

PENERAPAN METODE DATA MINING UNTUK PREDIKSI DINI KEGAGALAN MAHASISWA MENGGUNAKAN ALGORITMA DECISION TREE

Hendri Noviyanto¹, Arma Fauzi²

^{1,2}Sistem Komputer, Universitas Surakarta

hendrinoviyantoo@gmail.com¹, armafauzi.armafauzi@gmail.com²

Naskah diterima: 14 Nopember 2024 ; Direvisi : 26 Nopember 2024 ; Disetujui : 27 Nopember 2024

Abstrak

Kompleksitas permasalahan yang terdapat pada lingkungan Perguruan Tinggi yang dialami oleh Mahasiswa menyebabkan beberapa kemungkinan buruk seperti mahasiswa lulus tidak tepat waktu, *dropout* dan waktu tunggu mahasiswa yang cukup panjang. Hal tersebut merupakan salah satu faktor penghambat dalam pengembangan sumber daya manusia. Pada era digitalisasi pengembangan sumber daya manusia dituntut sangat cepat dan mutakhir. Faktor yang mempengaruhi kemungkinan terburuk cukup banyak, beberapa diantaranya seperti manajemen waktu, motivasi yang rendah, depresi karena tekanan personal dan sosial, kesalahan dalam memilih jurusan sesuai dengan passion, masalah keuangan, dan *resource* belajar. Oleh karena itu, dibutuhkan sebuah sistem pendukung keputusan proses prediksi dini kegagalan mahasiswa. Pada tahap proses belajar, beberapa mahasiswa membutuhkan sebuah bimbingan yang intens terkait dengan permasalahan internal maupun eksternal. Berdasarkan pada penelitian ini diharapkan informasi penting dapat diberikan kepada pemangku kepentingan dan dengan tepat serta cepat mengetahui kondisi mahasiswa untuk mengambil langkah-langkah strategis guna mencegah kegagalan akademis mahasiswa lebih lanjut. Metode pada penelitian ini menggunakan algoritma decision tree dengan menerapkan beberapa nilai K pada K-Fold Crossvalidation. Nilai K terbaik diperoleh pada nilai 8. Proses seleksi fitur diterapkan untuk mendapatkan fitur terbaik, terbukti bahwa seleksi fitur mampu meningkatkan nilai akurasi menjadi 90%.

Kata kunci: Data Mining, K-Fold Cross-validation, Prediksi, Seleksi Fitur, Decision Tree.

Abstract

The complexity of problems in the Higher Education environment experienced by students causes several bad possibilities such as students graduating late, dropping out and a long waiting time for students. This is one of the inhibiting factors in the development of human resources. In the digital era, human resource development is required to be very fast and sophisticated. There are quite a lot of factors that influence the worst possibilities, some of which are time management, low motivation, depression due to personal and social pressure, mistakes in choosing a major according to passion, financial problems, and learning resources. Therefore, a decision support system is needed for the early prediction process of student failure. At the learning process stage, some students need intense guidance related to internal and external problems. Based on this study, it is hoped that important information can be provided to stakeholders and accurately and quickly find out the condition of students to take strategic steps to prevent further student academic failure. The method in this study uses a decision tree algorithm by applying several K values to K-Fold Crossvalidation. The best K value is obtained at 8. The feature selection process is applied to obtain the best features, it is proven that feature selection can increase the accuracy value to 90%.

Keywords: Data Mining, K-Fold Cross-validation, Prediction, Feature Selection, Decision Tree.

PENDAHULUAN

Latar Belakang Kompleksitas permasalahan pada peserta didik di lingkungan sebuah perguruan tinggi sangat banyak, terlebih di perguruan tinggi swasta yang tidak begitu besar. Berdasarkan studi yang telah dilakukan dengan pendistribusian angket terdapat beberapa masalah yang dialami oleh mahasiswa seperti manajemen waktu, motivasi yang rendah, depresi karena tekanan personal dan sosial, kesalahan dalam memilih jurusan sesuai dengan passion, masalah keuangan, dan resource belajar [1]. Permasalahan di atas dapat menyebabkan kondisi mahasiswa lulus terlambat, kemampuan yang terbatas sehingga menyebabkan waktu tunggu menjadi semakin lama, dropout, dan tidak melanjutkan studi sampai akhir [2]. Hal seperti ini dapat merugikan kedua belah pihak karena memiliki kebutuhan yang sama (simbiosis mutualisme).

