Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 9(1) 2016: 15-29
http://ejournal.undip.ac.id/index.php/media_statistika

ANALISIS KLASIFIKASI MASA STUDI MAHASISWA PRODI STATISTIKA UNDIP DENGAN METODE *SUPPORT VECTOR MACHINE* (SVM) DAN ID3 *(ITERATIVE DICHOTOMISER 3)*

Dwi Ispriyanti[1], Abdul Hoyyi[2]
[1,2]Staf Pengajar Departemen Statistika FSM UNDIP

e-mail: dwiispriyanti@yahoo.com

DOI: 10.14710/medstat.9.1.15-29

Abstract

Graduation is the final stage of learning process activities in college. Undergraduate study period in UNDIP's academic regulations is scheduled in 8 semesters (4 years) or less and maximum of 14 semesters (7 years). Department of Statistics is one of six departments in the Faculty of Science and Mathematics UNDIP. Study period in this department can be influenced by many factors. Those factor are Grade Point Average (GPA) or IPK, gender, scholarship, parttime, organizations, and university entrance pathways. The aim of this paper is to determine the accuracy factors classification. We use SVM (Support Vector Machine method) and ID3 (Iterative Dichotomiser 3). The comparison of SVM and ID3 method, both for training and testing the data generate good accuracy, namely 90%. Especially ID3 training data gives better result than SVM.

Keywords: *SVM*, *ID3*

1. PENDAHULUAN

Pendidikan Tinggi merupakan jenjang pendidikan setelah pendidikan menengah yang mencakup program pendidikan diploma, sarjana, magister, spesialis dan doktor yang diselengggarkan oleh PerguruanTinggi. Kelulusan adalah hasil akhir dari proses kegiatan belajar mengajar selama mengikuti perkuliahan di perguruan tinggi. Universitas Diponegoro (UNDIP) adalah salah satu Universitas Negeri yang berada di Jawa Tengah memiliki 11 (sebelas) fakultas. Salah satu diantaranya adalah Fakultas MIPA, yang sekarang berganti nama menjadi Fakultas Sains dan Matematika (FSM). FSM terdiri dari 6 jurusan dengan 7 program S1 yaitu Matematika, Biologi, Kimia, Fisika, Statistika, dan Teknik Informatika serta D3 Instrumentasi dan Elektronika.

Prodi Statistika berdiri sejak tahun 2013 melalui surat Direktorat Jendral Pendidikan Tinggi No. 920/D/T/2003. Lama Studi Program Sarjana menurut Peraturan Akademik UNDIP dijadwalkan dalam 8 semester (4 tahun) atau dapat ditempuh kurang dari 8 semester (4 tahun) dan selambat lambatnya 14 semester (7 tahun). Setiap tahun UNDIP menyelenggarakan upacara wisuda dalam 4 periode yaitu periode Januari, April, Juli, dan Oktober. Dalam 4 periode kelulusan jumlah lulusan dengan jumlah mahasiswa baru tidak sebanding. Sehingga menimbulkan jumlah mahasiswa meningkat yang mengakibatkan rasio dosen dan mahasiswa setiap tahun akan bertambah. Lama masa studi mahasiswa kemungkinan dapat dipengaruhi oleh banyak faktor. Faktor-faktor yang diperkirakan mempengaruhi dalam kelulusan tepat waktu antara lain Indeks Prestasi Kelulusan (IPK), jenis kelamin, beasiswa, *part time*, organisasi, dan jalur masuk universitas.

Dari analisis deskritip data, peneliti ingin mengetahui faktor-faktor yang mempengaruhi lama masa studi mahasiswa prodi Statistika dan mengklasifikasikan kelulusan mahasiswa ke dalam dua kategori yaitu lulus tepat waktu untuk mahasiswa yang menempuh pendidikan S1 kurang dari sama dengan 4 tahun (8 semester) dan lulus tidak tepat waktu untuk mahasiswa yang menempuh pendidikan lebih dari 4 tahun (8 semester). Metode Statistika yang digunakan dalam penelitian ini adalah metode *Support Vektor Machine* (SVM) dan akan dibandingkan dengan metode dan ID3 (*Iterative Dichotomiser 3*). Sehingga tujuan dari penelitian ini adalah:

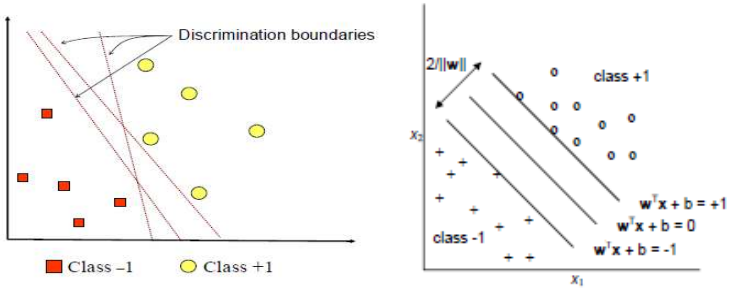
1. Memperoleh gambaran secara deskripsi faktor-faktor yang mempengaruhi masa studi mahasiswa prodi Statistika
2. Memperoleh model dan ketepatan klasifikasi lama studi mahasiswa kelulusan mahasiswa prodi Statistika dengan menggunakan metode *Support Vector Machine* (SVM).
3. Memperoleh model dan ketepatan klasifikasi lama studi mahasiswa kelulusan mahasiswa prodi Statistika dengan menggunakan metode algoritma *Iterative Dichotomiser 3* (ID3).
4. Membandingan ketepatan klasifikasi metode SVM dan ID3 pada klasifikasi lama studi mahasiswa prodi Statistika UNDIP.

