Polygon : Jurnal Ilmu Komputer dan Ilmu Pengetahuan Alam
Vol. 1 No. 5 September 2023
e-ISSN: 3046-5419; p-ISSN: 3032-6249, Hal 01-17
DOI : https://doi.org/10.62383/polygon.v1i5.855
Tersedia: https://journal.arimsi.or.id/index.php/Polygon

Penerapan Algoritma C4.5 untuk Prediksi Kelulusan Mahasiswa berdasarkan Data Akademik

Nuari Anisa Sivi[1*], Rudi Hartono[2], Putra Hanafi[3]
[1-3]Sistem Informasi, Universitas Nahdlatul Ulama Lampung, Indonesia
**Penulis Korespondensi: nuarianisasivi@gmail.com[1*], rudi.hartono1512@gmail.com[2], putrahanafi695@gmail.com[3]*

Abstract*. Data mining is a technology that plays an important role in supporting data-driven decision making, especially in complex and dynamic higher education environments. In the context of education management, the ability to predict student graduation is an essential aspect because it can help institutions plan strategic steps, intervene earlier, and optimize academic resources. This study aims to apply the C4.5 decision tree algorithm to build a student graduation prediction model based on academic data. The research dataset includes key variables such as Grade Point Average (GPA), total Semester Credit Units (SKS) taken, and student attendance rates during lectures. The analysis was conducted using the C4.5 algorithm, which is known for its high level of interpretability, making the model results easy to understand by policy makers. The test results showed an accuracy of 84.6%, indicating that this method has the potential to support data-based academic management systems. These findings are expected to serve as a basis for educational institutions to improve the effectiveness of monitoring and evaluating the student learning process.*

Keywords*: C4.5 Algorithm, Graduation Prediction, Academic Data, Data Mining, Classification*

Abstrak. Data mining merupakan teknologi yang memiliki peran penting dalam mendukung pengambilan keputusan berbasis data, terutama di lingkungan perguruan tinggi yang kompleks dan dinamis. Dalam konteks manajemen pendidikan, kemampuan memprediksi kelulusan mahasiswa menjadi aspek yang esensial karena dapat membantu institusi merencanakan langkah strategis, melakukan intervensi lebih awal, serta mengoptimalkan sumber daya akademik. Penelitian ini bertujuan menerapkan algoritma decision tree C4.5 untuk membangun model prediksi kelulusan mahasiswa berdasarkan data akademik. Dataset penelitian mencakup variabel utama seperti Indeks Prestasi Kumulatif (IPK), total Satuan Kredit Semester (SKS) yang telah diambil, dan tingkat kehadiran mahasiswa selama perkuliahan. Analisis dilakukan menggunakan algoritma C4.5 yang dikenal memiliki tingkat interpretasi yang baik, sehingga hasil model mudah dipahami oleh pihak pengambil kebijakan. Dari hasil pengujian diperoleh akurasi sebesar 84,6%, yang menunjukkan bahwa metode ini berpotensi mendukung sistem manajemen akademik berbasis data. Temuan ini diharapkan dapat menjadi landasan bagi institusi pendidikan dalam meningkatkan efektivitas pemantauan dan evaluasi proses belajar mahasiswa.

Kata kunci: Algoritma C4.5, Prediksi Kelulusan, Data Akademik, Data Mining, Klasifikasi

1. LATAR BELAKANG

Perguruan tinggi memiliki peran strategis dalam menghasilkan sumber daya manusia yang berkualitas dan mampu bersaing di tingkat global (Supriyanto & Hidayat, 2019). Seiring meningkatnya tuntutan dunia kerja, institusi pendidikan tinggi tidak hanya dituntut menyediakan proses pembelajaran yang berkualitas, tetapi juga memastikan mahasiswa mampu menyelesaikan studi tepat waktu (Azizah, 2020). Pemantauan performa akademik mahasiswa menjadi bagian penting dalam proses tersebut. Namun pada praktiknya, evaluasi akademik masih banyak dilakukan secara manual sehingga respons terhadap permasalahan akademik sering terlambat (Hasibuan, 2018).

Naskah Masuk: Maret 15,2022; Revisi: Juni 31, 2023; Diterima: Juli 27, 2023; Tersedia: September 30, 2023

Kemajuan teknologi informasi memberikan peluang besar untuk meningkatkan efektivitas evaluasi akademik, salah satunya melalui penerapan data mining *(Baker & Yacef, 2009)*. Teknik ini memungkinkan institusi mengidentifikasi pola tersembunyi dalam data akademik mahasiswa, seperti data nilai, jumlah SKS, dan kehadiran, yang dapat dimanfaatkan untuk memprediksi kelulusan atau risiko keterlambatan studi (Romero & Ventura, 2010). Algoritma klasifikasi, khususnya C4.5, menjadi salah satu metode yang sering digunakan karena dapat menghasilkan model yang mudah dipahami oleh pengguna non-teknis (Quinlan, 1993; Witten, Frank, & Hall, 2017).

Penelitian ini memanfaatkan algoritma C4.5 untuk membuat model prediksi kelulusan mahasiswa berdasarkan IPK, SKS yang telah ditempuh, dan kehadiran (Hernawati & Purwanto, 2021; Siregar, 2020). Ketiga variabel ini dipilih karena dianggap mewakili capaian akademik, progres perkuliahan, dan partisipasi mahasiswa selama proses pendidikan. Dengan dataset yang diperoleh dari institusi pendidikan, penelitian ini diharapkan mampu menghasilkan model prediktif yang tidak hanya akurat, tetapi juga dapat digunakan sebagai alat bantu dalam pengambilan keputusan akademik *(Kumar & Pal, 2011)*. Pengembangan model prediksi ini sekaligus mendorong penerapan manajemen akademik berbasis data yang lebih adaptif dan efektif (Delen, 2005).

