Prediksi Kelulusan Mahasiswa Prodi Informatika dengan Algoritma Decision Tree (C4.5) dan Naïve Bayes

Steven Gerrard, Ade Eviyanti*, Hamzah Setiawan, Ika Ratna

Sains dan Teknologi, Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Sidoarjo, Indonesia
Email: [1]ytbreven@gmail.com, [2,*]adeeviyanti@umsida.ac.id, [3]hamzah@umsida.ac.id,[4]ikaratna@umsida.ac.id
Email Penulis Korespondensi: adeeviyanti@umsida.ac.id

Abstrak−Parameter utama untuk mengukur kualitas perguruan tinggi, yang juga berdampak krusial pada proses akreditasi, adalah persentase mahasiswa yang lulus tepat waktu. Kendati demikian, realitas di lapangan menunjukkan bahwa banyak mahasiswa menghadapi kendala untuk menuntaskan masa studinya sesuai batas waktu ideal. Oleh karena itu, dibutuhkan sebuah strategi berbasis data guna memproyeksikan peluang kelulusan mahasiswa sejak dini. Riset ini bertujuan untuk mengomparasikan performa algoritma Decision Tree (C4.5) dan Naïve Bayes dalam melakukan klasifikasi potensi kelulusan secara tepat waktu. Data yang dimanfaatkan mencakup 161 entri mahasiswa Program Studi Informatika angkatan 2022 di Universitas Muhammadiyah Sidoarjo. Atribut yang dianalisis terbagi menjadi faktor akademik dan non-akademik, antara lain jenis kelamin, nilai IPS semester 1 hingga 6, IPK, skor serta status kelulusan PKMU, nilai BQ dan Ibadah, hingga akumulasi poin SKEK. Proses penelitian melewati beberapa fase: prapemrosesan, pelabelan kelas, pembentukan model, dan evaluasi kinerja melalui confusion matrix dan 5-fold cross-validation. Pengujian divalidasi dengan memilah data latih dan data uji pada rasio 70:30, 80:20, serta 90:10. Berdasarkan hasil uji coba, algoritma C4.5 mencetak tingkat akurasi puncak sebesar 100% pada seluruh skenario rasio, dengan rerata akurasi cross-validation mencapai 96,88%. Di sisi lain, Naïve Bayes mencatatkan akurasi maksimal 94,13% dengan rerata cross-validation 93,00%. Temuan ini mengindikasikan bahwa algoritma C4.5 memiliki keunggulan performa pada dataset spesifik ini. Luaran model prediksi ini diharapkan mampu menjadi landasan objektif bagi institusi dalam menetapkan kebijakan akademik yang proaktif.

Kata Kunci: Data Mining; Naive Bayes; C4.5; Akademik; Prediksi

Abstract− The primary parameter for measuring higher education quality, which also has a crucial impact on the accreditation process, is the percentage of students graduating on time. However, the reality on the ground shows that many students face obstacles in completing their studies within the ideal timeframe. Therefore, a data-driven strategy is needed to project students' chances of graduation early. This research aims to compare the performance of the Decision Tree (C4.5) and Naïve Bayes algorithms in classifying the potential for on-time graduation. The data utilized included 161 entries from the Informatics Study Program, class of 2022, at the University of Muhammadiyah Sidoarjo. The attributes analyzed were divided into academic and non-academic factors, including gender, first-semester social studies grades (IPS), GPA, PKMU (Community Service Program) graduation score and status, BQ and Ibadah scores, and accumulated SKEK points. The research process went through several phases: preprocessing, class labeling, model development, and performance evaluation through a confusion matrix and 5-fold cross-validation. The test was validated by separating the training and test data into ratios of 70:30, 80:20, and 90:10. Based on the test results, the C4.5 algorithm achieved a peak accuracy of 100% across all ratio scenarios, with an average cross-validation accuracy of 96.88%. Meanwhile, Naïve Bayes achieved a maximum accuracy of 94.13% with an average cross-validation of 93.00%. These findings indicate that the C4.5 algorithm has superior performance on this specific dataset. The output of this predictive model is expected to serve as an objective basis for institutions in establishing proactive academic policies.

Keywords: Data Mining; Naïve Bayes; C4.5; Akademik; Prediction

1. PENDAHULUAN

Kemajuan kualitas sumber daya manusia yang berdaya saing tinggi amat bergantung pada proses edukasi di jenjang perguruan tinggi. Indikator utama yang merepresentasikan pencapaian suatu institusi pendidikan tinggi dalam mengawal kegiatan akademiknya terlihat dari rasio kelulusan para pesertanya [1]. Terutama berkaitan dengan kapabilitas peserta didik guna menuntaskan masa belajarnya sesuai target waktu. Ketepatan jadwal tamat studi ini telah menjelma menjadi fokus utama bagi pihak kampus. Bukan semata-mata diukur dari segi efektivitas kegiatan belajar. Melainkan turut difungsikan selaku standar ukur untuk evaluasi akreditasi di level program studi maupun universitas.

Regulasi terkait tenggat waktu belajar beserta bobot SKS telah dibakukan di dalam Buku Pedoman Akademik Universitas Muhammadiyah Sidoarjo untuk periode 2023-2024. Panduan ini memaparkan bahwasanya peserta didik di jenjang Strata-1 diharuskan menuntaskan kewajiban akademik minimal sebanyak 144 SKS, yang memakan rentang waktu studi paling cepat 3 tahun hingga maksimal 7 tahun kalender akademik atau setara 14 semester. Realisasi atas kewajiban SKS ini diwujudkan lewat beraneka ragam metode instruksional yang selaras dengan regulasi resmi [2]. Penyelesaian studi yang tidak meleset dari jadwal merupakan sebuah parameter sentral guna menakar mutu pelayanan suatu universitas, terkhusus dalam tahapan akreditasi. Beragam regulasi edukasi turut menyoroti urgensi eksploitasi data kemahasiswaan secara terstruktur demi menyokong perumusan kebijakan yang dilandasi oleh objektivitas dan bukti nyata. Atas dasar tersebut, penelaahan atas profil rekam akademik peserta didik tidak sekadar berguna untuk keperluan tinjauan birokrasi, melainkan juga berperan sebagai jangkar dalam menyusun taktik eskalasi mutu pendidikan [3]. Oleh sebab itu, penjabaran data pendidikan mahasiswa tidak cuma bertugas sebagai penilaian administratif, melainkan turut menjadi fondasi perancangan regulasi strategis pihak kampus demi mendongkrak rasio penyelesaian studi sesuai jadwal.

Di lingkungan Universitas Muhammadiyah Sidoarjo (UMSIDA), sama halnya dengan fenomena di mayoritas kampus lain, penuntasan masa belajar persis pada kerangka waktu ideal (Paling lambat Tujuh Semester bagi Program Studi Sarjana) acap kali menghadirkan rintangan tersendiri. Bermacam-macam elemen bisa membawa dampak pada

durasi tamat belajar seseorang. Entah itu dari kacamata edukasi ataupun di luar akademis. Oleh karena itu, sangat diperlukan suatu mekanisme yang sanggup meramal peluang lulusnya peserta didik secara presisi lewat pemanfaatan histori data kemahasiswaan.

Beriringan dengan laju kemajuan teknologi informasi beserta ranah ilmu data, aneka pendekatan ekstraksi data telah marak diimplementasikan di ekosistem pendidikan, utamanya dalam mengolah rekam akademis demi memproduksi wawasan yang bersifat prediktif [4]. Merujuk pada pemaparan I. Diky Wardhani di dalam bukunya yang bertajuk Pengantar Data Mining [5], diuraikan bahwasanya data mining adalah serangkaian proses penelusuran pola berpadu dengan wawasan krusial dari tumpukan data yang masif. Mekanisme ini merangkum fase penghimpunan informasi, penarikan sampel, manipulasi data, hingga kalkulasi berbasis statistika. Data Mining berwujud sebagai sebuah inovasi teknologi yang dikaruniai kapasitas guna mengontrol data berskala raksasa secara serentak. Pada ranah akademis, praktik penambangan informasi ini acap dijuluki educational data mining (EDM), dengan sasaran utama untuk meninjau historis nilai demi memanen perspektif mutakhir yang berpotensi menyempurnakan skema pembelajaran [6].

