Implementasi Pemilihan Fitur Metode *Wrapper* dan *Embedded* dalam Prediksi Ketepatan Kelulusan Mahasiswa

Aria Hendrawan[1],[*], Lenny Margaretta Huizen[2], Agusta Praba Ristadi Pinem[3], Dinar Anggit Wicaksana[4]
[1,2,3,4] Program Studi Teknik Informatika, Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang, Semarang, 50196, Indonesia
[1]ariahendrawan@usm.ac.id*; [2]lenny@usm.ac.id, [3]pinem@usm.ac.id, [4]dinar@usm.ac.id

ABSTRACT

In the current era of Big Data, data sets have very diverse structures and high dimensions. This high dimension has independent attributes, dependent attributes, irrelevant and useless data which is a source of problems in getting the best accuracy in the process of applying the prediction classification method. Feature selection in each prediction classification case is very important because it is a component in machine learning and research workflows. In this study, the selection of data attribute features using wrapper and embedded methods was carried out on a dataset of students who had completed their bachelor's degree in college as many as 146 data which had 13 data attributes and 1 label. This feature selection method is carried out to obtain the appropriate data attributes which will later be used in the predictive classification process of students who have graduated on time or not. The results showed that using the Random Forest Classification algorithm, the accuracy value was 73% and there was an increase in accuracy after using the wrapper and embedded feature selection method by 100%.

Keywords: ***feature selection, classification, random forest, accuration, wrapper method, embedded method***

ABSTRAK

Era Big Data saat ini, kumpulan data memiliki struktur yang sangat beragam serta dimensi yang tinggi. Dimensi tinggi ini memiliki independen atribut, dependen atribut, data yang tidak relevan dan tidak berguna yang menjadi sumber masalah dalam mendapatkan akurasi yang terbaik di proses penerapan metode klasifikasi prediksi. Pemilihan fitur pada setiap kasus klasifikasi prediksi sangat penting untuk dilakukan karena merupakan komponen dalam pembelajaran mesin dan alur kerja penelitian. Dalam penelitian ini, pemilihan fitur atribut data menggunakan metode wrapper dan embedded yang dilakukan pada dataset mahasiswa yang telah menyelesaikan sarjana di perguruan tinggi sebanyak 146 data yang memiliki 13 atribut data dan 1 label. Metode pemilihan fitur ini dilakukan untuk mendapatkan atribut data yang sesuai yang nantinya digunakan dalam proses klasifikasi prediksi mahasiswa yang telah lulus tepat waktu atau tidak. Hasil penelitian menunjukkan dengan menggunakan algoritma Random Forest Classification, nilai akurasinya sebesar 73% dan terjadi peningkatan akurasi setelah menggunakan metode pemilihan fitur wrapper dan embedded sebesar 100%.

Kata kunci : pemilihan fitur, klasifikasi, random forest, akurasi, metode wrapper, metode embedded

PENDAHULUAN

Pembelajaran mesin membutuhkan proses analisis dalam kumpulan data yang besar untuk melatih dan menguji data tersebut. Hampir semua kumpulan data berisi banyak fitur, kurang lebih setengah dari fitur tersebut penting untuk memberi informasi dalam proses analisis. Oleh karena itu, untuk meminimalkan waktu pelatihan dan evaluasi, perlu memasukkan algoritma dengan mengambil fitur yang dibutuhkan. Pemilihan fitur terbaik diperlukan untuk mengaktifkan algoritma pembelajaran mesin dalam melatih dataset lebih cepat, membuat model yang mudah diinterpretasikan, waktu kompleksitas lebih sedikit, kinerja algoritma menjadi lebih efisien, memperoleh akurasi tinggi dan menghindari over fitting. Ada berbagai aplikasi untuk mengurangi dimensi kumpulan data berdimensi tinggi sebelum membuat model analisis apa pun (Thanoon et al., 2019)