2. KAJIAN TEORITIS

Pengertian Data Mining

Data mining merupakan salah satu cabang penting dalam bidang ilmu komputer yang berfokus pada proses penggalian informasi berharga dari kumpulan data dalam jumlah besar. Secara umum, data mining didefinisikan sebagai serangkaian prosedur sistematis untuk mengidentifikasi pola, tren, hubungan, atau struktur tersembunyi yang tidak dapat diamati secara langsung melalui pengolahan data konvensional. Menurut Han dan Kamber (2011), data mining tidak hanya berkaitan dengan kegiatan pencarian data, tetapi lebih jauh merupakan proses transformasi data mentah menjadi pengetahuan baru yang memiliki nilai strategis dalam mendukung pengambilan keputusan. Data yang digunakan dalam data mining biasanya tersimpan di dalam berbagai sistem seperti basis data (database), data warehouse, atau sistem informasi terintegrasi lainnya. Dengan perkembangan teknologi informasi yang semakin pesat, data mining telah berkembang menjadi alat analisis yang mampu memproses data secara cepat, akurat, dan efisien, bahkan ketika data memiliki volume yang sangat besar dan bersifat kompleks.

Dalam konteks dunia pendidikan, data mining banyak diterapkan dalam bidang yang dikenal sebagai Educational Data Mining (EDM). EDM merupakan subdisiplin ilmu yang secara khusus memanfaatkan teknik data mining untuk menganalisis data akademik mahasiswa, seperti nilai, tingkat kehadiran, interaksi selama pembelajaran, aktivitas pada platform e-learning, dan berbagai parameter pendidikan lainnya. Melalui EDM, institusi pendidikan dapat memperoleh pemahaman yang lebih mendalam mengenai faktor-faktor yang mempengaruhi keberhasilan akademik mahasiswa, termasuk pola belajar, kebiasaan menghadiri kelas, tingkat keterlibatan dalam kegiatan akademik, hingga variabel-variabel sosial dan perilaku yang mungkin berperan dalam menentukan performa mahasiswa. Hasil analisis ini dapat digunakan sebagai dasar untuk memperbaiki strategi pembelajaran, meningkatkan kualitas pengajaran, serta mengembangkan kebijakan akademik yang lebih tepat sasaran dan berbasis data. *(Romero & Ventura, 2010)*.

Selain itu, data mining dalam dunia pendidikan juga memberikan kontribusi penting dalam mengidentifikasi risiko akademik secara lebih awal. Misalnya, melalui analisis data secara otomatis, sistem dapat mendeteksi mahasiswa yang berpotensi mengalami kegagalan studi, keterlambatan kelulusan, atau penurunan performa akademik. Dengan demikian, lembaga pendidikan dapat mengambil langkah intervensi lebih cepat, seperti memberikan bimbingan akademik, konseling, atau strategi pendukung lainnya. Secara keseluruhan, penggunaan data mining dalam pendidikan tidak hanya membantu dalam memahami dinamika akademik mahasiswa, tetapi juga mendorong terciptanya proses pendidikan yang lebih adaptif, efektif, dan responsif terhadap kebutuhan individu mahasiswa. Pemanfaatan data yang sebelumnya hanya menjadi arsip pasif kini dapat diolah menjadi pengetahuan yang memberikan dampak langsung pada peningkatan mutu pendidikan. Dengan potensi yang sangat besar tersebut, data mining diakui sebagai salah satu teknologi kunci dalam menciptakan ekosistem pendidikan berbasis data (data-driven education) yang lebih modern dan berorientasi pada peningkatan kualitas layanan akademik. *(Delen, 2005)*.

Algoritma C4.5

Algoritma C4.5 merupakan salah satu algoritma klasifikasi yang berbasis pada pembentukan pohon keputusan (decision tree), dan dikembangkan oleh J. Ross Quinlan sebagai penyempurnaan dari algoritma sebelumnya, yaitu ID3. Algoritma ini bekerja dengan cara membangun struktur pohon yang terdiri dari simpul-simpul keputusan, di mana setiap simpul merepresentasikan suatu atribut, sedangkan cabang-cabangnya menggambarkan nilai dari atribut tersebut. Tujuan utama dari algoritma C4.5 adalah membentuk model klasifikasi yang mampu mengelompokkan data berdasarkan pola tertentu yang terkandung di dalam

dataset. Berbeda dengan algoritma ID3 yang menggunakan *information gain* sebagai pemilih atribut, C4.5 menggunakan *gain ratio* untuk menghindari bias terhadap atribut yang memiliki banyak kategori. Penggunaan *gain ratio* menjadikan pemilihan atribut pada C4.5 lebih seimbang dan akurat dalam membangun struktur pohon yang optimal. *(Quinlan, 1993).*

Salah satu keunggulan signifikan dari algoritma C4.5 adalah kemampuannya menangani jenis data yang lebih kompleks, seperti atribut dengan nilai kontinu. Pada prosesnya, nilai kontinu tersebut akan dikelompokkan menjadi interval tertentu sehingga tetap dapat digunakan sebagai pemisah dalam pohon keputusan. C4.5 juga mampu menangani data yang memiliki nilai kosong (missing values), di mana algoritma akan tetap melakukan perhitungan dengan menyesuaikan bobot data yang hilang tanpa harus menghapus seluruh data. Selain itu, algoritma ini dapat melakukan proses *pruning*, yaitu pemangkasan cabang pohon yang dianggap kurang relevan atau berpotensi menyebabkan overfitting. Teknik pruning membantu menghasilkan model pohon keputusan yang lebih sederhana, mudah dipahami, dan tetap memiliki tingkat akurasi yang baik. Karena sifatnya yang interpretatif dan mampu menyajikan aturan (rules) yang jelas dari setiap cabang pohon, C4.5 sering digunakan pada penelitian yang membutuhkan transparansi dalam proses pengambilan keputusan. *(Kotsiantis, 2013).*

Dalam bidang pendidikan, penggunaan algoritma C4.5 semakin populer karena metode ini mampu mengolah data akademik dalam jumlah besar dan menghasilkan model prediksi yang mudah dipahami oleh pihak non-teknis seperti dosen, tenaga kependidikan, maupun manajemen kampus. C4.5 telah banyak digunakan dalam berbagai penelitian untuk memprediksi kelulusan mahasiswa, mengidentifikasi risiko dropout, serta menganalisis performa akademik berdasarkan atribut-atribut seperti IPK, jumlah SKS, dan tingkat kehadiran. Kelebihan interpretasi visual dari pohon keputusan membuat pihak akademik dapat dengan mudah melihat faktor mana yang paling menentukan keberhasilan atau kegagalan seorang mahasiswa. Dengan demikian, algoritma C4.5 tidak hanya berfungsi sebagai alat klasifikasi, tetapi juga berperan sebagai alat analisis yang memberikan informasi berharga untuk mendukung perencanaan strategi pembelajaran, intervensi akademik, dan pengambilan kebijakan berbasis data. Secara keseluruhan, fleksibilitas, interpretabilitas, serta keakuratan yang ditawarkan oleh C4.5 menjadikannya salah satu algoritma paling relevan dan efektif dalam pemanfaatan data mining pada dunia pendidikan. *(Witten, Frank, & Hall, 2017).*

