikan dalam persamaan:

$f(X) = w^T x + b$(3) (2.1)

dimana *w* adalah n-dimensi bobot vektor dan *b* adalah pengali skalar atau nilai bias. Persamaan ini menemukan maksimum margin untuk memisahkan kelas dari kelas positif dari kelas negatif. Fungsi keputusan ditunjukkan dalam persamaan. Contoh untuk data linear terpisah ditunjukkan pada Gambar 2.2:

$y_i(w \cdot x_i + b) \geq 1 \; i = 1. ..k$ (2.2)

2.1.2.2 Parameter Selection

Pemilihan parameter memainkan peran sentral dalam *machine learning*. Gagasan utama pemilihan parameter adalah memilih subset dari parameter yang sesuai untuk membangun model pembelajaran yang kuat. Ada banyak potensi manfaat pemilihan parameter (Ludermir, de Souto, & Vellasco, 2012):

1. Memudahkan visualisasi data dan pemahaman tentang data,
2. mengurangi kebutuhan pengukuran dan penyimpanan,
3. mengurangi pelatihan dan penggunaan waktu,
4. menentang dimensi untuk meningkatkan kinerja prediksi.

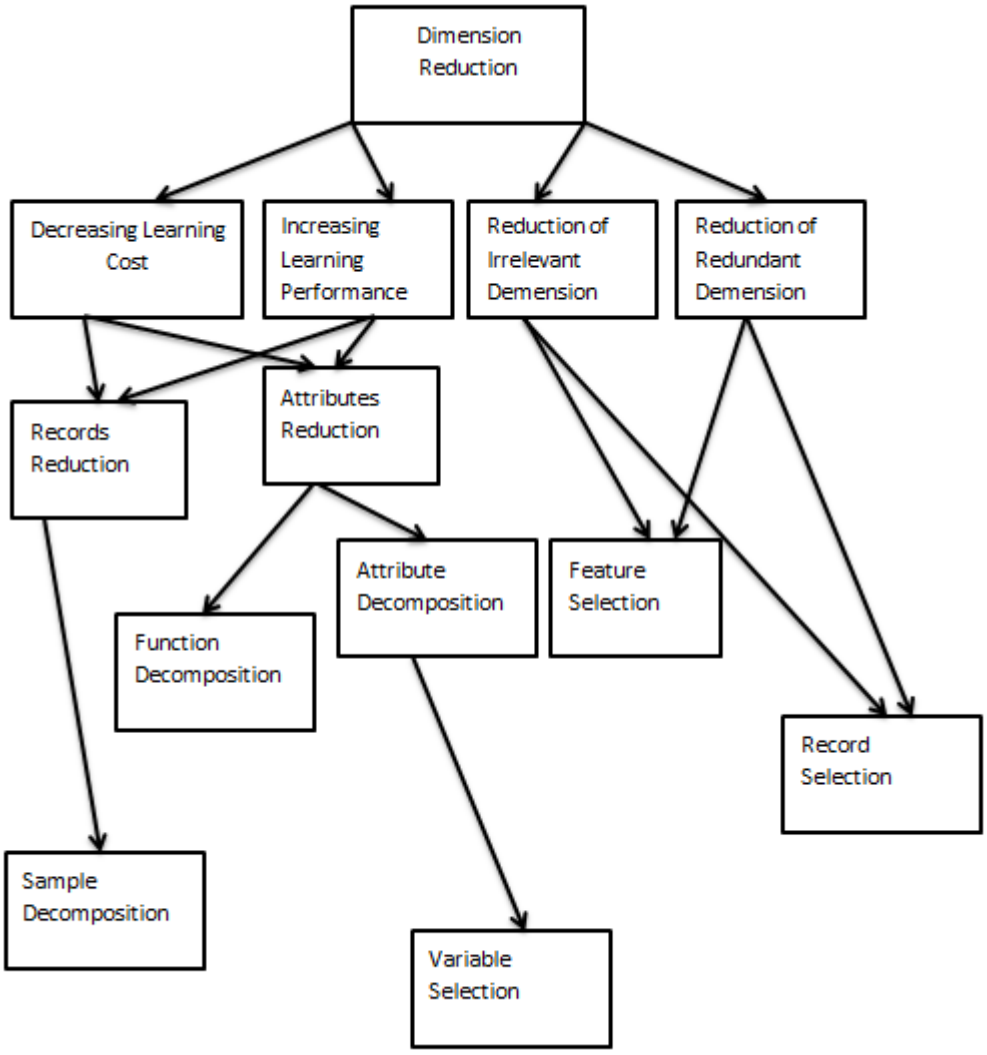
2.1.3 Feature Selection

Suatu observasi, example, pattern (pola) atau obyek biasanya ditandai oleh beberapa atribut. Atribut ini sering juga disebut dengan variabel. Juga ada yang menyebutnya dengan fitur. Dalam bahasan pengurangan dimensi data biasa disebut dengan *feature extraction* dan *feature selection.* Kedua istilah ini berarti mengurangi dimensi variabel atau kolom dari data (Santosa, 2007).

Tujuan dari seleksi fitur adalah untuk mengidentifikasi fitur dalam dataset yang sama pentingnya, dan membuang fitur lainnya sebagai informasi yang tidak relevan dan

berlebihan. Empat alasan utama untuk melakukan pengurangan dimensi (Oded Maimon, 2010):

1. Decreasing the learning (model) cost (penurunan pembelajaran (model) biaya)
2. Increasing the learning (model) performance (meningkatkan pembelajaran (model) kinerja)
3. Reducing irrelevant dimensions (mengurangi dimensi relevan)
4. Reducing redundant dimensions (mengurangi dimensi berlebihan)



Gambar 2.3 Taksonomi masalah pengurangan dimensi

(Oded Maimon, 2010)

2.1.4 Genetic Algorithm

Genetic Algorithm (Algoritma Genetika) adalah metodologi adaptif untuk mencari optimasi umum berdasarkan analogi langsung ke seleksi alam darwin dan genetika dalam sistem biologi. *Genetic Algorithm* bekerja dengan satu set solusi kandidat yang disebut populasi(C. L. Huang dan Wang 2006)

Proses inti algoritma genetika untuk memecahkan masalah ditunjukkan pada Algoritma 1, di mana n adalah individu penduduk, L adalah jumlah parameter diinginkan, N adalah iterasi, Range berbagai gen individu. Pada penentuan awal Algoritma Genetika, kita perlu menentukan ukuran populasi sesuai dengan jumlah parameter dan persyaratan akurasi, kode parameter dalam string biner, maka kita dapat menghasilkan secara acak populasi awal. Fungsi Pusat (Xi) digunakan untuk menghitung kebugaran dengan fungsi tujuan dan kendala.Fungsi Reproduksi (Xi, fiti), Crossover (Xi + 1), Mutasi (Xi + 1) menunjukkan tiga operasi utama: Reproduksi, Crossover dan Mutasi. Primer operator Reproduksi adalah untuk menekankan solusi yang baik dan menghilangkan solusi yang buruk dalam suatu populasi, sekaligus mempertahankan ukuran populasi konstan. Crossover dan Mutasi operator menghasilkan individu baru dengan konversi genkromosom menurut Crossover (Li, Ziangyang, 2015).

Algorithm 1
Genetic algorithm.

