PREDIKSI MAHASISWA DROP OUT MENGGUNAKAN METODE SUPPORT VECTOR MACHINE

Siti Nurhayati*[1], Kusrini[2], Emha Taufiq Luthfi[3]
[1]Universitas Yapis Papua
[2,3]Magister Teknik Informatika STMIK AMIKOM Yogyakarta
[1]stnurhayati37@yahoo.com ,[2]kusrini@amikom.ac.id, [3]emhataufiqluthfi@amikom.ac.id

Abstrak

Tingginya tingkat keberhasilan mahasiswa dan rendahnya tingkat kegagalan mahasiswa dapat mencerminkan kualitas dari suatu perguruan tinggi. Salah satu indikator kegagalan mahasiswa adalah kasus drop out. Untuk mengatasi permasalah, dilakukan prediksi menggunakan metode support vector machine. Support Vector Machine berusaha mencari hyperplane yang optimal dimana dua kelas pola dapat dipisahkan dengan maksimal, parameter yang di gunakan pada Support Vector Machine hanya parameter kernel dalam satu parameter C yang memberikan pinalti pada titik data yang di klasifikasikan secara acak. Dalam Support Vector Machine bobot (w) dan bias (b) merupakan solusi global optium dari quadratic programming sehingga cukup dengan sekali running akan menghasilkan solusi yang akan selalu sama untuk pilihan kernel dan parameter yang sama. Melalui penerapan support vector machine diharapakan untuk mendapatkan parameter Support Vector Machine yang digunakan tepat untuk memperoleh margin terbaik dalam memprediksi mahasiswa drop out.

Kata Kunci— *prediksi drop out, kernel, support vector machine, unified modeling language*

Abstract

High level of success and low level of success of college students can reflect the quality of a university. One of the indicators of college students' success is drop out case. To solve the problem, a prediction by support vector machine method was made. Support Vector Machine searched for optimal hyperplane where two pattern classes can be separated maximally. The parameter used in Support Vector Machine is kernel parameter in a C parameter which gives penalty to data points which are classified randomly. In Support Vector Machine, weight (w) and bias (b) are optimum global solutions of quadratic programming so one run is enough to produce a solution which will always be the same for the same kernel selection and parameter. The implementation of support vector machine is expected to produce Support Vector Machine parameter to obtain the best margin to predict dropped out college students.

Keywords*— drop out prediction, kernel, support vector machine, unified modeling language*

1. PENDAHULUAN

Perguruan tinggi merupakan penyelenggara pendidikan akademik bagi mahasiswa yang diharapkan dapat menyelenggarakan pendidikan yang berkualitas untuk menghasilkan sumberdaya manusia yang berilmu, cakap dan kreatif untuk mendukung tercapainya pembangunan nasional. Pendidikan dikatakan berhasil apabila telah memenuhi tujuan pendidikan nasional dan proses belajar mengajar dilaksanakan secara efektif dan efesien. Keberhasilan atau prestasi belajar mahasiswa sering dilihat sebagai kesuksesan dan keunggulan

pihak perguruan tinggi. Sebaliknya kegagalan atau rendahnya kualitas mahasiswa sering dilihat sebagai ketidakmampuan pihak perguruan tinggi menyelenggarakan proses pendidikan tinggi.

Drop out atau pemberhentian status mahasiswa adalah proses pencabutan status kemahasiswaan atas diri mahasiswa, yang disebabkan oleh hal-hal tertentu yang telah di tentukan oleh universitas yang bersangkutan. Tingginya jumlah mahasiswa drop out pada perguruan tinggi dapat diminimalisir dengan kebijakan dari perguruan tinggi untuk mengarahkan dan mencegah mahasiswa dari dropout [1] bahwa mendeteksi mahasiswa berisiko pada tahap awal pendidikan sangat penting dilakukan untuk menjaga mahasiswa dari dropout. Hal ini memungkinkan departemen penyelenggara pendidikan untuk memberikan pengarahan kepada mahasiswa yang membutuhkan. Oleh karena itu, perlu dilakukan kajian atau prediksi mahasiswa drop out sehingga dapat dijadikan informasi yang bermanfaat untuk memperkirakan tingkat drop out mahasiswa pada tahun-tahun yang akan datang dan mengurangi tingkat drop out mahasiswa. Prediksi drop out mahasiswa dapat membantu pihak institusi dalam pengambilan keputusan. Bagaimana menganalisa faktor penyebab drop out mahasiswa dan mendapatkan parameter terbaik dari metode yang digunakan untuk dapat memprediksi mahasiswa yang berpeluang drop out.

Prediksi drop out dapat dilakukan dengan serangkaian proses mendapatkan pengetahuan atau pola dari kumpulan data yang di sebut data mining. Data mining memecahkan masalah dengan menganalisis data yang telah ada dalam database. Beberapa algoritma klasifikasi data mining telah digunakan untuk memprediksi perilaku mahasiswa yang berpotensi drop out diantaranya *decision tree, neural network, naïve bayes,instance-based learning, logistic regression* dan *support vector machine.*

Berdasarkan penelitian sebelumnya yang dilakukan oleh [2] pada Universitas Dian Nuswantoro dengan melakukan analisis komparasi algoritma untuk prediksi mahasiswa Non Aktif menunjukkan bahwa algoritma decision tree merupakan algoritma yang paling akurat, namun demikian decision tree tidak dominan terhadap algoritma yang lain. Berdasarkan hasil penelitian tersebut juga, *logistic regression, decision tree, naïve bayes* dan *neural network* masuk dalam kategori excellent classification.

