KOMPARASI ANTARA *SUPPORT VECTOR MACHINE* DAN *K-NEAREST NEIGHBOR* DALAM PENENTUAN PEMBERIAN KREDIT TERHADAP KONSUMEN

TESIS

ESTER ARISAWATI
14000335

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
NUSA MANDIRI
JAKARTA
2012

KOMPARASI ANTARA *SUPPORT VECTOR MACHINE* DAN *K-NEAREST NEIGHBOR* DALAM PENENTUAN PEMBERIAN KREDIT TERHADAP KONSUMEN

TESIS

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

ESTER ARISAWATI
14000335

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
NUSA MANDIRI
JAKARTA
2012

SURAT PERNYATAAN ORISINALITAS

Yang bertanda tangan di bawah ini :

Nama	:	Ester Arisawati
NIM	:	14000335
Program Studi	:	Magister Ilmu Komputer
Jenjang	:	Strata Dua (S2)
Konsentrasi	:	*Management Information System*

Dengan ini menyatakan bahwa tesis yang telah saya buat dengan judul: "Komparasi antara *Support Vector Machine* dan *k-Nearest Neighbor* dalam Penentuan Pemberian Kredit Terhadap Konsumen" 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, 03 Mei 2012
Yang menyatakan,

Materai Rp. 6.000,-

Ester Arisawati

HALAMAN PENGESAHAN

Tesis ini diajukan oleh :

Nama	:	Ester Arisawati
NIM	:	14000335
Program Studi	:	Magister Ilmu Komputer
Jenjang	:	Strata Dua (S2)
Konsentrasi	:	*Management Information*
Judul Tesis	:	"Komparasi antara *Support Vector Machine* dan *k-Nearest Neighbor* dalam Penentuan Pemberian Kredit Terhadap Konsumen"

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, 03 Mei 2012
Pascasarjana Magister Ilmu Komputer
STMIK Nusa Mandiri
Direktur

H. Mochamad Wahyudi, MM, M.Kom

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	:	Dana Indra Sensuse, Ph.D	.....

KATA PENGANTAR

Puji syukur, penulis panjatkan kehadirat Tuhan Yesus Kristus, yang oleh karena kasih dan anugerahNya, 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 "Komparasi antara *Support Vector Machine* dan *k-Nearest Neighbor* dalam Penentuan Pemberian Kredit Terhadap Konsumen".

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 analisa kredit pada PT AEON Credit Service Indonesia Cabang Tangerang, Banten. Penulis juga lakukan mencari dan menganalisa berbagai macam sumber referensi, baik dalam bentuk jurnal ilmiah, buku-buku literatur, *internet*, dll 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. Bapak Dana Indra Sensuse Ph.D selaku pembimbing tesis yang telah menyediakan waktu, pikiran dan tenaga dalam membimbing penulis dalam menyelesaikan tesis ini.
2. Kepala Cabang PT AEON Credit Service Indonesia Cabang Tangerang, Banten yang telah mengijinkan penulis melakukan riset untuk mendapatkan data atau informasi yang penulis butuhkan.
4. Orang tua dan keluarga tercinta yang telah memberikan dukungan material dan moral kepada penulis.
5. Andreas Defrianto, suamiku tercinta yang telah memberikan dukungan dan semangat kepada penulis.

6. Seluruh staf pengajar (dosen) Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri yang telah memberikan pelajaran yang berarti bagi penulis selama menempuh studi.
7. Seluruh staf dan karyawan Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri yang telah melayani penulis dengan baik selama kuliah.
8. Teman-teman kuliah Angkatan V Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri .

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, 03 Mei 2012

Ester Arisawati

Penulis

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan di bawah ini, saya :

Nama	:	Ester Arisawati
NIM	:	14000335
Program Studi	:	Magister Ilmu Komputer
Jenjang	:	Strata Dua (S2)
Konsentrasi	:	*Management Information System*
Jenis Karya	:	Tesis

Demi pengembangan ilmu pengetahuan, dengan ini menyetujui untuk memberikan ijin kepada pihak Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Inbentukika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri) **Hak Bebas Royalti Non-Eksklusif** ***(Non-exclusive Royalti-Free Right)*** atas karya ilmiah kami yang berjudul : "Komparasi antara *Support Vector Machine* dan *k-Nearest Neighbor* dalam Penentuan Pemberian Kredit Terhadap Konsumen" 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, 03 Mei 2012

Materai Rp 6.000

Ester Arisawati

Penulis

ABSTRAK

Nama	:	Ester Arisawati
NIM	:	14000335
Program Studi	:	Magister Ilmu Komputer
Jenjang	:	Strata Dua (S2)
Konsentrasi	:	*Management Information System*
Judul	:	"Komparasi antara *Support Vector Machine* dan *k-Nearest Neighbor* dalam Penentuan Pemberian Kredit Terhadap Konsumen"

Pembiayaan konsumen *(Consumer finance)* adalah kegiatan pembiayaan untuk pengadaan barang berdasarkan kebutuhan konsumen dengan pembayaran secara angsuran. Sedangkan Perusahaan Pembiayaan adalah badan usaha yang khusus didirikan untuk melakukan sewa guna usaha, anjak piutang, pembiayaan konsumen, dan atau usaha kartu kredit. Perusahaan pembiayaan akan menyetujui kredit yang diajukan konsumen setelah melakukan analisa kredit terhadap kelayakan pemberian pembiayaan konsumen, apakah disetujui dan tidak disetujui.

Dalam proses analisa terhadap konsumen, terdapat beberapa yang tidak akurat, oleh karena itu konsumen tidak mampu membayar dengan tepat waktu yang mengakibatkan kredit macet. Untuk mengatasi permasalahan yang ada diperlukan suatu model yang mampu mengklasifikasikan dan memprediksi data konsumen yang bermasalah dan tidak bermasalah.

Dalam penelitian ini dilakukan komparasi dua algoritma klasifikasi yaitu *Support Vector Machine* dan *k-Nearest Neighbor* yang diaplikasikan terhadap data konsumen yang mendapat pembiayaan kredit baik yang konsumen yang bermasalah maupun tidak. Dari hasil pengujian dengan mengukur kinerja ketiga algoritma tersebut menggunakan metode pengujian *Cross Validation, Confusion Matrix* dan Kurva ROC, diketahui bahwa algoritma *k-Nearest Neighbor* memiliki nilai *accuracy* dan AUC paling tinggi dan yang paling rendah adalah metode *Support Vector Machine*.

Kata kunci : Analisa kredit, *Support Vector Machine*, *k-Nearest Neighbor*

ABSTRACT

Name	:	Ester Arisawati
NIM	:	14000335
Study of Program	:	Magister Ilmu Komputer
Levels	:	Strata Dua (S2)
Concentration	:	*Management Information System*
Title	:	"*Support Vector Machin*e Untuk Penentuan Kelayakan Pemberian Pembiayaan Konsumen"

Consumer finance is a financing activities for the procurement of goods based on the needs of consumers with payment in installments. While the Finance Company is a business entity specifically established to conduct leasing, factoring, consumer finance, and credit or business cards. Finance companies will approve the proposed consumer credit after a credit analysis on the feasibility of providing consumer financing, if approved and not approved.

In the analysis of the consumer, there are some that are not accurate, therefore consumers can not afford to pay in a timely manner which resulted in bad debts. To overcome the existing problems we need a model that can classify and predict consumer data is problematic and not problematic.

In this study a comparison of two algorithms, Support Vector Machine and k-Nearest Neighbor classification is applied to the data of consumers who receive both a consumer credit financing is problematic or not. From the test results to measure the performance of the three algorithms using the test method Cross Validation, Confusion Matrix and ROC curves, it is known that the k-Nearest Neighbor algorithm has the accuracy and AUC value of the highest and the lowest is the method of Support Vector Machine.

Key words: Credit Analysis, Support Vector Machine, k-Nearest Neighbor

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1. Latar Belakang Penulisan

Kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam-meminjam antara bank dengan pihak lain yang mewajibkan pihak peminjam untuk melunasi utangnya setelah jangka waktu tertentu dengan pemberian bunga (UU Perbankan No.10 Tahun 1998).

Pembiayaan konsumen adalah kegiatan pembiayaan untuk pengadaan barang berdasarkan kebutuhan konsumen dengan pembayaran secara angsuran. Sedangkan perusahaan pembiayaan adalah badan usaha yang khusus didirikan untuk melakukan sewa guna usaha, anjak piutang, pembiayaan konsumen, dan atau usaha kartu kredit (Peraturan Presiden Lembaga Pembiayaan No.9 Tahun 2009).

Penilaian kredit telah menarik banyak peneliti di keuangan dan industri perbankan. Studi terbaru menunjukkan bahwa Artificial Intelligence (AI) metode yang kompetitif untuk metode statistik untuk penilaian kredit. Kredit manajemen risiko telah memainkan peran kunci dalam industri keuangan dan perbankan (Li, Liu, Xu, & Shi, 2003).

Evaluasi resiko kredit adalah bagian penting dalam resiko keuangan manajemen, khususnya untuk lembaga pembiayaan dan kemampuan untuk membedakan pelanggan baik dari yang buruk juga sangat penting (Lai, Yu, Zhou, & Wang , 2006). Penilaian kredit melibatkan diskriminasi antar pembayar yang baik dan pembayar yang buruk (Ajith, Crina, & Vitorino, 2006).

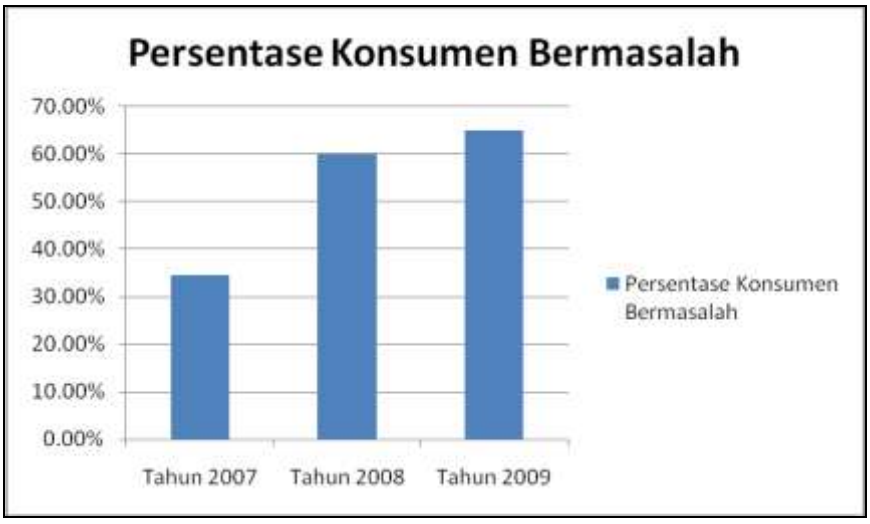
Penting bagi bank dan lembaga pembiayaan untuk mengevaluasi resiko kredit dilakukan dimuka bagi konsumen. Sebuah model yang baik bagi penilaian kredit akan membantu bank dan lembaga pembiayaan membuat keputusan yang tepat dalam rangka untuk menghindari potensi besarnya resiko (Zhang, Hifi, Chen, & Ye, 2008). Analisa resiko kredit merupakan topik penting dalam manajemen resiko keuangan (Kotsiantis, Kanellopoulus, Karioti, & Tampakas, 2009).

Penilaian kredit sebagai teknik penilaian yang sangat instrumen penting dalam industri keuangan dan perbankan (Wang, Lai, & Niu, 2011). Penilaian kredit telah menjadi isu yang sangat penting karena pertumbuhan terbaru dari industri kredit, sehingga kredit departemen bank menghadapi sejumlah besar data kredit konsumen untuk proses, tetapi tidak mungkin menganalisis ini sejumlah besar data baik dalam hal ekonomi dan tenaga kerja. Dalam studi ini kami terakhir karya-karya yang telah diterapkan metode data mining dalam masalah risiko kredit evaluasi (Keramati, & Yousefi, 2011).

Beberapa penelitian yang serupa untuk analisa penilaian kredit dengan membangun sebuah model model yang didasarkan pada klasifikasi pohon keputusan belajar data historis oleh beberapa peneliti seperti Zhang, Leung dan Ye (2008). Jiang (2008) mengusulkan model analisa penilaian kredit baru yang didasarkan pada *decision tree and simulated annealing algorithm*. Dong, Lai dan Zhou (2009) mengembangkan pengklasifikasian yang akurat untuk masalah penilaian kredit *dengan simulated annealing based rule extraction alogorithm* (SAREA). Juga Dima & Vasilache (2009) *the probit regression and the neural network model* disajikan sebagai alat pendukung keputusan untuk bank yang ingin menyaring perusahaan dalam mengajukan permohonan kredit.

Tabel 1.1. Laporan Data Konsumen Bermasalah dari Tahun 2007 s.d 2009 (sumber: laporan AEON cabang Tangerang)

Tahun	Jumlah Konsumen	Konsumen Buruk	Konsumen Baik	Persentase Kredit Macet
2007	1216	419	797	34.46%
2008	976	585	391	59.94%
2009	477	310	167	64.99%



Gambar 1.1. Grafik peningkatan persentase konsumen bermasalah dari Tahun 2007 s.d 2009

Untuk mengatasi permasalahan yang ada diperlukan suatu model yang mampu mengklasifikasikan dan memprediksi data konsumen yang bermasalah dan tidak bermasalah, maka penulis melalukan komparasi dengan menggunakan dua model klasifikasi *Support Vector Machine* dan *k-Nearest Neighbor*. Untuk mengetahui model mana yang tepat digunakan untuk memprediksi data konsumen yang bermasalah dan tidak bermasalah dalam pembayaran kredit.

