

## Deteksi Dini Mahasiswa *Drop Out* Menggunakan C5.0

Ulfi Saidata Aesy<sup>(1)\*</sup>, Alfirna Rizqi Lahitani<sup>(2)</sup>, Taufaldisatya Wijatama Diwangkara<sup>(3)</sup>,  
Riyanto Tri Kurniawan<sup>(4)</sup>

<sup>1,3,4</sup> Sistem Informasi S-1, Fakultas Teknik dan Teknologi Informasi, Universitas Jenderal Achmad Yani, Yogyakarta

<sup>2</sup> Sistem Informasi D-3, Fakultas Teknik dan Teknologi Informasi, Universitas Jenderal Achmad Yani, Yogyakarta

e-mail : {ulfiaesy,alfirnarizqi,t.wijatama.d,riyantotri14}@gmail.com.

\* Penulis korespondensi.

Artikel ini diajukan 7 Oktober 2020, direvisi 18 November 2020, diterima 19 November 2020, dan dipublikasikan 3 Mei 2021.

### Abstract

*The decline in the number of active students also occurred at the Faculty of Engineering and Information Technology, Universitas Jenderal Achmad Yani. This greatly affects the profile of study program graduates. So it is necessary to have a system that can detect students who are threatened with dropping out early. In this study, the attributes chosen were the student's GPA and the percentage of attendance. This attribute is used to classify students who are predicted to drop out. The research data use student data from the Faculty of Engineering and Information Technology, Universitas Jenderal Achmad Yani. This study uses the C5.0 algorithm to build a decision tree to assist data classification. The decision tree that was built with 304 data as training data resulted from a C5.0 decision tree which had an error rate of 5%. The accuracy results obtained from the 76 test data are 93%.*

**Keywords:** *Drop Out, Prediction, C5.0, Classification, Decision Tree*

### Abstrak

Penurunan jumlah mahasiswa aktif juga terjadi pada Fakultas Teknik dan Teknologi Informasi Universitas Jenderal Achmad Yani. Hal ini sangat mempengaruhi profil lulusan program studi. Sehingga perlu adanya sebuah sistem yang mampu mendeteksi mahasiswa yang terancam drop out secara dini. Pada penelitian ini, atribut yang dipilih adalah IPK mahasiswa dan persentase kehadiran. Atribut tersebut digunakan untuk mengklasifikasikan mahasiswa yang terprediksi drop out. Data penelitian menggunakan data mahasiswa Fakultas Teknik dan Teknologi Informasi Universitas Jenderal Achmad Yani. Penelitian ini menggunakan algoritma C5.0 untuk membangun pohon keputusan untuk membantu klasifikasi data. Pohon Keputusan yang dibangun dengan 304 data sebagai data latih menghasilkan pohon keputusan C5.0 yang memiliki tingkat error sebesar 5%. Hasil akurasi yang diperoleh dari 76 data uji adalah 93%.

**Kata Kunci:** *Drop Out, Prediksi, C5.0, Klasifikasi, Decision Tree*

## 1. PENDAHULUAN

Akreditasi program studi menjadi daya tarik dalam proses penerimaan mahasiswa. Oleh karena itu, program studi selalu berusaha untuk meningkatkan akreditasinya supaya dapat menjaring mahasiswa baru (Sutanto, 2017). Beberapa faktor terpenting dalam sebuah akreditasi program studi di perguruan tinggi adalah mahasiswa dan lulusan.

Di beberapa perguruan tinggi terutama di Universitas Jenderal Achmad Yani Yogyakarta pada Fakultas Teknik dan Teknologi Informasi mengalami penurunan jumlah mahasiswa aktif. Terdapat beberapa mahasiswa yang tidak melakukan Registrasi Ulang tanpa cuti selama lebih dari 2 Semester, sehingga dari pihak Program Studi harus melakukan tindakan berupa pemberian surat peringatan atau *drop out*. Pada proses perkuliahan terdapat beberapa mahasiswa yang kehadirannya sangat kurang sehingga jumlah kehadirannya tidak cukup untuk memenuhi minimal syarat mengikuti Ujian Tengah Semester dan Ujian Akhir Semester. Hal ini mempengaruhi hasil IPK mahasiswa dan rata-rata IPK lulusan yang masuk dalam salah satu elemen penilaian (Cahyo, 2018).