Pemilihan fitur adalah salah satu metode populer yang digunakan untuk meminimalkan dimensi dan kompleksitas komputasi kumpulan data. Metode pemilihan fitur saat ini adalah metode wrapper, filter, hybrid dan embedded (tertanam). Metode wrapper menggunakan algoritma pembelajaran saat ini untuk mengevaluasi fitur subset (Ebersbach et al., 2016). Metode seleksi filter menjalankan proses seleksi independen tanpa algoritma pembelajaran apapun (Sánchez-Maroño et al., 2007). Metode hybrid diperoleh dengan menggabungkan metode filter dan wrapper. Di Metode Tertanam (Embedded Method), Pemilihan fitur diintegrasikan sebagai bagian dari pembelajaran mesin. Metode tertanam tersebut menggabungkan kualitas metode filter dan wrapper. Di mana implementasinya algoritma memiliki metode pemilihan fitur sendiri di dalamnya. Sebuah algoritma pembelajaran mengambil keuntungan dari proses seleksi variabelnya sendiri dan melakukan seleksi fitur dan klasifikasi/regresi pada saat yang bersamaan. Teknik tertanam atau embedded method yang paling umum adalah algoritma pohon seperti Random Forest, Extra Tree dan sebagainya. (Lal Thomas Navin and Chap, 2006)

Efektivitas pembelajaran mesin dalam pengenalan pola, klasifikasi, prediksi dan temu kembali informasi dapat diperoleh dengan mengukur akurasi, recall dan presisi. Akurasi adalah persentase prediksi yang benar. Recall adalah persentase kasus positif yang sebenarnya. Ini adalah tingkatan antara item yang relevan untuk seluruh jumlah item yang relevan. Presisi adalah persentase kasus prediksi positif. Ini adalah tingkatan antara item yang relevan dengan item yang diambil. Ini berarti bagaimana kedua pengukuran itu dekat satu sama lain. Dalam tes, tingkat harmonik antara presisi dan recall tes disebut F-score. Kinerja classifier atau model klasifikasi dapat divisualisasikan dengan matriks konfusi sebagai tabel prediksi dan kelas yang sebenarnya.

Di sebagian besar kumpulan data, banyak fitur yang tidak relevan, karena fitur ini dapat memengaruhi akurasi sistem dan memaksimalkan biaya dan waktu komputasi. Oleh karena itu, penggunaan metode pemilihan fitur menjadi penting untuk menyederhanakan, mempercepat dan meningkatkan akurasi model sistem klasifikasi (Chandrashekar & Sahin, 2014). Banyak peneliti membahas metode wrapper sebagai metode pemilihan menggunakan teknik pembelajaran mesin untuk mempelajari fitur-fitur yang dipilih (Ebersbach et al., 2016). Metode wrapper telah digunakan dalam konteks yang luas, seperti kumpulan data bioinformatika. Konsep dan algoritma pemilihan fitur digunakan untuk menunjukkan aktivitas algoritma tersebut, dan berisi platform pemersatu yang diusulkan sebagai langkah perantara. Mereka mendemonstrasikan penggunaan seleksi fitur dalam penambangan data (Liang & Hu, 2015a). Metode pemilihan fitur diklasifikasikan menjadi pembungkus, metode filter, dan metode berbasis tertanam. Metode-metode ini menggunakan metode filter untuk melakukan praproses data dan mendapatkan peringkat setiap fitur dan menerapkan fitur peringkat tinggi ke sebuah prediktor. Sedangkan pada

metode pemilihan wrapper, fitur yang dipilih adalah fitur yang memberikan tingkat prediksi tertinggi (Liu et al., 2015). Perbandingan antara pemilihan fitur wrapper dan filter dilakukan oleh (Lal Thomas Navin and Chapelle, 2006; Sánchez-Maroño et al., 2007) menemukan bahwa wrapper dengan proses komputasi kritis lebih baik daripada sebuah filter. Peneliti lain menunjukkan bahwa menominasikan banyak fitur teratas yang memenuhi syarat untuk kemajuan pencapaian daripada mempertimbangkan satu fitur (Chandrashekar & Sahin, 2014). Klasifikasi pada pengklasifikasi tunggal atau kumpulan pengklasifikasi telah dieksplorasi, seperti membangun pohon keputusan dengan menghitung perolehan informasi (Made et al., 2016.). Pada penelitian yang dilakukan oleh (Green Arther Sandag, 2020) menyatakan bahwa Algoritma Random Forest memiliki nilai accuracy yang lebih baik dari pada algoritma Decision Tree, Logic Regression dan K-NN dengan nilai accuracy sebesar 86,27%.

