Prediksi Kelulusan Mahasiswa Fakultas Teknologi Informasi ISB Atma Luhur Menggunakan Algoritma C4.5

Ine Widyaningrum Mustama Putri[1], Rusdah Rusdah[2], Lis Suryadi[3], Dian Anubhakti[4]
Program Studi Sistem Informasi
Fakultas Teknologi Informasi Universitas Budi Luhur
Jakarta, Indonesia
1912500939@student.budiluhur.ac.id[1], rusdah@budiluhur.ac.id[2], lis.suryadi@budiluhur.ac.id[3],
dian.anubhakti@budiluhur.ac.id[4]

***Abstract*— Higher Education is a level of education after secondary education which includes diploma programs, undergraduate programs, master programs, doctoral programs, professional programs, and specialist programs organized based on the culture of the Indonesian nation. Student graduation is one of the important factors to improve university accreditation. Students who graduate above 5 years and the number of students who drop out are important indicators in determining accreditation which then causes the difficulty of accrediting a college to rise. This research aims as an early warning for students who graduate on time and graduate late from the Faculty of Information Technology, Institute of Science and Business Atma Luhur using the C4.5 decision tree algorithm by implementing the Cross-Industry Standard Process for Data Mining (CRISP- DM) method. The initial data of this research amounted to 1,015 which was taken through a query in the database of the Atma Luhur Institute of Science and Business. However, the data that will be used becomes 694 after preprocessing due to the large number of record contents that do not have a graduation year, with a total of 641 graduates graduating on time and 53 graduates graduating late. Based on the application of the model using the C4.5 decision tree algorithm and the Confusion Matrix method, the accuracy is 93.94%, Recall is 98.59%, and Precision is 95.03%. So it can be concluded that the C4.5 decision tree algorithm is the most effective algorithm for predicting student graduation, because it has a high level of accuracy.**

Keywords*— Prediction, Graduation, Decision Tree C4.5, Early Warning

***Abstrak*— Perguruan Tinggi merupakan jenjang pendidikan setelah pendidikan menengah yang mencakup program diploma, program sarjana, program magister, program doktor, program profesi, dan program spesialis yang diselenggarakan berdasarkan kebudayaan bangsa Indonesia. Kelulusan mahasiswa menjadi salah satu faktor penting untuk meningkatkan akreditasi perguruan tinggi. Mahasiswa yang lulus diatas 5 tahun dan jumlah mahasiswa yang mengalami drop out merupakan indikator penting dalam penentuan akreditasi yang kemudian menyebabkan sulitnya akreditasi sebuah perguruan tinggi naik. Penelitian ini bertujuan sebagai early warning atau peringatan dini untuk kelulusan mahasiswa yang lulus tepat waktu dan lulus terlambat fakultas teknologi informasi, Institut Sains dan Bisnis Atma Luhur menggunakan algoritma decision tree C4.5 dengan mengimplementasikan metode Cross-Industry Standard Process for Data Mining (CRISP- DM). Data awal penelitian ini berjumlah 1.015 yang diambil melalui query dalam database ISB Atma Luhur. Namun, data yang akan digunakan menjadi 694 setelah dilakukan preprocessing dikarenakan banyaknya isi record yang tidak memiliki tahun kelulusan, dengan total 641 wisudawan lulus tepat waktu dan 53 wisudawan lulus terlambat. Berdasarkan penerapan model menggunakan algoritma decision tree C4.5 dan metode Confusion Matrix diperoleh Akurasi sebesar 93.94%, Recall sebesar 98.59%, dan Presisi sebesar 95.03%. Maka dapat disimpulkan bahwa algoritma decision tree C4.5 adalah algoritma yang paling efektif untuk memprediksi kelulusan mahasiswa, karena memiliki tingkat akurasi yang tinggi.**

Kata Kunci*— Prediksi, Kelulusan, Decision Tree C4.5, Early Warning

I. PENDAHULUAN

Perguruan tinggi merupakan suatu instansi yang menyelenggarakan proses belajar, mengajar, penelitian serta pengabdian kepada masyarakat atau biasa disebut lembaga penyelenggara Tri Dharma Perguruan Tinggi. Setiap tahun perguruan tinggi menghasilkan lulusan mahasiswa sesuai bidang yang ditempuhnya, dalam kurun waktu penyelesaian studi tepat waktu. Lulus tepat waktu merupakan salah satu indikator keberhasilan mahasiswa dalam memperoleh gelar sarjana. Mahasiswa lulus tepat waktu apabila mampu menyelesaikan studinya di perguruan tinggi selama kurang dari atau sama dengan empat tahun, sedangkan mahasiswa dikatakan tidak lulus tepat waktu apabila menyelesaikan studinya lebih dari empat tahun. Namun tidak setiap mahasiswa dapat menyelesaikan pendidikan sarjana dalam kurun waktu 4 (empat) tahun di perguruan tingginya, mahasiswa yang telah menyelesaikan pendidikan sarjana kemudian mendaftar sebagai calon wisudawan baru kemudian bisa mengikuti wisuda. Berdasarkan Peraturan Menteri Riset Teknologi dan Pendidikan Tinggi Nomor 44 Tahun 2015 tentang Standar Nasional Pendidikan Tinggi terkait dengan beban belajar dan masa belajar mahasiswa program sarjana paling lama 7 (tujuh) tahun.