Ketika seorang mahasiswa hanya melakukan KRS tanpa mengikuti perkuliahan dan ujian, hasil IPS pada akhir semester adalah 0.0 sehingga mahasiswa tersebut membutuhkan waktu lebih lama untuk lulus. Jumlah mahasiswa aktif dan lama masa studi mahasiswa ini juga menjadi salah satu komponen penilaian akreditasi (Khasanah & Harwati, 2017).

Universitas menentukan beberapa kebijakan sebagai upaya untuk mengatasi penurunan akreditasi dengan melakukan *drop out* terhadap mahasiswa-mahasiswa yang nilai IPK nya dalam beberapa semester sangat rendah dan mahasiswa yang hanya melakukan KRS tanpa mengikuti perkuliahan (Gustian & Hundayani, 2017), namun kebijakan ini juga berpengaruh terhadap jumlah mahasiswa aktif. Oleh karena itu, diperlukan upaya untuk pencegahan mahasiswa *drop out* dengan mendeteksi secara dini mahasiswa *drop out* untuk mendapatkan penanganan dan perhatian khusus.

Mahasiswa akan diklasifikasikan berdasarkan hasil studi dan kehadirannya, sehingga dapat membantu pihak Program Studi untuk menentukan tindakan terhadap mahasiswa-mahasiswa yang terindikasi masuk ke klasifikasi mahasiswa yang terancam *drop out*. Pada deteksi dini mahasiswa *drop out* ini akan menggunakan algoritma C5.0. Di mana pengklasifikasian berdasarkan hasil studi dan kehadiran Algoritma C5.0 ini merupakan Algoritma modifikasi dari algoritma ID3 dan C4.5 (Mutrofin et al., 2019). Dengan menggunakan algoritma ini diharapkan akan dapat membantu dalam mendeteksi secara dini mahasiswa *drop out* di Universitas Jenderal Achmad Yani.

## 1.1. Tinjauan Pustaka

### 1.1.1. Mahasiswa dan Drop Out

Mahasiswa memiliki pengaruh yang cukup tinggi dalam akreditasi, sedangkan tidak sedikit Perguruan Tinggi yang menerapkan *drop out* untuk mengatasi permasalahan mahasiswa, misalnya IPK rendah, kurang serius dalam perkuliahan, hingga lama lulus. Hal ini yang membuat banyak peneliti yang melakukan penelitian terhadap faktor yang menyebabkan mahasiswa *drop out*. Penelitian yang sejenis dengan menggunakan algoritma *data mining* diterapkan pada solusi prediksi mahasiswa *drop out* dengan melihat dari sisi SKS perkuliahan, IPK dan jumlah semester yang telah dilalui (Utari et al., 2020). Dengan melakukan prediksi mahasiswa *drop out*, program studi dapat memantau mahasiswanya yang terprediksi *drop out* sehingga mahasiswa mendapatkan bimbingan secepatnya dan mahasiswa yang bersangkutan dapat lulus tepat waktu (Putra, 2017).

### 1.1.2. Algoritma C5.0

Algoritma C5.0 adalah salah satu algoritma *data mining* yang merupakan penyempurnaan dari algoritma ID3 dan C4.5 (Ahmadi et al., 2018). Algoritma ini cukup banyak diterapkan dalam beberapa penelitian, seperti contohnya pada penelitian mengenai diagnosa penyakit *Disk Hernia* dan *Spondilolisthesis* menggunakan algoritma C5.0. Penelitian ini menggunakan 310 data dari UCI *Machine Learning*, di mana terdapat tiga kelas klasifikasi yaitu Normal, *Disk Hernia*, dan *Spondylolisthesis*. Hasil penelitian menunjukkan bahwa algoritma C5.0 mampu melakukan identifikasi dengan akurasi sebesar 79%. Kemudian pohon keputusan yang dihasilkan Algoritma C5.0 dimaksimalkan dengan menggunakan algoritma *AdaBoost*, sehingga akurasi meningkat menjadi 83% (Aesyi et al., 2020).

