

## **Analisis Data Kelulusan Untuk Memprediksi Mahasiswa Berpotensi Dropout Menggunakan Algoritma C4.5**

**Ritham Tuntun<sup>1)</sup>, Oktaviar Rahmat<sup>2)</sup>, Mihuandayani<sup>3)</sup>, Michel Farrel Tomatala<sup>4)</sup>**

*<sup>1,2)</sup> Sistem Informasi, STMIK Multicom Bolaang Mongondow  
Jl. Sutoyo No.12, Kotamobagu, Sulawesi Utara  
Email : ritham@stmikmulticom.ac.id<sup>1)</sup>, oktaviar@stmikmulticom.ac.id<sup>2)</sup>,  
mihuandayani@stmikmulticom.ac.id<sup>3)</sup>, michel@stmikmulticom.ac.id<sup>4)</sup>*

### **Abstrak**

Pendidikan tinggi berperan penting dalam mencetak sumber daya manusia yang siap bersaing di dunia kerja. Namun, tingginya angka mahasiswa yang mengalami putus kuliah menjadi tantangan besar bagi perguruan tinggi, termasuk STMIK Multicom, yang dapat mempengaruhi reputasi dan akreditasi institusi. Penelitian ini bertujuan untuk menganalisis data kelulusan mahasiswa STMIK Multicom dari tahun 2019 hingga 2023 menggunakan algoritma pohon keputusan C4.5 untuk memprediksi potensi mahasiswa yang berisiko dropout. Algoritma C4.5, yang merupakan teknik data mining, memungkinkan pembentukan model prediksi yang mudah dipahami dengan menganalisis berbagai atribut mahasiswa, seperti status pekerjaan, pendapatan orang tua, status tempat tinggal, dan jarak tempat tinggal ke kampus. Hasil analisis menunjukkan bahwa model C4.5 menghasilkan akurasi 63%, recall 38%, dan presisi 62%, yang mengindikasikan potensi algoritma ini dalam memprediksi mahasiswa yang berisiko dropout. Penelitian ini diharapkan dapat memberikan kontribusi dalam menurunkan angka dropout di STMIK Multicom dan meningkatkan tingkat kelulusan mahasiswa.

**Kata Kunci : Data Mining; C4.5; Dropout; Klasifikasi; Orange Data Mining**

### **Abstract**

*Higher education plays a crucial role in developing human resources that are competitive in the workforce. However, the high dropout rate among students presents a significant challenge for universities, including STMIK Multicom, which can affect the institution's reputation and accreditation. This study aims to analyze the graduation data of STMIK Multicom students from 2019 to 2023 using the C4.5 decision tree algorithm to predict the potential for student dropout. The C4.5 algorithm, a data mining technique, enables the creation of an easily interpretable prediction model by analyzing various student attributes, such as employment status, parental income, residence status, and distance from home to campus. The analysis results show that the C4.5 model achieves an accuracy of 63%, a recall of 38%, and a precision of 62%, indicating the potential of this algorithm in predicting students at risk of dropping out. This study is expected to contribute to reducing dropout rates at STMIK Multicom and improving student graduation rates.*

**Keywords: Data Mining; C4.5; Dropout; Classification; Orange Data Mining**

### **1. PENDAHULUAN**

Pendidikan tinggi memainkan peran krusial dalam mempersiapkan sumber daya manusia yang siap menghadapi persaingan di dunia kerja. Meski demikian, salah satu tantangan utama yang dihadapi perguruan tinggi adalah tingginya jumlah mahasiswa yang terpaksa meninggalkan studi mereka. Hal ini tidak hanya berdampak negatif pada mahasiswa, tetapi juga pada reputasi dan kualitas akreditasi perguruan tinggi [1].