ISB (Institut Sains dan Bisnis) Atma Luhur adalah salah satu perguruan tinggi dalam bidang komputer dan bisnis digital di Provinsi Kepulauan Bangka Belitung khususnya di kota Pangkal Pinang. ISB Atma Luhur berdiri pada tahun 2020

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i3.1731, Copyright ©2023
Submitted : May 18, 2023, Revised : June 9, 2023, Accepted : August 29, 2023, Published : November 4, 2023

merupakan pengembangan atau rubah bentuk dari Sekolah Tinggi Manajemen Informatika dan Komputer Atma Luhur yang berdiri sejak tahun 2009. Saat ini telah berdiri dan berkembang beberapa sekolah tinggi dan universitas di Provinsi Kepulauan Bangka Belitung, sehingga lingkungan ISB Atma Luhur Fakultas Teknologi Informasi berada dalam lingkungan yang kompetitif.

Persentase kelulusan tepat waktu pada tahun ajaran 2015/2016 adalah 46%, sedangkan pada tahun ajaran 2016/2017 persentase kelulusan tepat waktu adalah 65% dan pada tahun ajaran 2017/2018 persentase kelulusan tepat waktu adalah 61%. Sehingga rata-rata persentase kelulusan tepat waktu tahun ajaran 2015/2016, 2016/217, 2017/2018 adalah 57.3%. Pada tahun ajaran 2015/2016 mahasiswa baru berjumlah 376, mahasiswa aktif sejumlah 257 sedangkan mahasiswa yang lulus tepat waktu tahun ajaran 2019/2020 berjumlah 233. Pada tahun ajaran 2016/2017 mahasiswa baru berjumlah 295, mahasiswa aktif sejumlah 209 sedangkan yang lulus tepat waktu tahun ajaran 2020/2021 berjumlah 194. Pada tahun ajaran 2017/2018 mahasiswa baru berjumlah 344, mahasiswa aktif sejumlah 342 sedangkan yang lulus tepat waktu tahun ajaran 2021/2022 berjumlah 214. Maka manajemen membutuhkan penelitian yang dapat memprediksi kelulusan mahasiswanya tepat waktu.

Salah satu teknik melakukan prediksi yang dapat dilakukan adalah dengan teknik penggalian data atau *data mining*. *Data mining* adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan. Dengan salah satu teknik dari *data mining* yaitu klasifikasi dengan algoritma C4.5 [1]. Algoritma C4.5 dipilih karena mampu memprediksi dengan memberikan tingkat nilai akurasi yang ideal dan hasil yang optimal untuk memprediksi yang diharapkan dapat menemukan informasi tingkat kelulusan dan persentase kelulusan mahasiswa sehingga dapat digunakan oleh pihak manajemen untuk mencari solusi dalam proses evaluasi pembelajaran di Fakultas Teknologi Informasi ISB Atma Luhur [2].

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metode prediksi yang berguna untuk mengeksplorasi informasi dari sebuah dataset, dalam penelitian ini digunakan dataset kelulusan mahasiswa Fakultas Teknologi Informasi ISB Atma Luhur tahun 2015-2017. Pada penelitian ini, penulis memprediksi kelulusan mahasiswa menggunakan metode data mining menggunakan *tools* RapidMiner untuk menunjang penelitian yang dilakukan.

Penelitian dengan menggunakan algoritma C4.5, tersebut menggunakan perbandingan algoritma C4.5 dan KNN dan menghasilkan data keakuratan yang tinggi diatas 80% [3].

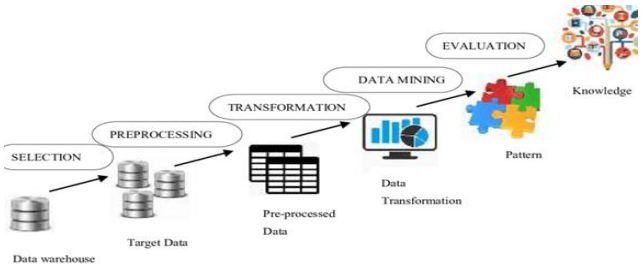
Penelitian lain yang menggunakan algoritma C4.5, tersebut uji coba untuk nilai akurasi dilakukan sebanyak 3 kali dengan jumlah *record data training* yang berbeda. Semakin banyak jumlah *record* data yang digunakan untuk proses data training maka semakin tinggi juga nilai akurasinya[4].

II. STUDI PUSTAKA

A. *Data Mining*

Data mining merupakan suatu proses penggalian atau penambangan informasi untuk menemukan informasi yang tidak ditemukan sebelumnya dari sekumpulan data besar (*big data*) atau *repository database* lainnya. *Data mining* bukan sekedar terkumpul data saja tetapi mencangkup analisis dan prediksi dari informasi yang ingin ditampilkan [5]. Data yang dikumpulkan disimpan dalam *database* kemudian diproses sehingga dapat dijadikan untuk pengambilan keputusan dalam melihat informasi yang akan digunakan [6].

*Knowledge Discovery In Databas*e (KDD) merupakan metode untuk memperoleh pengetahuan dari database yang ada. Dalam *database* terdapat tabel - tabel yang saling berhubungan / berelasi. Hasil pengetahuan yang diperoleh dalam proses tersebut dapat digunakan sebagai basis pengetahuan (*knowledge base*) untuk keperluan pengambilan keputusan. Istilah *Knowledge Discovery in Database* (KDD) dan *data mining* seringkali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain, dan salah satu tahapan dalam keseluruhan proses KDD adalah *data mining* [7].



Gambar 1. Proses KDD

B. Tahapan proses KDD ada 5 [8] yaitu :

1) *Data Selection:* pemilihan data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai.

2) *Preprocessing:* sebelum proses data mining dapat dilaksanakan perlu dilakukan proses *cleaning* dengan tujuan untuk membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak. Juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal [9].

3) *Transformation:* yaitu proses *coding* pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam *database*.