Sepasang instrumen klasifikasi yang sangat tenar serta masif dikaryakan mencakup algoritma Decision Tree (C4.5) berdampingan dengan Naïve Bayes. Algoritma Decision Tree (C4.5) beroperasi via penyusunan dahan keputusan yang berlandaskan pada besaran nilai gain dari sebuah variabel [7], sementara itu Naïve Bayes mengelompokkan data bersandar pada tingkat peluang kemunculan tiap-tiap atribut menyasar suatu kelas spesifik [8]. Kedua metode tersebut menyimpan keistimewaan tersendiri manakala memanipulasi masukan data demi kepentingan peramalan. Eksplorasi ini mengusung misi untuk membidani suatu arsitektur yang kapabel meramal rasio probabilitas kelulusan tepat jadwal dari murid UMSIDA dengan mengeksploitasi Algoritma Decision Tree (C4.5) berpadu dengan Naïve Bayes yang memanfaatkan sejumlah fitur semisal Indeks Prestasi Kumulatif (IPK), Indeks Prestasi Semester (IPS) dari 1 sampai VI, skor Baca Tulis Al-Qur'an (BTQ), Pendidikan Karakter Mahasiswa UMSIDA (PKMU), berikut akumulasi poin SKEK. Penentuan variabel-variabel ini dilandasi oleh ketersediaan informasi beserta tingkat keterkaitan data di sepanjang siklus belajar sang mahasiswa[9].

Kajian yang digarap oleh A. Wahyuni dkk. di bawah tajuk "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Decision Tree dan Naive Bayes" [10]. Pada riset tersebut, variabel yang dioperasikan demi keperluan kategorisasi data mining terbangun atas sepuluh elemen, meliputi NIM. Gender, status kemahasiswaan, umur, angka indeks prestasi di semester awal, indeks prestasi semester dua, indeks prestasi semester tiga, indeks prestasi semester empat, indeks prestasi kumulatif, beserta label keterangan kelulusan yang berkedudukan selaku variabel keluaran. Keseluruhan atribut yang diterjunkan ke proses pemilahan dalam data mining tersebut memang merangkum 10 komponen, yakni NIM, Jenis Kelamin, keaktifan mahasiswa, rentang usia, Indeks Prestasi semester 1, Indeks Prestasi semester 2, Indeks Prestasi semester 3, Indeks Prestasi semester 4, total IPK, ditambah status keterangan selaku luaran targetnya. Mengacu pada buah pengujian yang diuji coba via peranti lunak RapidMiner yang melibatkan kedua skema tersebut, instrumen pohon keputusan (C4.5) sukses merengkuh angka akurasi puncak di kisaran 88,92%, berbanding terbalik dengan Metode Naïve Bayes yang hanya sanggup mengamankan skor akurasi teratas di angka 84,98%.

Observasi seputar estimasi tamat studi mahasiswa secara disiplin waktu sejatinya sudah teramat sering diselenggarakan mengadopsi taktik penambangan data. Diska bersama Budayawan mendemonstrasikan implementasi instrumen *Naïve Bayes Classifier* guna menerka jadwal tuntas studi dari peserta didik Program Studi Pendidikan Teknik Informatika, yang mana melibatkan materi latih berjumlah 94 sampel berpadu dengan 46 sampel uji. "Di dalam aplikasi tersebut, klasifikasi *naïve bayes classifier* difungsikan guna memilah data bersandar pada hasil nilai selama 6 semester berjalan ditambah total sks. Luaran pemilahan dari aplikasi ini memvalidasi bahwasanya berbekal 46 data *testing* sanggup dikumpulkan capaian *accuracy* 82,61%, *precision* 91,66%, *recall* 61,11%." [11] Di samping pencapaian tersebut, penaksiran via kurva ROC menelurkan skor AUC senilai 0,915, yang mana diklasifikasikan oleh para perisetnya ke dalam taraf istimewa (*excellent*), alhasil arsitektur itu dianggap amat pantas dikaryakan demi keperluan menerka kelulusan disiplin waktu.

Karya ilmiah yang diramu oleh Kassymova dkk. merakit sebuah model demi menaksir ketepatan tamat belajar para murid calon pendidik di lingkungan IAIN Bone, yang mana mengaplikasikan ragam algoritma *data mining* semisal Decision Tree C4.5, Naïve Bayes, dipadukan dengan K-Nearest Neighbor (KNN). Kajian tersebut menyandarkan operasionalnya pada sejumlah fitur akademis berbaur dengan profil demografis meliputi jenis kelamin, umur, rentetan rekam akademis sejak semester awal sampai semester 4, didukung oleh IPK yang didapuk selaku parameter yang digunakan dalam fase klasifikasinya. Keluaran riset tersebut memaparkan keunggulan algoritma Decision Tree C4.5 yang menyumbangkan angka akurasi terdongkrak maksimal di posisi 93,90%, dibuntuti oleh KNN pada angka 92,07%, dan diakhiri oleh Naïve Bayes di kisaran 90,24%. Lebih lanjut lagi, buah dari pengetesan berbasis statistik membuktikan ketiadaan selisih yang terlalu mencolok di antara kemampuan akurasi ketiga instrumen itu saat memprediksi kelulusan sang murid. Penemuan ini menjadi indikator kuat bahwa intervensi *data mining* sanggup diberdayakan secara efektif guna menopang pihak lembaga pendidikan ketika memprediksi kelulusan peserta didik melalui pendekatan yang jauh lebih sistematis dan sarat akan bukti data [12].

Riset berbeda yang cukup relevan turut diprakarsai oleh Imam Riadi dkk, di mana mereka mempekerjakan metode Naïve Bayes guna menerka jadwal tamat ideal lewat serangkaian fase *data mining* yang komprehensif, bermula dari pemuatan data hingga ke jenjang pengujian model. "Fase-fase krusial di dalam kajian itu merangkum prosedur muat data, pembersihan data, pemilihan data, transformasi data, data pelatihan, pengujian data, serta hasil prediksi." Bersandar pada rekor pembuktian akurasi yang dieksekusi via *confusion matrix*, skor presisi yang berhasil dicapai menembus margin 72% dari keseluruhan 291 data (terdiri atas 273 data latih berdampingan dengan 18 data uji), alhasil sang penulis

memastikan bahwasanya mesin tebak ini pantas untuk dijadikan patokan oleh pihak fakultas manakala merumuskan keputusan yang berkaitan dengan status kelulusan [13].

Tak hanya Naïve Bayes, algoritma Decision Tree pun tak kalah laris dieksploitasi ketika memproyeksikan status kelulusan di kalangan mahasiswa. Eksperimen yang digawangi oleh Satrio Junaidi et al. mengeksekusi penelaahan komparasi menyasar empat metode pengelompokan (yakni Naïve Bayes, Random Forest, SVM, beserta ANN) demi menebak rasio tamat studi disiplin waktu dengan bermodalkan himpunan data berisi 101 mahasiswa dari Fakultas Sains dan Teknologi. Spesifik untuk unjuk kerja Naïve Bayes, tahap evaluasi memaparkan angka akurasi di level 0,87 (yang tertulis selaku *Accuracy: 0.8709677419354839*), di mana instrumen *confusion matrix* melahirkan rasio 25 tebakan Tepat Waktu yang valid, 4 tebakan Tepat Waktu yang melenceng, berikut 2 buah prediksi yang menyasar kelas Terlambat; Parameter ukuran lainnya mengabarkan bahwa *precision* untuk kategori Tepat Waktu mendarat sempurna di angka 1,00, *recall* di 0,86, disusul oleh metrik *precision* di kategori Terlambat yang menyentuh 0,33, berikut *recall* 1,00. Penemuan ini mengilustrasikan daya keandalan Naïve Bayes yang terpantau menonjol pada segelintir metrik biarpun ritme performa di tiap kelasnya berfluktuasi, oleh karenanya barisan penulis amat menyarankan adanya uji silang lintas-algoritma guna menunjuk mana model yang paling ideal selaras dengan tujuan institusi [14].