Di STMIK Multicom, masalah dropout mahasiswa menjadi perhatian serius. Tingginya tingkat mahasiswa yang tidak menyelesaikan studi tepat waktu dapat menurunkan tingkat kepercayaan masyarakat terhadap institusi ini. Selain itu, angka dropout yang tinggi juga mempengaruhi penilaian akreditasi, yang pada gilirannya berdampak pada tingkat kepercayaan publik dan calon mahasiswa yang akan menempuh pendidikan tinggi.

Untuk menyelesaikan masalah ini, dibutuhkan langkah-langkah yang efektif dan efisien

dalam mendeteksi mahasiswa yang berisiko mengalami putus kuliah. Adapun cara yang dapat digunakan yaitu melakukan analisis data mahasiswa dengan menggunakan teknik data mining. Data mining memungkinkan ekstraksi informasi penting dari data historis yang ada pada institusi. Adapun metode yang sering digunakan dalam melakukan analisis data dalam data mining yaitu pohon keputusan C4.5, yang mampu membangun model prediksi yang tepat dan akurat. Algoritma ini mampu menghasilkan pohon keputusan yang mudah dipahami dan diinterpretasikan, sehingga dapat menjadi alat bantu dalam pengambilan keputusan [2].

Algoritma pohon keputusan C4.5 pernah digunakan pada penelitian yang dilakukan oleh [3] pada kasus peminatan jurusan siswa berdasarkan kurikulum di SMA Negeri 1 Natar. Dalam penelitian tersebut algoritma C4.5 digunakan untuk menganalisis data peminatan siswa di SMA Negeri 1 Natar untuk jurusan IPA dan IPS. Analisis dilakukan berdasarkan pada zonasi, minat, prestasi, dan perpindahan tugas orang tua. Dengan digunakannya algoritma pohon keputusan C4.5 ini diharapkan adanya sebuah hasil klasifikasi dengan tingkat akurasi yang tinggi agar hasil klasifikasinya dapat memberikan hasil akhir pembelajaran yang maksimal bagi para siswa dalam proses kegiatan belajar mengajar. Implementasi algoritma C4.5 dalam penelitian ini menggunakan bantuan alat Weka. Hasil yang diperoleh pada penerapan algoritma C4.5 pada data peminatan siswa yaitu pada data a memperoleh tingkat akurasi 100%, dan pada data b memperoleh tingkat akurasi 80% sehingga disimpulkan algoritma C4.5 dapat digunakan serta dipercaya dalam menganalisis pola peminatan siswa.

Selanjutnya pernah ada penelitian yang telah dilakukan [4] berkaitan dengan algoritma C4.5 yang digunakan untuk menganalisis data untuk memberikan rekomendasi penerimaan beasiswa yayasan di SMK TI Airlangga. Dalam kasus penerimaan siswa baru yang akan menerima beasiswa, data mining digunakan untuk mendapatkan hasil yang bisa dipercaya yaitu dengan menggunakan algoritma klasifikasi C4.5. Dengan menggunakan teknik klasifikasi tersebut adanya hasil berupa rekomendasi untuk pihak sekolah agar dapat mempertimbangkan data calon siswa baru yang masuk untuk mendapat beasiswa. Data yang akan dianalisis berdasarkan nilai dari kriteria yang telah ditentukan pada masing-masing calon siswa yang mengajukan diri. Data-data yang sudah dikumpulkan kemudian akan dihitung menggunakan

metode C4.5 sehingga mendapatkan sebuah visual pohon keputusan yang akan dijadikan acuan untuk melakukan klasifikasi terhadap data calon penerima beasiswa. Dari hasil perhitungan menggunakan C4.5 didapatkan hasil berupa tingkat akurasi sebesar 57% dengan kriteria nilai yang paling berpeluang untuk mendapatkan rekomendasi untuk menerima beasiswa.