Menurut hasil penelitian dengan mengkomparatif 4 model yaitu *Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Artifical Neural Networks (ANN).* Dimana dalam simulasi komputer menunjukkan bahwa efektifitas klasifikasi *Support Vector Machine (SVM)* adalah algoritma yang terbaik pada umumnya. *Support Vector Machine (SVM)* menunjukkan lebih tinggi ketahanan dan kemampuan generalisasi dibandingkan dengan algoritma yang lain (Yu, Huang, Hu, & Cai, 2010).

Dan penelitian lain menganalisa dari dua metode klasifikasi yaitu *Naïve Bayes* dan *K-Nearest Neigbor* pada data data untuk persetujuan kartu kredit, dimana hasilnya *K-Nearest Neigbor* lebih tinggi tingkat akurasinya bila dibandingkan dengan metode klasifikasi *Naïve Bayes* (Islam, Wu, Ahmadi & Ahmed, 2007).

1.2. Identifikasi Masalah

Melihat dari latar belakang masalah diatas maka dapat mengindentifikasi masalah yaitu penentuan kelayakan pemberian pembiayaan kredit terhadap konsumen oleh perusahaan pembiayaan tidak akurat sehingga dapat menyebabkan masalah kredit macet dimana konsumen yang disetujui bermasalah dalam pembayaran angsurannya.

Sedangkan pertanyaan penelitian *(research question)* yang diangkat pada penelitian ini adalah bagaimana akurasi metode klasifikasi *data mining* antara *Support Vector Machine* dan *k-Nearest Neighbor*, metode mana yang paling akurat dalam penentuan kelayakan pemberian pembiayaan kredit terhadap konsumen.

1.3. Tujuan Penelitian

Tujuan Penelitian ini adalah melakukan analisis dan komparasi dua metode klasifikasi *data mining* antara *Support Vector Machine* dan *k-Nearest Neighbor*, metode mana yang paling akurat dalam penentuan kelayakan pemberian pembiayaan terhadap konsumen sehingga meminimalkan resiko kredit.

1.4. Ruang Lingkup Penelitian

Ruang lingkup penelitian, penulis hanya membatasi permasalahan pada analisa kredit yang dilakukan oleh lembaga pembiayaan terhadap konsumen. Dimana melakukan komparasi dengan menggunakan dua metode klasifikasi antara *Support Vector Machine (SVM)* dan *K-Nearest Neighbor* dengan cara menganalisis sejumlah atribut yang menjadi parameter dalam penentuan kelayakan konsumen untuk mengajukan pembiayaan kredit kemudian mengevaluasi hasil perbandingan untuk mengetahui metode klasifikasi *data mining* mana yang paling akurat.

1.5. Hipotesis

Dalam analisa kredit untuk penentuan kelayakan pemberian pembiayaan terhadap konsumen diduga *Support Vector Machine (SVM)* dan *k-Nearest*

Neighbor sama-sama mempunyai kemampuan paling efektif dalam penilaian klasifikasi kredit. Algoritma klasifikasi *Support Vector Machine (SVM)* dan *k-Nearest Neighbor* (k-NN) diduga dapat meningkatkan akurasi terutama dalam memprediksi data kredit dan mengklasifikasikan kredit yang bermasalah atau tidak bermasalah.

1.6. Sistematika Penulisan

Sistem penulisan ini terdiri dari :

Bab I : Pendahuluan

Bab ini membahas tentang latar belakang penulisan, permasalahan mengenai penentuan kelayakan pemberian pembiayaan , kemudian pemecahan masalah dan tujuan penelitian ini.

Bab II : Landasan Teori

Bab ini membahas tentang teori yang melandasi penelitian yaitu lembaga pembiayaan, kredit, analsis kredit serta metode klasifikasi *data mining*. Studi kasus dan penyelesaian juga disajikan untuk memberi contoh.

Bab III : Metode Penelitian

Bab ini membahas tentang metode pengumpulan data dan metode penelitian serta eksperimen. Eksperimen dilakukan dengan metode *Support Vector Machine (SVM)* dan *K-Nearest Neighbor* untuk meningkatkan akurasi terutama dalam memprediksi data kredit dan mengklasifikasikan kredit yang bermasalah atau tidak bermasalah.

Bab IV : Hasil dan Pembahasan

Bab ini membahas tentang pengujian model yang dihasilkan dari bab sebelumnya. Pengujian dilakukan dengan mengukur kinerja tiap metode menggunakan beberapa pengujian kemudian hasil pengukurannya dikomparasi untuk melihat akurasi dari kedua metode tersebut.

Bab V : Penutup

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

BAB II
LANDASAN DAN KERANGKA PEMIKIRAN

2.1. Tinjauan Pustaka

Dalam penulisan tesis ini, penulis melakukan tinjauan pustaka dengan menggunakan buku dan jurnal yang berkaitan dengan tema yang dipilih.

2.1.1. Kredit

Kredit berasal dari bahasa Yunani,*"credere"*, yang berarti, "kepercayaan" atau dalam bahasa Latin, *"creditum"* yang berarti "kepercayaan akan kebenaran". Menurut undang-undang nomor 10 tahun 1998 tentang perubahan atas undang-undang nomor 7 tahun 1992 tentang perbankan pasal 1 angka 11, pengertian kredit adalah penyedian yang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam-meminjam antara bank dengan pihak lain yang mewajibkan pihak peminjam untuk melunasi utangnya setelah jangka waktu tertentu dengan pemberian bunga.

Istilah kredit adalah penyerahan barang, jasa, atau uang dari satu pihak (kreditor atau pemberi pinjaman) atas dasar kepercayaan kepada pihak lain (nasabah atau pengutang/ *borrower*) dengan janji membayar dari penerima kredit kepada pemberi kredit pada tanggal yang telah disepakati kedua belah pihak (Rivai, 2006). Kesimpulan yang dapat di ambil, bahwa kredit adalah penyerahan nilai ekonomi sekarang atas kepercayaan akan kebenaran dengan harapan mendapatkan kembali suatu nilai ekonomi yang sama atau lebih di kemudian hari.

2.1.2. Analisis Kredit

Analisis kredit dilakukan supaya kredit yang diberikan dapat mencapai tujuan, harapan dan aman. Analisis kredit adalaha kajian yang dilakukan untuk mengetahui kelayakan dari suatu permasalahan kredit. Melalui hasil analisis kredit,dapat diketahui apakah usaha nasabah layak (*feasible*) dan hasil usaha dapat dipasarkan (*marketable*), dan menguntungkan (*profitable*), serta dapat dilunasi tepat waktu (Rivai, 2006).

Analisis kredit dilakukan oleh *account officer* yang dari sisi level jabatannya merupakan level seksi atau bagain, atau bahkan dapat berupa tim (*committee*) yang ditugaskan untuk melakukan analisis permohonan kredit (Rivai,2006). Dimana prinsip dasar dalam menganilisis kredit yang lazim dikenal dengan "Prinsip 6 C's", yaitu (Rivai, 2006):

1. *Character*

   Character adalah keadaan watak atau sifar dari nasabah, baik dalam kehidupan pribadi maupun dalam lingkungan usaha. Kegunaan dari penilaian terhadap karakter ini adalah untuk mengetahui sampai sejauh mana itikad atay kemauan nasabah untuk memenuhi kewajibannya (*willingness to pay*) sesuai dengan perjanjian yang telah ditetapkan.

2. *Capital*

   Capital adalah jumlah dana atau modal sendiri yang dimiliki oleh calon nasabah. Semakin besar modal sendiri dalam perusahaan, tentu semakin tinggi kesungguhan calon nasabah dalam menjalankan usahanya dan lembaga pemberi kredit akan merasa lebih yakin dalam memberikan kredit.

2. *Capacity*

   Capacity adalah kemampuan yang dimiliki calon nasabah dalam menjalankan usahanya guna memperoleh laba yang diharapkan. Kegunaan dari penilaian ini adalah untuk mengetahui atau mengukur sampai sejauh mana calon nasabah mampu untuk mengembalikan atau melunasi utang-utang (*ability to pay*) secara tepat dari usahanya yang diperolehnya.

3. *Collateral*

   Collateral adalah barang-barang yang diserahkan nasabah sebagai agunan terhadap kredit yang diterimanya. *Collateral* tersebut harus dinilai oleh bank untuk mengetahui sejauh mana dari usaha yang diperolehnya.

4. *Condition of Economic*

   Condition of Economic yaitu situasi dan kondisi politik, sosial, ekonomi, budaya yang memengaruhi keadaan perekonomian pada suatu saat yang kemungkinannya memengaruhi kelancaran perusahaan calon kreditur.

5. *Constrain*

   Constrain adalah batasan dan hambatan yang tidak memungkinkan suatu bisnis untuk dilaksanakan pada tempat tertentu, misalkan pendirian suatu usaha pompa bensin yang disekitarnya banyak bengkel las atau pembakaran batu baru.

2.1.3. *Data Mining*

Menurut Daryl Pregibons *Data mining* adalah perpaduan dari ilmu statistik, kecerdasan buatan, dan penelitian bidang *database* (Gorunescu, 2011). *Data mining* adalah bagian dari pengenalan pola, pengenalan pola adalah suatu disiplin ilmu yang mempelajari bagaimana kita mengelompokan obyek ke berbagai kelas dan bagaimana dari data bisa kita temukan kecenderungannya. *Data mining* memegang peran penting dalam bidang industri, keuangan, cuaca, ilmu dan teknologi (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).

Adapun definisi yang lain yaitu (Larose, 2005):

a. *Data mining* merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara yang berbeda dengan sebelumnya, yang dapat dipahami dan bermanfaat bagi pemilik data.
b. *Data mining* merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, *database,* dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar.

Kemajuan luar biasa yang terus berlanjut dalam bidang *data mining* didorong oleh beberapa faktor, antara lain (Larose, 2005):

1. Pertumbuhan yang cepat dalam kumpulan data.
2. Penyimpanan data dalam *data warehouse,* sehingga seluruh perusahaan memiliki akses ke dalam *database* yang andal.

3. Adanya peningkatan akses data melalui navigasi web dan intranet.
4. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
5. Perkembangan teknologi perangkat lunak untuk *data mining* (ketersediaan teknologi).
6. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

Dari definisi-definisi yang telah disampaikan, hal penting yang terkait dengan *data mining* adalah (Kusrini, 2009):

1. *Data mining* merupakan suatu proses otomatis terhadap data yang sudah ada.
2. Data yang akan diproses berupa data yang sangat besar.
3. Tujuan *data mining* adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

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*

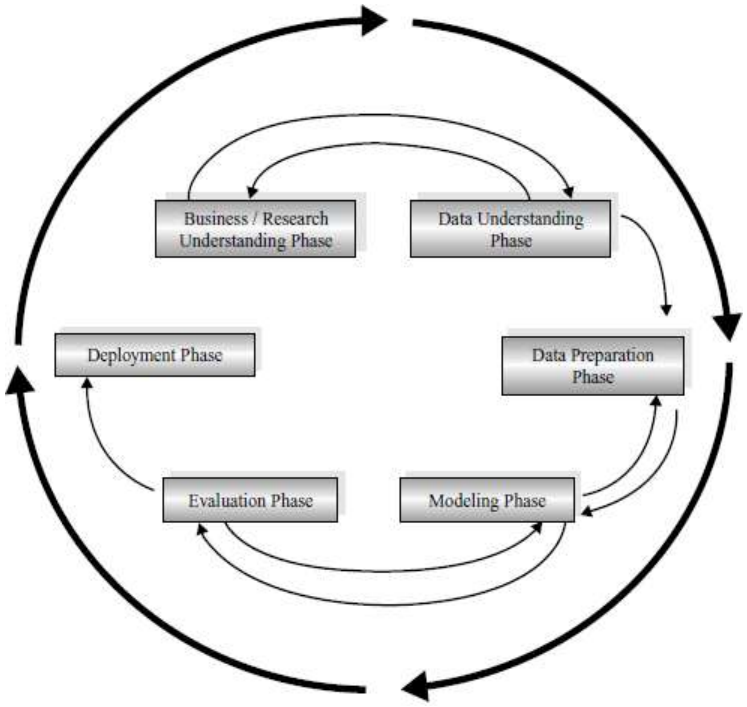
   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

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.

Cross-Industry Standard Process for Data Mining (CRISP-DM) yang dikembangkan tahun 1996 oleh analis dari beberapa industri seperti DaimlerChrysler, SPSS dan NCR. CRISP DM menyediakan standar proses *data mining* sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian.