2. TINJAUAN PUSTAKA

2.1. *Support Vector Machine* (SVM)

*Support Vector Machine (*SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 di *Annual Workshop on Computational Learning Theory*. Metode ini merupakan metode mesin pembelajaran (*learning machine*) dengan tujuan menemukan fungsi pemisah (*hyperplane*) terbaik yang memisahkan dua buah kelas pada *input space* (Nugroho dkk, 2003).

Konsep SVM suatu konsep untuk mencari fungsi pemisah terbaik yang berfungsi sebagai pemisah dua buah kelas pada *input space*. Pada Gambar 1. (a) memperlihatkan beberapa data yang merupakan anggota dari dua buah kelas +1 dan -1. Data yang tergabung pada kelas -1 disimbolkan dengan warna merah (kotak), sedangkan Data yang +1 disimbolkan dengan warna kuning (lingkaran). Fungsi pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur margin fungsi pemisah. Margin adalah jarak antara fungsi pemisah tersebut dengan data terdekat dari masing-masing kelas. Data yang paling dekat ini disebut sebagai *support vector* (Prasetyo, 2012). Usaha untuk mencari lokasi fungsi pemisah ini merupakan inti dari proses pembelajaran pada SVM (Cristanini dan Jhon, 2000).



(a) (b)

Gambar 1. (a) *Discrimination Boundaries* dan (b) Konsep Fungsi Pemisah

2.1.1. Klasifikasi Linier *Separable*

Misalkan diberikan himpunan $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$, dengan $\boldsymbol{x}_i \in \Re^p$, diketahui $\boldsymbol{X}$ memiliki pola tertentu, yaitu apabila $\boldsymbol{x}_i$ termasuk dalam suatu kelas maka diberi label y_i = +1, jika tidak termasuk diberi label y_i = -1 untuk itu label masing-masing dinotasikan $y_i \in \{-1,+1\}$ sehingga data berupa pasangan $(\boldsymbol{x}_1, y_1)$, $(\boldsymbol{x}_2, y_2)$,… $(\boldsymbol{x}_i, y_i)$ dimana i = 1, 2,….,n yang mana n adalah banyak data. Diasumsikan kedua kelas -1 dan +1 dapat terpisah secara sempurna oleh fungsi pemisah berdimensi p, yang didefinisikan $\boldsymbol{w}^T\boldsymbol{x} + b = 0$, $\boldsymbol{w}$ dan b adalah parameter model. Untuk mendapatkan fungsi pemisah terbaik adalah dengan mencari fungsi pemisah yang terletak ditengah-tengah antara dua bidang pembatas dengan memaksimalkan margin atau jarak antara dua set objek dari kelas yang berbeda (Santosa, 2007) .

Menurut Gunn (1998) fungsi pemisah optimal dihitung dengan memaksimalkan margin $\rho(\boldsymbol{w},b)$ untuk jarak x ke fungsi pemisah $(\boldsymbol{w},b)$ adalah:

$$d(\boldsymbol{w},b;\boldsymbol{x}) = \frac{\left|\boldsymbol{w}^T\boldsymbol{x} + b\right|}{\|\boldsymbol{w}\|}$$

$$\rho(\boldsymbol{w},b) = min_{\{\boldsymbol{x}_i;y_i=1\}}\, \boldsymbol{d}(\boldsymbol{w},b;\boldsymbol{x}_i) + min_{\{\boldsymbol{x}_i;y_i=-1\}}\, \boldsymbol{d}(\boldsymbol{w},b;\boldsymbol{x}_i) = \frac{2}{\|\boldsymbol{w}\|} \quad (1)$$

Optimalisasi dapat diselesaikan dengan fungsi *Lagrange Multiplier* (Prasetyo, 2012):

$$L(\boldsymbol{w},b,\alpha) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{n}\alpha_i\left(y_i\left[\boldsymbol{w}^T\boldsymbol{x}_i + b\right] - 1\right) \quad (2)$$

$$L(\boldsymbol{w},b,\alpha) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{n}\alpha_i\, y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i\right) - b\sum_{i=1}^{n}\alpha_i\, y_i + \sum_{i=1}^{n}\alpha_i \quad (3)$$

Dimana $\boldsymbol{w}^T\boldsymbol{w}$ dapat dijabarkan: $\boldsymbol{w}^T\boldsymbol{w} = \sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j\, \alpha_i\alpha_j\left(\boldsymbol{x}_i^T\boldsymbol{x}_j\right)$

maka, $$L_d = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\, y_i y_j\, \boldsymbol{x}_i^T\boldsymbol{x}_j \quad (4)$$

$$max_\alpha\, L_d = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\, y_i y_j\, \boldsymbol{x}_i^T\boldsymbol{x}_j \quad (5)$$

dengan batasan $\alpha_i \geq 0, i = 1, 2, ..., n$ dan $\sum_{i=1}^{n}\alpha_i\, y_i = 0$

Dari hasil perhitungan ini diperoleh α_i kebanyakan bernilai positif. Data yang berkorelasi dengan α_i yang positif disebut *support vector*. Penyelesaian persamaan (5) dapat digunakan untuk menentukan *Lagrange Multiplier* dan diperoleh fungsi pemisah terbaik. Setelah solusi permasalahan *quadratic programming* ditemukan (nilai α_i), maka kelas dari data yang akan diprediksi atau data testing dapat ditentukan berdasarkan fungsi sebagai berikut:

$$f\left(x_t\right) = \sum_{i=1}^{ns}\alpha_i\, y_i\, x_i\, .x_t + b \quad (6)$$

2.1.2. Klasifikasi Linear *Non-Separable*

Metode SVM juga dapat digunakan dalam kasus *non-separable* dengan memperluas formulasi yang terdapat pada kasus linier. Masalah optimasi sebelumnya baik pada fungsi

obyektif maupun kendala dimodifikasi dengan mengikutsertakan variabel *Slack* ξ > 0. Variabel *slack* merupakan sebuah ukuran kesalahan klasifikasi. Menurut Gunn (1998):