Menurut panelitian yang dilakukan oleh [3] mengenai Analisis Prediksi DO mahasiswa dalam *educational data mining* menggunakan jaingan saraf tiruan menunjukkan model yang di usulkan dapat digunakan untuk memprediksi potensi drop out dengan akurasi kemungkinan klasifikasi drop out sebesar 98,91% dengan tingkat signifikan sensifitas terbesar perilaku sosial sebesar 4,737. Hasil analisa sensifitas menunjukkan variabel iput untuk parameter sosial yang paling berpengaruh adalah kualitas interaksi dengan teman dan variabel hubungan keluarga, sedangkan variabel ipk untuk parameter studi lanjut dan motivasi untuk parameter akademik yang paling berpengaruh adalah SKS dan IPK.

Metode yang dapat digunakan untuk memprediksi adalah *Support Vector Machine* (SUPPORT VECTOR MACHINE). *Support Vector Machine* dikembangkan oleh Boser, Guyon, Vapnik pada tahun 1992. *Support Vector Machine* (SUPPORT VECTOR MACHINE) adalah metode klasifikasi yang bekerja dengan cara mencari *hyperplane* dengan margin terbesar. *Hyperplane* adalah garis batas pemisah data antar kelas, sedangkan margin adalah jarak antara *hyperplane* dengan data terdekat pada masing-masing kelas. Adapun data terdekat dengan *hyperplane* pada masing-masing kelas inilah yang disebut *support vector*.

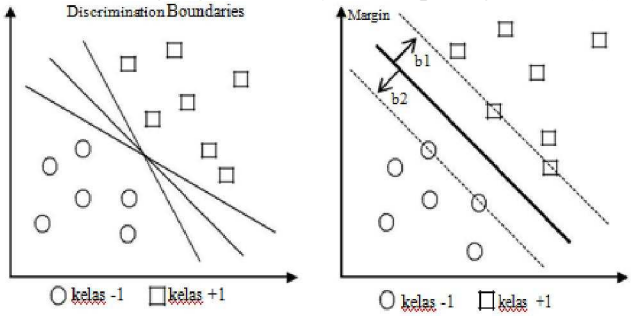
Penelitian [4] dengan judul Pemodelan Resiko Kredit dengan Pendekatan Support Vector Machine Penelitian ini menggunakan pendekatan Support Vector Machine untuk klasifikasi diaplikasi dalam manajemen resiko kredit. Klasifikasi dilakukan untuk memisahkan aplikasi kredit dari dua kelas yakni kelas baik dan kelas buruk. Data yang digunakan dari bank BPR Tasikmalaya dari bulan September 2005 sampai maret 2010. Dari 602 data yang di gunakan 402 data sebagai data training dan 200 data testing dengan fungsi polynomial kernel berderajat 2. Akurasi yang diperoleh jika di bandingkan dengan hasil klasifikasi berdasarkan bank Indonesia sebesar 71.5 % .

Support Vector Machine (SUPPORT VECTOR MACHINE) merupakan teknik klasifikasi yang *semi-eager learner*. Karena selain memerlukan proses pelatihan, Support Vector Machine juga menyimpan sebagian kecil data latih untuk digunakan kembali pada saat proses prediksi. Support Vector Machine memberikan model klasifikasi yang solusinya *global optimal,* yaitu selalu memberikan model yang sama dan solusi dengan margin maksimal. Tidak membutuhkan pemilihan parameter-parameter, hanya menentukan fungsi kernel yang harus digunakan (untuk kasus data yang distribusi kelasnya tidak dapat dipisakan secara linear). Penggunaan matriks kernel mempunyai keuntungan yaitu kinerja set data dengan dimensi besar tetapi jumlah datanya sedikit akan lebih cepat karena ukuran data pada dimensi baru berkurang banyak [5].

1.1. Support Vector Machine

Support Vector Machine (SUPPORT VECTOR MACHINE) adalah suau teknik yang relatif baru (1995) untuk melakkan prediksi, baik dalam kasus klasifikasi maupun regresi, yang sangat populer belakangan ini. Support vector machine berada dalam satu kelas dengan Neural Network dalam hal fungsi dan konsidi permasalahan yang bisa di selesaikan, keduanya masuk kedalam kelas *supervised learning*. Baik para ilmuwan maupun para praktisi telah banyak melakukan penelitian dengan menerapkan teknik ini untuk menyelesaikan masalah-masalah nyata dalam kehidupan. Baik dalam masalah gene expression analysis, finansial, cuaca hingga bidang kedokteran terbukti dalam banyak implemetasi support vector machine memberi hasil yang lebih baik dari neural network terutama dalam hal solusi yang di capai [6].

Secara teoritik Support Vector Machine dikembangkan untuk masalah klasifikasi dengan dua kelas sebagai usaha mencari *hyperplane* terbaik. *Hyperplane* merupakan fungsi pemisah antara dua kelas pada input space (Vapnik, 1992). Pada gambar 2.1 diperlihatkan beberapa data yang merupakan anggota dari dua buah kelas. Kelas positif dinotasikan dengan +1 dan kelas negatif di notasikan dengan -1. Data yang termasuk dalam kelas negatif di simbolkan dengan lingkaran. Proses pembelajaran problem klasifikasi di terjemahkan untuk menemukan garis (*hyperplane*) yang memisahkan antara dua kelompok tersebut. Pada gambar 1(a) ditunjukkan alternatif garis pemisah (discrimination boundaries), sedangkan pada gambar 1(b) diperlihatkan bahwa terdapat garis *hyperplane* yang tepat berada diantara dua buah kelas. Prinsip dasar dari analisis ini adalah menemukan hyperplane terbaik yakni dengan meminimalkan kesalahan klasifikasi dan memaksimalkan margin geometriknya seerti pada gambar 1(b) [7].



