AITI: Jurnal Teknologi Informasi, Volume 22 No. 1 Maret 2025, 1-14
ISSN 1693-8348 E-ISSN 2615-7128

JURNAL TEKNOLOGI INFORMASI

Meningkatkan kinerja SVM: Dampak berbagai teknik seleksi fitur pada akurasi prediksi

Lenny Margaretta Huizen [1)], Muhammad Basyier A [2)], Mochamad Idris [2)]
[1,2)]Universitas Semarang
Jl. Soekarno-Hatta, Tlogosari, Semarang 50196
Email: [1)]lenny@usm.ac.id

Riwayat Artikel		
Diterima:	Direvisi:	Disetujui:
08-08-2023	22-04-2024	25-04-2024

Abstract

In Higher Education accreditation, student graduation plays an important role as one of the assessment criteria. Graduation prediction is the main focus of helping institutions assess a student to graduate on time. This study takes historical data from students who have graduated, which is taken through a questionnaire from the Department of Information Systems and Informatics Engineering at Semarang University students. The feature selection selects the most relevant attributes in graduation prediction. The results of this selection are tested using the Support Vector Machine (SVM) Algorithm. The main objective of this study is to evaluate the impact of feature selection on graduation prediction. The results show that SVM with feature selection using weight by relief achieves an accuracy of 82%, a precision of 83.42%, and a recall of 80.83%. In contrast, SVM without weight by relief shows an accuracy of 69.23%, a precision of 70.83%, and a recall of 67.86%. The use of feature selection successfully reduces features from 27 to four of the most influential features..

Keywords: *feature selection, SVM, accuracy, data mining*

Abstrak

Pada akreditasi Perguruan Tinggi, kelulusan mahasiswa memainkan peran penting sebagai salah satu kriteria penilaian. Prediksi kelulusan merupakan fokus utama untuk membantu institusi dalam menilai seorang mahasiswa lulus tepat waktu. Penelitian ini mengambil data historis dari mahasiswa yang telah lulus, yang diambil melalui kuesioner dari mahasiswa Program Studi Sistem Informasi dan Teknik Informatika Universitas Semarang. Metode seleksi fitur digunakan untuk menyeleksi atribut yang paling relevan pada prediksi kelulusan. Hasil seleksi ini diujikan menggunakan Algoritma *Support Vector Machine* (SVM). Tujuan utama dari penelitian ini adalah untuk mengevaluasi dampak seleksi fitur terhadap prediksi kelulusan. Hasil pengujian menunjukkan bahwa SVM dengan seleksi fitur menggunakan *weight by relief* mencapai akurasi sebesar 82%, presisi 83,42%, dan recall 80,83%. Sebaliknya, SVM tanpa menggunakan *weight by relief* menunjukkan akurasi 69,23%, presisi 70,83%, dan *recall* 67,86%. Penggunaan seleksi fitur berhasil mengurangi fitur dari 27 menjadi empat fitur yang paling berpengaruh.

Kata kunci: seleksi fitur, SVM, akurasi, *data mining*

Pendahuluan

Bagi seorang lulusan, akreditasi sebuah Perguruan Tinggi (PT) lebih dipandang sebagai bentuk kualitas jika dibandingkan dengan Indeks Prestasi Kumulatif (IPK) [1]. Banyaknya mahasiswa yang dapat lulus dengan masa studi yang tepat menggambarkan kinerja yang baik pada sebuah perguruan tinggi, karena kelulusan menjadi satu dari beberapa kriteria penilaian akreditasi pada sebuah PT [2]. Penilaian mutu serta kelayakan sebuah institusi merupakan salah satu bentuk penilaian yang dilakukan dimana penilainya merupakan badan mandiri yang ada di luar lingkup PT.

Faktor yang berpengaruh terhadap tingkat keberhasilan mahasiswa dapat berasal dari faktor internal, yaitu dari mahasiswa sendiri, maupun faktor eksternal, yaitu dari luar mahasiwa. Faktor internal dapat berasal dari emosi, tingkat kepintaran mahasiswa, suasana hati, dan sebagainya, sedangkan faktor eksternal dapat berasal dari lingkungan mahasiswa, keluarga, atau kampus tempat berkuliah.

Untuk meningkatkan akurasi dan efisiensi prediksi kelulusan, dapat digunakan pemilihan fitur, yang bertujuan untuk memilih subset fitur yang paling relevan dan signifikan. Teknik seleksi fitur ini memungkinkan adanya pengurangan dimensi data, yang dapat meningkatkan keakuratan dan kecepatan algoritma. Dalam penelitian ini akan dilakukan peningkatan performa algoritma *Support Vector Machine* (SVM) dalam memprediksi kelulusan mahasiswa melalui pemilihan fitur yang tepat. SVM dipilih karena kemampuannya yang terkenal dalam menangani data non-linear dan berdimensi tinggi. Efektivitas dari SVM sangat bergantung pada jenis fungsi kernel yang digunakan, yang harus dipilih dan diterapkan berdasarkan karakteristik khusus dari *dataset* yang dianalisis.

Salah satu pendekatan pra-pemrosesan dalam *data mining* adalah pemilihan fitur. Pemilihan fitur dengan mengidentifikasi pola terkait berupaya meningkatkan algoritma pembelajaran yang mengkategorikan fitur ke dalam kelompok berbeda. Pemilihan karakteristik yang lebih relevan dan berdampak pada klasifikasi akan menghasilkan kelengkapan fitur. Untuk mengurangi jumlah fitur dan dimensi data yang tidak diperlukan, dilakukan pemilihan fitur. Untuk memperkecil dimensi model, maka diterapkan pemilihan fitur. Pemilihan fitur dapat memudahkan atau mempercepat proses pemodelan atau pembelajaran. Dengan metode ini, jumlah kualitas yang harus diproses dan diperiksa kompleksitasnya berkurang. Teknik ini dapat mengidentifikasi *subset* fitur yang signifikan dari *dataset* nilai mahasiswa. Selain itu, pemilihan fitur dapat digunakan untuk mengurangi dimensi model, membantu dalam mengurangi fitur domain, dan menghapus fitur yang berlebihan atau tidak relevan.

