Model Prediksi *Dropout* Mahasiswa Menggunakan Teknik *Data Mining*

Muchamad Taufiq Anwar[1], Lucky Heriyanto[2], Fadhla Fanini[3]
[1]*Program Studi Sistem Informasi Industri Otomotif, Politeknik STMI Jakarta*
Jl. Letjen Suprapto No. 26 Cempaka Putih, Jakarta Pusat, DKI Jakarta
E-mail : taufiq@stmi.ac.id[1], lucky.heri@kemenperin.go.id[2], fadlafanini@gmail.com[3]

Abstract— One of the problems that exist at XYZ College is the high number of students who drop out of study (DO), so that efforts are needed to minimize the number of students who drop out. This study aims to build a model that can predict whether a student will graduate or drop out. The data is taken from the academic data of students of the 2014-2019 class. Initial data processing was carried out in Python and modeling was carried out using the C4.5 / J48 algorithm on the WEKA (Waikato Environment for Knowledge Analysis) software. The results show that the attributes that most determine whether a student drop out or graduate are Semester 1 Achievement Index and Semester 2 Achievement Index, with an accuracy of the model reaching 90.6%.

Abstrak— Salah satu permasalahan yang ada di Perguruan Tinggi XYZ adalah tingginya jumlah mahasiswa yang putus studi (*dropout / DO*), sehingga diperlukan upaya untuk minimalisasi jumlah mahasiswa yang *dropout*. Penelitian ini bertujuan untuk membangun sebuah model yang dapat memprediksi apakah seorang mahasiswa akan lulus ataukah *dropout*. Data diambil dari data akademis mahasiswa angkatan 2014-2019. Pemrosesan awal data dilakukan dengan *Python* dan pemodelan dilakukan dengan menggunakan algoritma C4.5 / J48 pada perangkat lunak WEKA (*Waikato Environment for Knowledge Analysis*). Hasil menunjukkan bahwa atribut yang paling menentukan apakah seorang mahasiswa DO atau lulus adalah Indeks Prestasi Semester 1 dan Indeks Prestasi Semester 2, dengan akurasi model mencapai sebesar 90.6%.
Kata Kunci—model prediksi putus studi, *dropout*, *data mining*, C.45, J48

I. PENDAHULUAN

Salah satu permasalahan yang ada di Perguruan Tinggi XYZ adalah tingginya jumlah mahasiswa yang putus studi (*dropout* / DO). Pada data mahasiswa angkatan tahun 2014-2019, jumlah mahasiswa *dropout* mencapai 24,8%. Jumlah ini sangat tinggi sehingga diperlukan upaya minimalisasi jumlah mahasiswa yang *dropout*. *Data Mining* merupakan teknik penemuan pola di dalam data yang juga dapat menghasilkan informasi berharga dari sekumpulan data. Penelitian ini bertujuan untuk membangun sebuah model yang dapat memprediksi apakah seorang mahasiswa akan lulus ataukah *dropout*. Setelah diketahui terdapat potensi *dropout* pada setiap mahasiswa, maka pihak manajemen Perguruan Tinggi XYZ dapat melakukan *treatment* atau upaya pencegahan yang diperlukan untuk meminimalkan jumlah mahasiswa yang *dropout.*

Istilah putus studi menunjukkan pemutusan hubungan oleh seorang siswa dari sekolah, perguruan tinggi atau lembaga pendidikan lainnya tanpa memenuhi kursus terdaftar. Istilah putus sekolah menjelaskan bahwa "Setiap siswa yang meninggalkan sekolah atau lembaga pendidikan lain dengan alasan apa pun sebelum menyelesaikan program studi terdaftar tanpa pindah ke sekolah dasar atau institusi lain" [1]. *Data mining* (DM) membantu mengungkapkan penemuan pengetahuan dan mencari hubungan penting antara variabel / atribut yang berbeda di dalam data. DM dapat digunakan di berbagai bidang dunia nyata seperti perbankan, pendidikan, medis, telekomunikasi, deteksi penipuan, dll. *Educational Data Mining* (EDM) muncul karena meningkatnya aksesibilitas data pendidikan dan karenanya perlu menganalisis data yang sangat besar ini. EDM adalah bidang penelitian multidisiplin yang digunakan untuk menganalisis data pendidikan menggunakan teknik *data mining* [2].

