



Classification of Student Drop Out Risk Using Decision Tree C5.0 with Feature Selection

Klasifikasi Status Mahasiswa Berisiko *Drop Out* Menggunakan *Decision Tree C5.0* dengan Seleksi Fitur

Farisyah Lutfiah Hanis^{1*}, Ulfa Khaira², Daniel Arsa³

^{1,2,3}Program Studi Sistem Informasi, Fakultas Sains dan Teknologi,
Universitas Jambi, Indonesia

E-Mail: ¹farisyahlutfiahhanis01@gmail.com, ²ulfa.ilkom@gmail.com, ³danielarsa@unja.ac.id

Received Aug 07th 2025; Revised Sep 21th 2025; Accepted Sep 30th 2025; Available Online Oct 30th 2025

Corresponding Author: Farisyah Lutfiah Hanis

Copyright © 2025 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

This research compares the performance of the Decision Tree C5.0 algorithm with and without Mutual Information feature selection in classifying students at risk of dropping out. The data used includes undergraduate students from the 2018–2023 cohorts with attributes of academic performance each semester and student final project activities. The data processing stages include data selection, cleaning, labeling, and handling of imbalanced data using Random Undersampling. Model evaluation was performed using accuracy, precision, recall, and F1-score metrics. The results showed that for students from the 2022–2023 cohort, the model without feature selection produced an accuracy of 87.14%, which increased to 88.57% after applying feature selection. Meanwhile, for students enrolled in 2018–2021, the accuracy of the model without feature selection reached 93.04% and increased to 94.30% with the application of feature selection. The dominant factors affecting classification differed between the two groups, with the number of absences being the main indicator for students enrolled in 2022–2023, while the duration of the final project was more influential for students enrolled in 2018–2021.

Keyword: C5.0, Classification, Decision Tree, Drop Out, Feature Selection, Mutual Information

Abstrak

Penelitian ini bertujuan untuk membandingkan performa algoritma *Decision Tree C5.0* dengan dan tanpa seleksi fitur *Mutual Information* dalam klasifikasi mahasiswa berisiko *drop out*. Data yang digunakan mencakup mahasiswa S1 angkatan 2018–2023 dengan atribut performa akademik tiap semester serta aktivitas tugas akhir mahasiswa. Tahapan pengolahan data meliputi seleksi data, pembersihan, pelabelan, dan penanganan data tidak seimbang menggunakan *Random Undersampling*. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, dan f1-score. Hasil penelitian menunjukkan bahwa pada mahasiswa angkatan 2022–2023, model tanpa seleksi fitur menghasilkan akurasi sebesar 87,14%, dan meningkat menjadi 88,57% setelah penerapan seleksi fitur. Sementara itu, pada mahasiswa angkatan 2018–2021, akurasi model tanpa seleksi fitur mencapai 93,04% dan meningkat menjadi 94,30% dengan penerapan seleksi fitur. Faktor dominan yang memengaruhi klasifikasi berbeda pada masing-masing kelompok, di mana jumlah absen menjadi indikator utama pada mahasiswa angkatan 2022–2023, sedangkan durasi pengerjaan tugas akhir lebih berpengaruh pada mahasiswa angkatan 2018–2021.

Kata Kunci: C5.0, Decision Tree, Drop Out, Klasifikasi, Mutual Information, Seleksi Fitur

1. PENDAHULUAN

Dalam dunia pendidikan, termasuk Perguruan Tinggi saling berkompetisi untuk mencetak lulusan yang berkualitas [1]. Salah satu indikator kualitas perguruan tinggi dapat dilihat dari akreditasi yang baik karena memenuhi kriteria-kriteria yang telah ditentukan, salah satunya tingkat kelulusan tepat waktu [2]. Masa studi mahasiswa telah diatur dalam Permendikbud, yaitu selama 8-10 semester [3]. Namun dalam pelaksanaannya, banyak mahasiswa tidak mampu menyelesaikan studi tepat waktu, dan bahkan menghadapi kegagalan akademik berupa *drop out* (DO).



Fenomena ini juga terjadi di Universitas Jambi yang memiliki rata-rata masa studi lebih dari delapan semester. Berdasarkan data statistik [4] yang diakses Februari 2025, pada semester genap tahun akademik 2024/2025, masih terdapat 555 mahasiswa angkatan 2018 dan 1.128 mahasiswa angkatan 2019 yang belum menyelesaikan tugas akhir. Jumlah ini menunjukkan persentase yang cukup tinggi, yaitu 8,6% dari total mahasiswa angkatan 2018 dan 17,3% dari total mahasiswa angkatan 2019, sehingga dapat terlihat bahwa sebagian besar mahasiswa akhir belum dapat menyelesaikan studinya, sehingga berpotensi meningkatkan angka *drop out*.

Keterlambatan kelulusan dan *drop out* dapat menimbulkan dampak serius, tidak hanya bagi mahasiswa yang mengalami stres akademik dan stigma negatif [5], tetapi juga bagi orang tua dari segi finansial dan waktu [6]. Selain itu, masalah ini juga berpotensi memengaruhi akreditasi perguruan tinggi [7]. Kondisi tersebut menjadi hal penting yang perlu ditangani, salah satunya melalui penelitian untuk memprediksi potensi *drop out* mahasiswa.

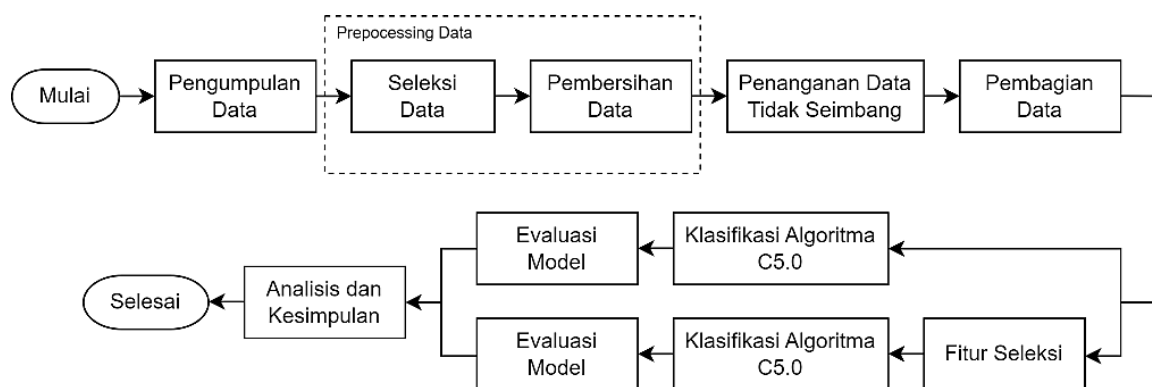
Sehubungan dengan hal tersebut, penelitian ini menggunakan metode klasifikasi dalam *data mining* untuk mengidentifikasi status risiko *drop out* mahasiswa. Metode klasifikasi yang dipilih adalah algoritma *decision tree* yang memiliki keunggulan menampilkan hasil prediksi melalui struktur pohon keputusan yang mudah dipahami [8]. Dalam membangun pohon keputusan, algoritma C5.0 dipilih karena merupakan pengembangan dari algoritma ID3 dan C4.5, dengan keunggulan tambahan pada proses *boosting* yang menjadikan C5.0 lebih unggul dari algoritma sebelumnya [9] [10]. Namun, algoritma ini memiliki kelemahan yaitu rentan terhadap *overlapping*, terutama ketika fitur yang digunakan jumlahnya sangat banyak [11]. Untuk mengatasi kelemahan tersebut, dilakukan penerapan teknik seleksi fitur dengan mengurangi jumlah fitur yang kurang relevan, sehingga dapat meningkatkan kinerja model [8]. Metode *Mutual Information* (MI) dipilih sebagai metode seleksi fitur karena kemampuannya yang dapat mengukur semua jenis hubungan antara atribut termasuk hubungan non linear. Keunggulan ini menjadikannya efektif untuk berbagai jenis data [12].