Input: $N, n, L, Range$
Output: X_{N-1}
1: Produce the initial population: a $n \times L$ matrix X_0, $-\text{Range} < X_0[k][j] < \text{Range}$.
2: $i = 0$.
3: **while** $i < N$ **do**
4: Calculate the fitness $fit_i = Fitness(X_i)$.
5: Reproduction operation $X_{i+1} = Reproduction(X_i, fit_i)$.
6: Crossover operation $X_{i+1} = Crossover(X_{i+1})$.
7: Mutation operation $X_{i+1} = Mutation(X_{i+1})$.
8: **end while**

Sumber:

Li, Ziangyang, 2015

Gambar 2.3

2.1.5 Pengujian Evaluasi dan Performa Prediksi

2.1.5.1 Pengujian K-*fold Cross-validation*

K-fold Cross-validation merupakan teknik validasi dengan membagi data awal secara acak kedalam *k* bagian yang saling terpisah atau "*fold*" (Han and Kamber 2007) . Sebuah pendekatan alternatif untuk training dan testing yang sering diterapkan ketika jumlah kasus yang kecil (dan yang banyak yang memilih menggunakan tanpa memandang ukurannya) dikenal sebagai *k-fold cross-validation.*.

2.1.5.2 Metode Evaluasi

Confusion matrix memberikan keputusan yang diperoleh dalam *traning* dan *testing*, *confusion matrix* memberikan penilaian *performance* klasifikas iberdasarkan objek dengan benar atau salah (Gorunescu, 2011). *Confusion matrix* berisi informasi aktual *(actual)* dan prediksi *(predicted)* pada sistem klasifikasi.

3.11. Tinjauan Studi

Model dan hasil penelitian sebelumnya yang dijadikan landasan dalam penelitian yang dilakukan. Hasil Penelitian Terkait yang pernah dilakukan oleh peneliti sebelumya yaitu :

2.1.6 Model Penelitian L.B Romdhane, N. Fadhel dan B. Ayeb.

(Romdhane, et al., 2010) meneliti tentang pendekatan yang efisien untuk membangun profile pelanggan dari data bisnis, menggunakan 18.918 data pelanggan, 15.000 akan digunakan untuk data pelatihan dan 3918 untuk data testing (prediksi), data pelanggan terdiri dari 23 atribut. Langkah pertama, mengelompokkan data dengan algoritma berbasis FCM untuk mengekstrak secara alami kelompok pelanggan. Pada langkah kedua, kita mengurangi jumlah atribut untuk masing-masing kelompok dihitung dari pelanggan dengan memilih hanya paling penting yang untuk kelompok itu.

Dengan menggunakan nilai informasi untuk mengukur pentingnya atribut. hasil dari kedua langkah ini, diperoleh satu set kelompok masing-masing dijelaskan oleh satu set yang berbeda dari atribut(atau karakteristik). Pada langkah ketiga dan terakhir dari model kami, kami membangun satu set profil setiap pelanggan dimodelkan oleh jaringan saraf (*neural*

network) *back propagation* dan dilatih dengan data pada kelompok yang sesuai pelanggan.Terlihat bahwa untuk 2923 pelanggan (yaitu, di 74,60% dari simulasi) model mampu mengenali profil(atau kategori); dan untuk 995 pelanggan(yaitu, di 25,39% dari simulasi) sistem tidak dapat mengenali profil mereka. Di antara profil diakui, 1464 profil(atau 37,37%) ditugaskan untuk profil yang tepat, dan1.459 profil (atau 37,24%) kesalahan klasifikasi. Hasil eksperimen pada sintetis dan besarset data dunia nyata mengungkapkan kinerja yang sangat memuaskan dari pendekatan kami.

2.1.7 Model Penelitian Wen-Chin Chen, Chiun- Chieh Hsu, dan Jing-Ning Hsu

(Chen, et al. 2011)meneliti tentang Seleksi Optimasi dari berbagai potensi pelanggan melalui kesatuan pola secara berurutan(sequential patern) dengan menggunakan model respon. Memperoleh pelanggan potensial yang melibatkan variabel perilaku dengan ponsel dan akses internet ponsel. Menggunakan data pelatihan 40% dan data pengujian 60% dari ideal potensial pelanggan yang dipilih secara acak. Dengan 3 model prediksi pelanggan potensial yaitu algoritma C5.0, Support Vector Machine dan Jaringan Saraf Tiruan (NN). SVM menunjukkan kinerja yang nyata lebih baik dibandingan dengan lainnya dengan akurasi 775.55%, NN akurasi 66.97% dan C5.0 62.06%. Menunjukkan bahwa akurasi dari model yang diusulkan dapat lebih ditingkatkan dengan memilih calon yang efektif berbagai pelanggan dengan menggunakan SVM.

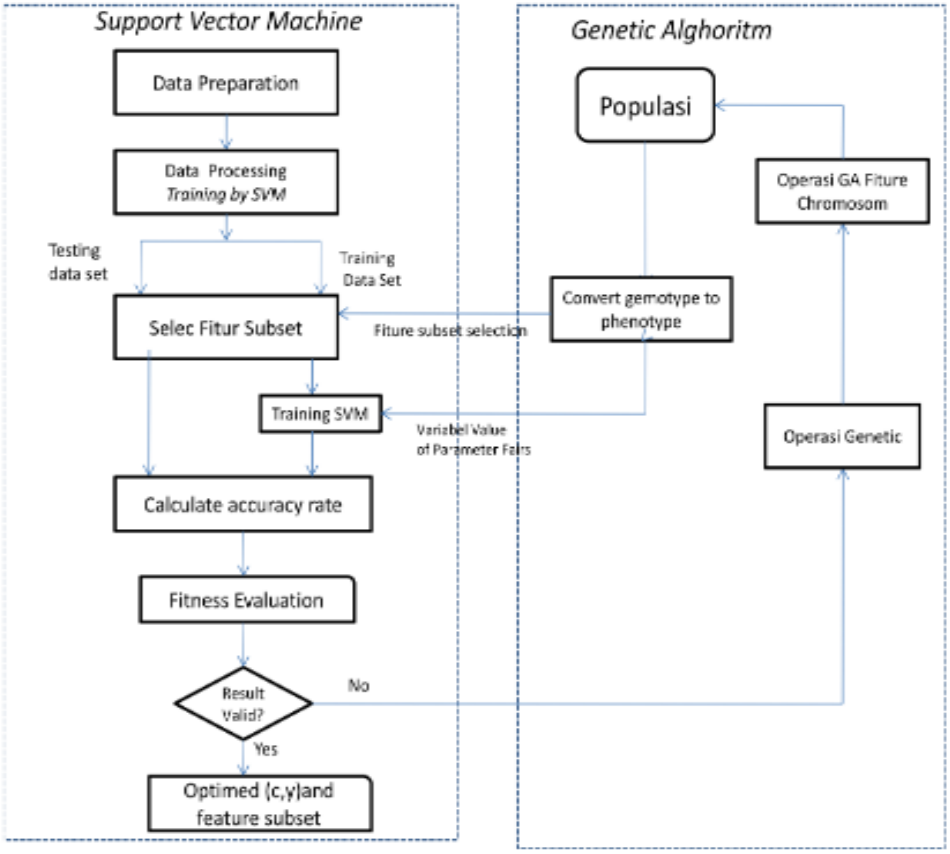
2.1.8 Model Penelitian Sergio Moro dan Raul M.S Laureano

(Moro dan Laureano, 2011) Yaitu meneliti tentang pelanggan deposito bank pada bank portugis, yang bertujuan untuk menemukan model yang mampu menjelaskan kesuksesan suatu kontak pelanggan. Model tersebut dapat meningkatkan efisiensi promosi dengan mengidentifikasi karakteristik utama yang mempengaruhi kesuksesan,membantu dalam manajemen yang lebih baik dari sumber daya yang tersedia(misalnya usaha manusia, panggilan telepon, waktu)dan seleksi yang berkualitas tinggi dan terjangkau pada sekumpulan potensi nasabah. Pada penelitian ini menggunakan tiga model untuk dijadikan perbandingan yaitu *Naïve Bayesian, Decision Tree* dan SVM. Setelah dilakukan pengujian didapat nilai AUC pada NB 0.870, DT sebesar 0.868, dan SVM sebesar 0.938. Model terbaik,diwujudkan oleh Support Vector Machine(SVM), mencapai prestasi prediksi yang tinggi. Menggunakan analisis sensitivitas, kami mengukur pentingnya masukan dalam model SVM dan pengetahuan tersebut dapat digunakan oleh para manajer untuk meningkatkan

pemasaran(misalnya dengan meminta agen untuk meningkatkan durasi panggilan telepon mereka atau pemasaran penjadwalan pada bulan tertentu).