Data yang digunakan adalah data kuesioner yang sudah diisi oleh mahasiswa yang sudah dinyatakan lulus. Pemrosesan awal data untuk penelitian ini dilakukan melalui pemilihan fitur, yaitu memilih sub kumpulan fitur dengan membuang item yang memiliki nilai prediktif minimal. Manfaat pemilihan fitur

antara lain menurunkan dimensi, menghilangkan karakteristik yang berlebihan, menyederhanakan data, menurunkan volume data, dan meningkatkan kebenaran algoritma [3]. Seseorang dapat memilih sub kumpulan informasi relevan dan non-*redundan* terbaik untuk memberikan jawaban ideal dengan pemilihan fitur tanpa membuat model atau algoritma menjadi lebih kompleks. Untuk akurasi maksimal (*confusion matrix*), algoritma harus mampu memberikan kinerja yang optimal.

Salah satu teknik pembelajaran untuk masalah regresi dan klasifikasi adalah *Support Vector Machines* (SVM). SVM adalah algoritma yang populer untuk klasifikasi dan regresi, yang baik digunakan untuk masalah linear maupun non-linear. Salah satu keunggulannya terletak pada kemampuannya untuk mengatasi pemisahan linear dalam data yang non-linear dan berdimensi tinggi, yang dicapai melalui penggunaan fungsi kernel yang sesuai. Efektivitas dari SVM sangat bergantung pada jenis fungsi kernel yang digunakan, yang harus dipilih dan diterapkan berdasarkan karakteristik khusus dari *dataset* yang dianalisis [4]. Untuk memaksimalkan hasil pada prosedur klasifikasi prediksi kelulusan mahasiswa digunakan seleksi fitur.

Sebuah penelitian telah dilakukan untuk mengevaluasi pengaruh pemrosesan data teks dengan metode seleksi fitur pada *classifier* Naive Bayes untuk ekstraksi kata kunci dari dokumen Bahasa Indonesia. Eksperimen ini melibatkan variasi kategori dokumen, bagian dokumen yang diolah, dan jumlah fitur yang dipotong. Penelitian ini menggunakan nilai *recall* sebagai parameter utama untuk mengukur performa model yang dihasilkan. Penelitian tersebut juga menunjukkan bahwa metode seleksi fitur *Chi-Square* dapat meningkatkan model dengan signifikan, dengan hasil nilai *recall* sekitar 3% pada dokumen dengan kategori yang berbeda, 72,50% saat hanya memproses bagian Abstrak, dan 68,57% saat memproses bagian Abstrak, Pendahuluan, dan Kesimpulan. Sementara itu, algoritma *Term Frequency Document Frequency* (TFDF) menghasilkan model yang baik hanya pada dokumen dengan bagian Abstrak, namun tidak efektif untuk dokumen yang memuat ketiga bagian [5].

Dalam penelitian lain telah dilakukan evaluasi terhadap penggunaan seleksi fitur dalam kombinasi dengan algoritma Naive Bayes untuk klasifikasi data penyakit jantung. Penelitian ini mencoba beberapa metode seleksi fitur, seperti *Forward Selection*, *Backward Elimination*, kombinasi *union* dari keduanya, serta *Information Gain* dan *Gain Ratio*. Data yang digunakan diambil dari UCI *Machine Learning Repository*. Hasil penelitian menunjukkan bahwa kombinasi *union* dari seleksi fitur *Information Gain* dan *Gain Ratio* menghasilkan akurasi tertinggi sebesar 91,80% dengan rasio data latih dan uji 80:20. Sementara itu, kombinasi *union* dari *Forward Selection* dan *Backward Elimination* hanya mencapai akurasi sebesar 83,61% [6]

SVM berbasis kernel hibrida *ensemble* digunakan sebagai metode pemilihan fitur dan diimplementasikan serta diuji pada *dataset benchmark* KYOTO 2006+. Metode yang diusulkan mencakup dua skenario, yaitu yang pertama dengan 11 fitur, dikenal sebagai EHK-SVM (11), dan yang kedua dengan sembilan fitur, dikenal sebagai EHK-SVM (9). Hasil observasi menunjukkan bahwa skenario dengan 11 fitur memberikan akurasi tertinggi, yang merupakan pilihan yang optimal untuk deteksi dengan tingkat keakuratan yang tinggi. Sementara itu, skenario dengan sembilan fitur memberikan keuntungan dalam waktu komputasi, memfasilitasi deteksi yang lebih cepat. EHK-SVM yang diusulkan dibandingkan dengan model-model eksisting yang menggunakan *dataset* yang sama untuk evaluasi, dan kesimpulan ditarik bahwa pendekatan yang diusulkan menunjukkan akurasi yang lebih baik [7].

Analisis sentimen populer digunakan untuk menilai sentimen masyarakat terhadap suatu objek, termasuk kinerja Lembaga Pemilihan Umum (KPU) dalam pelaksanaan PEMILU. Penelitian ini menggunakan data *retweet* dan komentar dari Twitter. Algoritma yang digunakan adalah SVM dengan optimasi bobot seleksi fitur korelasi. Hasil menunjukkan peningkatan signifikan dalam akurasi dan AUC dengan penggunaan FS [8].

