ANALISIS PERFORMA SELEKSI ATRIBUT UNTUK MENENTUKAN POTENSI MAHASISWA PUTUS STUDI

Vivi Nur Wijayaningrum[1], Ika Kusumaning Putri[2], Annisa Puspa Kirana[3], Muhammad Rizki Mubarok[4], Deatrisya Mirela Harahap[5], Berryl Radian Hamesha[6]

Jurusan Teknologi Informasi, Politeknik Negeri Malang
[1]vivinurw@polinema.ac.id, [2]ikakputri@polinema.ac.id, [3]puspakirana@polinema.ac.id,
[4]masbarok193@gmail.com, [5]mirelldee47@gmail.com, [6]berrylhamesha@gmail.com

Abstrak

Banyaknya kasus mahasiswa putus studi yang terjadi di sejumlah pendidikan tinggi menjadi perhatian khusus di berbagai negara. Efek yang ditimbulkan akibat masalah ini antara lain dapat menghambat perekonomian dan produktivitas di negara tersebut. Untuk mengatasi hal tersebut, beberapa algoritma telah digunakan untuk memprediksi potensi mahasiswa putus studi. Berbagai atribut data yang berkaitan dengan informasi mahasiswa, seperti data pribadi, riwayat akademik, dan latar belakang mahasiswa digunakan sebagai bahan pertimbangan mahasiswa tersebut berpotensi putus studi atau tidak. Namun, banyaknya atribut data yang digunakan pada proses prediksi memungkinkan terjadinya *overfitting*, menurunnya performa algoritma, dan menambah waktu komputasi. Pada penelitian ini, seleksi atribut data dilakukan dengan menggunakan Chi Square, Pearson Correlation Coefficient, dan Random Forest untuk selanjutnya dapat dilakukan prediksi potensi mahasiswa putus studi menggunakan Multi-Layer Perceptron. Hasil prediksi tersebut dievaluasi menggunakan akurasi dan F1-Score dengan menggunakan 9-*fold cross validation*. Melalui tiga skenario pengujian dengan menggunakan berbagai variasi banyaknya atribut data, terjadi peningkatan nilai akurasi dan F1-Score saat dilakukan seleksi atribut dengan nilai rata-rata di atas 0.8. Seleksi atribut menggunakan Chi Square menunjukkan akurasi tertinggi sebesar 0.834721 diperoleh saat penggunaan 16 atribut data, sementara F1-Score tertinggi diperoleh dari penggunaan 21 atribut data hasil penerapan Pearson Correlation Coefficient dan Random Forest dengan nilai sebesar 0.834568. Penerapan seleksi atribut untuk prediksi membuktikan bahwa nilai akurasi dan F1-Score lebih tinggi dibandingkan dengan tanpa prediksi, yaitu akurasi sebesar 0.809876 dan F1-Score sebesar 0.723456.

Kata kunci : akademik, chi square, pearson, pendidikan, prediksi, random forest

1. Pendahuluan

Pendidikan mempunyai peranan penting dalam kemajuan bangsa serta merupakan pusat demokrasi di berbagai negara (Roman, Davidse, Human-Hendricks, Butler-Kruger, & Sonn, 2022). Demi kesejahteraan warga negara, setiap warga negara setidaknya memperoleh pendidikan dasar. Lebih lanjut, pendidikan tinggi juga perlu ditempuh oleh setiap warga negara untuk mendapatkan kompetensi disiplin khusus dan keterampilan umum lainnya untuk menunjang kehidupan di masa mendatang (Chan, 2016).

Permasalahan yang saat ini sedang dialami oleh sejumlah pendidikan tinggi di berbagai negara adalah banyaknya mahasiswa yang putus studi (Wild & Heuling, 2020). Mahasiswa putus studi di perguruan tinggi dapat menghambat pertumbuhan ekonomi, daya saing, dan produktivitas yang berdampak tidak hanya pada mahasiswa tersebut, tetapi juga pada perguruan tinggi dan masyarakat (Realinho, Machado, Baptista, & Martins, 2022). Tingkat mahasiswa putus studi di tahun pertama mencapai 25% hingga 45% dengan berbagai latar belakang keputusan (Bäulke, Grunschel, & Dresel, 2021).