Gambar 1. Batas Keputusan untuk Set Data

Pengklasifikasi Support Vector Machine menggunakan sebuah fungsi atau hyperplane untuk memisahkan dua buah kelas pola. Support Vector Machine berusaha mencari hyperplane yang optimal dimana dua kelas pola dapat dipisahkan dengan maksimal [5]. Support Vector Machine hanya memerlukan parameter kernel (tergantung pada fungsi kernelnya) dalam satu parameter C yang memberikan pinalti pada titik data yang di klasifikasikan secara acak . Dalam Support Vector Machine bobot (w) dan bias (b) merupakan solusi globaloptium dari *quadratic programming* yang merupakan formulasi matematika dari Support Vector Machine sehingga cukup dengan sekali *running* akan menghasilkan solusi yang akan selalu sama untuk pilihan kernel dan parameter yang sama.

1. Karakteristik Support Vector Machine [7].
   a. Support Vector Machine sebenarnya bisa di katakan sebagai teknik klasifikasi yang semieager learner karena memerlukan proses pelatihan. Support Vector Machine menyimpan sebagian kecil data latih untuk digunakan kembali pada saat prediksi, sebagian data yang masih disimpan merupakan *support vector*.
   b. Untuk parameter yang sama digunakan untuk klasifikasi, Support Vector Machine memberikan model klasifikasi yang solusinya adalah global optimum, hal ini berarti Support Vector Machine selalu memberikan model yang sama dan solusi dengan margin maksimal.
   c. Proses pelatihan yang dilakukan Support Vector Machine tidak sebanyak ANN, tetapi sering kali memberikan kinerja yang lebih baik dapi pada ANN.
   d. Tidak membutuhkan pemilihan parameter-parameter seperti ANN. Pada Support Vector Machine hanya menentukan fungsi kernel yang harus digunakan (untuk kasus data yang distribusi kelasnya tidak dapat dipisahkan secara lienar)
   e. Support Vector Machine membutuhkan komputasi pelatihan dan predisi yang rumit karena data yang digunakan dalam proses pelatihan dan prediksi lebih besar dari pada dimensi sesungguhnya.
   f. Untuk set data berjumlah besar Support Vector Machine membutuhkan memori yang sangat besar untuk alokasi matriks kernel yang digunakan.
   g. Penggunaan matriks kernel mempunyai keuntungan lain, yaitu kinerja set data dengan dimensi besar tetapi jumlah datanya sedikit akan lebih cepat karena ukuran data pada dimensi baru berkurang banyak.
2. Kelebihan Support Vector Machine
   a. Support Vector Machine dapat digunakan untuk sampel data yang jumlahnya terbatas dengan jumlah variabel dependen yang besar.
   b. Algoritma dalam Support Vector Machine merupakan transformasi akhir optimasi dari *quadratic programming*. Secara teoritis dapat diperoleh nilai optimasi global untuk menghadapai masalah yang tidak dapat dihidari dalam masalah optimasi lokal.
   c. Algoritma Support Vector Machine berbentuk pemetaan non-linear dari space data original kedalam suatu feature space yang berdimensi tinggi dimana konstruksi dari fungsi diskriminasi linear untuk menggantikan fungsi non-liear dalam space data original. Hal ini menjamin bahwa Support Vector Machine memiliki kemampuan generalisasi.
3. Tujuan proses linear Support Vector Machine

Tujuan utama linear Support Vector Machine untuk menemukan fungsi pemisah (*klasifieer*) yang optial untuk memisahkan dua buah kelompok data dari dua kelas yang berbeda (Vapnik, 1995). Untuk menemukan persamaan garis terbaik yang memisahkan kelas secara benar yakni dengan meminimalkan kesalahan (*error*). Seperti analisis diskriminan linear, linear Support Vector Machine mempunyai beberapa tujuan yaitu :

a. Mengklasifikasi suatu observasi kedalam sebuah kelas.
b. Menguji apakan kelas tersebut berbeda secara signifikan

Proses linear Support Vector Machine (SUPPORT VECTOR MACHINE) untuk membentuk sebuah fungsi pemisah (*hyperpelane*) adalah :

a. Menentukan tujuan penelitian
b. Membuat desain linear Support Vector Machine Pemilihan variabel
c. Membentuk fungsi linear Support Vector Machine
d. Melakukan interpretasi terhadap fungsi linear Support Vector Machine yang terbentuk
e. Melakukan validasi fungsi linear Support Vector Machine.

1.1.1. Support Vector Machine Liner

Dalam linear Support Vector Machine permisah merupakan fungsi linear. Setiap data latih dinyatakan oleh (x_i, y_i) dengan $i = 1, 2, ..., N$, dan $x_i = \{x_{i1}, x_{i2}, ..., x_{iq}\}^T$ merupakan atribut

(fitur) set untuk data latih kelas ke-*i*. Untu yi € {-1, +1} menyatakan label kelas. Hyperplane klasifikasi Support Vector Machine, dinotasikan :

$w.x_i + b = 0$ (1)

w dan b adalah parameter model $w.x_i$ merupakan *inner product* antara *w* dan x_i.

Data x_i yang masuk kedalam kelas -1 adalah data yang memenuhi pertidaksamaan berikut :

$w.x_i + b \leq -1$ (2)

Sementara data x_i yang masuk kedalam kelas +1 adalah data yang memenuhi pertidaksamaan berikut :

$w.x_i + b \geq +1$ (3)

Jika data dalam kelas -1 (misal x_a) bertempat di hyperplane, maka untuk data kelas -1 dinotasikan :

$w.x_a + b = 0$ (4)

Sementara kelas +1 (misal x_b) akan memenuhi persamaan :

$w.x_b + b = 0$ (5)

Dengan mengurangkan persamaan (5) dengan (4), didapatkan :

x_a-x_b adalah vektor paralel di posisi hyperplane dan diarahkan dari x_a ke x_b.