Dalam CRISP-DM, Sebuah proyek *data mining* memiliki siklus hidup yang terbagi dalam enam fase (Gambar 2.1). Keseluruhan fase berurutan yang ada tersebut bersifat adaptif. Fase berikutnya dalam urutan bergantung kepada keluaran dari fase sebelumnya. Hubungan penting antarfase digambarkan dengan panah. Sebagai contoh, jika proses berada pada *fase modeling*. Berdasar pada prilaku dan karakteristik model, proses mungkin harus kembali kepada fase *data preparation* untuk perbaikan lebih lanjut terhadap data atau berpindah maju kepada fase *evaluation.*



Gambar 2.1. Proses *Data Mining* menurut CRISP-DM

Sumber: (Larose, 2005)

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 dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya.

   Contoh lain 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 nilai dari hasil akan di masa mendatang. Contoh prediksi dalam bisnis dan penelitian adalah:

   a. Prediksi harga beras dalam tiga bulan yang akan datang.
   b. Prediksi persentase kenaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikan.

Beberapa metode dan teknik yang di gunakan 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. Contoh lain klasifikasi dalam bisnis dan penelitian adalah :

   a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau bukan.
   b. Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.
   c. Mendiagnosis penyakit seorang pasien untuk mendapatkan termasuk kategori penyakit apa.

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. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target, Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), yang mana kemiripan *record* dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal. Contoh pengklusteran dalam bisnis dan penelitian adalah:

   a. Mendapatkan kelompok-kelompok konsumen untuk target pemasaran dari suatu produk bagi perusahaan yang tidak memiliki dana pemasaran yang besar.
   b. Untuk tujuan audit akuntansi, yaitu melakukan pemisahan terhadap perilaku financial dalam baik dan mencurigakan.

c. Melakukan pengklusteran terhadap ekspresi dari gen, untuk mendapatkan kemiripan perilaku dari gen dalam jumlah besar.

6. Asosiasi

Tugas asosiasi dalam *data mining* adalah menemukan yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja. Contoh asosiasi dalam bisnis dan penelitian adalah:

a. Meneliti jumlah pelanggan dari perusahaan telekomunikasi seluler yang diharapkan untuk memberikan respons positif terhadap penawaran *upgrade* layanan yang diberikan.
b. Menemukan barang dalam supermarket yang dibeli secara bersamaan dan barang yang tidak pernah dibeli secara bersamaan.

2.1.4. Klasifikasi

Klasifikasi adalah proses penemuan model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Han, 2006).

Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah *learning* (fase *training*), dimana algoritma klasifikasi dibuat untuk menganalisa data *training* lalu direpresentasikan dalam bentuk *rule* klasifikasi. Proses kedua adalah klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari *rule* klasifikasi (Han, 2006).

Klasifikasi adalah proses menempatkan objek tertentu atau konsep dalam satu set kategori, berdasarkan masing-masing sifat dari objek atau konsep tersebut (Gorunescu, 2011). Proses klasifikasi didasarkan pada empat komponen mendasar yaitu (Gorunescu, 2011):

a. *Class*

Variabel dependen yang berupa kategorikal yang mewakili sebuah ‘label’ yang diletakkan pada objek setelah klasifikasinya. Contoh dari *class*: resiko penyakit jantung, resiko kredit, *customer loyalty*, jenis gempa.

b. *Predictors*

Variabel independen yang diwakili oleh karakteristik (atribut) dari data dan harus diklasifikasikan berdasarkan klasifikasi yang dibuat. Contoh dari

predictors: merokok, minum alkohol, tekanan darah, status perkawinan, tabungan, aset, gaji.

c. *Training dataset*

Yang merupakan satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk pelatikan dalam menentukan kelas yang cocok berdasarkan *predictors.*

d. *Testing dataset*

Yang berisi data baru yang akan diklasifikasikan oleh model yang telah dibuat dan sehingga akurasi klasifikasi dapat dievaluasi.

Beberapa model algoritma klasifikasi yang populer digunakan secara luas, adalah (Gorunescu, 2011):

1. *Decision/classification trees*
2. *Bayesian classifiers/Naive Bayes classifiers*
3. *Neural networks*
4. *Statistical analysis*
5. *Genetic algorithms*
6. *Rough sets*
7. *k-nearest neighbor classifier*
8. *Rule-based methods*
9. *Memory based reasoning*
10. *Support vector machines*

2.1.5. *Support Vector Machine* (SVM)

2.1.5.1. Konsep *Support Vector Machine* (SVM)

Support Vector Machine dibuat oleh Vapnik, SVM adalah seperangkat model yang terkait untuk belajar mengawasi, berlaku untuk klasifikasi yang baik dan masalah regresi. Sebuah pengklasifikasian SVM menciptakan *hyperplane* maksimum margin yang terletak pada ruang input diubah dan membagi kelas, misalnya sekaligus memaksimalkan jarak ke contoh bersih terdekat (Maimon, 2010).

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).

Support Vector Machine (SVM) adalah suatu teknik yang relative baru (1995) untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi, yang sangat popular belakangan ini. SVM berada dalam satu kelas dengan ANN dalam hal fungsi dan kondisi permasalahan yang bias diselesaikan. Keduanya masuk dalam kelas *supervised learning*. Baik para ilmuwan maupun praktisi telah banyak menerapkan model ini dalam menyelesaikan masalah-masalah nyata seperti *gene expression analysis*, *financial*, cuaca hingga di bidang kedokteran. Terbukti dalam banyak implementasi, SVM member hasil yang lebih baik dari ANN, terutama dalam hal solusi yang dicapai. ANN menemukan solusi berupa *local optimal* sedangkan SVM menemukan solusi yang *global optimal*. Tidak heran bila kita menjalankan ANN solusi dari setiap training selalu berbeda. Hal ini disebabkan solusi *local optimal* yang dicapai tidak selalu sama. SVM selalu mencapai solusi yang sama untuk setiap *running*. Dalam model ini, kita berusaha untuk menemukan *fungsi* pemisah (klasifer) yang *optimal* yang bisa memisahkan dua set data dari dua kelas yang berbeda (Santosa, 2007).

Support Vector Machine (SVM) adalah sistem pembelajaran yang pengklasifikasiannya menggunakan ruang hipotesis berupa fungsi-fungsi *linear* dalam sebuah ruang fitur (*feature space*) berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan *learning bias* yang berasal dari teori pembelajaran statistik.

Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai rangkaian aqzharmonis konsep-konsep unggulan dalam bidang *pattern recognition*. Sebagai salah satu metode *pattern recognition*, usia SVM terbilang masih relatif muda. Walaupun demikian, evaluasi kemampuannya dalam berbagai aplikasinya menempatkannya sebagai *state of the art* dalam *pattern recognition*, dan dewasa ini merupakan salah satu tema yang berkembang

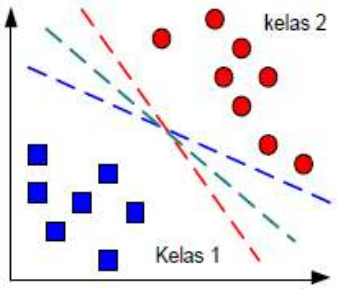
dengan pesat. SVM adalah metode *learning machine* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space* (Nugroho, Witarto, Handoko, 2003).

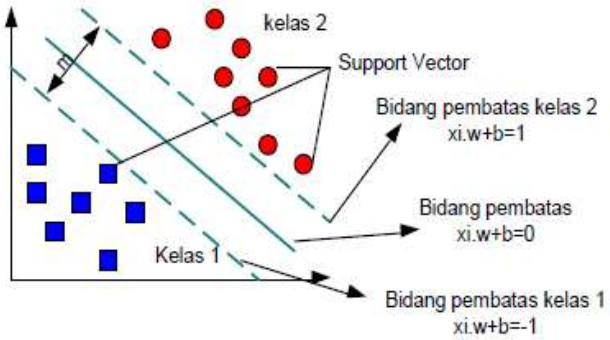
Karakteristik SVM sebagaimana telah dijelaskan pada bagian sebelumnya, dirangkumkan sebagai berikut (Nugroho, Witarto, Handoko, 2003):

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

2.1.5. 2. Studi Kasus *Support Vector Machine* (SVM)

Secara sederhana konsep SVM adalah sebagai usaha mencari *hyperlane* terbaik yang berfungsi sebagai pemisah dua buah *class* pada *input space*, dimana dapat dilihat pada gambar dibawah ini:





Gambar 2.2a. Bidang Pemisah **Gambar 2.2b. Bidang Pemisah Terbaik**

Pada gambar diatas memperlihatkan beberapa *pattern* yang merupakan anggota dari dua buah *class*: +1 dan -1. *Pattern* yang tergabung pada *class* -1 disimbolkan dengan warna biru (kotak), sedangkan *pattern* pada *class* +1, disimbolkan dengan warna merah (lingkaran). Problem klasifikasi dapat

diterjemahkan dengan usaha menemukan garis (*hyperplane*) yang memisahkan antara kedua kelompok tersebut. Berbagai alternatif garis pemisah (*discrimination boundaries*) ditunjukkan pada gambar 1. *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*. Garis solid pada gambar 2.2b. sebelah kanan menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua *class*, sedangkan titik biru dan merah yang berada dalam kotak dan lingkaran hitam adalah *support vector*. Usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada SVM.

Secara matematis, konsep dasar SVM adalah:

$$\min \frac{1}{2} \| w \|^2$$

$$y_i (wx_i + b) \geq 1, i = 1,...,,\lambda$$

Dimana x_i adalah data input, y_i adalah keluaran dari data x_i . w,b adalah parameter-parameter yang kita cari nilainya. Dalam formulasi diatas kita ingin meminimalisir fungsi tujuan (*object function*) $\frac{1}{2} \| w \|^2$ atau memaksimalkan kuantitas $\| w\|^2$ dengan memperhatikan pembatas y_i (wx_i + b) $\geq$ 1. Bila output data y_i = +1, maka pembatas menjadi (wx $_i$ b) $\geq$ 1 . Sebaliknya jika y_i = - 1 maka pembatas menjadi (wx_i + b) $\leq$ -1.

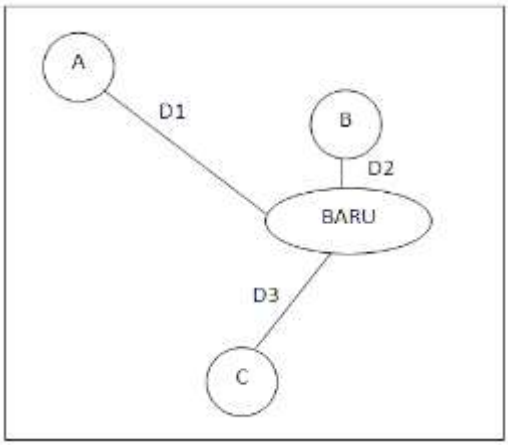
2.1.6. *k-Nearest Neighbor* (k-NN)

2.1.6.1. Konsep *k-Nearest Neighbor* (k-NN)

k-Nearest Neighbor (k-NN) adalah merupakan suatu metode yang paling sering digunakan untuk klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. k-NN merupakan salah satu metode pengklasifikasian data berdasarkan similaritas dengan label data, dimana tumpukan kumpulan data disimpan, sehingga klasifikasi bagi sebuah rekor aru dapat ditemukan hanya dengan membandingkan dengan catatan yang paling dekat atau paling mirip pada data *training set* (Larose, 2005).

k-Nearest Neighbor (k-NN) adalah teknik yang termasuk dalam kelompok klasifikasi nonparametic. Dimana tidak ada distribusi dari data yang ingin dikelompokkan. Algoritma ini sangat sederhana dan mudah diimplementasikan, mirip dengan teknik klastering, dengan mengelompokkan suatu data baru berdasarkan jarak data baru itu ke beberapa data atau tetangga (*neighbor*) terdekat. Dalam hal ini jumlah data atau tetangga terdekat ditentukan oleh user yang dinyatakan dengan *k* (Santosa, 2007).

Algoritma ini juga merupakan salah satu teknik *lazy learning*. Dengan mencari kelompok *k*, objek dalam data *training* yang paling dekat (mirip) dengan objek data baru (data *testing*) (Wu, 2009). Contoh kasus, misal diinginkan untuk mencari solusi terhadap masalah seorang pasien baru dengan menggunakan solusi dari pasien lama. Untuk mencari solusi dari pasien baru tersebut digunakan kedekatan dengan kasus pasien lama, solusi dari kasus lama yang memiliki kedekatan dengan kasus baru digunakan sebagai solusinya.



Gambar 2.3. Ilustrasi kasus algoritma k-NN

Ilustrasi pada gambar 2.3 diatas ada pasien baru dan 3 pasien lama (A, B, dan C). Ketika ada pasien baru maka yang diambil solusi adalah solusi dari kasus pasien lama yang memikili kedekatan terbesar. Misal D1 adalah jarak antara pasien baru dengan pasien A, D2 adalah jarak antara pasien baru dengan pasien B, D3 adalah jarak antara pasien baru dengan pasien C. Dari ilustrasi gambar terlihat bahwa D2 yang paling terdekat dengan kasus baru. Dengan demikian maka solusi dari kasus pasien B yang akan digunakan sebagai solusi dari pasien baru tersebut.

Dasar klasifikasi algoritma k-Nearest Neighbour adalah:

1. Temukan contoh pelatihan k yang paling dekat dengan contoh tak terlihat.
2. Ambil klasifikasi paling sering terjadi untuk contoh ini k.