Penelitian sebelumnya yang menggabungkan metode *decision tree* dan seleksi fitur dilakukan oleh Hariyanti et al. (2024), yang membandingkan seleksi fitur *Analysis of Variance* (ANOVA), *Univariate*, dan *chi-square*. Hasilnya menunjukkan bahwa ANOVA mampu meningkatkan performa model, meskipun membutuhkan waktu komputasi lebih lama [13]. Penelitian lain oleh Adnyana et al. (2021) menggunakan algoritma *Random Forest* dan ID3 dengan seleksi fitur *Principal Component Analysis* (PCA), namun hasilnya justru menurunkan akurasi model karena kelemahan PCA [14]. Sementara itu, Putri et al. (2024) mengaplikasikan algoritma C4.5 dengan seleksi fitur *chi-square* dan *mutual information* untuk klasifikasi status pasien, di mana seleksi fitur terbukti mampu meningkatkan akurasi [15].

Meskipun penelitian sebelumnya telah mengeksplorasi penggunaan algoritma *decision tree* dengan seleksi fitur, penelitian ini memiliki perbedaan signifikan. Penelitian ini secara spesifik menerapkan algoritma C5.0 dan seleksi fitur *mutual information* pada kasus mahasiswa berisiko *drop out* di Universitas Jambi. Selain itu, data yang digunakan tidak hanya mencakup histori akademik selama perkuliahan, tetapi juga informasi terkait proses dan aktivitas tugas akhir. Kombinasi kedua sumber data merupakan kebaharuan dalam penelitian ini dalam konteks klasifikasi mahasiswa berisiko *drop out*. Dengan demikian, penelitian ini bertujuan untuk menganalisis performa algoritma *decision tree* C5.0 dan seleksi fitur *mutual information* dalam mengklasifikasikan status mahasiswa, serta untuk mengetahui atribut apa saja yang paling berpengaruh terhadap klasifikasi mahasiswa yang berisiko *drop out*.

2. METODE PENELITIAN

Penelitian ini menggunakan kerangka *Knowledge Discovery in Databases* (KDD) yang merupakan proses standar dalam *data mining* untuk menemukan pola yang berguna dari data [11]. Tahapan metodologi penelitian ini diilustrasikan secara lebih rinci pada Gambar 1.



Gambar 1. Metodologi Penelitian

2.1. Pengumpulan Data

Proses pengumpulan data dilakukan dengan mengambil data akademik mahasiswa aktif S1 Universitas Jambi dari angkatan 2018-2023 yang diperoleh dari UPA TIK Universitas Jambi dan berasal dari Sistem Informasi Akademik (SIKAD) dan Sistem Elektronik Terintegrasi Tugas Akhir (ELISTA). Pada penelitian ini dibagi dua kategori data yaitu mahasiswa angkatan 2022 – 2023 dan mahasiswa angkatan 2018 – 2021. Setiap kategori mahasiswa menggunakan atribut yang berbeda, sesuai dengan ketersediaan data dan kondisi studi masing-masing.

2.2. Preprocessing Data

2.2.1. Seleksi Data

Pada tahap seleksi data, data yang relevan dipilih dari dataset operasional sebelum memulai tahap penambangan informasi. Data yang dipilih disimpan dalam file khusus yang terpisah dari *database* operasional untuk digunakan selama proses penambangan data [16].

2.2.2. Pembersihan Data

Tahap pembersihan ini mencakup tugas-tugas seperti menghapus data duplikat, mengidentifikasi dan mengoreksi kesalahan ketik dan kesalahan lainnya untuk memastikan kualitas data [16].

2.3. Penanganan Data Tidak Seimbang

Dataset yang tidak seimbang adalah kondisi dimana salah satu kelas memiliki jumlah data yang jauh lebih sedikit dibandingkan kelas lainnya [17]. Model yang dilatih dengan dataset tidak seimbang cenderung kurang efektif dalam mengklasifikasikan kelas minoritas, karena model belajar lebih banyak dari data mayoritas yang jumlahnya dominan [18]. Penanganan data tidak seimbang pada penelitian ini akan menggunakan *random undersampling*. *Random undersampling* mengatasi masalah data tidak seimbang dengan cara menghapus sebagian data dari kelas mayoritas secara acak, sehingga distribusi antar kelas menjadi lebih seimbang dan dapat meningkatkan kemampuan model dalam mengenali pola pada kelas minoritas dan mengurangi bias terhadap kelas mayoritas [19].

2.4. Pembagian Data

Pada tahap ini, data dibagi menjadi dua bagian yaitu *training* dan *testing*. Data *training* digunakan untuk melatih algoritma dalam mengenali pola, sementara data *testing* digunakan untuk mengevaluasi performa dari algoritma yang dilatih. Pembagian dilakukan dengan proporsi 80% untuk *training* dan 20% untuk *testing*.

2.5. Seleksi Fitur Mutual Information

Seleksi fitur adalah salah satu teknik yang bertujuan untuk mengurangi dimensi data dan mengurangi fitur yang dianggap kurang relevan [20]. Pada penelitian ini, seleksi fitur yang digunakan adalah *Mutual Information*. *Mutual Information* adalah teknik pemilihan fitur berbasis filter yang membantu mengevaluasi jumlah informasi yang terkandung dalam suatu fitur, sehingga dapat diketahui pengaruh dari suatu fitur yang dapat membantu menentukan keputusan klasifikasi yang akurat [21]. Untuk menghitung *mutual information* dapat menggunakan persamaan berikut.

$$MI(x, y) = H(y) + H(x) - H(y|x) \quad (1)$$

$$MI(x, y) = - \sum_{i,j} p(y_j, x_i) \log_2 \frac{p(y_j, x_i)}{p(y_j)p(x_i)} \quad (2)$$

2.6. Klasifikasi Data

Pada penelitian ini, klasifikasi data dilakukan dengan menggunakan algoritma *decision tree* C5.0. Dalam proses pembentukan pohon, algoritma C5.0 memilih atribut dengan nilai *information gain* tertinggi sebagai *root* dari pohon keputusan. Proses dimulai dengan menggunakan seluruh data sebagai akar, kemudian atribut yang dipilih membagi data menjadi *subset* berdasarkan nilai atribut tersebut [22]. Proses pembuatan pohon keputusan C5.0 mirip dengan algoritma C4.5, dimana kemiripan tersebut adalah perhitungan entropi dan *gain*. Jika dalam perhitungan algoritma C4.5 berhenti sampai pada perhitungan entropi dan *gain*, maka dalam algoritma C5.0 perhitungan tersebut masih dilanjutkan perhitungan *gain ratio* dengan menggunakan hasil dari *gain* dan entropi yang telah dihitung sebelumnya [23]. Berikut proses perhitungan algoritma C5.0 dengan menggunakan rumus persamaan berikut.