Di pihak berseberangan, terdapat kajian yang berjudul "Perbandingan algoritma c4.5 dan naive bayes dalam prediksi kelulusan mahasiswa", di mana material yang dieksploitasi meliputi histori mahasiswa purna studi pada tahun 2022 untuk jenjang S-1, yang diintegrasikan bersama rekam jejak indeks prestasi semester dari peranti lunak sistem informasi akademik. Pendekatan kalkulatif yang diaplikasikan pada riset tersebut mencakup metode C4.5 bersanding dengan Naïve Bayes, lantas pembuktian ini memberi sinyal bahwa Naïve Bayes memancarkan dominasi yang lebih terang ketimbang instrumen C4.5, hal tersebut terkonfirmasi dari perolehan skor akurasi yang lebih tinggi jika disandingkan dengan C4.5 saat diadu di lintasan pengujian 70%:30% dan 80%:20% yang mencetak skor 79,82%, 81,58%, lantas tatkala menyentuh proporsi data 90%:10% tingkat akurasi algoritma Naïve Bayes bertengger stabil sedangkan performa C4.5 justru mengalami penurunan [8].

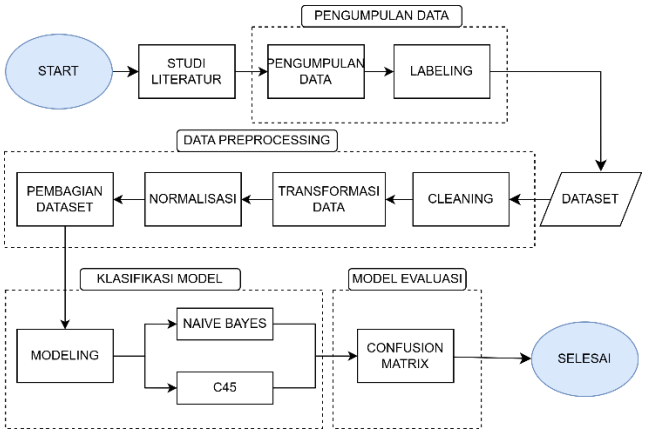
Walaupun sederetan riset di masa lampau telah memvalidasi fakta bahwasanya metode Naïve Bayes maupun Decision Tree mumpuni dikaryakan demi memprediksi lulusnya mahasiswa secara ideal, realitasnya dominasi kajian-kajian itu masih terfokus menyasar metode tunggal ataupun belum mengadu secara head-to-head kapabilitas dari kedua instrumen itu di atas bongkahan dataset yang sama. Atas landasan itulah, studi ini mengusung niat sentral guna membedah adu tangkas dari sisi performa algoritma Naïve Bayes dengan Decision Tree (C4.5) tatkala memproyeksikan keberhasilan kaum intelektual muda demi lulus patuh jadwal berbekal pendayagunaan data akademik kepunyaan mahasiswa, dengan begitu sanggup dicetak suatu sistem prediksi yang lebih mantap dan relevan demi menyokong rumusan keputusan di perguruan tinggi.

Melalui riset ini. Diharapkan akan didapatkan sebuah kerangka penaksir yang piawai mensuplai informasi awal kepada pengelola program studi maupun universitas supaya mengambil taktik preventif berpadu dengan kebijakan strategis menyasar mahasiswa yang diramalkan tidak sanggup menuntaskan masa belajarnya selaras dengan waktu, adapun buah dari kajian tersebut juga bisa ditunjuk selaku fondasi sewaktu membuat sistem informasi akademik bersenjata kecerdasan artifisial yang bertugas membantu skema pengambilan keputusan yang makin berkualitas di lanskap pendidikan tingkat tinggi.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Gambar 1 di bawah ini mengilustrasikan tahapan penelitian yang mencakup serangkaian proses pelaksanaan secara runtut.



Gambar 1. Tahapan Penelitian

Pendekatan yang diimplementasikan pada studi ini mencakup algoritma klasifikasi Naïve Bayes beserta C4.5, dengan rincian alur untuk masing-masing fase adalah sebagai berikut :

a. Studi Literatur.

Tinjauan pustaka dilaksanakan guna mendalami prinsip-prinsip fundamental dari algoritma Decision Tree (C4.5) maupun Naïve Bayes [15], beserta implementasinya di sektor pendidikan. Referensi yang ditelaah meliputi artikel jurnal, buku panduan, serta terbitan elektronik yang mengupas mengenai teknik klasifikasi, prapemrosesan data, hingga proses evaluasi model dengan memanfaatkan Confusion Matrix. Pemahaman mengenai konsep dasar algoritma Decision Tree (C4.5) dan Naïve Bayes [15], serta aplikasinya pada ranah pendidikan kembali ditekankan. Bahan bacaan yang dieksplorasi melibatkan jurnal sains, buku literatur, dan dokumen digital yang menjelaskan metode klasifikasi, tahapan preprocessing, serta penilaian model berbasis Confusion Matrix

Sebuah kajian yang membahas estimasi tingkat partisipasi pemilu dikaitkan dengan mutu pendidikan [16] memaparkan bahwa penggunaan metode SVM melalui kernel linear, RBF, serta polinomial terhadap dataset pendidikan mampu memberikan tingkat akurasi yang memuaskan (88,4-88,5%), sehingga algoritma SVM patut dipertimbangkan untuk keperluan prediksi yang bertumpu pada parameter pendidikan.

1. Objek Penelitian

   Objek penelitian pada studi ini adalah data akademik mahasiswa Program Studi Informatika Universitas Muhammadiyah Sidoarjo angkatan 2022. Dataset diperoleh dari Direktorat Akademik, Direktorat Kemahasiswaan dan Alumni, serta Lembaga Al-Islam dan Kemuhammadiyahan. Data yang digunakan berjumlah 161 mahasiswa dengan atribut akademik dan non-akademik yang meliputi jenis kelamin, indeks prestasi semester (IPS) semester 1 hingga semester 6, indeks prestasi kumulatif (IPK), nilai PKMU, status PKMU, nilai BQ, nilai ibadah, status BQ dan ibadah, serta total poin SKEK. Data tersebut digunakan sebagai variabel prediktor untuk menentukan status kelulusan mahasiswa tepat waktu atau tidak tepat waktu..

2. Naive Bayes

   Algoritma Naïve Bayes merupakan algoritma klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes dengan asumsi independensi antar fitur. Algoritma ini menghitung probabilitas posterior suatu kelas berdasarkan probabilitas prior dan likelihood dari setiap atribut yang diamati. Secara matematis, klasifikasi pada Naïve Bayes dapat dirumuskan sebagai berikut:.

$$P(C \mid X) = \frac{P(X \mid C) \, . P \, (C)}{P(X)} \tag{1}$$

   Naïve Bayes bekerja dengan asumsi independensi kondisional antar fitur sehingga estimasi densitas kondisional dapat dipisah per-fitur, sehingga pelatihan hanya perlu mengestimasi parameter univariat untuk setiap fitur dan kelas (mis. Bernoulli untuk fitur biner, multinomial untuk frekuensi kata, Gaussian untuk kontinu); posterior kelas dihitung dari kombinasi prior dan probabilitas kondisional, dan untuk stabilitas numerik praktik umum meliputi perhitungan dalam domain log serta penggunaan smoothing (Laplace/Dirichlet) untuk mengatasi zero counts. Pendekatan ini sangat efisien pada data berdimensi tinggi dan sparse seperti representasi bag-of-words, namun kalibrasi probabilitasnya bisa kurang akurat jika fitur saling berkorelasi kuat, sehingga sering disertai pemilihan atau transformasi fitur untuk meningkatkan performa[8].