Sebuah survei literatur mengenai prediksi *dropout* menemukan bahwa metode yang paling banyak digunakan adalah klasifikasi dan diikuti oleh *association rule mining* [3]. Penelitian lain juga menggunakan *Naïve Bayes* [4]. Sebuah survei literatur juga menemukan bahwa WEKA (*Waikato Environment for Knowledge Analysis*) merupakan salah satu *tools* paling populer untuk prediksi *dropout* [5]. Survei lain menemukan bahwa salah satu masalah yang ditemukan dalam prediksi *dropout* adalah masalah ketidakseimbangan data [6], sehingga dalam penelitian ini akan dilakukan eksperimen dengan data yang seimbang.

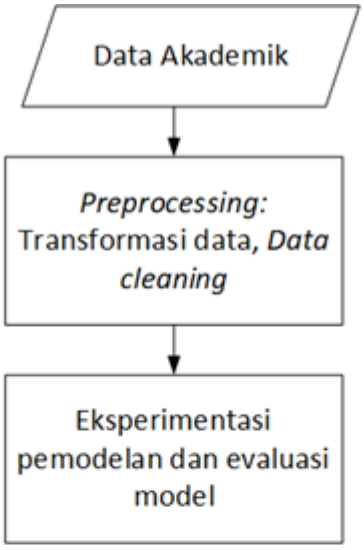
Model Decision Tree (DT) merupakan model dengan struktur hierarki seperti pohon yang berisi aturan-aturan untuk memecah sekumpulan data besar

menjadi grup kecil berdasarkan sebuah besaran tertentu. Terdapat tiga algoritma yang berbeda untuk besaran tersebut, yaitu Entropy Reduction, Gini, dan Chi-square. Penelitian terdahulu telah menunjukkan bahwa model decision tree memiliki performa yang lebih baik dibandingkan model data mining lain dalam berbagai keperluan termasuk untuk prakiraan cuaca [7]–[10]. Selain itu, J48 juga telah digunakan untuk memprediksi resistensi bakteri terhadap obat-obatan pada penderita Tuberculosis [11].

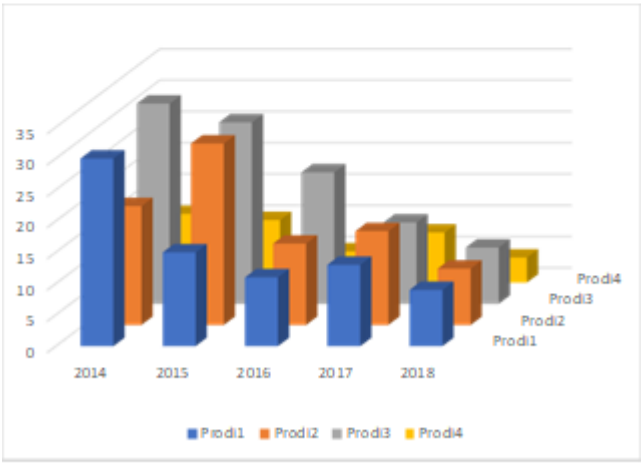
C4.5 adalah salah satu model DT yang merupakan penerus dari algoritma Iterative Dichotomiser 3 (ID3) yang dikembangkan oleh penulis yang sama, yaitu Ross Quinlan[12]. Algoritma ini memiliki beberapa peningkatan dari ID3 seperti kemampuan untuk menangani atribut kontinu dan nominal dan kemampuan untuk memangkas pohon setelah dibuat. C4.5 bekerja dengan membuat pohon berdasarkan entropi dan information gain untuk memilih atribut mana yang berguna dalam mengklasifikasikan data. Entropi adalah ukuran heterogenitas data, sedangkan information-gain adalah ukuran seberapa banyak informasi yang diperoleh dengan membandingkan entropi sebelum dan sesudah memisahkan dataset berdasarkan atribut tertentu. Rumus untuk entropi dan information-gain masing-masing ditunjukkan pada persamaan (1) dan (2). Implementasi terkenal dari C4.5 adalah fungsi J48 yang ditulis dalam bahasa Java yang disediakan dalam perangkat lunak WEKA [13]. Pseudocode untuk algoritma C4.5 ditunjukkan pada Algoritma 1 [14]. J48 akan menghasilkan pohon yang aturannya dapat dengan mudah dibaca oleh manusia. Metode J48 ini juga telah digunakan untuk mencari aturan kasus kebakaran hutan di Indonesia [15]. Penelitian [16] juga menunjukkan bahwa pohon keputusan sangat cocok untuk prediksi hujan.