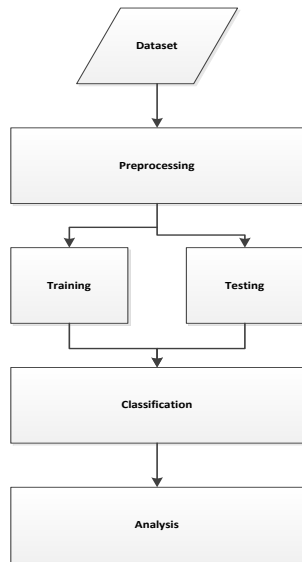
Proses prediksi merupakan sebuah teknik yang dapat dilakukan untuk mengetahui efek kemungkinan apa yang akan terjadi pada waktu mendatang [3][4]. Penelitian yang dilakukan oleh [5] dan [6] dengan memprediksi ketepatan waktu masa studi dirasakan mampu mengatasi sedikit permasalahan yang ada, namun permasalahan sebenarnya sangat kompleks sehingga membutuhkan beberapa *decision* secara menyeluruh. Penelitian [7][8] dan [9] memprediksi kesiapan kerja mahasiswa dari

dini, hal ini bertujuan untuk mengevaluasi kemampuan mahasiswa sebagai bentuk kesiapan kerja setelah lulus. Teknik yang digunakan pada penelitian ini adalah Data Mining atau penambangan data [10]. Teknik tersebut digunakan untuk mengeksplorasi data dengan cara mentransformasi data mentah menjadi pengetahuan praktis [11][6]. Teknik penambangan data yang dilakukan pada sebuah Perguruan Tinggi (PT) dapat digunakan untuk memecahkan masalah, menganalisis dampak masa depan dari sebuah keputusan dan meningkatkan kualitas PT [12].

METODE

Pada penelitian ini, terdapat beberapa langkah untuk mendapatkan hasil proses prediksi. Langkah awal yang dilakukan adalah mengumpulkan dataset untuk input data pada sistem. Dataset yang didapatkan dibagi menjadi 2 yaitu data *training* dan data *testing* [13].

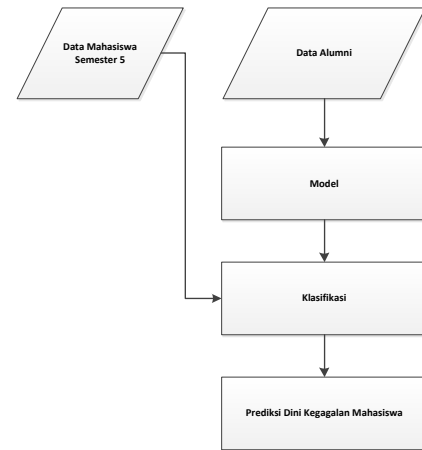
Langkah selanjutnya adalah proses klasifikasi untuk mendapatkan data akurasi dan langkah akhir adalah melakukan analisis terhadap hasil. Alur penelitian ini dapat dilihat pada Gambar di bawah ini.



Gambar 1. Alur Penelitian

Pada Gambar 1 Alur Penelitian proses prediksi diawali dengan mengumpulkan dataset. Dataset pada penelitian ini berasal dari database mahasiswa Universitas Surakarta dengan jumlah 200 data mahasiswa pada semester 5 dengan 6 attributes.

Tahap selanjutnya adalah tahap *preprocessing*, pada tahap ini, data dievaluasi dan diproses untuk menghilangkan *noise* termasuk *filtering attribut* yang digunakan. Hal ini, bertujuan untuk mendapatkan model yang maksimal atau terbaik. Menurut [5] dan [14] data input yang baik akan menghasilkan output yang baik, sehingga kualitas data input harus diperhatikan.



Gambar 2. Model yang digunakan

Pada Gambar 2 Model yang digunakan merupakan rancangan model yang di implementasikan pada sistem pendukung keputusan untuk menyelesaikan permasalahan kegagalan mahasiswa dalam menempuh studi. Sistem Pendukung keputusan di ilustrasikan pada Gambar 3.



Gambar 3. Ilustrasi Sistem

Pada Gambar 3 Ilustrasi Sistem dimulai dengan user interface untuk proses input data mahasiswa, selanjutnya data tersebut diklasifikasikan dengan model yang telah dibuat menggunakan algoritma *decision tree*. Output data mahasiswa adalah hasil akhir yang menjelaskan apakah mahasiswa tersebut bermasalah atau tidak. Jika bermasalah maka harus segera diberikan tindakan untuk membantunya.