Ada banyak cara untuk mengukur jarak kedekatan antara data baru dengan data lama (data *training*), diantaranya e*uclidean distance* dan *manhattan distance* (*city block distance*), yang paling sering digunakan adalah *euclidean distance* (Bramer,2007), yaitu :

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

Dimana a = a_1,a_2, …, a_n, dan b = b_1, b_2, …, b_n mewakili n nilai atribut dari dua *record*. Untuk atribut dengan nilai kategori, pengukuran dengan *euclidean distance* tidak cocok. Sebagai penggantinya, digunakan fungsi sebagai berikut (Larose, 2005):

different(a_i,b_i) { 0 jika $a_i = b_i$

= 1 selainnya

Dimana a_i dan b_i adalah nilai kategori. Jika nilai atribut antara dua *record* yang dibandingkan sama maka nilai jaraknya 0, artinya mirip, sebaliknya, jika berbeda maka nilai kedekatannya 1, artinya tidak mirip sama sekali. Misalkan atribut warna dengan nilai merah dan merah, maka nilai kedekatannya 0, jika merah dan biru maka nilai kedekatannya 1.

Untuk mengukur jarak dari atribut yang mempunyai nilai besar, seperti atribut pendapatan, maka dilakukan normalisasi. Normalisasi bisa dilakukan dengan *min-max normalization* atau *Z-score standardization* (Larose, 2005). Jika data *training* terdiri dari atribut campuran antara numerik dan kategori, lebih baik gunakan *min-max normalization* (Larose, 2005).

2.1.6.2. Studi Kasus *k-Nearest Neighbor* (k-NN)

Adapun rumus untuk menghitung kedekatan antara dua kasus adalah sebagai berikut (Kusrini, 2009):

$$Similarity(T, S) = \frac{\sum_{i=1}^{n} f(T_i, S_i) * w_i}{w_i}$$

Keterangan:

T : kasus baru

S : kasus yang ada dalam penyimpanan

n : jumlah atribut dalam setiap kasus

i : atribut individu antara 1 s/d n

f : fungsi *similarity* atribut i antara kasus T dan kasus S

w : bobot yang diberikan pada atribut ke-i

Kedekatan kasus biasanya berada nilai antara 0 dan 1. Nilai 0 artinya kedua kasus mutlak tidak mirip dan sebaliknya, jika nilai 1 artinya kedua kasus mutlak mirip.

Untuk algoritma *Nearest Neighbor* banyak kasus yang dapat selesaikan dan salah satunya adalah kasus tentang kemungkinan seorang nasabah bank akan bermasalah dalam pembayaran atau tidak (Kusrini, 2009).

Tabel 2.1. Kasus *k-Neareast Neighbor*

NO	JENIS KELAMIN	PENDIDIKAN	AGAMA	BERMASALAH
1	L	S1	ISLAM	YA
2	P	SMA	KRISTEN	TIDAK
3	L	SMA	ISLAM	TIDAK

Atribut **BERMASALAH** merupakan atribut tujuan

Selain atribut tujuan harus diberikan bobot dengan nilai yang berbeda-beda

Tabel 2.2. Bobot Atribut

ATRIBUT	BOBOT
JENIS KELAMIN	0.5
PENDIDIKAN	1
AGAMA	0.75

Kedekatan antar nilai dalam satu atribut juga harus didefinisikan

Tabel 2.3. Kedekatan Nilai Atribut Jenis Kelamin

NILAI 1	NILAI 2	KEDEKATAN
L	L	1
P	P	1
L	P	0.5
P	L	0.5

Tabel 2.4. Kedekatan Nilai Atribut Pendidikan

NILAI 1	NILAI 2	KEDEKATAN
S1	S1	1
SMA	SMA	1
S1	SMA	0.4
SMA	S1	0.4

Tabel 2.5. Kedekatan Nilai Atribut Agama

NILAI 1	NILAI 2	KEDEKATAN
ISLAM	ISLAM	1
KRISTEN	KRISTEN	1
ISLAM	KRISTEN	0.75
KRISTEN	ISLAM	0.75

Misalkan ada kasus nasabah baru dengan nilai atribut sebagai berikut:

1. Jenis Kelamin : L
2. Pendidikan : SMA
3. Agama : Kristen

Untuk memprediksi apakah nasabah tersebut akan bermasalah atau tidak, maka akan dilakukan langkah-langkah sebagai berikut:

1. Menghitung kedekatan kasus baru dengan **tabel kasus no. 1**, diketahui :

a : kedekatan nilai atribut JENIS KELAMIN (L dengan L) = 1

b : bobot atribut JENIS KELAMIN = 0.5

c : kedekatan nilai atribut PENDIDIKAN (SMA dengan S1) = 1

d : bobot atribut PENDIDIKAN = 1

e : kedekatan nilai atribut AGAMA (KRISTEN dengan ISLAM) = 0.75

f : bobot atribut AGAMA = 0.75

Dihitung :

$$Jarak = \frac{(a*b)+(c*d)+(e*f)}{b+d+f}$$

$$Jarak = \frac{(1*0.5)+(0.4*1)+(0.75*0.75)}{0.5+1+0.75}$$

$$Jarak = \frac{1.4625}{2.25} = 0.65$$

2. Menghitung kedekatan kasus baru dengan **tabel kasus no. 2**, diketahui :

   a : kedekatan nilai atribut JENIS KELAMIN (L dengan P) = 0.5

   b : bobot atribut JENIS KELAMIN = 0.5

   c : kedekatan nilai atribut PENDIDIKAN (SMA dengan S1) = 1

   d : bobot atribut PENDIDIKAN = 1

   e : kedekatan nilai atribut AGAMA (KRISTEN dengan ISLAM) = 0.75

   f : bobot atribut AGAMA = 0.75

   Dihitung :

$$Jarak = \frac{(a*b)+(c*d)+(e*f)}{b+d+f}$$

$$Jarak = \frac{(0.5*0.5)+(1*1)+(0.75*0.75)}{0.5+1+0.75}$$

$$Jarak = \frac{1.8125}{2.25} = 0.8$$

3. Menghitung kedekatan kasus baru dengan **tabel kasus no. 3**, diketahui :

   a : kedekatan nilai atribut JENIS KELAMIN (L dengan L) = 1

   b : bobot atribut JENIS KELAMIN = 0.5

   c : kedekatan nilai atribut PENDIDIKAN (SMA dengan SMA) = 1

   d : bobot atribut PENDIDIKAN = 1

   e : kedekatan nilai atribut AGAMA (KRISTEN dengan ISLAM) = 0.75

   f : bobot atribut AGAMA = 0.75

   Dihitung :

$$Jarak = \frac{(a*b)+(c*d)+(e*f)}{b+d+f}$$

$$Jarak = \frac{(1*0.5)+(1*1)+(0.75*0.75)}{0.5+1+0.75}$$

$$Jarak = \frac{2.0625}{2.25} = 0.9$$

4. Memilih kasus dengan kedekatan terdekat. Dari langkah 1,2 dan 3 diketahui bahwa nilai tertinggi adalah kasus 3. Berarti kasus yang terdekat dengan kasus baru ini adalah kasus 3. sehingga prediksi kemungkinan nasabah baru tersebut adalah **Tidak Bermasalah**.

2.1.7. Rapid Miner

Rapid Miner merupakan perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari Institute of Technology Blanchardstown dan Ralf Klinkenberg dari rapid-i.com dengan tampilan GUI (*Graphical User Interface*) sehingga memudahkan pengguna dalam menggunakan perangkat lunak ini. Perangkat lunak ini bersifat *open source* dan dibuat dengan menggunakan bahasa Java di bawah lisensi GNU Public License dan Rapid Miner dapat dijalankan di sistem operasi manapun. Dengan menggunakan Rapid Miner, tidak dibutuhkan kemampuan *coding* khusus, karena semua fasilitas sudah disediakan. Rapid Miner dikhususkan untuk penggunaan *data mining*. Model yang disediakan juga cukup lengkap, seperti model Bayesian Modelling, Tree Induction,Neural Network dan lain-lain. Banyak metode yang disediakan oleh Rapid Miner mulai dari klasifikasi, klustering, asosiasi dan lain-lain. Jika tidak ada metode atau model algoritma yang tidak ada dalam Weka, pengguna boleh menambahkan modul lain, karena Weka bersifat *open source*, jadi siapapun dapat ikut mengembangkan perangkat lunak ini.

2.1.8. Evaluasi dan Validasi Klasifikasi *Data Mining*

Pada penelitian ini dalam menguji model, digunakan metode *Cross Validation*, *Confusion Matrix*, dan kurva ROC (*Receiver Operating Characteristic*).

1. *Cross Validation*

   Cross Validation adalah pengujian standar yang dilakukan untuk memprediksi *error rate*. Data *training* dibagi secara random ke dalam beberapa bagian dengan perbandingan yang sama kemudian *error rate* dihitung bagian demi bagian, selanjutnya hitung rata-rata seluruh *error rate* untuk mendapatkan *error rate* secara keseluruhan. Sekarang perhatikan apa yang harus dilakukan ketika jumlah data untk pelatihan dan pengujian sangat terbatas. Tentu saja harus menggunakan metode ketidaksepakatan cadangan dalam jumlah tertentu untuk pengujian dan menggunakan sisanya untuk pelatihan dan menetapkan bagian itu selain untuk validasi, jika diperlukan (Witten, 2011). Cara yang lebih umum untuk mengurangi bias yang disebabkan oleh sampel tertentu

dipilih untuk ketidaksepakatan adalah mengulang seluruh proses, pelatihan dan pengujian, beberapa kali dengan berbeda acak sampel. Dalam setiap iterasi proporsi tertentu, misalnya dua pertiga dari data yang dipilih secara acak untuk pelatihan, mungkin denagn stratifikasi dan sisanya digunakan untuk pengujian. Tingkat kesalahan pada ietrasi yang berbeda dirata-ratakan untuk menghasilkan tingkat kesalahan secara keseluruhan. Ini adalah metode ketidaksepakatan berulang tingkat kesalahan estimasi.

Cara standar untuk memprediksi tingkat kesalahan teknik belajar diberikan tunggal, sampel tetap data adalah dengan menggunakan bertingkat sepuluh kali lipat cross-validasi atau yang sering disebut 10-*Fold Cross Validation*. Data tersebut dibagi secara acak menjadi 10 bagian di mana kelas diwakili di sekitar sama proporsi seperti pada dataset lengkap. Setiap bagian mengulurkan pada gilirannya dan skema belajar dilatih pada sisa sembilan per sepuluh, kemudian tingkat kesalahan yang dihitung pada set ketidaksepakatan. Dengan demikian, prosedur pembelajaran dilaksanakan sebanyak 10 kali pada pelatihan yang berbeda set (setiap set memiliki banyak kesamaan dengan yang lain). Akhirnya, 10 perkiraan kesalahan dirata-ratakan untuk menghasilkan perkiraan kesalahan secara keseluruhan. Mengapa 10? Ekstensif tes pada dataset yang berbeda-beda, dengan belajar yang berbeda teknik, telah menunjukkan bahwa 10 adalah tentang jumlah hak lipatan untuk mendapatkan yang terbaik memperkirakan kesalahan, dan ada juga beberapa bukti teoritis yang mendukung hal ini. Meskipun argumen ini tidak berarti konklusif, dan perdebatan terus kemarahan dalam pembelajaran mesin dan lingkaran data mining tentang apa adalah skema terbaik untuk evaluasi, sepuluh kali lipat cross-validasi telah menjadi metode standar dalam praktis hal. Tes juga menunjukkan bahwa penggunaan stratifikasi meningkatkan hasil sedikit. Dengan demikian, teknik evaluasi standar dalam situasi di mana hanya terbatas. Data yang tersedia adalah bertingkat sepuluh kali lipat cross-validasi. Perhatikan bahwa baik stratifikasi maupun pembagian ke dalam 10 lipatan memiliki tepatnya: Cukuplah untuk membagi data ke dalam 10 set kira-kira sama di mana nilai-nilai berbagai kelas yang diwakili di sekitar proporsi yang

tepat. Sebuah sepuluh kali lipat tunggal lintas validasi mungkin tidak cukup untuk mendapatkan kesalahan yang dapat diandalkan memperkirakan. Berbeda sepuluh kali lipat cross-validasi percobaan dengan pembelajaran yang sama skema dan dataset sering menghasilkan hasil yang berbeda karena efek acak.

Gambar 2.4. Ilustrasi 10-*Fold Cross Validation*

2. *Confusion matrix*

Confusion Matrix adalah alat (*tool*) visualisasi yang biasa digunakan pada *supervised learning*. Tiap kolom pada matriks adalah contoh dalam kelas prediksi, sedangkan setiap baris mewakili kejadian di kelas yang sebenarnya. Satu keuntungan dari *Confusion Matrix* adalah mudah untuk mengetahui jika data ada diantara dua kelas (*mislabeling*). *Confusion Matrix* berisi informasi tentang aktual (*actual*) dan prediksi (*predicted*) pada sistem klasifikasi. Kinerja sistem seperti ini biasanya dievaluasi dengan menggunakan data pada matriks (Gorunescu, 2011).

Confusion matrix adalah metode yang menggunakan tabel matriks seperti pada **Tabel 2.1**, jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif (Bramer, 2007).