3. Decision Tree C4.5

   Algoritma Decision tree c4.5 adalah cara untuk mengklasifikasikan data dengan membuat struktur pohon berdasarkan pilihan atribut yang paling informatif. Algoritma C4.5 memanfaatkan Rasio Gain untuk menemukan atribut yang paling baik, sehingga bisa mengurangi kecenderungan pada atribut yang memiliki banyak nilai dan menghasilkan model yang mudah dipahami dalam format aturan if-then. [17]. Entropi mengukur ketidakpastian sebuah himpunan data, dengan rumus sebagai berikut:

$$Entropy(S) = \sum_{i} P_i Log_2 P_i \tag{2}$$

   Sedangkan information gain mengukur pengurangan entropy setelah membagi himpunan data berdasarkan atribut A. Dengan rumus sebagai berikut :

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{3}$$

   C4.5 memilih atribut dengan nilai Gain Ratio tertinggi karena atribut tersebut paling efektif mengurangi ketidakpastian kelas setelah memperhitungkan distribusi pembagian. Entropy dan Gain menunjukkan seberapa baik atribut memisahkan kelas, sedangkan Gain Ratio menormalkan Gain untuk menghindari preferensi terhadap atribut dengan banyak nilai. [17]

b. Pengumpulan data.

Dalam fase ini, kumpulan data dihimpun bersumber dari Direktorat Akademik, Direktorat Kemahasiswaan dan Alumni, beserta Lembaga Al-Islam dan Kemuhammadiyahan lewat prosedur permohonan surat izin riset kepada tiap instansi yang bersangkutan. Berkas informasi yang didapatkan berformat ekstensi CSV, kemudian seusai melewati tahapan integrasi dari beragam referensi data tersebut, terkumpul sebanyak 161 entri riwayat mahasiswa. Kendati demikian, pada masa penghimpunan ini kondisi data masih tergolong kotor atau bercampur aduk, sehingga menuntut

adanya langkah penanganan lebih lanjut agar bisa dieksploitasi dengan maksimal. Materi data yang dieksekusi pada studi ini berasal dari rekam jejak mahasiswa Program Studi Informatika untuk tahun masuk 2022, yang merangkum keterangan spesifik perihal akademik serta non-akademik terkait perjalanan aktivitas studi mereka. Guna menyajikan ilustrasi terkait komposisi draf data di awal perolehannya, sampel dataset orisinal yang difungsikan pada kajian ini dicantumkan dalam Tabel 1

Tabel 1. Dataset Mentah

NO	0	-	160
NIM	221080200006	-	221080200098
Jenis Kelamin	L	-	Tidak Ditemukan
IPS1	3.47	-	Tidak Ditemukan
IPS2	3.58	-	Tidak Ditemukan
IPS3	3.89	-	Tidak Ditemukan
IPS4	3.92	-	Tidak Ditemukan
IPS5	3.81	-	Tidak Ditemukan
IPS6	3.27	-	Tidak Ditemukan
IPK	3.69	-	Tidak Ditemukan
NILAI_PKMU	A	-	NaN
STATUS_PKMU	LULUS	-	NaN
BQ	325.0	-	NaN
IBADAH	4.0	-	NaN
STATUS_BQ_IBADAH	LULUS	-	NaN
TOTAL_POINT	95	-	0

Merujuk pada Tabel 1, susunan data mentah ini terbukti masih menyimpan sejumlah anomali berupa isian yang rumpang atau tidak terlacak, semisal kemunculan label NaN serta kekosongan nilai pada berbagai fiturnya. Di samping itu, ditemukan pula ketidaksamaan format penyajian data antara tipe numerik dan jenis kategorikal akibat perbedaan titik asal pengambilannya. Situasi ini mengisyaratkan bahwa data yang sudah terkumpul belum layak untuk langsung diaplikasikan pada fase pemodelan; oleh sebab itu, dibutuhkan sebuah tahapan prakondisi data yang meliputi proses pembersihan, sinkronisasi, hingga modifikasi struktur agar dataset tersebut tampil lebih sistematis serta ideal untuk kebutuhan analisis dan pembentukan model klasifikasi.

c. Labeling.

Proses Pelabelan dilakukan dengan menetapkan kelas target berpa stattus kelulusan mahasiswa, yaitu " LULUS TEPAT WAKTU" dan "TIDAK LULUS TEPAT WAKTU" penentuan label didasarkan pada dua kategori data , yaitu:

1. Data Akademik.
   Riwayat akademik Mahasiswa yang meliputi IPS (indeks prestasi semester) semester 1 hingga semester 6 serta ipk. Berdasarkan Pedoman Akademik [2], apabila IPS kurang dari 2.00 maka jumlah maksimum SKS yang dapat diambil adalah 12 SKS. Selain itu, IPK dipengaruhii oleh total SKS yang ditempuh. Mahasiswa dinyatakan lulus dengan predikat memuaskan apabila indeks prestasi kumulatif (IPK) sebesar 2,76 hingga 3,00, Oleh karena itu, penulis menetapkan nilai minimum IPK sebesar 2,76 sebagai acuan dalam penentuan label akhir.
2. Data Non-Akademik.
   Riwayat non-akademik mahasiswa, khususnya status kelulusan PKMU dan Total poin. Berdasarkan Pedoman Akademik [2], salah satu persyaratan ujian proposal adalah penyelesaian tugas akhir, skripsi, atau tesis. Untuk program S1 atau D-IV, skor Skek Minimal yang harus dicapai adalah 185 angka kredit. Dengan demikian, total poin Skek sebesar 120 ditetapkan sebagai batasan label akhir untuk data non akademik

d. Data Preprocessing.

Fase prapemrosesan data memegang peranan krusial dalam kajian *data mining*, di mana tujuannya adalah guna mengondisikan kumpulan data agar mumpuni untuk diterjunkan ke dalam tahapan pemodelan. Dalam tahapan ini, terselenggara serangkaian mekanisme manipulasi data demi mendongkrak mutu informasinya, sehingga mampu melahirkan sebuah sistem prediksi yang lebih presisi dan dapat dipercaya keandalannya. Mekanisme *preprocessing* yang dipraktikkan pada riset ini mengusung kegiatan pembersihan data (*cleaning*), konversi format, serta penormalan alias penyandian (*encoding*) data.

1. Cleaning
   Dalam sub-tahap ini, dieksekusi suatu prosedur penyaringan data guna menjamin bahwa dataset yang dikaryakan telah bersih dari segala jenis kekeliruan maupun ketidakselarasan informasi. Operasi cleaning ini mencakup tindakan eliminasi pada entri yang ganda, penyelesaian terhadap sel yang tak terisi (missing value), berbarengan dengan pembetulan atas format data yang berantakan. Langkah ini sangat vital untuk menggaransi bahwa material data yang diumpankan ke fase pembentukan model membawa kualitas yang prima sekaligus tidak memunculkan bias interpretasi pada kesimpulan analisisnya.
2. Transformasi Data

Konversi data dijalankan guna menyelaraskan tatanan formatnya sehingga kompatibel untuk dikalkulasi oleh metode klasifikasi yang diterapkan. Melalui fase ini, sejumlah variabel yang semula mengusung tipe susunan yang berlainan diseragamkan sedemikian rupa agar membentuk satu struktur yang padu. Proses pengubahan ini pun dilaksanakan demi meyakinkan bahwa masing-masing fitur telah memegang karakteristik tipe data yang linier dengan tuntutan kebutuhan analisisnya.

3. Normalisasi
   Pada tahapan ini direalisasikan sebuah mekanisme encoding, yakni mengonversi data-data yang berwujud kategorikal ke dalam rupa numerik agar bisa dicerna oleh algoritma machine learning. Pendekatan ini bernilai urgensi tinggi mengingat sebagian besar metode klasifikasi hanya dibekali kapabilitas untuk mengoperasikan masukan yang berwujud bilangan. Teknik penerjemahan kode yang diaplikasikan pada kajian ini direpresentasikan secara jelas pada Tabel 2.