4) *Data Mining:* proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu [10].

5) *Interpretation / Evaluation:* pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i3.1731, Copyright ©2023
Submitted : May 18, 2023, Revised : June 9, 2023, Accepted : August 29, 2023, Published : November 4, 2023

bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya atau tidak.

C. *Metode Decision Tree*

Pohon (*tree*) merupakan sebuah struktur data yang terdiri dari simpul (*node*) dan rusuk (*edge*). Simpul pada sebuah pohon dibedakan menjadi tiga, yaitu simpul akar, simpul percabangan dan simpul daun. Pohon keputusan merupakan presentasi sederhana dari teknik klasifikasi untuk sejumlah kelas berhingga, simpul internal maupun simpul akar ditandai dengan nama atribut, rusuk – rusuknya diberi label nilai atribut dan simpul daun ditandai dengan kelas – kelas yang berbeda [11].

Pohon keputusan merupakan sebuah tampilan grafis proses pengambilan keputusan yang mengidentifikasi alternatif yang ada, kondisi alamiah dan peluangnya dan juga imbalan bagi setiap kombinasi alternatif keputusan. Cara kerja pohon keputusan yaitu mengubah bentuk data tabel menjadi model pohon, mengubah model pohon menjadi *rule* (aturan), dan menyederhanakan *rule* (aturan).

D. Algoritma C4.5

Algoritma C4.5 merupakan salah satu algoritma yang dapat digunakan untuk mengkonstruksi sebuah pohon keputusan. Algoritma C4.5 merupakan pengembangan algoritma ID3 (Quinlan, 1993), menentukan seberapa informatif sebuah masukan atribut dimana kekurangan yang dimiliki algoritma ID3 ditutupi oleh algoritma C4.5. Empat hal yang membedakan algoritma C4.5 dengan ID3 antara lain: tahan (*robust*) terhadap *data noise*, mampu menangani variabel dengan tipe diskrit maupun kontinu, mampu menangani variabel yang memiliki *missing value*, dan dapat memangkas cabang dari pohon keputusan [12].

Secara umum Algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

- Pilih atribut sebagai akar
- Buat cabang untuk masing – masing nilai
- Bagi kasus dalam cabang
- Ulangi proses untuk masing – masing cabang sampai semua kasus pada cabang memiliki kelas yang sama. Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut – atribut yang ada, untuk menghitung gain digunakan rumus seperti yang tertera pada persamaan (1).

$$Gain(S,A) = Entropy(\mathrm{S}) - \sum_{n=1}^{n} \frac{|S_i|}{|S|} * Entropy(Si) \qquad \textbf{(1)}$$

Keterangan:

S : Himpunan Kasus
A : Atribut
n : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke i
|S| : Jumlah kasus dalam S

Sebelum mendapatkan nilai *Gain* adalah dengan mencari nilai Entropi. Entropi digunakan untuk menghasilkan sebuah atribut. Rumus dasar dari Entropi seperti terdapat pada persamaan (2).

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2\, pi \qquad \textbf{(2)}$$

Keterangan:

S : Himpunan kasus
n : Jumlah partisi S
pi : Proporsi dari Si terhadap S

E. *Cross Validation*

Cross *Validation* atau dapat disebut estimasi rotasi adalah sebuah teknik validasi model untuk menilai bagaimana hasil statistik analisis akan menggeneralisasi kumpulan data independen, teknik ini utamanya digunakan untuk melakukan prediksi model dan memperkirakan seberapa akurat sebuah model prediktif ketika dijalankan dalam prakteknya. Dalam sebuah masalah prediksi, sebuah model biasanya diberikan kumpulandata (*dataset*) yang diketahui untuk digunakan dalam menjalankan pelatihan (*data training*), serta kumpulan data yang tidak diketahui (*data testing*)terhadap model yang diuji [13].

F. Presisi, *Recall* dan Akurasi

Precission (presisi) dan *recall* digunakan untuk mengukur kinerjasistem. Presisi adalah kecocokan antara bagian data yang diambil dengan informasi yang dibutuhkan. *Recall* merupakan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. *Accuracy* (akurasi) adalah tingkat kedekatan antara nilai yang didapat terhadap nilai sebenarnya. Presisi,*recall* dan akurasi dapat dihitung menggunakan *confusion matrix*. Ukuran besaran presisi, *recall*, dan akurasi biasanya diberi nilai dalam bentuk presentase antara 1 sampai 100%. Sebuah sistem akan dianggap baik jika tingkat presisi, *recall*, dan akurasi-nya tinggi [14].

G. *Tools* yang Digunakan

Dalam penelitian ini, penulis menggunakan *software* atau *tools*, antara lain

1) RapidMiner versi 9.10: RapidMiner dalam penelitian ini digunakan untuk melakukan tahap *modeling* data. RapidMiner adalah aplikasi atau perangkat lunak yang berfungsi sebagai alat pembelajaran dalam ilmu data mining. Platfrom dikembangkan oleh perusahaan yang didedikasikan untuk semua langkah yang melibatkan sejumlah besar data dalam bisnis komersial, penelitian, pendidikan, pelatihan, dan pembelajaran [15].

2) Microsoft Excel 2011: Program aplikasi Microsoft Excel ini memiliki banyak fitur dan fungsi yang digunakan untuk mengolah angka. Fitur Fungsi dan Formula atau yang lebih dikenal dengan rumus Excel menjadikannya terkenal dan banyak digunakan dalam berbagai bidang dan persoalan seperti membuat, mengedit, mengurutkan, menganalisa, serta meringkas data [16].