3.12. Kerangka Pemikiran

Model kerangka pemikiran yang digunakan adalah adalah *method improvement* (perbaikan metode), yang sering digunakan pada penelitian di bidang sains dan teknik, termasuk bidang computing didalamnya. Pada peneilitian ini, data set yang digunakan adalah *Bank Marketing* UCI dataset.



Gambar 2.10 Kerangka Pemikiran

BAB 3
METODE PENELITIAN

3.1 Desain Penelitian

Menurut (Berndtsson et al., 2008) ada empat metode penelitian yang digunakan yaitu tindakan penelitian, studi kasus, eksperimen dan survey. Dalam penelitian ini dilakukan beberapa langkah yang dilakukan dalam proses penelitian yaitu :

1. Pengumpulan data

   Pada bagian ini dijelaskan tentang bagaimana dan darimana data dalam penelitian ini didapatkan, ada dua tipe dalam pengumpulan data, yaitu pengumpulan data primer dan pengumpulan data sekunder. Pada tahap ini ditentukan data yang akan diproses. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua data dalam data set, termasuk variabel yang diperlukan dalam proses.

2. Pengolahan data awal

   Pada bagian dijelaskan tentang tahap awal *data mining.* Pengolahan awal data meliputi proses input data ke format yang dibutuhkan, pengelompokan dan penentuan atribut data, serta pemecahan data *(split*) untuk digunakan dalam proses pembelajaran *(training)* dan pengujian *(testing).*

3. Metode yang diusulkan

   Pada tahap ini data dianalisis, dikelompokan variabel mana yang berhubungan dengan satu sama lainnya. Setelah data dianalisis lalu diterapkan model-model yang sesuai dengan jenis data. Pembagian data kedalam data latihan *(training data)* juga diperlukan untuk pembuatan model.

4. Eksperimen dan pengujian metode

   Pada bagian ini dijelaskan tentang langkah-langkah eksperimen meliputi cara pemilihan arsitektur yang tepat dari model atau metode yang diusulkan sehingga didapatkan hasil yang dapat membuktikan bahwa metode yang digunakan adalah tepat.

5. Evaluasi dan validas

   Pada tahap ini dilakukan evaluasi dan validasi hasil penerapan terhadap model penelitian yang dilakukan untuk mengetahui tingkat keakurasian model.

3.2 Pengumpulan Data

Pada penelitian ini, data yang digunakan adalah data sekunder karena diperoleh dari data *Bank Marketing* dalam UCI *machine learning repository*. Dimana data tersebut berisi 45212 data nasabah.

Parameter yang memenuhi kriteria nasabah yang digunakan dalam penelitian ini berjumlah 15 parameter dengan kriterianya yaitu *age* (umur), *job* (pekerjaan nasabah), *education* (pendidikan), *marital status* (status perkawinan), *annual balance* (saldo tahunan), *housing* (kepemilikan rumah), *loans in delay* (tunggakan pinjaman), *contact* (jenis kontak yang dapat dihubungi), *day* (tanggal marketing), *month* (bulan marketing), *duration* (lamanya dihubungi), *campaign* (promosi), *pdays* (promosi perhari), *previous* (promosi sebelumnya) dan *poutcome* (hasil sebelumnya).

Tabel 3.1

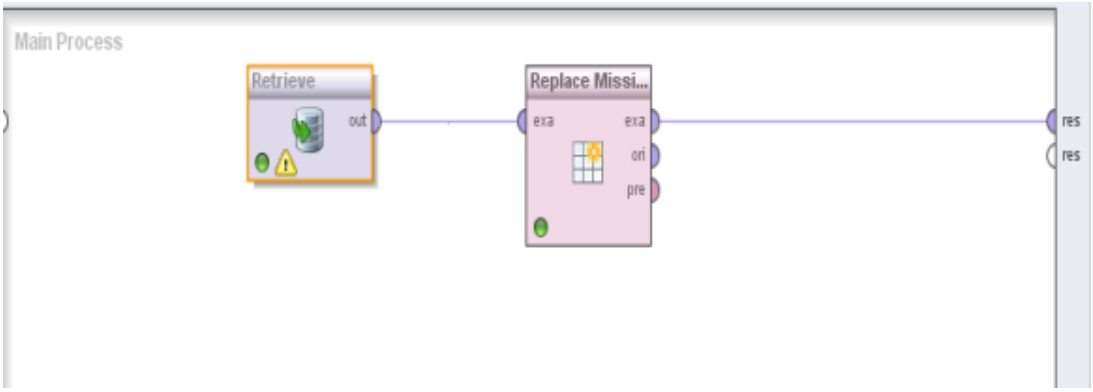
Bank Marketing Data

No	age	job	marital	education	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	Y
1	58	anageme	married	tertiary	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
2	44	technician	single	secondary	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
3	33	ntreprene	married	secondary	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
4	47	blue-colla	married	unknown	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
5	33	unknown	single	unknown	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
6	35	anageme	married	tertiary	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
7	28	anageme	single	tertiary	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
8	42	ntreprene	divorced	tertiary	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
9	58	retired	married	primary	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
10	43	technician	single	secondary	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
11	41	admin.	divorced	secondary	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
12	29	admin.	single	secondary	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
13	53	technician	married	secondary	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
14	58	technician	married	unknown	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
15	57	services	married	secondary	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no

Data pada setiap dataset yang tidak memiliki nilai akan dihapus dan tidak digunakan. Untuk mendapatkan data yang berkualitas, beberapa teknik yang dilakukan sebagai berikut (Vercellis, 2009):

1. Data validation, kualitas *input* data dapat membuktikan tidak memuaskan karena ketidak lengkapan, kebisingan dan inkonsistensi. Dengan cara mengidentifikasi, memperbaiki dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten dan data yang tidak lengkap (*Missing Value*).

2. Data *integration and transformation*, untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai numeric. Data ditransformasikan kedalam *software Rapidminer*.
3. *Data size reduction and discritization*, untuk memperoleh data set dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat informative.