Penelitian mengenai *social distancing* di Twitter dianalisis menggunakan algoritma SVM dengan seleksi fitur *Weight by Correlation.* Hasil penelitian menunjukkan peningkatan akurasi dari 67,00% menjadi 70,33% dan AUC dari 0,709 menjadi 0,838 [9].

Heterogenitas dan dimensi tinggi dari kumpulan data sering menjadi tantangan dalam mendapatkan akurasi optimal dalam proses klasifikasi prediksi. Pemilihan fitur yang tepat menjadi kunci penting dalam pembelajaran mesin dan alur penelitian. Penelitian telah dilakukan untuk pemilihan fitur menggunakan metode *wrapper* dan *embedded* pada *dataset* 146 mahasiswa sarjana dengan 13 atribut data dan satu label, untuk memprediksi apakah mahasiswa tersebut lulus tepat waktu atau tidak. Hasil penelitian menunjukkan bahwa dengan menggunakan algoritma *Random Forest Classification*, akurasi awal sebesar 73% dapat ditingkatkan menjadi 100% setelah penerapan metode pemilihan fitur *wrapper* dan *embedded* [10].

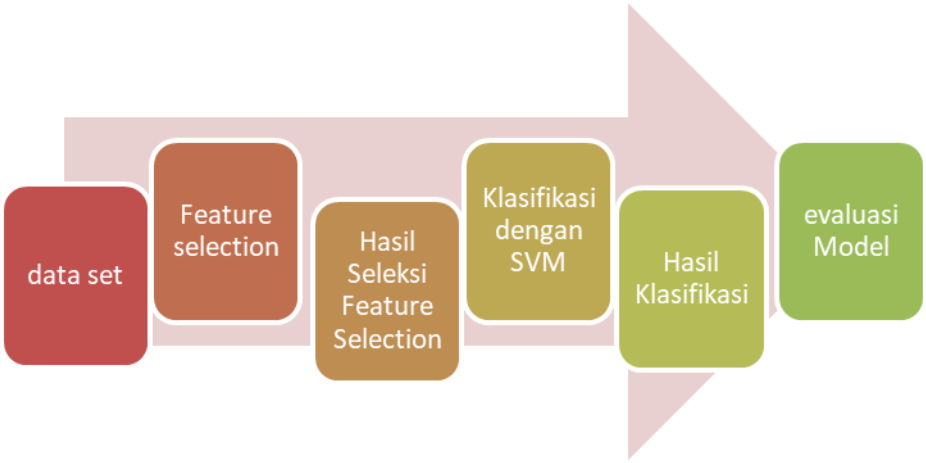
Seleksi fitur merupakan proses pemilihan sebagian subset dari fitur atau variabel dalam suatu *dataset*. Tujuannya adalah untuk mengurangi *noise* atau data yang tidak relevan guna meningkatkan hasil atau performa secara keseluruhan [11]. Contoh penerapan seleksi fitur dapat ditemukan dalam penelitian mikrogenetik [12]. Dalam domain genetika, seringkali *dataset* berjumlah ratusan atau ribuan data yang saling terkait. Namun, tidak semua variabel atau fitur dalam *dataset* tersebut memberikan informasi yang relevan terhadap kelas data tersebut.

Dengan melakukan seleksi fitur, variabel yang tidak memberikan kontribusi yang signifikan terhadap klasifikasi data dapat diidentifikasi dan dihapus. Dengan

mengurangi jumlah variabel yang digunakan, kita juga secara efektif mengurangi jumlah data yang harus diproses. Hal ini tidak hanya membantu mengurangi beban komputasi, tetapi juga meningkatkan performa klasifikasi data, karena fokus diberikan pada variabel yang benar-benar penting dalam mendukung tujuan analisis [13].

Metode

Salah satu teknik *data mining* yang paling populer dan banyak digunakan adalah seleksi fitur [14], yang merupakan strategi yang digunakan dalam penelitian ini. Tahap pra-pemrosesan bertujuan untuk mengurangi kompleksitas kualitas yang sedang diproses dan dianalisis. Tujuannya adalah untuk mengidentifikasi subset dari fitur yang paling relevan. Proses seleksi fitur ini sering digunakan untuk mengurangi dimensi dari model, yang mengurangi jumlah fitur yang tidak perlu dalam domain dan menghilangkan fitur yang berlebihan [15]. Proses pembelajaran atau pemodelan menjadi lebih efisien dan cepat. Hasil dari pemilihan fitur inilah yang digunakan untuk proses klasifikasi dengan menggunakan algoritma SVM. Alur seleksi fitur dapat dilihat pada Gambar 1.



Gambar 1 Alur seleksi fitur

SVM merupakan algoritma yang pertama kali dikembangkan dalam *Annual Workshop on Computational Learning Theory* pada tahun 1992 oleh Boser, Guyon, dan Vapnik. SVM menggabungkan tiga konsep utama, yaitu pencarian *hyperplane* optimal, operasi *dot-product* yang diperluas (untuk mengakomodasi solusi *linear* dan *non-linear*), dan konsep *soft-margin* (dengan kemungkinan adanya kesalahan pada *dataset* pelatihan) [16]. Terlepas konsepnya yang sederhana, SVM dapat menunjukkan performa yang baik.

Pengoperasian SVM dijelaskan sebagai upaya untuk menemukan *hyperplane* terbaik yang berfungsi sebagai pemisah antar kelas pada ruang input. *Hyperplane* ditemukan dengan mengukur margin antara *hyperplane* dan model data terdekat dari setiap kelas. Margin sendiri merupakan jarak antara *hyperplane* dan data yang terdekat dari kedua kelas. Salah satu perbedaan utama antara SVM dan

jaringan syaraf (*neural network*) lainnya adalah bahwa SVM berfokus pada pencarian *hyperplane* terbaik dalam ruang input, khususnya dalam pemisahan antara kelas-kelas data.