Penelitian Terkait

Penelitian terkait mengenai penerapan algoritma C4.5 dalam dunia pendidikan telah memberikan kontribusi besar dalam menunjukkan efektivitas algoritma ini sebagai alat

prediksi berbasis data. Salah satu penelitian yang cukup dikenal adalah penelitian yang dilakukan oleh Kumar dan Pal (2011), yang mengevaluasi kemampuan algoritma C4.5 dalam memprediksi performa akademik siswa berdasarkan berbagai parameter seperti nilai harian, hasil ujian, serta tingkat kehadiran. Hasil penelitian tersebut menunjukkan bahwa algoritma C4.5 memiliki tingkat akurasi yang tinggi dan mampu menghasilkan model klasifikasi yang stabil. Selain itu, struktur pohon keputusan yang dihasilkan dapat memberikan gambaran yang jelas mengenai faktor-faktor yang memiliki pengaruh paling besar terhadap performa siswa, sehingga algoritma ini dinilai cocok untuk diterapkan pada konteks akademik yang membutuhkan analisis berbasis data secara komprehensif.

Penelitian berikutnya dilakukan oleh Suryani (2020), yang memanfaatkan algoritma C4.5 untuk melakukan klasifikasi status mahasiswa aktif dan tidak aktif pada sebuah institusi pendidikan. Dalam penelitian tersebut, C4.5 digunakan untuk menganalisis data akademik seperti IPK, riwayat pengambilan SKS, dan kehadiran mahasiswa. Hasil penelitian tersebut memperlihatkan bahwa algoritma C4.5 mampu mengolah data pendidikan dengan jumlah besar dan menghasilkan model yang mudah dipahami oleh pihak akademik. Model yang dihasilkan tidak hanya berfungsi sebagai alat klasifikasi, tetapi juga membantu institusi dalam memetakan mahasiswa yang membutuhkan perhatian atau bimbingan lebih intensif. Penelitian Suryani ini memperkuat pandangan bahwa C4.5 memiliki kemampuan kuat dalam mendukung proses evaluasi akademik secara sistematik dan terukur.

Selain kedua penelitian tersebut, penelitian lain yang sering dijadikan acuan adalah penelitian yang dilakukan oleh Delen (2005), yang menerapkan berbagai teknik data mining, termasuk pohon keputusan, untuk memprediksi risiko dropout mahasiswa di tingkat perguruan tinggi. Delen menemukan bahwa metode decision tree, termasuk algoritma C4.5, dapat mengidentifikasi pola-pola tertentu yang tidak terlihat dalam proses evaluasi manual. Atribut seperti nilai mata kuliah dasar, frekuensi kehadiran, dan aktivitas akademik lainnya memiliki kontribusi signifikan dalam menentukan apakah seorang mahasiswa berpotensi mengalami kegagalan atau keluar dari program studi. Model yang dihasilkan Delen tidak hanya digunakan untuk analisis prediksi, tetapi juga berfungsi sebagai sistem peringatan dini (early warning system) yang membantu institusi pendidikan dalam memberikan intervensi akademik. Temuan ini memperlihatkan bahwa algoritma C4.5 bukan hanya efektif dalam klasifikasi, tetapi juga relevan dalam pengembangan kebijakan strategis berbasis data.

Secara keseluruhan, penelitian-penelitian terdahulu tersebut memperlihatkan bahwa algoritma C4.5 memiliki performa yang konsisten dan sangat layak digunakan dalam konteks analisis akademik. Keberhasilan algoritma ini dalam berbagai penelitian sebelumnya menjadi

landasan kuat yang mendukung pemilihannya dalam penelitian ini. Melalui penerapan C4.5, penelitian ini berupaya memperluas pemahaman mengenai pola-pola akademik mahasiswa, serta menghadirkan model prediksi yang akurat, interpretatif, dan dapat digunakan oleh institusi pendidikan sebagai alat bantu dalam meningkatkan kualitas pengelolaan akademik. Temuan-temuan dari penelitian sebelumnya menjadi acuan penting yang membuktikan bahwa C4.5 merupakan algoritma yang relevan, efektif, dan memiliki potensi besar untuk dikembangkan lebih lanjut dalam berbagai bidang analisis data pendidikan.

Kerangka Berfikir

Teori

Kerangka berpikir dalam penelitian ini menggambarkan alur logis bagaimana data akademik mahasiswa dapat dimanfaatkan untuk membangun model prediksi kelulusan menggunakan algoritma C4.5. Penelitian ini berangkat dari permasalahan utama bahwa institusi pendidikan sering mengalami kesulitan dalam memantau perkembangan mahasiswa secara efektif, terutama dalam mengidentifikasi mahasiswa yang berpotensi terlambat lulus. Selama ini, proses evaluasi akademik lebih banyak dilakukan secara manual dan bersifat reaktif sehingga intervensi akademik sering kali diberikan terlambat. Oleh karena itu, diperlukan suatu sistem prediksi otomatis yang mampu memberikan gambaran lebih awal mengenai status kelulusan mahasiswa berdasarkan data akademik yang telah tersedia.

