Penerapan *Feature Selection* untuk Prediksi Lama Studi Mahasiswa

I Made Budi Adnyana
STIKOM Bali
Jln Raya Puputan No 86, Renon, Denpasar, Telp. (0361) 244445
e-mail: budi.adnyana@stikom-bali.ac.id

Abstrak

Kelulusan tepat waktu merupakan permasalahan yang sering dialami oleh institusi perguruan tinggi. Beberapa faktor dapat menjadi penyebabnya. Pada penelitian ini diterapkan teknik data mining feature selection untuk menganalisis pengaruh mata kuliah terhadap lama studi mahasiswa. Teknik feature selection yang digunakan yaitu Correlation Based, Information Gain Based, dan Learner Based. Akurasi dari masing-masing metode seleksi fitur diukur menggunakan algoritma klasifikasi Naïve Bayes. Hasil uji coba menunjukkan penerapan teknik feature selection mampu meningkatkan akurasi klasifikasi dari algoritma Naïve Bayes. Hasil uji coba terhadap dataset nilai mahasiswa menunjukkan teknik Learner Based menggunakan model Wrapper menghasilkan akurasi paling tinggi. Akurasi paling rendah diperoleh menggunakan teknik Information Gain.

Kata kunci: *Klasifikasi, Prediksi, Feature Selection.*

Abstract

Student graduation on time is the problems that often experienced by college institution. A number of factors can be a cause. In this paper proposed feature selection data mining techniques for evaluating the impacts of course subjects over student graduation time. Feature selection techniques that used is Correlation Based, Information Gain Based, and Learner Based. Accuracy of each feature selection methods measured using Naïve Bayes classification algorithm. Experiment result that implementation of feature selection techniques can improve classification accuracy of Naïve Bayes algorithms. Highest accuracy was obtained using Learner Based techniques using Wrapper model. Lowest accuracy was obtained using Information Gain Based techniques.

Keywords*: Classification, Prediction, Feature Selection.*

1. Pendahuluan

Salah satu faktor yang menentukan kualitas perguruan tinggi adalah persentase kemampuan mahasiswa untuk menyelesaikan studi tepat waktu. Saat ini, masalah kegagalan studi mahasiswa dan faktor-faktor penyebabnya menjadi topik yang menarik untuk diteliti [1]. Salah satu faktor yang mempengaruhi kelulusan tepat waktu mahasiswa adalah nilai mata kuliah yang telah ditempuh oleh mahasiswa itu sendiri. Perguruan tinggi perlu menganalisis mata kuliah mana yang mempunyai pengaruh besar terhadap kelulusan mahasiswa dan mata kuliah mana yang kurang berpengaruh sehingga dapat diketahui faktor-faktor penyebab kegagalan mahasiswa dalam hal lulus tepat waktu. Beberapa penyebab adanya perbedaan pengaruh mata kuliah terhadap kelulusan mahasiswa di antaranya adalah tingkat kesulitan mata kuliah, kesesuaian mata kuliah terhadap topik skripsi yang diambil mahasiswa dan faktor lainnya.

Database perguruan tinggi menyimpan data akademik mahasiswa, khususnya data mengenai mata kuliah, nilai mahasiswa, dan data kelulusan mahasiswa. Data tersebut apabila digali dan dianalisis dengan tepat maka akan dapat diketahui pola atau korelasinya sehingga dapat digunakan untuk mengambil keputusan. Serangkaian proses mendapatkan pengetahuan atau pola dari kumpulan data disebut dengan *data mining*. *Data mining* memecahkan masalah dengan menganalisis data yang telah ada dalam *database*. *Data mining* dalam dunia pendidikan dikenal dengan *Educational Data Mining* [2]. EDM mengembangkan metode untuk menggali data pendidikan dan menggunakan metode tersebut untuk lebih memahami mahasiswa. EDM dapat membantu pendidik untuk menganalisis cara belajar, mendeteksi mahasiswa yang memerlukan dukungan dan memprediksi kinerja mahasiswa.

Teknik *data mining* yang digunakan pada penelitian ini teknik *feature selection.* Seleksi fitur adalah salah satu teknik *data mining* yang umum digunakan pada tahapan *pre-processing*. Teknik ini digunakan untuk mengurangi kompleksitas atribut yang akan dikelola pada processing dan analisis. Teknik ini dilakukan untuk mengetahui *subset* fitur yang paling signifikan dari data set nilai mahasiswa. Pemilihan fitur sering digunakan untuk pengurangan dimensi model. Pemilihan fitur membantu mengurangi fitur domain, menghilangkan fitur yang berlebihan. Dengan cara ini akan membantu mempercepat proses pembelajaran/pemodelan [3]. Pada penelitian ini digunakan tiga teknik seleksi fitur, yaitu *Correlation Based*, *Information Gain Based*, dan *Learner Based*.