Gambar 3.1
Model Replace Missing

Tabel 3.2

Model Replace Missing

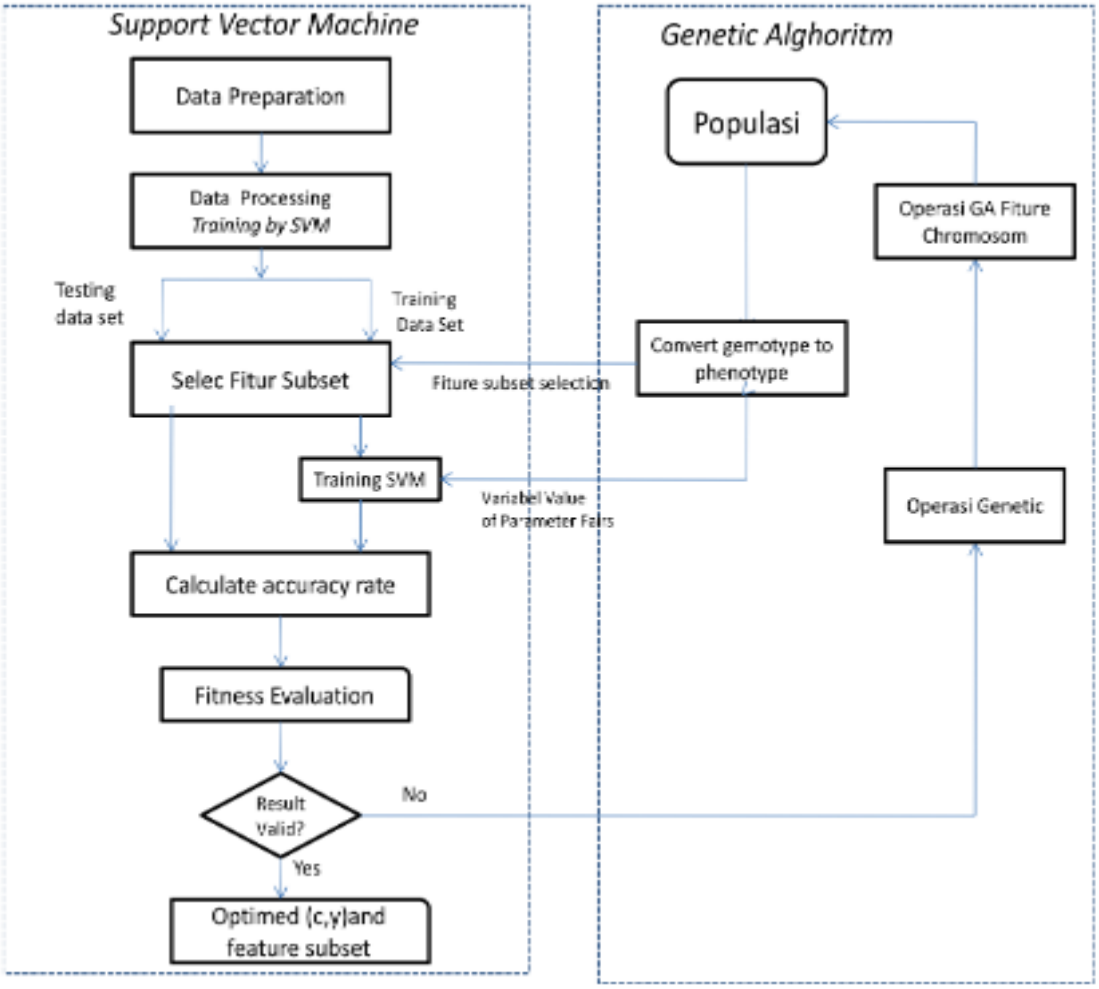
Meta Data View | Data View | Plot View | Advanced Charts | Annotations

ExampleSet (200 examples, 1 special attribute, 16 regular attributes)

Role	Name	Type	Statistics	Range	Missings
label	Y	binominal	mode = no (185), least = yes ('	no (185), yes (15)	0
regular	age	integer	avg = 45.505 +/- 10.526	[23.000 ; 78.000]	0
regular	job	polynominal	mode = blue-collar (43), least =	management (34), technician (	0
regular	marital	polynominal	mode = married (130), least =	married (130), single (41), divo	0
regular	education	polynominal	mode = secondary (111), least	tertiary (40), secondary (111), u	0
regular	default	binominal	mode = no (198), least = yes (2	no (198), yes (2)	0
regular	balance	integer	avg = 560.585 +/- 1430.697	[-674.000 ; 12223.000]	0
regular	housing	binominal	mode = yes (167), least = no (2	yes (167), no (33)	0
regular	loan	binominal	mode = no (169), least = yes (2	no (169), yes (31)	0
regular	contact	binominal	mode = unknown (191), least =	unknown (191), cellular (9)	0
regular	day	integer	avg = 5.485 +/- 2.199	[5.000 ; 16.000]	0
regular	month	binominal	mode = may (191), least = sep	may (191), sep (9)	0
regular	duration	integer	avg = 308.350 +/- 313.640	[13.000 ; 2033.000]	0
regular	campaign	integer	avg = 1.235 +/- 0.549	[1.000 ; 5.000]	0
regular	pdays	integer	avg = 9.985 +/- 81.865	[-1.000 ; 792.000]	0
regular	previous	integer	avg = 0.075 +/- 0.593	[0.000 ; 7.000]	0
regular	poutcome	binominal	mode = unknown (199), least =	unknown (199), other (1)	0

3.3 Metode yang diusulkan

Pada tahap modeling ini dilakukan pemprosesan data traning (90%) dan data testing (10%) sehingga akan membahas metode algoritma yang diuji dengan memasukan data bank marketing kemudian di analisa dan dikomparasi. Berikut ini bentuk gambaran tahapan pengujian.



Sumber:Devos, Downey, dan Duponchel,2014

Gambar 3.2

Tahapan Penelitian

3.4 Eksperimen dan Pengujian Metode

Pada penelitian ini dilakukan proses eksperimen dan pengujian model menggunakan dataset *Bank Marketing* UCI *Repository* pada aplikasi *RapidMiner*5. Dalam penelitian eksperimen digunakan spesifikasi software dan hardware sebagai alat bantu dalam penelitian ini pada Tabel 3.2

Tabel 3.3.

Spesifikasi software dan hardware

Software	Hardware
Sistem Operasi : Microsoft Windows 7	CPU : Intel Pentium Core i3
Data Mining : Rapid Miner Versi 5	Memory : 2 GB
	Hardisk : 320 GB

3.5 Evaluasi dan Validasi Hasil

Hasil akhir dari penelitian ini adalah kegiatan valiadasi terhadap model yang digunakan, validasi ini dilakukan untuk menguji terhadap model prediksi yang dianggap paling optimal dengan prediksi kesalahan pada peningkatan nilai akurasi.

BAB 4
HASIL PENELITIAN DAN PEMBAHASAN

4.1 Tahapan Pengujian

Berikut merupakan tahapan dari pengujian yang dilakukan:

1) Data Preparation

Fase ini merupakan proses penyiapan data yang akan di olah di SVM dan semua data harus lengkap dengan variable dan datanya. Data diambi dari UCI Dataset *Bank Marketing.* Data yang sudah siap akan di import ke rapid miner dan kemudian di olah dengan SVM

2) Data Processing

Tahapan ini akan dilakukan proses pembersihan data yang sudah di import ke rapid miner, kemudian data ini akan di testing dan di training dengan SVM apakah data tersebut layak untuk di uji atau tidak, jika masih ada data yang missing maka harus di cleaning data missing tersebut dengan SVM Setelah melakukan testing data sampai data tersebut layak untuk dilanjutkan pengujian selanjutnya.

3) Select Fitur Subset

Tahapan ini akan dilakukan penyeleksian set fitur terlebih dengan menkonversi gemotype dan phenotype dan testing 10% data set untuk seleksi fitur didalam GA dan kemudian 90% dataset akan di training didalam untuk seleksi fitur dalam SVM, kemudian menghitung akurasi yang didapakan.

4) Variable Value Of Parameter pair

Pada tahapan ini akan dilakukan pengoptimalan parameter C dan y untuk mendapatkan nilai parameter dari variable/fitur yang paling fair. Dan kemudian setelah parameter dan y di optimalkan akan di training kembali di dalam SVM untuk melihat performance dan akurasi yang di dapatkan.