Penelitian serupa dengan melakukan kombinasi beberapa algoritma diaplikasikan pada sistem deteksi intrusi (IDS) dalam jaringan komputer dengan tujuan untuk meningkatkan kualitas deteksi. Sulit untuk membedakan koneksi yang tidak sah dari yang resmi karena penyusup bertindak mirip dengan pengguna normal. Dalam algoritma yang diusulkan, integrasi *Tree Augmented Naive Bayes* (TAN) dalam *Bayesian Network* (BN) dan *Boosting* dalam struktur pohon keputusan C5.0 digunakan untuk mengambil keuntungan dan menghindari kelemahan. Hasil percobaan menunjukkan bahwa algoritma yang diusulkan tidak hanya mencapai hasil yang memuaskan dalam akurasi dan mampu mengurangi tingkat alarm palsu (presentasi kejadian yang



diklasifikasikan sebagai jenis serangan dalam jaringan komputer tetapi tingkat serangan masih termasuk normal), tetapi juga meningkatkan pekerjaan yang ada (Nia & Khalili, 2015).

Algoritma C5.0 pernah digunakan untuk penilaian Kinerja Pegawai Negeri Sipil dengan hasil akurasi 98.08% (Kastawan et al., 2018). Kedudukan dan peranan pegawai negeri sipil sebagai abdi masyarakat mengharuskan pegawai untuk memberikan pelayanan yang adil kepada masyarakat. Peran tersebut membuat kinerja pegawai menjadi hal yang penting. Penggunaan algoritma C5.0 untuk memproses data kinerja pegawai, mampu menghasilkan prediksi atau masukan dalam memberikan rekomendasi jabatan, kepangkatan maupun pemberian tunjangan gaji.

Selain itu, C5.0 juga dapat digunakan untuk menemukan atribut penting dalam kumpulan data. Kumpulan data mungkin mengandung banyak atribut yang bekerja semua. Akan tetapi dengan semua atribut kemungkinan hasil dapat tidak efisien, karenanya hanya atribut penting yang harus dipertimbangkan untuk dipilih dalam proses klasifikasi dan pengelompokan pola data yang besar. Studi kasus dalam karya ini adalah eksplorasi atribut dengan menemukan akurasi dalam *dataset* kanker payudara (Ojha et al., 2017).

Tahap penentuan *root node* algoritma C5.0 (Kastawan et al., 2018):

- 1) Menghitung *entropy* yang digunakan untuk mengklasifikasikan label kelas dari *tuple* acak di D. Pada variable  $pi$  merupakan peluang yang bukan nol dari *tuple* acak di D. Penggunaan log basis 2 dalam menghitung *entropy* adalah untuk pengkodean ke dalam bit.

$$Info(D) = - \sum_{i=1}^m pi \log_2(pi) \quad (1)$$

- 2) Hasil perhitungan *entropy* digunakan untuk menghitung *information gain* pada atribut A.

$$Info_A(D) = \sum_{j=1}^Y \frac{|D_j|}{D} \times Info(D_j) \quad (2)$$

- 3) Setelah itu dilanjutkan dengan menghitung *gain* pada atribut A.

$$GAIN(A) = Info(D) - Info(D_j) \quad (3)$$

Hasil *information gain* tertinggi akan dipilih menjadi *node*.

## 2. METODE PENELITIAN

Tahap awal penelitian ini adalah dengan melakukan analisis permasalahan yang ada kemudian melakukan pemetaan solusi dengan menggunakan data yang sesuai dan yang tersedia (Morales et al., 2017). Pada metode penelitian dalam penelitian ini berupa pembuatan aplikasi dan pengujian terhadap aplikasi tersebut. Adapun tujuan yang dicapai untuk membantu mengklasifikasikan mahasiswa *drop out* secara dini berdasarkan hasil studi dan kehadiran menggunakan algoritma C5.0.