$$entropy(s) = - \sum_{i=1}^n P_i * \log_2 p_i \quad (3)$$

Dalam persamaan ini, S adalah himpunan kasus, n adalah jumlah banyaknya partisi S , dan p_i adalah porsi dari setiap partisi S_i terhadap S . Nilai ini dihitung dengan mengalikan proporsi setiap kelas dengan logaritma basis 2 dari proporsinya, lalu menjumlahkan seluruh hasilnya untuk mendapatkan nilai entropy secara keseluruhan. Setelah nilai entropy dari setiap kelas didapatkan, langkah selanjutnya adalah mencari information gain dengan menggunakan persamaan berikut:

$$\text{information gain}(S, A) = \text{entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{entropy}(S_i) \quad (4)$$

Nilai information gain ini didapatkan dengan mengurangi nilai entropy (S) dengan total entropy dari setiap atribut (A). Dalam persamaan tersebut, $|S_i|$ adalah jumlah kasus pada nilai atribut, dan $|S|$ adalah jumlah total kasus. Kemudian setelah kedua tahap tersebut selesai, maka dapat mencari nilai *Gain Ratio* dengan menggunakan perhitungan dengan rumus persamaan berikut ini:

$$\text{GainRatio} = \frac{\text{information gain}(S, A)}{\sum_{i=1}^n \text{entropy}(S_i)} \quad (5)$$

Dari hasil perhitungan, *gain ratio* terbesar akan terpilih sebagai akar (*root*) dan yang rendah akan menjadi cabang. Kemudian proses *gain ratio* diulang sampai masing-masing cabang pada semua kelas memiliki kelasnya.

2.7. Evaluasi Model

Pada tahap evaluasi ini, penilaian dilakukan berdasarkan hasil *confusion matrix* yang digunakan untuk menghitung *performance matrix*. *Confusion matrix* adalah alat dalam melakukan mengevaluasi kinerja model klasifikasi dalam data *mining* [24]. Matriks ini mengukur kinerja atau tingkat kebenaran proses klasifikasi model dengan menghitung nilai akurasi, presisi, *recall*, dan *F1-Score* [25].

1. Akurasi, adalah rasio prediksi yang benar terhadap total data.

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

2. Presisi, adalah rasio prediksi benar positif terhadap keseluruhan hasil prediksi positif.

$$\text{precision} = \frac{TP}{TP+FP} \quad (7)$$

3. *Recall*, merupakan rasio prediksi positif yang benar terhadap keseluruhan data sebenarnya yang positif.

$$\text{recall} = \frac{TP}{TP+FN} \quad (8)$$

4. *F1-Score*, perbandingan rata-rata presisi dan recall.

$$F1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah rekap data pada semester ganjil 2024/2025 angkatan 2018 – 2023 yang diperoleh dari UPA TIK Universitas Jambi. Data dibagi menjadi 2 kategori yaitu data mahasiswa angkatan 2022-2023 dan mahasiswa angkatan 2018-2021. Mahasiswa angkatan 2022–2023 terdiri dari 11.053 data yang mencakup 17 atribut akademik dari SIAKAD. Sementara itu, data mahasiswa angkatan 2018–2021 terdiri dari 7.717 data dan 22 atribut yang berasal dari SIAKAD dan ELISTA.

3.2. Preprocessing Data

3.2.1. Seleksi Data

Pada tahap ini terdiri dari beberapa proses, yaitu penambahan atribut baru yang berasal dari kombinasi atribut yang telah ada, pengelompokan atribut untuk nilai yang memiliki makna serupa, dan penghapusan atribut yang tidak diperlukan.

1. Mahasiswa Angkatan 2022-2023

Tabel 1 merupakan rincian atribut setelah proses seleksi data pada mahasiswa angkatan 2022-2023 yang terdiri dari 15 atribut.

Tabel 1. Atribut Mahasiswa 2022-2023

Nama Atribut	Keterangan Singkat
angkatan	Informasi tahun masuk mahasiswa
asal_sekolah, pekerjaan_ortu, penghasilan_ortu, ukt, jenis_tinggal	Informasi latar belakang keluarga, riwayat pendidikan, biaya kuliah dan tempat tinggal
ipk_2, ipk_3, ipk_4, total_sks, jml_nilai_d, jml_mk_lulus, jml_mk_mengulang, jml_absen, jml_izin	Performa dan aktivitas akademik pada semester tertentu

2. Mahasiswa Angkatan 2018-2021

Tabel 2 merupakan rincian atribut setelah proses seleksi data pada mahasiswa angkatan 2022-2023 yang terdiri dari 22 atribut.

Tabel 2. Atribut Mahasiswa 2018-2021

Nama Atribut	Keterangan Singkat
angkatan	Informasi tahun masuk mahasiswa
asal_sekolah, pekerjaan_ortu, penghasilan_ortu, ukt, jenis_tinggal	Informasi latar belakang keluarga, riwayat pendidikan, biaya kuliah dan tempat tinggal
ipk_2, ipk_3, ipk_4, ipk_8, sks_4, sks_8, jml_nilai_d, jml_mk_mengulang, jml_absen, jml_izin, jarak_bulan_pengajuan_judul_ke_sempurna, rata_rata_bimbingan, durasi_hari_pengerjaan_ta, mengulang_seminar_sidang, jml_bimbingan, jml_kontrak_ta,	Peforma dan aktivitas akademik pada semester tertentu
	Aktivitas tugas akhir mahasiswa

Pemilihan atribut ini didasarkan pada dua faktor utama yaitu relevansi teoretis dari penelitian terdahulu dan ketersediaan data pada sistem akademik dan tugas akhir. Secara umum, atribut-atribut tersebut dikelompokkan ke dalam empat kategori utama.

1. Atribut Angkatan

Atribut merupakan salah satu atribut yang digunakan dalam proses pelabelan status mahasiswa.

2. Atribut Latar Belakang Demografi

Pemilihan atribut seperti asal_sekolah, pekerjaan_ortu, penghasilan_ortu, ukt, dan jenis_tinggal didasarkan pada temuan dari beberapa penelitian yang menunjukkan bahwa faktor-faktor ini memiliki korelasi signifikan dengan risiko *drop out* mahasiswa [26], [27], [28].

3. Atribut Performa dan Aktivitas Akademik

Atribut ini menjadi inti dalam mengukur performa mahasiswa dan dipilih berdasarkan indikator utama evaluasi studi [29]. Atribut seperti ipk, total_sks, jml_nilai_d, jml_mk_lulus, dan jml_mk_mengulang telah terbukti berpengaruh terhadap status kelulusan [30], [31], [32], [33], [34]. Selain itu, jumlah absen dan jumlah izin juga digunakan sebagai indikator kehadiran yang berkorelasi dengan risiko drop out [35].