Tabel 2. Encoding

Data Asli		Encoding
Jenis Kelamin	L	0
	P	1
status_pkmu	LULUS	0
	TIDAK LULUS	1
status_bq_ibadah	LULUS	0
	TIDAK LULUS	1
status_final	LULUS TEPAT WAKTU	0
	TIDAK LULUS TEPAT WAKTU	1

Merujuk visualisasi Tabel 2, masing-masing fitur berskala kategori dialihwujudkan ke bentuk nilai matematis dengan menerapkan taktik label encoding. Prosedur ini diupayakan guna meringankan beban komputasi algoritma saat menjalankan operasi matematis di sepanjang periode pelatihan modelnya. Berkat dilaksanakannya rangkaian prapemrosesan ini, kumpulan materi yang mulanya berwujud mentah berubah menjadi jauh lebih tertata dan matang untuk disuplai ke jenjang pemodelan dengan memanfaatkan algoritma Decision Tree (C4.5) bersama Naïve Bayes.

4. Pembagian Dataset
   Kumpulan entri dipisahkan ke dalam dua segmen yakni data pelatihan beserta data pengujian melalui rasio pembagian spesifik; di mana riset ini mengimplementasikan komparasi rasio 70:30, 80:20, serta 90:10 demi keperluan melatih dan mengetes keandalan model.

e. Klasifikasi Model.
   Model Klasifikasi dibangun menggunakan dua algoritma sebagai berikut :
   1. Naive bayes.
      Metode algoritma tersebut mengadopsi prinsip probabilistik yang bersumber dari Teorema Bayes [17] , dengan memegang dalil bahwa tiap-tiap atribut berkedudukan secara independen. Algoritma ini dinilai sangat ideal bagi tipe data berspesifikasi sebaran probabilitas yang teratur, [18] penelitian tersebut turut menjabarkan langkah-langkah kalkulasi Naïve Bayes (mencakup prior, likelihood, hingga posterior) berbarengan dengan metode perhitungan konvensional via Excel untuk dijadikan materi komparasi terhadap besaran akurasi yang dicetak oleh perangkat lunak RapidMiner. Kajian tersebut memegang peranan krusial selaku referensi operasional guna mendeskripsikan proses implementasi beserta rumusan probabilistik di dalam suatu rancangan model klasifikasi.
   2. C4.5
      Pendekatan Decision Tree menyusun cabang-cabang keputusannya dengan bersandar pada parameter nilai gain ratio [9]. Keunggulan C4.5 terletak pada kemampuannya untuk mengelola masukan bertipe kategori maupun angka sekaligus menoleransi kemunculan data yang bernilai hampa.

f. Model Evaluasi.
   Model evaluasi dibangun dengan tujuan untuk mengidentifikasi performa model klasifikasi menggunakan confusion matrix dan Cross validation,
   1. Confusion Matrix
      Instrumen Confusion Matrix merupakan salah satu teknik guna menguji tingkat kesuksesan operasional model klasifikasi dengan cara menyandingkan keluaran prediksi model terhadap label aktualnya secara presisi [19]. Pendekatan ini menyodorkan penjabaran yang lebih mendetail terkait unjuk kerja suatu model lewat presentasi nominal besaran True Positive (TP), True Negative (TN), False Positive (FP), hingga False Negative (FN) yang dicetak oleh algoritma pengelompokan bersangkutan. Berbekal analisis Confusion Matrix ini, pengkaji sanggup mendeteksi tingkatan presisi berbarengan dengan anomali klasifikasi yang menimpa model, sehingga metrik tersebut berkapasitas untuk difungsikan selaku landasan baku dalam menilai mutu dari sistem yang telah dibentuk.
   2. Cross Validation
      Guna mengamankan kepastian bahwasanya model klasifikasi yang dibangun mengantongi kapabilitas generalisasi yang mumpuni, para praktisi data mining lazimnya mempekerjakan taktik cross validation.

Mekanisme kerja metode ini bertumpu pada pemecahan kumpulan data ke dalam beraneka porsi (fold), yang mana setiap irisannya bakal dirotasi untuk berperan secara bergilir sebagai material uji sekaligus materi latih. Strategi pemecahan ini dinilai manjur dalam menekan angka bias saat berlangsungnya proses penaksiran model, seraya menyuguhkan perkiraan efektivitas yang jauh lebih stabil ketimbang sekadar bertumpu pada satu pola pembagian statis antara data latih dan uji tunggal [20].

3. HASIL DAN PEMBAHASAN

Data penelitian ini berasal dari 161 mahasiswa Program Studi Informatika, Fakultas Sains dan Teknologi, Angkatan 2022. Data yang digunakan masih berupa data mentah sebelum melalui proses preprocessing. Tahapan preprocessing terdiri atas empat langkah, yaitu cleaning, transformasi data, normalisasi, dan pembagian dataset, Pada tahap cleaning, atribut yang tidak relevan dihapus sehingga tersisa atribut: Jenis Kelamin, Indeks Prestasi Semester (IPS) 1–6, nilai_pkmu, status_pkmu, bq, ibadah, status_bq_ibadah, Total Point, serta status_final. Kemudian Tahap transformasi dilakukan dengan mengubah beberapa atribut bertipe alfanumerik menjadi numerik agar sistem dapat lebih konsisten dalam melakukan prediksi. Tahapan selanjutnya Tahap normalisasi mencakup penyesuaian data, misalnya nilai IPS dalam bentuk desimal, atribut bq dalam bentuk numerik, serta Total Point yang juga dinyatakan dalam numerik dan Tahap terakhir adalah pembagian dataset menjadi data latih dan data uji dengan rasio 90:10, 80:20, dan 70:30. Secara keseluruhan, jumlah data yang diolah dalam penelitian ini adalah 161 mahasiswa.

Tabel 3. Dataset

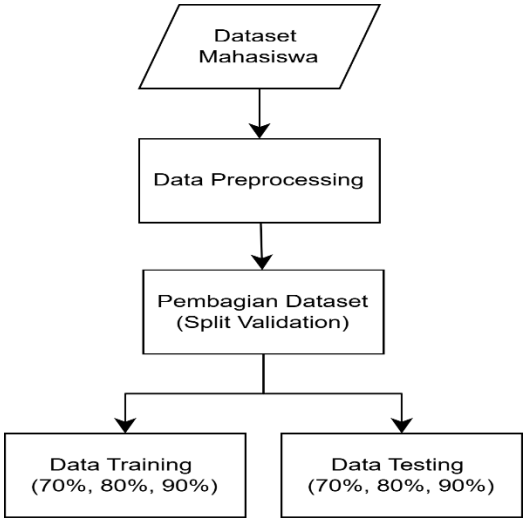
NO	0	1	-	159	160
Jenis Kelamin	0	0	-	0	0
IPS1	3.47	3.67	-	3.58	3.13
IPS2	3.58	3.8	-	3.5	2.12
IPS3	3.89	3.72	-	3.59	0.0
IPS4	3.92	3.06	-	3.56	0.0
IPS5	3.81	3.36	-	3.51	0.0
IPS6	3.27	3.6	-	3.33	0.0
IPK	3.69	3.52	-	3.51	0.0
NILAI_PKMU	4.0	4.0	-	4.0	0.0
STATUS_PKMU	0	0	-	0	1
BQ	325	3	-	3	0
IBADAH	4	3	-	3	0
STATUS_BQ_IBADAH	0	0	-	0	1
TOTAL_POINT	95	130	-	190	10
STATUS_FINAL	1	0	-	0	1

Mengacu pada Tabel 3, dataset penelitian ini mencakup atribut akademik dan non-akademik yang digunakan untuk memprediksi status kelulusan mahasiswa Program Studi Informatika. Atribut Jenis Kelamin direpresentasikan secara numerik, dengan kode 0 untuk Laki-laki dan 1 untuk Perempuan. Sementara itu, atribut IPS1 hingga IPS6 beserta IPK menggambarkan capaian akademik mahasiswa, baik per semester maupun secara kumulatif. Adapun atribut non-akademik meliputi nilai_pkmu, status_pkmu, BQ, IBADAH, status_bq_ibadah, Total_Point, serta status_final.