Microsoft Excel dalam penelitian ini digunakan untuk mengolah data *reprocessing*. Pengelolaan atau pengolahan data

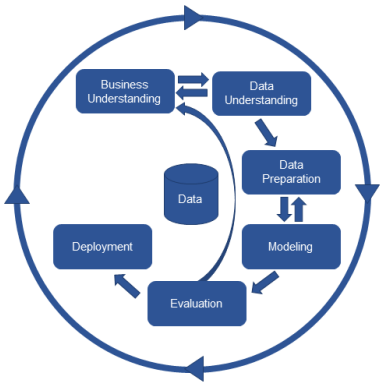
p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i3.1731, Copyright ©2023
Submitted : May 18, 2023, Revised : June 9, 2023, Accepted : August 29, 2023, Published : November 4, 2023

berfungsi untuk pengelolaan *database* statistik ,mencari nilai tengah, rata-rata, pencarian nilai maksimum [17].

III. METODOLOGI PENELITIAN

Pada tahapan penelitian dijelaskan apa saja yang dilakukan beserta luaran dari setiap tahapan. Dijelaskan juga bagaimana data dikumpulkan dan luaran dari data tersebut, bagaimana data yang diperoleh dianalisis atau diolah untuk diproses pada tahapan berikutnya.

Pada penelitian ini menggunakan metodologi yakni CRISP-DM untuk melakukan analisis dan mengolah data sebagai pemecah masalah yang umum untuk bisnis dan penelitian. Metodologi ini terdiri dari enam tahapan yaitu *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment* [18]. Proses metodologi ini terdiri dari 6 tahapan yang dapat dijelaskan sebagai berikut :



Gambar 2. Diagram CRISP-DM

A. Diagram CRISP-DM

1) Pemahaman Bisnis *(Business Understanding)*: pada tahap *business understanding,* dilakukan wawancara langsung dengan narasumber dari Biro Sistem Informasi ISB Atma Luhur. Wawancara ini bertujuan untuk menemukan identifikasi masalah. Hasil dari tahapan ini adalah diketahuinya atau diidentifikasinya rumusan masalah dan tujuan.

2) Pemahaman Data *(Data Understanding)*: pada tahap ini, berdasarkan tujuan penelitian maka dilakukan permintaan data kelulusan mahasiswa tahun ajaran 2015/2016, 2016/2017, dan 2017/2018 melalui *query* pada Biro Sistem Informasi ISB Atma Luhur, setelah itu didapatkan beberapa atribut yang diperlukan oleh penulis untuk penelitian. Data wisudawan yang terkumpul sebanyak sebanyak 1015 *record* dan 57atribut. Pada tahap ini penulis mempelajari data agar dapat mengenal seperti apa data yang akan digunakan untuk keperluan penelitian.

Atribut yang didapatkan adalah program studi, NIM, nama, jenis kelamin, asal SMA, tahun ajaran Lulus, tanggal lulus, keterangan keaktifan perkuliahan semester 1-12, IPS semester 1-12, jumlah SKS yang diambil pada semester 1-12, IPK semester 1-12, jenis kelas, SKS lulus semester 1-12.

3) Persiapan Data *(Data Preparation)*: pada tahap *data preparation*, data yang telah dikumpulkan akan dilakukan *preprocessing* menggunakan Microsoft Excel sebagai berikut:

a) *Data Reduction:* kegiatan *preprocessing* yang dilakukan pada tahapan ini adalah reduksi data yaitu penghapusan atribut yaitu atribut NIM, nama, SKS yang diambil semester 3-12, IPK semester 3-12, keterangan aktif kuliah semester 3-12. Proses penghapusan ini dilakukan karena atribut tersebut sudah tidak dapat digunakan sebagai *early warning* dalam menentukan tepat atau terlambat nya lama masa studi mahasiswa.

b) *Data Cleaning*: terhadap 6 *record missing value* pada atribut SKS semester 2 dan SKS lulus semester 2.

c) *Data Transformation:* tahap ini menggabungkan semua *sheet* tahun ajaran 2015/2016, 2016/2017, dan 2017/2018 ke dalam satu *sheet*, kemudian membentuk data baru yaitu masa studi diperoleh dari tahun lulus dikurangi tahun masuk, dari masa studi ini akan dibuat atribut keterangan lulus.

d) Penentuan Data *Training* dan Data *Testing*: data hasil *preprocessing* dipecah menjadi data *training* (data latih) dan data *testing* (data uji) menggunakan *split* data dengan rasio 60:40, 70:30, dan 80:20.

4) Pemodelan (*Modeling*): pada tahap ini penelitian melakukan pemodelan terhadap dataset yang sudah dilakukan *Preprocessing* dan telah ditentukan *data training* dan *data testing* serta tanpa *split* yaitu *cross validation.* Pada tahapan ini akan dieksplorasi beberapa algoritma untuk klasifikasi untuk mendapatkan model yang terbaik.

5) Evaluasi (*Evaluation)*: pada tahap evaluasi, dilakukan pengukuran kinerja algoritma model, algoritma ini dapat melakukan klasifikasi terhadap prediksi kelulusan mahasiswa menggunakan *confusion matrix* yaitu presisi, *recall*, dan akurasi.

6) Penyebaran (*Deployment*): Tahapan ini dilakukan dengan pembuatan laporan dan artikel jurnal menggunakan model yang dihasilkan.

B. Perbedaan dari penelitian sebelumnya

Dalam penelitian sebelumnya oleh Rizki Muliono, J. H. Lubis, dan Nurul Khairina yang menggunakan *algoritma K-Nearest Neighbor* (KNN) tingkat Akurasi yang didapatkan level K3 lebih baikdari level K1 dan K2 yaitu 98.5% [19].