4. Atribut Aktivitas Tugas Akhir

Poin kebaruan dalam penelitian ini adalah penggunaan atribut terkait aktivitas tugas akhir (jml_kontrak_ta, rata_rata_bimbingan, durasi_pengerjaan_ta, mengulang_seminar_sidang, jml_bimbingan). Atribut-atribut ini, yang berasal dari sistem ELISTA, sangat relevan untuk memprediksi keterlambatan studi, terutama pada mahasiswa tingkat akhir [36]. Penggunaan data ini memungkinkan analisis yang lebih komprehensif dibandingkan penelitian sebelumnya.

3.2.2. Pembersihan Data

Pada tahap ini, dilakukan identifikasi dan penanganan terhadap nilai yang hilang (*missing value*) agar tidak memengaruhi kualitas analisis dan proses pelatihan model. Penanganan *missing value* dilakukan dengan cara yang sama pada kedua kategori mahasiswa. *Missing value* pada atribut kategorikal diisi menggunakan modus, sedangkan atribut numerik menggunakan median. Data dengan *missing value* pada atribut penting untuk pelabelan seperti Indeks Prestasi Kumulatif (IPK) dan Satuan Kredit Semester (SKS) dihapus guna menjaga keakuratan label. Seluruh proses dilakukan menggunakan R. Tabel 3 merupakan rincian jumlah data setelah dibersihkan.

3.2.3. Pelabelan Data

Pada tahap ini ditentukan atribut target yaitu status mahasiswa yang dikelompokkan kedalam dua kelas yaitu Berisiko DO dan Tidak Berisiko DO. Pelabelan ini berdasarkan Peraturan Rektor Universitas Jambi Nomor 31 Tahun 2024 tentang Evaluasi Pendidikan. Pelabelan ditentukan berdasarkan atribut angkatan, IPK, dan SKS, dengan menggunakan kode R yang disesuaikan untuk masing-masing kelompok angkatan. Setelah pelabelan selesai, atribut tersebut dihapus dari dataset untuk mencegah data *leakage*, sehingga model tidak memperoleh informasi langsung dari atribut yang menjadi dasar pelabelan. Tabel 4 jumlah data pelabelan pada masing masing kategori.

Tabel 3. Jumlah Data Setelah Pembersihan

	Mahasiswa Angkatan 2022 – 2023	Mahasiswa Angkatan 2018 – 2021
Data awal	11.053	7.717
Data yang dihapus	0	2.894
Data bersih	11.053	4.823

Tabel 4. Jumlah Data Hasil Pelabelan

Kelas	Mahasiswa Angkatan 2022 – 2023	Mahasiswa Angkatan 2018 – 2021
Tidak Berisiko DO	10.876	4.425
Berisiko DO	177	398
Total Data	11.053	4.823

Berdasarkan Tabel 4, terlihat bahwa jumlah mahasiswa yang dikategorikan Berisiko DO jauh lebih sedikit dibandingkan dengan mahasiswa yang Tidak Berisiko DO. Hal ini menunjukkan bahwa data bersifat tidak seimbang, sehingga diperlukan penanganan pada tahap selanjutnya untuk menghindari bias dalam proses pemodelan.

3.3. Penanganan Data Tidak Seimbang

Pelabelan data sebelumnya menunjukkan adanya ketidakseimbangan antara jumlah mahasiswa yang tidak berisiko DO dan yang berisiko DO. Ketidakseimbangan ini dapat membuat model bias terhadap kelas mayoritas, sehingga dilakukan penyeimbangan data menggunakan metode *Random Undersampling* di R dengan bantuan *library dplyr*. Tabel 5 merupakan rincian distribusi data sebelum dan sesudah dilakukan *undersampling*:

Tabel 5. Jumlah Data Sebelum dan Sesudah *Undersampling*

Angkatan	Kelas	Jumlah Data Awal	Jumlah Data Setelah <i>Undersampling</i>
2022-2023	Tidak Berisiko DO	10.876	177
	Berisiko DO	177	177
	Total	11.053	354
2018-2021	Tidak Berisiko DO	4.425	398
	Berisiko DO	398	398
	Total	4.823	796

3.4. Pembagian Data

Pada tahap ini, data dibagi menjadi dua bagian yaitu *training* dan *testing*. Pembagian dilakukan dengan proporsi 80% untuk *training* dan 20% untuk *testing*. Tabel 6 adalah jumlah data setelah dilakukan pembagian.

Tabel 6. Jumlah Data Setelah Pembagian

Angkatan	Kelas	Jumlah Data	Data <i>Training</i>	Data <i>Testing</i>
2022-2023	Tidak Berisiko DO	177	142	35
	Berisiko DO	177	142	35
	Total	354	284	70
2018-2021	Tidak Berisiko DO	398	319	79
	Berisiko DO	398	319	79
	Total	796	638	158

3.5. Klasifikasi *Decision Tree C5.0*

Algoritma 1 digunakan dalam proses ini adalah *decision tree C5.0*. Pada tahap ini, data *training* digunakan untuk mempelajari pola-pola yang terdapat dalam data, sehingga menghasilkan model klasifikasi. Selanjutnya, performa model diuji menggunakan data testing untuk melihat akurasi. Proses pemodelan ini menggunakan R dengan memanfaatkan *library C5.0*.

Algoritma 1

```
# Pemodelan C5.0
library(C50)
features <- setdiff(names(trainData), "status")

model_c50 <- C5.0(x = trainData[, features],
                  y = trainData$status)

pred_test <- predict(model_c50, testData[, features], type = "class")
conf_matrix_test <- confusionMatrix(pred_test, testData$status, positive = "Berisiko DO")
print(conf_matrix_test)
```

Algoritma 1 digunakan untuk proses klasifikasi terhadap data pengujian menggunakan *model decision tree C5.0*. Model (*model_c50*) yang sebelumnya telah dilatih pada data *training* digunakan untuk memprediksi kelas atau status mahasiswa pada data uji (*testData*). dengan menyertakan seluruh fitur kecuali atribut target. Hasil prediksi disimpan dalam variabel (*pred_test*). Selanjutnya, performa model dievaluasi dengan membandingkan hasil prediksi dengan label sebenarnya menggunakan *confusion matrix*. Tabel 7 merupakan hasil performa model *decision tree C5.0* pada masing-masing kategori mahasiswa:

Tabel 7. Peforma Model *Decision Tree C5.0*

Angkatan	Akurasi	Presisi	Recall	F1-Score
2022 – 2023	87,14%	90,62%	82,86%	86,57%
2018 – 2021	93,04%	91,46%	94,94%	93,17%

3.6. Klasifikasi *Decision Tree C5.0* dengan Seleksi Fitur

Seleksi fitur dilakukan untuk mengidentifikasi atribut yang paling relevan dalam memengaruhi prediksi potensi *drop out* mahasiswa. Metode yang digunakan adalah *Mutual Information*, yaitu teknik seleksi fitur yang mengukur ketergantungan antara masing-masing fitur dengan kelas target. Nilai *Mutual Information* yang tinggi menunjukkan bahwa suatu fitur memiliki kontribusi yang besar terhadap proses klasifikasi. Tahap seleksi fitur ini menggunakan *library infotheo* pada R.