Beberapa penelitian sebelumnya telah dilakukan untuk memprediksi terjadinya kasus putus studi pada mahasiswa dengan menggunakan berbagai pendekatan. Berens, dkk, mendeteksi dini mahasiswa berisiko putus studi di Universitas Jerman menggunakan 21 atribut data meliputi data demografis dan prestasi akademik mahasiswa. Selanjutnya, algoritma AdaBoost untuk mengombinasikan analisis regresi, Jaringan Syaraf Tiruan, dan Decision Tree pada keseluruhan atribut data tersebut (Berens, Schneider, Görtz, Oster, & Burghoff, 2019).

Pada penelitian lainnya, Realinho, dkk, menggunakan 34 atribut data yang meliputi demografis, sosioekonomi, dan prestasi mahasiswa, baik saat awal pendaftaran sebagai mahasiswa, maupun di akhir semester pertama dan kedua perkuliahan di Institut Politeknik Portalegre, Portugal. Pemodelan menggunakan pembelajaran mesin dilakukan untuk memprediksi kinerja akademik dan potensi putus studi mahasiswa

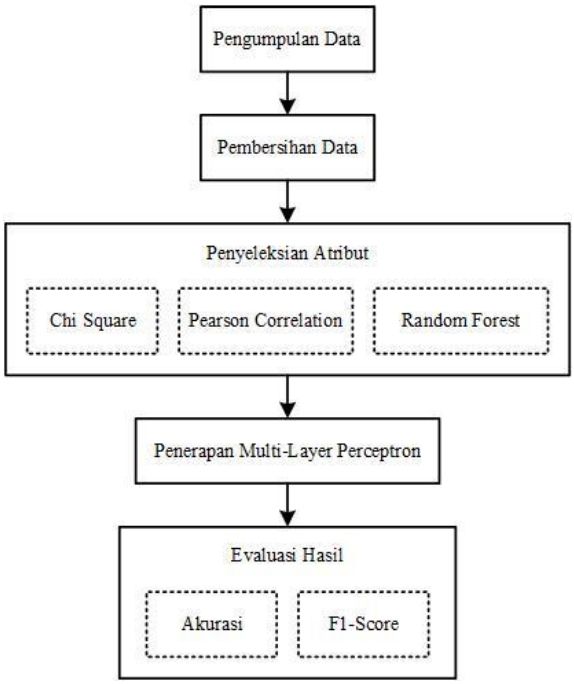
berdasarkan data mahasiswa tersebut (Realinho et al., 2022).

Sejumlah algoritma mesin pembelajaran seperti Jaringan Syaraf Tiruan, regresi logistik, Naive Bayes, Support Vector Machine, dan Random Forest digunakan oleh Dasi untuk memprediksi kemungkinan mahasiswa putus studi pada semester berikutnya. Dengan menggunakan data pribadi mahasiswa, riwayat akademik, dan status akademik mahasiswa dari semester berikutnya, kinerja dari setiap algoritma dibandingkan melalui metrik evaluasi. Data yang digunakan pada penelitian tersebut bersumber dari Kaggle dengan data set terdiri dari 261 mahasiswa dan 10 atribut data. Hasil penelitian menunjukkan bahwa Random Forest, regresi logistik, Decision Tree, dan Jaringan Syaraf Tiruan memberikan hasil yang paling akurat dibandingkan dua algoritma lainnya (Dasi & Kanakala, 2022).