Dengan memberikan label -1 untuk kelas pertama dan kelas +1 kelas kedua, maka untuk prediksi semua data uji menggunakan formula :

$$y = \begin{cases} +1, jika\ w.z + b > 0 \\ -1, jika\ w.z + b < 0 \end{cases} \quad (6)$$

Hyperlane untuk kelas -1 adalah data support vector yang memenuhi persamaan :

$w.x_a + b = -1$ (7)

Sementara kelas +1 adalah data support vector yang memenuhi persamaan :

$w.x_b + b = +1$ (8)

Dengan demikian maka margin dapat dihitung dengan mengurangkan persamaan (8) dengan (7) didapatkan :

$w.(x_b - x_a) = 2$ (10)

Margin hyperplane diberikan oleh jarak antara dua hyperplane dari dua kelas tersebut. Notasi diatas diringkas menjadi :

$$\|w\| x\ d = 2\ atau\ d = \frac{2}{\|d\|} \quad (11)$$

1.1.2. Support Vector Machine Non Linear

Dalam masalah klasifikasi kebanyakan sampel data tidak terpisah secara linier sehingga jika digunakan Support Vector Machine linier maka hasil yang diperoleh tidak optimal dan mengakibatkan hasil klasifikasi yang buruk. Support Vector Machine linier dapat diubah menjadi Support Vector Machine non-linier dengan menggunakan metode kernel. Metode ini bekerja dengan cara memetakan data input ke ruang *feature* yang dimensinya lebih tinggi. Diharapkan data input hasil pemetaan ke ruang *feature* akan terpisah secara linier sehingga dapat dicari *hyperplane* yang optimal. Pendekatan ini berbeda dengan metode klasifikasi pada umumnya yang justru mengurangi dimensi awal untuk menyederhanakan proses komputasi dan memerikan akurasi prediksi yang lebih baik[5].

Fungsi kernel yang digunakan dalam literatur suport vector machine [7].

1. Linear $K = (x, y) = x.y$
2. Polynomial $K(x, y) = (x.y + c)^d$
3. Gaussian RBF $K(x, y) = exp\left(\frac{-\|x-y\|^2}{2.\sigma^2}\right)$
4. Sigmon (tangen hiperbolik) $K\ (x, y) = \tanh(\sigma(x.y) + c)$
5. Invers multiquadric $K\ (x, y) = \frac{1}{\sqrt{\|x-y\|^2 + c^2}}$

x,y adalah pasangan dua data dari semua bagain data latih. Parameter σ, *c, d,* > 0, merupakan konstanta $\|x - y\|^2$ merupakan kuadrat jarak antara vektor x dan y. Fungsi kernel mana yang harus digunakan untuk substitusi *dot product* dari fitur dimensi lama ke dimensi baru sangat bergantung pada kondisi data [5]. Biasanya metode validasi silang (Hstie *et al,* 2001) digunakan untuk pemilihan fungsi kernel. Pemilihan fungsi kernel yang tepat merupakan hal yang sangat penting karena fungsi kernel akan menentukan fitur baru (dimesi tinggi) dimana fungsi klasifikasi (*hyperplane)* akan dicari.

Sepanjang fungsi kernelnya logis, Support Vector Machine akan bekerja secara benar meskipun tidak diketahui seperti apa pemetaan yang dilakukan. Fungsi kernel yang logis diberikan oleh teori Mercer (Vapnik, 1995 dan Haykin, 1999), dimana fungsi kernel harus memenuhi syarat kontinyu dan positif. Lebih mudah menemukan fungsi kernel daripada mencari peta Φ seperti apa yang tepat untuk melakukan pemetaan fitur lama ke fitur baru. Dengan kata lain, pada penerapan metode kernel tidak perlu diketahui pemetaan seperti apa yang digunakan untuk satu persatu data, tetapi lebih penting mengetahui bahwa *dot-product* dua data di fitur baru bisa digantikan oleh fungsi kernel [7].

1.2. K-Fold Cross Validation

Cross validation adalah salah satu cara menemukan parameter terbaik dari suatu model dengan cara menguji besarnya error pada data test. Dalam cross validation data dibagi ke dalam k sampel dengan ukuran yang sama. k-1 sampel digunakan untuk training dan 1 sampel sisanya untuk testing. Ini sering di sebut dengan validasi *k–fold* [6]. Metode cross validation membagi data menjadi dua yaitu data training dan data testing. Setelah pengujian dilakukan proses silang dimana data pengujian dijadikan data pelatihan dan sebaliknya data pelatihan sebelumnya kini menjadi data pengujian.

Ciri-ciri *k-fold* cross validation sebagai berikut :

1. Mempartisi data secara random kedalam k buah himpunan atau fold yaitu D1, D2, ..., Dk. Setiap kelompok mempunyai jumlah yang hampir sama.
2. Pada perulangan pertama, digunakan sebagai data uji dan himpunan lainnya sebagai pelatih.
3. Melakukan training dan pengujian sebanyak k kali
4. Menghitung keakuratan dengan rumus

1.2 Matlab (MATrix Laboratory)

Matlab singkatan dari MATrix Laboratory, merupakan bahasa pemrograman yang dikembangkan oleh the Mathwork .Inc (http://www.mathworks.com). Bahasa pemrograman ini banyak digunakan untuk perhitungan numeric teknikal, komputasi, simbolik, visualisasi, grafis, analisis dan matematika, statistika, simulasi pemodelan dan desain GUI [8]. Guide atau GUI builder merupakan sebuah graphical user interface (GUI) yang dibangun dengan obyek grafik seperti tombol (button), kotak teks, slider, menu dan lain-lain. Aplikasi yang menggunakan GUI umumnya lebih mudah dipelajari dan digunakan menggunakan GUI umumnya lebih mudah dipelajari dan digunakan karena orang yang menjalankannya tidak perlu mengetahui perintah yang ada dan bagaimana kerjanya.