$$y_i\left[\left(\boldsymbol{w}^T\boldsymbol{x}_i\right)-b\right]\geq 1-\xi_i, i=1,2,...,n \tag{7}$$

$$\Phi\left(\boldsymbol{w},\xi\right)=\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+C\sum_{i=1}^{n}\xi_i \tag{8}$$

dimana parameter *C* berfungsi untuk mengontrol hubungan antara variabel *slack* dengan margin. Semakin besar nilai *C (cost)*, maka semakin besar pula pelanggaran yang dikenakan untuk tiap klasifikasi (Prasetyo, 2012). Menurut Kecman (2005) model optimasi (8) dapat diselesaikan dengan menggunakan fungsi *Lagrange*:

$$L(\boldsymbol{w},b,\alpha,\xi,\beta)=\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}-C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i\left(y_i\left[\left(\boldsymbol{w}^T\boldsymbol{x}_i\right)+b\right]-1+\xi_i\right)-\sum_{i=1}^{n}\beta_i\xi_i \tag{9}$$

dengan α, β adalah fungsi *Lagrange Multiplier*. Nilai optimal dari persamaan (9) dapat dihitung dengan meminimalkan terhadap $\boldsymbol{w}$, ξ, b dan memaksimalkan terhadap α, β. Menurut Kecman (2005), untuk menyederhanakan harus ditransformasi ke dalam fungsi *lagrange multiplier* itu sendiri (dualitas masalah), sehingga menjadi sebuah persamaan

$$max_\alpha\ L_d=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\ y_i y_j\ \boldsymbol{x}_i^T\boldsymbol{x}_j \tag{10}$$

dengan batas $0\leq\alpha_i\leq C, i=1,2,...,n$ dan $\sum_{i=1}^{n}\alpha_i\ y_i$.

2.1.3 Klasifikasi Non-Linear

Pada umumnya masalah dalam dunia nyata (*real world problem*) jarang yang bersifat liniear *separable* (tidak terpisahkan secara linear), tetapi bersifat non-linear (Nugroho dkk, 2003). Untuk menyelesaikan problem non-linear, SVM dimodifikasi dengan memasukkan fungsi kernel. Kernel dapat didefenisikan sebagai suatu fungsi yang memetakan fitur data dari dimensi awal (rendah) ke fitur yang lebih tinggi (bahkan jauh lebih tinggi). Pendekatan ini berbeda dengan metode klasifikasi pada umumnya yang justru mengurangi dimensi awal untuk menyederhanakan proses komputasi dan memberikan akurasi prediksi yang lebih baik (Prasetyo, 2012). Misalkan untuk *n* sampel data $\left(\left(\Phi(\boldsymbol{x}_1),y_1\right);\left(\Phi(\boldsymbol{x}_2),y_2\right);...;\left(\Phi(\boldsymbol{x}_n),y_n\right)\right)$, *dot product* dua buah vektor ($\boldsymbol{x}_i$) dan ($\boldsymbol{x}_j$) dinotasikan sebagai $\Phi\left(\boldsymbol{x}_i\right).\Phi\left(\boldsymbol{x}_t\right)$. Nilai *dot product* tersebut dapat dihitung tanpa mengetahui fungsi transformasi Φ dengan memakai komponen kedua buah vektor tersebut di ruang dimensi asal, seperti berikut:

$$K\left(\boldsymbol{x}_i,\boldsymbol{x}_t\right)=\Phi\left(\boldsymbol{x}_i\right).\Phi\left(\boldsymbol{x}_t\right)$$

Nilai $K\left(\boldsymbol{x}_i,\boldsymbol{x}_t\right)$ merupakan fungsi kernel yang menunjukkan pemetaan non-liner pada *feature space*. Prediksi himpunan data dengan dimensi fitur yang baru diformulasikan dengan:

$$f\left(\Phi(\boldsymbol{x})\right)=\text{sign}\left(\boldsymbol{w}.\Phi\left(\boldsymbol{x}_t\right)+b\right)=\text{sign}\left(\sum_{i=1}^{ns}\alpha_i\ y_i\ \Phi\left(\boldsymbol{x}_i\right).\Phi\left(\boldsymbol{x}_t\right)+b\right)$$

$$=\text{sign}\left(\sum_{i=1}^{ns}\alpha_i\ y_i\ \text{K}\left(\boldsymbol{x}_i,\boldsymbol{x}_t\right)+b\right)$$

dengan *ns* : jumlah data yang menjadi *support vector*

$\boldsymbol{x}_i$: *support vector*

$\boldsymbol{x}_t$: data testing yang akan diprediksi

Klasifikasi dengan Fungsi kernel ada beberapa nilai C (*cost*) yang digunakan untuk menentukan ketepatan klasifikasi, yaitu 0.1,1,10 dan 100 , dan parameter d (*degree*) yang masing-masing berbeda untuk tiap jenis Kernel.

Tabel 1. Fungsi Kernel yang dipakai dalam penelitian

Jenis Kernel	Definisi Fungsi
Linear	$K(\boldsymbol{x}_i, \boldsymbol{x}_t) = \boldsymbol{x}_i^T \boldsymbol{x}_t$
Polynomial	$K(\boldsymbol{x}_i, \boldsymbol{x}_t) = \left(\boldsymbol{x}_i^T \boldsymbol{x}_t + 1\right)^d$
Radial Basic Function (RBF)	$K(\boldsymbol{x}_i, \boldsymbol{x}_t) = \exp\left(-\frac{1}{2\sigma^2} \| \boldsymbol{x}_i - \boldsymbol{x}_t \|^2\right)$