Dalam penelitian sebelumnya oleh Abdul Rohman dan Anief Rofiyanto yang menggunakan algoritma DecisionTree C4.5 rata-rata tingkat Akurasi nya 65.98% [20].

Dalam penelitian sebelumnya oleh Nurul Khasanah et al. yang menggunakan algoritma *Naïve Bayes* tingkat Akurasi nya rata-rata sebesar 88.16% [21] .

C. *Data Preprocessing*

Data yang digunakan harus melewati *data preprocessing*, karena sumber data yang diperoleh dari database masih bersifat kotor, artinya terdiridari beberapa data yang tidak lengkap, data hilang atau kosong, kekurangan atribut tertentu atau atribut yang sesuai, *data noise*, dan data yang tidak konsisten.

Preprocessing terdiri dari 4 jenis yaitu *data cleaning, data integration, data transformation,* dan *data reduction.* Jumlah data tahun kelulusan 2019, 2020, dan 2021 yang diperoleh sebanyak 694 data. Pada penelitian ini,*preprocessing data* yang akan dilakukan adalah *cleaning data* yaitu digunakan untuk melengkapi atau menghilangkan data yang tidak lengkap, menghapus data atribut yang tidak diperlukan dalam proses

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i3.1731, Copyright ©2023
Submitted : May 18, 2023, Revised : June 9, 2023, Accepted : August 29, 2023, Published : November 4, 2023

klasifikasi, dan memperbaiki data yang tidak konsisten, membentuk atribut baru. *Data preprocessing* dilakukan dengan memfilter secara manual menggunakan microsoft excel.

D. Teknik Pengujian

Pada tahap ini pengujian dilakukan dengan metode *confusion matrix* mempresentasikan hasil evaluasi model. Evaluasi menggunakan *confusion matrix* menghasilkan nilai akurasi, presisi dan *recall*. Akurasi dalam klasifikasi merupakan presentasi ketepatan *record data* diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi. Presisi merupakan proposikasi yang diprediksi positif yang juga positif benar pada data sebenarnya. *Recall* merupakan proporsi kasus positif yang sebenarnya diprediksi positif secara benar [22].

IV. ANALISIS DAN PEMBAHASAN

Data yang digunakan adalah data sekunder yang didapatkan dari Biro Sistem Informasi ISB Atma Luhur. Data ini ditarik menggunakan *query* dan *function* di Oracle. Data keseluruhan yang digunakan pada penelitian adalah 694. Atribut yang digunakan dari data penelitian adalah program studi, jenis kelamin, asal SMA, keterangan masa aktif kuliah semester 1 dan 2, IPS semester 1 dan 2, jumlah SKS yang diambil pada semester1 dan 2, IPK semester 1 dan 2, jenis kelas,jumlah SKS lulus semester 1 dan 2, dan keterangan lulus.

A. *Data Preprocessing* atau Pra Pemrosesan Data

1) Data *Reduction:* kegiatan *preprocessing* yang dilakukan pada tahapan ini adalah reduksi data yaitu penghapusan atribut yaitu atribut NIM, nama, SKS yang diambil Semester 3-12, SKS Lulus Semester 3-12, IPK Semester 3-12, IPS 3-12, keterangan aktifkuliah semester 3-12. Proses penghapusan ini dilakukan karena atribut tersebut sudah tidak dapat digunakan sebagai *early warning* dalam menentukan tepat atau terlambatnya lama masa studi mahasiswa.

Kemudian dilakukan proses penghapusan data mahasiswa tahun ajaran tahun ajaran 2015/2016, 2016/2017, dan 2017/2018 yang tidak memiliki data tanggal TA dan tanggal lulus. Artinya mahasiswa tersebut tidak lulus, sejumlah 320 mahasiswa. Terdapat 1 *record* yang tidak memiliki nilai pada atribut SKS lulussemester 2, namun memiliki nilai IPS semester 2, sehingga diputuskan untuk menghapus satu *record* tersebut. Sehingga *dataset* setelah proses reduksi data adalah 16 atribut, 694 *record*.

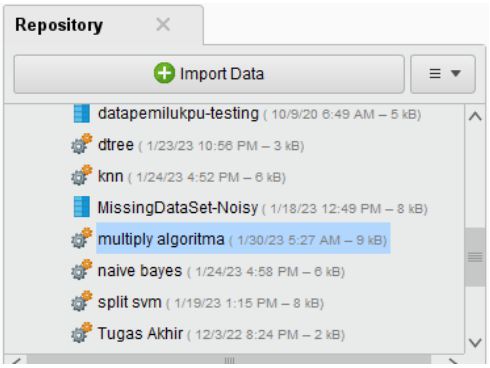
2) Data Cleaning: terhadap 6 *record missing value* pada atribut SKS semester 2 dan SKS lulus semester 2. Karena atribut aktif semester 2 bernilai tidak, maka atribut SKS semester 2 dan SKS lulus semester 2 diisi dengan nilai 0 (nol).

3) Data Transformation: tahap ini menggabungkan semua sheet tahun ajaran 2015/2016, 2016/2017, dan 2017/2018 ke dalam satu sheet, kemudian membentuk data baru yaitu masa studi diperoleh dari tahun lulus dikurangi tahun masuk. Sehingga jumlah atribut menjadi 17 atribut. Berdasarkan atribut masa studi dibuat atribut keterangan lulus. Bila masa studi = 4 tahun, maka keterangan lulus berisi tepat. Bila masa studi > 4 tahun, maka keterangan lulus berisi telat.