Dalam beberapa kasus dengan menggunakan data-data primer, algoritma C4.5 pohon keputusan juga pernah dibandingkan dengan algoritma lain. Seperti yang dilakukan oleh [5] yang membandingkan algoritma C4.5 dengan algoritma lain seperti NBC dan KNN pada data penerima bantuan PKH (program keluarga harapan). Program PKH ini merupakan salah satu program dari pemerintah untuk keluarga yang masuk dalam kategori keluarga kurang mampu sehingga dalam proses penentuan penerimanya dibutuhkan suatu pertimbangan agar mendapat hasil yang tepat dan akurat. Maka dari itu pada penelitian ini melakukan penerapan 3 algoritma yang salah satunya adalah C4.5 dengan tujuan menentukan algoritma yang memiliki hasil akurasi tertinggi sehingga dapat dijadikan algoritma yang melakukan klasifikasi terhadap data-data calon penerima program keluarga harapan. Terdapat 33 atribut pada data yang digunakan serta 544 jumlah data yang akan dianalisis. Setelah dilakukan penerapan masing-masing algoritma terhadap data yang sudah ada, didapatkan hasil tingkat akurasi C4.5 adalah yang paling tinggi yaitu sebesar 80,16%, diikuti secara berurutan NBC dan KNN masing-masing secara berurutan 77,51% dan 76,72%.

Penelitian yang dilakukan oleh [6] juga melakukan perbandingan antara algoritma C4.5 dan Naive bayes pada data penerima kredit. Algoritma dibandingkan untuk melihat algoritma manakah yang memperoleh kinerja tertinggi dalam melakukan klasifikasi penerima kredit. Beberapa masalah terkait penerima kredit yaitu terdapat banyaknya kesalahan penentuan penerima kredit yang dapat mengakibatkan kerugian pada pihak yang memberikan kredit kepada pengguna kredit. Berdasarkan data yang sudah ada di masa lampau maka data mining diharapkan mampu memberikan rekomendasi penerima kredit yang paling akurat. Maka dipilihlah 2 algoritma salah satunya adalah C4.5. Pada data dalam kasus ini terdapat 4 atribut pada data yang sudah disediakan dan 15 baris data dengan hasil keputusan berupa 'ditolak' atau 'diterima'. Setelah dilakukan penerapan kedua algoritma tersebut pada data yang sudah ada dengan mengambil hasil akhir berupa rekomendasi

'diterima', didapatkan hasil berupa tingkat akurasi pada algoritma naive bayes sebesar 86,67% sedangkan algoritma C4.5 memperoleh akurasi yang lebih tinggi yaitu sebesar 100%.

Penelitian ini bertujuan untuk menganalisis data kelulusan mahasiswa STMIK Multicom dari tahun 2019 hingga 2023 dengan memanfaatkan algoritma pohon keputusan C4.5. Melalui analisis ini, diharapkan dapat ditemukan pola-pola signifikan yang dapat digunakan untuk memprediksi potensi dropout mahasiswa. Hasil penelitian ini diharapkan dapat memberikan sumbangsih yang signifikan dalam usaha menurunkan angka putus kuliah dan meningkatkan tingkat kelulusan di STMIK Multicom.

## 2. METODE PENELITIAN

Pada poin ini berisi pembahasan terkait data mining, algoritma yang digunakan, metode pengujian, serta dataset yang digunakan pada penelitian ini.

### 2.1 Data Mining

Menurut [2] dalam buku Algoritma Data Mining, data mining merupakan proses untuk mengungkap atau membagi pengetahuan yang tersembunyi pada sekumpulan data. Proses ini memanfaatkan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi serta mengidentifikasi informasi yang bermanfaat dan pengetahuan terkait dari berbagai dataset besar. Salah satu teknik dalam data mining adalah algoritma klasifikasi, yang bertujuan untuk mengelompokkan data berdasarkan variabel target yang telah ada atau yang sudah ditentukan. Salah satu algoritma yang telah digunakan pada banyak penelitian dengan teknik data mining untuk menganalisis dan memprediksi pola adalah algoritma C4.5.