Tabel 2.6. Model *Confusion Matrix*

Sumber: (Bramer, 2007)

Klasifikasi yang benar	**Diklasifikasikan sebagai**	
	+	–
+	*true positive*	*false positive*
–	*false positive*	*true positive*

True positives adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positives* adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *false negatives* adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* adalah jumlah *record* negatif yang diklasifikasikan sebagai negative, kemudian masukkan data uji. Setelah data uji dimasukkan ke dalam *confusion matrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *sensitivity* (*recall*), *specificity*, *precision* dan *accuracy*. *Sensitivity* digunakan untuk membandingkan jumlah TP terhadap jumlah *record* yang positif sedangkan *specificity* adalah perbandingan jumlah TN terhadap jumlah *record* yang negatif. Untuk menghitung digunakan persamaan di bawah ini (Han, 2007):

1. $sensitivity = \frac{TP}{P}$
2. $specificity = \frac{TN}{N}$
3. $precision = \frac{TP}{TP + FP}$
4. $accuracy = sensitivity\frac{P}{(P + N)} + specificity\frac{N}{(P + N)}$

Keterangan:

TP = jumlah *true positives*

TN = jumlah *true negatives*

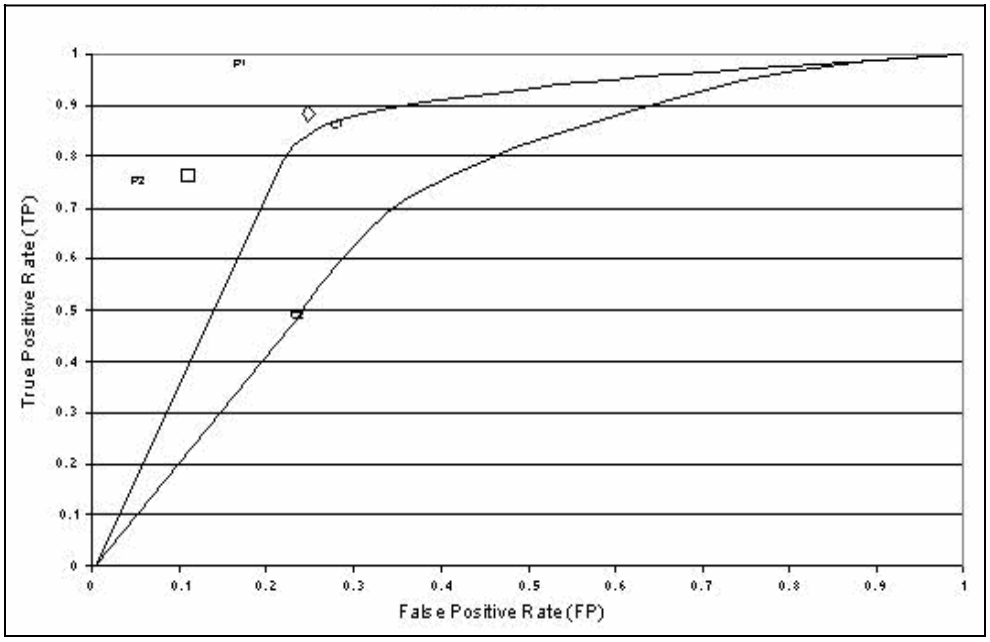
P = jumlah *record positives*

N = jumlah *tupel negatives*

FP = jumlah *false positives*

3. Kurva ROC

ROC Curve adalah cara lain untuk menguji kinerja pengklasifikasi. Sebuah grafik ROC adalah plot dengan tingkat positif salah (FP) pada sumbu *X* dan tingkat positif benar (TP) pada sumbu *Y*. Titik (0,1) adalah klasifikasi sempurna yang mengklasifikasikan semua kasus positif dan kasus negatif dengan benar, karena tingkat positif salah (FP) adalah 0 (tidak ada), dan tingkat positif benar (TP) adalah 1. Titik (0,0) merupakan sebuah klasifikasi yang memprediksi setiap kasus menjadi negatif, sedangkan titik (1,1) sesuai dengan sebuah klasifikasi yang memprediksi setiap kasus menjadi positif. Titik (1,0) adalah klasifikasi yang tidak benar untuk semua klasifikasi. Dalam banyak 33 kasus, klasifikasi memiliki parameter yang dapat disesuaikan untuk meningkatkan *TP* atau penurunan *FP*. Setiap pengaturan parameter menyediakan pasangan *FP* dan *TP* dan serangkaian pasangan tersebut dapat digunakan untuk memetakan kurva ROC. Klasifikasi non-parametrik diwakili oleh titik ROC tunggal, sesuai dengan pasangannya *(FP, TP)* (Gorunescu, 2011).



Gambar 2.5. Kurva ROC

Gambar 2.5 menunjukkan sebuah contoh dari grafik ROC dengan dua kurva ROC berlabel C1 dan C2, dan dua ROC poin berlabel P1 dan P2. Algoritma

non-parametrik menghasilkan titik ROC tunggal untuk data tertentu yang ditetapkan. Pada fitur *ROC Curve* dijelaskan bahwa:

a. *ROC Curve* adalah independen dari distribusi kelas.
b. *ROC Curve* merangkum semua informasi yang terdapat dalam C*onfusion Matrix*, karena *FN* adalah komplemen dari *TP* dan *TN* adalah komplemen dari *FP*.
c. *ROC Curve* menyediakan alat visual untuk memeriksa *tradeoff* antara kemampuan klasifikasi untuk benar mengidentifikasi kasus positif dan mengklasifikasikan jumlah kasus negatif yang salah.

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horisontal dan *true positives* sebagai garis vertikal (Vecellis, 2009).*The area under curve* (AUC) dihitung untuk mengukur perbedaan performansi metode yang digunakan. AUC dihitung menggunakan rumus: (Liao, 2007)

$$\theta^r = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\psi\left(\mathrm{x}i^r, \mathrm{x}j^r\right)$$

Dimana:

$$(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases}$$

K = jumlah algoritma klasifikasi yang dikomparasi

X = output *positif*

Y = output *negatif*

2.2. Tinjauan Studi

Beberapa penelitian terdahulu yang menggunakan model klasifikasi *Support Vector Machines* (SVM) dan *K-Nearest Neighbor*, secara garis besar adalah sebagai berikut:

1. ***Support Vector Machines Approach to Credit Assessment* (Li, Liu, Xu, & Shi, 2003).** Penelitian ini membandingkan model *Support Vector Machines* (SVM) dengan kriteria dasar bank yang digunakan pada aplikasi pemberian kartu kredit pemohon untuk mengklasifikasikan sampel menjadi dua kelas yatiu baik dan buruk. Hasil klasifikasi penilaian dengan menggunakan SVM memiliki keunggulan yang jelas bila dibandingankan dengan metode bank. Hasil percobaan awal menunjukkan bahwa model SVM ternyata menjadi alat yang efektif untuk penilaian klasifikasi kredit.
2. ***A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation* (Yu, Huang, Hu, & Cai, 2010).** Penelitian ini membandingkan empat model dalam *data mining* yaitu *Logistic Regression* (LR), *Decision Trees* (DT), *Support Vector Machine* (SVM), dan *Artificial Neural Networks* (ANN) dalam menangani masalah evaluasi resiko kredit. Evaluasi resiko kredit pada *individu* merupakan hal yang penting dalam *data mining* dan menantang masalah dalam analisis keuangan yang dominan. Dari hasil penelitian menunjukan SVM menghasilkan nilai akurasi klasfikasi terbaik dan SVM menunjukkan lebih tinggi ketahanan dan kemampuan generalisasi bila dibandingkan dengan algoritma yang lainnya.
3. ***Investigating the Performance of Naive-Bayes Classifiers and K-Nearest Neighbor Classifiers* (Islam, Wu, Ahmadi & Ahmed, 2007).** Penelitian ini menganalisa dua metode klasifikasi yaitu *Naive-Bayes* dan *K-Nearest Neighbor* yang diimplementasikan dan diterapkan pada data set aplikasi persetujuan kartu kredit. Hasil kesimpulan dari analisa menyatakan bahwa *K-Nearest Neighbor* tingkat akurasi kesalahan klasifikasi lebih kecil bila dibandingkan dengan *Naive-Bayes*, maka *K-Nearest Neighbor* lebih unggul.

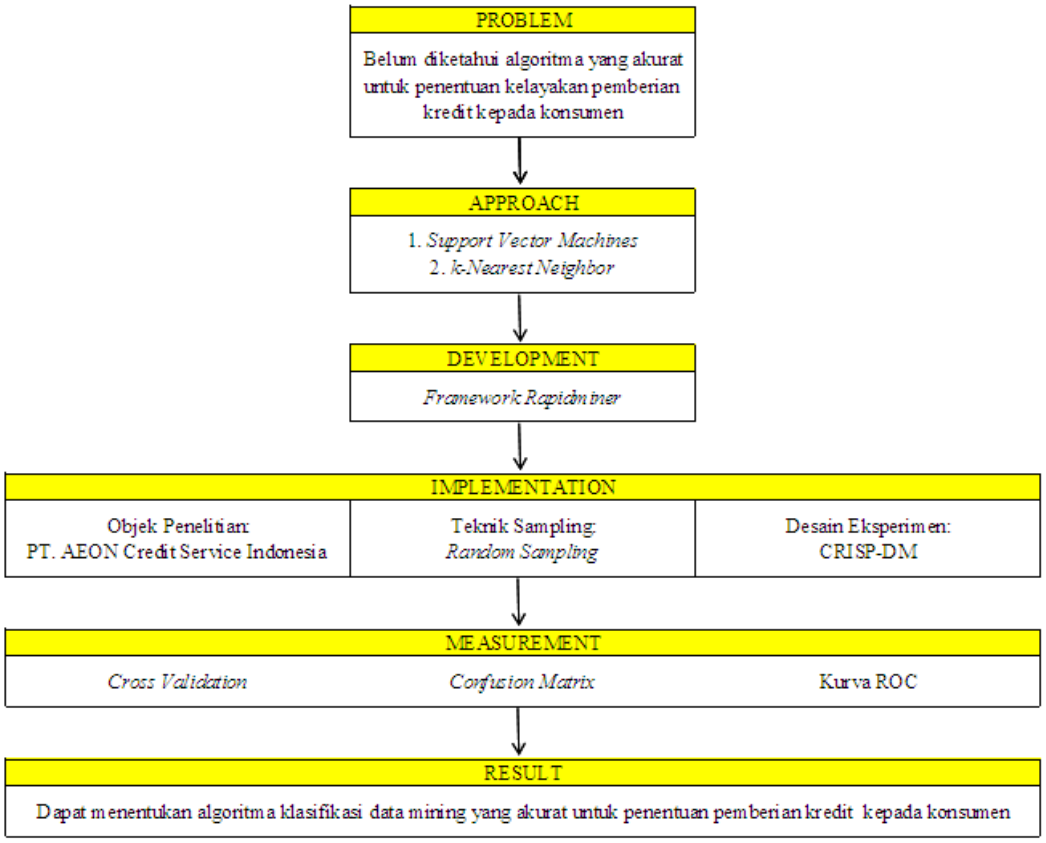
2.3. Tinjauan Organisasi

Dimana perusahaan AEON CREDIT SERVICE CO.,LTD. didirikan di Tokyo, Jepang pada tanggal 20 Juni 1981, dibawah pimpinan President dan C.E.O yaitu Mr. Kazuhide Kamitani. Terhitung per 20 Februari 2011 memiliki 28.07 juta jumlah pemegang kartu (termasuk 8.07 juta pemegang kartu diluar negeri). AEON CREDIT SERVICE CO.,LTD. Memiliki banyak anak perusahaan dibeberapa negara di Asia diantaranya Hong Kong, Thailand, Malaysia, Taiwan, China, Indonesia, Vietnam, dan Philippines.

Di Indonesia sendiri, yang menawarkan jumlah penduduk terbesar di antara negara-negara ASEAN dan dengan pertumbuhan ekonomi yang cukup tinggi, AEON CREDIT SERVICE CO.,LTD mendirikan anak perusahaan pada bulan Mei 2006 yang bernama PT. AEON Credit Service Indonesia. PT. AEON Credit Service Indonesia adalah perusahaan pembiayaan konsumen, grup AEON CREDIT SERVICE CO.,LTD. yang berasal dari Jepang dan memiliki berbagai cabang di Hong Kong, Thailand, Filipina dan Malaysia. PT AEON Credit Service Indonesia menyediakan jasa pembiayaan untuk produk elektronik, perlengkapan rumah tangga, furnitur, alat musik, komputer, telepon seluler dan banyak lainnya. PT AEON bekerja sama dengan lebih dari 1500 merchants, di antaranya Giant, Carrefour, Electronic City, Lotte Mart, Erafone, Best Denki, Yamaha Music Dealers, dan banyak lagi.