5) Fitness Evaluation

Tahapan ini akan dilakukan evaluasi dengan cara mem-validasi fitur yang diseleksi yaitu dengan *K-fold Cross-validation* yang ada didalam SVM untuk mengetahui performance dan akurasi dari pengujian yang telah dilakukan

6) Operasi *Genetic Feature Chromosom* dan SVM

Tahapan ini akan dilakukan pengujian antara seleksi fitur dan seleksi parameter yang ada dalam algoritma GA dan kemudian akan di training kembali

dengan SVM kemudian di validasi dan akan menghasilkan performance dan akurasi hasil akhir dari pengujian dengan SVM dengan GA

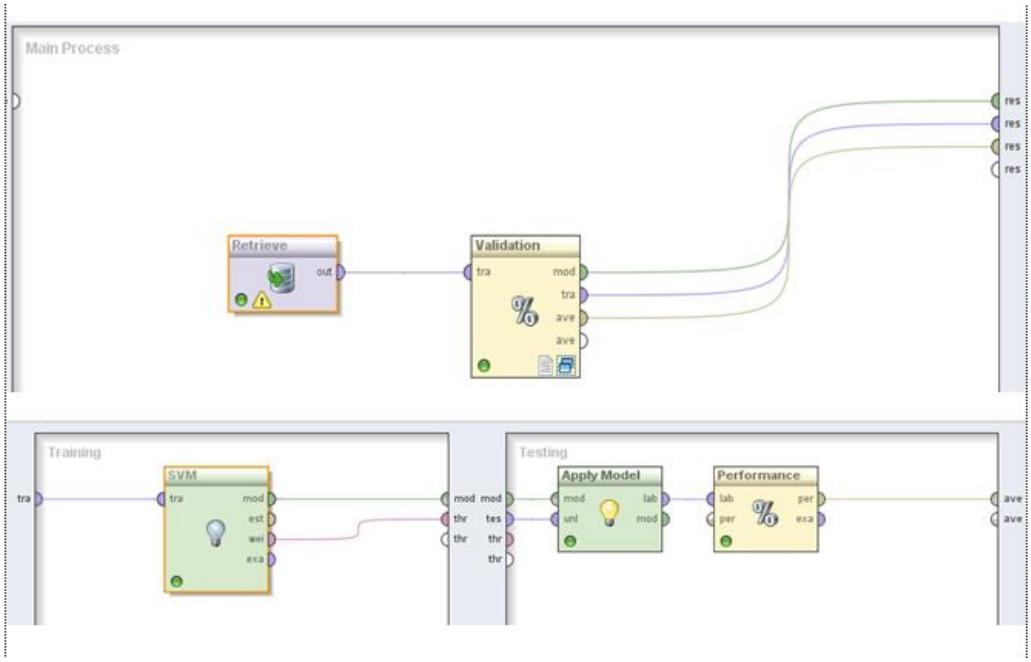
7) Result/optimed (C,y) and feature subset

Tahapan ini akan terihat hasil akhir dari hasil dari pengujian data yang sudah dilakukan setelah parameter C,y di optimalkan dan akan terlihat hasil performance/akurasi hasil akhir dari pengujian data dengan SVM berbasis GA.

4.2 Hasil Eksperimen dan Pengujian Metode

4.2.1 Metode Support Vector Machine

Pengujian pertama adalah menguji data dengan menggunakan metode Support Vector Machine, data yang digunakan adalah data testing. Software yang digunakan untuk pengujian adalah software *RapidMiner*.



Gambar 4.1

Pengujian *K-Fold Cross Validation* algoritma *Support Vector Machine*

Setelah dilakukan pengujian dengan metode support vector machine maka didapat nilai akurasi sebesar 85%.

accuracy: 85.50% +/- 10.83% (mikro: 85.50%)

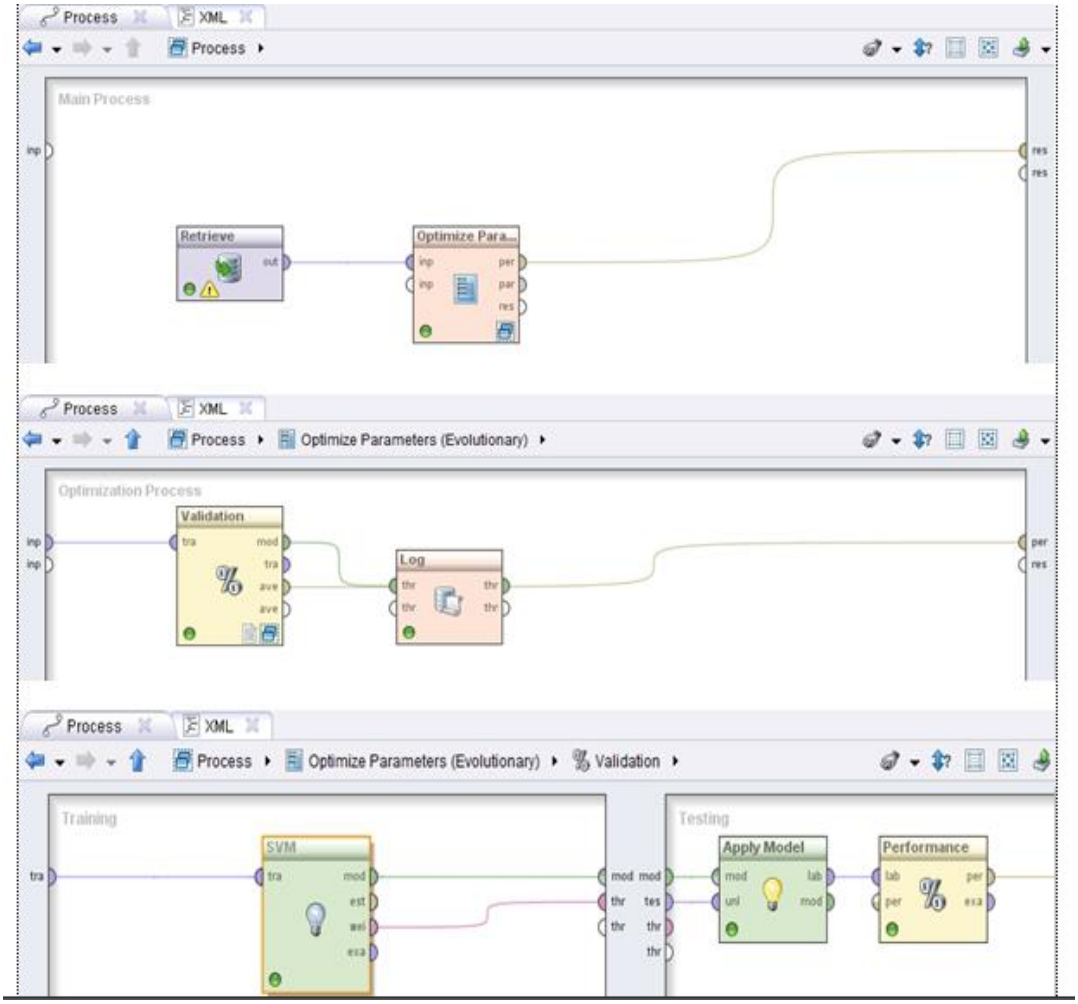
	true no	true yes	class precision
pred. no	161	5	96.99%
pred. yes	24	10	29.41%
class recall	87.03%	66.67%	

Gambar 4.2

Hasil Akurasi Pengujian Metode Support Vector Machine

4.2.2 Parameter Support Vector Machines berbasis Genetic Algorithm

Berikut gambar pengujian algoritma *Support Vector Machine* berbasis *Genetic Algorithm* untuk optimasi parameter menggunakan *K-Fold Cross Vlidation* dengan menggunakan *Rapid Miner* :