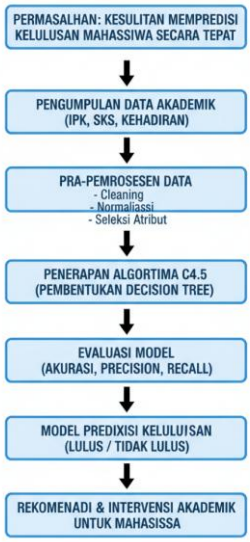
Dalam konteks tersebut, data akademik seperti Indeks Prestasi Kumulatif (IPK), jumlah Satuan Kredit Semester (SKS) yang telah diselesaikan, serta tingkat kehadiran mahasiswa menjadi variabel penting yang dapat dianalisis. Ketiga variabel ini mencerminkan performa akademik mahasiswa dari berbagai aspek, yaitu pencapaian nilai, progres pengambilan mata kuliah, dan keterlibatan dalam proses pembelajaran. Data tersebut kemudian dikumpulkan dari Sistem Informasi Akademik Universitas Nahdlatul Ulama Lampung. Namun sebelum digunakan untuk pemodelan, data harus melalui tahapan pra-pemrosesan agar kualitasnya terjamin dan bebas dari kesalahan, seperti missing values, data rangkap, atau ketidakkonsistenan nilai.

Setelah data dibersihkan, tahap berikutnya adalah menerapkan algoritma C4.5 menggunakan perangkat lunak WEKA. Algoritma C4.5 bekerja dengan memilih atribut yang paling berpengaruh berdasarkan gain ratio dan membentuk model klasifikasi berbentuk pohon keputusan. Pohon keputusan ini menggambarkan jalur pengambilan keputusan dari setiap atribut hingga menghasilkan prediksi akhir berupa status kelulusan mahasiswa. Dengan demikian, model yang dibangun tidak hanya memberikan hasil prediksi, tetapi juga menyajikan penjelasan visual mengenai faktor-faktor yang paling memengaruhi kelulusan. Hal ini menjadi

kelebihan utama C4.5 karena hasilnya mudah dipahami, bahkan oleh pihak akademik non-teknis.

Model yang telah dihasilkan kemudian dievaluasi menggunakan confusion matrix yang mencakup nilai akurasi, precision, dan recall. Evaluasi ini penting untuk memastikan bahwa model tersebut bekerja dengan baik dan mampu memberikan prediksi yang konsisten apabila diterapkan pada data nyata. Setelah model diuji dan diketahui performanya, hasil prediksi dapat dimanfaatkan oleh institusi sebagai dasar dalam memberikan intervensi dini kepada mahasiswa yang terdeteksi berisiko tidak lulus tepat waktu. Dengan demikian, penelitian ini diharapkan mampu memberikan kontribusi nyata dalam meningkatkan efektivitas manajemen akademik berbasis data. Secara keseluruhan, hubungan antarvariabel, proses pengolahan data, serta pengujian model membentuk kerangka berpikir yang sistematis dan terstruktur dalam menghasilkan model prediksi kelulusan mahasiswa.

Diagram Alur



Gambar 1. Diagram Alur

3. METODE PENELITIAN

Waktu dan Tempat Penelitian

Penelitian ini dilakukan pada Bulan September – Februari 2024 di Universitas Universitas Nahdlatul Ulama Lampung

Jenis Penelitian

Jenis penelitian yang digunakan dalam studi ini adalah penelitian kuantitatif dengan pendekatan deskriptif dan eksploratif berbasis data mining. Pendekatan kuantitatif dipilih karena penelitian ini berfokus pada pengolahan data numerik mahasiswa, seperti IPK, jumlah SKS, dan tingkat kehadiran, yang kemudian dianalisis menggunakan algoritma klasifikasi C4.5

untuk memperoleh model prediksi kelulusan. Pendekatan deskriptif digunakan untuk menggambarkan karakteristik data akademik mahasiswa serta memaparkan pola-pola yang muncul berdasarkan hasil analisis. Sementara itu, pendekatan eksploratif dilakukan untuk menyelidiki sejauh mana variabel-variabel akademik tersebut dapat memengaruhi hasil prediksi kelulusan mahasiswa melalui teknik klasifikasi dalam data mining.

Penelitian ini juga dikategorikan sebagai penelitian terapan (applied research) karena bertujuan menghasilkan model prediksi yang dapat digunakan secara langsung dalam lingkungan akademik sebagai alat bantu pengambilan keputusan. Dengan memanfaatkan algoritma C4.5, penelitian ini tidak hanya menganalisis data historis, tetapi juga mengembangkan model yang mampu mengidentifikasi pola kelulusan mahasiswa secara lebih akurat. Oleh karena itu, jenis penelitian yang digunakan sangat relevan untuk menghasilkan solusi berbasis teknologi yang dapat meningkatkan efektivitas manajemen akademik.

Instrumen Penelitian

Instrumen yang digunakan dalam penelitian sebagai berikut :

a. Dataset Akademik Mahasiswa

   Dataset berupa data historis mahasiswa Universitas Nahdlatul Ulama Lampung yang mencakup variabel Indeks Prestasi Kumulatif (IPK), jumlah Satuan Kredit Semester (SKS) yang telah diselesaikan, tingkat kehadiran perkuliahan, serta status kelulusan. Dataset ini menjadi bahan utama untuk proses analisis dan pemodelan.

b. Sistem Informasi Akademik (SIAKAD)

   Sistem informasi internal kampus yang digunakan sebagai sumber pengambilan data akademik mahasiswa. SIAKAD menjadi instrumen penting dalam memperoleh data yang valid dan ter-update.

c. Perangkat Komputer / Laptop

   Digunakan untuk menjalankan software analisis, melakukan pengolahan data, membangun model klasifikasi, serta menyusun laporan penelitian. Spesifikasi perangkat disesuaikan dengan kebutuhan pemrosesan data.

d. Perangkat Lunak WEKA (Waikato Environment for Knowledge Analysis)

   Tools utama dalam penerapan algoritma C4.5, yang digunakan untuk proses klasifikasi, pembentukan decision tree, evaluasi model, serta visualisasi hasil analisis.

e. Software Pengolah Data (Excel/Google Sheets)

   Digunakan untuk cleaning data, normalisasi, pengecekan missing values, serta menyiapkan dataset sebelum dimasukkan ke WEKA.

f. Bahasa Pemrograman Opsional (Python/R) – jika diperlukan
   Digunakan untuk eksplorasi data tambahan, visualisasi, atau evaluasi model lanjutan. Tidak wajib tetapi dapat mendukung pemrosesan data jika diperlukan.