2.4. Kerangka Pemikiran

Kerangka pemikiran pada penelitian ini terdiri dari beberapa tahap seperti terlihat pada Gambar 2.4. Permasalahan (*problem*) pada penelitian ini adalah “ini adalah Belum diketahui algoritma yang akurat untuk penentuan kelayakan pemberian kredit kepada konsumen. Untuk itu dibuat *approach* (model) yaitu algoritma *Support Vector Machine* dan *k-nearest neighbor* untuk memecahkan permasalahan kemudian dilakukan pengujian terhadap kinerja dari ketiga metode tersebut. Pengujian menggunakan metode *Cross Validation*, *Confusion Matrix* dan kurva ROC. Untuk mengembangkan aplikasi (*development*) berdasarkan model yang dibuat, digunakan Rapid Miner. Desain ekperimennya digunakan CRISP-M. Dibawah ini adalah kerangka pemikiran dalam bentuk bagan:



Gambar 2.6. Kerangka Pemikiran

BAB III
METODOLOGI PENELITIAN

3.1. Metode Penelitian

Metode penelitian yang digunakan dalam penelitian ini adalah menggunakan metode penelitian eksperimen. Sedangkan data yang digunakan dalam penelitian ini adalah data primer, penulis mendapatkan secara langsung *database* konsumen dari PT AEON Credit Service Indonesia cabang Tangerang. Metode penelitian yang digunakan penulis dalam penelitian eksperimen ini dengan menggunakan metode *Cross-Industry Standard Process for Data Mining* (CRISP-DM) terdiri dari enam tahap yang merupakan sebuah proses siklis yaitu (David Olson & Yong Shi, 2008) :

a. *Business Understanding* (Pemahaman Bisnis) adalah:

   Pemahaman bisnis meliputi penetapan tujuan bisnis, penilaian situasi terkini, penetapan tujuan bisnis, penetapan tujuan penggalian data, dan pengembangan rencana proyek.

b. *Data Understanding* (Pemahaman Data) adalah:

   Begitu tujuan bisnis dan rencana proyek ditetapkan, pemahaman data mempertimbangkan data yang dibutuhkan. Langkah ini bisa meliputi pengumpulan data awal, deskripsi data, eksplorasi data, dan verfikasi kualitas data. Eksplorasi data seperti peninjauan statistik rangkuman (yang meliputi tampilan visual variabel-variabel kategorik) bisa terjadi pada akhir tahap ini. Model-model seperti analisis pengelompokan (*cluster analysis*) dapat pula diterapkan dalam tahap ini, dengan tujuan mengidentifikasi pola dalam data tersebut.

c. *Data Preparation* (Persiapan Data) adalah:

   Setelah sumber data yang tersedia diidentifikasi, sumber data tersebut perlu diseleksi, dibersihkan, dibangun ke dalam wujud yang dikehendaki dan dibentuk. Pembersihan dan transformasi data dalam persiapan model data perlu dilakukan pada tahap ini. Eksplorasi data secara lebih mendalam juga dapat diterapkan dalam tahap ini, dan penggunaan model-model tambahan sekali lagi

memberikan peluang untuk meihat berbagai pola berdasarkan pemahaman bisnis.

d. *Modeling* (Pembuatan Model) adalah:

   Metode penggalian data, seperti visualisasi (penggambaran data dan penetapan hubungan) serta analisis pengelompokan (untuk mengidentifikasi variabel mana yang berhubungan satu sama lain) bermanfaat bagi analisis awal. Alat bantu seperti induksi aturan yang digeneralisasikan dapat mengembangkan aturan-aturan asosiasi awal. Begitu pemahaman data yang lebih luas diperoleh (sering kali melalui pengenalan pola yang dipicu dengan melihat output model), model-model lebih terinci yang sesuai dengan jenis data tersebut dapat diterapkan. Pembagian data ke dalam data latihan dan data uji juga diperlukan untuk pembuatan model.

e. *Evaluation* (Evaluasi) adalah:

   Hasil model sebaiknya dievaluasi dalam konteks tujuan bisnis yang ditetapkan pada tahap awal (pemahaman bisnis). Hal ini akan mengarahkan pada identifikasi kebutuhan lain (kerap kali melalui pengenalan pola), sering kali kembali ke tahap-tahap CRISP-DM sebelumnya. Perolehan pemahaman bisnis merupakan prosedur berulang dalam penggalian data, di mana hasil dari beragam visualisasi, fakta statistik, dan metode kecerdasaan buatan menunjukan hubungan-hubungan baru kepada pengguna yang memberikan pemahaman yang lebih mendalam mengenai operasi perusahaan.

f. *Deployment* (Pelaksanaan) adalah:

   Pengalian data dapat digunakan baik untuk membuktikan hipotesis sebelumnya, ataupun untuk penemuan pengetahuan (pengidentifikasian hubungan yang tidak terduga dan bermanfaat), Melalui pengetahuan yang ditemukan dalam tahap awal proses CRISP-DM, model yang kuat dapat diperoleh yang mungkin kemudian dapat diterapkan pada kegiatan bisnis untuk berbagai keperluan, termasuk memprediksi atau mengidentifikasikan situasi-situasi penting. Model-model ini perlu dipantau untuk mengawasi adanya perubahan dalam operasi, karena apa yang mungkin tepat untuk saat ini mungkin tidak lagi tepat satu tahun ke depan. Jika perubahan besar benar-benar terjadi, model tersebut sebaiknya dibuat ulang. Merupakan hal yang bijaksana

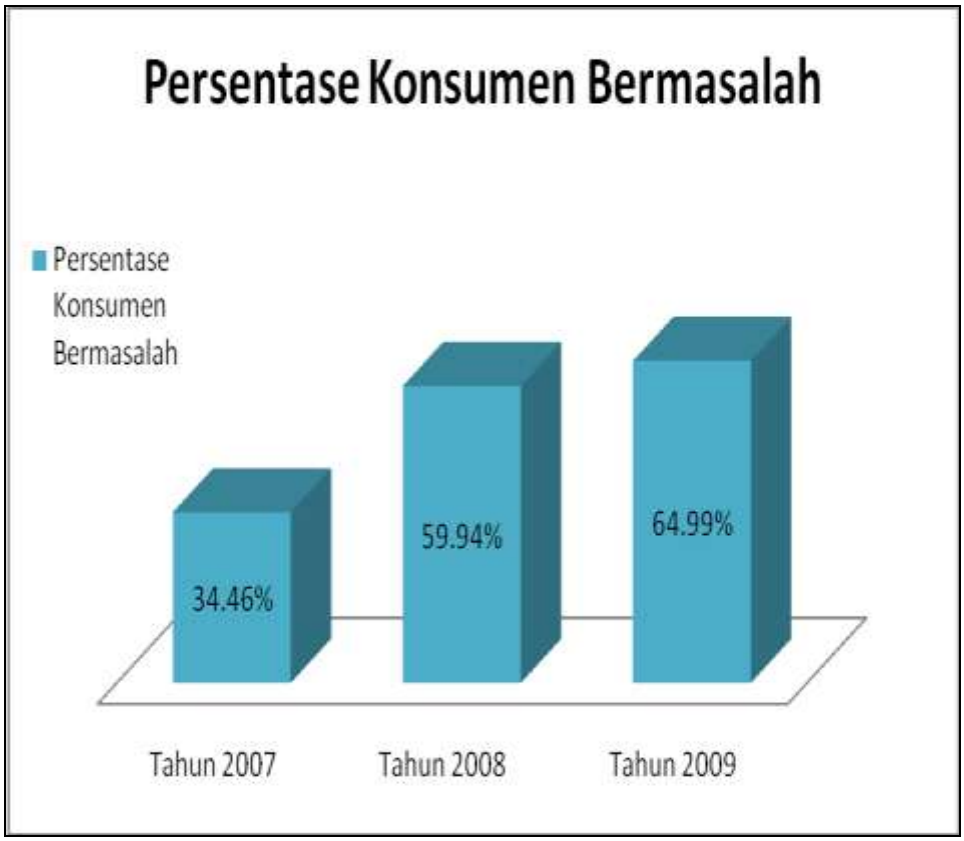
untuk mencatat hasil proyek penggalian data agar bukti-bukti yang terdokumentasi tersedia untuk penelitian di masa mendatang.

Gambar 3.1. Proses Penggalian Data CRISP-DM

3.2. *Business Understanding* (Pemahaman Bisnis)

Bagian penting dari suatu penelitian penggalian data adalah dimana mengetahui untuk apa penelitian tersebut dilakukan. Berawal dari kebutuhan manajerial akan pengetahuan baru, menentukan tujuan akhir, dan membuat rencana untuk mendapatkan pengetahuan seperti itu perlu dikembangkan, berkenan dengan mereka yang bertanggung jawab untuk menggumpulkan data, menganilis data, dan membuat laporan.

Dengan mengumpukan data konsumen kredit yang didapat dari PT AEON Credit Service Indonesia cabang Tangerang diketahui bahwa jumlah konsumen bermasalah tiap tahunnya meningkat. Jumlah konsumen bermasalah tahun 2007 adalah 419 dari 1216 konsumen, tahun 2008 adalah 585 dari 976 konsumen dan tahun 2009 adalah 310 dari 477 konsumen. Dari data tahun 2007 sampai tahun 2009 didapat tingkat tingginya persentasi kredit macet yang menjadi permasalahan.



Gambar 3.2. Grafik Peningkatan Persentase Konsumen Bermasalah

Dengan melihat gambar 3.2 permasalaha meningkatnya jumlah persentase konsumen bermasalah yang terjadi pada PT AEON dalam pembayaran angsuran ini diakibatkan dari analisa yang kurang akurat. Bertujuan untuk mengurangi jumlah konsumen bermasalah akan diterapkan suatu metode klasifikasi pada *data mining.* Berdasarkan jurnal dan referensi lainnya yang sudah dilakukan oleh peneliti-peneliti sebelumnya maka dilakukuan komparasi antara metode klasifikasi *Support Vector Machine* dan *k-Nearest Neighbor* untuk menangani permasalahan yang terjadi sehingga dapat mengurangi tingginya presentase jumlah konsumen bermasalah.

3.3. *Data Understanding* (Pemahaman Data)

Penggalian data berorientasi pada tugas, tugas bisnis yang berbeda membutuhkan kelompok data yang berbeda pula. Hal yang pertama dilakukan dalam proses penggalian data adalah memilih data yang berkaitan dari banyak *database* yang tersedia untuk menggambarkan pemahaman bisnis yang diberikan dengan tepat. Ada tiga hal yang penting untuk dipertimbangkan dalam pemilihan data. Hal yang pertama adalah menetapkan deskripsi masalah dengan ringkas dan jelas. Hal yang kedua adalah mengidentifikasi data yang relevan untuk mendeskripsikan masalah. Dan yang terakhir hal yang ketiga adalah variabel yang terpilih untuk data yang relevan sebaiknya independen satu sama lain.

Independesi variabel-variabel tersebut tidak berisi informasi yang tumpang tindih. Seleksi variabel indenpeden yang teliti dapat mempermudah algoritma penggalian data untuk menemukan pola-pola pengetahuan yang bermanfaat dengan segera. Dimana sumber data untuk pemilihan data bisa beragam, namun data yang didapat berupa data transaksi (*transactional data*). Jenis data dapat dikategorikan sebagai data kuantitatif dan data kualitatif.

Berdasarkan *database* data yang didapat adalah jenis data kuanlitatif (*quantitative data*) menggunakan nilai numerik. Data tersebut berupa data diskret (bilangan bulat) atau kontinu (bilangan rill). Yang berisi dari sejumlah data-data kredit konsumen yang telah diketahui statusnya baik dan buruk. Dalam menentukan kelayakan konsumen penerima kredit, tiga belas atribut predictor dan satu atribut kelas. Dibawah ini adalah atribut-atribut yang menjadi parameter terlihat pada Tabel 3.1.:

Tabel 3.1. Atribut, Nilai dan Keterangan

No	Atribut	Nilai	Keterangan
1	*Age* (umur)	20 tahun	umur konsumen
2	*Sex* (jenis kelamin)	1	laki-laki
		2	perempuan
3	*Marry Status*	1	belum menikah
		2	menikah
4	*Education Level*	1	SLTP
	(tingkat pendidikan)	2	SLTA
		3	Diploma 1/2/3
		4	S1
		5	S2
		6	S3
		7	Master
		8	yang lainnya
5	*Live Year* (lama tinggal)	10 tahun	lama tinggal per tahun
6	*Now House Owner Relate Type*	1	milik sendiri
	(status dari kepemilikan	2	milik keluarga
	rumah tinggal)	3	milik saudara
		4	sewa atau kontrak
		5	kredit
		6	milik kantor atau perusahaan
		7	hipotek bank
7	*Live Person* (jumlah tanggungan)	3 orang	jumlah tanggungan
8	*Work Year* (lama bekerja)	3 tahun	lamanya bekerja
9	*Salary* (gaji)	Rp 3.000.000	besarnya gaji per bulan
10	*Othering* (pendapatan lainnya)	Rp 1.000.000	besarnya pendapatan lainnya per bulan
11	*Business Type*	1	perusahaan swasta
	(Jenis Usaha/ Pekerjaan)	2	pemerintah
		3	perusahaan negara
		4	kaum *professional*
		5	lainnya
12	*Employments Type*	1	pekerja tetap
	(Jenis Pekerjaan)	2	pekerja kontrak
13	*Status*	Good	konsumen yang tidak bermasalah
		Bad	konsumen yang bermasalah

3.4. *Data Preparation* (Persiapan Data)

Merapikan data terpilih untuk mendapatkan kualitas yang lebih baik adalah tujuan dari prapengolahan data. Dimana ada beberapa data yang terpilih mungkin mempunyai format-format yang berbeda karena mereka dipilih dari sumber data yang berbeda-beda. Dapat dikatakan secara umum, pembersihan data berarti menyaring, menggabungkan, dan mengisi kembali nilai-nilai yang hilang (imputasi – *imputation).*

Dengan penyaringan data, data yang terpilih dicari pencilan (*outlier*-nilai yang jauh berbeda dibanding nilai-nilai lainnya dalam data) dan redundansinya. *Outlier* berbeda jauh dari sebagian besar data, atau data yang jelas-jelas berada di luar kisaran kelompok data terpilih. Dengan penghalusan data, nilai-nilai yang hilang dari data terpilih ditemukan dan nilai-nilai yang baru atau masuk akal kemudian ditambahkan. Sebuah nilai yang hilang sering kali menyebabkan tidak adanya solusi ketika algoritma penggalian data diterapkan untuk menemukan pola-pola pengetahuan.