Metode seleksi fitur pernah diterapkan dalam bidang EDM pada penelitian yang berjudul "Fitur Seleksi *Forward Selection* untuk Menentukan Atribut yang Berpengaruh pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma *Naïve Bayes*". Pada penelitian ini digunakan metode *Naïve Bayes* dengan memanfaatkan fungsi seleksi fitur dari *Forward Selection* untuk pemilihan atribut data dengan karakteristik data itu sendiri, dan meningkatkan ketepatan klasifikasi *Naïve Bayes*. *Forward Selection* berbasis *Naïve Bayes* lebih akurat dan efektif dalam mengklasifikasikan status kelulusan mahasiswa dengan hasil akurasi 97,14% dan termasuk dalam kategori "*excellent classification*" dan memperoleh atribut yang berpengaruh yaitu: status pekerjaan dan IPK semester 4 [4].

Pada penelitian ini teknik seleksi fitur berguna untuk memilih atribut mata kuliah mana saja yang memiliki pengaruh yang relevan terhadap lama studi mahasiswa dan mengesampingkan mata kuliah yang kurang berpengaruh. Hasil dari proses ini diharapkan dapat membantu perguruan tinggi dalam membuat kebijakan akademis agar dapat mengoptimalkan tingkat kelulusan mahasiswa pada tahun-tahun berikutnya.

2. Metode Penelitian

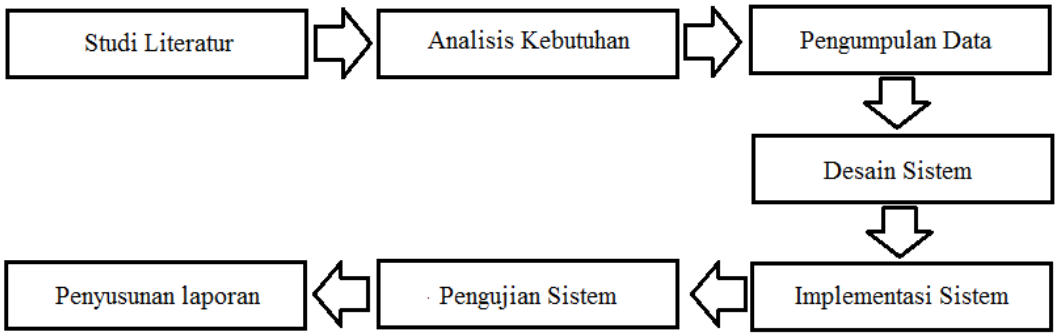
Model konseptual pada penelitian ini berdasarkan teknik *data mining* yang diterapkan pada data set nilai mahasiswa. Penelitian ini dilakukan untuk mengetahui mata kuliah apa saja yang mempunyai pengaruh relevan terhadap lama studi mahasiswa. Proses *data mining* yang digunakan pada penelitian ini menggunakan *feature selection* dengan tiga buah model yang populer digunakan, yaitu *Correlation Based*, *Information Gain Based*, dan *Learner Based*.

Data set yang digunakan pada penelitian ini adalah sekumpulan data nilai mahasiswa STIKOM Bali program studi Sistem Informasi. Data nilai yang digunakan yaitu nilai dari semester I sampai dengan semester VII (tidak termasuk mata kuliah konsentrasi). Data set terdiri dari 40 atribut dan 1246 *row*. Data mahasiswa yang digunakan adalah yang telah lulus 5 tahun terakhir dan bukan termasuk mahasiswa transfer. Contoh format data set nilai mahasiswa yang digunakan ditunjukkan pada Tabel 1.

Tabel 1. Format data set nilai mahasiswa.

ID Mahasiswa	Nilai Matakuliah #1	Nilai Matakuliah #2	...	Nilai Matakuliah #N	Lama Studi

Dalam penelitian ini penulis memerlukan waktu 9 bulan untuk menyelesaikan penelitian ini dan untuk tempat penelitian penulis akan melakukan penelitian di STMIK STIKOM Bali.