Karakteristik Matlab sebagai berikut [10] :

1. Bahasa pemrogramannya didasarkan pada matriks yaitu baris dan kolom
2. Lambat (dibandingkan dengan Fortran atau C) karena bahasanya langsung diartikan.
3. *Automatic memory management,* misalnya kita tidak harus mendeklarasikan arrays terlebih dahulu.
4. Tersusun rapi
5. Memiliki waktu pengembangan program yang lebih cepat dibandingkan bahasa pemrograman tradisional seperti fortran atau C.
6. Dapat diubah ke bahasa C lewat matlab compiler untk efisiensi yang lebih baik.
7. Tersedia banyak toolbox untuk aplikasi-aplikasi khusus.
8. Bersama dengan Maple untuk komputasi-komputasi simbolik.
9. Dalam *share memory parallel-computers,* seperti SIG Origin2000, beberapa opersi secara otomatis dapat diproses bersama.

1.3. UML (*Unified Modeling Language)*

UML *(Unified Modeling Language)* adalah notasi yang lengkap untuk membuat visualisasi model suatu sistem. Sistem berisi informasi dan fungsi, tetapi secara normal digunakan untuk memodelkan sistem komputer. Di dalam pemodelan objek guna menyajikan sistem yang berorientasi objek kepada orang lain, akan sangat sulit di lakukan dalam bentuk kode bahasa pemrograman [11].

UML disebut sebagai bahasa pemodelan bukan metode. Bahasa pemodelan (sebagian besar grafik) merupakan notasi model yang digunakan untuk mendesain secara cepat. Bahasa pemodelan merupakan bagian terpenting dari metode. UML merupakan bahasa standar untuk penulisan *blueprint software* yang digunakan untuk visualisasi, spesifikasi, pembentukan dan pendokumentasian alat-alat dari sistem perangkat lunak. UML biasanya disajikan dalam diagram atau gambar yang meliputi *class* beserta atribut dan operasinya, serta hubungan antar kelas. UML terdiri dari banyak diagram diantaranya *use case diagram, activity diagram*, *class diagram* dan *sequence diagram.*

2. METODE PENELITIAN

Penelitian ini menggunakan metode penelitian tindakan (*action research)* yaitu metode yang bertujuan untuk mengembangkan keterangan baru untuk mengatasi dalam dunia kerja atau kebutuhan praktis manusia lainnya. Untuk menemukan dasar-dasar dan langkah-langkah yang tepat untuk melakukan tindakan perbaikan secara praktis [12]. Action research dilakukan dalam lima tahapan yaitu *diagnosing*, *action planning, action taking, evaluation* dan *reflection*.

2.1. Pengumpulan Data

Metode pengumpulan data yang dilakukan pada penelitian ini yaitu :

1. Wawancara
   Wawancara di lakukan secara langsung terhadap pihak Fakultas Kesehatan Masyarakat Unhas untuk mendapatkan informasi dan data-data yang dibutuhkan dalam memprediksi mahasiswa drop out.
2. Observasi
   Mengadakan pengamatan dan pengumpulan data history berupa data internal dan external yang telah tersusun dalam database maupun kelengkapan data yang di butuhkan.

2.2. Analisis Data

Metode analisis data pada penelitian ini menggunakan algoritma Suport Vector Machine dengan melakukan proses *preprocessing* dengan mengubah representasi data, normalisasi data, dan transformasi serta proses pembelajaran Support Vector Machine dengan melakukan proses pelatihan dan pengujian data untuk mencari *hyperplane* dengan margin terbesar. Data yang digunakan untuk penelitian ini adalah data Mahasiswa dan data Evaluasi IP dan IPK mahasiswa fakultas kesehatan masyarakat Unhas tahun 2007 – 2011 yang bersumber dari BAAK. Data terdiri dari dua program studi yaitu Kesehatan Masyarakat (Kesmas) dan Ilmu Gizi.

a. Preprocessing data
   Preprocessing yaitu untuk data yang bersifat noise dan missing value dan normalisasi. Dengan pembersihan dan pengintegrasian, data noise dan informasi yang tidak relevan dari dataset akan berkurang. Didalam preprocessing dilakukan cleaning data yaitu proses yang digunakan untuk menghapus data ganda, memeriksa data yang tidak konsisten, penanganan data missing dan merapikan data noise, dan normalisasi data. Pada tahap ini *missing value* akan dihapus, jumlah data yang memiliki *missing value* sangat kecil dibandingkan jumlah keseluruhan data. Proses penghapusan *missing value* dilakukan secara manual. Dengan cara mengecek satu persatu data. Ketik ditemukan data yang memiliki *missing value* maka akan dihapus, sampai *missing value* tidak ditemukan lagi. Tujuan penerapan metode normalisasi adalah untuk membuat rentang nilai atribut kedalam skala tertentu. Hal tersebut bertujuan agar proses perhitungan lebih mudah dan atribut yang rentang nilainya besar tidak mendominasi fitur yang rentang nilainya kecil.