Data yang diperoleh untuk penelitian ini sebanyak 2973 *record* transaksi kredit konsumen baik yang bermasalah maupun yang tidak bermasalah. Untuk mendapatkan data yang berkualitas lebih baik, beberapa teknik *preprocessing* digunakan, yaitu (Vercellis, 2009):

1. *Data validation*

   Kualitas *input* data dapat membuktikan tidak memuaskan karena ketidaklengkapan, 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 transformation*

   Yang paling penting dari analisa *data mining* adalah untuk menerapkan transformasi beberapa untuk dataset sehingga dapat meningkatkan akurasi dan kategorikal sebuah model pembelajaran.

3. *Data reduction*

   Ketika dihadapkan *dataset* kecil, transformasi yang dijelaskan biasanya cukup untuk mempersiapkan data masukan untuk analisa *data mining*. Data yang

berulang (redundansi) adalah informasi yang sama yang tercatat dalam beberapa cara yang berbeda. Ada tiga kriteria utama untuk menentukan apakah suatu teknik reduksi data harus digunakan adalah efisiensi, akurasi, dan kesederhanaan model yang dihasilkan.

Setelah dilakukan *preprocssing* data yang didapat maka jumlah data akan berkurang menjadi 2669 *record* dengan cara direduksi yaitu menghilangkan duplikasi data, seperti yang ada pada Tabel 3.2. merupakan sampel data *training*:

Tabel 3.2. Contoh Data *Training*

NO	AGE	SEX	MARRY STATUS	LIVE YEAR	NOWHOUSEOWNER RELATETYPE	LIVE PERSON	LIVE PERSON NUM	WORK YEAR	SALARY	OTHER INCOME	BUSINESS TYPE	EMPLOYMENTS TYPE	STATUS
1	19	1	1	5	3	3	5	2	1550000	0	1	1	GOOD
2	20	2	1	2	4	3	3	2	1200000	0	1	1	GOOD
3	21	1	1	2	2	1	5	2	855000	0	1	1	GOOD
4	21	2	1	21	2	1	7	3	950000	0	1	1	BAD
5	21	2	1	5	3	3	4	1	1400000	0	1	2	BAD
6	22	1	1	22	2	1	4	2	750000	600000	1	1	GOOD
7	22	2	1	4	2	1	6	4	808000	0	1	2	GOOD
8	22	2	2	3	3	2	2	2	900000	0	1	1	BAD
9	22	1	1	10	2	1	5	3	900000	0	1	1	BAD
10	22	2	1	15	2	1	5	2	1000000	300000	1	1	GOOD
11	22	1	1	3	2	1	5	1	1100000	125000	1	1	BAD
12	22	2	1	12	2	1	5	1	1100000	0	1	1	GOOD
13	22	2	1	3	3	3	3	3	1175000	0	1	1	GOOD
14	22	2	1	15	2	1	5	0	1500000	0	1	2	BAD
15	22	2	1	22	2	1	5	2	1500000	0	1	1	GOOD
16	22	1	1	27	3	1	4	1	1800000	0	1	1	GOOD
17	22	2	1	20	1	1	5	2	2500000	0	1	1	GOOD
18	22	1	1	19	1	1	3	4	5000000	10000000	3	2	BAD
19	23	2	1	12	2	1	5	1	500000	0	1	1	GOOD
20	23	2	1	9	2	1	5	2	600000	0	1	1	GOOD

3.5. *Modeling* (Pembuatan Model)

Modeling adalah suatu tahapan dimana peranti lunak penggalian data digunakan untuk memproduksi hasil untuk berbagai situasi. Analisis pengelompokan dan eksplorasi *visual* data biasanya diterapkan lebih dulu. Bergantung pada jenis data, berbagai model baru kemudian diterapkan, dengan tujuan memungkinkan pengguna untuk bekerja dengan data guna memperoleh pemahaman. Teknik yang digunakan dalam penggalian data ini adalah klasifikasi.

Klasifikasi (*Classification*), metode-metode ditujukan untuk pembelajaran fungsi-fungsi berbeda yang memetakan masing-masing data terpilih ke dalam salah satu dari kelompok kelas yang telah ditetapkan sebelumnya. Penelitian ini menggunakan model klasifikasi yaitu *Support Vector Machines* yaitu metode yang mencari fungsi pemisah (*hyperplane*) terbaik untuk memisahkan data-data

dengan kelas-kelas yang berbeda dan metode *k-Nearest Neighbor* (k-NN) yang adalah merupakan suatu metode yang paling sering digunakan untuk klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.

3.6. *Evaluation* (Evaluasi)

Tahap evaluasi data sangatlah penting, dengan mengasimilasikan pengetahuan dari data yang telah digali. Untuk mengevaluasi pola dengan menggunakan *software* RapidMiner. Evaluasi dan validas menggunakan metode *cross validation, confusion matrix* dan kurva ROC. Evaluasi yang baik mengarahkan pada keputusan-keputusan bisnis yang produktif, sementara analisis evaluasi yang buruk mungkin melewatkan informasi yang bermanfaat.

3.7. *Deployment* (Pelaksanaan)

Setelah melewati tahap *modeling* dan *evaluation* dan tahap-tahap sebelumnya, selanjutnya pada tahap ini ditetapkan model yang dianggap paling akurat untuk diterapkan dalam penentuan kelayakan pemberian kredit terhadap konsumen.

3.8. Jadwal Penelitian

Dalam melakukan penelitian ada beberapa tahapan yang dilakukan yaitu:

1. Identifikasi masalah dan analisa kebutuhan

   Tahap ini adalah dengan dimulai dengan mengindentifikasi masalah yang berkaitan dengan pembiayaan kredit kepada konsumen. Dan setelah melakukan identifikasi kemudian merumuskan masalah yang terjadi. Dari masalah yang ditemukan selanjutnya dilakukan analisa kebutuhan untuk memecahkan masalah tersebut.

2. Pengumpulan data

   Tahap ini dilakukan mulai dari pengumpulan data yang dibutuhkan dilakukan dengan cara studi literatur, observasi, dan melakukan tanya jawab kepada bagian analisa kredit. Data yang dikumpulkan kemudian dianalisa untuk menentukan atribut dan *record* mana yang diperlukan dan tidak diperlukan untuk tahap selanjutnya.

3. Eksperimen

   Tahap ini adalah menentukan model yang digunakan untuk dilakukan pengujian dengan memasukkan data *training* ke dalam model menggunakan *software* RapidMiner.

4. Implementasi

   Tahap ini menerapkan model yang dihasilkan ke dalam sistem untuk menganalisa konsumen yang termasuk *good* atau *bad* sehingga dapat dipakai oleh pengguna.

5. Evaluasi

   Tahap ini dilakukan evaluasi untuk mengukur berhasil atau tidaknya model yang telah dikembangkan paling akurat dalam menentukan kelayakan konsumen dan mengukur keakuratan hasil yang telah dicapai.

6. Penulisan

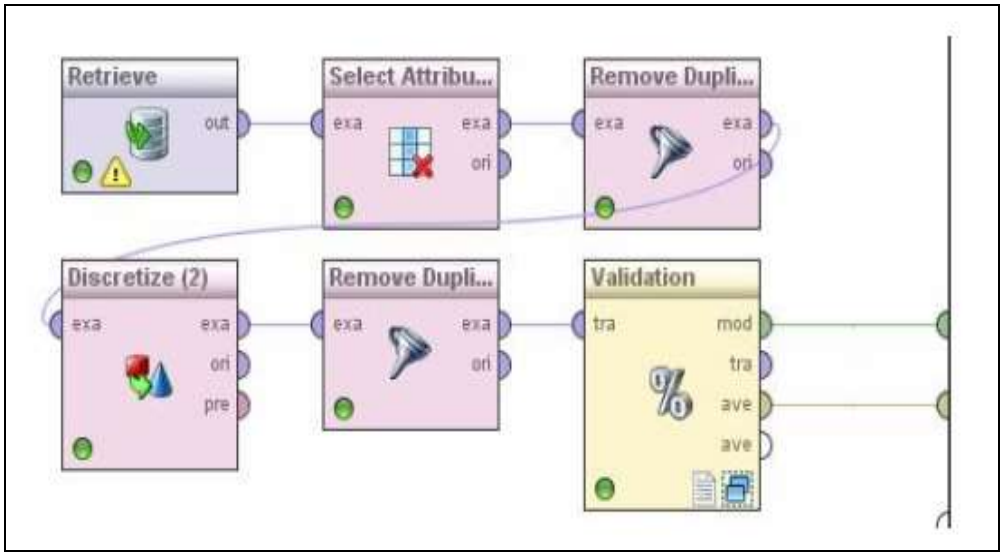
   Tahap ini penelitian dituangkan dalam bentuk laporan (tesis). Supaya lebih efisien, pembuatan laporan dilakukan sejalan dengan tahap lain yang dilakukan dalam penelitian.

BAB IV
HASIL PENELITIAN DAN PEMBAHASAN

4.1. Hasil Penelitian

Tujuan dari penelitian ini adalah mengetahui tingkat keakurasian dari dua algoritma klasifikasi *data mining* pada data konsumen dalam bentuk kredit yaitu semua data yang telah disetujui oleh pihak perusahaan pembiayaan. Untuk menentukan tingkat keakurasian maka hasil dari analisis algoritma *Support Vector Machine* dan *k-Nearest Neighbor* akan dibandingkan atau dikomparasikan.

Sebelum melakukan komparasi, masing-masing algoritma akan dilakukan pengujian kinerjanya. Cara standar untuk memprediksi tingkat kesalahan pada sampel menggunakan *10-Fold Cross Validation*. Data tersebut dibagi secara acak menjadi 10 bagian dimana kelas diwakili disekitar sama proporsi seperti *data set* lengkap. *10-Fold Cross Validation* telah menjadi metode standar dan cukup untuk mendapatkan perkiraan kesalahan yang dapat diandalkan (Witten, 2011). Dengan desain modelnya seperti pada gambar 4.1.



Gambar 4.1. Desain Model Validasi

4.2. Pengujian Model

4.2.1. Pengujian Model *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 *accuracy* = 67.59%, *precission* = 67.94%, *recall* = 98.80% seperti pada bagan 4.1. dibawah ini:

Gambar 4.2. *Performance Vector Support Vector Machine*

PerformanceVector

```
PerformanceVector:
accuracy: 67.59% +/- 0.57% (mikro: 67.59%)
ConfusionMatrix:
True:   BAD     GOOD
BAD:    17      20
GOOD:   780     1651
precision: 67.94% +/- 0.60% (mikro: 67.91%) (positive class: GOOD)
ConfusionMatrix:
True:   BAD     GOOD
BAD:    17      20
GOOD:   780     1651
recall: 98.80% +/- 3.59% (mikro: 98.80%) (positive class: GOOD)
ConfusionMatrix:
True:   BAD     GOOD
BAD:    17      20
GOOD:   780     1651
AUC (optimistic): 0.758 +/- 0.043 (mikro: 0.758) (positive class: GOOD)
AUC: 0.758 +/- 0.043 (mikro: 0.758) (positive class: GOOD)
AUC (pessimistic): 0.758 +/- 0.043 (mikro: 0.758) (positive class: GOOD)
```

1. *Confusion Matrix*

   Tabel 4.1. adalah perhitungan berdasarkan data *training* pada Tabel 4.1., diketahui dari 2468 data, 17 diklasifikasikan *bad* sesuai dengan prediksi yang dilakukan dengan metode *Support Vector Machine*, lalu 780 data diprediksi *bad* tetapi ternyata *good*, 1651 data *class good* diprediksi sesuai, dan 780 data diprediksi *good* ternyata *bad*.