Terdapat banyak faktor atau atribut yang memberikan dampak pada terjadinya mahasiswa putus studi, misalnya nilai pendaftaran masuk kuliah yang rendah, Indeks Prestasi Kumulatif (IPK) yang rendah, dan jenis mata kuliah yang ditempuh setiap semester (Wild & Heuling, 2020). Pada setiap penelitian yang telah dilakukan sebelumnya, jumlah dan jenis atribut yang digunakan untuk menentukan kemungkinan mahasiswa putus studi bervariasi sesuai dengan studi kasus pada setiap perguruan tinggi. Nyatanya, terlalu banyaknya atribut data yang digunakan dapat menurunkan performa algoritma dalam melakukan prediksi, menambah waktu komputasi, dan memungkinkan terjadinya *overfitting* (Anasanti, Hilyati, & Novtariany, 2022). Oleh karena itu, pada penelitian ini dilakukan perbandingan kinerja sejumlah algoritma dalam melakukan seleksi atribut untuk mendapatkan hasil prediksi yang akurat secara efisien. Dengan 23 atribut data yang telah digunakan pada penelitian sebelumnya untuk memprediksi potensi mahasiswa putus studi (Wijayaningrum, Kirana, Putri, & Satrio, 2022), penyeleksian atribut data menggunakan sejumlah metode yang terdiri dari Chi Square, Pearson Correlation, dan Random Forest untuk meningkatkan kinerja algoritma Multi-Layer Perceptron (MLP) dalam memprediksi potensi mahasiswa putus studi, serta mengurangi waktu komputasi secara keseluruhan.

2. Metodologi

Beberapa tahapan penelitian yang dilakukan untuk menyeleksi atribut guna memprediksi potensi mahasiswa putus studi ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

Data mahasiswa yang diperoleh berasal dari Sistem Informasi Akademik Mahasiswa (SIAKAD) dan Learning Management System (LMS) untuk mata kuliah Praktikum Dasar Pemrograman di Jurusan Teknologi Informasi, Politeknik Negeri Malang. Informasi yang diperoleh melalui SIAKAD berkaitan dengan data diri mahasiswa, antara lain jenis kelamin, umur, program studi, jalur seleksi masuk, asal sekolah, ada atau tidaknya beasiswa, serta banyaknya kehadiran dan ketidakhadiran di kelas. Sementara itu, data yang diperoleh melalui LMS berkaitan dengan aktivitas mahasiswa selama proses pembelajaran berlangsung. Data pendukung lainnya juga diperoleh melalui survei mahasiswa untuk mendapatkan informasi mengenai latar belakang mahasiswa, seperti riwayat kesehatan, kepemilikan perangkat pembelajaran, kondisi internet, metode dan frekuensi belajar di rumah, aktivitas selain kuliah, dan pendidikan orang tua. Tabel 1 menunjukkan rincian atribut data mahasiswa yang digunakan pada penelitian ini.

Tabel 1. Data Mahasiswa

No	Atribut	Keterangan	Nilai
1	PS	Program Studi	Kategori
2	Sekolah	Asal sekolah	Kategori
3	Jalur	Jalur seleksi masuk kuliah	Kategori
4	Beasiswa	Beasiswa kuliah	Kategori
5	Orang tua	Orang tua menempuh jenjang perguruan tinggi	Kategori
6	Asal	Daerah tempat tinggal asal	Kategori

7	Kesehatan	Kondisi kesehatan	Kategori
8	Laptop	Kepemilikan laptop	Kategori
9	Pengalaman	Pengalaman membuat kode program	Kategori
10	Kemahasiswaan	Partisipasi dalam kegiatan kemahasiswaan	Kategori
11	Aktivitas	Aktivitas lain selain kuliah	Kategori
12	Tempat tinggal	Tempat tinggal selama kuliah	Kategori
13	Koneksi	Jenis koneksi internet	Kategori
14	Kondisi jaringan	Kondisi jaringan internet	Kategori
15	Metode belajar	Metode belajar mata kuliah praktikum	Kategori
16	Praktikum	Kehadiran dalam perkuliahan praktikum	Kategori
17	Tugas	Persentase pengumpulan tugas	Numerik
18	Keterlambatan	Persentase keterlambatan pengumpulan tugas dan ujian	Numerik
19	JK	Jenis kelamin	Kategori
20	Umur	Umur	Numerik
21	Kehadiran	Persentase kehadiran di kelas	Numerik
22	Nilai MID	Nilai tengah semester	Kategori
23	Kelas Target	Potensi putus studi atau tidak	Kategori

2.2 Pembersihan Data

Pada tahap selanjutnya, pengecekan data dilakukan untuk menghindari adanya informasi yang tidak lengkap atau tidak sesuai dengan format data lainnya. Data yang mengandung informasi berulang (ganda) akan dihapus dari data set. Sementara data yang mengandung informasi tidak lengkap akan diberi nilai sesuai dengan rata-rata nilai dari atribut tersebut. Dalam tahap ini, diperoleh data set berisikan 84 baris data yang menyatakan banyaknya mahasiswa dengan 22 atribut data.