2.2. Algoritma *Iterative Dichotomiser* 3 (ID3)

Iterative Dichotomiser 3 (ID3) adalah algoritma *decision tree learning* (algoritma pembelajaran pohon keputusan) yang paling dasar. Algoritma ini melakukan pencarian secara menyeluruh (*greedy*) pada semua kemungkinan pohon keputusan. Salah satu algoritma induksi pohon keputusan yaitu ID3 (*Iterative Dichotomiser 3*). ID3 dikembangkan oleh J. Ross Quinlan. Algoritma ID3 dapat diimplementasikan menggunakan fungsi *rekursif* (fungsi yang memanggil dirinya sendiri). Algoritma ID3 berusaha membangun *decision tree* (pohon keputusan) secara *top-down* (dari atas ke bawah), mulai dengan pertanyaan "atribut mana yang pertama kali harus dicek dan diletakkan pada *root*?" pertanyaan ini dijawab dengan mengevaluasi semua atribut yang ada dengan menggunakan suatu ukuran statistik (yang banyak digunakan adalah *information gain*) untuk mengukur efektivitas suatu atribut dalam mengklasifikasikan kumpulan sampel data (David, 2004).

Secara garis besar pohon keputusan terdiri atas tiga bagian, yaitu simpul akar, simpul keputusan atau simpul pengujian dan simpul daun. Menurut Han, *et.al.* (2011) simpul akar merupakan simpul teratas dalam suatu pohon keputusan, simpul keputusan merupakan simpul untuk menguji atribut dimana setiap cabang mewakili hasil pengujian pada simpul keputusan dan simpul daun merupakan simpul terakhir yang dilabeli kelas klasifikasi. Dalam konstruksi pohon keputusan data terbagi menjadi sampel pelatihan (*training sample*) dan sampel pengujian (*testing sample*). Sampel pelatihan digunakan untuk konstruksi pohon keputusan dan sampel pengujian digunakan untuk menguji ketepatan kelas.

2.2.1. *Entropy*

Entropy bisa dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas (Santosa, 2007). Di dalam bidang *Information Theory*, sering digunakan *entropy* sebagai suatu parameter untuk mengukur heterogenitas (keberagaman) dari suatu kumpulan sampel data. Jika kumpulan sampel data semakin heterogen, maka nilai *entropy*-nya semakin besar. Secara matematis, *entropy* dirumuskan sebagai berikut:

$$Entropy\ (S) = \sum_i^c - p_i log_2 p_i$$

dimana c adalah jumlah kelas klasifikasi, p_i adalah proporsi untuk kelas *i*.

2.2.2. Information Gain

Setelah mendapatkan nilai *entropy* untuk suatu kumpulan sampel data, maka dapat diukur efektivitas suatu atribut dalam mengklasifikasikan data. Ukuran efektivitas ini disebut juga sebagai *information gain*. Secara matematis, information gain dari suatu atribut A, dituliskan sebagai berikut:

$$Gain\,(S,A) = Entropy\,(S) - \sum_{v \in Values(A)} \frac{S_v}{S} Entropy\,(S_v)$$

$$Entropy\,(S) = \sum_i^c - p_i log_2 p_i$$

Entropy (S_v) adalah *entropy* untuk sampel-sampel yang memiliki nilai *v*

dimana *c* adalah jumlah kelas klasifikasi, p_i adalah proporsi untuk kelas *i*
A adalah Atribut
v menyatakan suatu nilai yang mungkin untuk atribut A
Values (A) adalah himpunan nilai-nilai yang mungkin untuk atribut A
S_v adalah ukuran sampel untuk nilai *v*
S adalah ukuran sampel
Entropy (S_v) adalah *entropy* untuk sampel-sampel yang memiliki nilai *v*

2.3 Ketepatan Pohon Klasifikasi

Menurut Prasetyo, klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target *f* yang memetakan setiap set atribut (fitur) *x* ke satu dari sejumlah label kelas *y* yang tersedia (Prasetyo, 2012). Algoritma klasifikasi menggunakan *data training* untuk membuat sebuah model. Model yang sudah dibangun tersebut kemudian digunakan untuk memprediksi label kelas data baru yang belum diketahui.

Salah satu pengukur kinerja klasifikasi adalah tingkat akurasi. Sebuah sistem dalam melakukan klasifikasi diharapkan dapat mengklasifikasi semua set data dengan benar, tetapi tidak dipungkiri bahwa kinerja suatu sistem tidak bisa 100% akurat. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi.Matriks konfusi merupakan tabel pencatat hasil kerja klasifikasi (Prasetyo, 2012).

Matriks konfusi merupakan tabel pencatat hasil kerja klasifikasi. Tabel 1 merupakan matriks konfusi yang melakukan klasifikasi masalah biner. Setiap sel f_{ij} dalam matriks menyatakan jumlah rekord (data) dari kelas *i* yang hasil prediksinya masuk ke kelas *j*. Misalnya, sel f_{11} adalah jumlah data dalam kelas A yang secara benar dipetakan ke kelas A, dan f_{10} adalah data dalam kelas A yang dipetakan secara salah ke kelas B. Matriks Konfusi dapat dilihat pada Tabel 2 berikut:

Tabel 2. Matriks Konfusi

f_{ij}		**Kelas Hasil Prediksi (*j*)**	
		Kelas = A	**Kelas = B**
Kelas asli (*i*)	Kelas = A	f_{11}	f_{10}
	Kelas = B	f_{01}	f_{00}

dengan f_{11} : Banyaknya observasi kelas 1 yang tepat diklasifikasikan sebagai kelas 1.
f_{10} : Banyaknya observasi kelas 1 yang salah diklasifikasikan sebagai kelas 0.
f_{01} : Banyaknya observasi kelas 0 yang salah diklasifikasikan sebagai kelas 1.
f_{00} : Banyaknya observasi kelas 0 yang tepat diklasifikasikan sebagai kelas 0.