b. Transformasi data dan seleksi data.
   Transformasi data digunakan untuk mengubah dataset sehingga konten informasi terbaik diambil dan dengan melakukan pengurangan atau pengubahan tipe data standar sehingga data siap digunakan untuk di presentasikan ke teknik data mining. Seleksi data bertujuan untuk mengidentifikasi variabel-variabel relevan yang digunakan dalam penelitian.
c. Seleksi Variabel
   Proses seleksi variabel data set menggunakan metode Principal Component Analisis (PCA). Seluruh variabel *dataset* akan diseleksi untuk mendapatkan variabel-variabel yang relevan. hasil seleksi variabel untuk program studi Kesehatan masyarakat tahun 2007 dari 174 record data terseleksi 15 variabel. Tahun 2008 dari 369 record data terselesksi 17 variabel. Tahun 2009 dari 229 record terseleksi 17 variabel. Tahun 2010 dari 246 record terseleksi 15 variabel. Tahun 2011 dari 237 record terseleksi 12 variabel. Pada program studi gizi tahun 2007 dari 63 record data terseleksi 11 variabel. Tahun 2008 dari 110 record data terselesksi 13 variabel. Tahun 2009 dari 85 record terseleksi 11 variabel. Tahun 2010 dari 93 record terseleksi 10 variabel. Tahun 2011 dari 37 record terseleksi 8 variabel.
d. Pelatihan dan Pengujian
   Pada tahap ini data dilatih sebagai proses pembelajaran untuk model pembelajaran Support Vector machine untuk data uji. Tahap pengujian merupakan tahapan yang digunakan untuk menguji data data penelitian. Pengujian berfungsi untuk menguji keakuratan metode Support Vector Machine.

2.3 Proses Prediksi

Metode yan digunakan untuk memprediksi mahasiswa drop out adalah Support Vector Machine (SUPPORT VECTOR MACHINE). Variabel input yang digunakan dalam penelitian ini yaitu data mahasiswa dan data evaluasi IP dan IPK. Jumlah kategori yang akan di prediksi memiliki 2 (dua) kategory output adalah mahasiswa drop out dan non drop out. Tahapan algoritma Support Vector Machine sebagai berikut :

a. Melakukan proses transformasi data sesuai dengan format Support Vector Machine
b. Menentukan fungsi kernel yang akan digunakan
   Penelitian ini menggunakan tipe kernel polynomial dan Radial basis function (RBF)
c. Menentukan nilai-nilai parameter kernel dan parameter cost (C) untuk melakukan optimasi
d. Memilih parameter tebaik untuk optimasi data training dan untuk prediksi data testing
e. Menghitung ketepatan prediksi

3. HASIL DAN PEMBAHASAN

Pengujian Support Vector Machine untuk memprediksi mahasiswa drop out dilakukan dengan beberapa tahapan yaitu proses pembagian data, optimasi parameter dan pembelajaran Support Vector Machine.

a. Pembagian data
   Data set mahasiswa dibagi menjadi dua bagian yaitu data training dan data testing. Kedua data tersebut dipilih secara bebas menggunakan teknik stratified random. Data training harus berbeda dengan data testing untuk memperoleh performa yang baik dalam tahap optimisasi.
b. Optimasi parameter Support Vector Machine
   Beberapa algoritma klasifikasi terkadang membutuhkan pengaturan pada parameter tertentu. Support Vector Machine merupakan salah satu algoritma yang membutuhkan pengaturan fungsi kernel. Kernel Support Vector Machine yang digunakan pada penelitian ini adalah kernel Polynomial dan RBF. Kernel Polynomial memiliki satu parameter degree (p), degree bernilaa 1, 2, 3, 4, ..., n. Dan kernel RBF menghendaki pengaturan parameter C dan γ, bernilai lebih dari 0. Degree (*P*) yang digunakan adalah 1, 2, 3,..., 10 Nilai Cost (C) 0.1, 0.2, .., 0.5 dan Nilai Gamma (γ) 0.0001, 0.001, 0.1, 0.2 ..., 0,8. Oleh karena itu, pada tahap optimasi parameter Support Vector Machine dilakukan metode 10-flod cross validation

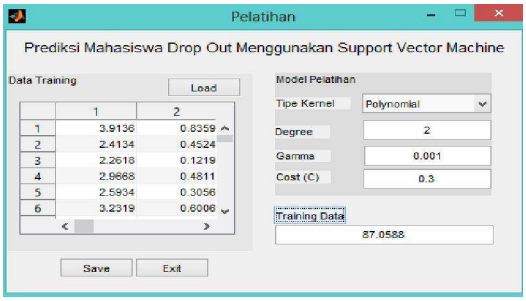
c. Pembelajaran Support Vector Machine

Pada tahap ini parameter optimum p, C dan γ digunakan pada pembelajaran Support Vector Machine pada data latih untuk menghasilkan predictor model dan kemudian diujikan pada data uji. Langkah ini diterapkan pada setaip subset fitur (atribut) hasil seleksi metode seleksi variabel. Prediksi Support Vector Machine dilakukan dengan bantuan perangkat lunak matlab. Metode evaluasi performa yang digunakan adalah 10-flod cross validation. *K-fold* cross validation dipilih karena lebih akurat dalam mengestimasi performa. Cross-validation adalah metode evaluasi yang membagi data kedalam dua segmen, satu digunakan untuk pembelajaran atau training model dan lainnya digunakan sebagai model validasi. Data yang diambil secara stratified random kemudian dibagi K buah partisi dengan ukuran yang sama, dan selanjutnya akan dilakukan iterasi sebanyak K. Pada setiap iterasi digunakan satu partisi untuk pengujian dan k-1 untuk pelatihan.

Implementasi antarmuka yang menggambarkan tampilan prototype yang dibangun yaitu implementasi antarmuka prediksi mahasiswa drop out. Berikut ini adalah implementasi antarmuka dari prototype yang dibangun:

1. Menu data Training

Menu data training digunakan untuk melakukan proses training. Proses dimulai dengan menampilkan data training, lalu menentukan nilai parameter Support Vector Machine berupa nilai degree, gamma dan cost, kemudian diakhiri dengan menjalankan proses training. Tampilan input data training terdapat pada gambar 1 di bawah ini.