Dengan melihat penelitian yang sebelumnya, deteksi *drop out* banyak menggunakan metode *Naïve Bayes*, ID3 dan C4.5. Sedangkan dalam beberapa pengujian pohon keputusan, algoritma C5.0 mencapai hasil akurasi yang paling baik dibanding ID3 dan C4.5 (Rajeswari & Suthendran, 2019). Oleh karena itu penelitian ini mengakomodasi algoritma C5.0 sebagai solusi pohon keputusan untuk melakukan deteksi dini mahasiswa *drop out* dengan menggunakan fitur diantaranya adalah hasil studi dan jumlah kehadiran.

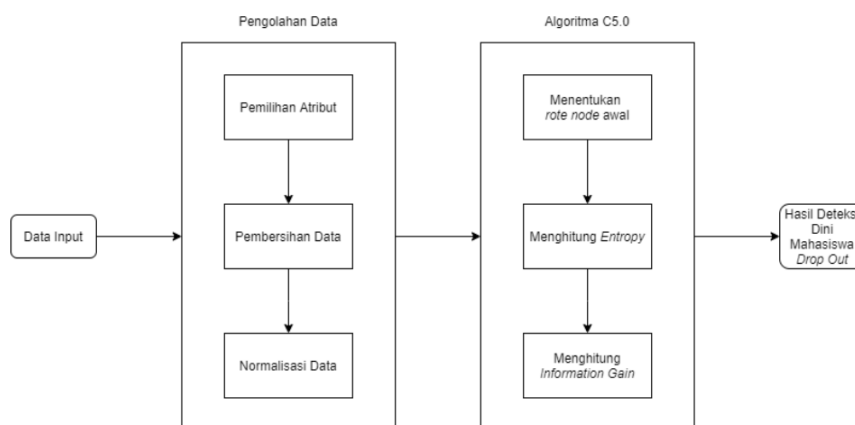
Penelitian dilakukan pada Fakultas Teknik dan Teknologi Informasi Universitas Jenderal Achmad Yani Yogyakarta. Tahap awal peneliti akan mengumpulkan data terkait hasil studi dan presensi mahasiswa. Selanjutnya data akan diolah dengan dilakukan pemilihan atribut yang sesuai dengan penelitian yang ditunjukkan oleh Gambar 1. Kemudian data yang atributnya sudah terpilih akan dibersihkan yang nantinya data akan dinormalisasikan.



Setelah data dinormalkan, maka data akan masuk ke proses pengklasifikasian menggunakan algoritma C5.0. Langkah awal dari algoritma ini adalah dengan menentukan *root node* awal dari atribut yang ada. Kemudian setelah *root node* sudah ditentukan, maka menghitung *entropy* dari atribut-atribut yang akan dicari *node* selanjutnya. Setelah semua atribut telah dihitung *entropy*-nya, maka langkah selanjutnya adalah dengan menghitung *Information Gain*. Nilai *information gain* tertinggi akan dijadikan *node* selanjutnya. Kemudian kembali lagi dengan menghitung *entropy* dan *information gain* dari atribut tersisa. Proses ini berlangsung hingga *node* akhir (Susanti et al., 2019).

Kemudian tahap berikutnya adalah melakukan analisis terhadap data dengan melakukan klasifikasi dengan menggunakan metode algoritma C5.0. Hasil dari tahap analisis ini akan dilakukan pembahasan terhadap hasil klasifikasi dari data yang ada yang nantinya akan mengelompokkan mahasiswa yang masuk dalam kelas *Drop out* (Alban & Mauricio, 2019).

Setelah hasil klasifikasi keluar, maka data mahasiswa yang terprediksi DO akan diperoleh. Tahap akhir dari penelitian ini adalah melakukan pengujian dan evaluasi sistem. Hal ini dilakukan untuk melihat performa sistem yang dibangun.



Gambar 1. Desain Sistem.