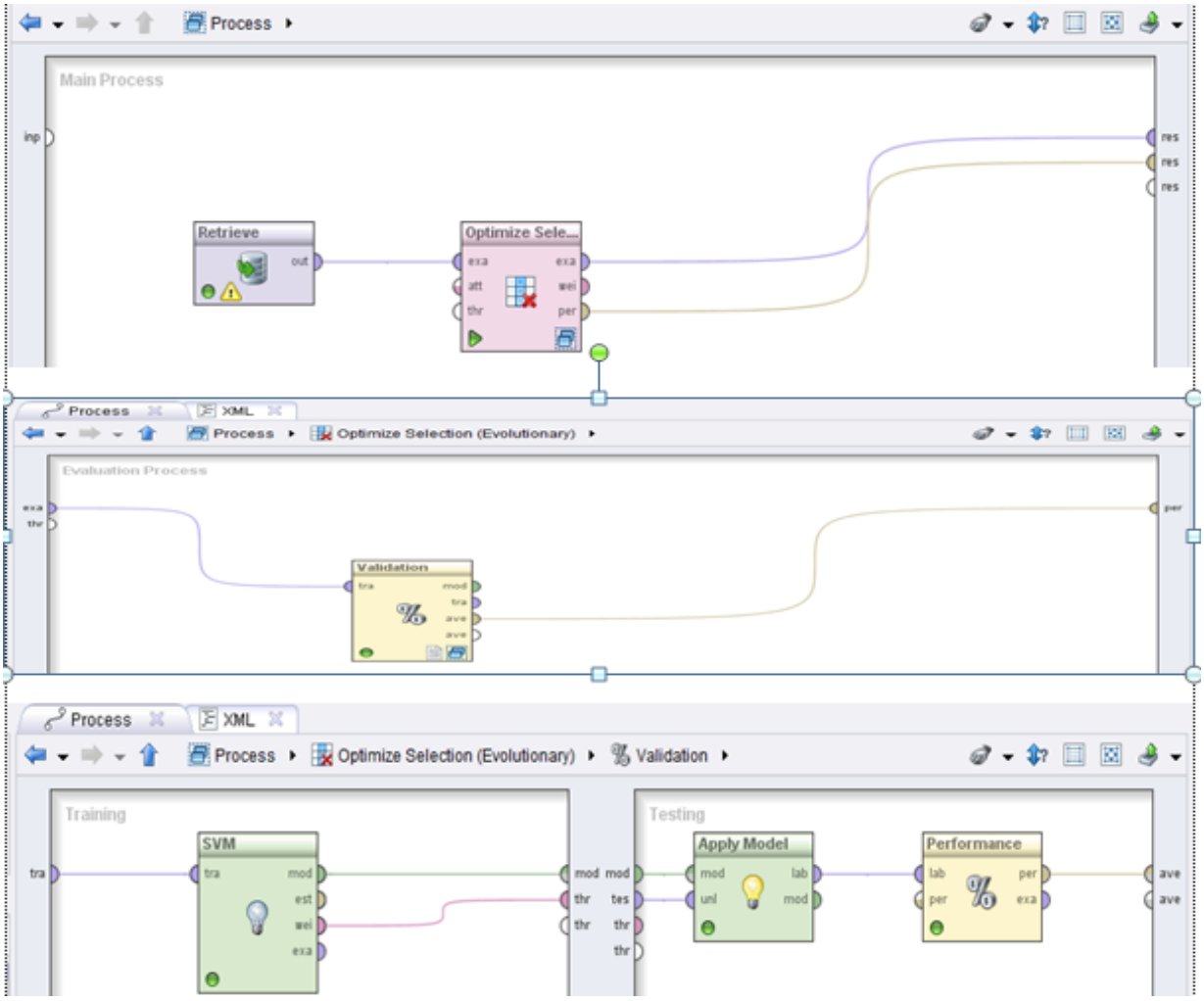
Gambar 4.3

Pengujian Optimasi Parameter *Support Vector Machine* berbasis Genetic Algorithm

Untuk meningkatkan kinerja prediksi maka dilakukan uji coba untuk pemilihan parameter dengan menggunakan Genetic Algorithm yaitu paramater C, γ dan Epsilon. Setelah dilakukan pengujian maka didapatkan nilai akurasi sebesar 87%.

4.2.3 Seleksi Fitur Support Vector Machines berbasis Genetic Algorithm

Berikut gambar pengujian algoritma *Support Vector Machine* berbasis *Genetic Algorithm* untuk seleksi fitur menggunakan *K-Fold Cross Vlidation* dengan menggunakan *Rapid Miner* :



Gambar 4.4

Pengujian *Feature Selection Support Vector Machine* berbasis Genetic Algorithma

Dalam penelitan ini penulis melakukan seleksi fitur atau atribut yang digunakan yaitu 15 parameter dengan kriterianya yaitu *age* (umur), *job* (pekerjaan nasabah), *education* (pendidikan), *marital status* (status perkawinan), *annual balance* (saldo tahunan), *housing* (kepemilikan rumah), *loans in delay* (tunggakan pinjaman), *contact* (jenis kontak yang dapat dihubungi), *day* (tanggal marketing), *month* (bulan marketing), *duration* (lamanya dihubungi), *campaign* (promosi), *pdays* (promosi perhari), *previous* (promosi sebelumnya) dan *poutcome* (hasil sebelumnya).

Meta Data View | Data View | Plot View | Advanced Charts | Annotations

ExampleSet (200 examples, 1 special attribute, 16 regular attributes)

Role	Name	Type	Statistics	Range	Missings
label	Y	binominal	mode = no (185), least = yes ('	no (185), yes (15)	0
regular	age	integer	avg = 45.505 +/- 10.526	[23.000 ; 78.000]	0
regular	job	polynominal	mode = blue-collar (43), least =	management (34), technician (	0
regular	marital	polynominal	mode = married (130), least =	married (130), single (41), divo	0
regular	education	polynominal	mode = secondary (111), least	tertiary (40), secondary (111), u	0
regular	default	binominal	mode = no (198), least = yes (2	no (198), yes (2)	0
regular	balance	integer	avg = 560.585 +/- 1430.697	[-674.000 ; 12223.000]	0
regular	housing	binominal	mode = yes (167), least = no (1	yes (167), no (33)	0
regular	loan	binominal	mode = no (169), least = yes (1	no (169), yes (31)	0
regular	contact	binominal	mode = unknown (191), least =	unknown (191), cellular (9)	0
regular	day	integer	avg = 5.485 +/- 2.199	[5.000 ; 16.000]	0
regular	month	binominal	mode = may (191), least = sep	may (191), sep (9)	0
regular	duration	integer	avg = 308.350 +/- 313.640	[13.000 ; 2033.000]	0
regular	campaign	integer	avg = 1.235 +/- 0.549	[1.000 ; 5.000]	0
regular	pdays	integer	avg = 9.985 +/- 81.865	[-1.000 ; 792.000]	0
regular	previous	integer	avg = 0.075 +/- 0.593	[0.000 ; 7.000]	0
regular	poutcome	binominal	mode = unknown (199), least =	unknown (199), other (1)	0

Gambar 4.5

***Feature Selection* sebelum dilakukan pengujian**

Dari 7 variabel prediktor dilakukan seleksi atribut atau fitur sehingga menghasilkan terpilihnya pendidikan (education), saldo tahunan (balance), kepemilikan (housing), tunggakan pinjaman (loan), lamanya komunikasi (duration), promosi perhari (pdays), promosi sebelumnya (previous), hasil sebelumnya(poutcome), dengan nilai akurasi yaitu 88%. Sedangkan atribut atau fitur lainnya seperti tidak berpengaruh terhadap bobot atribut.