Pada SVM, data yang tersedia dapat dinyatakan sebagai $x_i \in R^d$, sedangkan label data dapat dinyatakan sebagai $y_i \in \{-1+1\}$ dengan i = 1, 2, ..., n dimana n adalah banyaknya data [17]. Oleh karena itu, dapat diasumsikan bahwa kedua kelas -1 dan +1 dapat dipisahkan secara sempurna oleh *hyperplane* berdimensi *d*, persamaan ini digunakan untuk mendefinisikan *hyperplane* yang memisahkan dua kelas data yang berbeda, dimana $\vec{w}$ merupakan vektor bobot yang menentukan orientasi *hyperplane* dan b merupakan bias yang menentukan posisi *hyperplane* dan $\vec{x}$ untuk lebih jelasnya dapat dilihat pada Persamaan (1)

$$\vec{w} * \vec{x} + b = 0 \tag{1}$$

Pattern $\vec{x}$ termasuk dalam *class* -1 sehingga dapat dirumuskan sebagai *pattern* yang memenuhi pertidaksamaan, menunjukkan bahwa vektor $\vec{x}$ berada di sisi tertentu dari sebuah *hyperplane* dalam ruang vektor, khususnya di sisi yang mengindikasikan kelas negatif pada Pertidaksamaan (2).

$$\vec{w} * \vec{x} + b \leq -1 \tag{2}$$

Pattern $\vec{x}$ merupakan *class* +1 menunjukkan bahwa titik $\vec{x}$ berada di sisi lain dari batas pemisah, yang umumnya mengindikasikan kelas positif dalam klasifikasi linier sehingga dapat dirumuskan pada Pertidaksamaan (3).

$$\vec{w} * \vec{x} + b \leq +1 \tag{3}$$

Margin terbesar diperoleh dengan memaksimalkan jarak antara *hyperplane* dan titik terdekatnya. Ini termasuk dalam *Quadratic Programming* yang berupaya mencari titik minimum dengan memperhatikan *constraint* Persamaan (4) dan Pertidaksamaan (5).

$$\min_{\ddot{w}} \tau(w) = \frac{1}{2}\left|\left|\ddot{w}\right|\right|^2 \tag{4}$$

dimana w merupakan vektor bobot yang dioptimalkan, $\tau(w)$ merepresentasikan fungsi objektif dan

$$y_i \overrightarrow{x_i}.\overrightarrow{w_j} + b \leq -1, \forall i \tag{5}$$

dimana y_i merupakan label kelas untuk data ke i yang nilainya bisa $+1$ atau -1, $\overrightarrow{x_i}$ merupakan vektor fitur data ke i.

Teknik *Lagrange Multiplier* menggabungkan minimisasi margin dan pemenuhan kendala klasifikasi, untuk lebih jelasnya dapat dilihat pada Persamaan (6).

$$L(\overrightarrow{w}, b, \alpha) = \frac{1}{2}\left|\left|\ddot{w}\right|\right|^2 - \sum_{i,j=1}^{l} \alpha_j (y_i((\vec{x}_i \overrightarrow{w}_{j+b)} - 1)) \tag{6}$$

$$(i = 1,2, \dots l)$$

di mana α_i merupakan *Lagrange multiplier* yang bernilai nol atau positif ($\alpha_i \geq 0$). Nilai optimal diperoleh dengan meminimalkan L terhadap $\vec{w}$ dan b, serta memaksimalkan L terhadap α, seperti terlihat pada Persamaan (7).

$$\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \, \vec{x}_i \vec{x}_j \tag{7}$$

menjadi seperti yang ditunjukkan pada Persamaan (8)

$$\alpha_i \geq 0 (i = 1, 2, \dots l) \qquad \sum_{i,j=1}^{l} a_i y_i = 0 \tag{8}$$

Dalam hasil perhitungan, nilai positif α_i didapatkan. Data yang memiliki korelasi dengan nilai α_i yang positif ini disebut sebagai vektor pendukung (*support vector*). Secara singkat, SVM adalah algoritma yang berfokus pada optimisasi fungsi matematika dengan memperhatikan data yang telah diberikan, dengan vektor pendukung memegang peran penting dalam proses ini.

Hasil dan Pembahasn

Proses pengumpulan data dimulai dari mencari variabel-variabel yang digunakan untuk membuat kuesioner. Variabel tersebut didapatkan dari hasil studi literatur. Dari variabel-variabel yang sudah terpilih, dibuat sebuah kuesioner berisi pertanyaan-pertanyaan. Untuk lebih jelasnya dapat dilihat pada Tabel 1.

Tabel 1 Pertanyaan kuesoner

Kode	Pertanyaan
A1	Jenis kelamin
A2	Jurusan
A3	Bagaimanakah pelayanan administrasi akademik dan kemahasiswaan dari fakultas Saudara?
A4	Di mana asal sekolah Anda?
A5	Bagaimanakah pelayanan administrasi akademik dan kemahasiswaan dari Universitas Saudara?
A6	Bagaimanakah proses pelayanan pendaftaran kerja praktek atau K3 atau magang atau tugas akhir di Fakultas?
A7	Bagaimanakah sistem pendaftaran kerja praktek atau K3 atau magang atau tugas akhir di Fakultas yang memanfaatkan Teknologi Informasi seperti *website* atau *mobile system*?
A8	Bagaimanakah ketepatan jadwal perkuliahan selama semester ini bagi mahasiswa?
A9	Bagaimanakah profesionalitas dosen dalam pengajaran?
A10	Bagaimana komunikasi anda dengan Dosen Pembimbing?
A11	Bagaimanakah profesionalitas dosen dalam bimbingan kerja praktek ataupun tugas akhir?
A12	Bagaimanakah kenyamanan dalam berinteraksi dengan dosen dalam hal akademik, seperti proses bimbingan kerja praktek atau tugas akhir?
A13	Bagaimanakah proses koreksi revisi bimbingan kerja praktek atau tugas akhir?