3.1 Analisa Data menggunakan Algoritma Decision Tree C4.5

Algoritma Decision Tree C4.5 telah banyak digunakan dalam berbagai penelitian klasifikasi dan menunjukkan performa yang baik, termasuk pada penelitian yang dilakukan oleh Arif Senja Fitrani dkk [21]. Dalam memprediksi partisipasi pemilih pada pemilu dengan tingkat akurasi yang tinggi. Oleh karena itu, pada penelitian ini algoritma C4.5 digunakan untuk menganalisis dan memprediksi kelulusan mahasiswa berdasarkan atribut akademik dan non-akademik yang tersedia.

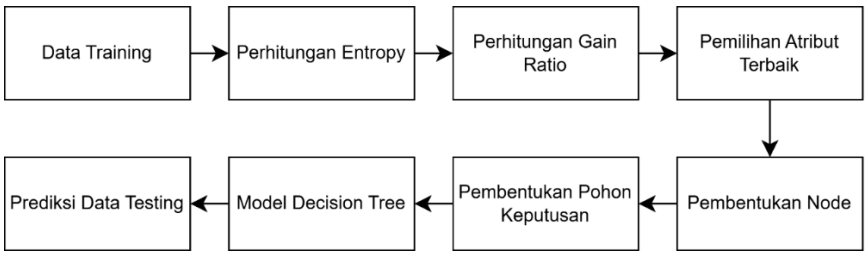
Setelah melalui tahap preprocessing, data kemudian dianalisis menggunakan algoritma C4.5 dengan bantuan platform Google Colabs. Model C4.5 diuji menggunakan metode split validation, yaitu teknik yang membagi data menjadi dua bagian: data latih (training) dan data uji (testing).



Gambar 2. Splitting data

Proses pembagian dataset yang tertera pada Gambar 2 dilakukan menggunakan metode split validation, di mana dataset dibagi menjadi data latih dan data uji dengan beberapa rasio pembagian yaitu 70:30, 80:20, dan 90:10. Data latih digunakan untuk membangun model klasifikasi, sedangkan data uji digunakan untuk mengevaluasi performa model yang dihasilkan.

Setelah proses pembagian dataset menjadi data latih dan data uji dilakukan pada tahap sebelumnya, langkah berikutnya adalah melakukan proses pemodelan untuk membangun model prediksi kelulusan mahasiswa menggunakan algoritma yang telah ditentukan. Proses pemodelan ini bertujuan untuk membentuk model klasifikasi yang mampu mengidentifikasi pola dari data yang digunakan dalam penelitian.



Gambar 3. Proses Modeling C4.5

Proses pembangunan model menggunakan Algoritma Decision Tree C4.5 dalam memprediksi kelulusan mahasiswa ditunjukan pada Gambar 3. Pada tahap ini, data latih digunakan untuk membentuk model klasifikasi berdasarkan atribut akademik dan non-akademik yang tersedia. Pada algoritma C4.5, proses pemodelan dilakukan dengan menghitung nilai entropy dan gain ratio untuk menentukan atribut terbaik sebagai node pada pohon keputusan. Proses ini dilakukan secara berulang hingga terbentuk struktur decision tree yang dapat digunakan untuk mengklasifikasikan data. Untuk memperoleh hasil yang optimal, pengujian model dilakukan dengan beberapa skenario pembagian dataset, yaitu 70:30, 80:20, dan 90:10 antara data latih dan data uji. Kinerja model kemudian dievaluasi menggunakan confusion matrix untuk mengukur tingkat akurasi prediksi yang dihasilkan oleh algoritma C4.5. Hasil evaluasi pada pembagian data dengan rasio 70:30 disajikan pada Tabel berikutnya.

Tabel 4. Confusion Matrix C4.5 70:30

	True Terlambat	True Tepat
False Terlambat	29	0
False Tepat	0	20

Berdasarkan Tabel 4, confusion matrix menampilkan hasil evaluasi model C4.5 pada data uji dengan rasio 30%. Model mampu memprediksi seluruh data secara tepat, dengan 29 mahasiswa terklasifikasi sebagai terlambat dan 20 mahasiswa terklasifikasi sebagai tepat waktu. Tidak terdapat kesalahan dalam proses klasifikasi, sehingga tingkat akurasi mencapai 100%. Hasil ini menunjukkan efektivitas model yang sangat tinggi, meskipun tetap diperlukan pengujian pada dataset yang lebih besar untuk memastikan kemampuan generalisasi. Adapun hasil akurasi pada rasio pembagian data 80:20 dan 90:10 disajikan pada Tabel 8.

Selain menggunakan pembagian data latih dan uji, evaluasi dilakukan dengan 5-Fold Cross Validation. Teknik ini membagi dataset menjadi lima bagian, di mana setiap bagian secara bergantian digunakan sebagai data uji, sementara sisanya digunakan sebagai data latih. Hasil akurasi tiap fold ditunjukkan pada Tabel berikut.

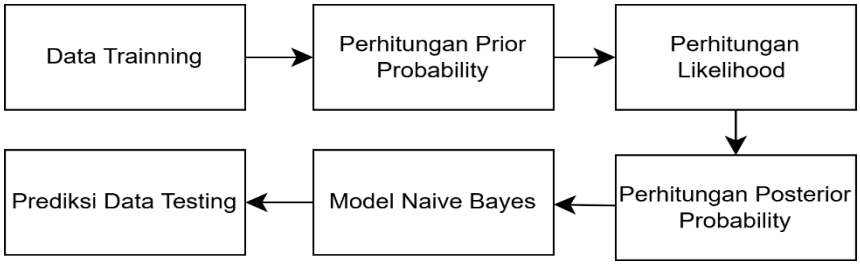
Tabel 5. Cross Validation C4.5

Fold	Akurasi
Fold 1	100.00%
Fold 2	96.88%
Fold 3	96.88%
Fold 4	90.62%
Fold 5	96.88%
Rata-rata	96.88%

Berdasarkan Tabel 5, model C4.5 menunjukkan performa yang sangat baik dan konsisten pada setiap fold. Akurasi tertinggi dicapai pada Fold 1 dengan nilai 100%, sedangkan akurasi terendah terjadi pada Fold 4 dengan nilai 90.62%. Rata-rata akurasi keseluruhan sebesar 96.88%, yang menunjukkan bahwa model memiliki kemampuan generalisasi yang kuat dan tidak mengalami overfitting secara signifikan.

3.2 Analisa Data menggunakan Naïve Bayes

Setelah proses pemodelan menggunakan algoritma Decision Tree C4.5 dilakukan, tahap selanjutnya adalah membangun model prediksi menggunakan algoritma pembanding, yaitu Naïve Bayes. Algoritma ini digunakan untuk mengklasifikasikan data kelulusan mahasiswa berdasarkan probabilitas dari setiap atribut yang digunakan dalam penelitian.



Gambar 4. Proses Modeling Naïve Bayes

Proses pembangunan model menggunakan Naive Bayes dalam memprediksi kelulusan mahasiswa ditunjukan pada Gambar 4. Pada tahap ini, data latih digunakan untuk menghitung probabilitas masing-masing atribut terhadap kelas target, yaitu status kelulusan mahasiswa. Algoritma Naïve Bayes bekerja berdasarkan teorema Bayes dengan asumsi bahwa setiap atribut bersifat independen satu sama lain. Model kemudian menghitung probabilitas posterior dari setiap kelas berdasarkan atribut yang dimiliki oleh data uji, sehingga dapat menentukan kelas dengan probabilitas tertinggi sebagai hasil prediksi. Untuk memastikan kinerja model yang optimal, pengujian dilakukan dengan beberapa skenario pembagian dataset, yaitu 70:30, 80:20, dan 90:10 antara data latih dan data uji. Evaluasi kinerja model dilakukan menggunakan confusion matrix untuk mengetahui tingkat akurasi prediksi yang dihasilkan oleh algoritma Naïve Bayes. Hasil pengujian pada pembagian data dengan rasio 70:30 disajikan pada Tabel berikutnya.