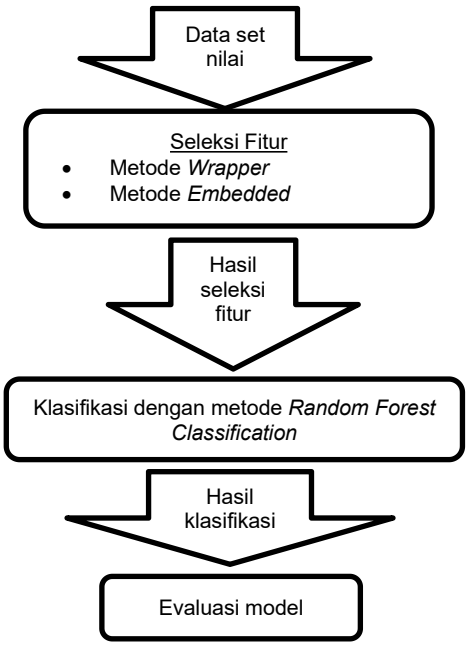
Pada penelitian (Lal Thomas Navin and Chapelle, 2006; Xu et al., 2017) peneliti berfokus pada pemilihan fitur yang sesuai untuk melakukan langkah-langkah serangan dengan memberi peringkat beberapa fitur menurut entropi perolehan informasinya yang tinggi. Dengan metode pemilihan fitur, kita dapat menemukan cara untuk memilih bagian penting dari kumpulan data tanpa kehilangan informasi apa pun. Hasilnya menunjukkan aktivitas metode pemilihan yang disebutkan di atas untuk mengurangi dimensi kumpulan data dan mengurangi jumlah fitur yang dipilih menjadi setengah dari jumlah fitur sebenarnya dengan akurasi tinggi dan operasi komputasi yang lebih sedikit.

Tujuan dari penelitian ini adalah fokus pada pemilihan fitur penting yang memiliki informasi efektif dari keseluruhan dataset. Masalah penelitian terletak pada menemukan faktor-faktor mahasiswa yang lulus tidak dengan tepat waktu atau dengan tepat waktu, selanjutnya mencoba untuk membuat prediksi klasifikasi seakurat mungkin menggunakan teori-teori dari metode wrapper dan metode embedded untuk membuktikan manfaat menggunakan teori-teori ini.

METODE PENELITIAN

Penelitian ini dilakukan untuk mengetahui faktor – faktor apa saja yang mempengaruhi ketepatan lulus mahasiswa. Proses pengolahan data yang digunakan pada penelitian ini menggunakan feature selection dengan dua buah model yang populer digunakan, yaitu metode wrapper dan metode embedded.

Data set yang digunakan pada penelitian ini adalah dataset yang diambil dari 146 mahasiswa yang lulus dengan tepat waktu atau tidak tepat waktu, di mana tepat waktu disini memiliki arti lulus dalam waktu 4 tahun untuk mendapatkan gelar Sarjana Komputer di Fakultas Teknologi Informasi dan Komunikasi Universitas Semarang. Dataset ini berupa hasil kuesioner yang terdapat 13 fitur, yaitu diantaranya pelayanan terhadap mahasiswa dari segi Tata Usaha, ketersediaan Sistem Informasi Akademik, Komunikasi antara mahasiswa dengan Dosen dalam penyelesaian bimbingan skripsi maupun perkuliahan, keadaan ekonomi keluarga mahasiswa, manajemen waktu prioritas mahasiswa. Dan terdapat 1 label yaitu mahasiswa lulus tepat waktu atau tidak tepat waktu. Alur penerapan seleksi fitur terdapat pada Gambar 1.



Dari data yang sudah didapatkan seperti dalam tabel 1. Proses selanjutnya adalah merubah data tersebut menjadi numerik, seperti yang ditunjukkan dalam Tabel 1.