Sistem informasi persediaan barang akan menggunakan internetdengan mendaftarkan hosting dan domain dari Rumah WEB

Teknik Pengumpulan Data

Untuk memperoleh suatu data yang diperlukan dalam penelitian ini, maka penulis menggunakan beberapa metode sebagai berikut :

1. Dokumentasi Data Akademik
   a. Mengumpulkan data historis mahasiswa yang telah tersimpan dalam Sistem Informasi Akademik (SIAKAD) Universitas Nahdlatul Ulama Lampung.
   b. Data yang diperoleh meliputi IPK, jumlah SKS yang telah ditempuh, tingkat kehadiran, serta status kelulusan mahasiswa.
   c. Teknik ini digunakan karena data akademik sudah terdokumentasi secara digital dan dapat diakses sebagai bahan penelitian.
2. Wawancara (Interview)
   a. Melakukan komunikasi dengan pihak akademik atau administrasi fakultas untuk memperoleh konfirmasi mengenai keaslian data, penjelasan atribut, atau kebijakan akademik yang berkaitan dengan kelulusan.
   b. Wawancara dilakukan secara informal untuk mendapatkan informasi pendukung mengenai proses penilaian dan ketentuan akademik kampus.
3. Studi Pustaka
   a. Mengumpulkan referensi dari jurnal, buku, skripsi, dan penelitian terdahulu yang relevan dengan data mining, algoritma C4.5, serta prediksi kelulusan mahasiswa.
   b. Studi literatur digunakan untuk memperkuat landasan teori serta membandingkan hasil penelitian sebelumnya dengan penelitian ini.

Pra-Pemrosesan Data

Pra-pemrosesan data merupakan salah satu tahapan paling penting dalam proses data mining karena kualitas data yang digunakan sangat menentukan akurasi dan keandalan model prediksi yang dihasilkan. Tahap ini dilakukan untuk memastikan bahwa dataset berada dalam kondisi bersih, lengkap, dan sesuai untuk diolah menggunakan algoritma C4.5. Proses pra-pemrosesan melibatkan beberapa langkah, dimulai dari pemeriksaan kelengkapan data, penanganan data yang hilang, normalisasi, hingga seleksi atribut.

Langkah pertama dalam pra-pemrosesan adalah mengidentifikasi dan menangani data yang hilang (missing values). Pada dataset akademik, data yang hilang dapat berupa nilai IPK yang belum tersedia, SKS yang belum diperbarui, atau kehadiran yang tidak tercatat. Jika data yang hilang jumlahnya kecil, maka data tersebut dapat dihapus untuk menjaga konsistensi. Namun jika jumlahnya cukup besar dan berpotensi menghilangkan banyak informasi penting, maka dilakukan metode pengisian seperti imputasi rata-rata (mean imputation) atau teknik estimasi lainnya agar dataset tetap representatif. Penanganan ini penting karena algoritma C4.5 dapat menerima data yang hilang, tetapi kualitas model akan lebih baik apabila data sebelumnya dibersihkan.

Langkah selanjutnya adalah proses normalisasi data, terutama pada atribut yang memiliki rentang nilai yang sangat berbeda seperti IPK, kehadiran, dan SKS. Normalisasi membantu menyamakan skala nilai sehingga atribut dengan skala lebih besar tidak mendominasi proses perhitungan. Misalnya, IPK memiliki rentang nilai 0–4, sedangkan SKS dapat memiliki rentang mulai dari 0–144 SKS. Tanpa normalisasi, atribut SKS berpotensi memberikan pengaruh lebih besar pada pemilihan atribut dalam decision tree. Normalisasi dilakukan agar setiap atribut memiliki kontribusi yang proporsional dalam proses klasifikasi.

Selain itu, dilakukan pula pembersihan data (data cleaning) untuk memastikan tidak ada data duplikat, nilai ekstrem (outlier) yang tidak logis, atau inkonsistensi dalam data. Contohnya, data mahasiswa dengan kehadiran lebih dari 100% atau nilai IPK di luar rentang 0–4 harus diperiksa dan diperbaiki. Proses ini dilakukan agar model tidak menghasilkan pola yang keliru akibat adanya kesalahan pencatatan atau nilai yang tidak valid dalam dataset.

Tahap terakhir adalah seleksi atribut (attribute selection). Seleksi ini dilakukan untuk menentukan variabel mana saja yang relevan dalam membangun model prediksi menggunakan algoritma C4.5. Pada penelitian ini, atribut yang digunakan adalah IPK, jumlah SKS yang telah diselesaikan, dan tingkat kehadiran. Atribut lain yang tidak berpengaruh secara signifikan terhadap prediksi kelulusan dihilangkan agar model lebih sederhana, efisien, dan memiliki tingkat interpretasi yang lebih baik. Seleksi atribut juga membantu mengurangi risiko overfitting dan meningkatkan performa algoritma.

Secara keseluruhan, pra-pemrosesan data merupakan proses krusial dalam penelitian ini karena memastikan bahwa data yang digunakan memiliki kualitas yang baik dan siap diolah untuk membentuk model prediksi yang akurat. Tanpa tahap ini, hasil analisis dapat menjadi bias, tidak akurat, atau bahkan menyesatkan, sehingga berdampak langsung pada efektivitas model prediksi kelulusan mahasiswa.

Implementasi Algoritma

Implementasi algoritma C4.5 pada penelitian ini dilakukan dengan menggunakan perangkat lunak WEKA (Waikato Environment for Knowledge Analysis), sebuah aplikasi yang banyak digunakan dalam bidang data mining dan machine learning. WEKA dipilih karena menyediakan berbagai algoritma klasifikasi, termasuk C4.5 yang diimplementasikan dalam modul J48, serta memiliki antarmuka yang mudah digunakan untuk analisis data terstruktur. Proses implementasi dimulai dengan mempersiapkan dataset hasil pra-pemrosesan dalam format yang kompatibel, yaitu ARFF atau CSV. Dataset kemudian dimasukkan ke dalam WEKA melalui menu Explorer, di mana setiap atribut dikonfirmasi kembali untuk memastikan bahwa tipe datanya telah sesuai, misalnya atribut numerik untuk IPK, SKS, dan kehadiran, serta tipe nominal untuk status kelulusan.