Algoritma 2

```
# Seleksi Fitur dengan Mutual Information
X_train <- trainData[, -which(names(trainData) == "status")]
y_train <- trainData$status

mi_scores <- sapply(X_train, function(x) {
  if (!is.factor(x)) {
    x_disc <- discretize(x)
  } else {
    x_disc <- x
  }
  y_disc <- discretize(as.numeric(y_train))
  mutinformation(x_disc, y_disc)
})

feature_importance <- data.frame(
  attributes = names(mi_scores),
  importance = mi_scores
)
feature_importance <- feature_importance[order(feature_importance$importance,
                                              decreasing = TRUE), ]
print(feature_importance)
```

Algoritma 2 ini digunakan untuk menghitung skor *Mutual Information* dari setiap atribut dalam data training terhadap variabel target status. Pertama, fitur prediktor (*X_train*) dipisahkan dari variabel target (*y_train*). Kemudian, setiap atribut dalam *X_train* dihitung skor *Mutual Information*-nya terhadap target menggunakan fungsi (*mutinformation*). Jika suatu atribut bukan bertipe faktor, maka akan didiskretisasi terlebih dahulu agar kompatibel dengan perhitungan *Mutual Information*. Nilai *Mutual Information* dari

setiap atribut kemudian dikumpulkan dalam bentuk data frame bernama *feature_importance*, yang berisi nama atribut dan nilai pentingnya. Berikut merupakan hasil perhitungan nilai *Mutual Information* terhadap atribut pada masing-masing kategori mahasiswa:

1. Mahasiswa Angkatan 2022-2023

Tabel 8. Nilai *Mutual Information* Mahasiswa 2022-2023

Atribut	Nilai <i>Mutual Information</i>
jml_absen	0.314353275
jml_mk_lulus	0.263536682
jml_mk_nilai_d	0.090164504
ukt	0.051858529
penghasilan_ortu	0.030852992
jml_mk_mengulang	0.030658137
pekerjaan_ortu	0.014352290
jenis_tinggal	0.011617976
jml_izin	0.004255623
asal_sekolah	0.004103960

Berdasarkan hasil perhitungan *Mutual Information* terhadap masing-masing atribut pada Tabel 8, diperoleh bahwa atribut jumlah absen, jumlah mata kuliah lulus, dan jumlah nilai D memiliki kontribusi informasi paling tinggi terhadap label kelas. Untuk menentukan fitur mana yang dipertahankan, dilakukan pengujian terhadap berbagai nilai ambang batas (*threshold*). Berikut merupakan hasil klasifikasi berdasarkan *threshold* yang digunakan:

Tabel 9. *Thershold* Mahasiswa 2022-2023

<i>Thershold</i>	Akurasi	Presisi	<i>Recall</i>	<i>F1-Score</i>
>0.03	88,57%	90,91%	85,71%	88,24%
>0.05	87,14%	90,26%	82,86%	86,57%
>0.09				
>0.2	82,86%	89,66%	74,29%	81,25%

Berdasarkan hasil perbandingan pada berbagai nilai *threshold* Tabel 9, dipilih *threshold* >0,03 karena memberikan performa paling optimal dengan akurasi sebesar 88,57%, presisi 90,91%, *recall* 85,71%, dan *F1-score* 88,24%. Pada *threshold* ini, terdapat 6 atribut yang dipertahankan, yaitu jumlah absen (*jml_absen*), jumlah mata kuliah lulus (*jml_mk_lulus*), jumlah nilai D (*jml_mk_nilai_d*), UKT, penghasilan orang tua (*penghasilan_ortu*), dan jumlah mata kuliah mengulang (*jml_mk_mengulang*).

2. Mahasiswa Angkatan 2018-2021

Tabel 10. Nila *Mutual Information* Mahasiswa 2018-2021

Atribut	Nilai <i>Mutual Information</i>
durasi_hari_pengerjaan_ta	0.447796283
jml_kontrak_ta	0.399868985
jml_mk_mengulang	0.395299550
pengajuan_judul_ke_semprom	0.179905567
ukt	0.096227976
jenis_tinggal	0.063286163
rata_rata_bimbingan	0.062123574
jml_mk_nilai_d	0.060729682
jml_absen	0.051989456
penghasilan_ortu	0.037189585
jml_izin	0.017184517
mengulang_seminar_sidang	0.013809198
jml_bimbingan	0.007413686
pekerjaan_ortu	0.005837672
asal_sekolah	0.003903298

Tabel 10 menunjukkan bahwa atribut durasi pengerjaan tugas akhir, jumlah kontrak tugas akhir, dan jumlah mata kuliah mengulang memiliki nilai *Mutual Information* tertinggi, yang mengindikasikan tingkat pengaruh yang cukup besar terhadap label kelas. Selanjutnya dilakukan pengujian terhadap beberapa nilai ambang batas (*threshold*) untuk menentukan batas yang optimal dalam seleksi fitur.

Berdasarkan hasil pengujian dengan berbagai nilai *threshold* Tabel 11, dipilih *threshold* >0,1 karena menghasilkan performa yang tinggi dan seimbang dengan akurasi sebesar 94,30%, presisi 96,05%, *recall* 92,41%, serta *F1-score* 94,19%. Pada *threshold* ini, jumlah fitur yang dipertahankan hanya 3 atribut, yaitu *durasi_hari_pengerjaan_ta*, jumlah kontrak tugas akhir (*jml_kontrak_ta*), dan jumlah mata kuliah mengulang (*jml_mk_mengulang*).

Tabel 11. *Thershold* Mahasiswa 2018-2021

<i>Thershold</i>	Akurasi	Presisi	<i>Recall</i>	<i>F1-Score</i>
>0.03	93,67%	92,59%	94,94%	93,75%
>0.05				
>0.06				
>0.09	93,67%	94,81%	92,41%	93,59%
>0.1	94,30%	96,05%	92,41%	94,19%
>0.3	94,30%	93,75%	94,94%	94,34%

3.7. Evaluasi Model

Berikut merupakan perbandingan performa model algoritma *decision tree* C5.0 antara model dengan seleksi fitur dan tanpa seleksi fitur berdasarkan metrik akurasi, presisi, *recall*, dan *f1-score*:

1. Mahasiswa Angkatan 2022-2023

Tabel 12. Peforma Model Mahasiswa 2022-2023

Jenis Klasifikasi	Akurasi	Presisi	<i>Recall</i>	<i>F1-Score</i>
C5.0 tanpa Seleksi Fitur	87,14%	90,62%	82,86%	86,57%
C5.0 dengan Seleksi Fitur	88,57%	90,91%	85,71%	88,24%

Berdasarkan hasil evaluasi pada Tabel 12, model C5.0 tanpa seleksi fitur menghasilkan akurasi sebesar 87,14%, presisi 90,62%, *recall* 82,86%, dan *F1-score* 86,57%. Sementara itu, model C5.0 dengan seleksi fitur memberikan peningkatan performa dengan akurasi mencapai 88,57%, presisi 90,91%, *recall* 85,71%, serta *F1-score* 88,24%. Hasil ini menunjukkan bahwa penerapan seleksi fitur mampu meningkatkan kinerja model pada seluruh metrik, terutama pada *recall* dan *F1-score*, sehingga model menjadi lebih seimbang dalam mendeteksi mahasiswa yang berisiko *drop out* maupun yang tidak.