Tabel 1. Hasil Transformasi Dataset

	f1	f2	f3	f4	...	f11	f12	f13	label
0	4	4	4	4	...	4	1	3	0
1	3	3	3	3	...	3	4	4	1
2	3	3	3	3	...	3	4	4	1
3	4	4	4	4	...	4	4	4	1
4	3	3	2	2	...	2	3	3	0
...	...	...	...	...	...	...	...	...	...
141	3	3	3	3	...	3	1	3	0
142	3	3	3	3	...	3	2	3	0
143	3	3	2	2	...	3	3	3	0
144	3	3	3	3	...	2	0	2	0
145	4	4	4	4	...	4	2	4	1

Terdapat 2 metode yang digunakan dalam penelitian ini, yaitu:

1. Metode Wrapper

Metode wrapper adalah metode yang dapat menemukan atribut terbaik dengan menggunakan algoritma dari Data Mining target. Metode ini mengevaluasi kombinasi karakteristik penggunaan algoritma pembelajaran. Tahapan pengelolaan diawali dengan penyeleksian subset, setelah itu dilanjutkan dengan mengevaluasi karakteristik menggunakan metode klasifikasi. Selanjutnya menambahkan fitur secara bertahap ke setiap tahap sebelum memilihi nilai yang terbaik. Tahap terakhir, mengintegrasikan ciri – ciri yang dipilih pada fase sebelumnya dengan ciri – ciri yang tersisa. Proses tersebut diulang sampai semua karakteristik model telah digunakan dan opsi terbaik dari kombinasi telah dipilih untuk memberikan nilai kinerja terbaik. (Anggraeni et al., 2021)

2. Metode Embedded

Metode embedded adalah metode yang dapat melakukan penyeleksian dengan fitur dan pembelajaran secara bersamaan. Penyeleksian dilakukan dengan menerapkan pembatasan sparsity ke dalam algoritma. (Liang & Hu, 2015b).

HASIL DAN PEMBAHASAN

Tahapan pertama dari penelitian ini adalah merubah dataset menjadi data numerik, selanjutnya membagi data tersebut dengan pembagian data training 80% dan data testing 20%.

Tahapan kedua selanjutnya adalah menerapkan metode wrapped dengan algoritma forward selection yang mengambil 5 top fitur dari 13 pemilihan fitur lainnya. Serta menerapkan metode embedded dengan algoritma random forest importance yang juga mengambil 5 top fitur dari 13 pemilihan fitur lainnya. Pada Tabel 2 merupakan hasil 5 top fitur yang terpilih.

Tabel 2. Hasil Pemilihan Fitur Terpilih

Metode Seleksi Fitur	Fitur Terpilih
Forward Selection	['f1', 'f3', 'f4', 'f7', 'f13']
Random Forest Importance	['f4', 'f5', 'f11', 'f12', 'f13']

Pembahasan penelitian yang dilakukan dalam hal ini adalah dari 13 fitur yang didapatkan dari dataset mahasiswa yang terdiri dari faktor – faktor ketepatan kelulusan, setelah dilakukan pemilihan fitur dengan menggunakan metode forward selection terdapat 5 buah fitur yang dipilih yakni f1, f3, f4, f7 dan f13. Di dalam Forward-Sequential Feature Selection prosedurnya melakukan secara pengulangan untuk menemukan fitur baru terbaik lalu ditambahkan ke kumpulan fitur yang dipilih. Dalam hal ini adalah memaksimalkan skor validasi silang ketika penaksir dilatih pada fitur tunggal ini. Setelah fitur pertama dipilih, selanjutnya mengulangi prosedur dengan menambahkan fitur baru ke kumpulan fitur yang dipilih. Prosedur berhenti ketika jumlah fitur terpilih yang diinginkan tercapai. Selanjutnya dengan menggunakan Random Forest Importance didapatkan 5 buah fitur terpilih dari rangking teratas yakni f13, f12, f11, f4, dan f5.