Setelah dataset terunggah dengan benar, langkah berikutnya adalah memilih algoritma J48 sebagai representasi dari algoritma C4.5. WEKA kemudian akan menjalankan proses perhitungan *gain ratio* untuk menentukan atribut mana yang memiliki informasi paling tinggi dalam memisahkan data ke dalam kelas tertentu, dalam hal ini antara mahasiswa yang lulus dan tidak lulus. Pemilihan atribut dengan gain ratio tertinggi dilakukan secara otomatis oleh sistem, dan nilai tersebut digunakan sebagai simpul akar (root) dalam pohon keputusan. Selanjutnya, algoritma membagi dataset ke dalam cabang-cabang berdasarkan nilai atribut tersebut, kemudian melanjutkan proses yang sama pada setiap subset data hingga terbentuk pohon keputusan lengkap. Setiap simpul daun (leaf node) yang terbentuk merupakan hasil klasifikasi akhir yang menunjukkan prediksi status kelulusan mahasiswa.

WEKA juga menyediakan opsi untuk melakukan *pruning* atau pemangkasan pohon. Pemangkasan ini berfungsi menghilangkan cabang-cabang pohon yang dianggap tidak signifikan dan berpotensi menyebabkan overfitting. Dengan melakukan pruning, model pohon keputusan yang terbentuk menjadi lebih sederhana, lebih mudah dipahami, dan memiliki kemampuan generalisasi yang lebih baik ketika digunakan pada data baru yang tidak termasuk dalam dataset pelatihan. Pada tahap ini, peneliti dapat mengatur parameter seperti confidence factor, minimum number of instances per leaf, dan lainnya untuk mengoptimalkan performa model. Pemilihan parameter yang tepat sangat menentukan kualitas pohon keputusan yang dihasilkan.

Setelah model selesai dibangun, langkah berikutnya adalah melakukan evaluasi model menggunakan data uji atau melalui metode cross-validation yang disediakan oleh WEKA. Evaluasi yang dilakukan meliputi penghitungan akurasi, precision, recall, serta analisis confusion matrix. WEKA secara otomatis menyediakan laporan lengkap mengenai performa

model, termasuk struktur pohon keputusan yang dapat divisualisasikan secara grafik maupun dalam bentuk aturan (rules) yang lebih mudah dipahami. Visualisasi ini sangat membantu dalam menginterpretasikan pola-pola keputusan yang ditemukan oleh algoritma C4.5, misalnya bagaimana IPK atau tingkat kehadiran memengaruhi prediksi kelulusan.

Secara keseluruhan, implementasi algoritma C4.5 dengan menggunakan WEKA mempermudah proses analisis karena sistem bekerja secara otomatis dalam memilih atribut terbaik, membangun pohon keputusan, serta mengevaluasi performa model. Selain menghasilkan model klasifikasi yang akurat, proses implementasi ini juga menghasilkan representasi visual yang dapat digunakan sebagai alat bantu dalam memahami faktor-faktor yang memengaruhi kelulusan mahasiswa. Dengan demikian, implementasi algoritma C4.5 melalui WEKA tidak hanya memberikan hasil prediksi, tetapi juga memberikan wawasan analitis yang penting bagi pihak akademik dalam meningkatkan pengelolaan dan monitoring kinerja mahasiswa.

Evaluasi Model

Evaluasi model merupakan tahapan penting dalam penelitian ini karena bertujuan menilai seberapa baik algoritma C4.5 dalam mengklasifikasikan data akademik mahasiswa ke dalam kategori lulus atau tidak lulus. Evaluasi dilakukan menggunakan confusion matrix, yaitu sebuah metode yang membandingkan hasil prediksi model dengan kondisi aktual data untuk mengukur tingkat keberhasilan klasifikasi. Confusion matrix terdiri dari empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). TP adalah jumlah mahasiswa yang diprediksi lulus dan benar-benar lulus, TN adalah jumlah mahasiswa yang diprediksi tidak lulus dan benar-benar tidak lulus. Sebaliknya, FP menunjukkan mahasiswa yang diprediksi lulus tetapi sebenarnya tidak lulus, sedangkan FN adalah mahasiswa yang diprediksi tidak lulus namun sebenarnya lulus. Keempat komponen ini memberikan gambaran mendetail mengenai kualitas prediksi yang dihasilkan oleh model C4.5.

Berdasarkan nilai-nilai dalam confusion matrix tersebut, beberapa metrik evaluasi penting kemudian dihitung untuk mengukur performa model, yaitu akurasi, precision, dan recall. Akurasi merupakan ukuran yang menunjukkan persentase keseluruhan prediksi yang benar dibandingkan jumlah total prediksi. Akurasi sering digunakan sebagai indikator umum performa model, namun pada data yang tidak seimbang (misalnya jumlah mahasiswa lulus jauh lebih banyak daripada yang tidak lulus), akurasi saja tidak cukup untuk menggambarkan kualitas model secara menyeluruh. Precision digunakan untuk mengukur tingkat ketepatan prediksi kelulusan, yaitu seberapa besar mahasiswa yang diprediksi lulus benar-benar lulus dalam kenyataan. Precision sangat penting dalam konteks penelitian ini, karena prediksi

kelulusan yang salah dapat memberikan dampak negatif pada proses pengambilan keputusan akademik, misalnya memberi indikasi yang keliru tentang kondisi mahasiswa.

Di sisi lain, recall mengukur sejauh mana model mampu mengenali seluruh mahasiswa yang benar-benar lulus dari keseluruhan data yang tersedia. Dengan kata lain, recall menunjukkan kemampuan model dalam mendeteksi kelas positif secara menyeluruh. Nilai recall yang tinggi menandakan bahwa model tidak melewatkan banyak mahasiswa yang seharusnya terdeteksi sebagai lulus. Precision dan recall saling melengkapi, dan keduanya bersama dengan akurasi memberikan pemahaman yang lebih komprehensif mengenai performa model C4.5. Melalui evaluasi ini, peneliti dapat mengidentifikasi apakah model telah bekerja dengan baik atau masih terdapat kesalahan prediksi yang signifikan dan perlu diperbaiki.