Kode	Pertanyaan
A14	Bagaimanakah kenyamanan dalam berinteraksi dengan dosen dalam hal non akademik seperti mengikuti perlombaan, bimbingan dalam komunitas yang diselenggarakan dosen, atau mengikuti kepanitian acara fakultas bersama dosen?
A15	Bagaimanakah keadaan ekonomi Saudara untuk bisa menuntaskan kuliah dengan tepat waktu?
A16	Apakah membuat program menjadi hambatan dalam menyelesaikan Tugas Akhir?
A17	Apakah faktor keluarga juga memengaruhi kelulusan Anda?
A18	Apakah Anda lulus tepat waktu (3-4 tahun)?
A19	Di mana asal daerah Anda?
A20	Apakah Anda mengikuti organisasi?
A21	Apakah Anda mengikuti perlombaan?
A22	Apakah Anda menulis artikel ilmiah?
A23	Apakah Anda penerima beasiswa?
A24	Apakah Anda asisten praktikum?
A25	Berapa jumlah SKS yang diambil rata-rata tiap semester?
A26	Apakah Anda bekerja?
A27	Berapa IPK akhir Saudara?

Dari pertanyaan-pertanyaan yang ada pada Tabel 1, dibuatlah sebuah kuesioner dengan hasil yang dapat dilihat pada Tabel 2.

Tabel 2 Hasil kuesioner

No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	⋮	A25	A26	A27
1	L	TI	C	SMK	C	C	C	C	C	C	C	C	C	C	C	⋮	20	T	3,5
2	P	TI	B	SMA	B	B	B	B	B	SB	SB	SB	SB	SB	C	⋮	22	Y	3,94
3	P	TI	B	SMA	B	B	B	B	B	B	B	B	B	B	B	⋮	21	Ya	3,57
4	L	TI	B	SMK	SB	B	C	B	B	B	SB	B	B	B	B	⋮	23	T	3,66
5	L	TI	B	SMA	B	SB	SB	B	B	B	B	B	B	B	B	⋮	20	Y	3,94
6	P	TI	C	SMK	C	C	C	B	B	B	B	B	B	B	TB	⋮	22	Y	3,5
7	L	TI	SB	SMK	B	B	B	B	B	B	B	B	B	B	C	⋮	20	Y	3,69
8	L	TI	SB	SMK	SB	B	B	C	B	SB	SB	B	B	B	C	⋮	20	Y	3,86
9	P	TI	B	SMK	B	B	B	B	B	B	B	B	B	B	B	⋮	24	Y	3,75
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	P	TI	B	SMK	B	B	B	B	B	C	B	B	B	B	C	⋮	24	T	3,45
62	P	TI	B	SMK	B	B	B	B	B	SB	SB	B	B	B	B	⋮	20	T	3,9

Pada hasil kuesioner Tabel 2, L=Laki-laki, P=Perempuan, B=Baik, C= Cukup, SB= Sangat Baik, T= Tidak, Y= Ya, TB=Tidak Baik. Hasil kuesioner tersebut kemudian diolah lagi menjadi data numerik atau data angka. Hasilnya dapat dilihat pada Tabel 3.

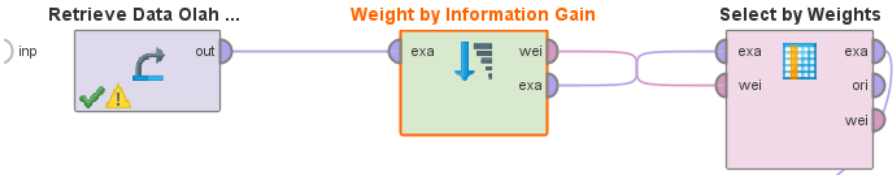
Tabel 3 Data numerik hasil kuesioner

No.	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	⋮	A25	A26	A27
1	2	1	3	2	3	3	3	3	3	3	3	3	3	3	3	⋮	20	2	3,5
2	1	1	2	1	2	2	2	2	2	1	1	1	1	1	3	⋮	22	1	3,94
3	1	1	2	1	2	2	2	2	2	2	2	2	2	2	2	⋮	21	1	3,57
4	2	1	2	2	1	2	3	2	2	2	1	2	2	2	2	⋮	23	2	3,66
5	2	1	2	1	2	1	1	2	2	2	2	2	2	2	2	⋮	20	1	3,94
6	1	1	3	2	3	3	3	2	2	2	2	2	2	2	4	⋮	22	1	3,69
7	2	1	1	2	2	2	2	2	2	2	2	2	2	2	3	⋮	20	1	3,69
8	2	1	1	2	1	2	2	3	2	1	1	2	2	2	3	⋮	20	1	3,86
9	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	⋮	24	1	3,75
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	1	1	2	2	2	2	2	2	2	3	2	2	2	2	3	⋮	24	2	3,45
62	1	1	2	2	2	2	2	2	2	1	1	2	2	2	2	⋮	20	2	3,9

Data diolah dengan menggunakan tool *Rapidminer*. *Tool* ini akan membaca data dari file Excel, proses pembacaan dapat dilihat pada Gambar 2. Pembacaan data dilakukan dengan melalukan upload file *Excel* yang berisi data hasil kuesioner yang sudah diubah menjadi data numerik (Tabel 3). Untuk lebih jelasnya dapat dilihat pada Gambar 2.