II. METODE PENELITIAN

Data diambil dari data akademis mahasiswa angkatan 2014-2019. Data mentah memiliki 7220 baris dan 16 kolom. Data ini berisi atribut seperti NIM, Indeks Prestasi Semester (IPS), jumlah SKS yang diambil, dsb. Data mentah berupa data per semester per mahasiswa sehingga perlu ditransformasi menjadi data per mahasiswa. Setelah dilakukan transformasi, beberapa data memiliki data semester (IPS) yang tidak lengkap, sehingga entri data tersebut tidak dipakai. Setelah dilakukan *data cleaning*, data bersih memiliki 979 baris dan 13 kolom yang terdiri dari Prodi, Angkatan, nilai IPS 1 sampai IPS 10, serta Status DO / Lulus. Dalam data ini, jumlah data mahasiswa DO = 305 (31%) dan Lulus = 674 (69%). Pemodelan kemudian dilakukan dengan menggunakan algoritma C4.5 pada *software* WEKA



Gambar. 1. Metode penelitian



Gambar. 2. Jumlah mahasiswa DO untuk masing-masing prodi dan angkatan.

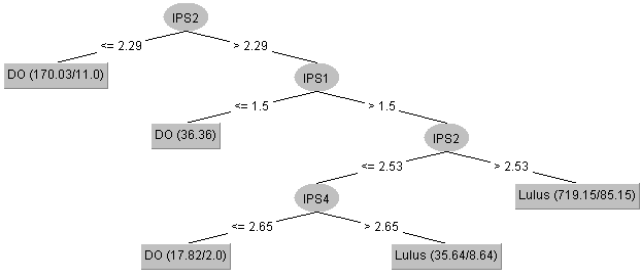
III. HASIL DAN PEMBAHASAN

Jumlah mahasiswa DO untuk masing-masing prodi dan angkatan ditunjukkan pada Gambar 2. Secara umum, jumlah total mahasiswa DO mengalami penurunan dari tahun ke tahun. Namun demikian, patut diperhatikan bahwa angkatan lebih muda mungkin datanya belum “final” sehingga jumlah mahasiswa DO yang disebutkan di sini masih mungkin untuk bertambah.

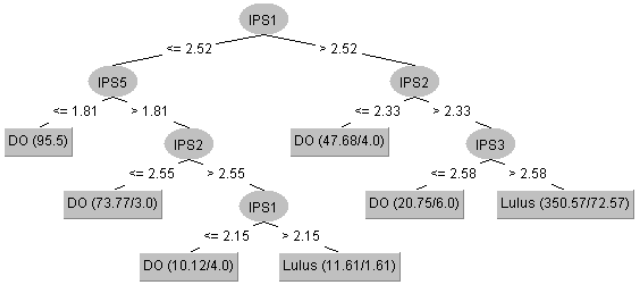
Tabel 1.
Atribut data

Atribut	Keterangan	Tipe data
Prodi	Program Studi	Nominal
Angkatan	Angkatan	Nominal
IPS1	Indeks Prestasi Semester 1	*Float*
IPS2	Indeks Prestasi Semester 2	*Float*
IPS3	Indeks Prestasi Semester 3	*Float*
IPS4	Indeks Prestasi Semester 4	*Float*
IPS5	Indeks Prestasi Semester 5	*Float*
IPS6	Indeks Prestasi Semester 6	*Float*
IPS7	Indeks Prestasi Semester 7	*Float*
IPS8	Indeks Prestasi Semester 8	*Float*
IPS9	Indeks Prestasi Semester 9	*Float*