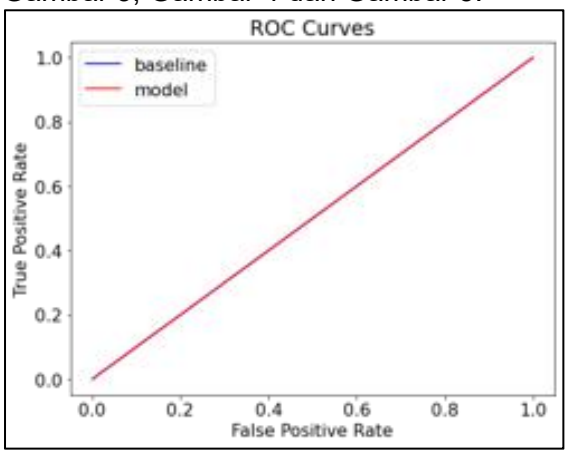
Hasil kedua metode tersebut juga harus diujikan nilai akurasi dengan menggunakan metode Random Forest

Classification untuk menunjukkan pentingnya pemilihan fitur seleksi ini. Hasil akurasi menggunakan *random forest classification* ditunjukkan pada Tabel 3.

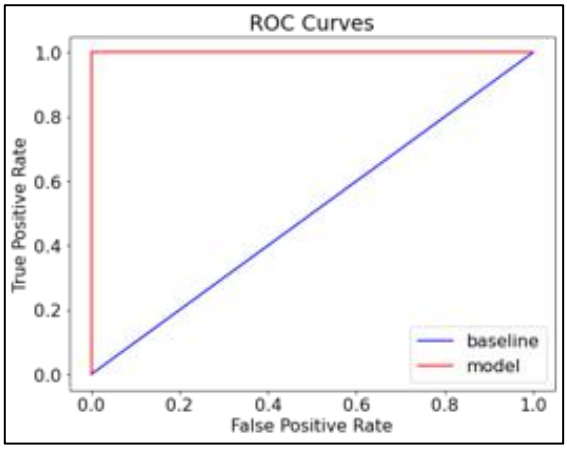
Tabel 3. Hasil Akurasi menggunakan Random Forest dan Random Forest dengan Pemilihan Fitur

	Random Forest (RF) Classification	RF + *Forward Selection (Wrapper Method)*	RF + *Random Forest Importance (Embedded Method)*
Akurasi	73%	100%	100%

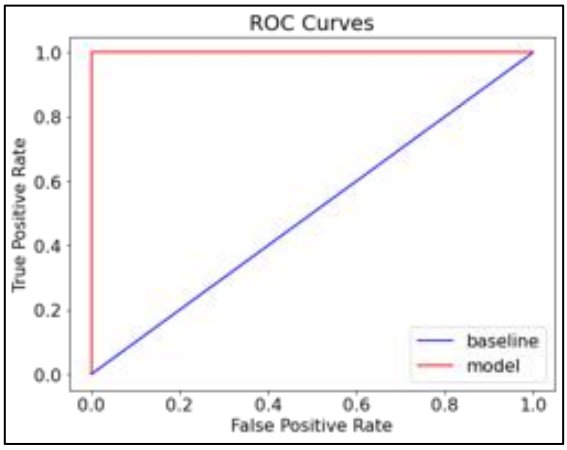
Hasil pengujian akurasi sebelum dilakukan pemilihan fitur terdapat nilai akurasi 73% sedangkan setelah dilakukan pemilihan fitur terdapat peningkatan menjadi 100% baik itu menggunakan wrapper method ataupun embedded method. Hal ini dilakukan analisa kembali dengan melihat kurva dan nilai AUC (Area Under The Curve) ROC (Receiver Operating Characteristics). Kurva AUC - ROC adalah pengukuran kinerja untuk masalah klasifikasi pada berbagai pengaturan ambang batas. ROC adalah kurva probabilitas dan AUC mewakili derajat atau ukuran keterpisahan. Ini memberitahu seberapa besar model mampu membedakan antar kelas. Semakin tinggi AUC, semakin baik model dalam memprediksi 0 kelas sebagai 0 dan 1 kelas sebagai 1. Dengan analogi, semakin tinggi AUC, semakin baik model dalam membedakan. Kurva ROC dalam hasil penelitian ini dapat ditunjukkan pada Gambar 3, Gambar 4 dan Gambar 5.