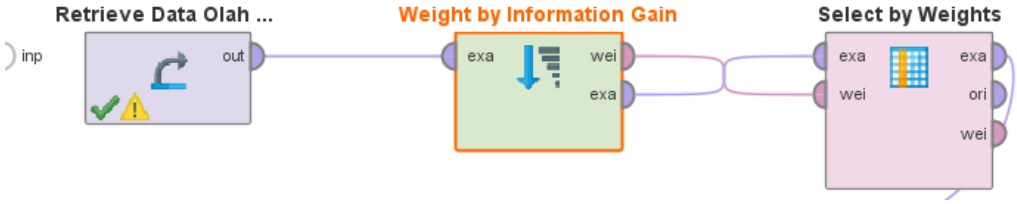
Gambar 2 Proses pembacaan data dengan RapidMiner

Implementasi algoritma seleksi fitur dapat dilakukan dengan menguji *dataset* mahasiswa menggunakan *tools* RapidMiner, tujuannya adalah untuk memilih atribut data yang memiliki karakteristik yang sama dengan data itu sendiri sehingga meningkatkan ketepatan klasifikasi. Algoritma seleksi fitur yang akan diujikan adalah *weight by information gain*, *weigt by relief*, dan *weight by correlation*. *Weight by information gain* menghitung nilai *gain* pada tiap fitur yang ada [18]. Untuk lebih jelasnya dapat dilihat pada Gambar 3.



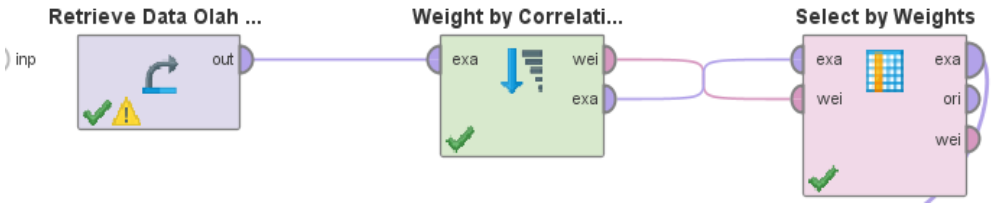
Gambar 3 Seleksi fitur *weight by information gain*

Penggunaan *weight by relief* dikarenakan metode ini secara efektif menghindari pengaruh fitur yang berkorelasi lemah atau fitur yang tidak berkorelasi pada hasil klasifikasi SVM [19] untuk lebih jelasnya dapat dilihat pada Gambar 4.



Gambar 4 Seleksi fitur *weight by relief*

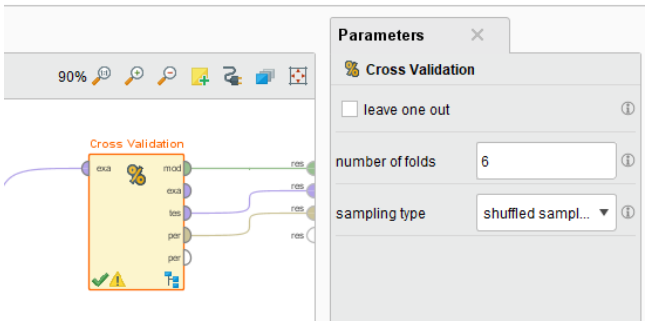
Weight by Correlation berfungsi untuk menghitung bobot berdasarkan korelasi antara satu fitur dengan fitur yang ada dalam *dataset* dengan mempertimbangkan nilai label [20], untuk lebih jelasnya dapat dilihat pada Gambar 5.



Gambar 5 *Weight by correlation*

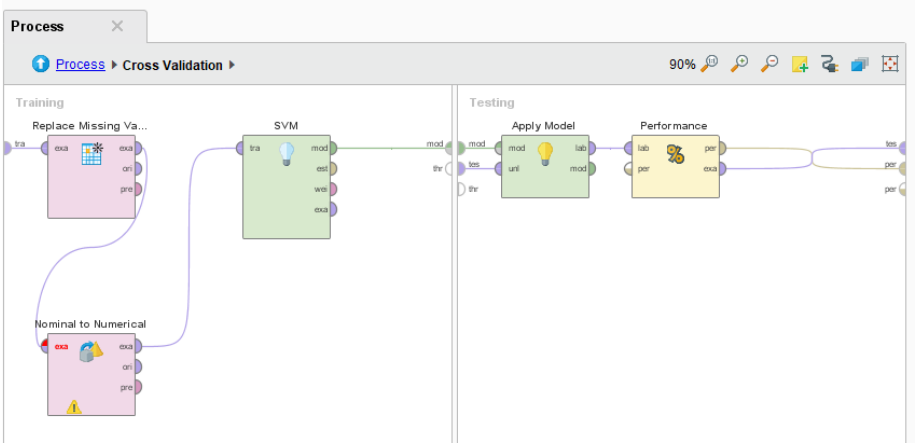
Data awal dibagi menjadi dua bagian, yaitu data latih dan data uji. Bagian data latih dan data uji diambil secara acak dari *dataset* yang telah melalui tahap *preprocessing*. Metode *Cross Validation* digunakan untuk mengukur kinerja model yang dibangun, yang bertujuan untuk menilai akurasi model berdasarkan

perbandingan tertentu antara data latih dan data uji. Pemisahan ini dapat dilakukan secara relatif menggunakan rasio pemisahan yang disesuaikan dengan rasio yang telah ditentukan, dan jenis pengambilan sampel yang digunakan adalah *shuffle sampling*. Proses ini dapat dilihat secara detail dalam Gambar 6 yang menggambarkan langkah-langkah *Cross Validation.*



Gambar 6 *Cross Validation*

Dalam klasifikasi data penelitian dapat menggunakan SVM, dan langkah selanjutnya adalah menerapkan model ini untuk melakukan prediksi pada data. Untuk mengukur sejauh mana model ini dapat memprediksi data dengan tepat, dilakukan evaluasi menggunakan metrik kinerja. Untuk detail lebih lanjut, dapat dilihat pada Gambar 7 yang menunjukkan implementasi algoritma SVM dengan lebih rinci.