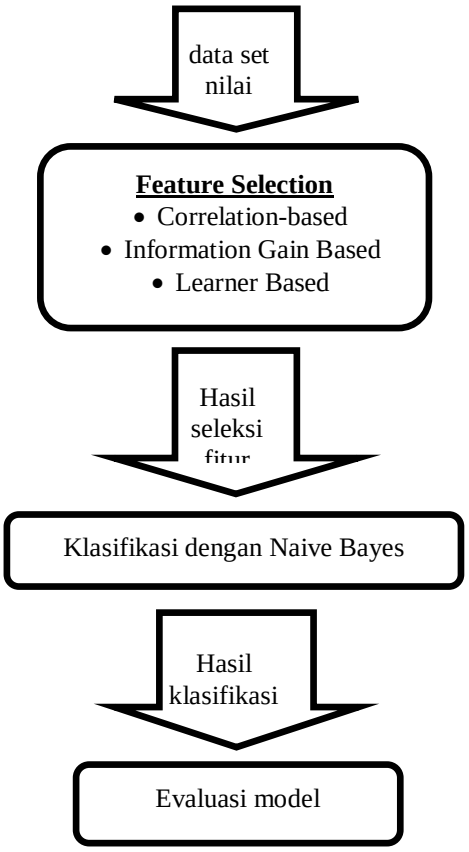


Gambar 1. Sistematika penelitian.

Studi literatur dilakukan dengan pencarian referensi diperoleh melalui internet, buku, karya tulis, dan media informasi lainnya yang berhubungan dengan objek penelitian, yaitu yang berkaitan dengan *data*

mining, *feature selection*, dan *classification*. Analisis kebutuhan dilakukan analisis terhadap kebutuhan-kebutuhan sistem yang akan dirancang agar nantinya dapat memenuhi target yang telah ditentukan. Hasil dari analisis kebutuhan sistem ini kemudian akan digunakan dalam melakukan proses desain sistem. Setelah mendapat sumber referensi, maka tahap selanjutnya adalah mengumpulkan data yang sesuai dengan objek penelitian. Data diperoleh dengan melakukan wawancara dan observasi di STIKOM Bali, khususnya pada divisi Akademik. Data yang dikumpulkan berupa data mahasiswa, data kelulusan, dan data nilai mata kuliah yang akan digunakan dalam proses *data mining*.

Berdasarkan hasil analisis kebutuhan sistem, maka tahap selanjutnya adalah mendesain sistem yang dilakukan dengan perancangan alur metode *feature selection* untuk diterapkan pada permasalahan analisis mata kuliah yang berpengaruh terhadap kelulusan mahasiswa. Metode *feature selection* yang akan digunakan pada penelitian ini adalah teknik *Correlation Based*, *Information Gain Based*, dan *Learner Based*. Setelah dirancang, langkah selanjutnya adalah melakukan implementasi dan uji coba menggunakan aplikasi WEKA dengan memanfaatkan fungsi "*Select Attribute*". Evaluasi dilakukan dengan menggunakan algoritma *Naïve Bayes* untuk melakukan klasifikasi berdasarkan atribut yang terpilih pada tahap seleksi fitur. Nilai yang diukur berdasarkan *output* dari aplikasi WEKA adalah persentase *Correctly Classified Instances*, *Incorrectly Classified Instances*, *Mean absolute error*, dan *Relative absolute error*.