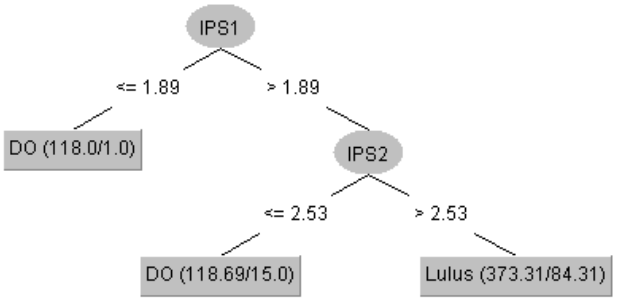
Hasil pemodelan dengan menggunakan fungsi klasifikasi J48 pada software WEKA dengan setting minObj = 10 (minObj adalah jumlah minimum data untuk setiap "daun" pada pohon J48) menunjukkan bahwa yang menentukan apakah seorang mahasiswa DO atau lulus adalah IPS 1, IPS2, dan IPS 4 seperti ditunjukkan pada Gambar 3. Dengan menggunakan 10-fold cross-validation, akurasi model ini adalah 87.7%. Confusion matrix untuk model ini ditunjukkan pada Tabel 2. Ketika minObj = 5, akurasi = 87.95% (hanya sedikit lebih baik) sementara pohon menjadi lebih kompleks. Dengan demikian, kita akan gunakan hasil pohon yang lebih sederhana. Hasil eksperimen pemodelan dengan J48 menunjukkan bahwa aribut 'Prodi' tidak relevan terhadap prediksi DO mahasiswa. Sementara atribut 'Angkatan' memiliki sedikit pengaruh (jika 'Prodi' >= 2016, maka DO) tetapi hal ini tentu saja bias (karena belum ada data mahasiswa yang lulus untuk tahun 2016 ke atas), sehingga atribut 'Angkatan' tidak digunakan.



Gambar. 3. Hasil model decision tree menggunakan J48.



Gambar. 4. Hasil model decision tree menggunakan J48 setelah dilakukan *class-balancing*.



Gambar. 5. Hasil model *decision tree* menggunakan J48 dengan atribut IPS 1 dan IPS 2 untuk data berimbang.

Berdasarkan Gambar 3, rules yang dihasilkan adalah:

a. Jika IPS 2 <= 2.29, maka DO.

b. Jika IPS 1 <= 1.5 dan IPS 2 > 2.29, maka DO.

c. Jika IPS 1 > 1.5 dan IPS 2 <= 2.53 dan IPS 4 <= 2.64, maka DO.

d. Jika IPS 1 > 1.5 dan IPS 2 > 2.53, maka Lulus.

e. Jika IPS 1 > 1.5 dan IPS 2 <= 2.53 dan IPS 4 > 2.64, maka Lulus.

Secara umum, rules tersebut menyiratkan bahwa mahasiswa dengan IPS 1, IPS 2, dan IPS 4 yang rendah akan DO (rules a, b, c). Mahasiswa dengan IPS yang baik akan lulus (rule d). Sementara jika ada mahasiswa dengan nilai IPS awal rendah (IPS 2) tetapi nilai IPS di semester berikutnya (IPS 4) membaik, maka dia akan lulus (rule e). Perlu diperhatikan bahwa pnelitian ini hanya menggunakan data IPS. Untuk memperbaiki akurasi model, dalam penelitian berikutnya dapat ditambahkan atribut lain yang dapat menjelaskan DO atau lulusnya mahasiswa. Untuk saat ini, penggalian dan eksplorasi atribut tersebut dapat dilakukan dengan wawancara.