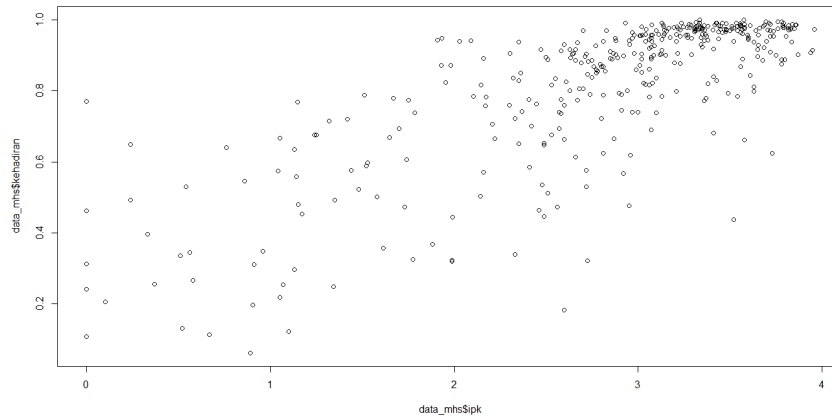
### 3. HASIL DAN PEMBAHASAN

Penelitian ini mengambil data pada Fakultas Teknik dan Teknologi Informasi Universitas Jenderal Achmad Yani Yogyakarta. Data yang diperoleh adalah data mahasiswa angkatan 2012, angkatan 2013, angkatan 2015, angkatan 2017, dan angkatan 2018. Pengarsipan data yang belum rapi menjadi kendala dalam pengumpulan data, sehingga hanya terkumpul data sebanyak 380 data.

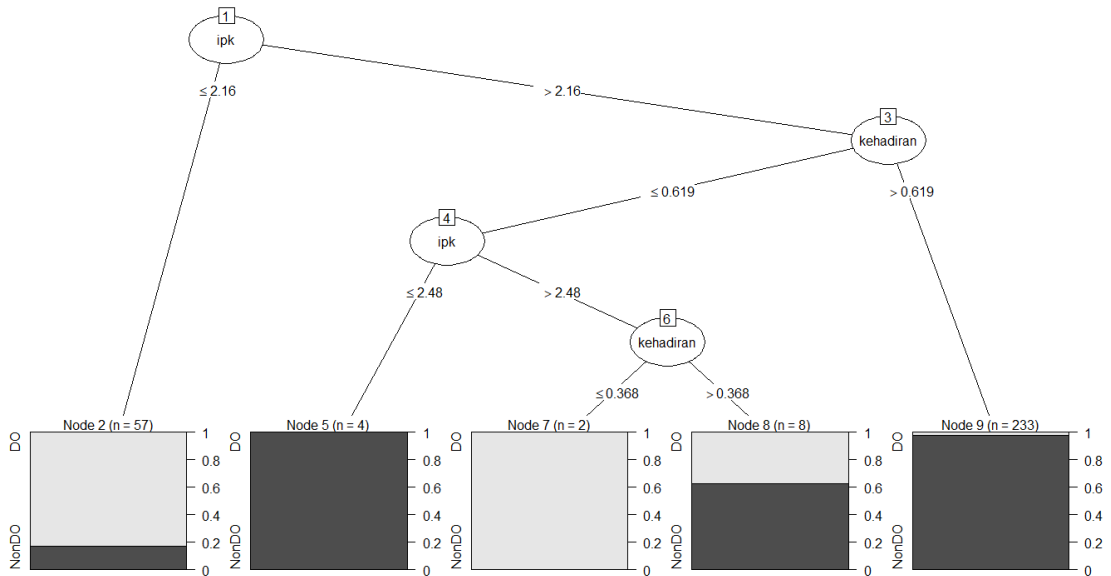
Atribut yang digunakan dalam penelitian ini adalah IPK dan kehadiran. Dari seluruh data yang diperoleh, dilakukan proses pembersihan data sehingga data yang dapat digunakan untuk penelitian sebanyak 380 data. Dari data yang diperoleh kemudian dilakukan proses normalisasi data termasuk pengelompokan data berdasarkan label DO dan NonDO yang digunakan sebagai hasil prediksi. Hasil dari normalisasi dan pelabelan data menghasilkan 308 data mahasiswa NonDO dan 72 data mahasiswa DO. Distribusi data berdasarkan atribut IPK dan kehadiran dari 380 data yang dapat dilihat pada Gambar 2.