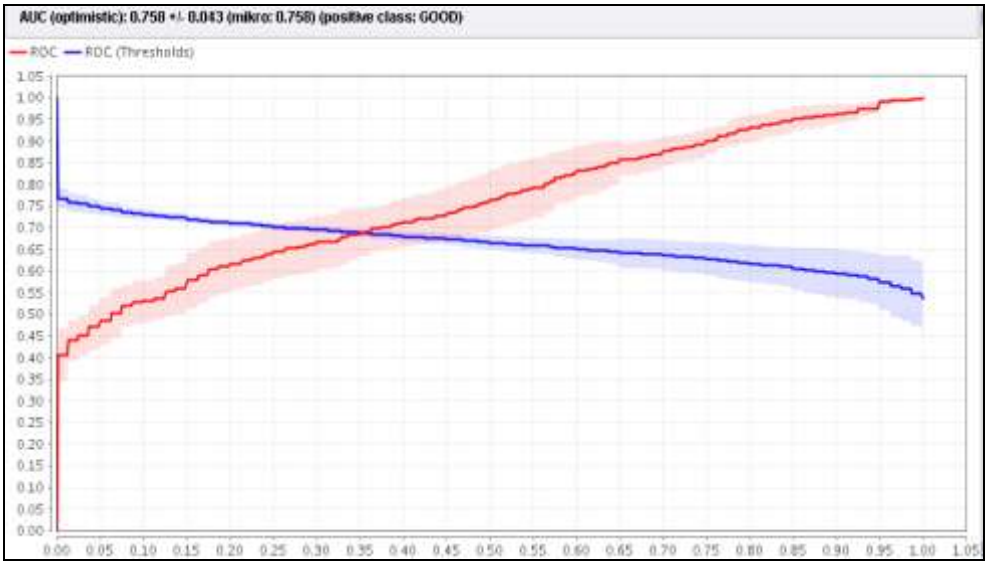
Tabel 4.1 Model *Confusion Matrix* untuk Metode *Support Vector Machine*

accuracy: 67.59% +/- 0.57% (mikro: 67.59%)

	true BAD	true GOOD	class precision
pred. BAD	17	20	45.95%
pred. GOOD	780	1651	67.91%
class recall	2.13%	98.80%	

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 Machines*. 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 Machines*

4.2.2. Pengujian Model *k-Nearest Neighbor*

Nilai *accuracy*, *precision*, dan *recall* dari *data training* dapat dihitung dengan menggunakan RapidMiner. Hasil pengujian dengan menggunakan model *k-Nearest Neighbor* didapatkan hasil *accuracy* = 81.60%, *precission* = 87.90%, *recall* = 84.50% seperti pada bagan 4.2 dibawah ini:

Gambar 4.4. *Performance Vector k-Nearest Neighbor*

PerformanceVector

```
PerformanceVector:
accuracy: 81.60% +/- 1.84% (mikro: 81.60%)
ConfusionMatrix:
True:   BAD     GOOD
BAD:    602     259
GOOD:   195     1412
precision: 87.90% +/- 1.66% (mikro: 87.87%) (positive class: GOOD)
ConfusionMatrix:
True:   BAD     GOOD
BAD:    602     259
GOOD:   195     1412
recall: 84.50% +/- 2.60% (mikro: 84.50%) (positive class: GOOD)
ConfusionMatrix:
True:   BAD     GOOD
BAD:    602     259
GOOD:   195     1412
AUC (optimistic): 0.962 +/- 0.008 (mikro: 0.962) (positive class: GOOD)
AUC: 0.500 +/- 0.000 (mikro: 0.500) (positive class: GOOD)
AUC (pessimistic): 0.638 +/- 0.032 (mikro: 0.638) (positive class: GOOD)
```

1. *Confusion Matrix*

   Tabel 4.2. adalah perhitungan berdasarkan data *training* pada Tabel 4.2., diketahui dari 2468 data, 602 diklasifikasikan *bad* sesuai dengan prediksi yang dilakukan dengan metode *k-Nearest Neighbor*, lalu 259 data diprediksi *bad* tetapi ternyata *good*, 1412 data *class good* diprediksi sesuai, dan 195 data diprediksi *good* ternyata *bad*.

Tabel 4.2. Model *Confusion Matrix* untuk Metode *k-Nearest Neighbor*

accuracy: 81.60% +/- 1.84% (mikro: 81.60%)

	true BAD	true GOOD	class precision
pred. BAD	602	259	69.92%
pred. GOOD	195	1412	87.87%
class recall	75.53%	84.50%	

2. Kurva ROC

Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua metode komparasi bisa dilihat pada Gambar 4.3 yang merupakan kurva ROC untuk algoritma *k-Nearest Neighbor*. Kurva ROC pada gambar 4.3 mengekspresikan *confusion matrix* dari Tabel 4.2. Garis horizontal adalah *false positives* dan garis vertikal *true positives*.

Gambar 4.5. Kurva ROC dengan Metode *k-Nearest Nieghbor*

4.3. Analisis Hasil Komparasi

Dari hasil analisis model yang dihasilkan dengan metode algoritma *Support Vector Machine* dan *k-Nearest Neighbor* diuji menggunakan metode *Cross Validation* maka dapat dirangkumkan seperti tabel 4.3:

Tabel 4.3 Komparasi Nilai A*ccuracy* dan AUC

	Support Vector Machine (SVM)	*k-Nearest Neighbor* (KNN)
Accuracy	67.59%	81.60%
AUC	0.758	0.962

Dari tiga table *confusion matrix*, selanjutnya dilakukan perhitungan nilai *accuracy, precision, sensitivity*, dan *recall*. Perbadingan nilai *accuracy, preMcision, sensitivity*, dan *recall* yang telah dihitung untuk metode *Support Vector Machine* dan *k-Nearest Neighbor* dapat dilihat pada Tabel 4.4.

Tabel 4.4. Komparasi Nilai A*ccuracy, Precision* dan *Recall*

	Support Vector Machine (SVM)	*k-Nearest Neighbor* (KNN)
Accuracy	67.59%	81.60%
Precision	67.94%	87.90%
Recall	98.80%	84.50%

Tabel 4.3 membandingkan *accuracy* dan AUC dari tiap metode. Terlihat bahwa nilai *accuracy k-Nearest Neighbor* paling tinggi begitu pula dengan nilai AUC-nya. Untuk metode *Support Vector Machine* juga menunjukan nilai yang cukup sesuai. Untuk keakuransian nilai AUC dalam klasifikasi *data mining* dibagi menjadi lima kelompok (Gorunescu, 2011), yaitu:

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

Dilihat dari pengelompokkan di atas dan Tabel 4.3 maka dapat disimpukan bahwa metode *Support Vector Machine* termasuk klasifikasi yang cukup karena memiliki nilai AUC antara 0.70-1.0 sedangkan *k-Nearest Neighbor* termasuk klasifikasi yang sangat baik karena memiliki nilai AUC antara 0.90-1.00.

4.4. Implikasi Penelitian

Hasil evaluasi menunjukkan ternyata algoritma *k-Nearest Neighbor* terbukti paling akurat dibanding *Support Vector Machine*. Kedua metode klasifikasi tersebut diterapkan pada data konsumen yang mendapatkan kredit pembiayaan. Dengan hasil ini, menunjukkan bahwa metode *k-Nearest Neighbor* merupakan metode yang cukup baik dalam pengklasifikasian data, dengan demikian

algoritma *k-Nearest Neighbor* dapat memberikan pemecahan untuk permasalahan penentuan kelayakan konsumen yang mendapatkan kredit pembiayaan.

Dengan hasil ini maka algoritma model *k-Nearest Neighbor* dapat mendukung pengambilan keputusan dan pengembangan sistem informasi manajemen strategik, model ini dapat diterapkan pada perusahan pembiayaan atau *leasing* dengan menggunakan *software* RapidMiner.

Penelitian ini diharapkan bisa digunakan pada perusahaan pembiayaan untuk lebih meningkatkan akurasi analisa kelayakan kredit bagi konsumen yang hendak mengajukan kredit.Dalam mendukung pengambilan keputusan dan pengembangan sistem informasi manajemen strategik, model ini dapat diterapkan pada perusahaan pembiayaan dengan menerapkan sistem yang menggunakan perangkat keras dan perangkat lunak, disertai dengan pembuatan *Standard Operational Procedure* dan pelatihan bagi *end-user*.

BAB V
PENUTUP

5.1. Kesimpulan

Dalam penelitian ini dilakukan pembuatan dengan menggunakan dua model algoritma yaitu S*upport Vector Machines* dan *k-Nearest Neighbor* menggunakan data konsumen yang mendapat kredit pembiayaan. Model yang dihasilkan, dikomparasi untuk mengetahui algoritma yang paling baik dalam penentuan resiko kredit konsumen. Untuk mengukur kinerja keduaa algoritma tersebut digunakan metode pengujian *Cross Validation, Confusion Matrix* dan Kurva ROC, diketahui bahwa algoritma *k-Nearest Neighbor* memiliki nilai *accuracy* dan AUC paling tinggi dan yang paling rendah metode S*upport Vector Machines*.

Dari hasil analisis tersebut dapat disimpulkan bahwa metode *k-Nearest Neighbor* merupakan metode yang cukup baik dalam pengklasifikasian data. Dengan demikian algoritma *k-Nearest Neighbor* dapat memberikan pemecahan untuk permasalahan penentuan kelayakan konsumen untuk mendapatkan kedit pembiayaan pada perusahaan pembiayaan.

5.2. Saran

Meskipun telah dilakukan komparasi dengan menggunakan dua metode klasfikasi, dan mendapatkan hasil bahwa metode *k-Nearest Neighbor* merupakan metode yang cukup baik dalam pengklasifikasian data, namun ada beberapa hal yang dapat ditambahkan agar penelitian ini bisa ditingkatkan, berikut adalah saran-saran yang diusulkan:

1. Penelitian ini diharapkan bisa digunakan pada perusahaan pembiayaan untuk lebih meningkatkan akurasi analisa kelayakan kredit bagi konsumen yang hendak mengajukan kredit.
2. Dalam mendukung pengambilan keputusan dan pengembangan sistem informasi manajemen strategik, model ini dapat diterapkan pada perusahaan pembiayaan dengan menerapkan sistem yang menggunakan perangkat keras dan perangkat lunak, disertai dengan pembuatan *Standard Operational Procedure* dan pelatihan bagi *end-user*.

3. Menambahkan lagi beberapa algoritma klasifikasi *data mining* untuk dikomparasikan seperti *K-Means. Artifical Neural Network, AdaBoost, Naïve Bayes*, CART dan lain-lainya.

DAFTAR PUSTAKA

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

Bramer, Max. (2007). *Principles of Data Mining*. London: Springer

David Olson & Yong Shi (2008). *Pengantar Ilmu Penggalian Data Bisnis*. Jakarta: Penerbit Salemba Empat.

Dima, A. M., & Vasilache, S., (2009) *ANN Model for Corporate Credit Risk Assessment*. IEEE.

Dong, G., Kin, K.L., & Zhou, L., (2009). *Simulated Annealing Based Rule Extraction Algorithm Credit Scoring Problem*. Hong Kong. IEEE.

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

Han, J.,&Kamber, M. (2006).*Data Mining Concept and Tehniques*.San Fransisco: Morgan Kauffman.

Islam, M. J., Wu, Q. M. J., Ahmadi, M., Ahmed, S., (2007). *Investigating the Perpormance of Naïve-Bayes Classifiers and K-Nearest Neighbor Classifiers*. IEEE.

Jiang, Yi., (2009). *Credit Scoring Model Based on the Decision Tree and the Simulated Annealing Algorithm*. China. IEEE.

Keramati, A., & Yousefi, N., (2011). *A Proposed Classification of Data Mining Techniques in Credit Scoring*. Malaysia

Kotsiantis, S., Kanellopoulos, D., Karioti, V., & Tampakas, V. (2009). *An Ontology-based Portal for Credit Risk Analysis*. IEEE.

Kusrini, & Luthfi, E. T. (2009). Algoritma Data Mining. Yogyakarta: Andi Publishing.

Lai, K. K., Yu, L., Zhou, L., & Wang , S. (2006). *Credit Risk Evaluation with Least Square Support Vector Machine*. China.

Larose, D. T. (2005).*Discovering Knowledge in Data*. New Jersey: John Willey & Sons, Inc.

Li, J., Liu, J., Xu, Weixuan., & Shi, Yong. (2003). *Support Vector Machine Approach to Credit Assessment*. China.

Liao,Warrwn T ., (2007). *Recent Advances in Data Mining of Enterprise Data: Algoritms and Applicaions*. Singapore: World Scientific Publishing

Maimon, Oded & Rokach, Lior.(2005). *Data Mining and Knowledge Discovery Handbook*. New York: Springer

Nurgroho, A. S., Witarto, A. B., & Handoko, D., (2003). *Support Vector Machine*. IlmuKomputer.Com

Rivai, Veithzal., & Veithzal, Andria Permata. (2006). *Credit Management Handbook*. Jakarta: Raja GrafindoPersada.

Santosa, B. (2007). *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Jogyakarta: Graha Ilmu.

Wang, Q., Lai, K. K., & Niu, D., (2011). *Green Credit Scoring System and its Risk Assessement Model with Support Vector Machine*. China. IEEE.

Vercellis, Carlo (2009). *Business Intelligent: Data Mining and Optimization for Decision Making*. Southern Gate, Chichester, West Sussex: John Willey & Sons, Ltd.

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

Wu, Xindong& Kumar, Vipin. (2009). The Top Ten Algorithms in Data Mining. Boca Raton: CRC Press

Yu, H., Huang, X., Hu, X., & Cai. H,. (2010). *A Comparative Study in Data Mining Algorithms for Individual Credik isk Evaluation*.

Zhang, D., Hifi, M., Chen, Q., & Ye, W., (2008). *A Hybrid Credit Scoring Model Based on Genetic Programming and Support Vector Machines*. China.IEEE

Zhang. D., Leung, S. C. H., Ye, Zhime., (2008) *A Decision TreeScoring Model Based on Genetic Algoritm and K-means Algorithm*. IEEE.