ExampleSet (200 examples, 1 special attribute, 8 regular attributes)

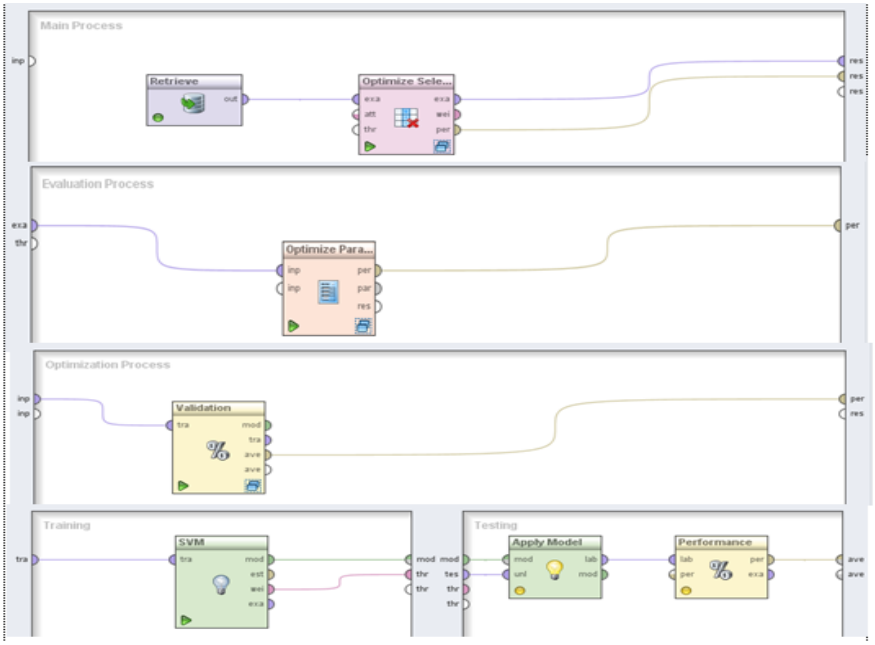
Role	Name	Type	Statistics	Range	Missings
label	Y	binominal	mode = no (126), least :	no (126), yes (74)	0
regular	education	polynominal	mode = secondary (101	tertiary (80), secondary (	0
regular	balance	integer	avg = 1313.860 +/- 2257	[-887.000 ; 17023.000]	0
regular	housing	binominal	mode = no (113), least :	yes (87), no (113)	0
regular	loan	binominal	mode = no (181), least :	no (181), yes (19)	0
regular	duration	integer	avg = 276 +/- 237.603	[9.000 ; 1556.000]	0
regular	pdays	integer	avg = 37.475 +/- 86.638	[-1.000 ; 530.000]	0
regular	previous	integer	avg = 0.900 +/- 2.134	[0.000 ; 12.000]	0
regular	poutcome	polynominal	mode = unknown (150),	unknown (150), succes:	0

Gambar 4.6

***Feature Selection* sesudah dilakukan pengujian**

4.2.4 Seleksi Fitur dan Parameter Support Vector Machines berbasis Genetic Algorithm

Setelah dilakukan seleksi fitur dan parameter, untuk mendapatkan nilai *RMSE* yang optimal langkah berikutnya dilakukan pengujian terhadap penerapan genetika algoritma untuk seleksi fitur dan parameter. Berikut gambar pengujian algoritma *Support Vector Machine* berbasis *Genetic Algorithm* untuk seleksi fitur dan parameter menggunakan dengan menggunakan *Rapid Miner*



Gambar 4.7

Pengujian *Feature Selection* dan *Parameter Support Vector Machine* berbasis

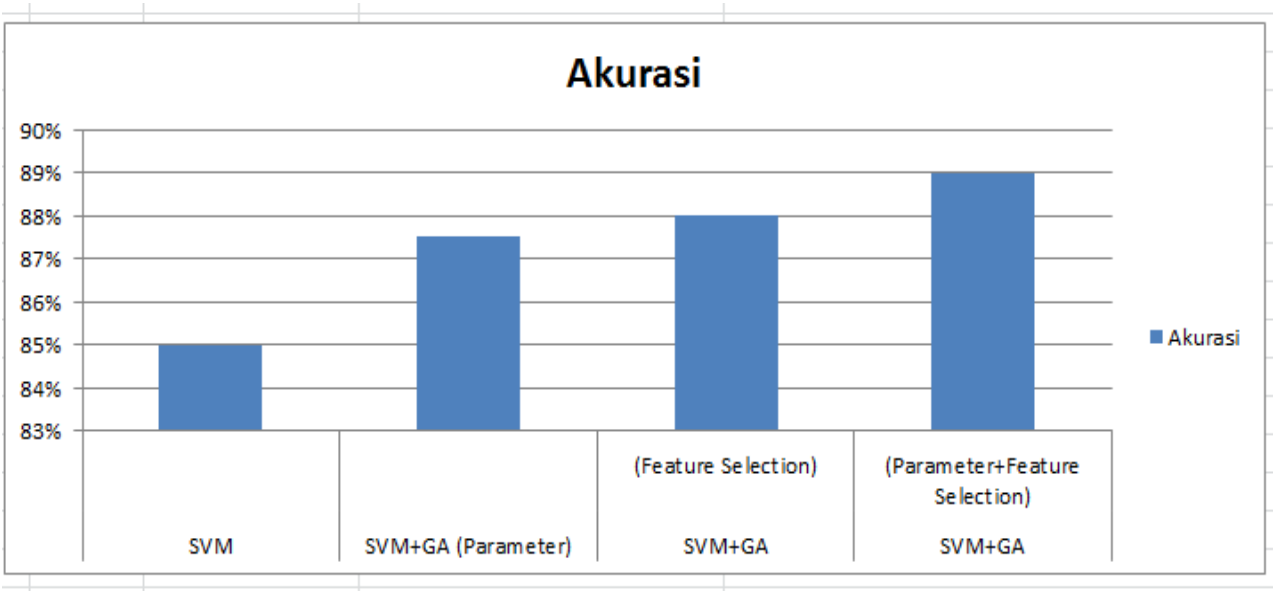
Genetic Algorithm

Setelah dilakukan pengujian maka didapatkan nilai akurasi sebesar 88.5% Dari hasil pengujian diatas dengan mengukur nilai kesalahan atau *RMSE* terbukti bahwa hasil pengujian dengan *support vector machine* berbasis algoritma genetika memiliki *RMSE* dengan nilai error kesalahan yang paling kecil kurang dibandingan dengan nilai algoritma SVM. Dapat dilihat pada tabel 4.1 dibawah ini :

Tabel 4.1 Tabel Hasil Pengujian Metode

	SVM	SVM+GA (Parameter)	SVM+GA (Feature Selection)	SVM+GA (Parameter+Feature Selection)
Akurasi	85%	88%	88%	89%

Perbandingan akurasi *RMSE* berdasarkan tabel 4.1 dapat diilustrasikan dalam grafik maka akan tampak pada gambar 4.7 sebagai berikut :



Gambar 4.8 Grafik Hasil Pengujian Metode

Dengan demikian algoritma support vector machine berbasis algoritma genetika dapat memberikan solusi permasalahan dalam memprediksi nasabah telemarketing yang mempunyai parameter potensial terhadap bank. Dengan nilai akurasi lebih tinggi.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Pengujian dengan mengunakan *support vector machine* berbasis *Genetic Algorithm* untuk seleksi atribut dengan penyesuaian pada parameter C dan γ sebagai berikut:

1. Hasil eksperimen pengujian data *bank marketing* UCI data set,dengan *support vector machine* sebelum dan sesudah dilakukan seleksi atribut dan optimasi parameter dengan algoritma genetika, menunjukkan bahwa ada perbedaan yang signifikan pada akurasinya.
2. Penerapan model SVM dengan Algoritma Genetika untuk seleksi fitur dan optimasi parameter terbukti meningkatkan akurasi dalam prediksi nasabah pada bank telemarketing dengan peningkatan nilai akurasi yang semula dari 85% menjadi 89%.