Gambar 1 Menu Data Training

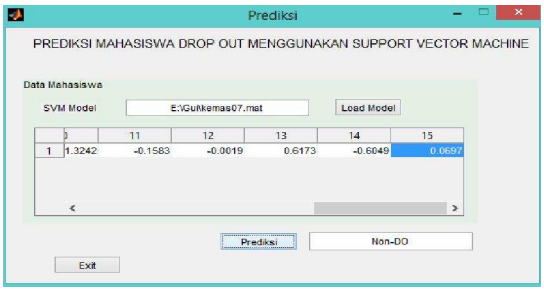
2. Menu Data Testing

Menu data testing digunakan untuk melakukan proses pengujian data. Proses dimulai dengan menginput data testing dari parameter Support Vector Machine Yaitu dengan cara *load* data yang sudah di simpan ketika proses *training* dan diakhiri dengan menjalankan poses testing dan akan menampilkan hasil dari testing Tampilan menu data testing pada gambar 2 di bawah ini :

Gambar 2 Menu Data Testing

3. Menu Data Prediksi

Input data prediksi dilakukan oleh user. Data yang digunakan sebagai masukan adalah data baru yang sebelumnya belum pernah digunakan. Pada tahap ini sistem menggunakan data akhir hasil proses pelatihan sebelumnya. Hasil atau keluaran yang ditampilkan berupa prediksi mahasiswa drop out dan non drop out. Tampilan input data prediksi pada gambar 3 di bawah ini :



Gambar 3. Menu Prediksi Drop Out

Parameter Support Vector Machine (SVM)

Penelitian ini menggunakan tipe kernel polynomial dan Radial basis function (RBF). Parameter nilai degree (P) 1,2,3, ..., 10, nilai cost (C) berdasarkan 0.1, 0.2, ..., 0.3 nilai gamma 0.0001, 0.001, 0.1, ..., 0.8, dan *k-fold* (number of validation) k=10. Berikut hasil percobaan pada tabel 1da tabel 2 yang telah dilakukan dengan beberapa fungsi kernel dan memasukan nilai cost (C) serta nilai range (*k-fold)* yang telah ditentukan untuk menguji masing-masing program studi.

1. Program Studi Kesehatan Masyarakat

Tabel 1. Pengujian Prediksi Support Vector Machine 2007-2011

k-fold = 10					
Tahun	Tipe kernel	Parameter			Akurasi
		C	P	γ	
2007	2	0.2	5	0.001	95.23%
2008	1	0.1	3	0.001	87.27%
2009	1	0.1	5	0.1	90.58 %
2010	2	0.7	6	0.1	90.23%
2011	1	0.1	3	0.001	89.18%

Data yang diuji dengan memasukan nilai model testing Support Vector Machine untuk tahun 2007-2011. Tahun 2007 kernel Radial basic funtion (RBF) dengan nilai cost (C) 0.2, degree (P) 2 dan gamma (γ) 0.001 menghasilkan akurasi 95.23 %. Tahun 2008 kernel polynomial dengan nilai cost (C) 0.1, degree (P) 3 dan gamma (γ) 0.001 menghasilkan akurasi 87.27%. Tahun 2009 kernel polynomial cost (C) 0.1, degree (P) 5 dan gamma (γ) 0.1 menghasilkan akurasi 90.58%, tahun 2010 kernel Radian basic function (RBF) dengan nilai cost (C) 0.7, degree (P) 6 dan gamma (γ) 0.1 menghasilkan akurasi 90.23%, dan tahun 2011 kernel polynomial dengan nilai cost (C) 0.1, degree (P) 3 dan gamma (γ) 0.001 dan akurasi 89.18%.

2. Program Studi Gizi

Tabel 2. Pengujian Prediksi Support Vector Machine 2007-2011

k-fold = 10					
Tahun	Tipe kernel	Parameter			Akurasi
		C	P	γ	
2007	2	0.4	7	0.5	89.08%
2008	2	0.1	6	0.1	86.17%
2009	2	0.2	7	0.001	98.25 %
2010	2	0.4	7	0.5	88.21%
2011	1	0.1	3	0.001	91.57 %

Data yang diuji dengan memasukan nilai model testing Support Vector Machine untuk tahun 2007 -2011. Tahun 2007 kernel Radial basic function (RBF) dengan cost (C) 0.4, degree (P) 7 dan gamma (γ) 0.5 menghasilkan akurasi 89.08 %. Tahun 2008 kernel Radial basic function (RBF) dengan nilai cost (C) 0.1, degree (P) 6 dan gamma (γ) 0.1 menghasilkan akurasi 86.17%. Tahun 2009 kernel RBF nilai cost (C) 0.2, degree (P) 7 dan gamma (γ) 0.001 menghasilkan akurasi 98.25%, tahun 2010 kernel Radial basic function (RBF) dengan nilai cost (C) 0.4, degree (P) 7 dan gamma (γ) 0.5 menghasilkan akurasi 88.21%, dan tahun 2011 kernel polynomial dengan nilai cost (C) 0.1, degree (P) 3 dan gamma (γ) 0.001 menghasikan akurasi 91.57%.

Analisis hasil pengujian Support Vector Machine

Untuk mengevaluasi performa model Suppor Vector Machine (SVM), perlu dilakukan pengukuran performa, yaitu pengukuran tingkat akurasi (*accuracy)* dan tingkat kesalahan (*error rate)*. Pengukuran performa Support Vector Machine pada penelitian ini menggunakan data mahasiswa tahun 2011 dengan 310 recor data, dan menggunakan tipe kernel polynomial, *k-fold =10* nilai cost 0,5 gamma 0,001 dan degree (P) 8.