Gambar 7 Implementasi algoritma SVM

Pada penelitian ini pengujian dibagi menjadi dua jenis yaitu, *data training* dan *data testing*. Data yang digunakan sebagai data training dan data testing diambil secara acak dari 62 data yang akan dianalisis. Pada pembagiannya 70% data training, 30% data testing (rasio 70:3). Untuk detailnya dapat dilihat pada Tabel 4.

Tabel 4 Hasil perbandingan implementasi algoritma SVM dengan seleksi fitur

	Accuracy	*Precision*	*Recall*
Weight by Information Gain	67,50%	63,83%	64,33%
Weight by Relief	82%	83,42%	80,83%
Weight by Correlation	80,95%	74,01%	77,64%
SVM tanpa seleksi fitur	69,23%	70,83%	67,86%

SVM dengan seleksi fitur dan *weight by relief* mempunyai nilai *accuracy* sebesar 82%, *precision* 83,42%, *recall* 80,83%. Nilai ini merupakan nilai yang tertinggi dibandingkan dengan hasil pembobotan yang lain. Penggunaan SVM tanpa seleksi fitur mempunyai nilai *accuracy* sebesar 69,23%, *precision* 70,83%, *recall* 67,86%. Hal ini merupakan nilai yang lebih rendah jika dibandingkan dengan SVM dengan *weight by relief.* Pada empat pengujian, jumlah variasi fitur yang digunakan adalah 27 fitur, akan tetapi setelah dilakukan dengan empat pengujian *feature* yang berbeda, jumlah variasi fitur berkurang menjadi enam fitur. Hal ini membuktikan bahwa seleksi fitur mempunyai pengaruh terhadap fitur-fitur yang ada. Hasil seleksi fitur *weight by relief* menghasilkan enam fitur yang dapat dilihat pada Gambar 8 dan Tabel 5.

Attribute ↓	Weight
A7	-0.410
A6	-0.056
A27	-0.571
A22	0.508
A19	-0.019
A17	0.892

Gambar 8 Variasi fitur pada seleksi fitur dengan *weight by relief*

Tabel 5 Hasil seleksi fitur

Kode	Pertanyaan
A6	Bagaimanakah proses pelayanan pendaftaran kerja praktek atau K3 atau magang atau tugas akhir di Fakultas ?
A7	Bagaimanakah sistem pendaftaran kerja praktek atau K3 atau magang atau tugas akhir di Fakultas yang memanfaatkan Teknologi Informasi seperti *website* atau *mobile system*?
A17	Apakah faktor keluarga juga memengaruhi kelulusan Anda?
A19	Di mana asal daerah Anda?
A22	Apakah Anda menulis artikel ilmiah?
A27	Berapa IPK akhir Saudara?

Simpulan

Dengan mengidentifikasi pola yang bersangkutan, pemilihan fitur berupaya untuk meningkatkan proses pembelajaran yang mengelompokkan suatu atribut ke dalam kategori tertentu. Dengan memilah fitur akan membantu dalam meminimalisir fitur domain serta menghapus fitur yang tidak diperlukan. Cara pemilihan fitur ini dapat mempercepat proses pembelajaran atau pemodelan. Hasil yang didapatkan dari pengujian dengan menggunakan empat variasi algoritma dari seleksi fitur dan pengujian tanpa melakukan seleksi fitur adalah SVM dengan seleksi fitur *weight by relief* mempunyai nilai *accuracy* sebesar 82%, *precision* 83,42%, *recall* 80,83%. Nilai ini merupakan nilai yang tertinggi dibandingkan dengan hasil pembobotan yang lain. Penggunaan SVM tanpa *weight by relief* mempunyai nilai *accuracy* sebesar 69,23%, *precision* 70,83%, dan *recall* 67,86%, yang menunjukkan nilai yang lebih rendah jika dibandingkan dengan SVM dengan

weight by relief. Penggunaan seleksi fitur tersebut memengaruhi jumlah fitur yang digunakan, yaitu dari yang semula 27 fitur, setelah menggunakan algoritma seleksi fitur, maka dipilihlah empat fitur yang paling relevan atau paling berpengaruh terhadap hasil akurasi. Untuk penelitian selanjutnya adalah menerapkan metode *data mining* yang lain, seperti *Decision Tree, Random Forest,* atau *Neural Networks*, untuk membandingkan dan mengoptimalkan model prediksi kelulusan. Eksplorasi metode-metode tersebut dapat memberikan wawasan baru tentang bagaimana berbagai pendekatan dalam *data mining* dapat memengaruhi akurasi, presisi, dan *recall* dalam prediksi kelulusan.

Daftar Pustaka

[1] M. Rashif *et al.*, "Fungsi Penilaian Instrumen Akreditasi bagi Institusi Perguruan Tinggi Baru." *Al'Adl: Jurnal Hukum*, vol. 13, no. 2, pp. 391-409, 2021.