Proses pembuatan pohon keputusan dimulai dengan pembagian data menjadi data latih dan data uji secara *random* dengan perbandingan 80% dan 20% yaitu 304 data latih dan 76 data uji. Dari data latih tersebut, dibangun model pohon keputusan dengan menentukan *root node* terlebih dahulu. *Root node* ditentukan berdasarkan *information gain* tertinggi dari atribut. Pada penelitian *information gain* tertinggi adalah atribut IPK. Kemudian *node* lainnya akan terbentuk sehingga membentuk pohon keputusan seperti pada Gambar 3.





Gambar 2. Distribusi Data.



Gambar 3. Pohon Keputusan Menggunakan Algoritma C5.0.

Pohon keputusan yang terbentuk memiliki *error* sebesar 5.9%. Klasifikasi data dari pohon keputusan dibagi menjadi 2, yaitu Do dan NonDo. Penggunaan atribut IPK sebesar 100% dan atribut kehadiran sebesar 81.25%. Hasil klasifikasi ditunjukkan pada Tabel 1.

Tabel 1. Hasil Klasifikasi

	DO	NonDO
DO	49	8
NonDO	10	237

Pada tahap selanjutnya dilakukan untuk mengetahui seberapa tepat sistem melakukan prediksi. Pengujian menggunakan data uji sebanyak 76 data. Dari hasil pengujian diperoleh data 71 terprediksi benar yang dapat dilihat pada Tabel 2.



Tabel 2. Hasil Prediksi

	DO	NonDO
DO	11	3
NonDO	2	60

Dari hasil prediksi dapat dihitung dengan melakukan perbandingan jumlah diagnosis benar (x) oleh sistem dengan jumlah data (n) (Aesy & Wardoyo, 2019):

$$Akurasi = \frac{x}{n} \times 100\% \quad (4)$$

Perhitungan akurasi menggunakan rumus tersebut menghasilkan akurasi sebesar 93%.

#### 4. KESIMPULAN

Berdasarkan hasil penelitian tersebut, maka dapat disimpulkan:

- 1) Penelitian ini telah membangun sistem yang mampu menghasilkan pohon keputusan menggunakan algoritma C5.0 untuk membantu mengklasifikasikan mahasiswa yang terprediksi *drop out* secara dini. Pohon keputusan tersebut memiliki tingkat *error* sebesar 5%.
- 2) Penggunaan data latih 80% dan data uji 20% secara *random* menghasilkan akurasi sebesar 93%. Hal ini menunjukkan bahwa performa dari pohon keputusan tersebut cukup baik untuk melakukan prediksi dini mahasiswa *drop out*.

#### UCAPAN TERIMA KASIH

Ucapan terimakasih ini ditujukan kepada Direktorat Penelitian dan Pengabdian kepada Masyarakat Direktorat Jenderal Penguatan Penelitian dan Pengembangan (DRPM) Kementerian Riset dan Pendidikan Tinggi (Kemristekdikti) Republik Indonesia yang telah memberikan kesempatan kepada tim peneliti untuk melakukan penelitian pada tahun 2020 skema Penelitian Dosen Pemula (PDP). Tim penelitian juga mengucapkan terima kasih kepada Fakultas Teknik dan Teknologi Informasi Universitas Jenderal Achmad Yani Yogyakarta yang telah membantu tim peneliti dalam penyediaan data penelitian.