Dalam penelitian ini, evaluasi model dilakukan menggunakan fitur evaluasi otomatis yang disediakan oleh perangkat lunak WEKA. WEKA menghitung confusion matrix secara langsung dan menampilkan nilai akurasi, precision, dan recall bersama dengan metrik lainnya seperti F-measure dan ROC area. Selain itu, evaluasi menggunakan metode validasi silang (cross-validation) juga dilakukan untuk memastikan bahwa model tidak mengalami overfitting, yaitu kondisi di mana model terlalu menyesuaikan diri dengan data pelatihan sehingga performanya menurun saat digunakan pada data baru. Validasi silang ini memberikan gambaran yang lebih realistis mengenai kemampuan model ketika diterapkan pada dataset mahasiswa di luar data pelatihan.

Secara keseluruhan, tahap evaluasi model ini memastikan bahwa algoritma C4.5 tidak hanya menghasilkan prediksi, tetapi juga memberikan performa klasifikasi yang dapat dipertanggungjawabkan secara ilmiah. Evaluasi menggunakan confusion matrix dan metrik turunannya memberikan dasar kuat bagi peneliti untuk menyimpulkan apakah model yang dibangun dapat digunakan sebagai alat bantu prediksi yang efektif dalam lingkungan akademik. Dengan nilai akurasi, precision, dan recall yang tinggi, model prediksi kelulusan mahasiswa dapat memberikan manfaat nyata bagi institusi pendidikan dalam melakukan monitoring, perencanaan akademik, dan intervensi dini terhadap mahasiswa yang membutuhkan perhatian khusus.

4. HASIL DAN PEMBAHASAN

Analisis Kinerja Model Klasifikasi Algoritma C4.5

Model klasifikasi prediksi kelulusan mahasiswa yang dibangun menggunakan algoritma C4.5 menunjukkan performa yang cukup baik berdasarkan hasil evaluasi yang

diperoleh. Berdasarkan hasil pengujian menggunakan confusion matrix, model menghasilkan akurasi sebesar 84,6%, yang berarti bahwa dari seluruh data mahasiswa yang diuji, 84,6% di antaranya berhasil diprediksi dengan benar oleh model. Angka ini menunjukkan bahwa algoritma C4.5 mampu mengenali pola-pola akademik mahasiswa secara efektif dan memetakannya dengan tepat ke dalam kategori *lulus* dan *tidak lulus*. Selain akurasi, model juga memiliki nilai precision sebesar 82%**,** yang menunjukkan bahwa sebagian besar mahasiswa yang diprediksi *lulus* oleh model memang benar-benar lulus dalam data aktual. Precision yang tinggi menandakan bahwa model tidak banyak melakukan kesalahan prediksi ketika menentukan suatu mahasiswa termasuk ke dalam kategori *lulus*.

Sementara itu, nilai recall sebesar 86% mengindikasikan bahwa model mampu mendeteksi sebagian besar mahasiswa yang benar-benar lulus dalam data. Dengan kata lain, model memiliki kemampuan yang baik dalam menangkap atau mengenali mahasiswa dengan status lulus dari seluruh populasi mahasiswa yang benar-benar lulus. Kombinasi antarnilai akurasi, precision, dan recall tersebut menunjukkan bahwa model yang dihasilkan tidak hanya akurat tetapi juga stabil dan konsisten dalam mengklasifikasikan data mahasiswa pada berbagai kondisi data pelatihan.

Struktur pohon keputusan (decision tree) yang dihasilkan oleh algoritma C4.5 juga memberikan gambaran visual mengenai atribut-atribut yang paling berpengaruh dalam menentukan status kelulusan mahasiswa. Pada pohon keputusan tersebut, Indeks Prestasi Kumulatif (IPK) muncul sebagai atribut pertama atau *root*, yang berarti IPK merupakan indikator paling dominan dalam memprediksi kelulusan. Mahasiswa dengan IPK tinggi cenderung memiliki peluang lulus yang lebih besar dibandingkan mahasiswa dengan IPK rendah. Setelah IPK, atribut kehadiran muncul sebagai variabel penting berikutnya. Tingkat kehadiran yang baik menunjukkan partisipasi aktif mahasiswa dalam proses pembelajaran. Atribut jumlah SKS yang telah ditempuh kemudian menjadi faktor pendukung yang memperkuat prediksi kelulusan mahasiswa, karena mahasiswa yang telah menyelesaikan sebagian besar SKS wajib umumnya berada pada tahap akhir studi.

Pengujian model juga dilakukan menggunakan metode k-fold cross-validation untuk memverifikasi konsistensi performa model. Hasil validasi silang menunjukkan bahwa model tetap memberikan performa yang stabil pada pembagian data yang berbeda, sehingga model C4.5 dinilai cukup andal dan robust untuk diterapkan pada dataset mahasiswa lainnya dalam skala yang lebih besar. Stabilitas performa ini membuktikan bahwa algoritma C4.5 tidak hanya bekerja baik pada satu set data, tetapi juga dapat mempertahankan kualitas prediksinya pada data baru.

Interpretasi Faktor-Faktor Penentu Kelulusan Berdasarkan Pohon Keputusan

Hasil penelitian ini memperlihatkan bahwa data akademik mahasiswa memiliki hubungan yang kuat terhadap kelulusan mereka, terutama pada atribut IPK, tingkat kehadiran, dan jumlah SKS. Atribut IPK yang menjadi akar pada pohon keputusan menunjukkan bahwa prestasi akademik merupakan indikator utama dalam menentukan kelulusan. Hal ini sejalan dengan sistem evaluasi akademik perguruan tinggi yang menggunakan IPK sebagai parameter utama dalam menilai capaian pembelajaran mahasiswa. Mahasiswa dengan IPK tinggi umumnya memiliki pemahaman materi kuliah yang baik, kemampuan menyelesaikan tugas tepat waktu, serta konsistensi dalam mengikuti proses pembelajaran, sehingga peluang kelulusannya juga lebih tinggi.