Berdasarkan isi matriks konfusi, dapat diketahui jumlah data dari masing-masing kelas yang diprediksi secara benar, yaitu ($f_{11} + f_{00}$), dan data yang diklasifikasikan secara salah, yaitu ($f_{10} + f_{01}$). Maka dapat dihitung tingkat akurasi dan tingkat kesalahan prediksi:

1. Akurasi $= \dfrac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$
2. Tingkat kesalahan prediksi $= \dfrac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$

3. METODOLOGI PENELITIAN

3.1. Lokasi dan Variabel Penelitian

Lokasi Penelitian adalah Jurusan Statistika Fakultas Sains dan Matematika Universitas Diponegoro Semarang Jawa Tengah. Data diperoleh secara random berjumlah 119 dari berbagai angkatan dan beberapa periode wisuda.

Variabel penelitian terdiri dari variabel dependen dan independen. Variabel dependen pada penelitian ini yaitu masa (lama) studi mahasiswa Prodi Statistika Fakultas Sains dan yang terdiri dari 2 kategori lama studi , yaitu dikatagorikan tepat waktu adalah mahasiswa yang lulus <= 4 tahun dan tidak tepat waktu adalah mahasiswa yang lulus > 4 tahun. Sedangkan variabel independennya adalah jenis kelamin, IPK, beasiswa (pernah atau tidak menerima beasiswa), *part time* (pernah atau tidak pernah melakukan *part time*), jalur masuk ke Undip (PSSB, Mandiri, atau SNMPTN), dan Organisasi (aktif atau tidak dalam berorganisasi).

Tabel 3. Variabel Penelitian

No	Variabel	Kategori	Skala
1	Lama Studi (Y)	0 = ≤ 4 1 = > 4	Nominal
2	Jenis Kelamin (X1)	0 = Perempuan 1 = Laki-laki	Nominal
3	IPK (X3)	0 = < 3 1 = 3-3,5 2 = > 3,5	Ordinal
4	Beasiswa (X4)	0 = Tidak 1 = Ya	Nominal
5	*Part time* (X5)	0 = Tidak 1 = Ya	Nominal
6	Organisasi (X6)	0 = Tidak 1 = Ya	Nominal
7	Jalur MasukUndip (X7)	0 = PSSB 1 = SNMPTN 2 = Mandiri	Nominal

3.2. Langkah-langkah Analisis

Langkah-langkah yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Membuat deskripsi data.
2. Membagi data menjadi sampel pelatihan dan sampel pengujian dengan melakukan beberapa kali percobaan dengan melihat hasil akurasi yang paling tinggi.
3. Mengkonstruksikan pohon keputusan dengan SVM:

a. Menentukan fungsi kernel permodelan
b. Menetukan nilai-nilai parameter kernel (d) dan parameter *cost* (C) untuk optimasi
c. Menentukan nilai parameter terbaik untuk optimasi data training untuk klasifikasi data testing

4. Mengkonstruksikan pohon keputusan dengan algoritma ID3 dengan menghitung nilai *entropy* dan *information gain* dari masing-masing atribut.
5. Melakukan analisis terhadap hasil pohon keputusan yang terbentuk dan menghitung nilai akurasi pohon.
6. Menguji pohon keputusan menggunakan sampel pengujian.
7. Membandingkan hasil SVM dan algoritma ID3

4. HASIL DAN PEMBAHASAN

4.1. Analisis Deskriptif

Analisis deskriptif digunakan untuk memperoleh gambaran data secara umum. Dari data secara random didapat 119 dari berbagai angkatan dan beberapa periode didapat persentase 58% diantaranya adalah mahasiswa yang lama studi dengan waktu kurang dari sama dengan 48 bulan (tepat waktu), sedangkan sisanya yaitu 42% merupakan mahasiswa yang lama studinya lebih dari 48 bulan (tidak tepat waktu). Berdasarkan jenis kelamin jumlah mahasiswa yang lulus dengan tepat waktu dengan jenis kelamin wanita adalah 75% dan pria 25%. Sedangkan untuk tidak tepat waktu jenis kelamin wanita adalah 62% dan pria 38%. Persentase berdasarkan IPK dengan lama studi tepat waktu dengan IPK kurang dari tiga tidak ada, IPK tiga sampai dengan tiga koma lima adalah 38%, dan IPK lebih dari tiga koma lima adalah 62%. Sedangkan, mahasiswa yang lulus tidak tepat waktu dengan IPK kurang dari tiga adalah 12%, IPK tiga sampai dengan tiga koma lima adalah 82%, dan IPK lebih dari tiga koma lima adalah 6%. Berdasarkan mahasiswa yang mendapatkan beasiswa untuk mahasiswa yang lulus tepat waktu dan pernah mendapatkan beasiswa adalah 20%, sedangkan yang tidak pernah sebanyak 80%. Jumlah mahasiswa yang lulus dengan lama studi tidak tepat waktu dan pernah mendapatkan beasiswa adalah sebanyak 26%, sedangkan yang tidak pernah sebanyak 74%. Untuk variabel *part time*, dari data yang diambil untuk mahasiswa dengan lama studi kurang dari sama dengan 4 tahun (tepat waktu) dan pernah melakukan pekerjaan *part time* adalah 39%, sedangkan, yang tidak pernah melakukan pekerjaan *part time* adalah 61%. Jumlah mahasiswa yang lulus dengan lama studi lebih dari 4 tahun (tidak tepat waktu) dan pernah melakukan pekerjaan *part time* adalah 58% sedangkan, yang tidak pernah melakukan pekerjaan *part time* adalah 42%.