2. Mahasiswa Angkatan 2018-2021

Tabel 13. Peforma Model Mahasiswa 2018-2021

Jenis Klasifikasi	Akurasi	Presisi	<i>Recall</i>	<i>F1-Score</i>
C5.0 tanpa Seleksi Fitur	93,04%	91,46%	94,94%	93,17%
C5.0 dengan Seleksi Fitur	94,30%	93,75%	94,94%	94,34%

Pada Tabel 13 model C5.0 tanpa seleksi fitur menghasilkan akurasi sebesar 93,04%, dengan presisi 91,46%, *recall* 94,94%, dan *F1-score* 93,17%. Sementara itu, penerapan seleksi fitur meningkatkan performa model, ditunjukkan dengan akurasi yang naik menjadi 94,30%, presisi 93,75%, *recall* tetap stabil di 94,94%, serta *F1-score* yang meningkat menjadi 94,34%. Peningkatan ini menunjukkan bahwa seleksi fitur tidak hanya menyederhanakan jumlah atribut yang digunakan, tetapi juga mampu memperbaiki keseimbangan antara presisi dan *recall*, sehingga menghasilkan model yang lebih efektif dalam mengklasifikasikan mahasiswa berisiko *drop out*.

3.8. Visualisasi Hasil Klasifikasi

Dalam proses klasifikasi menggunakan algoritma *decision tree* C5.0, dihasilkan pohon keputusan dan aturan keputusan (*decision rules*) yang terbentuk dari struktur cabang pohon.

1. Mahasiswa Angkatan 2022-2023

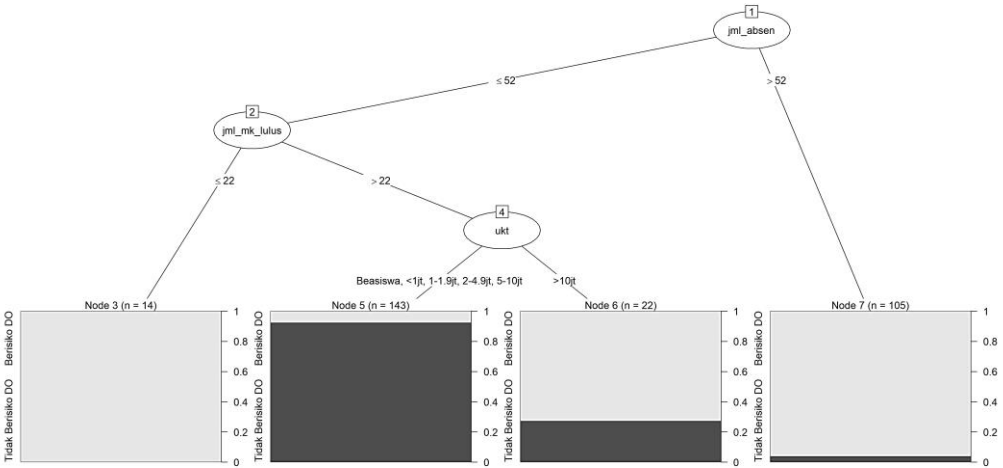
Gambar 2 menunjukkan pohon keputusan dari model dengan seleksi fitur, di mana jumlah absen (*jml_absen*) menjadi atribut utama dalam mengklasifikasikan mahasiswa angkatan 2022–2023 yang berisiko DO.

Rules Pohon Keputusan

Berdasarkan hasil pemodelan menggunakan pohon keputusan, diperoleh 4 aturan keputusan (*decision rules*) yang terbentuk dari struktur cabang pohon, yaitu pada Tabel 14.

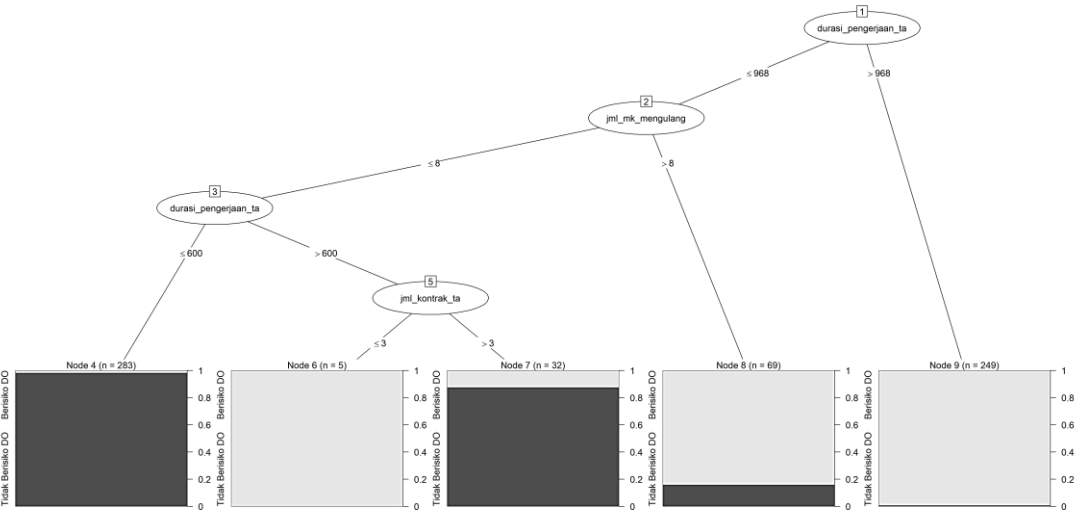
Tabel 14. Rules Mahasiswa 2022-2023

Rules	Kondisi
Rule 1	IF jml_absen > 52 THEN status = Berisiko DO
Rule 2	IF jml_absen ≤ 52 AND jml_mk_lulus ≤ 22 THEN status = Berisiko DO
Rule 3	IF jml_absen ≤ 52 AND jml_mk_lulus > 22 AND UKT IN {Beasiswa, Kurang dari Rp1.000.000, Rp1.000.000 -Rp1.999.999, Rp2.000.000 – Rp4.999.999, Rp5.000.000 – Rp10.000.000} THEN status = Tidak Berisiko DO
Rule 4	IF jumlah absen ≤ 52 AND jml_mk_lulus > 22 AND UKT = Lebih dari Rp 10.000.000 THEN status = Berisiko DO



Gambar 2. Pohon Keputusan Mahasiswa 2022-2023

2. Mahasiswa Angkatan 2018-2021



Gambar 3. Pohon Keputusan Mahasiswa 2018-2021

Gambar 3 menunjukkan hasil pohon keputusan dari model klasifikasi dengan seleksi fitur, dimana atribut durasi pengerjaan tugas akhir (durasi_pengerjaan_ta) merupakan akar dari pohon keputusan, yang menunjukkan bahwa atribut ini memiliki pengaruh paling besar dalam menentukan risiko DO pada mahasiswa angkatan 2018 – 2021.

Rules Pohon Keputusan

Sedangkan pada pemodelan kategori mahasiswa angkatan 2018-2021, diperoleh 5 aturan keputusan (decision rules) yang terbentuk dari struktur cabang pohon, yaitu pada Tabel 15.