Data set yang telah siap selanjutnya dibagi menjadi dua bagian yaitu data latih dan data uji dengan menggunakan *k*-fold cross validation. Pemilihan subset secara acak pada *k*-fold cross validation dianggap sebagai mekanisme efektif untuk menguji keberhasilan algoritma (Widodo, Brawijaya, & Samudi, 2022).

2.3 Penyeleksian Atribut

Penyeleksian atribut bertujuan untuk meningkatkan kinerja algoritma dalam memprediksi sehingga diharapkan dapat memberikan hasil evaluasi yang baik, serta mengurangi waktu komputasi secara keseluruhan (Jain & Singh, 2018). Pada penelitian ini, tiga metode seleksi atribut yang terdiri dari Chi Square, Pearson Correlation, dan Random Forest digunakan untuk menentukan atribut-atribut penting yang mempengaruhi berpotensi atau tidaknya mahasiswa putus studi.

Chi Square sebagai metode seleksi atribut yang pertama, bekerja dengan membandingkan nilai yang diperoleh dari frekuensi suatu kelas karena pemisahan dengan frekuensi yang diharapkan dari kelas tersebut, sehingga Chi Square akan mengevaluasi setiap gen secara individual yang berkaitan dengan setiap kelas (Sulistiani & Tjahyanto, 2017). Persamaan 1 menunjukkan perhitungan Chi Square pada sebuah gen yang disimbolkan dengan X^2. N_{ij} merupakan banyaknya sampel dari kelas C_i di dalam interval ke-j, M_{Ij} merupakan banyaknya sampel dalam interval ke-j, dan I menunjukkan banyaknya interval. Nilai frekuensi yang diharapkan pada N_{ij} sesuai dengan Persamaan 2.

$$X^2 = \sum_{i=1}^{C}\sum_{j=1}^{I}\frac{\left(N_{ij} - E_{ij}\right)^2}{E_{ij}} \quad (1)$$

$$E_{ij} = M_{Ij}\frac{|C_i|}{N} \quad (2)$$

Metode seleksi atribut kedua, yaitu Pearson Correlation Coefficient digunakan untuk menghitung nilai koefisien korelasi antara dua variabel. Apabila dua variabel mempunyai nilai korelasi yang tinggi, maka kedua variabel tersebut mempunyai kovarian yang besar. Hal ini menandakan bahwa perubahan salah satu variabel dapat diketahui dari perubahan variabel lainnya (Mei, Tan, Yang, & Shi, 2022). Persamaan 3 menunjukkan cara menghitung Pearson Correlation Coefficient yang disimbolkan dengan r (Profillidis & Botzoris, 2019), dengan n menunjukkan banyaknya atribut data, x_i dan y_i adalah variabel ke-i yang dihitung korelasinya.

$$r_{xy} = \frac{n\sum_{i=1}^{n}x_i y_i - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}x_i^{\,2} - \left(\sum_{i=1}^{n}x_i\right)^2\sqrt{n\sum_{i=1}^{n}y_i^{\,2} - \left(\sum_{i=1}^{n}y_i\right)^2}} \quad (3)$$

Metode ketiga yang digunakan pada penelitian ini adalah Random Forest. Random Forest sebagai salah satu algoritma klasifikasi, terdiri dari kombinasi sejumlah pohon keputusan, yang menggabungkan metode agregasi bootstrap dan mengacak pemilihan node data selama proses pembentukan pohon keputusan. Random Forest terbagi ke dalam dua jenis, yaitu pohon regresi untuk variabel respons berupa data kontinu dan pohon klasifikasi untuk variabel

respons berupa data kategori (Chen, Dewi, Huang, & Caraka, 2020).