Tabel 6. Confusion Matrix Naïve Bayes 70:30

	True Terlambat	True Tepat
False Terlambat	29	0
False Tepat	3	17

Berdasarkan Tabel 6, hasil confusion matrix menunjukkan bahwa jumlah mahasiswa yang diprediksi terlambat dengan benar adalah 29, tanpa adanya kesalahan klasifikasi. Sementara itu, pada kategori tepat waktu, terdapat 17 prediksi yang benar dan 3 prediksi yang keliru. Untuk menghitung akurasi, digunakan perbandingan antara jumlah prediksi yang benar dengan total data yang diuji[22]. Rumus akurasi dapat dituliskan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TN + TP + FP + FN} \tag{4}$$

$$Precision = \frac{TP + TN}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Hasil pengujian menunjukkan bahwa algoritma Naive Bayes menghasilkan tingkat akurasi sebesar 93,87%, dengan tingkat kesalahan (error) sebesar 6,13%. Selain menggunakan pembagian data latih 70% dan data uji 30%,

pengujian akurasi juga dilakukan dengan rasio 80:20 dan 90:10. Rincian hasil akurasi dari kedua rasio tersebut dapat dilihat pada Tabel 8. Kemudian hasil juga diuji dengan cross validation sebagaimana yang diuji pada metode algoritma decision tree(C4.5) dengan hasil sebagai berikut;

Tabel 7. Cross Validation Naïve Bayes

Fold	Akurasi
Fold 1	94.00%
Fold 2	91.00%
Fold 3	92.00%
Fold 4	95.00%
Fold 5	93.00%
Rata-rata	93.00%

Berdasarkan hasil pada Tabel 7 algoritma Naive Bayes menunjukkan akurasi yang cukup tinggi dengan rata-rata sebesar 93.00%. Meskipun terdapat variasi antar fold, nilai akurasi tetap berada di atas 90%, yang menandakan bahwa model memiliki kemampuan generalisasi yang baik. Namun, dibandingkan dengan algoritma C4.5, performa Naive Bayes masih sedikit lebih rendah, terutama karena adanya beberapa kesalahan klasifikasi pada data uji. Hal ini menunjukkan bahwa Naive Bayes cukup efektif dalam memprediksi ketepatan waktu kelulusan mahasiswa, tetapi kurang konsisten dibandingkan C4.5.

3.3 Pengujian Hasil

Setelah seluruh tahapan pengujian dengan algoritma C4.5 dan Naive Bayes dilakukan, penulis kemudian membandingkan hasil analisis dalam bentuk tabel. Perbandingan difokuskan pada nilai akurasi kedua algoritma tersebut dengan menggunakan Google Colabs dalam memprediksi kelulusan mahasiswa. Rincian hasil akurasi dapat dilihat pada Tabel 8 berikut,

Tabel 8. Perbandingan Hasil Akurasi Algoritma

No	Algoritma	Data Trainning (%)	Data Testing (%)	Akurasi
1	C.45	70	30	100%
		80	20	100%
		90	10	100%
2	Naïve Bayes	70	30	93,88%
		80	20	90,91%
		90	10	94,13%

Berdasarkan Tabel 8, algoritma C4.5 menunjukkan akurasi sempurna sebesar 100% pada seluruh rasio pembagian data. Hal ini mengindikasikan bahwa pohon keputusan yang dibentuk mampu memisahkan kelas secara presisi terhadap dataset yang digunakan. Namun, akurasi sempurna perlu diinterpretasikan secara hati-hati karena berpotensi menandakan adanya overfitting atau feature leakage, terutama jika terdapat atribut yang secara deterministik memisahkan kelas, seperti Total Point. Dengan kata lain, hasil yang terlalu sempurna bisa jadi bukan mencerminkan kemampuan generalisasi model, melainkan ketergantungan pada satu fitur yang sudah merepresentasikan label.

Sebaliknya, algoritma Naive Bayes menghasilkan akurasi yang tinggi namun lebih realistis, yaitu 93.88%, 90.91%, dan 94.13% pada masing-masing rasio. Performa ini menunjukkan bahwa Naive Bayes lebih tahan terhadap dominasi satu fitur dan lebih sesuai untuk data dengan distribusi probabilistik. Walaupun akurasinya sedikit lebih rendah dibanding C4.5, hasil yang diperoleh lebih konsisten dan dapat dianggap lebih representatif terhadap kondisi nyata.

Jika dibandingkan dengan penelitian Mursidil Arif, dkk [23] . yang menggunakan algoritma Naive Bayes pada data mahasiswa Program Studi Informatika Universitas Muhammadiyah Sidoarjo dan memperoleh tingkat akurasi sebesar 68%, hasil penelitian ini menunjukkan peningkatan performa yang signifikan, di mana algoritma Naive Bayes mencapai akurasi hingga 94,13%. Peningkatan tersebut diduga dipengaruhi oleh penggunaan atribut akademik yang lebih lengkap hingga semester 6 serta penambahan atribut non-akademik berupa Total Point (SKEK) yang memiliki korelasi kuat terhadap status kelulusan mahasiswa.

Secara keseluruhan, hasil pengujian ini memperkuat pentingnya evaluasi model dengan pendekatan yang lebih ketat, seperti k-fold cross-validation, serta perlunya audit terhadap fitur yang digunakan agar model tidak hanya menghafal pola, tetapi benar-benar belajar dari data yang representatif. Dengan demikian, penelitian ini menegaskan bahwa meskipun C4.5 tampak unggul dari sisi akurasi, Naive Bayes memberikan gambaran performa yang lebih realistis dan dapat dijadikan alternatif dalam membangun sistem prediksi kelulusan mahasiswa.

4. KESIMPULAN

Penelitian ini bertujuan untuk membangun dan membandingkan model prediksi kelulusan mahasiswa tepat waktu pada Program Studi Informatika Universitas Muhammadiyah Sidoarjo menggunakan algoritma Decision Tree (C4.5) dan

Naïve Bayes. Berdasarkan pengolahan terhadap 161 data mahasiswa angkatan 2022 yang mencakup atribut akademik dan non-akademik melalui tahapan preprocessing, pelabelan, pemodelan, serta evaluasi menggunakan split validation dan 5-Fold Cross Validation, diperoleh hasil bahwa algoritma Decision Tree (C4.5) memiliki performa klasifikasi yang sangat tinggi dengan akurasi mencapai 100% pada skenario pembagian data 70:30, 80:20, dan 90:10 serta rata-rata akurasi sebesar 96,88% pada pengujian cross validation. Sementara itu, algoritma Naïve Bayes juga menunjukkan performa yang baik dengan akurasi tertinggi sebesar 94,13% dan rata-rata cross validation sebesar 93,00%, meskipun masih terdapat beberapa kesalahan klasifikasi dibandingkan dengan C4.5. Hasil analisis menunjukkan bahwa atribut Total Point (SKEK) menjadi faktor paling dominan dalam menentukan status kelulusan mahasiswa, diikuti oleh atribut akademik seperti IPS dan IPK, yang menunjukkan bahwa faktor non-akademik juga berkontribusi terhadap ketepatan waktu kelulusan mahasiswa. Namun demikian, penelitian ini masih memiliki keterbatasan pada jumlah dataset yang relatif terbatas dan hanya mencakup satu angkatan serta satu program studi, sehingga penelitian selanjutnya disarankan untuk memperluas jumlah data, menambahkan variabel lain yang relevan, serta menguji algoritma klasifikasi lainnya agar model yang dihasilkan dapat memiliki kemampuan generalisasi yang lebih baik dan berpotensi dikembangkan sebagai sistem pendukung keputusan dalam pemantauan kelulusan mahasiswa di perguruan tinggi.