Dataset

Dataset yang digunakan pada penelitian dapat dilihat pada tabel 1. Dataset

Tabel 1. Dataset

| Y1 | Y2 | Y3 | Y4 | Y5 | Y6 | X |
|-----|-----|-----|-----|-----|-----|------------|
| 3 | 3 | 1 | 1 | 3 | 1 | Tidak |
| 3 | 3 | 1 | 1 | 3 | 1 | Tidak |
| ... | ... | ... | ... | ... | ... | ... |
| 3 | 2 | 0 | 0 | 2 | 0 | Bermasalah |

Keterangan :

Y1 : Manajemen (Baik | Cukup | Kurang)

Y2 : Motivasi (Baik | Cukup | Kurang)

Y3 : Psikologi (Bermasalah | Tidak)

Y4 : Jurusan (Sesuai | Tidak)

Y5 : Keuangan (Baik | Cukup | Kurang)

Y6 : Sumber Belajar (Baik | Cukup | Kurang)

X : Kelas (Bermasalah | Tidak)

HASIL DAN PEMBAHASAN

Pada penelitian ini, kami menggunakan Algoritma *Decision Tree* untuk melakukan proses prediksi dini kegagalan mahasiswa dalam menempuh masa studi. Penelitian ini dibantu menggunakan *software* WEKA [15][16] untuk membantu melakukan proses prediksi dini kegagalan mahasiswa pada proses pembelajaran guna membentuk model yang diharapkan.

K-Fold Cross Validation

Proses pengujian dengan memanfaatkan nilai K pada *K-Fold Cross-Validation*. *K-Fold Cross-Validation* merupakan sebuah teknik validasi dalam *machine learning* yang digunakan untuk membuat kinerja model dengan lebih akurat, menghindari *overfitting*

dan memastikan bahwa model memiliki kemampuan generalisasi yang baik [17].

Tabel 2. Pengujian K-Fold

| | K-10 | K-9 | K-8 | K-7 |
|-----|------|------|------|------|
| J48 | 88 | 88 | 89.5 | 88 |
| | K-6 | K-5 | K-4 | K-3 |
| J48 | 89.5 | 88.5 | 88.5 | 87.5 |

Pada Tabel 2 Pengujian *K-Fold* telah dilakukan proses pengujian menggunakan varian nilai K dari 3 sampai 10. Nilai terbaik terdapat pada nilai K = 8 dengan perolehan nilai Akurasi sebesar 89,5%.

Penerapan Seleksi Fitur pada Atribut

Penerapan seleksi fitur pada fitur atribut digunakan untuk mendapatkan fitur terbaik dari beberapa fitur yang tersedia. Fitur terbaik ini dapat meningkatkan nilai akurasi pada proses klasifikasi [18]. Proses seleksi fitur menggunakan tekni *wrapper* [3] dengan metode J48 sebagai *evaluatornya* dan *search* menggunakan *geneticsearch* [19] menghasilkan nilai akurasi meningkat sebesar 90%. Fitur yang lolos dari proses seleksi adalah Y2, Y3, Y4, Y5 dan Y6.

Confusin Matrix

=== Confusion Matrix ===

a b <-- classified as

107 7 | a = Tidak

13 73 | b = Bermasalah

Hasil di atas merupakan tabel *confusion matrix* dari penerapan seleksi fitur. Data pada a dapat diklasifikasikan dengan benar sebanyak 107 dan tidak dapat diklasifikasikan sebanyak 7 data. Pada data b dapat dilihat data yang dapat diklasifikasikan sebanyak 73 data dan tidak dapat diklasifikasikan sebanyak 13 data.

PENUTUP

Hasil dari penelitian prediksi dini kegagalan mahasiswa dalam proses belajar di universitas menggunakan algoritma *decision tree* memiliki hasil yang cukup bagus dengan nilai akurasi sebesar 90%. Proses *K-Fold Cross-Validation* dan fitur yang tepat sangat berpengaruh terhadap hasil akurasi. Proses implementasi model ke dalam sistem pendukung keputusan diharapkan mampu mengatasi permasalahan yang terjadi di lingkungan universitas.