Gambar 2. Alur penerapan feature selection

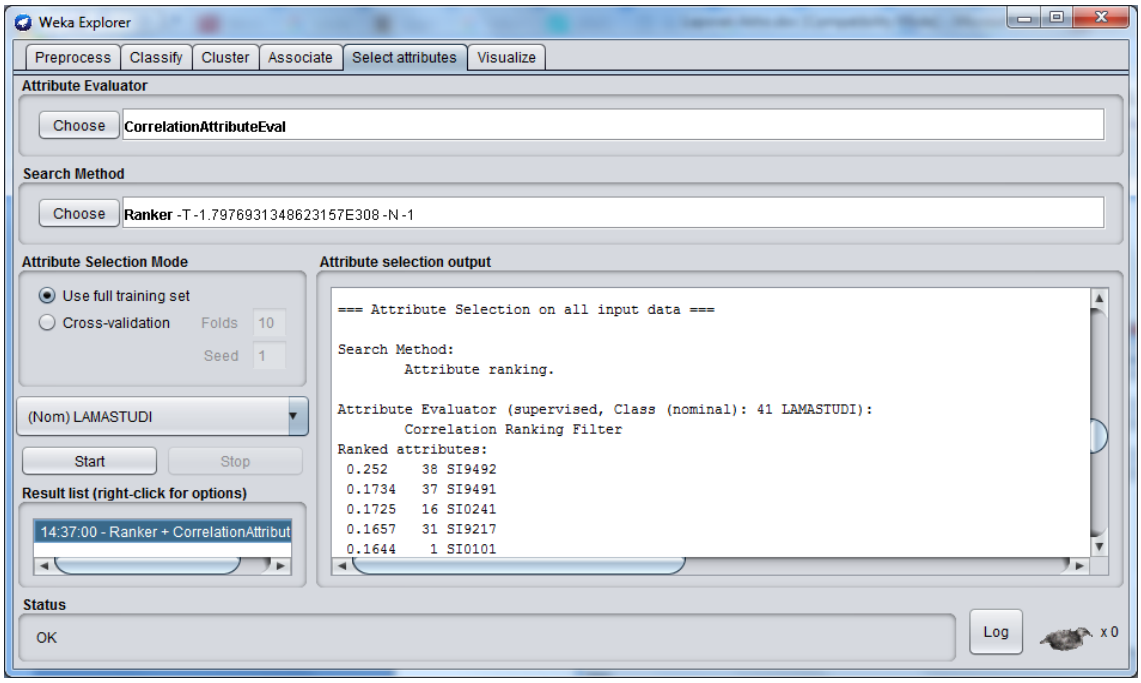
3. Hasil dan Pembahasan

Tahapan pertama dari penelitian ini adalah melakukan proses *feature selection* atau seleksi fitur. Seleksi fitur adalah salah satu teknik *data mining* yang umum digunakan pada tahapan *pre-processing*. Teknik ini digunakan untuk mengurangi kompleksitas atribut yang akan dikelola pada *processing* dan analisis. Teknik ini dilakukan untuk mengetahui *subset* fitur yang paling signifikan dari data set nilai mahasiswa. Pemilihan fitur sering digunakan untuk pengurangan dimensi model. Pemilihan fitur membantu mengurangi fitur domain, menghilangkan fitur yang berlebihan. Dengan cara ini akan membantu mempercepat proses pembelajaran/pemodelan [3]. Pada penelitian ini digunakan tiga teknik seleksi fitur, yaitu *Correlation Based*, *Information Gain Based*, dan *Learner Based*.

Teknik *Correlation Based* yang populer digunakan untuk melakukan seleksi terhadap fitur yang paling relevan dalam data set adalah teknik *Correlation Feature Selection* (CFS). Teknik ini menghitung korelasi antara masing-masing atribut dan variabel *output*, lalu memilih atribut yang mempunyai nilai korelasi menengah ke atas (mendekati 1) dan membuang atribut yang memiliki nilai korelasi rendah (mendekati 0). CFS menggunakan kinerja prediktif dan inter-korelasi fitur untuk mencari sekumpulan fitur yang bagus. Eksperimen yang dilakukan pada data set *discrete* dan *continuous* menunjukkan CFS dapat menurunkan dimensi data set secara drastis dengan tetap menjaga atau meningkatkan kinerja dari *learning algorithm* [5].

Teknik populer lainnya untuk seleksi fitur adalah *Information Gain Feature Selection*. Teknik ini menghitung *information gain* atau *entropy* dari masing-masing atribut berdasarkan variabel output. Nilai *output* bervariasi antara 0 (informasi minimum) sampai dengan 1 (informasi maksimum). Atribut-atribut yang memberikan lebih banyak informasi akan memiliki nilai *information* gain yang lebih tinggi dan dapat dipilih, sedangkan atribut yang kurang memberikan informasi akan mempunyai nilai gain *information* yang rendah dan dapat dibuang.

Teknik ketiga yang digunakan pada penelitian ini adalah *Learner Based Feature Selection*. Teknik ini mengevaluasi kinerja algoritma pada data set dengan *subset* atribut yang berbeda-beda. *Subset* yang menghasilkan kinerja terbaik akan dijadikan *subset* terpilih. Uji coba proses seleksi fitur pada penelitian ini menggunakan aplikasi WEKA dengan memanfaatkan fasilitas "*Select Attribute*". *Attribute evaluator* yang digunakan adalah *CorrelationAttributEval* dengan *Ranker Search Method*, *InfoGainAttributEval* dengan *Ranker Search Method*, dan *WrapperSubsetEval* dengan *GreedyStepWise Search Method*. Masing-masing menggunakan 10 *folds cross-validation*.