### 2.2 Algoritma C4.5

Metode C4.5 termasuk dalam kategori pohon keputusan, untuk membangun struktur yang menyerupai diagram alur (flowchart), di mana setiap simpul internal melakukan pengujian terhadap atribut tertentu, setiap cabang mewakili hasil dari pengujian tersebut, dan simpul daun (atau simpul terminal) menunjukkan label kelas yang ditentukan berdasarkan hasil pengujian. Simpul yang terletak di bagian paling tinggi atau terlebih dahulu dari pohon keputusan ini disebut sebagai root, yang merupakan titik awal untuk proses klasifikasi. Algoritma C4.5 merupakan algoritma pembelajaran terawasi (supervised learning), yang memerlukan

klasifikasi pra-variabel sasaran agar proses klasifikasi dapat dilakukan dengan benar. Untuk membangun model yang efektif, data pelatihan perlu dipersiapkan dengan baik agar mengandung nilai dari variabel target yang relevan. Algoritma ini kemudian mengevaluasi atribut-atribut serta penggunaan pengukuran statistik yang cermat, seperti informasi nilai gain serta hasil perhitungan entropy. Informasi nilai gain sendiri berfungsi sebagai ukuran untuk menilai sejauh mana efektivitas atribut dalam mengklasifikasikan data secara akurat dan efisien [7].

Algoritma C4.5 termasuk dalam kategori algoritma pohon keputusan yang memanfaatkan dua jenis input utama, yaitu data pelatihan dan data uji (testing samples) [4]. Dalam penerapannya, data pelatihan digunakan untuk membangun pohon keputusan yang nantinya akan diuji untuk memastikan validitas dan keakuratan model yang dibentuk. Sementara itu, data sampel digunakan sebagai parameter untuk melakukan proses klasifikasi terhadap data yang belum terklasifikasi. Algoritma C4.5 sendiri merupakan salah satu metode yang efektif dalam membangun pohon keputusan dengan menggunakan data pelatihan yang tersedia untuk menciptakan aturan klasifikasi yang jelas dan tepat. Algoritma ini dikembangkan oleh Ross Quinlan sebagai penyempurnaan dari algoritma ID3 yang sebelumnya juga diciptakan oleh Quinlan. Beberapa perbaikan yang dibawa oleh C4.5 antara lain kemampuannya untuk menangani nilai yang hilang (missing values), kemampuan untuk memproses data kontinu, serta penerapan teknik pruning yang bertujuan untuk mengurangi kompleksitas pohon keputusan dan meningkatkan akurasi hasil klasifikasi [8].

Menurut [2], ada beberapa tahap dalam pembuatan pohon keputusan menggunakan algoritma C4.5, yaitu:

- 1) Mempersiapkan data latih yang umumnya berasal dari data historis yang telah melalui proses klasifikasi sebelumnya
- 2) Penentuan root (akar) dari pohon keputusan, yang diperoleh dari atribut yang dipilih dengan didasarkan pada perhitungan nilai gain untuk setiap atribut. Atribut dengan nilai gain tertinggi akan dipilih sebagai root pohon keputusan. Sebelum melakukan perhitungan gain, terlebih dahulu perlu dihitung nilai entropy menggunakan rumus berikut.

$$Entropy(s) = \sum_{i=1}^n - p_i * \log_2 p_i$$

Keterangan:

S : jumlah kasus

A : atribut pada data

n : banyaknya jumlah partisi

pi : proporsi Si pada S

- Menentukan nilai gain dengan menggunakan metode information gain, yang mengukur seberapa efektif atribut dalam membagi data berdasarkan kategori yang ada:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^n \frac{S_i}{S} * Entropy(S_i)$$

Keterangan :

S : kumpulan kasus

A : atribut

n : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

- Mengulangi proses pemilihan root untuk setiap subpohon hingga seluruh kasus terpartisi dengan sempurna, sehingga setiap data masuk ke dalam kategori yang sesuai
- Proses partisi *Decision tree* akan berhenti ketika:
  - Semua kasus dalam node N mendapat kelas yang sama.
  - Tidak ada atribut di dalam kasus yang dipartisi lagi.
  - Tidak ada kasus di dalam cabang yang kosong

### 2.3 Cross Validation

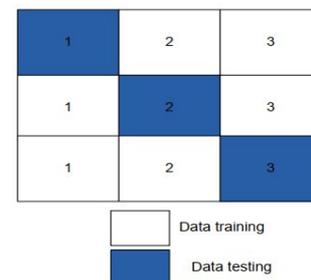
Menurut [9], Cross Validation, atau yang sering disebut sebagai estimasi rotasi, adalah teknik yang digunakan untuk memvalidasi model dengan tujuan mengevaluasi dan mengetahui sejauh mana hasil statistik dari analisis dapat digeneralisasi pada himpunan data yang independen. Teknik ini umumnya diterapkan untuk memprediksi klasifikasi model serta untuk memperkirakan akurasi model prediktif ketika diterapkan pada data baru dalam praktik. Salah satu varian dari teknik Cross Validation adalah K-Fold Cross Validation, di mana data dibagi menjadi 'k' bagian yang masing-masing memiliki ukuran yang sama. Penggunaan metode K-Fold Cross Validation bertujuan untuk mengurangi potensi bias dalam data dan meningkatkan ketepatan evaluasi model.

Proses pelatihan (training) dan pengujian (testing) dilakukan sebanyak jumlah k yang telah ditentukan. Pada tahap pertama, subset S1 digunakan sebagai data pengujian, sementara subset lainnya digunakan sebagai data pelatihan. Pada tahap kedua, subset S2 digunakan sebagai data

pengujian, dan subset lainnya, termasuk subset sebelumnya yang kini berfungsi sebagai data pelatihan, yaitu S1, digunakan untuk pelatihan. Proses ini diulang dengan mengganti subset yang digunakan sebagai data pengujian hingga semua subset telah digunakan. Dalam 3-Fold Cross Validation, setiap data akan dieksekusi sebanyak tiga kali, sesuai dengan nilai k yang ditentukan, sehingga setiap subset data memiliki kesempatan yang sama untuk berfungsi sebagai data pengujian maupun data pelatihan. Model pengujian ini dapat digambarkan dengan pembagian data yang dinamakan D1, D2, dan D3, seperti yang ditunjukkan dalam ilustrasi berikut.:

- Tahap ke-1 data yang ada pada blok D1 sebagai data uji sedangkan D2 dan D3 sebagai data latih
- Tahap ke-2 data D2 sebagai data uji sedangkan pada blok data D1 serta D3 digunakan sebagai data latih
- Tahap ke-3 data D3 sebagai data uji sedangkan D1 dan D2 sebagai data latih.

Berikut adalah ilustrasi penggunaan 3 K-Fold Cross Validation.



**Gambar 1.** Bentuk penggunaan K Fold Cross Validation

Sebagai pengukuran kinerja dari hasil classification adalah dengan cara komparasi seluruh data testing yang memiliki hasil prediksi klasifikasi dengan benar terhadap banyaknya data testing. Berikut adalah rumus yang digunakan yaitu:

$$Akurasi = \frac{\sum \text{klasifikasi benar}}{\sum \text{data uji}} \times 100\%$$

Selain kinerja, standar deviasi juga dihitung. Standar deviasi adalah ukuran yang menggambarkan sebaran data, yang menunjukkan jarak antara nilai rata-rata dengan titik data tertentu. Semakin besar nilai standar deviasi, semakin besar pula penyebaran data dari median, dan sebaliknya. Tujuan perhitungan standar deviasi dalam penelitian ini adalah untuk mengetahui sejauh mana perbedaan antara rata-rata akurasi dengan akurasi yang diperoleh pada setiap tahap percobaan yang dilakukan [10]. Rumus yang digunakan untuk

menghitung standar deviasi, yang merupakan salah satu ukuran statistik untuk menggambarkan sebaran atau variasi data, adalah sebagai berikut::

$$S = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

Keterangan:

n : jumlah tahap atau percobaan

μ : mean atau rata-rata

X : Percobaan ke-i

i : index setiap percobaan

## 2.4 Dataset

Dataset yang digunakan untuk penerapan algoritma data mining dalam penelitian ini yaitu data lulusan dari tahun 2019 sampai 2023 di STMIK Multicom. Dataset ini diperoleh dari admisi bagian data mahasiswa di STMIK Multicom. Alasan penggunaan dataset lulusan tahun 2019 sampai 2023 pada penelitian ini yaitu untuk mengetahui hasil atau label akhir yaitu dropout (ya) atau lulus (tidak). Dataset lulusan STMIK Multicom tahun 2019 sampai 2023 dapat dilihat pada Tabel 1 Dataset berikut ini:

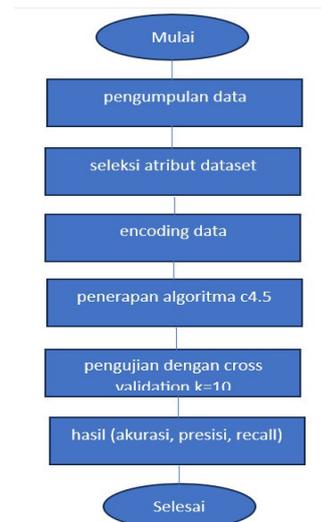
Tabel 1 Dataset

No	Nama	Progra m Studi	Jenis Kelamin	Status Pekerjaan
1	Mhs001	TI	L	Bekerja
2	Mhs002	SI	L	Bekerja
3	Mhs003	SI	P	Bekerja
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
19 8	Mhs198	SI	L	Bekerja
19 9	Mhs199	SI	P	Tidak
20 0	Mhs200	SI	P	Tidak

Data lulusan terdiri dari 200 record data serta 9 kolom yaitu nama, program studi, jenis kelamin, status pekerjaan, pendapatan orang tua, status tempat tinggal, jarak dari tempat tinggal ke kampus, lulusan sekolah (SMA/SMK), dan status (dropout/lulus).

## 2.5 Alur Pelaksanaan

Alur Pelaksanaan terdiri dari 6 tahapan yng dapat dilihat pada Gambar 2. Alur pelaksanaan berikut ini:

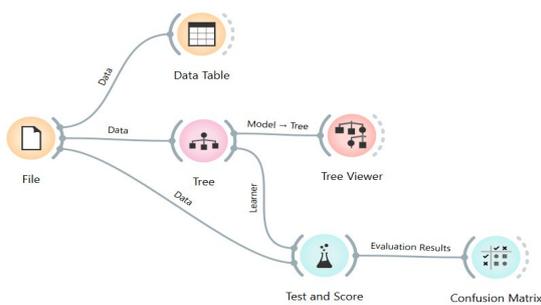


Gambar 2. Tahapan pelaksanaan

Tahapan pertama yaitu pengumpulan data dengan cara mengumpulkan data lulusan STMIK Multicom tahun 2019, 2020, 2021, 2022, dan 2023 dalam bentuk file excel. Selanjutnya kelima file tersebut digabung dan dijadikan 1 file yang berisi data lulusan dari tahun 2019 sampai 2023. Tahapan selanjutnya yaitu melakukan proses seleksi atribut dengan cara menghapus atribut yang tidak digunakan seperti nomor urut, nim, dan nama mahasiswa, kemudian mempertahankan atribut sisanya. Selanjutnya data dikategorikan sesuai urutan kepentingan dan diubah ke dalam bentuk angka. Tahapan selanjutnya melibatkan penerapan algoritma C4.5 untuk membangun pohon keputusan, diikuti dengan pengujian menggunakan metode K-Fold Cross Validation dengan jumlah k=10. Proses ini dilakukan dengan menggunakan bantuan tools Orange Data Mining, yang memudahkan dalam analisis data dan evaluasi model klasifikasi.. Setelah tahapan-tahapan tersebut maka didapatkan hasil berupa akurasi, presisi, dan recall.

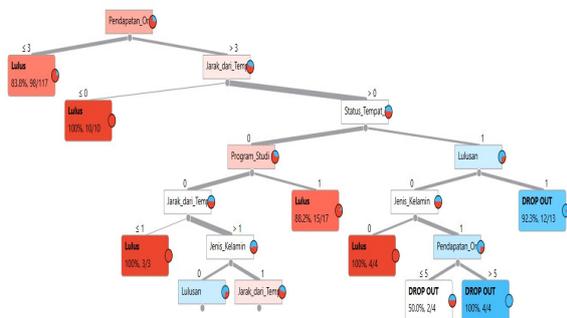
## 3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, pemodelan algoritma C4.5 dilakukan dengan memanfaatkan alat bantu Orange Data Mining. Model algoritma C4.5 yang digunakan dalam penelitian ini dapat dilihat pada Gambar 3, yang menunjukkan struktur pohon keputusan yang terbentuk berdasarkan data yang dianalisis.



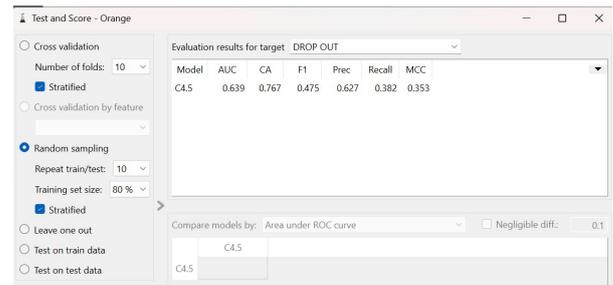
**Gambar 3.** Model C4.5

Berdasarkan Gambar 3. File digunakan untuk menyimpan dataset yang akan digunakan yaitu dataset lulusan STMIK Multicom 2019 sampai 2023. Data table digunakan untuk melihat dan memeriksa isi dari dataset yang ada pada file. Tree digunakan untuk menerapkan algoritma C4.5 serta Tree Viewer digunakan untuk melihat bentuk dari visualisasi pohon keputusan yang dihasilkan. Test and Score digunakan untuk melakukan training dan testing sekaligus melakukan evaluasi dengan menggunakan metode K Fold Cross Validation dengan K=10. Hasil dari penerapan model tersebut menghasilkan sebuah decision tree yang bisa dilihat di Gambar 4. Visualisasi pohon keputusan berikut ini menggambarkan bagaimana model membuat keputusan berdasarkan atribut-atribut yang ada dalam data.



**Gambar 4.** Visualisasi pohon keputusan

Berdasarkan Gambar 4, dapat diketahui bahwa atribut Pendapatan orang tua menjadi root node, diikuti jarak dari tempat tinggal ke kampus, status tempat tinggal, dan lulusan (SMA/SMK). Setelah penerapan metode C4.5 dan evaluasi yang digunakan yaitu k fold cross validation serta penggunaan jumlah yaitu K=10, diperoleh hasil yang bisa dilihat di dalam Gambar 5. Hasil evaluasi ini menunjukkan kinerja model dalam melakukan klasifikasi berdasarkan data yang digunakan dalam penelitian..



**Gambar 5.** Hasil evaluasi

Berdasarkan hasil evaluasi pada Gambar 5, evaluasi menggunakan bantuan alat Orange Data Mining dengan menerapkan metode cross validation dengan jumlah k=10 dapat diketahui tingkat kinerja algoritma C4.5 pada kasus prediksi mahasiswa berpotensi dropout pada dataset lulusan STMIK Multicom 2019 sampai 2023 yaitu tingkat akurasi sebesar 0,639 atau sebesar 63%, recall sebesar 0,382 atau 38%, presisi 0,627 atau 62%.