Tabel 15. Rules Mahasiswa 2018-2021

Rules	Kondisi
Rule 1	IF durasi_pengerjaan_ta > 968 THEN status = Berisiko DO
Rule 2	IF durasi_pengerjaan_ta ≤ 968 AND jml_mk_mengulang > 8 THEN status = Berisiko DO
Rule 3	IF durasi_pengerjaan_ta ≤ 968 AND jml_mk_mengulang ≤ 8 AND durasi_pengerjaan_ta ≤ 600 THEN status = Tidak Berisiko DO
Rule 4	IF durasi_pengerjaan_ta ≤ 968 AND jml_mk_mengulang ≤ 8 AND durasi_pengerjaan_ta > 600 AND jml_kontrak_ta ≤ 3 THEN status = Berisiko DO
Rule 5	IF durasi_pengerjaan_ta ≤ 968 AND jml_mk_mengulang ≤ 8 AND durasi_pengerjaan_ta > 600 AND jml_kontrak_ta > 3 THEN status = Tidak Berisiko DO

3.9. Diskusi

Hasil klasifikasi menunjukkan bahwa penerapan seleksi fitur mampu meningkatkan kinerja model C5.0 pada kedua kelompok mahasiswa. Peningkatan akurasi, presisi, *recall*, dan *F1-score* membuktikan bahwa seleksi fitur berhasil mengeliminasi atribut-atribut yang kurang relevan, sehingga model mampu mengidentifikasi pola dengan lebih efektif. Temuan ini sejalan dengan studi sebelumnya Putri et al., 2024 yang juga menunjukkan bahwa seleksi fitur secara signifikan dapat meningkatkan akurasi dan efisiensi model klasifikasi [15].

Dari hasil pohon keputusan, diketahui bahwa atribut kunci untuk setiap kelompok mahasiswa berbeda. Pada mahasiswa tingkat awal (angkatan 2022-2023), atribut jumlah absen menjadi faktor dominan dalam menentukan risiko *drop out*. Hal ini mengindikasikan bahwa pada masa-masa awal perkuliahan, tingkat kehadiran terhadap kegiatan akademik menjadi faktor paling kuat. Namun, pada mahasiswa tingkat akhir (angkatan 2018-2021), atribut durasi pengerjaan tugas akhir menjadi akar pohon keputusan, menegaskan bahwa pada tahap akhir studi, proses penyelesaian tugas akhir menjadi hambatan utama. Hal ini juga membuktikan bahwa faktor risiko *drop out* bersifat dinamis dan dapat berubah seiring masa studi mahasiswa.

Meskipun demikian, penelitian ini memiliki keterbatasan karena hanya menganalisis data dari satu universitas. Oleh karena itu, hasil ini perlu didukung dengan analisis kondisi lapangan yang lebih luas agar kebijakan yang dibuat bisa lebih tepat sasaran. Penelitian selanjutnya dapat memperluas cakupan dengan mengaplikasikan model ini pada universitas lain atau mengintegrasikan variabel lain yang lebih mendalam untuk mengidentifikasi pola yang lebih kompleks.

4. KESIMPULAN

Berdasarkan hasil penelitian, algoritma C5.0 terbukti mampu memberikan performa klasifikasi yang baik pada kedua kelompok data mahasiswa. Pada angkatan 2022–2023, model tanpa seleksi fitur menghasilkan akurasi 87,14%, sedangkan penerapan seleksi fitur dengan *threshold* >0,03 mampu meningkatkan akurasi menjadi 88,57% dengan keseimbangan presisi dan *recall* yang lebih baik. Sementara itu, pada angkatan 2018–2021, model tanpa seleksi fitur memiliki akurasi 93,04%, dan meningkat menjadi 94,30% setelah dilakukan seleksi fitur dengan *threshold* >0,1. Hasil ini menunjukkan bahwa seleksi fitur tidak hanya membuat model lebih efisien, tetapi juga memperkuat performa klasifikasi. Atribut yang paling berpengaruh juga berbeda pada masing-masing kelompok, di mana jumlah absen menjadi indikator utama pada mahasiswa angkatan 2022–2023, sedangkan durasi pengerjaan tugas akhir lebih dominan pada mahasiswa angkatan 2018–2021.

REFERENSI

- [1] A. Kurniasih and K. Isyara, "Penggunaan Metode SMOTE pada Naïve Bayes Gaussian untuk Klasifikasi Mahasiswa Drop Out," 2023, Accessed: Dec. 16, 2024. [Online]. Available: <https://conference.upnvj.ac.id/index.php/senamika/article/view/2585/1937>
- [2] BANPT, BANPT. 2019. Accessed: Dec. 16, 2024. [Online]. Available: https://www.banpt.or.id/wp-content/uploads/2019/09/Lampiran-02-PerBAN-PT-3-2019-Kriteria-dan-Prosedur-IAPT-3_0.pdf
- [3] Permendikbud No. 49 Tahun 2014 tentang Standar Nasional Pendidikan Tinggi (SNPT). 2014. Accessed: Dec. 02, 2024. [Online]. Available: <https://peraturan.go.id/id/permendikbud-no-49-tahun-2014>
- [4] "https://dashboard.unja.ac.id/." Accessed: Feb. 19, 2025. [Online]. Available: <https://dashboard.unja.ac.id/>
- [5] M. C. Pascoe, S. E. Hetrick, and A. G. Parker, "The impact of stress on students in secondary school and higher education," *Int J Adolesc Youth*, vol. 25, no. 1, pp. 104–112, Dec. 2020, doi: 10.1080/02673843.2019.1596823.
- [6] A. W. Abdullah and A. Muhid, "Social Support, Academic Satisfaction, and Student Drop Out Tendency/ Dukungan Sosial, Academic Satisfaction, dan Kecenderungan Drop Out pada Mahasiswa," *Psikoislamika : Jurnal Psikologi dan Psikologi Islam*, vol. 18, no. 1, pp. 174–187, Jun. 2021, doi: 10.18860/psikoislamika.v18i1.11546.