Gambar 3. Seleksi fitur dengan aplikasi WEKA

Berdasarkan hasil uji coba diketahui bahwa dari 41 fitur yang terdapat pada data uji direduksi menjadi 11 fitur dengan menggunakan teknik *CorrelationAttributEval* dan *InfoGainAttributEval*, 5 buah fitur menggunakan teknik *WrapperSubsetEval* sebagaimana ditunjukkan pada Tabel 2. Hasil seleksi fitur ini akan digunakan untuk eksperimen selanjutnya, yaitu proses klasifikasi dengan menggunakan beberapa metode *Naïve Bayes*.

Tabel 2 Fitur terpilih menggunakan teknik *feature selection*.

Metode Seleksi Fitur	Fitur Terpilih
CorrelationAttributEval	SI9492, SI9491, SI0241, SI9217, SI0101, SI9215, SI0211, SI0213, SI9203, SI0212, SI9209
InfoGainAttributEval	SI9492, SI9217, SI0213, SI9491, SI0201, SI9215, SI0241, SI0211, SI0303, SI9501, SI0302
WrapperSubsetEval	SI0104, SI0214, SI0218, SI9106, SI9491

Hasil seleksi fitur ini akan diuji pada algoritma klasifikasi *Naïve Bayes* untuk mengetahui kinerja dan akurasi yang dihasilkan dari masing-masing teknik seleksi fitur. Teknik validasi yang digunakan adalah 10 *folds cross-validation*. Hasil uji coba proses klasifikasi dengan *Naïve Bayes* dapat dilihat pada Tabel 3.

Tabel 3. Hasil uji klasifikasi dengan *Naïve Bayes* (NB).

	NB	NB + Correlation Attribut Eval	NB + Info Gain Attribut Eval	NB + Wrapper Subset Eval
Correctly Classified Instances	69.95%	75.74%	73.65%	77.83%
Incorrectly Classified Instances	30.05%	24.26%	26.35%	22.17%
Kappa statistic	0.3785	0.4199	0.3913	0.3672
Mean absolute error	0.2066	0.1866	0.196	0.2173
Root mean squared error	0.4168	0.3616	0.3701	0.3406
Relative absolute error	74.08%	66.91%	70.26%	77.91%
Root relative squared error	111.71%	96.90%	99.18%	91.27%

Berdasarkan hasil uji coba klasifikasi menggunakan *Naïve Bayes* pada Tabel 5.5 dapat diketahui bahwa penerapan *feature selection* secara umum meningkatkan akurasi dari algoritma klasifikasi. Akurasi tertinggi diperoleh dengan menerapkan teknik *Wrapper* memperoleh akurasi sebesar 77.83%. Akurasi terendah diperoleh dengan teknik *Information Gain* sebesar 73.65%. Tingkat kesalahan pada tiap-tiap metode dalam eksperimen adalah cukup rendah dilihat dari angka *Mean Absolute Error* yang berkisar antara 0.1 – 0.2. Paling tinggi adalah teknik *Wrapper* yaitu 0.2173 dan paling rendah adalah teknik *Correlation* yaitu 0.188.

4. Kesimpulan

Penerapan *feature selection* secara umum dapat meningkatkan akurasi dari algoritma klasifikasi *Naïve Bayes*. Akurasi tertinggi diperoleh dengan menerapkan teknik *Wrapper* memperoleh akurasi sebesar 77.83%. Akurasi terendah diperoleh dengan teknik *Information Gain* sebesar 73.65%. Tingkat kesalahan pada tiap-tiap metode dalam eksperimen adalah cukup rendah dilihat dari angka *Mean Absolute Error* yang berkisar antara 0.1 – 0.2. Paling tinggi adalah teknik *Wrapper* yaitu 0.2173 dan paling rendah adalah teknik *Correlation* yaitu 0.188.

Daftar Pustaka

[1] C. Marquez-Vera, C. Romero, and S. Ventura, "Predicting School Failure Using Data Mining," *EDM*, 2011.
[2] R.S.J.D. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.
[3] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*. Waltham, MA: Elsevier, 2012.
[4] M.F. Nugroho and S. Wibowo, "Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naïve Bayes," *Jurnal Informatika UPGRIS*, vol. 3, no. 1, pp. 63–70, 2017.
[5] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," in *17th Int. Conf. on Machine Learning*, Stanford, 2000, pp. 359–366.