DAFTAR PUSTAKA

- [1] Santika, "FAKTOR-FAKTOR YANG MEMPENGARUHI PRESTASI AKADEMIK MAHASISWA," *Jurnal Pendidikan*, vol. 21, pp. 1–17, Mar. 2020, doi: 10.33830/jp.v21i1.704.2020.
- [2] S. Nurhayati, K. kusrini, and E. T. Luthfi, "Prediksi Mahasiswa Drop Out Menggunakan Metode Support Vector Machine," *Sisfotenika*, vol. 5, no. 1, 2015, doi: 10.30700/jst.v5i1.25.
- [3] H. Noviyanto, T. B. Adji, and N. A. Setiawan, "Seleksi Fitur Menggunakan Metode Kombinasi Algoritme Genetika Dan Sequential Minimal Optimization Untuk Klasifikasi Halaman Web," *KNASTIK*, 2016.
- [4] A. López-García, O. Blasco-Blasco, M. Liern-García, and S. E. Parada-Rico, "Early detection of students' failure using Machine Learning techniques," *Operations Research Perspectives*, vol. 11, p. 100292, 2023, doi: <https://doi.org/10.1016/j.orp.2023.100292>.
- [5] H. Noviyanto and B. Mukti, "Period Study Accuracy Prediction using Sequential Minimal Optimization Algorithm," *Sinkron*, vol. 5, no. 1, pp. 164–169, 2020, doi: 10.33395/sinkron.v5i1.10621.
- [6] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.
- [7] H. Noviyanto and B. Mukti, "Prediksi Kesiapan Kerja Mahasiswa menggunakan Algoritme K-Means dan C4.5," *JOSTECH: Journal of Science and Technology*, vol. 2, no. 2, pp. 179–188, 2022, doi: 10.15548/jostech.v2i2.4422.
- [8] R. P. S. Putri and I. Waspada, "Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika," *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, vol. 4, no. 1, p. 1, 2018, doi: 10.23917/khif.v4i1.5975.
- [9] E. P. K. Orpa, E. F. Ripanti, and T. Tursina, "Model Prediksi Awal Masa Studi Mahasiswa Menggunakan Algoritma Decision Tree C4.5," *Jurnal Sistem dan*

- Teknologi Informasi (JUSTIN)*, vol. 7, no. 4, p. 272, 2019, doi: 10.26418/justin.v7i4.33163.
- [10] B. H. Pangestu, Y. Permanasari, and ..., "Data Mining Menggunakan Algoritma Naive Bayes Classifier Untuk Evaluasi Kinerja Karyawan," *Bandung Conference ...*, pp. 177–184, 2023.
- [11] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining*, Third Edit. United States: Elsevier, 2011. doi: 10.1016/B978-008045405-4.00153-1.
- [12] Amazone, "What is Data Mining," aws.amazon.com. Accessed: Feb. 15, 2024. [Online]. Available: <https://aws.amazon.com/id/what-is/data-mining/>
- [13] D. W. Sugiarto, E. Erni, and W. mualim, "SISTEM PENDUKUNG KEPUTUSAN UNTUK PENILAIAN AKHIR MAHASISWA PADA SIDANG SKRIPSI DI ITB YADIKA PASURUAN MENGGUNAKAN METODE SIMPLE MULTI ATTRIBUTE RATING TECHNIQUE (SMART)," *Jurnal Sistem Informasi Aplikasi Teknologi Informasi*, vol. 1, no. 1, Jan. 2024, doi: 10.53567/JOSIATI.V1I1.4.
- [14] A. Sudrajat, "Penerapan Metode Naïve Bayes Untuk Menentukan Penilaian Kinerja Karyawan PT.Sinergi Guna Solusindo," vol. 99, no. 99, pp. 1596–1606, 2022.
- [15] S. R. Garner, "WEKA: The Waikato Environment for Knowledge Analysis," *Proc New Zealand Computer Science Research Students Conference*, pp. 57–64, 1995.
- [16] G. Galih and M. Eriyadi, "Perbandingan Model NBC, SVM, dan C4.5 dalam Mengukur Kinerja Karyawan Berprestasi Pasca Pandemi Covid-19," *Jurnal Informatika*, vol. 9, no. 2, pp. 123–130, 2022, doi: 10.31294/inf.v9i2.13772.
- [17] M. A. Shahid, M. M. Alam, and M. M. Su'ud, *Improved accuracy and less fault prediction errors via modified sequential minimal optimization algorithm*, vol. 18, no. 4 APRIL. 2023. doi: 10.1371/journal.pone.0284209.
- [18] A. I. Pratiwi and Adiwijaya, "On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis," *Applied Computational Intelligence and Soft Computing*, vol. 2018, 2018, doi: 10.1155/2018/1407817.
- [19] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm," *Procedia Technology*, vol. 25, pp. 326–332, 2016, doi: 10.1016/j.protcy.2016.08.114.