- [7] N. Ulinuha and A. Fanani, "Klasifikasi Status Drop Out Mahasiswa Menggunakan Naïve Bayes dengan Seleksi Fitur Information Gain," *Techno.Com*, vol. 22, no. 4, pp. 1014–1025, Nov. 2023, doi: 10.33633/tc.v22i4.9004.
- [8] P. W. Rahayu *et al.*, "Buku Ajar Data Mining," 2024. [Online]. Available: <https://www.researchgate.net/publication/377415198>
- [9] E. Prasetyo, *Data mining mengolah data menjadi informasi menggunakan matlab*. 2014.
- [10] L. Karlitasari, I. W. Sriyasa, I. Wahyudi, and H. B. Santosi, "Prediksi Morfologi Jamur Menggunakan Algoritma C5.0," 2023.
- [11] N. L. W. S. R. Ginantra, F. N. Arifah, and A. H. Wijaya, *Data Mining dan Penerapan Algoritma*. Yayasan Kita Menulis, 2021. Accessed: Jan. 01, 2025. [Online]. Available: <http://repo.handayani.ac.id/148/1/1.2%20FullBook%20Data%20Mining%20dan%20Penerapan%20Algoritma.pdf>
- [12] J. R. Vergara and P. Estévez, "A review of feature selection methods based on mutual information," Jan. 2014. doi: 10.1007/s00521-013-1368-0.
- [13] E. Hariyanti, D. P. Hostiadi, Anggreni, Yohanes Priyo Atmojo, I Made Darma Susila, and I. Tangkawarow, "Analisis Perbandingan Metode Seleksi Fitur pada Model Klasifikasi Decision Tree untuk Deteksi Serangan di Jaringan Komputer," *Jurnal Sistem dan Informatika (JSI)*, vol. 18, no. 2, pp. 208–217, May 2024, doi: 10.30864/jsi.v18i2.615.
- [14] I. G. N. D. Adnyana, R. M. Arjuna, A. N. Indraini, and D. S. Pasvita, *Pengaruh Seleksi Fitur Pada Algoritma Machine Learning Untuk Memprediksi Pembatalan Pesanan Hotel*. 2021.
- [15] Putri, E. Sulistianingsih, N. Imro'ah, and N. N. Debatara, "Application Of C4.5 Algorithm With Feature Selection In Classification Of Discharge Status Of Head Injury Patients," *VARIANCE: Journal of Statistics and Its Applications*, vol. 6, no. 2, pp. 165–174, Oct. 2024, doi: 10.30598/variancevol6iss2page165-174.
- [16] Kusrini and E. T. Luthfi, *Algoritma Data Mining*. Penerbit Andi, 2009.
- [17] N. A. Azhar, M. S. Mohd Pozi, A. Mohamed Din, and A. Jatowt, "An Investigation of SMOTE based Methods for Imbalanced Datasets with Data Complexity Analysis," *IEEE Trans Knowl Data Eng*, pp. 1–1, 2022, doi: 10.1109/TKDE.2022.3179381.
- [18] V. H. Barella, L. P. F. Garcia, M. C. P. de Souto, A. C. Lorena, and A. C. P. L. F. de Carvalho, "Assessing the data complexity of imbalanced datasets," *Inf Sci (N Y)*, vol. 553, pp. 83–109, Apr. 2021, doi: 10.1016/j.ins.2020.12.006.
- [19] A. F. Pulungan and D. Selvida, "Kombinasi Metode Sampling pada Pengklasifikasian Data Tidak Seimbang Menggunakan Algoritma Support Vector Machine (SVM)," 2022.
- [20] I. Colanus, R. Drajana, and A. Bode, "Prediksi Status Penderita Stunting Pada Balita Provinsi Gorontalo Menggunakan K-Nearest Neighbor Berbasis Seleksi Fitur Chi Square," *Jurnal Nasional Komputasi dan Teknologi Informasi*, vol. 5, no. 2, 2022.
- [21] A. Hanafi, A. Adiwijaya, and W. Astuti, "Klasifikasi Multi Label pada Hadis Bukhari Terjemahan Bahasa Indonesia Menggunakan Mutual Information dan k-Nearest Neighbor," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 9, no. 3, pp. 357–364, Sep. 2020, doi: 10.32736/sisfokom.v9i3.980.
- [22] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier Science, 2011.
- [23] M. Zainuri, M. H. Fahmi, and R. A. Hamdhana, "Komparasi Metode Klasifikasi Algoritma C5.0 dan Naïve Bayes untuk Menentukan Jurusan Siswa," *Jurnal Sistem Informasi dan Informatika (JUSIFOR)*, vol. 1, no. 1, pp. 1–8, May 2022, doi: 10.33379/jusifor.v1i1.1277.
- [24] M. R. Galela, "Tidak perlu confused dengan Confusion Matrix." Accessed: Jan. 01, 2025. [Online]. Available: <https://klc2.kemenkeu.go.id/kms/knowledge/tidak-perlu-confused-dengan-confusion-matrix-728befa0/detail/>
- [25] N. Nurainun, E. Haerani, F. Syafria, and L. Oktavia, "Penerapan Algoritma Naïve Bayes Classifier Dalam Klasifikasi Status Gizi Balita dengan Pengujian K-Fold Cross Validation," *Journal of Computer System and Informatics (JoSYC)*, vol. 4, no. 3, pp. 578–586, May 2023, doi: 10.47065/josyc.v4i3.3414.
- [26] J. Fredricka, L. Elfianty, and J. Wahyudi, "Implementasi Algoritma C-Means dan Algoritma Mixture dalam Pengclusteran Data Mahasiswa Drop Out," *JUKI: Jurnal Komputer dan Informatika*, vol. 4, 2022.
- [27] B. G. Sudarsono and A. U. Bani, "Prediksi Mahasiswa Berpotensi Berhenti Kuliah Secara Sepihak Menggunakan Data Mining Algoritma C4.5," 2020.
- [28] D. M. A. M. Sanjaya, A. A. I. I. Paramitha, and N. W. Utami, "Penerapan Data Mining untuk Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Algoritma C4.5: Studi Kasus STMIK Primakara," 2022.
- [29] Universitas Jambi, *Peraturan Rektor No 31 Tahun 2024 tentang Peraturan Akademik*. 2024.

-
- [30] A. Ramadhani, R. F. Noor, D. Vernanda, and T. Herdiawan, “Klasifikasi Mahasiswa Berpotensi Drop Out Menggunakan Algoritma C4.5 di Politeknik Negeri Subang,” *Jurnal Tekno Kompak*, vol. 18, no. 1, p. 101, Feb. 2024, doi: 10.33365/jtk.v18i1.3439.
 - [31] A. S. B. Lomi, A. A. Pekuwali, and R. T. Abineno, “Pengelompokan Mahasiswa Berpotensi Drop Outpada Program Studi Teknik Infomatika Menggunakan Metode K-Means Clustering,” pp. 340–351, Aug. 2024.
 - [32] H. Hermanto, “Prediksi Kelulusan dan Putus Studi Mahasiswa dengan Pendekatan Bertingkat pada Perguruan Tinggi,” *SIMADA (Jurnal Sistem Informasi dan Manajemen Basis Data)*, vol. 3, no. 2, pp. 140–148, Jan. 2021, doi: 10.30873/simada.v3i2.2359.
 - [33] T. A. Marzuqi, E. Kristiani, and Marcel, “Prediksi Mahasiswa Drop-Out Di Universitas XYZ,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 6, pp. 1345–1350, Dec. 2024, doi: 10.25126/jtiik.2024118689.
 - [34] N. Y. L. Gaol, “Prediksi Mahasiswa Berpotensi Non Aktif Menggunakan Data Mining dalam Decision Tree dan Algoritma C4.5,” *Jurnal Informasi & Teknologi*, pp. 23–29, Mar. 2020, doi: 10.37034/jidt.v2i1.22.
 - [35] A. Andri and P. Paulus, “Prediksi Mahasiswa Berisiko Drop Out (DO) dengan ADTree dan NNge,” *Jurnal SIFO Mikroskil*, vol. 22, no. 1, pp. 47–58, Aug. 2021, doi: 10.55601/jsm.v22i1.794.
 - [36] N. Wakhidah and Saemuri, “Faktor Terlambatnya Penyelesaian Studi Di Perguruan Tinggi Keagamaan Islam: Studi Penerapan Kebijakan Masa Studi Mahasiswa Di Indonesia,” vol. 6, Mar. 2025.