#### DAFTAR PUSTAKA

- Aesy, U. S., Diwangkara, T. W., & Kurniawan, R. T. (2020). Diagnosa Penyakit Disk Hernia Dan Spondylolisthesis Menggunakan Algoritma C5. *Telematika*, 16(2), 81. <https://doi.org/10.31315/telematika.v16i2.3181>
- Aesy, U. S., & Wardoyo, R. (2019). Prediction of Length of Study of Student Applicants Using Case Based Reasoning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 13(1), 11. <https://doi.org/10.22146/ijccs.28076>
- Ahmadi, E., Weckman, G. R., & Masel, D. T. (2018). Decision making model to predict presence of coronary artery disease using neural network and C5.0 decision tree. *Journal of Ambient Intelligence and Humanized Computing*, 9(4), 999–1011. <https://doi.org/10.1007/s12652-017-0499-z>
- Alban, M., & Mauricio, D. (2019). Predicting University Dropout through Data Mining: A systematic Literature. *Indian Journal of Science and Technology*, 12(4), 1–12. <https://doi.org/10.17485/ijst/2019/v12i4/139729>
- Cahyo, P. W. (2018). Klasterisasi Tipe Pembelajar Sebagai Parameter Evaluasi Kualitas Pendidikan Di Perguruan Tinggi. *Teknomatika*, 11(1), 49–55.
- Gustian, D., & Hundayani, R. D. (2017). Combination of AHP Method with C4.5 in the level classification level out students. *2017 International Conference on Computing, Engineering, and Design (ICCED)*, 1–6. <https://doi.org/10.1109/ICED.2017.8308098>
- Kastawan, P. W., Wiharta, D. M., & Sudarma, M. (2018). Implementasi Algoritma C5.0 pada Penilaian Kinerja Pegawai Negeri Sipil. *Majalah Ilmiah Teknologi Elektro*, 17(3), 371. <https://doi.org/10.24843/MITE.2018.v17i03.P11>



- Khasanah, A. U., & Harwati. (2017). A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques. *IOP Conference Series: Materials Science and Engineering*, 215, 012036. <https://doi.org/10.1088/1757-899X/215/1/012036>
- Morales, A. C., Amir, O., & Lee, L. (2017). Keeping It Real in Experimental Research—Understanding When, Where, and How to Enhance Realism and Measure Consumer Behavior. *Journal of Consumer Research*, 44(2), 465–476. <https://doi.org/10.1093/jcr/ucx048>
- Mutrofin, S., Khalimi, A. M., Kurniawan, E., Ginardi, R. V. H., Fatichah, C., & Sari, Y. A. (2019). Detection of Potentially Students Drop Out of College in Case of Missing Value Using C4.5. *2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC)*, 349–354. <https://doi.org/10.1109/ICSECC.2019.8907014>
- Nia, F. Y., & Khalili, M. (2015). An efficient modeling algorithm for intrusion detection systems using C5.0 and Bayesian Network structures. *2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, 1117–1123. <https://doi.org/10.1109/KBEI.2015.7436203>
- Ojha, U., Jain, M., Jain, G., & Tiwari, R. K. (2017). Significance of important attributes for decision making using C5.0. *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–4. <https://doi.org/10.1109/ICCCNT.2017.8204031>
- Putra, A. (2017). SOLUSI PREDIKSI MAHASISWA DROP OUT PADA PROGRAM STUDI SISTEM INFORMASI FAKULTAS ILMU KOMPUTER UNIVERSITAS BINA DARMA. *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, 8(1). <https://doi.org/10.24176/simet.v8i1.893>
- Rajeswari, S., & Suthendran, K. (2019). C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud. *Computers and Electronics in Agriculture*, 156, 530–539. <https://doi.org/10.1016/j.compag.2018.12.013>
- Susanti, Y., Respatiwan, Handajani, S. S., Pratiwi, H., Slamet, I., Hartatik, & Istiqomah, F. (2019). Classification of teak wood production in Central Java using the C5.0 algorithm. *AIP Conference Proceedings*, 2202(1), 020094. <https://doi.org/10.1063/1.5141707>
- Sutanto, E. M. (2017). The influence of organizational learning capability and organizational creativity on organizational innovation of Universities in East Java, Indonesia. *Asia Pacific Management Review*, 22(3), 128–135. <https://doi.org/https://doi.org/10.1016/j.apmrv.2016.11.002>
- Utari, M., Warsito, B., & Kusumaningrum, R. (2020). Implementation of Data Mining for Drop-Out Prediction using Random Forest Method. *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 1–5. <https://doi.org/10.1109/ICoICT49345.2020.9166276>