Untuk variabel Organisasi, jumlah mahasiswa yang lulus dengan tepat waktu dan pernah mengikuti organisasi (aktif dalam berorganisasi) sebanyak 89% sedangkan, yang tidak pernah mengikuti organisasi sebanyak 11%. Jumlah mahasiswa yang lulus dengan lama studi tidak tepat waktu dan pernah mengikuti organisasi (aktif dalam berorganisasi) sebanyak 92% orang sedangkan yang tidak pernah mengikuti organisasi sebanyak 8%. Untuk variabel jalur masuk, jumlah mahasiswa yang lulus tepat waktu dengan jalur masuk ke Undip melalui jalur PSSB sebanyak 7%, jalur SNMPTN sebanyak 33%, dan jalur UM sebanyak 60% . Jumlah mahasiswa yang lulus dengan tidak tepat waktu dengan jalur masuk ke Undip melalui jalur PSSB sebanyak 6%, jalur SNMPTN sebanyak 68%, dan jalur UM sebanyak 26%.

4.2. Klasifikasi Lama Studi Mahasiswa dengan Metode *Support Vector Machine* (SVM)

Klassifikasi dengan metode SVM, digunakan fungsi kernel Linier, *Polynomial*, dan *Radial Basis Function* (RBF). Untuk melakukan klasifikasi dengan menggunakan SVM, ada beberapa nilai C (*cost*) yang digunakan untuk mengetahui ketepatan klasifikasi terbaik.

4.2.1 Klasifikasi SVM Menggunakan Fungsi Kernel Linier

Pada SVM dengan fungsi kernel linier, nilai C yang digunakan yaitu 0,1; 1; 10 dan 100. Nilai-nilai C tersebut kemudian diterapkan pada data *training* untuk mendapatkan nilai eror pada masing-masing model klasifikasi. Berikut adalah nilai eror dengan data *training* dan *testing*. Dari data yang diambil sebanyak 119 dibagi untuk data training sebanyak 89 dan testing 30. Hasil dari nilai C yang berbeda-beda diperoleh nilai eror klasifikasi seperti terlihat pada Tabel 4. Dari tabel tersebut, terlihat nilai parameter C (*cost*) terbaik untuk digunakan pada fungsi pemisah SVM dengan menggunakan fungsi kernel linier adalah C = 1.

Tabel 4. Nilai Eror klasifikasi dengan Menggunakan Fungsi Kernel Linier

Parameter *cost* (C)	Eror Klasifikasi
0,1	0,303371
1	0,269663
10	0,269663
100	0,325843

Dengan menggunakan parameter C = 1, nilai parameter kemudian diterapkan pada klasifikasi data testing yang kemudian dievaluasi dengan menghitung akurasi ketepatan klasifikasinya. Berikut adalah matriks konfusi dengan menggunakan data testing sebanyak 30 data.

Tabel 5. Matriks Konfusi dengan Menggunakan Fungsi Kernel Linier

Kelas	Prediksi	
	Tepat waktu	Tidak tepat waktu
Tepat waktu	15	5
Tidak tepat waktu	2	8

dengan akurasi $= \dfrac{15+8}{15+5+2+8} = 0{,}7667$

4.2.2 Klasifikasi SVM Menggunakan Fungsi Kernel *Polynomial*

Pada SVM dengan fungsi kernel *polynomial*, terdapat parameter d (*degree*) dan C (*cost*). Penentuan parameter d (*degree*) untuk fungsi pemisah dengan fungsi kernel *polynomial* ini dicobakan beberapa nilai parameter dengan rentang 2 sampai dengan 5 dan nilai C yang digunakan yaitu 0,1; 1; 10 dan 100. Nilai-nilai C dan *d* tersebut kemudian diterapkan pada data training untuk mendapatkan nilai eror pada masing-masing model klasifikasi. Nilai eror dengan data training dan menggunakan parameter d dan C yang berbeda-beda disajikan pada Tabel 6.

Tabel 6. Nilai Eror klasifikasi dengan Menggunakan Fungsi Kernel *Polynomial*

Parameter C (*cost*)	Parameter d (*degree*)	Eror klasifikasi
0,1	2	***0,157303***
	3	0,168539
	4	0,179775
	5	0,179775
1	2	0,179775
	3	0,179775
	4	0,179775
	5	0,179775
10	2	0,168539
	3	0,179775
	4	0,179775
	5	0,179775
100	2	0,168539
	3	0,179775
	4	0,179775
	5	0,179775

Dari Tabel 6, nilai parameter C (*cost*) dan parameter d (*degree*) terbaik untuk digunakan pada model fungsi pemisah SVM dengan menggunakan fungsi kernel *polynomial* adalah C = 0,1 dan d = 2.

Dengan menggunakan parameter C = 0,1 dan d = 2, nilai parameter tersebut kemudian diterapkan pada klasifikasi data training dan *testing* yang kemudian dievaluasi dengan menghitung akurasi ketepatan klasifikasinya. Matriks konfusi dengan menggunakan data *testing* sebanyak 30 data dapat dilihat pada Tabel 7.

Tabel 7. Matriks Konfusi dengan Menggunakan Fungsi Kernel *Polynomial*

Kelas	Prediksi	
	<= 4 tahun	> 4 tahun
<= 4 tahun	15	1
> 4 tahun	2	12

dengan akurasi = $\frac{15+12}{15+1+2+12} = 0{,}90$

4.2.3 Klasifikasi SVM Menggunakan Fungsi Kernel RBF

Pada SVM dengan fungsi kernel *Radial Basis Function* (RBF), terdapat parameter γ(*gamma)* dan C (*cost*). Penentuan parameter γ(*gamma)* untuk fungsi pemisah dengan fungsi kernel RBF ini dicobakan beberapa nilai parameter yaitu 0,003; 0,007; 0,015 dan 0,031 dan nilai C yang digunakan yaitu 0,1; 1; 10 dan 100. Nilai-nilai C dan γ tersebut kemudian diterapkan pada data *training* untuk mendapatkan nilai eror pada masing-masing model klasifikasi. Tabel Nilai error klasifikasi dapat dilihat pada Tabel 8. Dari tabel tersebut, nilai parameter C (*cost*) dan parameter γ (*gamma*) terbaik untuk digunakan pada model fungsi pemisah SVM dengan menggunakan fungsi kernel RBF adalah C = 100 dan γ = 0,031. Dengan menggunakan parameter C = 100 dan γ = 0,031, nilai parameter tersebut kemudian diterapkan pada klasifikasi data *testing* yang kemudian dievaluasi dengan menghitung akurasi ketepatan klasifikasinya.