Untuk pengukuran akurasi dapat menggunakan persamaan :

$$Accuraccy = \frac{Jumlah\ prediksi\ yang\ benar}{Jumlah\ prediksi\ keseluruan}$$

$$Accuraccy = \frac{304}{310} = 98,06$$

Sedangkan pengukuran tingkat kesalahan (*error rate*) menggunakan persamaan :

$$Error\ = \frac{Jumlah\ prediksi\ yang\ \ salah}{Jumlah\ prediksi\ keseluruan}$$

$$Error\ = \frac{6}{310} = 0.0193$$

Tabel 3. Hasil Pengujian Support Vector Machine

Prediksi Menggunakan Support Vector Machine		
Output	KS	HU
Non Drop Out	283	290
Drop Out	27	20
Total Data	310	310

Keterangan :

KS :Kondisi Sebenarnya

HU : Hasil Uji

KSVM :Ketepatan Support Vector Machine

Berdasarkan hasil pengujian performa pada tabel 3.3 dengan menggunakan data uji tahun 2011 sebanyak 310 set data dengan mahasiswa dengan mahasiswa non drop out 283 orang (KS) dan drop out 27 (KS) orang menghasilkan prediksi mahasiswa drop out 20 (HU) orang dan 290 orang (HU) non drop out. Diperoleh jumlah prediksi yang benar (sama) sebanyak 304 set data dan jumlah prediksi yang salah (tidak sama) sebanyak 6 set data. Untuk pengukuran tingkat akurasi diperoleh akurasi sebesar 98,06% dan nilai error sebesar 0.0193.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan dapat disimpulkan bahwa :

a. Algoritma Support Vector Machine (SVM) dapat digunakan untuk melakukan prediksi mahasiswa drop out dengan menggunakan variabel input data individu dan evaluasi IP dan IPK mahasiswa dengan variabel output mahasiswa drop out dan non drop out. Faktor-faktor yang mempengaruhi penyebab mahasiswa drop out yaitu faktor non akademik dan akademik.

b. Berdasarkan hasil pengujian data mahasiswa program studi kesehatan masyarakat parameter terbaik algoritma Support Vector Machine dengan nilai *k-fold* =10, tipe kernel polynomial, diperoleh nilai *error rate* = 0.094, nilai cost (C) =0.1, degree (P) =5, dan gamma (γ)= 0.1 dan tipe kernel RBF diperoleh nilai *error rate* = 0,047, nilai cost (C) =0.2, degree (P) =5, dan gamma (γ) =0.001. Untuk data program studi gizi dengan nilai *k-fold* =10, tipe kernel polynomial, diperoleh nilai *error rate* =0.084, nilai cost (C) =0.1, degree (P) =3, dan gamma (γ) =0.001 dan tipe kernel RBF diperoleh nilai *error rate* =0.017, nilai cost (C) =0.2, degree (P) =7, dan gamma (γ) = 0.001.

5. SARAN

Saran untuk penelitian selanjutnya perlu dilakukan :

a. Analisis lebih lanjut terhadap faktor-faktor yang mempengaruhi drop out dan menggunakan data kemahasiswaan yang lebih lengkap.
b. Pemilihan parameter Support Vector Machine yang optimal dapat menggunakan metode Particle Swarm Optimization (PSO).
c. Untuk fungsi kernel Radial Basis Funciton (RBF) pencarian nilai pasangan cost dan gamma (C,γ) terbaik dapat menggunakan metode *grid search*.
d. Mengevaluasi atribut-atribut mana yang nantinya akan sangat berpengaruh pada nilai keluaran.
e. Menambahkan beberapa algoritma klasifikasi data mining untuk di komparasi dengan metode Support Vector Machine.

DAFTAR PUSTAKA

[1] Dekker, G.W., 2009, *Prediction student drop out: A case study,* USA, Academic Press, 2nd international Conference On Educational Data Mining, Cordoba, Spain.

[2] Hastuti, K., 2012, *Analisis Komparasi Algoritma Klasifikasi Data Mining untuk Prediksi Mahasiswa Non Aktif*, Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2012, ISBN 979-26-0255-0, 23 Juni 2012.

[3] Hidayat M. M., Purwitasari D., dan Ginardi H., 2013*, Analisis Prdiksi DO Mahasiswa dalam Educational Data Mining menggunakan JaringanSyaraf Tiruan*, Jurnal IPTEK Vol 17 No.2 Desember 2013.

[4] Sumarni, A., 2014, *Kalsifikasi Data Nap (Nota Analisis Pembiayaan) untuk Prediksi Tingkat Keamanan Pemberian Kredikt* (Studi Kasus : Bank Syariah Mandiri Cabang Luwuk Sulawesi Tengah), Tesis, Ilmu Komputer, Universitas Gajah Mada, Yogyakarta.

[5] Nugroho, 2008, *SUPPORT VECTOR MACHINE: Paradigma Baru dalam softcomputing dan Aplikasinya,* Konferensi Nasional Sistem & Informatika, Bali, 2008

[6] Santoso, B., 2007, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Graha Ilmu, Yogyakarta.

[7] Prasetyo, E.,2014, *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*, Andi, Yogyakarta..

[8] Prasetyo, T. W. (2002). *Analisis dan Desain Sistem Kontrol dengan MATLAB* . Andi, Yogyakarta.

[9] Sugiarti, Y, 2013, *Analisis & Perancangan UML (Unified Modeling Language) Generated VB.6*, Graha Ilmu, Yogyakarta.

[10] Prasetyo, T. W. (2002). Analisis dan Desain Sistem Kontrol dengan MATLAB . Andi, Yogyakarta.

[11] Yasin, V., 2012, Rekayasa Perangkat Lunak Berorientasi Objek Pemodelan, Arsitektur, dan Perancangan (Modeling, Architecture and Desaign), Mitra Wacana Media.

[12] Darmawan, D, 2013, *Metode Penelitian Kuantitatif*, Remaja Rosdakarya, Bandung.