2.4 Penerapan Algoritma Klasifikasi

Setelah penyeleksian atribut dilakukan menggunakan sejumlah algoritma, setiap skenario hasil seleksi atribut tersebut diklasifikasikan menggunakan Multi-Layer Perceptron (MLP) untuk memprediksi potensi mahasiswa putus studi. Pada penelitian ini, nilai setiap parameter MLP yang digunakan adalah sebagai berikut:

- Banyaknya layer : 2
- Banyaknya neuron : 5
- Learning rate : 0.01
- Banyaknya epoch : 300
- Threshold : 0.0001
- Momentum : 0.9

2.5 Evaluasi Hasil

Evaluasi hasil penerapan algoritma seleksi atribut dan algoritma klasifikasi untuk memprediksi potensi mahasiswa putus studi dilakukan menggunakan dua metrik evaluasi yaitu akurasi dan F1-score. Perhitungan akurasi dan F1-score ditunjukkan pada Persamaan 4 dan 5.

$$akurasi = \frac{tp + tn}{tp + tn + fp + fn} \quad (4)$$

$$f1 - score = \frac{2 \times \frac{tp}{tp + fp} \times \frac{tp}{tp + fn}}{\frac{tp}{tp + fp} + \frac{tp}{tp + fn}} \quad (5)$$

Pada Persamaan 4 dan 5, *tp* menyatakan *true positive*, *tn* menyatakan *true negative*, *fp* menyatakan *false positive*, dan *fn* menyatakan *false negative*.

3. Hasil dan Pembahasan

Sejumlah skenario pengujian terpisah dilakukan untuk mendapatkan hasil prediksi potensi mahasiswa putus studi menggunakan akurasi dan F1-Score. Hasil seleksi atribut dari ketiga metode seleksi atribut yaitu Chi Square, Pearson Correlation, dan Random Forest digunakan sebagai atribut masukan pada MLP guna mendapatkan keluaran hasil prediksi.

Hasil penerapan Chi Square untuk menyeleksi atribut ditunjukkan pada Tabel 2 dalam bentuk urutan keterkaitan atribut dengan kelas yang ditandai dengan nilai X^2.

Tabel 2. Hasil Perhitungan Chi Square

No	Atribut ke	Nama Atribut	X^2
1	18	Keterlambatan	168.765488
2	22	Nilai MID	35.684445
3	17	Tugas	15.635015
4	21	Kehadiran	4.618496
5	3	Jalur	3.516926
6	9	Pengalaman	1.779412
7	2	Sekolah	1.700000
8	15	Metode belajar	1.697326
9	14	Kondisi jaringan	1.518027
10	12	Tempat tinggal	1.286765
11	16	Praktikum	0.936851
12	8	Laptop	0.933529
13	5	Orang tua	0.585600
14	4	Beasiswa	0.417781
15	19	JK	0.411765
16	7	Kesehatan	0.233957
17	20	Umur	0.231727
18	1	PS	0.220588
19	11	Aktivitas	0.212096
20	6	Asal	0.158983
21	13	Koneksi	0.063809
22	10	Kemahasiswaan	0.001131

Semakin besar nilai X^2, maka semakin informatif atribut tersebut, yang artinya atribut tersebut mempunyai pengaruh yang besar terhadap kelas yang dihasilkan, dalam hal in kelas yang dimaksud adalah berpotensi atau tidaknya mahasiswa putus studi. Pada Tabel 2 tersebut, terlihat bahwa atribut keterlambatan mempunyai X^2 terbesar, yang artinya keterlambatan mahasiswa dalam mengumpulkan tugas dan ujian memberikan lebih banyak pengaruh terhadap hasil akhir mahasiswa dinyatakan berpotensi putus studi atau tidak, dibandingkan dengan atribut-atribut lainnya.

Hasil penerapan Pearson Correlation untuk menyeleksi atribut ditunjukkan pada Tabel 3 dalam bentuk urutan keterkaitan atribut dengan kelas yang ditandai dengan nilai *r*.