Gambar 3. Nilai AUC = 0.5



Gambar 4. Nilai AUC = 1



Gambar 5. Nilai AUC = 1

SIMPULAN

Dari hasil penelitian yang sudah dilakukan menggunakan dataset faktor – faktor ketepatan kelulusan mahasiswa sebanyak 146 mahasiswa dengan jumlah 13 fitur dan 1 label. Pemilihan fitur menggunakan metode wrapper memilih 5 buah fitur yaitu f1, f3, f4, f7, dan f13. Dan Metode embedded memilih 5 buah fitur yaitu f13, f12, f11, f4, dan f5. Hasil pengujian akurasi dengan menggunakan algoritma Random Forest Classification tanpa pemilihan fitur memiliki nilai 73% dan nilai AUC 0,5. Penggunaan algoritma Random Forest Classification dan pemilihan fitur baik itu metode wrapper dan metode embedded memiliki nilai akurasi sebesar 100% yang sama – sama memiliki nilai AUC sebesar 1.

SARAN

Dari penelitian ini, bisa dilakukan penelitian selanjutnya yaitu membandingkan dengan algoritma klasifikasi lainnya untuk mendapatkan analisa yang baik dan tepat dalam penggunaan pemilihan fitur.

DAFTAR PUSTAKA

Anggraeni, A. N., Mustofa, K., & Priyanta, S. (2021). Comparison of Filter and Wrapper Based Feature Selection Methods on Spam Comment Classification. 15(3), 245–254.

Chandrashekar, G., & Sahin, F. (2014). A Survey on Feature Selection Methods. Comput. Electr. Eng., 40(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Ebersbach, M., Herms, R., Lohr, C., & Eibl, M. (n.d.). Wrappers for Feature Subset Selection in CRF-based Clinical Information Extraction.

Lal Thomas Navin and Chapelle, O. and W. J. and E. A. (2006). Embedded Methods. In M. and G. S. and Z. L. A. Guyon Isabelle and Nikravesh (Ed.), Feature Extraction: Foundations and Applications (pp. 137–165). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-35488-8_6

Liang, M., & Hu, X. (2015a). Feature selection in supervised saliency prediction. IEEE Transactions on Cybernetics, 45(5), 900–912. https://doi.org/10.1109/TCYB.2014.2338893

Liang, M., & Hu, X. (2015b). Feature selection in supervised saliency prediction. IEEE Transactions on Cybernetics, 45(5), 900–912. https://doi.org/10.1109/TCYB.2014.2338893

Liu, Y., Tang, F., & Zeng, Z. (2015). Feature selection based on dependency margin. IEEE Transactions on Cybernetics, 45(6), 1209–1221. https://doi.org/10.1109/TCYB.2014.2347372

Made, I., Adnyana, B., Jln, S. B., & Puputan, R. (n.d.). Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa.

Sánchez-Maroño, N., Alonso-Betanzos, A., & Tombilla-Sanromán, M. (2007). Filter methods for feature selection - A comparative study. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4881 LNCS, 178–187. https://doi.org/10.1007/978-3-540-77226-2_19

Sandag, G. A. (2020). Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest. Cogito Smart Journal, 6(2), 167-178. http://cogito.unklab.ac.id/index.php/cogito/article/download/270/153.

Thanoon, M. A., Zedan, M. J. M., & Hameed, A. N. (2019). Feature Selection Based on Wrapper and Information Gain. NICST 2019 - 1st Al-Noor International Conference for Science and Technology, 32–37. https://doi.org/10.1109/NICST49484.2019.9043805

Xu, J., Tang, B., He, H., & Man, H. (2017). Semisupervised Feature Selection Based on Relevance and Redundancy Criteria. IEEE Transactions on Neural Networks and Learning Systems, 28(9), 1974–1984. https://doi.org/10.1109/TNNLS.2016.2562670