REFERENCES

[1] A. M. Limbong and M. Asbari, "Transformasi Standar Nasional dan Akreditasi Pendidikan Tinggi," *JOURNAL OF INFORMATION SYSTEMS AND MANAGEMENT*, vol. 03, no. 01, pp. 101–105, 2024, doi: https://doi.org/10.4444/jisma.v3i1.905.

[2] Tim Penulis UMSIDA, *Pedoman Akademik Universitas Muhammadiyah Sidoarjo Tahun 2023-2024*. Sidoarjo: UMSIDA Press, 2023.

[3] E. F. Wati, L. Indriyani, E. Sunita, and A. D. Kuswanto, "Optimasi Machine Learning dalam Memprediksi Kelulusan Mahasiswa," *JURNAL REKAYASA PERANGKAT LUNAK*, vol. 6, no. 2, pp. 137–142, Nov. 2025, doi: https://doi.org/10.31294/reputasi.v6i2.11067.

[4] J. Khatib Sulaiman, I. Attyyatullatifah, M. Kamayani, U. D. Muhammadiyah HAMKA, K. Kunci Prediksi Kelulusan, and P. Algoritma, "Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa Teknik Informatika dengan Orange Data Mining," *Indonesian Journal of Computer Science*, vol. Vol. 13 No. 2, Apr. 2024, doi: https://doi.org/10.33022/ijcs.v13i2.3796.

[5] D. Wardhani *et al.*, *Pengantar Data Mining PT. MIFANDI MANDIRI DIGITAL*. Deli Serdang: PT. Mifandi Mandiri Digital, 2023.

[6] N. Khasanah, D. Uki, E. Saputri, T. Hidayat, F. Aziz, and U. N. Mandiri, "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma C4.5 dengan RapidMiner: Studi Kasus Data Akademik Perguruan Tinggi XYZ," *Journal Computer Science*, vol. 4, no. 2, pp. 100–107, 2025, doi: https://doi.org/10.31294/ijcs.v4i2.9647.

[7] Z. Fatah and M. Hasanah, "Penerapan Decision Tree Pada Klasifikasi Kelulusan Mahasiswa," *Jurnal Mahasiswa Teknik Informatika*, vol. 4, pp. 225–230, Aug. 2025, doi: https://doi.org/10.35473/jamastika.v4i2.4487.

[8] Rovidatul, Y. Yunus, and G. W. Nurcahyo, "Perbandingan algoritma c4.5 dan naive bayes dalam prediksi kelulusan mahasiswa," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 4, no. 1, pp. 193–199, Apr. 2023, doi: 10.37859/coscitech.v4i1.4755.

[9] Dyah Ardyani Rizqi Azizah Adha, Aulia Noveesa Allanda, Diah Ayu Fatmasari, and Siska Narulita, "Performansi Algoritma C4.5 untuk Prediksi Kelulusan Mahasiswa," *Jurnal Cakrawala Informasi*, vol. 3, no. 2, pp. 9–17, Dec. 2023, doi: 10.54066/jci.v3i2.339.

[10] A. Wahyudi, Kusrini, and F. Wahyu, "PREDIKSI KELULUSAN MAHASISWA TEPAT WAKTU MENGGUNAKAN METODE DECISION TREE DAN NAÏVE BAYES," *Jurnal Permata Indonesia*, vol. 14, no. 2, pp. 132–138, Nov. 2023, doi: 10.59737/jpi.v14i2.276.

[11] K. R. Diska and K. Budayawan, "Sistem Informasi Prediksi Kelulusan Menggunakan Metode Naive Bayes Classifer (Studi Kasus: Prodi Pendidikan Teknik Informatika)," *JURNAL PENDIDIKAN TAMBUSAI*, vol. 7 no 1, pp. 936–943, Feb. 2023, doi: https://doi.org/10.31004/jptam.v7i1.5375.

[12] G. K. Kassymova and O. Ndayizeye, "Prediction model of teacher candidate student graduation status: Decision Tree C4.5, Naive Bayes, and k-NN," *Jurnal Penelitian Hukum dan Pendidikan*, vol. 21, no. 2, pp. 1407–1418, 2022, doi: https://doi.org/10.30863/ekspose.v21i2.3407.

[13] Imam Riadi, Rusydi Umar, and Rio Anggara, "Prediksi Kelulusan Tepat Waktu Berdasarkan Riwayat Akademik Menggunakan Metode Naïve Bayes," *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 4, no. 1, pp. 191–203, Jan. 2024, doi: 10.51454/decode.v4i1.308.

[14] Satrio Junaidi, R. Valicia Anggela, and D. Kariman, "Klasifikasi Metode Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa dengan Algoritma Naïve Bayes, Random Forest, Support Vector Machine (SVM) dan Artificial Neural Nerwork (ANN)," *Journal of Applied Computer Science and Technology*, vol. 5, no. 1, pp. 109–119, Jun. 2024, doi: 10.52158/jacost.v5i1.489.

[15] I. Penulis, N. Kholik Afandi UINSI Samarinda, and I. Artikel, "Literature Review is A Part of Research," *Sulawesi Tenggara Educational Journal*, vol. 1, no. 3, pp. 64–71, Dec. 2021, doi: https://doi.org/10.54297/seduj.v1i3.203.

[16] A. W. Anggraeni, A. S. Fitrani, and A. Eviyanti, "Penerapan Algoritma Support Vector Machine untuk Memprediksi Tingkat Partisipasi Pemilu terhadap Kualitas Pendidikan," *Edumatic: Jurnal Pendidikan Informatika*, vol. 8, no. 1, pp. 21–27, Jun. 2024, doi: 10.29408/edumatic.v8i1.24838.

[17] Z. Setiawan *et al.*, *BUKU AJAR DATA MINING*. Yogyakarta: PT. Sonpedia Publishing Indonesia, 2023. [Online]. Available: www.buku.sonpedia.com

[18] R. Limia Budiarti and N. Kahar, "ANALISIS PREDIKSI KELULUSAN MAHASISWA FAKULTAS ILMU KOMPUTER UNIVERSITAS NURDIN HAMZAH MENGGUNAKAN METODE NAIVE BAYES," *JOURNAL OF INFORMATION TECHNOLOGY*, vol. Vol. 8 No.2, pp. 48–55, Dec. 2024, doi: https://doi.org/10.53564/fortech.v8i2.1312.

[19] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, vol. 27(4s), pp. 4023–4031, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.

[20] J. Allgaier and R. Pryss, "Cross-Validation Visualized: A Narrative Guide to Advanced Methods," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, pp. 1378–1388, Jun. 2024, doi: 10.3390/make6020065.

[21] A. S. Fitrani, M. A. Rosid, C. Taurusta, and I. Fauzia, "Classification Using C4.5 Algorithm in Election Participation Prediction," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, pp. 1–9, Jul. 2020, doi: 10.1088/1757-899X/874/1/012016.

[22] W. A. Firmansyach, U. Hayati, and Y. A. Wijaya, "ANALISA TERJADINYA OVERFITTING DAN UNDERFITTING PADA ALGORITMA NAIVE BAYES DAN DECISION TREE DENGAN TEKNIK CROSS VALIDATION," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, no. 1, pp. 262–269, Feb. 2023, doi: https://doi.org/10.36040/jati.v7i1.6329.

[23] M. M. Arif, H. Setiawan, A. S. Fitrani, F. Sains, and D. Teknologi, "Penggunaan Datamining Untuk Memprediksi Masa Studi Mahasiswa di Universitas Muhammadiyah Sidoarjo Dengan Algoritma Naive Bayes," vol. 4, no. 3, pp. 622–629, Jul. 2023, doi: https://doi.org/10.30645/kesatria.v4i3.210.g209.