Tabel 2.
Confusion matrix untuk data tidak berimbang dengan *minObj* = 10

		Diprediksi	
		DO	Lulus
Aktual	DO	200	105
	Lulus	15	659

Tabel 3.
Confusion matrix untuk data berimbang dengan *minObj* = 10

		Diprediksi	
		DO	Lulus
Aktual	DO	229	76
	Lulus	20	285

Tabel 4.
Ringkasan perbandingan performa model untuk beberapa *setting* eksperimen yang berbeda

Setting	KTB*			KB*		
	Aks*	FN*	FP*	Aks	FN	FP
minObj = 5, *10-fcv**	87.95%					
minObj = 10, *10-fcv*	87.7%	0.344	0.22	84.3%	0.249	0.066
Atribut IPS1 dan IPS2, *10-fcv*	86.5%			81.2%	0.325	0.052
Atribut IPS1 dan IPS2, *FTD**				83.3%	0.282	0.052

Dari confusion matrix pada Tabel 2, terlihat bahwa ada lebih banyak (105 lawan 15) mahasiswa yang sebenarnya DO tetapi salah diprediksikan menjadi Lulus (False Negative lebih tinggi daripada False Positive dengan FP = 15/674 = 0.22 dan FN = 105/305 = 0.344). Dalam kasus deteksi DO ini, sebaiknya besarnya FN dapat dibuat lebih kecil. Ketimpangan ini dapat diakibatkan oleh tidak seimbangnya sebaran data untuk kedua kelas. Sehingg dalam ekseprimen berikutnya, dilakukan penyeimbangan kelas dengan teknik random undersampling pada kelas "Lulus". Random undersampling dilakukan dengan menggunakan filter 'SpreadSubSample' pada WEKA sehingga jumlah data pada kelas 'Lulus' sama dengan jumlah data pada kelas 'DO' yaitu 305. Setelah dilakukan class-balancing, akurasi berkurang menjadi 84.3%. Hasil decision tree tertampil pada Gambar 4. Namun demikian, perlu diingat bahwa hasil ini merupakan hasil dari random sampling, sehingga jika dilakukan sampling pada waktu yang percobaan lain, hasil bisa saja berbeda.

Setelah dilakukan class-balancing ini, angka FP dan FN sedikit berkurang dari hasil sebelumnya dengan FP = 0.066 dan FN = 0.249. Dengan demikian, maka usaha untuk memperbaiki model adalah dengan mengeksplorasi variabel lain di luar IPS. Dari kedua eksperimen, disimpulkan bahwa faktor yang paling mempengaruhi model prediksi DO ini adalah IPS 1 dan IPS 2. Kesimpulan ini didukung dengan hasil eksperimen dengan hanya menggunakan atribut IPS 1 dan IPS 2 yang menghasilkan akurasi model sebesar 86.5% pada data tidak berimbang dan sebesar 81.2% pada data berimbang (akurasi 83.3% dengan full training dataset). Hasil decision tree untuk model ini ditunjukkan pada Gambar 5. Ketika menggunakan keseluruhan data training dan menggunakan kelas berimbang, model memiliki akurasi sebesar 84.9% - hanya sedikit lebih baik dari ketika menggunakan 10-fold cross-validation. Ringkasan perbandingan performa model untuk beberapa setting eksperimen yang berbeda ditunjukkan pada Tabel 4. Akurasi tertinggi sebesar 90,6% dicapai ketika keseluruhan dataset digunakan. Penelitian terdahulu yang juga menggunakan jumlah data yang hampir sama dengan penelitian ini dan menggunakan J48 memiliki akurasi 80% [17], sehingga penelitian ini memiliki akurasi yang lebih baik daripada penelitian terdahulu