4) *Data Reduction* tahap kedua: pada tahap ini atribut tahun masuk, tahun lulus dan masa studi dihapus. Sehingga dataset terdiri atas 14 atribut.

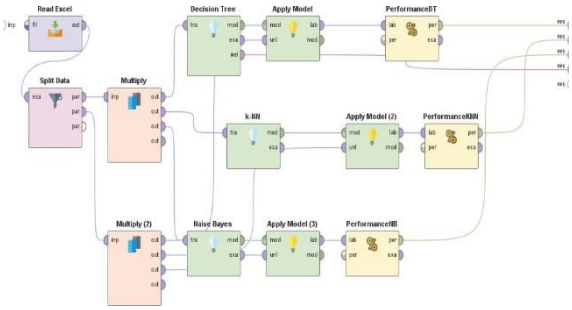
B. Penentuan Data Latih dan Data Uji

Setelah *preprocessing* data, selanjutnya akan dilakukan penentuan data *training* (data latih) dan data *testing* (data uji). Data *training* adalah data yangdigunakan untuk melatih mesin agar dapat mengenali pola, sedangkan data testing adalah data yang digunakan untuk menguji hasil dari pelatihan yang telah dilakukan terhadap mesin [23].



Gambar 3. Penentuan Data Latih dan Data Uji

Pemodelan dilakukan dengan memanfaatkan aplikasi RapidMiner. Proses diawali dengan mengambil data training yang telah disediakan sebelumnya. Ditambahkan dengan operator *Read Excel*, untuk melakukan import data karena data training berupa file excel; operator *Split Data* untuk membagi data; operator *Multiply*, untuk membuat salinan objek data; operator Algoritma C4.5, KNN dan *Naïve Bayes*; operator *Apply Model*, untuk menerapkan model yang telah dilatih sebelumnya menggunakan *data training* pada *unlabeled data* (*data testing*) dan operator *Performance*, untuk mengevaluasi kinerja model [24].



Gambar 4.Proses Pemodelan menggunakan Teknik Split Data.

Penentuan *data training* dan *data testing* dilakukan dengan cara membagi data *(split data)* dengan perbandingan.Dalam penelitian ini data *training* dan data *testing* dibagi menjadi tiga perbandingan yaitu 80:20, 70:30, dan 60:40 [25]. Sebaran data *training* seperti pada tabel 1.

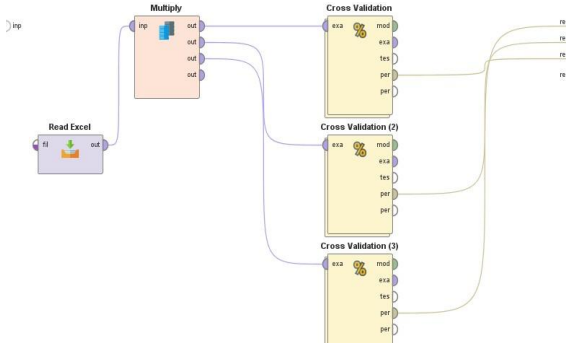
Tabel 1. Tabel Sebaran Data Training

Komposisi *Dataset*	**Data *Training***	**Data *Testing***

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i3.1731, Copyright ©2023
Submitted : May 18, 2023, Revised : June 9, 2023, Accepted : August 29, 2023, Published : November 4, 2023

60 : 40	416	278
70 : 30	486	208
80 : 20	555	139

Teknik *sampling* yang dipilih adalah *stratified random sampling*. Setelah menentukan pembagian perbandingan untuk data *training* dan data *testing*, kemudian membuat persebaran data untuk masing-masing bagian tersebut. Sebagai contoh perbandingan 80:20 yang berarti 80% merupakan data *training* dan 20% merupakan data *testing*.



Gambar 5. Eksplorasi dengan parameter uji akurasi

Pada Gambar 5 dieksplorasi menggunakan teknik penentuan data *training* dan *testing* menggunakan *10-fold cross validation* dengan teknik *stratified random sampling.* Hasil eksplorasi menggunakan *split* data dan *10-fold cross validation* disajikan pada Tabel 2 dengan parameter uji yaitu akurasi.

Tabel 2. Parameter Uji Akurasi

Dataset	**Akurasi**		
	C4.5	**KNN**	***Naïve Bayes***
60:40	92.45%	94.60%	92.09%
70:30	91.83%	93.75%	91.35%
80:20	92.81%	94.24%	92.81%
10-Fold Cross Validation	93.95%	93.67%	91.79%

Setelah melakukan eksplorasi, dapat dilihat pada tabel II bahwa komposisi data *training* dan *testing* yang menghasilkan akurasi tertinggi dengan *classifier* C4.5 diperoleh dari dataset *cross validation* yaitu sebesar 93,95%. Sedangkan pada *classifier* KNN diperoleh dari dataset dengan komposisi 60:40, yaitu sebesar 94,60%. Dan pada *classifier Naïve Bayes* akurasi tertinggi diperoleh dari dataset dengan komposisi 80:20, yaitu sebesar 92,81%.