#### 4. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa algoritma C4.5 berhasil menghasilkan tingkat akurasi sebesar 63% dan presisi sebesar 62%, yang menunjukkan bahwa model ini memiliki kemampuan prediksi yang cukup baik dalam memproyeksikan potensi dropout pada mahasiswa. Meskipun demikian, mengingat kompleksitas dan pentingnya presisi dalam konteks prediksi akademik, diperlukan penelitian dan pengembangan lebih lanjut untuk meningkatkan kinerja algoritma, khususnya dalam hal akurasi dan efektivitasnya dalam menggunakan dataset primer yang relevan. Upaya optimasi dan penyesuaian terhadap karakteristik data akan menjadi langkah penting dalam meningkatkan performa model ini.

#### 5. SARAN

Penelitian ini tentu memiliki keterbatasan, oleh karena itu, terdapat beberapa saran untuk penelitian selanjutnya, sebagai berikut:

- 1) Menggunakan algoritma data mining yang lain atau membandingkan algoritma yang digunakan dengan algoritma lain
- 2) Menggunakan jumlah k yang lebih variatif pada cross validation untuk mendapatkan tingkat kinerja yang lebih baik
- 3) Menambah jumlah data serta menggunakan pembagian proporsi data dengan jumlah rasio 9:1 atau 7:3.

## DAFTAR PUSTAKA

- [1] I. Fadjrian, S. Budi, and H. W. Aji, "Identifikasi Faktor-Faktor Yang berhubungan Dengan Mahasiswa Putus Kuliah di IPB Angkatan 2008 Menggunakan Analisis Survival," *Xplore*, vol. 2, no. 1, 2013.
- [2] Kusrini and L. Taufiq Emha, *Algoritma Data Mining Yogyakarta*, no. February. 2009. [Online]. Available: <https://books.google.co.id/books?id=-Ojclag73O8C&printsec=frontcover#v=onepage&q&f=false>
- [3] V. Anestiviya, A. Ferico, and O. Pasaribu, "ANALISIS POLA MENGGUNAKAN METODE C4.5 UNTUK PEMINATAN JURUSAN SISWA BERDASARKAN KURIKULUM (STUDI KASUS : SMAN 1 NATAR)," *Jurnal Teknologi dan Sistem Informasi (JTSI)*, vol. 2, no. 1, pp. 80–85, 2021, [Online]. Available: <http://jim.teknokrat.ac.id/index.php/JTSI>
- [4] A. Azahari and N. Nursobah, "Rekomendasi Penerimaan Beasiswa Yayasan Untuk Siswa Baru SMK TI Airlangga dengan Algoritma C4.5," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 609, Apr. 2021, doi: 10.30865/mib.v5i2.2943.
- [5] A. Dina, I. Permana, F. Muttakin, and I. Maita, "Perbandingan Algoritma NBC, KNN, dan C4.5 Untuk Klasifikasi Penerima Bantuan Program Keluarga Harapan," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 3, p. 1079, Jul. 2023, doi: 10.30865/mib.v7i3.6316.
- [6] D. Librado and A. H. Nasyuha, "Penerapan Data Mining Untuk Klasifikasi Penerima Kredit Dengan Perbandingan Algoritma Naïve Bayes dan Algoritma C4.5," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 4, p. 1952, Oct. 2023, doi: 10.30865/mib.v7i4.6907.
- [7] Dr. Suyanto, *Data Mining Untuk Klasifikasi Dat*, no. x. Bandung: Informatika, 2019.
- [8] Y. D. Atma and A. Setyanto, "Perbandingan algoritma c4.5 dan k-nn dalam identifikasi mahasiswa berpotensi drop out," *Metik Jurnal ISSN : 2580-1503*, vol. 2, no. 2, pp. 31–37, 2018.
- [9] M. Bramer, *Principles of Data Mining*. 2007. doi: 10.1007/978-1-84628-766-4.
- [10] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara Knn Dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation," vol. 5, no. 5, pp. 577–584, 2018, doi: 10.25126/jtiik20185983.