Tabel 3. Hasil Perhitungan Pearson Correlation

No	Atribut ke	Nama Atribut	*r*
1	22	Nilai MID	0.578982
2	21	Kehadiran	0.446494
3	18	Keterlambatan	0.382293
4	3	Jalur	0.344674
5	17	Tugas	0.333910
6	8	Laptop	0.322040
7	20	Umur	0.306150
8	9	Pengalaman	0.225478
9	12	Tempat tinggal	0.194541
10	7	Kesehatan	0.182818
11	15	Metode belajar	0.180008
12	16	Praktikum	0.168492
13	14	Kondisi jaringan	0.155667
14	2	Sekolah	0.136630
15	11	Aktivitas	0.127731
16	5	Orang tua	0.100799
17	1	PS	0.095871
18	19	JK	0.085749
19	4	Beasiswa	0.082088
20	6	Asal	0.058160
21	13	Koneksi	0.050521
22	10	Kemahasiswaan	0.005946

Sama halnya seperti hasil perhitungan Chi Square, nilai *r* yang diberikan pada setiap atribut menunjukkan korelasinya dengan kelas sebagai target keluaran. Oleh karena itu, atribut yang mempunyai pengaruh paling besar terhadap kelas target akan mempunyai nilai *r* yang paling tinggi. Pada Tabel 3, diketahui bahwa atribut yang mempunyai korelasi terbesar dengan kelas target adalah atribut nilai MID, diikuti dengan atribut kehadiran, dan seterusnya hingga atribut dengan nilai r terkecil adalah atribut kemahasiswaan.

Hasil penerapan Random Forest untuk menyeleksi atribut ditunjukkan pada Tabel 4 dalam bentuk urutan keterkaitan atribut dengan kelas yang ditandai dengan nilai *importance*.

Tabel 4. Hasil Perhitungan Random Forest

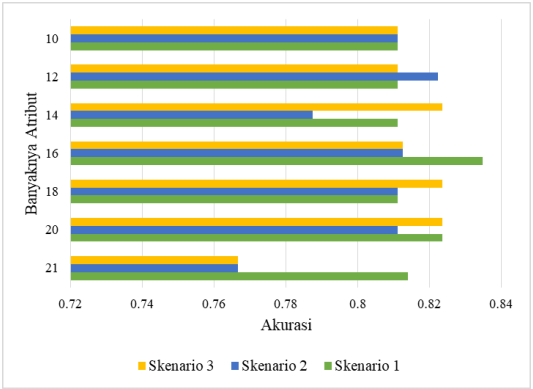
No	Atribut ke	Nama Atribut	*Importance*
1	22	Nilai MID	0.211840
2	18	Keterlambatan	0.126525
3	21	Kehadiran	0.082452
4	17	Tugas	0.078467
5	3	Jalur	0.058776
6	8	Laptop	0.048936
7	15	Metode belajar	0.047345
8	12	Tempat tinggal	0.046667
9	7	Kesehatan	0.038962
10	20	Umur	0.031967
11	5	Orang tua	0.028908
12	14	Kondisi jaringan	0.028508
13	2	Sekolah	0.025193
14	9	Pengalaman	0.024010
15	10	Kemahasiswaan	0.021852
16	11	Aktivitas	0.021324
17	13	Koneksi	0.017641
18	16	Praktikum	0.014225
19	1	PS	0.013724
20	4	Beasiswa	0.012179
21	19	JK	0.010872
22	6	Asal	0.009627

Selanjutnya, MLP diterapkan untuk melakukan prediksi potensi mahasiswa putus studi melalui beberapa skenario. Dalam hal ini, nilai *k* yang digunakan pada *k*-fold cross validation sebesar 9. Pemilihan nilai *k* ini berdasarkan hasil penelitian sebelumnya (Wijayaningrum et al., 2022), yang menyatakan bahwa 9-fold dapat memberikan nilai akurasi terbaik. Hasil penerapan MLP dengan beberapa skenario pengujian ditunjukkan pada Tabel 5. Skenario 1 menunjukkan penerapan MLP dengan seleksi atribut menggunakan Chi Square, skenario 2 menunjukkan penerapan MLP dengan seleksi atribut menggunakan Pearson Correlation Coefficient, dan skenario 3 menunjukkan penerapan MLP dengan seleksi atribut menggunakan Random Forest.