C. Hasil Komparasi Model

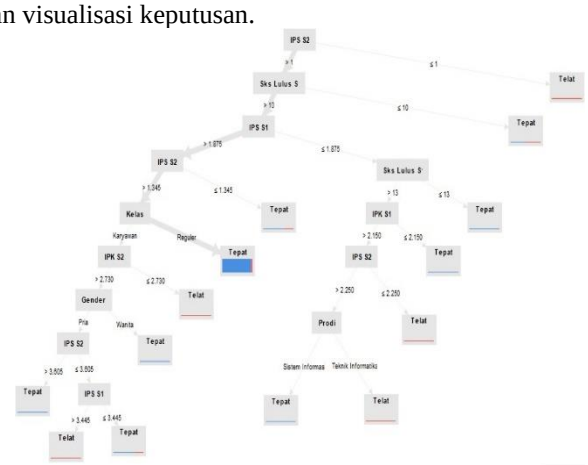
Pada tahap awal pemodelan, dilakukan eksplorasi beberapa algoritma untuk mengetahui performa model terbaik. Parameter uji yang digunakan adalah nilai akurasi, presisi, *recall* dan AUC. Tabel 3 merupakan hasil akurasi algoritma *decision tree C4.5* menggunakan metode 10-*fold cross validation.*

Tabel 3. Hasil Akurasi

Algoritma	**Akurasi**	**Presisi**	***Recall***	**AUC**
C4.5	93.95%	69.00%	39.67%	68.2%
KNN	93.67%	68.33%	30.33%	71.5%
Naïve Bayes	91.79%	46.78%	40.00%	78.5%

D. Hasil Komparasi Model

Setelah dilakukan hasil komparasi model, selanjutnya pada bagian ini, disajikan hasil pemrosesan dari model atau algoritma *Decision Tree* C4.5. Implementasi algoritma *Decision Tree* C4.5 terdiri dari, pembentukan *decision rules* dan visualisasi keputusan.



Gambar 6. Visualisasi Pohon Keputusan *Decision Tree*

Dalam visualisasi pohon keputusan terdapat 2 kesimpulan yaitu:

- Atribut atau faktor yang paling berpengaruh adalah IPS semester 1 dan 2, SKS lulus semester 1, kelas, IPK semester 1 dan 2, jenis kelamin, program studi.
- Atribut atau faktor yang tidak berpengaruh adalah asal SMA, keterangan aktif semester 1 dan 2, SKS lulus semester 2.

E. Pengujian

Pada tahap ini pengujian menggunakan metode *confusion matrix* dan parameter uji yang terdiri dari akurasi, presisi, *recall*. Hasil pengujian disajikan dalam bentuk tabulasi, menggunakan perbandingan *confusion matrix*. Pengujian *confusion matrix* untuk dataset yang diolah menggunakan algoritma *decision tree* yang dapat dilihat pada tabel 4.

Tabel 4. Pengujian *confusion matrix*

	***True* Tepat Waktu**	***True* Telat**	***Class Precission***
Prediction Tepat Waktu	631	33	95.03%
Prediction Telat	9	21	70.00%
Class *Recall*	98.59%	38.89%	

Pada *prediction* tepat waktu menghasilkan nilai *true* tepat waktu senilai 631, *true* telat senilai 33 akan menghasilkan nilai *class precission* 95.03% sedangkan untuk *prediction* telat menghasilkan *true* tepat waktu senilai 9, *true* telat senilai 21 akan di dapatkan hasil *class precission* 70.00%. Hasil dari *class recall* true tepat waktu 98.59%, *true* telat 38.89%. Perhitungan

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i3.1731, Copyright ©2023
Submitted : May 18, 2023, Revised : June 9, 2023, Accepted : August 29, 2023, Published : November 4, 2023

akurasi, presisi, dan *recall* yang dilakukan juga secara manual untuk membandingkan hasil perhitungan RapidMiner seperti pada persamaan (3), (4), dan (5) :

$$Akurasi = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} = \frac{631 + 21}{631 + 21 + 33 + 9} = 93.94\% \quad \textbf{(3)}$$

$$Presisi = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{631}{631 + 33} = 95.03\% \quad \textbf{(4)}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{631}{631 + 9} = 9 \quad \textbf{(5)}$$

Pengolahan data yang dilakukan pada tahap *preprocessing* adalah memilih atribut yang tidak memiliki *record*, atribut yang tidak diperlukan sebagai *early warning*, dan atribut yang terdapat *missing value*. Dataset yang awal mulanya berjumlah 1.015 setelah dilakukannya *preprocessing* menjadi berjumlah 694. terkumpul selanjutnya ditentukan data *training* dan data *testing* menggunakan *split* data dan *10-fold cross validation* dengan *tools* RapidMiner. Selanjutnya dilakukan pengujian untuk mengetahui hasil uji akurasi, presisi, dan *recall* menggunakan *confussion matrix*. Setelah mendapatkan hasil uji akurasi, maka dapat diambil kesimpulan bahwa algoritma C4.5 mermpunyai hasil uji akurasi paling tinggi dibandingkan dengan algortima KNN dan *Naïve Bayes.*

V. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan untuk memprediksi kelulusan mahasiswa pada Fakultas teknologi Informasi ISB Atma Luhur maka dapat ditarik kesimpulan bahwa algoritma C4.5 terbukti lebih akurat dalam memprediksi kelulusan mahasiswa karena mempunyai nilai akurasi paling tinggi dibandingkan dengan algoritma KNN dan *Naïve Bayes* senilai 93.94%. Algoritma C4.5 merupakan algoritma yang paling akurat mengingat kelebihan algortima ini paling efisien dalam menangani banyak tipe atribut diskret dab numerik dalam dataset. Untuk penelitian selanjutnya sebaiknya menggunakan atribut yang lebih banyak agar menghasilkan data yang lebih akurat.

Penelitian selanjutnya dapat dilakukan dengan menambahkan metode *feature selection* untuk mendapatkan hasil yang lebih baik dalam klasifikasi.

REFERENCES

[1] A. K. Wahyudi, N. Azizah, and H. Saputro, "Data Mining Klasifikasi Kepribadian Siswa SMP Negeri 5 Jepara Menggunakan Metode Decision Tree Algoritma C4.5," *JISTER (Journal of Information System and Computer Science),* vol. 2, no. 2, 2022.