[2] L. Setiyani, M. Wahidin, D. Awaludin, and S. Purwani, "Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes: Systematic Review," *Faktor Exacta*, vol. 13, no. 1, p. 35, Jun. 2020, doi: 10.30998/faktorexacta.v13i1.5548.

[3] I. Made, B. Adnyana, S. B. Jln, and R. Puputan, "Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa" *Jurnal Sistem dan Informatika*, vol. 13, no. 2, pp. 72-76, 2019.

[4] S. Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti *et al.*, "Terakreditasi SINTA Peringkat 2 Penggunaan Feature Selection di Algoritma Support Vector Machine untuk Sentimen Analisis Komisi Pemilihan Umum," *masa berlaku mulai*, vol. 1, no. 3, pp. 364–370, 2017.

[5] H. M. Lumbantobing and R. A. Marcellino, "Penerapan Metode Feature Selection pada Algoritma Naïve Bayes dalam Kasus Keyword Extraction." *CITEE 2020*, 2020.

[6] S. R. Azizah, R. Herteno, A. Farmadi, D. Kartini, and I. Budiman, "Kombinasi Seleksi Fitur Berbasis Filter dan Wrapper Menggunakan Naïve Bayers pada Klasifikasi Penyakit Jantung", *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 6, pp. 1361-1368, 2023, doi: 10.25126/jtiik.2023107467.

[7] G. V. Gopal and G. R. M. Babu, "An ensemble feature selection approach using hybrid kernel based SVM for network intrusion detection system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, pp. 558–565, Jul. 2021, doi: 10.11591/ijeecs.v23.i1.pp558-565.

[8] F. Abdusyukur, "Penerapan Algoritma Support Vector Machine (SVM) untuk Klasifikasi Pencemaran Nama Baik di Media Sosial Twitter," *KOMPUTA : Jurnal Ilmiah Komputer dan Informatika*, vol. 12, no. 1, 2023.

[9] S. Nurhasanah Nugraha, T. Rivanie, S. Rahayu, W. Gata, R. Pebrianto, and S. Nusa Mandiri, "Sentimen Analisis Penerapan Social Distancing Menggunakan Feature Selection Pada Algoritma Support Vector Machine," *Jurnal Teknik Komputer AMIK BSI*, vol. VI, no. 2, 2020, doi: 10.31294/jtk.v4i2.

[10] A. Hendrawan, L. M. Huizen, A. Praba, R. Pinem, and A. Wicaksana, "Implementasi Pemilihan Fitur Metode Wrapper dan Embedded dalam Prediksi Ketepatan Kelulusan Mahasiswa". *Seminar Nasional Penelitian dan Pengabdian Kepada Masyarakat (SNPPKM)*, 2021.

[11] D. Setiawan, "Komparasi Teknik Feature Selection Dalam Klasifikasi Serangan IoT Menggunakan Algoritma Decision Tree," *Jurnal Media Informatika Budidarma*, vol. 8, no. 1, pp. 83-93, 2024, doi: 10.30865/mib.v8i1.6987.

[12] H. Alshamlan, S. Omar, R. Aljurayyad, and R. Alabduljabbar, "Identifying Effective Feature Selection Methods for Alzheimer's Disease Biomarker Gene Detection Using Machine Learning," *Diagnostics*, vol. 13, no. 10, May 2023, doi: 10.3390/diagnostics13101771.

[13] D. Setiawan, "Komparasi Teknik Feature Selection Dalam Klasifikasi Serangan IoT Menggunakan Algoritma Decision Tree," *Jurnal Media Informatika Budidarma*, vol. 8, no. 1, pp. 83-93, 2024, doi: 10.30865/mib.v8i1.6987.

[14] R. Wibowo and H. Indriyawati, "Top-k-Feature Selection untuk Deteksi Penyakit Hepatitis Menggunakan Algortime Naïve Bayes 1 Fakultas Teknologi Informasi dan Komunikasi." *Jurnal Buana Informatika*, vol. 11, no. 1, pp. 1-9, 2020.

[15] I. Maulida, A. Suyatno, and H. R. Hatta, "Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain," *Jurnal SIFO Mikroskil*, vol. 17, no. 2, pp. 249–258, Oct. 2016, doi: 10.55601/jsm.v17i2.379.

[16] Y. Reswan, "DESAIN APLIKASI PENGENALAN POLA TANDA TANGAN MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM)," *Jurnal Media Infotama*, vol. 17, no. 1, pp. 8-12, 2021.

[17] A. Tzotsos and D. Argialas, "Support vector machine classification for object-based image analysis," *Lecture Notes in Geoinformation and Cartography*, vol. 0, no. 9783540770572, pp. 663–677, 2008, doi: 10.1007/978-3-540-77058-9_36.

[18] A. Budianita and F. I. Pratama, "Penerapan Algoritma Klasifikasi Dengan Fitur Seleksi Weight By Information Gain Pada Pemodelan Prediksi Kelulusan Mahasiswa," *Infotekmesin*, vol. 11, no. 2, pp. 80–86, Aug. 2020, doi: 10.35970/infotekmesin.v11i2.255.

[19] J. Zhou, J. Huang, and L. Zheng, "Application of an Improved FSVM Algorithm in Breast Cancer Diagnosis," *Journal of Applied Mathematics and Computation*, vol. 4, no. 2, pp. 18–25, May 2020, doi: 10.26855/jamc.2020.06.002.

[20] I. Santoso, Windu Gata, and Atik Budi Paryanti, "Penggunaan Feature Selection di Algoritma Support Vector Machine untuk Sentimen Analisis Komisi Pemilihan Umum," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 3, pp. 364–370, Dec. 2019, doi: 10.29207/resti.v3i3.1084.