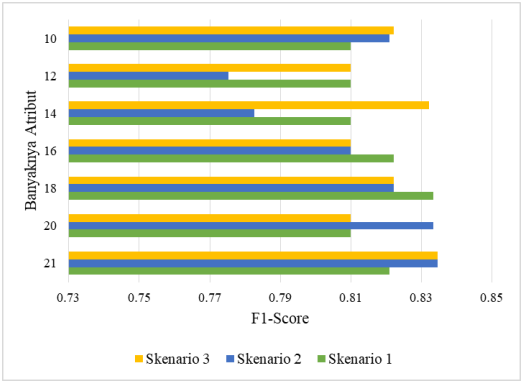
Tabel 5. Hasil Pengujian Algoritma

No	Atribut	Skenario	Akurasi	F1-Score
1	22	Penelitian sebelumnya	0.809876	0.723456
2	21	1	0.813889	0.813889
		2	0.766667	0.834568
		3	0.766667	0.834568
3	20	1	0.823611	0.809876
		2	0.811111	0.833333
		3	0.823611	0.809876
4	18	1	0.811111	0.833334
		2	0.811111	0.822223
		3	0.823611	0.822223
5	16	1	0.834721	0.822223
		2	0.812500	0.809876
		3	0.812500	0.809876
6	14	1	0.811111	0.809876
		2	0.787500	0.782716
		3	0.823611	0.832098
7	12	1	0.811111	0.809876
		2	0.822222	0.775308
		3	0.811111	0.809876
8	10	1	0.811111	0.809876
		2	0.811111	0.820987
		3	0.811111	0.822222

Hasil perbandingan penerapan tiga metode seleksi atribut yang ditunjukkan pada Tabel 5 dapat diilustrasikan ke dalam bentuk grafik seperti yang ditunjukkan pada Gambar 1 untuk evaluasi menggunakan Akurasi dan Gambar 2 untuk evaluasi menggunakan F1-Score.



Gambar 1. Hasil Perbandingan dengan Akurasi



Gambar 2. Hasil Perbandingan dengan F1-Score

Grafik perbandingan hasil menggunakan akurasi yang ditampilkan pada Gambar 1 menunjukkan bahwa berkurangnya atribut dapat meningkatkan akurasi yang dihasilkan karena atribut-

atribut yang digunakan untuk memprediksi potensi mahasiswa putus studi merupakan atribut-atribut yang mempunyai korelasi tinggi dengan kelas target. Akurasi tertinggi diperoleh dari hasil seleksi atribut menggunakan Chi Square yaitu sebanyak 16 atribut data.

Pada Gambar 2, F1-Score tertinggi dihasilkan dari penggunaan 21 atribut data hasil seleksi atribut menggunakan Pearson Correlation dan Random Forest. Namun, dengan mempertimbangkan waktu komputasi yang dibutuhkan untuk memprediksi potensi mahasiswa putus studi menggunakan 21 atribut data, hasil yang diperoleh tidak jauh berbeda saat digunakan 14 atribut data hasil seleksi atribut menggunakan Random Forest dengan waktu komputasi yang lebih sedikit.

Nilai F1-Score cenderung tidak stabil dibandingkan evaluasi menggunakan Akurasi, yakni mengalami penurunan dan peningkatan. Hal ini disebabkan karena pada dasarnya evaluasi menggunakan akurasi baik digunakan saat distribusi kelas positif dan negatif seimbang, sebaliknya F1-Score digunakan saat distribusi kelas positif dan negatif tidak seimbang. Pada penelitian ini, distribusi kelas mahasiswa berpotensi putus studi dan kelas mahasiswa tidak berpotensi putus studi tidak seimbang, karena data set yang diperoleh menyatakan realitanya tidak banyak mahasiswa yang putus studi.

Secara keseluruhan, penggunaan ketiga metode seleksi atribut terbukti memberikan akurasi dan F1-Score yang lebih tinggi dibandingkan dengan penggunaan 22 atribut data dari hasil penelitian sebelumnya untuk memprediksi potensi mahasiswa putus studi.

4. Kesimpulan dan Saran

Pada penelitian ini, prediksi potensi mahasiswa putus studi dengan seleksi atribut menggunakan Chi Square, Pearson Correlation Coefficient, dan Random Forest terbukti dapat meningkatkan akurasi dan F1-Score dibandingkan dengan prediksi yang dilakukan menggunakan semua atribut data. Melalui sejumlah skenario pengujian, Secara keseluruhan, semua skenario pengujian dengan menggunakan berbagai variasi banyaknya atribut memberikan akurasi dan F1-Score rata-rata di atas 0.8. Nilai tertinggi akurasi sebesar 0.834721 diperoleh saat prediksi potensi mahasiswa putus studi dilakukan menggunakan 16 atribut data, sementara nilai tertinggi F1-Score diperoleh saat prediksi potensi mahasiswa putus studi dilakukan menggunakan 21 atribut data.