[2] R. H. Pambudi and B. D. Setiawan, "Penerapan Algoritma C4.5 Untuk Memprediksi Nilai Kelulusan Siswa Sekolah Menengah Berdasarkan Faktor Eksternal," vol. 2, no. 7, 2018.

[3] E. Purwanto, "Prediksi Kelulusan Tepat Waktu Menggunakan Metode C4.5 Dan K-NN (Studi Kasus : Mahasiswa Program Studi S1 Ilmu Farmasi, Fakultas Farmasi, Universitas Muhammadiyah Purwokerto)," vol. 20, no. 2, pp. 131–142, 2019.

[4] A. F. A. Rahman, S. Sorikhi, and S. Wartulas, "Prediksi Kelulusan Mahasiswa Menggunakan Aalgoritma C4.5 (Studi Kasus Di Universitas Peradaban)," vol. 1, no. 2, 2020.

[5] A. S. Sungae, "Optimasi Algoritma C4.5 Dalam Prediksi Web Phishing Menggunakan Seleksi Fitur Genetic Algoritma," *Paradigma,* vol. XX, no. 2, pp. 27–32, 2018.

[6] E. Elisa, " Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Konstruksi PT.Arupadhatu Adisesanti," *Jurnal Online Informasi*, vol. 2, no. 1, 2017.

[7] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Jurnal Edik Informatika*, vol. 2.i2 (213-219), no. 2, 2016.

[8] S. Z. Harahap, A. Nastuti "Teknik Data Mining Untuk Penentuan Paket Hemat Sembako Dan Kebutuhan Harian Dengan Menggunakan Algoritma FP-GROWTH (Studi Kasus di Ulfamart Lubuk Alung)," *Jurnal Ilmiah Fakultas Sains dan Teknologi*, vol. 7, no. 3, 2019.

[9] F. Elfaladonna and A. Rahmadani, "Analisa Metode Classification-Decission Tree Dan Algoritma C.45 Untuk Memprediksi Penyakit Diabetes Dengan Menggunakan Aplikasi Rapid Miner," *SINTECH (Science and Information Technology) Journal*, vol. 2, no. 1, 2019.

[10] A. Y. Saputra and Y. Primadasa, "Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbour," vol. 17, no. 4, 2018.

[11] S. Bahri and A. Lubis, "Metode Klasifikasi Decision Tree Untuk Memprediksi Juara English Premier League," vol. 2, no. 1, 2020.

[12] P. B. N. Setio, D. R. S. Saputro, and B. Winarno, "Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C4.5," *PRISMA (Prosiding Seminar Nasional Matematika)*, vol. 3, pp. 64–71, 2020.

[13] R. N. Dessy et al, "Kajian Machine Learning Dengan Komparasi Klasifikasi Prediksi Dataset Tenaga Kerja Non-Aktif," vol. 7, no. 1, 2019.

[14] F. Satria, Zamhariri, M. Apun Syaripudin, "Prediksi Ketepatan Waktu Lulus Mahasiswa Menggunakan Algoritma C4.5 Pada Fakultas Dakwah Dan Ilmu Komunikasi UIN Raden Intan Lampung," *Jurnal Ilmiah MATRIK* , vol. 22, no. 1, 2020.

[15] V. R. Prasetyo, H. Lazuardi, A. A. Mulyono, and C. Lauw, "Penerapan Aplikasi RapidMiner Untuk Prediksi Nilai Tukar Rupiah Terhadap US Dollar Dengan Metode Linear Regression," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 7, no. 1, pp. 8–17, 2021.

[16] M. O. Odja, F. J. Likadja, W. T. Ina, and S. I. Pella, "Penggunaan Microsoft Excel untuk Kemudahan Pengolahan Data Nilai Hasil Belajar Siswa," Vol. XV, no. 2, 2021.

[17] A. R. Putri, "Optimalisasi Penggunaan Microsoft Excel Untuk Pengolahan Nilai Raport Di SMAN 1 Ngunut Tulungagung," vol. 3, no. 1, 2015.

[18] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," vol.5, no. 22, 2021.

[19] R. Muliono, J. H. Lubis, and N. Khairina, "Analysis K-Nearest Neighbor Algorithm for Improving Prediction Student Graduation Time," *SinkrOn*, vol. 4, no. 2, p. 42, 2020.

[20] A. Rohman and A. Rufiyanto, "Implementasi Data Mining Dengan Algoritma Decision Tree C4.5 Untuk Prediksi Kelulusan Mahasiswa Di Universitas Pandanaran", vol. 3, 2019.

[21] N. Khasanah, A. Salim, and et al, "Prediksi Kelulusan Mahasiswa Dengan Metode Naive Bayes," vol. 13, 2022.

[22] D. Marlina and M. Bakrie, "Penerapan Data Mining Untuk Memprediksi Transaksi Nasabah Dengan Algoritma C4.5," *Jurnal Teknologi dan Sistem Informasi (JTSI)*, vol. 2, no. 1, 2021.

[23] W. Musu and A. Ibrahim, H. Heriadi "Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5," vol. 10, no.1, 2021.

[24] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data Untuk Klasifikasi Wne Menggunakan Algoritma K-NN," vol. 4, no. 1, 2019.

[25] R. S. Putri and I. Waspada, "Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika," *Khazanah Informatika*, vol. 4, no. 1, 2018.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i3.1731, Copyright ©2023
Submitted : May 18, 2023, Revised : June 9, 2023, Accepted : August 29, 2023, Published : November 4, 2023