Tabel 8. Nilai Eror Klasifikasi dengan Menggunakan Fungsi Kernel RBF

Parameter C (*cost*)	Parameter γ	Eror klasifikasi
0,1	0,003	0,41573
	0,007	0,41573
	0,015	0,41573
	0,031	0,41573
1	0,003	0,41573
	0,007	0,41573
	0,015	0,382023
	0,031	0,280899
10	0,003	0,213483
	0,007	0,179775
	0,015	0,224719
	0,031	0,280899
100	0,003	0,292135
	0,007	0,269663
	0,015	0,213483
	0,031	***0,146067***

Tabel 9. Matriks Konfusi dengan Menggunakan Fungsi Kernel RBF

Kelas	Prediksi	
	Tepat waktu	**Tidak tepat waktu**
Tepat waktu	15	1
Tidak tepat waktu	2	12

dengan akurasi = $\frac{15+12}{15+1+2+12} = 0{,}90$

Tabel 10. Hasil Klasifikasi dengan menggunakan SVM

	Fungsi Kernel		
	Linear	**RBF**	***Polynomial***
Training	0,7078	**0,8539**	0,8427
Testing	0,7667	**0,9000**	**0,9000**

4.3. Klasifikasi Lama Studi Mahasiswa Jurusan Statistika dengan Metode ID3 (*Iterative Dichotomiser 3*)

Hasil Algoritma ID3 untuk mengidentifikasi data studi kasus lama studi mahasiswa jurusan Statistika dengan menggunakan sampel sebanyak 119 lulusan. Setiap simpul daun dilabeli kelas yang ditunjukkan dengan jumlah kasus pada simpul tersebut dan jumlah kesalahan klasifikasi apabila terdapat kesalahan klasifikasi. Berikut ini informasi yang dapat diperoleh dari hasil klasifikasi menggunakan Algorithma ID3:

1. Simpul
   Simpul menunjukkan hasil pengujian nilai suatu atribut, simpul yang terbentuk meliputi simpul akar, simpul keputusan dan simpul daun. Pada penelitian ini banyak dari seluruh simpul yang terbentuk adalah mencapai 11 simpul.

2. Simpul Daun
   Simpul daun merepresentasikan kelas yang terbentuk. Pada penelitian ini terbentuk sebanyak 7 simpul daun, ini artinya terdapat 7 karakteristik status lama studi mahasiswa jurusan Statistika FSM Undip.
3. Simpul Akar
   Simpul akar merupakan simpul yang terpilih pertama kali dalam suatu konstruksi pohon keputusan berdasarkan nilai *information gain* terbesar. Berdasarkan konstruksi pohon yang terbentuk, dapat dilihat pada penelitian ini mengenai status lama studi mahasiswa Jurusan Statistika Fakultas Sains dan Matematika, bahwa atribut IPK terpilih sebagai simpul akar.

4.3.1. Konstruksi Algoritma ID3

1. Hitung nilai *entropy* kelas yang disimbolkan *Entropy* (S)

$$Entropy\ (S) = \sum_i^c - p_i log_2 p_i = -\left(\frac{52}{89}\right)\log_2\left(\frac{52}{89}\right) - \left(\frac{37}{89}\right)\log_2\left(\frac{37}{89}\right) = 0,97941$$

2. Hitung nilai *Entropy pada* atribut IPK

Entropy (sangat memuaskan) = $-\left(\frac{21}{50}\right)\log_2\left(\frac{21}{50}\right) - \left(\frac{29}{50}\right)\log_2\left(\frac{29}{50}\right) = 0,981454$

Entropy (memuaskan) = $-\left(\frac{0}{5}\right)\log_2\left(\frac{0}{5}\right) - \left(\frac{5}{5}\right)\log_2\left(\frac{5}{5}\right) = 0$

Entropy (dengan pujian) = $-\left(\frac{31}{34}\right)\log_2\left(\frac{31}{34}\right) - \left(\frac{3}{34}\right)\log_2\left(\frac{3}{34}\right) = 0,430552$

Gain (S, IPK) = $Entropy\ (S) - \sum_{v \in Values(A)} \frac{S_v}{S} Entropy\ (S_v)$

Gain (S, IPK) $= 0,97941 - \left(\frac{50}{89}\text{ x } 0,981454\right) - \left(\frac{5}{89}\text{ x } 0\right) - \left(\frac{31}{89}\text{ x } 0,430552\right) = 0,263553$

Pilih atribut dengan *information gain* terbaik sebagai pemilah. Berikut ini nilai *information gain* seluruh atribut pada simpul akar secara lengkap dengan proses perhitungan yang sama dengan proses mendapatkan nilai *information gain* atribut IPK.

Tabel 11. Nilai *Informartion Gain* pada Simpul Akar

No.	Atribut	Gain
1	Jenis Kelamin	0,006862
2	**IPK**	**0,263553**
3	Beasiswa	0,000023
4	Part time	0,035002
5	Organisasi	0,009775
6	Jalur Masuk	0,102298

Dapat dilihat pada Tabel 11 bahwa atribut IPK adalah atribut dengan nilai *information gain* terbesar, maka atribut IPK terpilih sebagai pemilah. Selanjutnya pembentukan cabang untuk mengkonstruksikan simpul anak berdasarkan informasi pada simpul akar yang sekaligus sebagai simpul induk.