Data yang digunakan pada penelitian ini mempunyai distribusi kelas positif dan negatif yang tidak seimbang karena banyaknya data set mahasiswa yang benar-benar putus studi lebih kecil dibandingkan dengan mahasiswa yang tidak putus studi. Oleh karena itu, pada penelitian selanjutnya, penanganan data tidak seimbang dapat dilakukan sebelum penerapan algoritma klasifikasi diterapkan sehingga performa algoritma menjadi lebih baik.

Daftar Pustaka:

Anasanti, M. D., Hilyati, K., & Novtariany, A. (2022). Exploring feature selection techniques on Classification Algorithms for Predicting Type 2 Diabetes at Early Stage. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(5), 832–839. https://doi.org/10.29207/resti.v6i5.4419

Bäulke, L., Grunschel, C., & Dresel, M. (2021). Student dropout at university: a phase-orientated view on quitting studies and changing majors. *European Journal of Psychology of Education*, *37*(3), 853–876. https://doi.org/10.1007/s10212-021-00557-x

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods. *Journal of Educational Data Mining*, *11*(3), 1–41. https://doi.org/10.2139/ssrn.3275433

Chan, R. Y. (2016). Understanding the purpose of higher education: An analysis of the economic and social benefits for completing a college degree. *Journal of Education Policy, Planning and Administration*, *6*(5), 1–40.

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1), 52. https://doi.org/10.1186/s40537-020-00327-4

Dasi, H., & Kanakala, S. (2022). Student Dropout Prediction Using Machine Learning Techniques. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(4), 408–414.

Jain, D., & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, *19*(3), 179–189. https://doi.org/10.1016/j.eij.2018.03.002

Mei, K., Tan, M., Yang, Z., & Shi, S. (2022). Modeling of Feature Selection Based on Random Forest Algorithm and Pearson Correlation Coefficient. In *Journal of Physics: Conference Series* (Vol. 2219, p. 012046). IOP Publishing. https://doi.org/10.1088/1742-6596/2219/1/012046

Profillidis, V. A., & Botzoris, G. N. (2019). Statistical Methods for Transport Demand Modeling. In *Modeling of Transport Demand* (pp. 163–224). Elsevier. https://doi.org/10.1016/b978-0-12-811513-8.00005-4

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, *7*(11), 146.

https://doi.org/10.3390/data7110146

Roman, N. V., Davidse, P. E., Human-Hendricks, A., Butler-Kruger, L., & Sonn, I. K. (2022). School Dropout: Intentions, Motivations and Self-Efficacy of a Sample of South Africa Youth. *Youth*, *2*(2), 126–137. https://doi.org/10.3390/youth2020010

Sulistiani, H., & Tjahyanto, A. (2017). Comparative Analysis of Feature Selection Method to Predict Customer Loyalty. *IPTEK Journal of Engineering*, *3*(1), 1–5. https://doi.org/10.12962/joe.v3i1.2257

Widodo, S., Brawijaya, H., & Samudi, S. (2022). Stratified K-fold cross validation optimization on machine learning for prediction. *Sinkron*, *7*(4), 2407–2414. https://doi.org/10.33395/sinkron.v7i4.11792

Wijayaningrum, V. N., Kirana, A. P., Putri, I. K., & Satrio, T. O. (2022). Prediction of Student Academic Performance in Practicum Courses Based on Activity Logs and Student Background. In *2022 International Conference on Electrical and Information Technology* (pp. 366–371). IEEE. https://doi.org/10.1109/IEIT56384.2022.9967888

Wild, S., & Heuling, L. S. (2020). Student dropout and retention: An event history analysis among students in cooperative higher education. *International Journal of Educational Research*, *104*, 101687. https://doi.org/10.1016/j.ijer.2020.101687