IV. KESIMPULAN

Penelitian ini mencoba membuat model prediksi apakah seorang mahasiswa akan DO atau lulus dengan menggunakan data IPS. Pemodelan dilakukan menggunakan algoritma C4.5. Hasil menunjukkan bahwa faktor yang paling menentukan apakah seorang mahasiswa DO atau lulus adalah IPS 1 dan IPS 2. Nilai IPS yang rendah pada semester awal memberikan risiko mahasiswa untuk DO, kecuali jika ada perbaikan IPS di semester yang akan datang (IPS 3, IPS 4, dan IPS 5). Akurasi model ini adalah mencapai 90,6% dan lebih baik dari penelitian sebelumnya yaitu sebesar 80%. Penelitian berikutnya hendaknya mengeksplorasi variabel lain yang dapat mempengaruhi DO untuk meningkatkan akurasi model.

DAFTAR PUSTAKA

[1] K. Bonneau, "Brief 3: What is a dropout," *North Carolina Educ. Res. Data Center, Cent. Child Fam. Policy. Retrieved Novemb.*, vol. 30, p. 2011, 2006.

[2] R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *JEDM| J. Educ. Data Min.*, vol. 1, no. 1, pp. 3–17, 2009.

[3] M. Kumar, A. J. Singh, and D. Handa, "Literature survey on educational dropout prediction," *Int. J. Educ. Manag. Eng.*, vol. 7, no. 2, p. 8, 2017.

[4] V. Hegde and P. P. Prageeth, "Higher education student dropout prediction and analysis through educational data mining," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 2018, pp. 694–699.

[5] M. Alban and D. Mauricio, "Predicting university dropout through data mining: A Systematic Literature," *Indian J. Sci. Technol.*, vol. 12, no. 4, pp. 1–12, 2019.

[6] N. Mduma, K. Kalegele, and D. Machuve, "A survey of machine learning approaches and techniques for student dropout prediction," 2019.

[7] A. Joshi, B. Kamble, V. Joshi, K. Kajale, and N. Dhange, "Weather forecasting and climate changing using data mining application," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 3, pp. 19–21, 2015.

[8] F. Sheikh, S. Karthick, D. Malathi, J. S. Sudarsan, and C. Arun, "Analysis of data mining techniques for weather prediction," *Indian J. Sci. Technol.*, vol. 9, no. 38, 2016.

[9] M. T. Anwar, W. Hadikurniawati, E. Winarno, and W. Widiyatmoko, "Performance Comparison of Data Mining Techniques for Rain Prediction Models in Indonesia," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2020, pp. 83–88.

[10] M. T. Anwar, S. Nugrohadi, V. Tantriyati, and V. A. Windarni, "Rain Prediction Using Rule-Based Machine Learning Approach," *Adv. Sustain. Sci. Eng. Technol.*, vol. 2, no. 1, 2020.

[11] W. Hadikurniawati, M. T. Anwar, D. Marlina, and H. Kusumo, "Predicting tuberculosis drug resistance using machine learning based on DNA sequencing data," in *Journal of Physics: Conference Series*, 2021, vol. 1869, no. 1, p. 12093.

[12] J. Quinlan, *C4. 5: programs for machine learning*. 2014.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[14] P. Nevlud, M. Bures, L. Kapicak, and J. Zdralek, "Anomaly-based network intrusion detection methods," *Adv. Electr. Electron. Eng.*, vol. 11, no. 6, pp. 468–474, 2013.

[15] M. T. Anwar, H. D. Pumomo, S. Y. J. Prasetyo, and K. D. Hartomo, "Decision Tree Learning Approach To Wildfire Modeling on Peat and Non-Peat Land in Riau Province," in *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2018, pp. 409–415.

[16] R. S. Kumar and C. Ramesh, "A study on prediction of rainfall using datamining technique," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, 2016, vol. 3, pp. 1–9.

[17] R. L. S. do Nascimento, R. B. das Neves Junior, M. A. de Almeida Neto, and R. A. de Araújo Fagundes, "Educational data mining: An application of regressors in predicting school dropout," in *International Conference on Machine Learning and Data Mining in Pattern Recognition*, 2018, pp. 246–257.