4.3.2 Identifikasi Lama Studi Mahasiswa Jurusan Statistika FSM Undip

Algoritma ID3 menghasilkan 11 lama studi mahasiswa yang tepat waktu dan tidak tepat waktu. Berikut ini adalah profil lama studi mahasiswa yang lulus tepat waktu dan tidak tepat waktu:

a. Lulus Tepat Waktu

Pada penelitian ini algoritma ID3 menghasilkan 4 profil lama studi mahasiswa jurusan Statistika FSM Undip yang diidentifikasikan lulus tepat waktu:

1. Mahasiswa dengan IPK tiga sampai dengan tiga koma lima, tidak *part time*, jalur masuk melalui SNMPTN, dan tidak mengikuti organisasi.
2. Mahasiswa dengan IPK tiga sampai dengan tiga koma lima, tidak *part time*, jalur masuk melalui UM.
3. Mahasiswa dengan IPK tiga sampai tiga koma lima, tidak *part time*, jalur masuk melalui PSSB.
4. Mahasiswa dengan IPK lebih dari tiga koma lima.

b. Tidak Tepat Waktu

Pada penelitian ini algoritma ID3 menghasilkan 3 profil lama studi mahasiswa Jurusan Statistika FSM Undip yang diidentifikasikan tidak lulus tepat waktu:

1. Mahasiwa dengan IPK tiga sampai dengan tiga koma lima, tidak *part time*, jalur masuk melalui SNMPTN, mengikuti organisasi.
2. Mahasiswa dengan IPK tiga sampai tiga koma lima dan ikut *part time*.
3. Mahasiswa dengan IPK kurang dari tiga.

4.3.3 Pengukuran Ketepatan Hasil Klasifikasi Algoritma ID3

Setelah didapatkan secara utuh hasil klasifikasi Algoritma ID3 berupa pohon keputusan, langkah selanjutnya adalah mengukur ketepatan hasil klasifikasi yang terbentuk. Ketepatan klasifikasi maupun kesalahan klasifikasi dirangkum dalam tabel matriks konfusi. Berikut ini tabel matrik konfusi pada konstruksi pohon Algoritma ID3:

Tabel 12. Matriks Konfusi Algoritma ID3

Kelas	Prediksi	
	Tepat waktu	**Tidak tepat waktu**
Tepat waktu	32	5
Tidak tepat waktu	6	46

dengan akurasi = $\frac{32+46}{89} \text{x } 100\% = 87{,}64\%$

4.3.4 Pengujian Hasil Pohon Keputusan.

Pohon konstruksi Algoritma ID3 tersebut diujikan dengan memasukkan data testing kedalam pohon konstruksi. Ukuran sampel pelatihan dalam pelatihan ini adalah sebanyak 30 kasus. Tabel matriks konfusi pada sampel pengujian sebagai berikut:

Tabel 13. Matriks Konfusi Sampel Pengujian

Kelas	Prediksi	
	Tepat waktu	**Tidak tepat waktu**
Tepat waktu	12	1
Tidak tepat waktu	2	15

dengan akurasi = $\frac{12+15}{30} \text{x } 100\% = 90\%$

4.4. Perbandingan Ketepatan Klasifikasi dengan Metode SVM(*Support Vector Machine*) dan Metode ID3 (*Iterative Dichotomiser 3*)

Setelah dilakukan perhitungan dari kedua metode tersebut , baik untuk data training maupun testing diperoleh hasil sebagai berikut:

Tabel 14. Perbandingan Klasifikasi SVM dan ID3

Akurasi	Training	Testing
SVM dengan Kernel RBF	85,39%	90%
Algoritma ID3	87,64%	90%

Berdasarkan perbandingan akurasi dari kedua metode dapat diketahui bahwa kedua metode menghasilkan akurasi yang baik.

5. KESIMPULAN

Dari hasil dan pembahasan dapat disimpulkan:

1. Dari gambaran deskriptif :
   a. Jumlah yang lulus tepat waktu lebih banyak perempuan dibanding laki-laki.
   b. IPK lulusan baik yang tepat waktu maupun tidak tepat waktu, lebih dari 60% IPK diatas 3
   c. Organisasi maupun bea siswa tdk mempengaruhi ketepatan waktu lulusan, banyak yang lulus tepat waktu aktif dalam organisasi.
   d. Pekerjaan *part time* mengganggu ketepatan waktu lulusan.
2. Ketepatan klasifikasi dengan menggunakan metode SVM untuk data testing, diperoleh hasil yang sama yaitu mencapai 90 % untuk metode SVM dengan fungsi kernel polynomial dan RBF.
3. Dengan metode ID3, IPK terpilih sebagai informasi Gain
4. Perbandingan kedua metode SVM dan ID3, baik untuk data training maupun testing menghasilkan akurasi yang baik, yaitu 90%. Khususnya unt data training ID3 lebih unggul dibanding SVM.

DAFTAR PUSTAKA

Cristanini, N and Jhon S. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge. Cambrige University Press.

David, Mcg. 2004. *Tutorial*: *The ID3 Decision Tree Algorithm*. Monash University Faculty of Information Technology.

Gunn, S. 1998. *Support Vector Machine for Classification and Regression*. University of Southampton. Southampton.

Han, J., Kamber, M. and Pei, J. 2011. *Data Mining Conce*pt and *Technique Third edition*. Massachusetts. Elservier Inc.

Kecman, V. 2005. *Support Vector Machines – an Introduction*. Springer-Verlag Berlin Heidelberg. Netherlands.

Nugroho, A.S., Witarto, A.B, dan Handoko, D. 2003. *Support Vector Machine – Teori dan Aplikasinya dalam Bioinformatika*. URL: http://asnugroho.net/papers/ikcsvm.pdf

Prasetyo, E. 2012. *Data Mining*: *Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta. C.V Andi Offset.

Santosa, B. 2007. *Data Mining*: *Teknik Pemanfaatan Data Untuk Keperluan Bisnis*. Yogyakarta. Graha Ilmu.

Lampiran : Pohon Keputusan Data Training.