5.2 Saran

Pada penelitian ini secara umum penerapan model GA-SVM dapat meningkatkan akurasi prediksi potensi nasabah, akan tetapi karena keterbatasan penelitian ini perlu disarankan untuk melakukan penelitian lanjutan yang berkaitan dengan prediksi untuk mendapatkan akurasi yang lebih baik. Adapun saran-saran yang perlu diberikan yaitu:

Bank Marketing data merupakan data yang diambil dari UCI Repository dimana penerapanya disesuaikan dengan kondisi parameter nasabah didunia, penelitian ini diharapkan dapat dijadikan acuan bagi pemilihan nasabah sebagai strategi untuk mendapatkan nasabah potensial yang datanya dapat bermanfaat bagi marketing bank.

DAFTAR PUSTAKA

Aydin, Ilhan, Mehmet Karakose, and Erhan Akin. 2011. *A Multi-Objective Artificial Immune Algorithm for Parameter Optimization in Support Vector Machine*. *Applied Soft Computing* 11(1): 120–29.

Berndtsson, Hansson, Olsson, and Lundell. 2008. *Thesis Projects*.

Chen, Wen-Chin, Chiun-Chieh Hsu, and Jing-Ning Hsu. 2011. *Optimal Selection of Potential Customer Range through the Union Sequential Pattern by Using a Response Model*. *Expert Systems with Applications* 38(6): 7451–61. http://linkinghub.elsevier.com/retrieve/pii/S0957417410014417 (October 25, 2013).

Dehuai, Zeng et al. 2012. *Wick Sintered Temperature Forecasting Based on Support Vector Machines with Simulated Annealing*. *Physics Procedia* 25: 427–34. http://linkinghub.elsevier.com/retrieve/pii/S1875389212005238.

Devos, Olivier, Gerard Downey, and Ludovic Duponchel. 2014. *Simultaneous Data Pre-Processing and SVM Classification Model Selection Based on a Parallel Genetic Algorithm Applied to Spectroscopic Data of Olive Oils*. *Food Chemistry* 148: 124–30. http://linkinghub.elsevier.com/retrieve/pii/S0308814613014520.

Han, Jiawei, and Micheline Kamber. 2007. *Data Mining Concepts and Techniques*.

Huang, Cheng Lung, and Chieh J. Wang. 2006. "A GA-Based Feature Selection and Parameters Optimizationfor Support Vector Machines." *Expert Systems with Applications* 31(2): 231–40.

Huang, Cheng-Lung, and Chieh-Jen Wang. 2006. "A GA-Based Feature Selection and Parameters Optimizationfor Support Vector Machines." *Expert Systems with Applications* 31(2): 231–40.

Ilhan, Ilhan, and Gülay Tezel. 2013. "A Genetic Algorithm-Support Vector Machine Method with Parameter Optimization for Selecting the Tag SNPs." *Journal of Biomedical Informatics* 46(2): 328–40.

Liao, Shu-hsien, Yin-ju Chen, and Hsin-hua Hsieh. 2011. "Mining Customer Knowledge for Direct Selling and Marketing." *Expert Systems with Applications* 38(5): 6059–69. http://linkinghub.elsevier.com/retrieve/pii/S0957417410012443 (October 25, 2013).

Lin, Shih-Wei, Shih-Chieh Chen, Wen-Jie Wu, and Chih-Hsien Chen. 2009. "Parameter Determination and Feature Selection for Back-Propagation Network by Particle Swarm Optimization." *Knowledge and Information Systems* 21(2): 249–66.

Ludermir, Teresa B., Marcilio C.P. de Souto, and Marley Vellasco. 2012. "Automatic Parameters Selection in Machine Learning." *Neurocomputing* 75(1): 1–2.

Machairas, Vasileios, Aris Tsangrassoulis, and Kleo Axarli. 2014. “Algorithms for Optimization of Building Design: A Review.” *Renewable and Sustainable Energy Reviews* 31(1364): 101–12. http://linkinghub.elsevier.com/retrieve/pii/S1364032113007855.

Moro, Sérgio, and Raul M S Laureano. 2011. *Using Data Mining for Bank Direct Marketing : An Application of the CRISP-DM Methodoody*

Oded Maimon, Lior Rokach. 2010. *Data Mining & Knowledge*.

Oded Maimon;Lior Rokach. 2010. Data Mining and Knowledge Discovery Handbook *Data Cleansing Data Mining and Knowledge*.

Romdhane, L.B., N. Fadhel, and B. Ayeb. 2010. *An Efficient Approach for Building Customer Profiles from Business Data*. *Expert Systems with Applications* 37(2): 1573–85.

Vajiramedhin, Chakarin. 2014. *Feature Selection with Data Balancing for Prediction of Bank Telemarketing*. 8(114): 5667–72.

Wang, Shuzhou, and Bo Meng. 2011. *Parameter Selection Algorithm for Support Vector Machine. Procedia Environmental Sciences* 11: 538–44.

Weiwen, Xiong, Chen Liang, Zhang Zhiyong, and Qiu Zhuqiang. 2008. *RFM Value and Grey Relation Based Customer Segmentation Model in the Logistics Market Segmentation*. *2008 International Conference on Computer Science and Software Engineering* (1): 1298–1301. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4723147.

Witten, Frank, Hall. 2011. *Data Mining*.

Wu, Jia et al. 2015. *Self-Adaptive Attribute Weighting for Naive Bayes Classification*. *Expert Systems with Applications* 42(3): 1487–1502. http://www.sciencedirect.com/science/article/pii/S0957417414005582.

Xing, Bi, and Wang Xin-feng. 2010. *The Evaluation of Customer Potential Value Based on Prediction and Cluster Analysis*. *2010 International Conference on Management Science & Engineering 17th Annual Conference Proceedings*: 613–18. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5719866.

Yukun Bao, Zhongyi Hu, Tao Xiong. 2013. *A PSO and Pattern Search Based Memetic Algorithm 1 for SVMs Parameters Optimization. Journal of Chemical Information and Modeling* 53: 160.

Yuxia, Hu, and Zhang Hongtao. 2012. *Chaos Optimization Method of SVM Parameters Selection for Chaotic Time Series Forecasting*. *Physics Procedia* 25: 588–94.

Zameer, Aneela, Sikander M. Mirza, and Nasir M. Mirza. 2014. *Core Loading Pattern Optimization of a Typical Two-Loop 300 MWe PWR Using Simulated Annealing (SA), Novel Crossover Genetic Algorithms (GA) and Hybrid GA(SA) Schemes. Annals of Nuclear Energy* 65: 122–31. http://dx.doi.org/10.1016/j.anucene.2013.10.024.

Zhao, Mingyuan et al. 2011. *Feature Selection and Parameter Optimization for Support Vector Machines: A New Approach Based on Genetic Algorithm with Feature Chromosomes*. *Expert Systems with Applications* 38(5): 5197–5204.