Selain IPK, kehadiran juga terbukti memiliki pengaruh signifikan terhadap status kelulusan. Kehadiran menjadi salah satu indikator keterlibatan aktif mahasiswa dalam lingkungan akademik. Mahasiswa dengan kehadiran tinggi cenderung mendapatkan pengalaman belajar yang lebih baik, memahami materi perkuliahan secara langsung, dan lebih disiplin dalam mengikuti kegiatan akademik. Tingkat kehadiran yang rendah sering kali berhubungan dengan rendahnya pemahaman materi, ketertinggalan dalam mengikuti perkembangan perkuliahan, hingga kemungkinan munculnya masalah akademik lainnya. Oleh karena itu, munculnya atribut kehadiran sebagai faktor kedua paling berpengaruh dalam pohon keputusan memperkuat temuan bahwa keterlibatan langsung dalam proses pembelajaran memiliki kontribusi besar terhadap kesuksesan akademik mahasiswa.

Atribut jumlah SKS yang telah ditempuh juga memberikan kontribusi penting dalam prediksi kelulusan. Mahasiswa yang telah menyelesaikan sebagian besar SKS yang diperlukan untuk kelulusan biasanya berada pada tahap akhir studi dan memiliki progres akademik yang baik. Sebaliknya, mahasiswa dengan jumlah SKS rendah atau progres lambat cenderung memiliki risiko keterlambatan studi yang lebih tinggi. Hal ini karena jumlah SKS berkaitan erat dengan lama masa studi dan pemenuhan persyaratan kurikulum. Dengan demikian, jumlah SKS menjadi indikator struktural yang dapat menunjukkan posisi mahasiswa dalam perjalanan studinya.

Struktur pohon keputusan yang dihasilkan oleh model C4.5 bukan hanya memberikan prediksi, tetapi juga menyajikan informasi yang mudah dipahami mengenai pola-pola akademik mahasiswa. Pohon keputusan tersebut dapat digunakan sebagai alat bantu visual dalam proses pengambilan keputusan oleh pihak kampus. Sebagai contoh, mahasiswa yang terklasifikasi ke dalam cabang pohon dengan risiko tidak lulus dapat diberikan perhatian lebih awal, seperti bimbingan akademik, program percepatan, atau konseling belajar. Dengan cara

ini, institusi dapat melakukan intervensi lebih cepat dan efektif dibandingkan dengan evaluasi manual yang sering dilakukan pada akhir semester.

Penggunaan algoritma C4.5 dalam penelitian ini juga memberikan nilai tambah karena model yang dihasilkan bersifat interpretatif, yaitu memberikan aturan-aturan keputusan (rules) yang mudah dipahami. Aturan tersebut dapat digunakan oleh pihak akademik sebagai landasan dalam merancang strategi akademik, meningkatkan monitoring mahasiswa, serta mengembangkan sistem pendukung keputusan berbasis data. Secara keseluruhan, temuan penelitian ini menunjukkan bahwa penggunaan data akademik dalam analisis prediksi kelulusan dapat memberikan gambaran yang lebih jelas mengenai faktor-faktor yang mempengaruhi keberhasilan mahasiswa, serta menawarkan solusi praktis berbasis teknologi yang dapat meningkatkan kualitas manajemen akademik di perguruan tinggi.

5. KESIMPULAN DAN SARAN

Kesimpulan

Penelitian ini menunjukkan bahwa algoritma C4.5 efektif digunakan untuk memprediksi kelulusan mahasiswa berdasarkan IPK, SKS, dan kehadiran. Dengan akurasi mencapai 84,6%, model ini mampu mengidentifikasi mahasiswa yang berpotensi lulus atau tidak lulus secara cukup akurat. Selain itu, model yang dihasilkan mudah dipahami sehingga cocok digunakan dalam pengambilan keputusan akademik. Secara keseluruhan, algoritma C4.5 memiliki potensi besar dalam mendukung pengelolaan akademik berbasis data.

Saran

Institusi pendidikan disarankan mulai mengimplementasikan sistem prediksi kelulusan berbasis data mining untuk membantu mendeteksi mahasiswa yang memiliki risiko keterlambatan studi. Untuk penelitian selanjutnya, disarankan menggunakan jumlah data yang lebih besar serta menambahkan variabel seperti kondisi sosial ekonomi atau aktivitas organisasi. Perbandingan dengan metode lain seperti Random Forest atau SVM juga dapat dilakukan untuk mengetahui model dengan performa terbaik. Selain itu, pengembangan dashboard visual interaktif akan memudahkan penyajian hasil prediksi bagi pihak akademik.

DAFTAR REFERENSI

Azizah, N. (2020). Analisis faktor penyebab keterlambatan studi mahasiswa. *Jurnal Pendidikan Tinggi*, 5(2), 112–120.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining. *Journal of Educational Data Mining*, 1(1), 3–17.

Delen, D. (2005). Predicting student attrition with data mining methods. *Computers & Education*, 50(3), 879–894.

Delen, D. (2005). *Predicting student attrition with data mining methods*. Computers & Education, 50(3), 879–894.

Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Hasibuan, M. (2018). *Manajemen pendidikan dan prestasi mahasiswa*. Jakarta: Prenada Media.

Hernawati, E., & Purwanto, A. (2021). Pengaruh kehadiran terhadap prestasi akademik mahasiswa. *Jurnal Evaluasi Pendidikan*, 12(1), 45–54.

Kotsiantis, S. (2013). *Decision Trees: A Recent Overview*. Artificial Intelligence Review, 39(4), 261–283.

Kumar, M., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, 2(6), 63–69.

Kumar, M., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, 2(6), 63–69.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state-of-the-art. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(6), 601–618.

Romero, C., & Ventura, S. (2010). *Educational Data Mining: A Review of the State-of-the-Art*. IEEE Transactions on Systems, Man, and Cybernetics, 40(6), 601–618.

Siregar, E. (2020). Hubungan penyelesaian SKS dengan masa studi mahasiswa. *Jurnal Pendidikan dan Kebijakan*, 9(1), 21–29.

Supriyanto, A., & Hidayat, T. (2019). Analisis tingkat kelulusan mahasiswa sebagai indikator mutu perguruan tinggi. *Jurnal Manajemen Pendidikan*, 7(4), 301–309.

Suryani, N. (2020). Penerapan Algoritma C4.5 untuk Klasifikasi Status Akademik Mahasiswa. *Jurnal Teknologi dan Sistem Informasi*, 8(2), 155–162.

Witten, I. H., Frank, E., & Hall, M. A. (2017). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.

Witten, I. H., Frank, E., & Hall, M. A. (2017). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufman.